

## 12.2 Comparison of Competitor Drugs

In the research, there tends to be a lot of information about individual drugs. However, there is not much information about how drugs perform in comparison to their competitors. There needs to be more drug comparative information so that physicians are better informed about the true benefits of prescribing a more costly medication as compared to a less expensive or generic drug [4]. Big data technology can play a role in making such comparisons easier to accomplish.

## 12.3 Pharmaceutical Research and Development

Big Data can help to streamline the Pharmaceutical Research and development process. As a result, important drugs can be delivered to the market more quickly and the cost of drug development will be reduced.

Big data can enhance the process of identifying appropriate patients to enroll in the clinical trials. First, multiple sources are now available from which to select patients. For example, social media can be incorporated into the selection process and used in addition to physician information. Secondly, the participate selection criteria can include more inclusive factors, such as genetic information. This will enable better targeting of potential trial subjects which will result in more pertinent information, while at the same time shorting trail times and reducing expenses [24].

Trial can be monitored and tracked in real time. Real time trial monitoring can decrease the number of safety and operational issues. The result is the avoidance of potentially costly issues such as adverse events or unnecessary delays [24].

Electronically captured data can improve communication. Information can be shared easily between functions and external parties. All interested individuals can have access to the data at the same time including all departments, external partners, physicians, and contract research organizations (CROs). This will replace the issue of having rigid departmental data silos that hinder interaction [24].

Genomic and proteomic data can be used to speed drug development by providing the capability to better target treatments based upon genetic indicators [17].

## 13 ADMINISTRATIVE COSTS

According to the Institute of Medicine (IOM), the United States spends 361 billion annually on health care administration. This is more than twice our total spending on heart disease and three times our spending on cancer. Also according to the IOM, fully half of these expenditures are unnecessary [9].

One way that providers can save money is to digitize billing processes such as benefit verification, denial management, and claims submission. A benefit verification that is done electronically costs 49 cents per patient. Comparatively, the same process done manually costs 8 dollars. It is estimated that providers could save 9.4 dollars annually by transitioning to electronic processing [21].

One example in which digitized processes are being used to streamline billing processes effectively is at the Phoenix Childrens Hospital in Arizona. They use a tool that automatically converts the clinical notes in the electronic health record (EHR) system to billable diagnostic codes [21].

## 14 FRAUD AND ABUSE

Common types of fraud and abuse include: billing for services that are not rendered, billing for more expensive procedures than were actually delivered, and the performance of unnecessary services.

In the past, the process of identifying misrepresented claims was tedious and time consuming. Big Data analytics makes it possible to easily identify and tag such claims. According to an article by RevCycle Intelligence, when there is repeated misrepresentation of some key fact or event, patterns are created in the data that can be detected by comparing the information to legitimate claims [10]. Anthem Health Insurance, one of the nations biggest insurance payers, uses big data and machine learning algorithms to tag suspicious claims as the claims are being processed. Tagged claims are then sent to clinical coding experts for review. The objective is to identify and address fraudulent claims before they are actually paid [10].

The Center for Medicare and Medicaid Services used predictive data analytics to identify and recover 210.7 million [22] in health care fraud in 2015. They did this by assigning risk scores to claims and providers via algorithms. This enabled the identification of abnormal billing patterns in claim submissions [10].

United Healthcare realized a 2200 percent return on their investment in a Hadoop Big Data platform that was used to identify and tag inaccurate claims using a systemic and repeatable methodology [22].

Other uses of Big Data analytics in fighting fraud and abuse include: identifying links between providers to access whether an identified unethical activity is being practiced by related providers, identification of a hospitals overutilization of services in a short time period, recognizing patients who are receiving health care services from different hospitals in different locations at the same time, and detecting prescriptions that are filled for the same patient in multiple locations at the same time. Big Data analytics can also utilize machine learning algorithms combined with historical information to detect trends in anomalies and suspicious data patterns.

## 15 GENOMICS ANALYTICS

Big data is playing a major role in the field of genomics and precision medicine. These technologies are helping clinicians choose the best treatment plan for individuals based upon their genetic makeup. Combining data from electronic health records (EHRs), clinical trials, and genetic testing gives researchers information to develop more effective treatments for complex diseases such as cancer and diabetes [25], and HIV. Genetic testing that has been made possible by the mapping of the human genome will cut costs and improve survival rates [1].

One area in which genomics can have a dramatic impacts is in pharmaceuticals management. In the United States, 300 million dollars are spent annually on pharmaceuticals. Studies indicate that between 20 to 75 percent of patients are not responsive to prescribed drug therapies. This can often be contributed to incorrect dosing or drug mismatches. However, 50 percent of the time it is because of a molecular mismatch between the patient and the drug. According to Alan Mertz, president of the American Clinical Laboratory Association, an estimated 30 to 110 billion can be saved

by using genetic test to select a drug that is a precise match for the genetics of the patient. By using each patients unique genomic profile, therapy can become more targeted and the instances of inappropriate care will be reduced [1].

For breast cancer patients, genetic testing can identify which 30 percent of women of an overabundance of the HER2 protein. Regular chemotherapy will not help these women, but a drug called Herceptin does. Having this information not only provides doctors with the information they need to prescribe the correct medication, it enables thousands of women avoid needless harsh, expensive chemotherapy treatment. As a result, genetic testing has been shown to reduce the risk of death by 33 percent and the risk of recurrence by 52 percent for breast cancer patients. The resulting savings are estimated to be 24 thousand dollars per patient [1].

Genetic tests can help physicians select the appropriate drug for patients with metastatic colon cancer. According to one estimate, 700 million dollars could be saved annually be obtaining this information before administering treatment [1].

According to a 2006 Brookings/AEI estimate, using genetic tests to determine the appropriate dose of the blood thinner, warfarin, could save the United States 1.1 billion dollars annually. According to a study in June 2010 by the Journal of American College of Cardiology, this test could reduce hospital admissions that are caused by inaccurate dosages by 31 percent [1].

Genomic technology is also good for the United States economy. According to Battelle, a global research organization, human genome sequencing projects generated 796 billion in economic output, 244 billion in personal income and 3.8 million job-years of employment in the United States [1].

The process of gene sequencing continues becomes more efficient and cost effective. It is expected to become a regular part of medical care in the near future [15].

## 16 TELEMEDICINE

Telemedicine is receiving medical treatment and advice remotely, on a computer over the internet with a physician [12]. Telemedicine has been in the market for 40 years, but the with availability of internet connected technology such as smartphones, wireless devices, and video conferences, it is becoming commonplace. It is primarily used for initial diagnosis, remote patient monitoring, and medical education. However, it is also being used for more complicated care such as telesurgery. Telesurgery is a technique in which doctors perform surgery via robots with the assistance of high speed real time data delivery technology [34].

Telemedicine is especially beneficial to patients who live in rural communities who may have to travel long distances to see a doctor or specialist. Telemedicine also gives doctors who are located in multiple locations the ability to discuss and share information. Telemedicine facilitates medical education by giving caregivers the ability to observe and be trained by subject experts no matter where their location.

Telemedicine has the potential to significantly reduce costs by reducing the number of outpatient and hospital visits [38].

## 17 USE CASES

Valence Health has built a data lake that they use as their primary data repository using a MapR Converged Data Platform. The system includes 3000 inbound data feeds and contains 45 different types of data including: lab test results, patient vitals, prescriptions, immunizations, pharmacy benefits, claims information from doctors and hospitals. The system reports dramatically better system performance than legacy system technology. For example, previously, it took 22 hours to process 20 million laboratory records. Now the processing time for the same number of records is 20 minutes. In addition, the new system requires less hardware [22].

The National Institute of Health developed a data lake which combines data sets from separate institutions. Now that all of the data is housed in the same location, analysis is more efficient and can be more easily shared [22].

United Healthcare uses Hadoop to maintain a platform with tools that they use to analyze information generated from claims, prescriptions, provider contracts, plan subscriber, and review information [22].

Novartis, a global healthcare company, uses Hadoop and Apache Spark to build a workflow system that aids in the integration, processing, and analysis of Next Generation Sequencing research as it relates to Genomic Analytics [22].

## 18 CHALLENGES

One of the most compelling challenges is clinicians willingness and ability to change behavior based upon the information provided by the data. Studies have shown that it takes more than a decade of compelling clinical evidence before a new finding becomes common clinical practice. Therefore, we need to do a better job of working with clinicians on finding ways to use the data to provide higher quality care [17].

In health care, the privacy, security, and confidentiality of the patient is paramount [15]. Big data technology has inconsistent security technology. The Health Insurance Portability and Accountability Act (HIPPA) is a federal law that was passed in 1996 that sets a national standards to protect the confidentiality of medical records and personal health information. The HIPAA law is applicable to any component of the information can be used to identify a person. The protections apply to both electronic and non-electronic forms of information [32]. HIPAA regulations make it a federal offense to breach patient security. It is important to work with vendors who understand the importance of security [15]. Liason Technologies is one company that provides solutions to the healthcare and life sciences industry that has experience meeting the HIPAA security requirements [22].

Health care data has inconsistent formatting and definitional issues [17]. There is proliferation of data formats and data representations. There are inconsistent variable definitions. A value may have different meanings for different groups. For example, a cohort definition for an asthmatic patient often differs from one group of clinicians to another [16]. Big data has the challenge of bringing all of this information together.

Another issue is lack of technical experts. The manipulation and extraction of data from often unstructured data sets require special knowledge. There have been some recent changes in tooling that

will make it easier for individualized with less specialized skills to manipulate the data. For example, Big data is starting to use include SQL as a tools for querying and data manipulation. Examples are Microsoft Polybase, Impala, and SQL Hadoop [15].

## 19 CONCLUSION

Big data analytics has huge potential to save the United States billions of dollars in health care costs while drastically improving health outcomes. Vast amounts of information is being captured, stored and combined in ways that offer insights have never before been possible. Innovative Big data tools are reducing medical waste, decreasing medical errors, fighting fraud, and keeping people healthier. Value based reimbursement solutions have the potential to revolutionize the health delivery system in the United States by motivating providers to find ways to deliver the best possible medical care with the most economical use of resources. The development of most of these tools is only in the preliminary stage. Therefore, we are only beginning to realize some of the potential benefits. Big data really does have the potential to bend the cost curve. Big data in health care is here to stay.

## ACKNOWLEDGMENTS

The author would like to thank Dr. Gregor von Laszewski and the teaching assistants in the Data Science department at Indiana University for their support and suggestions to write this paper.

## REFERENCES

- [1] American Clinical Library Association. 2011. Genetic Testing Can Help the United States Cut Costs and Improve Care. Web page as article. (July 2011). <https://www.prnewswire.com/news-releases/genetic-testing-can-help-the-us-cut-costs-and-improve-health-care-126105103.html>
- [2] American Economic Association. 2017. Would Price Transparency Lower Health-care Costs. Web page as article. (Feb. 2017). <https://www.aeaweb.org/research/health-care-price-transparency>
- [3] Tanya Bentley. 2018. Waste in the US Health System - A conceptual framework. *The Milbank Quarterly* 86 (Dec. 2018), 629–659. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2690367/>
- [4] Business Insider. 2016. Why the Price of Prescription drugs in the US is Out of Control. Web page as article. (Aug. 2016). <http://www.businessinsider.com/why-the-us-pays-more-for-prescription-drugs-2016-8>
- [5] Christian Ofori Boateng. 2016. Top 3 Ways Big Data Helps Decrease the Cost of Health Care. Web page as Article. (Nov. 2016). <https://go.christiansteven.com/top-3-ways-big-data-helps-decrease-the-cost-of-health-care>
- [6] CIO. 2015. How Big Data can save 400 billion in healthcare costs. Web page as Article. (Oct. 2015). <https://www.cio.com/article/2993986/big-data-how-big-data-can-help-save-400-billion-in-healthcare-costs.html>
- [7] CMS Centers for Medicare and Medicaid Services. 2017. Accountable Care Organizations. Web page. (Nov. 2017). <https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/ACO/>
- [8] Consumer Reports. 2014. Why is Healthcare so Expensive. Web page. (Sept. 2014). <https://www.consumerreports.org/cro/magazine/2014/11/it-is-time-to-get-mad-about-the-outrageous-cost-of-health-care/index.htm>
- [9] DataFloq. 2016. Five ways Big Data in reducing healthcare costs. Web page as article. (March 2016). <https://datafloq.com/read/5-ways-big-data-reducing-healthcare-costs/89>
- [10] Datameer. 2017. The Role of Big Data in Preventing Healthcare Fraud, Waste, and Abuse. Web page as article. (Sept. 2017). <https://www.datameer.com/company/datameer-blog/role-big-data-preventing-healthcare-fraud-waste-abuse/>
- [11] Digitalist. 2016. Can Big Data Analytics Save Billions in Healthcare Costs. Web page as Article. (Feb. 2016). <http://www.digitalistmag.com/resource-optimization/2016/02/29/big-data-analytics-save-billions-in-healthcare-costs-04037289>
- [12] Forbes. 2015. How Big Data is changing Healthcare. Web page as Article. (April 2015). <https://www.forbes.com/sites/bernardmarr/2015/04/21/how-big-data-is-changing-healthcare/#427274d12873>
- [13] Forbes. 2016. How an Electronic Health Record can Save Time, Money and Lives. Web page. (Dec. 2016). <https://www.forbes.com/sites/robertpearl/2016/12/01/how-an-electronic-health-record-can-save-time-money-and-lives/2/#4445b8275f57>
- [14] Harvard Business Review. 2014. How Cities are Using Analytics to Improve Public Health. Web page as article. (Sept. 2014). <https://hbr.org/2014/09/how-cities-are-using-analytics-to-improve-public-health>
- [15] Health Catalyst. 2017. Big Data in Healthcare Made Simple: Where it Stands Today and Where its Going. Web page as Article. (Oct. 2017). <https://www.healthcatalyst.com/big-data-in-healthcare-made-simple>
- [16] Health Catalyst. 2017. Five Reasons Healthcare Data is so Complex. Web page as article. (Nov. 2017). <https://www.healthcatalyst.com/>
- [17] Health Catalyst. 2017. Hadoop in Healthcare A no nonsense Q and A. Web page as article. (Nov. 2017). <https://www.healthcatalyst.com/Hadoop-in-healthcare>
- [18] Kayyali, Basel, Knott, David, Kuiken, Steve Van. 2013. McKinsey on Healthcare. Web page as Article. (April 2013). <http://healthcare.mckinsey.com/big-data-revolution-us-healthcare>
- [19] KQED Science. 2015. How San Diego is Using Big Data to Improve Public Health. Web page as article. (Aug. 2015). <https://ww2.kqed.org/futureofyou/2015/08/19/how-san-diego-is-using-big-data-to-improve-public-health/>
- [20] Liaison. 2017. Value Based Healthcare - The patient is the Center but Data is the Key. Web page as blog. (June 2017). <https://www.liaison.com/blog/2017/06/22/value-based-healthcare-patient-center-data-key/>
- [21] Managed Healthcare Executive. 2017. Five ways to reduce healthcare administrative costs. Web page as article. (April 2017). <http://managedhealthcareexecutive.modernmedicine.com/managed-healthcare-executive/news/five-ways-reduce-healthcare-administrative-costs>
- [22] McDonald, Carol. 2016. How Big Data is Reducing Costs and Improving Outcomes in Healthcare. Web page as Article. (June 2016). <https://mapr.com/blog/reduce-costs-and-improve-health-care-with-big-data/>
- [23] McKinsey and Company. 2013. The Trillion Dollar Prize. Web page as article. (Feb. 2013). <https://healthcare.mckinsey.com/sites/default/files/the-trillion-dollar-prize.pdf>
- [24] McKinsey and Company. 2017. How Big Data can Revolutionize pharmaceutical R and D. Web page as article. (Nov. 2017). <https://www.mckinsey.com/industries/pharmaceuticals-and-medical-products/our-insights/how-big-data-can-revolutionize-pharmaceutical-r-and-d>
- [25] Pacient. 2017. How Big Data Can Improve Health Care. Web page as article. (Nov. 2017). <https://pacient.care/decks/privacy-technology/health-technology/how-big-data-can-improve-healthcare>
- [26] PBSO News Hour. 2012. Health Costs: How the US Compares with Other Countries. Web page as Article. (Oct. 2012). <https://www.pbs.org/newshour/health/health-costs-how-the-us-compares-with-other-countries>
- [27] Practice Fusion. 2017. EHR Adoption Rates 20 Must see stats. Web page as Article. (March 2017). <https://www.practicefusion.com/blog/ehr-adoption-rates/20-must-see-stats>
- [28] OECD Publishing. 2017. *Health at a Glance 2017*. OECD, Paris. [http://dx.doi.org/10.1787/health\\_glance-2017-en](http://dx.doi.org/10.1787/health_glance-2017-en)
- [29] Raghupathi Viju Raghupathi, Wullianallur. 2014. Big Data Analytics in Healthcare Promise and Potential. *Springer Health Information Science and Systems* 2 (Feb. 2014), 2–3. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4341817/>
- [30] Rock Health. 2017. The Future of Personalized Healthcare: Predictive Analytics. Web page. (Nov. 2017). <https://rockhealth.com/reports/predictive-analytics/>
- [31] Search Technologies. 2017. Using Big Data Predictive Analytics to Improve Healthcare. Web page as article. (Sept. 2017). <https://www.searchtechnologies.com/blog/predictive-analytics-in-healthcare>
- [32] Stephen B Thacker. 2003. HIPAA Privacy Rule and Public Health. *CDC* 52 (April 2003), 1–12. <https://www.cdc.gov/mmwr/preview/mmwrhtml/m2e411a1.htm>
- [33] The Common Wealth Fund. 2017. The Affordable Care Act Payment and Delivery System reforms: A progress report. Web page as article. (Feb. 2017). <http://www.commonwealthfund.org/publications/issue-briefs/2015/may/aca-payment-and-delivery-system-reforms-at-5-years>
- [34] The datapine blog. 2017. Nine examples of Big Data Analytics in Healthcare that Can Save People. Web page. (May 2017). <https://www.datapine.com/blog/big-data-examples-in-healthcare/>
- [35] The Wall Street Journal. 2017. How to Research Medical Prices. Web page as article. (Nov. 2017). <http://guides.wsj.com/health/health-costs/how-to-research-health-care-prices/>
- [36] US Department of Health and Human Resources. 2017. EHR Basics. Web page. (Sept. 2017). <https://www.healthit.gov/providers-professionals/learn-ehr-basics>
- [37] Wikipedia. 2017. Evidence Based Medicine. Web page. (Nov. 2017). [https://en.wikipedia.org/wiki/Evidence-based\\_medicine](https://en.wikipedia.org/wiki/Evidence-based_medicine)
- [38] Wikipedia. 2017. Telemedicine. Web page. (Nov. 2017). <https://en.wikipedia.org/wiki/Telemedicine>
- [39] Zane Benefits. 2017. FAQ - How much does Individual Insurance cost. Web page. (Nov. 2017). <https://www.zanebenefits.com/blog/bid/97380/faq-how-much-does-individual-health-insurance-cost>

```
bibtext report
```

```
=====
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib.bib
```

```
bibtext _ label error
```

```
=====
bibtext space label error
```

```
=====
bibtext comma label error
```

```
=====
latex report
```

```
=====
[2017-12-11 13.30.51] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflatex)
Typesetting of "report.tex" completed in 1.0s.
./README.yml
53:20      error      no new line character at the end of file  (new-line-at-end-of-file)
```

```
=====
Compliance Report
```

```
=====
name: Judy Phillips
hid: 332
paper1: Oct 31 2017 100%
paper2: 100%
project: 100%
```

```
yamlcheck
```

wordcount

---

10  
wc 332 project 10 8954 report.tex  
wc 332 project 10 9324 report.pdf  
wc 332 project 10 1543 report.bib

find "

---

passed: True

find footnote

---

passed: True

find input{format/i523}

---

4: \input{format/i523}

passed: True

find input{format/final}

---

passed: False

floats

---

figures 0  
tables 0  
includegraphics 0  
labels 0  
refs 0  
floats 0

True : ref check passed: (refs >= figures + tables)  
True : label check passed: (refs >= figures + tables)  
True : include graphics passed: (figures >= includegraphics)  
True : check if all figures are referred to: (refs >= labels)

Label/ref check  
passed: True

When using figures use columnwidth  
[width=1.0\columnwidth]  
do not change the number to a smaller fraction

find textwidth

---

passed: True

below\_check

---

bibtex

---

label errors

bibtex errors

This is BibTeX, Version 0.99d (TeX Live 2016)  
The top-level auxiliary file: report.aux  
The style file: ACM-Reference-Format.bst  
Database file #1: report.bib.bib

bibtex\_empty\_fields

---

entries in general should not be empty in bibtex

find ""

---

passed: True

ascii

---

```
=====
The following tests are optional
=====
```

Tip: newlines can often be replaced just by an empty line

```
find newline
-----
```

```
passed: True
cites should have a space before \cite{} but not before the {
```

```
find cite {
-----
```

```
passed: True
```

# Using Machine Learning Classification of Opioid Addiction for Big Data Health Analytics

Sean M. Shiverick

Indiana University Bloomington

smshiver@indiana.edu

## ABSTRACT

Classification of opioid misuse and abuse can identify important features relevant for predicting drug addiction and overdose death. Machine learning procedures were applied to data from a large National Survey of Drug Use and Health (NSDUH-2015) to classify individuals for illicit opioid use according to demographic characteristics and mental health attributes (e.g., depression). Classification models of opioid addiction can be extended for big data health analytics to include high-dimensional datasets, data collected over previous years, or expanded to the larger population of patients taking prescription opioid medication. The results seek to raise awareness of risk factors related to opioid addiction among patients and medication prescribers, and help decrease the risk of opioid overdose death.

## KEYWORDS

Big Data, Health Analytics, Classifier Algorithms, Opioid Addiction, i523, hid335

## 1 INTRODUCTION

Big Data offers tremendous potential to fuel innovation and transform society. Can this momentum be harnessed to address a serious health crisis such as the opioid overdose epidemic? [7] Health informatics is generating huge amounts of data at a rapid pace, from electronic medical records (EMRs), clinical research data, to population-level public health data [5]. This project considers health analytics from two levels, the research questions being addressed and the data used to answer them. The question of interest in this project is whether opioid dependency and addiction can be predicted from demographic attributes and psychological characteristics. Survey research provides data on a wide range of issues that people may be reluctant to disclose, including mental health disorders, personal medical health concerns, prescription medications, and illicit drug use. Responses to surveys may be biased to some degree, but measures of confidentiality and anonymity help to assure more accurate disclosures. The goal of this project is to use machine learning procedures to classify individuals susceptible to opioid abuse and dependence. Understanding the features that contribute to opioid addiction can identify underlying risk factors and increase awareness of potential opioid abuse for patients and health care providers. The results could be extended to big data from previous years of the opioid crisis and to the larger population of patients taking prescription opioid mediation. Different machine learning classification methods are discussed.

## 1.1 Opioid Overdose Epidemic

The abuse of prescription opioid medication in the U.S. has become a major health crisis of epidemic proportions [26]. Over 2 million Americans were dependent or abused prescription opioids such as oxycodone or hydrocodone in 2014[3]. Overdose deaths from prescription opioids have quadrupled since 1999, resulting in more than 180,000 deaths between 1999 to 2015 [11]. Drug overdose deaths increased significantly for males and females, between 25-44 years, ages 55 and older, for Non-Hispanic Whites and Blacks, in the Northeast, Midwest, and Southern regions of the U.S. [7]. Mobile health applications can monitor patient medication consumption and provide an early warning system for potential abuse, detecting sudden changes in medications, higher dosages, or rapid escalation of a prescribed dosage [25]. Reliable information about medication dosages can be difficult to obtain based on self-reports. Individuals dependent or addicted to prescription opioids may obtain synthetic opioids such as fentanyl or illicit drugs such as heroin. Because the dosage levels and potency of illicit opioids are largely unknown, there is greater risk of drug overdose death. The sharp increase in overdose deaths due to synthetic opioids (other than methadone) has coincided with the increased availability of illicitly manufactured fentanyl, which is indistinguishable from prescription fentanyl. The findings indicate the opioid overdose epidemic is getting worse, and requires urgent action to prevent opioid dependence, abuse and overdose death. The target group for this project is individuals who reported misusing or abusing prescribed opioid medication who also used heroin, shown in Figure 1.

## 1.2 Machine Learning Approaches

Machine learning is a set of procedures and automated processes for extracting knowledge from data. The two main branches of machine learning are supervised learning and unsupervised learning. Supervised learning problems involve prediction about a specific target variable or outcome of interest. If a given dataset has no target outcome, unsupervised learning methods can be used to discover underlying structure in unlabeled data. The goal of this project is to classify opioid addiction and focuses on supervised learning. Supervised learning is used to predict a certain outcome from a given input, when examples of input/output pairs are available [10]. A machine learning model is constructed from the training set of input-output pairs, to predict new test data not previously seen by the model. The two major approaches to supervised learning problems are regression and classification. When the target variable to be predicted is continuous, or there is continuity between the outcome (e.g., home values, or income), a regression model is used to test the set of features that predict the target variable. If the target is a class label, set of categorical or binary outcomes (e.g., spam or ham, benign or malignant), then classification is used to

predict which class or category label that new instances will be assigned to.

### 1.3 Classification Algorithms

Comparing the performance of different learning algorithms can be helpful for selecting the best model for a given problem [14]. One of the simplest classification algorithms is K-Nearest-Neighbors (KNN) which takes a set of data points and classifies a new data point based on the distance (e.g., Euclidean, by default) to its nearest neighbors. The main parameter for KNN is the number of neighbors, and k of 3 or 5 neighbors works well. The advantage of the KNN classifier is that it provides a solution that is easy to understand. A limitation of KNN is that it does not perform well with a large number of features (100 or more) or sparse datasets. Several different classification algorithms are considered below.

**1.3.1 Logistic Regression Classifier.** Logistic regression is a commonly used linear model for classification problems. The decision boundary for the logistic regression classifier is a linear function of the input; a binary classifier separates two classes using along a line, plane, or hyperplane. Linear classification models differ in terms of (1) how they measure how well a particular combination of coefficients and intercept fit the training data, and (2) the type of regularization used [10]. The main parameter for linear classification models is the regularization parameter C. High values of C correspond to less regularization and the model will fit the training set as best as possible, stressing the importance of each individual data point to be classified correctly. By contrast, with low values of C, the model puts more emphasis on finding coefficient vectors (i.e., weights) that are close to zero, trying to adjust to the majority of data points. In addition, the penalty parameter influences the coefficient values of the linear model. The L2 penalty (Ridge) uses all available features, but pushes the coefficient values toward zero. The L1 penalty (Lasso) sets the coefficient values for most features to zero, and uses only a subset of features for improved interpretability. This analysis used a logistic regression classifier to predict Heroin use from demographic attributes, mental health, prescription opioids, medication use, misuse, and illicit drug use.

**1.3.2 Tree Based Models.** Decision tree models are widely used for classification and regression. Tree models “learn” a hierarchy of if-else questions that are represented in the form of a decision tree. Building decision trees proceeds from a root node as the starting point and continues through a series of decisions or choices. Each node in the tree either represents either a question or a terminal node (i.e., leaf) that contains the outcome. Applied to a binary classification task, the decision tree algorithm *learns* the sequence of if-else questions that arrives at the outcome most quickly. For data with continuous features, the decisions are expressed in the form of, “Is feature x larger than value y?” [10] In constructing the tree, the algorithm searches through all possible decisions or tests, and find a solution that is most informative about the target outcome. A decision tree classifier is used for binary or categorical targets, and decision tree regression is used for continuous target outcomes. The recursive branching process of tree based models yields a binary tree of decisions, with each node representing a test that considers a single feature. This process of recursive partitioning

is repeated until each leaf in the decision tree contains only a single target. Prediction for a new data point proceeds by checking which region of the partition the point falls in, and predicting the majority in that feature space. The main advantage of tree based models is that they require little adjustment and are easy to interpret. A drawback is that they can lead to very complex models that are highly overfit to the training data. A common strategy to prevent overfitting is *pre-pruning*, which stops tree construction early by limiting the maximum depth of the tree, or the maximum number of leaves. One can also set the minimum number of points in a node required for splitting. Another approach is to build the tree and then remove or collapse nodes with little information, which is called *post-pruning*. Decision trees work well with features measured on very different scales, or with data that has a mix of binary and continuous features.

**1.3.3 Random Forests Classifier.** A random forest is a collection of decision trees that are slightly different from the others, which each overfits the data in different ways. The idea behind random forests is that overfitting can be reduced by building many trees and averaging their results. This approach retains the predictive power of trees while reducing overfitting. Randomness is introduced into the tree building process in two ways: (a) selecting a bootstrap sample of the data, and (b) selecting features in each node branch [10, 14]. In building the random forest, we first decide how many trees to build (e.g., 10 or 100), and the algorithm makes different random choices so that each tree is distinct. The bootstrapping method repeatedly draws random samples of size n from the dataset (with replacement). The decision trees are built on these random samples that are the same size as the original data, with some points missing and some data points repeated. The algorithm also selects a random subset of p features, repeated separately each node in the tree, so that each decision at the node branch is made using a different subset of features. These two processes help ensure that all of the decision trees in the random forest are different. The important parameters for the random forests algorithm are the number of sampled data points and the maximum number of features; the algorithm could look at all of the features in the dataset or a limited number. A high value for *maximum-features* will produce trees in the random forest that are very similar and will fit the data easily based on the most distinctive features, whereas a low value will produce trees that are very different from each other, and reduces overfitting. Random forests is of the most widely used ML algorithms that works well without very much parameter tuning or scaling of data. A limitation of this approach is that Random forests do not perform well with very high-dimensional, data that is sparse data, such as text data.

### 1.4 Project Goals

The general idea of the project is that prescription opioid dependency and addiction will in many cases lead to the use of illicit opioids such as heroin or fentanyl. According to this reasoning, it was hypothesized that individuals who report using heroin may also be susceptible to misusing or abusing prescription opioid medications. The goal of the study was to identify the set of features important for predicting opioid addiction. The data used in the project is from the National Survey on Drug Use and Health from 2015 (NSUH-

2015) [1], which is the most recent year available. The NSDUH-2015 is a comprehensive survey that covers all aspects of substance use, misuse, dependency, and abuse, including questions related to both prescription medications (opioids, tranquilizers, sedatives) and illicit drugs (e.g., heroin, cocaine, methamphetamine), drug dependency, addiction, and treatment, demographic measures of education and employment, physical health, depression, and mental health treatment. Several classification models were constructed to classify heroin use in the sample by demographics attributes and mental health characteristics (e.g., adult depression). This method addresses the following issues related to opioid dependency and addiction: (i) Identify factors related to illicit opioid use, (ii) Identify factors related to prescription opioid misuse and abuse, and (iii) Examine the relationship between prescription opioid misuse, abuse and heroin use.

## 2 METHOD

The project workflow pipeline is outlined in a readme markdown file in the project folder [22]. The steps included in the workflow were (1) Download and Extract the Data, (2) Data Cleaning and Preparation, (3) Exploratory Data Analysis, (4) Data Visualization, (5) Analysis of Classification Models for Heroin Use, and (6) Analysis of Classification Models for Prescription Opioid Pain Reliever Misuse.

### 2.1 Data

Data from the 2015 NSUH was downloaded from the Substance Abuse and Mental Health Data Archive (SAMHDA) [1] URL using the get-data.py function written to unzip the data files, extract the data as a Pandas data frame, and write the file to CSV file [4]. The dataset consists of 57,146 observations with 2,666 features representing individual-level responses from a survey of the U.S. population. According to the NSDUH codebook, sampling was weighted across states by population size for a representative distribution selected from 6,000 area segments. The sample design used five state sample size groups drawing more heavily from the eight states with the largest population (e.g., CA, FL, IL, MI, NY, OH, PA, TX) which together account for 48 percent of total U.S. population aged 12 or older. All identifying information was collapsed (e.g., age categories) and state identifiers were removed from the public use file to ensure confidentiality. The NSDUH public-use files do not include geographic location, or demographic variables related to ethnicity or immigration status. The weighted survey screening response rate was 81.94 percent and the weighted interview response rate was 71.2 percent.

### 2.2 Data Cleaning and Preparation

**2.2.1 Data Cleaning.** All steps of this analysis was completed in a python interactive notebook [16] based following examples from *Python for Data Analysis* [9]. After saving the NSDUH-2015 as a data frame object, the dataset was subset by columns to include demographic characteristics (e.g., age category, sex, marital status, education, employment status, and category of metropolitan area), measures of physical health (e.g., overall health, STDs, Hepatitis, HIV, Cancer, hospitalization), mental health (e.g., Adult Depression, Emotional Distress, Suicidal Thoughts, Plans), Suicide Attempts,

Pain Reliever Medication Use, Misuse, and Abuse (over past year, past month), Prescription Opioid Medications Taken in Past year (e.g., Hydrocodone, Oxycodone, Tramadol, Morphine, Fentanyl, Oxymorphone, Demerol, Hydromorphone), Heroin Use, Abuse (over past year, past month), Tranquilizer Use, Sedative Use, Cocaine Use, Amphetamine and Methamphetamine Use, Hallucinogen Use, Drug Treatment (e.g., Inpatient, Outpatient, Hospital, Mental Health Clinic, ER, Drug Treatment Status), and Mental Health Treatment History. A codebook was created to provide a complete list of variables included with summaries of response categories [19]. The following steps were taken to detect and remove inconsistencies in the data [13]:

- (1) Remove missing values (i.e., NaN)
- (2) Recode blanks, non-responses, or legitimate skips (e.g., 99, 991, 993) to zero
- (3) Recode dichotomous responses (e.g., Yes=1 / No=2) so that No=0
- (4) Recode categorical variables to be consistent with amount or degree (e.g., 1=low, 2=med, 3=high)
- (5) Rename selected variables for better description (e.g., Adult Major Depressive Episode Lifetime changed from AMDELT to DEPMELT)

**2.2.2 Aggregated Variables.** Because the majority of features were represented as dichotomous Yes / No variables, related features were summed to create aggregated variables. For example, overall health, STD, Hepatitis, HIV, Cancer, and hospitalization were aggregated to create a single health measure. The health measure was recoded so that higher scores indicated better health. Questions related to depression, emotional distress, and suicidal thoughts were summed to create a single variable for mental health (MENTHLTH) with scores ranging from 0 to 9. Responses to pain reliever medication use, misuse, abuse, or dependency, were aggregated to create a single variable of pain reliever misuse or abuse (PRLMISAB). All prescription painkiller medications used in the past year were summed. Similarly, all related responses were summed to create single variables for Tranquilizers, Sedatives, Cocaine, Amphetamines, Hallucinogens, Drug Treatment, and Mental Health Treatment. The target outcome of interest for classification, lifetime heroin use (i.e., “Have you ever used heroin before, at any time?”) is a dichotomous variables. The demographic characteristics and aggregated variables were subset and saved to a new data frame consisting of 2 features and 57,146 observations, which was exported to CSV file.

## 3 RESULTS

### 3.1 Exploratory Data Analysis

Of the total sample of N=57,146 respondents, 26,736 were male and 30,410 female; 6,343 individuals reported misusing pain medication at some point (570 males, 386 females), but only 956 respondents had used heroin (570 males, 386 females). Table 1 shows the raw counts of individual substance use by age group (with the sample size for each age group), listing the ten most commonly used opioid pain medications, self-reported misuse of prescription opioid pain relievers (i.e., PRL Misuse Ever), use of prescription Tranquilizers, Sedatives, and Methadone. In addition, self-reported use of illicit drugs such as heroin, cocaine, amphetamines, methamphetamine,

Hallucinogens, including LSD and Ecstasy (MDMA). This summary table shows that substance use seems to be highest for individuals between the ages of 18 to 25 and from 35 to 49 years. Of the prescription relievers, Hydrocodone use (e.g., Vicodan) was almost double the rate of Oxycodone use (e.g., Oxycodone) for each age group, and was significantly higher than any other prescription opioid medication. Use of prescription Fentanyl and Demerol, two powerful opioids, and synthetic morphines such as Oxymorphone and Hydromorphone, was very low. The rate of prescription Tranquilizer use was several orders of magnitude higher than Sedative use or Methadone use. Compared to other illicit drugs such as Cocaine, Amphetamines, Hallucinogens, heroin use was not very common in this sample. The highest rates of heroin use were seen between the ages of 18 to 49, and was lowest for respondents in the youngest age group 12 to 17, and individuals over 50.

[Table 1 about here.]

Table 2 shows the frequency of individuals reporting that they had experienced mental health issues such as depression, suicidal thoughts, whether they had received mental health treatment, received treatment from a private therapist, or believed that they needed drug treatment, but had not sought treatment, across each age category. Frequency of depression was not included for respondents between 12 to 17 years, because the survey measure was for adult depression.

[Table 2 about here.]

Figure 1 shows the proportion of individuals who reported misusing prescription opioid pain relievers and who reported using heroin. The left column of the Figure 1 shows the majority of respondents (89 percent) stated they had never misused prescription opioid pain medication or used heroin, although 10 percent reported misusing opioid pain medication at some point. The right panel of Figure 1 shows that, of those individuals who reported using heroin, the proportion who also reported misusing opioid pain medication was almost twice as large as the proportion of those who only used heroin. This is consistent with the hypothesis that misuse of prescription opioids is linked with heroin use for some individuals.

[Figure 1 about here.]

Figure 2 shows the aggregated measure of Opioid Pain Reliever misuse and abuse plotted against the aggregated measure of Heroin use (which includes misuse, abuse, lifetime use, past year use, 30 day use), with weighted regression lines grouped by size of City/Metropolitan region (from none to large). The largest proportion of the sample who report prescription opioid misuse, abuse, and heroin use is represented by observations from large metropolitan areas (red circles) with large population size. However, a small number of observations from rural or small metropolitan regions (blue and green circles) showed very high rates of prescription opioid misuse and abuse. Regression lines (i.e., line of best fit) shown are weighted by the City/Metro region attribute, with a steeper slope shown for smaller metropolitan regions than large metropolitan regions. The difference in slope may be due to the influence of the small number of outliers who had high degrees of prescription opioid misuse, and heroin use. The plot also shows a clear divide on the y-axis, which separates the sample according to high and low or no prescription

opioid misuse, although the continuum of heroin use from no, low, to high is distributed fairly evenly along the x-axis.

[Figure 2 about here.]

Figure 3 shows the pairplots of demographic features including mental health (higher scores equal to more depression), Prescription Opioid Pain Reliever (PRL) Medication (aggregated), Heroin Use (aggregated measure), and Size of City/Metropolitan region. The top row shows that the majority of the sample reported no mental health concerns, whereas a small proportion of the sample reported depression, emotional distress, or suicidal thoughts. Only few people self-described as high in depression reported low Prescription Opioid PRL misuse and abuse. The plot also reveals that prescription opioid misuse and heroin use were distributed approximately evenly for individuals reporting either low, moderate, or high levels of depression, which suggests that depression was not a factor in predicting opioid misuse. The second row shows a small number of individuals from rural areas or small cities who reported very high levels of prescription opioid misuse, although the majority of respondents misusing or abusing prescription opioid were from large metropolitan areas. As described above, the majority of respondents (about 90 percent of the sample) reported they had never misused prescription opioids. In the second row and third and fourth columns, a natural break is seen between individuals who reported high levels of prescription opioid misuse and abuse and those who reported very low or no opioid misuse. A very small proportion of the entire sample reported both misusing and abusing prescription opioids and using heroin, but this is a group of interest. The last column of the second row shows the individuals reporting high levels of opioid misuse and abuse were distributed evenly across city/metropolitan areas of different sizes, with only slightly higher numbers for small cities or rural areas. As stated above, only few participants reported using heroin, and of these, the majority were from large metropolitan areas. Finally, the sample seems to have slightly higher proportions from small and large metropolitan areas, which is likely due to weighted sampling, which drew more from heavily populated regions.

[Figure 3 about here.]

### 3.2 Classifier Models of Heroin Use

This analysis classified individuals according to whether they had ever used heroin (i.e., "Heroin Use Ever"). All classifier models were constructed using SciKit Learn [10] using an interactive python jupyter notebook [17]. The features of interest were demographic characteristics, health, mental health (adultdepression), prescription opioid misuse and abuse (PRLMISEVR, PRLMISAB, PRLANY), prescription tranquilizers use and sedatives use (TRQLZRS, SE-DATVS), use of illicit drugs (COCAINE, AMPHETMN), drug treatment (TRTMNT), and mental health treatment (MHTRTMT). The target variable was Heroin Use (HEROINEVR). Next, the dataset was split into the training set and test sets using the train-test-split() function in sklearn. Model accuracy for the training set and test set are reported, with different parameter values, and features importance.

**3.2.1 Logistic Regression Classifier.** Logistic Regression Classification is based on a linear equation that calculates the relative

weight of each feature for a categorical target or binary outcome (yes / no) [14]. The logistic regression classifier was fit to the training data in Scikit-Learn, and the model was validated on the test data. By default, the model applies L2 penalty (Ridge). The training set accuracy was 0.983 and the test set accuracy was 0.984. The parameter ‘C’ determines the strength of regularization, with higher values of C providing greater regularization. The L1 penalty (Lasso) limits the values of most coefficients to zero, creating a more interpretable model that uses only a few features. Figure 4 plots the coefficients of logistic regression classifier for heroin use with the L1 Penalty (Lasso) under different values of parameter C. The default setting, C=1.0, provides good performance for train and test sets, but the model is very likely underfitting the test data. Using a higher value of C fits a more flexible model and generally gives improved accuracy for both training and tests sets. Using a value of C=100 yielded training set accuracy of 0.98 and test set accuracy of 0.98. Figure 4 shows that the features coefficient values did not change much according to the values of parameter C, and the accuracy values were approximately the same for all values of C. Examination of the coefficients from the logistic regression classifier revealed the three features which were most closely associated with Heroin use were: Prescription Opioid Pain Reliever (PRL) Misuse ever (as predicted), Cocaine Use, and Amphetamine use, respectively.

[Figure 4 about here.]

**3.2.2 Decision Tree Classifier.** The following analysis used the *Decision Tree Classifier* package in Scikit-Learn, which only does pre-pruning. First, the decision model was build using the default setting of a fully developed tree until all leaves are pure. The random state’ features is fixed to break ties internally. Accuracy on the training set was 0.99 and test set accuracy was 0.974. Without restricting their depth, decision trees can become complex; unpruned trees are prone to overfitting and do not generalize well to new data. Limiting the depth of tree decreases overfitting, which results in lower training set accuracy, but improved performance on the test set. Next, pre-pruning was applied, with a maximum depth of 4, which means the algorithm split on four consecutive questions. Training set accuracy of the pruned tree was 0.985 and test set accuracy was 0.984. Even with a depth of 4, the tree can become a bit complex. Figure 5 shows a partial view of the decision tree classifier of heroin use (the entire tree was too wide to include as a legible Figure), and the full tree image is available in the notebook BDA-Analytics-Classifier-Heroin.ipynb [17]. The decision tree shows the top features that the algorithm split on to classify heroin use. One way to interpret a decision tree it by following the sample numbers represented at the test split for each node. The classifier algorithm selected Cocaine Use (aggregated score) as the root node of the decision tree. The branch to the left side of the tree represents samples with a score equal to or less than 1.5 (n=40956), whereas the branch to the right represents samples with a Cocaine Use score greater than 1.5 (n=1903). The second split on the right occurs for Any Prescription Opioid Pain Reliever Use (PRLANY), with n=1443 having a score less than or equal to 3.5, and n=460 respondents with a PRL score greater than 3.5. In other words, of those respondents who reported relatively high Cocaine use, a small portion also reported relatively high Prescription Opioid PRL

use. Instead of looking at the whole tree, features importance is a common summary function that rates how important each feature is for the classification decisions made in the algorithm. Each feature is assigned an importance value between 0 and 1; with a value of 1 indicating the feature perfectly predicts the target and a value of 0 meaning that the feature was not used at all. Feature importance values also always sum to 1. A feature may have a low feature importance value because another feature encodes the same information. The top two important features for classifying Heroin Use were Cocaine Use and Any Prescription Opioid PRL Use, with smaller importance given to Opioid PRL Misuse Ever and Prescription Opioid PRL Misuse and Abuse.

[Figure 5 about here.]

**3.2.3 Random Forests Classifier.** Random forests is an ensemble approach that builds many trees and averages their results to reduce overfitting. The model was build using the *Random Forest Classifier* package in Scikit-Learn. The parameters of interest for building random forests are: (a) the number of trees (n-estimators), (b) the number of data points for bootstrap sampling (n-samples), and (c) the maximum number of features considered at each node (max-features). The max-features parameter determines how random each tree is, with smaller values of max-features resulting in trees in the random forest that are very different from each other. This analysis applied a random forest consisting of 100 trees to classify Heroin Use, and the random state was set to zero. The training set accuracy was 0.999 and the test set accuracy was 0.984. Often the default settings for random forests work well, but we can apply pre-pruning as with a single tree, or adjust the maximum number of features. Feature importance for random forests is computed by aggregating the feature importance over trees in the random forest, and random forests gives non-zero importance to more features than a single tree. Typically random forests provide a more reliable measure of feature importance than the feature importance for a single tree. Figure 6 shows the feature importance of the random forests classifier for heroin use with 100 trees. Similar to the single tree, the random forest selected Cocaine Use as the most informative feature in the model, followed by Any PRL Use, which is an aggregated measure of prescription opioid medication use. Following after that, several features were tied for third place of importance, namely Education Level, Overall Health, Age Category, and Pain Reliever Misuse and Abuse. Random forests provides much of the same benefit as decision trees, while compensating for some of their shortcomings of overfitting. Single trees are still useful for visually representing the decision process.

[Figure 6 about here.]

**3.2.4 Gradient Boosting Classifier Tree.** Gradient boosting machines is another ensemble method that combines multiple decision trees for regression or classification by building trees in a serial fashion, where each tree tries to correct for mistakes of the previous one [10]. Gradient boosted regression trees use strong pre-pruning, with shallow trees of a depth of one to five. Each tree only provides a good estimate of part of the data, but combining many shallow trees (i.e., “weak learners”), the use many simple models iteratively improves performance. In addition to pre-pruning and the number of trees, an important parameter for gradient boosting is the

learning rate, which determines how strongly each tree tries to correct for mistakes of previous trees. A high learning rate produces stronger corrections, allowing for more complex models. Adding more trees to the ensemble also increases model complexity. Gradient boosting and random forests perform well on similar tasks and data; it is common to first try random forests and then include gradient boosting to attain improvements in accuracy of the learning model. This analysis used the *Gradient Boosting Classifier* from Scikit-Learn to classify Heroin Use, with the default setting of 100 trees of maximum depth of 3, and a learning rate of 0.1. The model was built on the training set and evaluated on the test set, with both training set and test set accuracy equal to 0.984. To reduce overfitting, pre-pruning could be implemented by reducing the maximum depth, or by reducing the learning rate. Figure 7 shows that the feature importance for the gradient boosting classifier tree looks similar to the feature importance for random forests, but the gradient boosting has decreased the importance of many features to zero. Again Cocaine is selected as the most informative features, followed by Any Opioid PRL Use. In addition to Prescription Opioid PRL Misuse and Abuse, the gradient boosting classifier selected Amphetamine Use as an informative feature of Heroin Use.

[Figure 7 about here.]

### 3.3 Classifier Models of Prescription Opioid Pain Reliever (PRL) Misuse

This section reports results from the same set of classification analyses described above using *Prescription Opioid Pain Reliever Misuse* (PRLMISEVR) as the target variable. Attributes related to Heroin Use were now included as features (e.g., HEROINEVR, HEROINUSE, HEROINFQY). The classifier models were built using SciKit Learn in a python notebook [18]. The dataset was split into the training set and test sets using the train-test-split function in sklearn and the target variables were designated. Model accuracy for the training set and test set are reported, for different parameter values, with feature importance.

**3.3.1 Logistic Regression Classifier.** The logistic regression classifier was fit to the training data using the L1 penalty (Lasso), using different values of the regularization parameter C, and the model was validated on the test data. Higher value of parameter C typically gives improved accuracy for both training and tests sets; however, in this case, the training set accuracy was 0.901 and test set accuracy was 0.903, and these values were consistent for all values of parameter C. Figure 8 plots the coefficients of logistic regression classifier for Prescription Opioid PRL Misuse under different values of C. As shown in Figure 8, the features with the highest coefficient values were Treatment (for substance use), Heroin Use (as predicted), as well as Cocaine and Amphetamine use. This result indicates that Prescription Opioid Misuse is positively related to Drug Treatment, meaning that respondents who reported higher levels of opioids misuse were also in treatment, but that people who were misusing opioid medications were also more likely to have used illicit drugs such as heroin, cocaine, and amphetamine.

[Figure 8 about here.]

**3.3.2 Decision Tree Classifier.** The Decision Tree Classifier package in Scikit-Learn was used to build the tree model, pre-pruning

was applied with a maximum depth of 4, which means the algorithm split on four consecutive questions. The training set accuracy of the pruned tree was 0.902 and test set accuracy was 0.902. Figure 9 shows a partial view of the decision tree classifier of prescription opioid misuse (the full tree is included in the BDA-Analytics-Classifier-PRL.ipynb notebook) [18]. As Figure 9 shows, the decision tree classifier selected Cocaine Use as the root note, that branched by the test score equal to or less than 0.5 (any Cocaine Use). At the second node, on the branch to the right n=5015 samples were further divided according to heroin use, with n=1913 having a score greater than 0.5 (any Heroin Use). At the third node on the right branch, samples were selected according to Tranquilizer medication use, with n=1419 scoring positively. On the left branch, the second node selected was Drug Treatment, with n=2844 respondents scoring positively that they had received Drug Treatment. Feature importance of the decision tree classifier selected Cocaine Use as the most informative feature for Prescription Opioid PRL Misuse. Following afterwards, Tranquilizer Use, Drug Treatment, and Heroin Use were tied for second place.

[Figure 9 about here.]

**3.3.3 Random Forests Classifier.** The Random Forest Classifier package in Scikit-Learn was used to classify Prescription Opioid PRL Misuse as the target variable, with 100 trees. The model accuracy for the training set was 0.955 and the test set accuracy was 0.896, which suggests that the model overfit the data. Figure 10 shows the feature importance of the random forests classifier for Prescription Opioid PRL Misuse. As Figure 10 shows, several features were identified as important for classifying Prescription Opioid PRL Misuse. The random forest selected Overall Health as the most informative feature in the model, followed by Cocaine Use, Education Level, Age Category, and Size of City Metropolitan region. Because of the additional features included as important, gradient boosting was performed to clarify the feature importance.

[Figure 10 about here.]

**3.3.4 Boosted Gradient Classifier.** The Gradient Boosting Classifier from Scikit-Learn was used to classify Prescription Opioid PRL Misuse, using the default setting of 100 trees, of maximum depth of 3, and a learning rate of 0.1. The model accuracy for the training set was 0.894 and accuracy for the test set was 0.893. Gradient boosting typically improves test set accuracy by using many simple models iteratively. In this case, model accuracy for gradient boosting was no better than random forests, and this is because the default parameter settings were used; further parameter tuning is needed to improve model performance. Feature importance was a primary interest for identifying features related to 'prescription opioid abuse. Figure 11 shows the feature importance for the gradient boosting classifier tree. As Figure 11 shows, several features were important for classifying prescription opioid misuse, and contrary to the random forests, gradient boosting selected Tranquilizer use as the most informative feature. Following closely in importance were Heroin Use and Age Category. Tied for fourth place were Cocaine Use and Treatment, with Mental Health (depression) coming in fourth in terms of feature importance. This result illustrates that several features are important for understanding Prescription Opioid Misuse, and the relations among features may be complex.

[Figure 11 about here.]

## 4 DISCUSSION

The results show that rates of prescription opioid use, misuse, and abuse are much higher than use of illicit opioids such as heroin and fentanyl. The use of Hydrocodone (Vicodan) was double the rate of Oxycodone use (Oxycodone) across almost all age groups. The use of traditional prescription opioids was greater than reported use of synthetic opioids. Illicit drug use was highest for respondents between the ages of 18 to 25. In terms of mental health, more individuals between 18 to 25 years reported experiencing a major depressive episode (in adulthood) than any other age group. In terms of the so-called *treatment gap*, almost twice as many respondents between 18 to 25 years who felt a need for substance use treatment, had not received treatment, than younger individuals between 12 to 17 years. The large majority of respondents (approximately 90 percent) had not misused prescription opioid pain relievers or used heroin. However, of those individuals who reported misusing prescription opioid pain relievers, almost twice as many had also used heroin than had not (see Figure 1), which partially supports the hypothesis that prescription opioid use is associated with use of illicit opioids such as heroin. Prescription opioid misuse and heroin use was also higher in large metropolitan areas than smaller cities or rural areas, but a small portion of individuals in non-metropolitan regions reported very high levels of prescription opioid misuse. These data points may represent outliers, but a large sample would allow for analysis of how opioid misuse and addiction differ for smaller rural regions versus large urban areas.

### 4.1 Comparison of Classifier Models

Several classifier algorithms were used to identify relevant features for predicting heroin use and prescription opioid misuse. Comparing the performance of different algorithms is helpful for selecting the best model. Test set accuracy was comparable across models for both Heroin Use (0.98) and Prescription Opioid PRL Misuse (0.89-0.90). Logistic Regression provided the feature coefficients for different values of the regularization parameter C. The Decision Tree classifier provided an easy to use, interpretable visual of the decisions involved at each step of classification. Random forests provides a more reliable indication of features importance than a single tree, whereas the gradient boosting classifier included additional tuning parameter for a more powerful model and more interpretable analysis of feature importance. Each classifier method provides a different level of analysis. For classifying heroin use, the logistic regression classifier showed that Prescription Opioid PRL Misuse had the highest coefficient value, but the tree-based classifiers each identified Cocaine Use as the most informative feature for predicting heroin use. For classifying Prescription Opioid PRL Misuse, logistic regression showed that Treatment had the highest coefficient value, but the tree based models each differed in selecting the most important features. Decision trees indicated that Cocaine Use was most informative, the random forests classifier selected health as the most important feature, and the gradient boosting model selected Tranquillizer use as most informative of prescription opioid PRL misuse. The different model each have

their advantages and limitations, logistic regression provides the coefficients, but random forests and gradient boosting are helpful for identified sets of important features.

### 4.2 Study Limitations

The main goal of this project was to identify features relevant for predicting opioid addiction by classifying cases according to heroin use. Only a small proportion of the sample reported having used heroin, and scores for mental health issues were very low. A limitation of survey data is that responses may be biased by under-reporting or minimizing the use of illicit or illegal substances. People may also be reluctant to disclose mental health issues or health problems (e.g., STDs, HIV status, suicide attempts). It is possible that this sample is representative of the frequency of opioid use and misuse in the larger population. Recent statistics from the CDC show that heroin use has increased among most demographics groups, with an average estimated rate of approximately 2.6 percent between 2011-2013 [7]. The rate of heroin use reported in the NSDUH-2015 sample was 1.6 percent. Therefore, it seems that the actual rate of heroin use in the U.S. population may not be accurately reflected in this sample. Another limitation is that the project dataset was constructed as a subset of features from the NSDUH-2015 data. Ninety attributes out of 2666 features in the original data were selected, and many features were combined to create aggregated variables for health, mental health, prescription opioid misuse and abuse, drug treatment, mental health treatment. Future research could include a more comprehensive selection of features to identify the set of features relevant for predicting opioid dependency and addiction. An important challenge for making sense of big data is developing analytic tools adequate to handle large volumes of data.

### 4.3 Extension to Big Data

A general tenet of big data is that, “More data is always better.” The methods used in this project could be extended to better approximate big data for predicting opioid use in the following ways: (1) Include a larger selection of features from the attributes in the NSDUH-2015 dataset; (2) Include survey data from previous years (e.g., 2005-2015) for a larger sample; and (3) Obtain a broader sample from the population of patients who are taking prescribed opioid medications. The most immediate step would be to include additional features for use with the classifier models. Additional data from the NSDUH was downloaded from previous years (2012 to 2014); preliminary examination of the data revealed inconsistencies in questions and prescription opioid medications that would need to be resolved in order to combine data from multiple years. Data cleaning can be a time consuming process, but important for obtaining usable data. Unfortunately, owing to constraints of time for completing the project, it was not possible to integrate data from previous years into the project dataset. In working with big data, there are several steps involved in the consolidation of data from multiple sources into a single dataset (in addition to data cleaning), which include extraction, integration, and aggregation of features [13]. A future study could integrate data from different years, using a broader set of features, with more inclusive sample

representative of the larger population, and integrate data from multiple sources.

#### 4.4 Opioid Addiction and Epidemic Spreading

Drug addiction has many similar characteristics to other chronic medical illnesses, but there are unique challenges to the treatment of addiction [8, 23]. In drug rehabilitation treatment programs, patients undergo intense detoxification that reduces their drug tolerance, but are then released back into the environments associated with their drug use, putting them at high risk for relapse and potential drug overdose [6]. If the prescription opioid crisis is a genuine epidemic, we must consider the process of spreading or diffusion of contagion. Epidemic spreading is a dynamic process based on networks of direct person-to-person contact and indirect exposure via transportation pathways [2]. Epidemics are quantified in terms of the proportion of the population infected, those yet to be infected, and the rate of transmission. Potentially everyone is at risk of becoming dependent or addicted to prescription medications or illicit opioids. In terms of the opioid epidemic, rather than labeling persons as infected or uninfected, it is more useful to consider people as either susceptible to dependence and addiction or less susceptible. Furthermore, the structure of the contact network can influence epidemic spreading [12]. For example, in the case of simple contagion, weak ties among acquaintances or infrequent associations provide shortcuts between distant nodes that reduce distance within the network [?] which can facilitate the spread of contagion, or in this case drug use. Furthermore, contact networks for drug use may have “small world” properties where a small number of nodes have a high number of connections that can rapidly transmit contagion throughout the network [?]. Network analysis may help to identify the underlying structure of the contact network of opioid use, to examine pathways and points of contact in the misuse and abuse of prescription opioid medications. According to a classical conditioning model of addiction, situational cues or events can elicit a motivational state underlying relapse to drug use. Addictive behavior can be also be reinstated after extinction of dependency by exposure to drug-related cues or stressors in the environment [15]. Future research could use social network modeling to explore how drug dependency and addiction are subserved by patterns of social interaction.

### 5 CONCLUSION

This project compared several classification algorithms to predict heroin use and prescription opioid misuse and abuse. The results provided partial support for the hypothesis that prescription opioid misuse is associated with the use of illicit opioids such as heroin. Several features were identified as important for classifying heroin use, including Cocaine Use, Amphetamine Use, and any prescription opioid medication use. In regards to predicting heroin use, it appears the use of other illicit drugs such as Cocaine and Amphetamine was perhaps more informative than any prescription opioid use or misuse. Heroin use was selected as important for classifying prescription opioid pain reliever misuse, but additional factors also played a role, including tranquilizer use, age category, overall health, cocaine use. Substance treatment had the largest regression coefficient, suggesting that people who are misusing

prescription opioid pain medication are also more likely to be in drug treatment programs. The direction of these effects cannot be determined owing to the nature of the analyses. On the one hand individual misusing or abusing prescription opioids may also be using heroin. Alternatively, individuals with a susceptibility for opioid use may be equally likely to have used heroin and also to have misused prescription opioids. A general conclusion is that of those individuals who reported misusing prescription opioid medications, twice as said they had used heroin than reported they had not used heroin. The results do not provide sufficient evidence to rule out alternative hypotheses. Given the relatively low rates of opioid and heroin in this sample, additional evidence is needed to resolve this question. The study can provide information to raise awareness about the risk factors for prescription opioid addiction and may help reduce opioid overdose deaths.

### ACKNOWLEDGMENTS

The author would like to thank Dr. Gregor von Laszewski, the Teaching Assistants, Juliette Zurick, Miao Jiang, Hungri Lee, Grace Li, Saber Sheybani Moghadam, and others who helped to improve this project and report.

### REFERENCES

- [1] Substance Abuse, Center for Behavioral Health Statistics Mental Health Services Administration, and Quality. 2016. *National Survey on Drug Use and Health (NSDUH) 2015*. Online data archive. United States Department of Health and Human Services, Ann Arbor, MI. <https://doi.org/10.3886/ICPSR50011.v1>
- [2] Vittoria Colizza, Alain Barrat, Marc Barthélémy, and Alessandro Vespignani. 2006. The role of the airline transportation network in the prediction and predictability of global epidemics. *Proceedings of the National Academy of Sciences of the United States of America* 103, 7 (2006), 2015–2020. <https://doi.org/10.1073/pnas.0510525103> arXiv:<http://www.pnas.org/content/103/7/2015.full.pdf>
- [3] Centers for Disease Control and Prevention. 2017. Prescription Opioid Overdose Data. online. (Oct. 2017). <https://www.cdc.gov/drugoverdose/data/overdose.html>
- [4] hd1 and yoavram. 2016. Python: Download Returned Zip file from URL. Online. (Feb. 2016). <https://stackoverflow.com/questions/9419162/python-download-returned-zip-file-from-url> Stackoverflow.com.
- [5] M. Herland, T. M. Khoshgoftaar, and R. Wald. 2014. A review of data mining using big data in health informatics. *Journal Of Big Data* 1, 2 (2014). <https://doi.org/10.1186/2196-1115-1-2>
- [6] K. Johnson, A. Isham, D.V. Shah, and D.H. Gustafson. 2011. Potential Roles for New Communication Technologies in Treatment of Addiction. *Current psychiatry reports*, (2011). <https://doi.org/10.1007/s11920-011-0218-y>
- [7] Rose A Judd, Noah Aleshire, Jon E. Zibbell, and R. Matthew Gladden. 2016. *Increases in Drug and Opioid Overdose Deaths, United States, 2000–2014*. techreport 64(50). Centers for Disease Control and Prevention, Atlanta, GA. <https://www.cdc.gov/mmwr/preview/mmwrhtml/mm6450a3.htm> Morbidity and Mortality Weekly Report (MMWR).
- [8] Lisa A. Marsch. 2012. Leveraging technology to enhance addiction treatment and recovery. *Journal of Addictive Diseases* 31, 3 (2012), 313–318. <https://doi.org/10.1080/10550887.2012.694606>
- [9] Wes McKinney. 2017. *Python for Data Analysis*. O'Reilly Media Inc., Sebastopol, CA. <https://github.com/wesm/pydata-book>
- [10] Andreas C. Müller and Sarah Guido. 2017. *Introduction to Machine Learning*. O'Reilly, Sebastopol, CA. [https://github.com/amueller/introduction\\_to\\_ml\\_with\\_python/](https://github.com/amueller/introduction_to_ml_with_python/)
- [11] National Institute on Drug Abuse (NIDA). 2017. *Overdose Death Rates*. Summary. National Institutes of Health (NIH), Washington D.C. <https://www.drugabuse.gov/related-topics/trends-statistics/overdose-death-rates>
- [12] Romualdo Pastor-Satorras and Alessandro Vespignani. 2001. Epidemic Spreading in Scale-Free Networks. *Phys. Rev. Lett.* 86 (Apr 2001), 3200–3203. Issue 14. <https://doi.org/10.1103/PhysRevLett.86.3200>
- [13] E. Rahm and H. Hai Do. 2000. *Data cleaning: Problems and current approaches*. techreport 23(4). Bulletin of the Technical Committee on Data Engineering, 1730 Massachusetts Avenue, Washington D.C. <https://s3.amazonaws.com/academia.edu.documents/41858217/A00DEC-CD.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1511155930&Signature=VWRM7u4KwP6ZxX5jB%2Bh6wMCbpg%3D&>

- response-content-disposition=inline%3B%20filename%3DAutomatically-extracting.structure\_.from.pdf#page=5
- [14] Sebastian Raschka and Vahid Mirjalili. 2017. *Python Machine Learning, Second Edition*. Packt, Birmingham, UK. <https://github.com/rasbt/python-machine-learning-book-2nd-edition>
  - [15] Yavin Shaham, Uri Shalev, Lin Lu, Harriet de Wit, and Jane Stewart. 2003. The reinstatement model of drug relapse: history, methodology and major findings. *Psychopharmacology* 168, 1 (01 Jul 2003), 3–20. <https://doi.org/10.1007/s00213-002-1224-x>
  - [16] S.M. Shiverick. 2017. BDA Project Data. github. (Nov. 2017). <https://github.com/bigdata-i523/hid335/blob/master/project/BDA-Project.ipynb>
  - [17] S.M. Shiverick. 2017. Classification Models of Heroin Use. github. (Nov. 2017). <https://github.com/bigdata-i523/hid335/blob/master/project/BDA-Analytics-Classifier-Heroin.ipynb> Interactive Python Jupyter Notebook.
  - [18] S.M. Shiverick. 2017. Classification Models of Prescription Opioid Pain Relievers Misuse. github. (Nov. 2017). <https://github.com/bigdata-i523/hid335/blob/master/project/BDA-Analytics-Classifier-PRL.ipynb> Interactive Python Jupyter Notebook.
  - [19] S.M. Shiverick. 2017. Project Codebook for Data Variables from NSDUH-2015. github. (Nov. 2017). <https://github.com/bigdata-i523/hid335/blob/master/project/project-data-codebook.txt>
  - [20] S. M. Shiverick. 2017. Exploratory Data Analysis. Github. (Dec. 2017). <https://github.com/bigdata-i523/hid335/blob/master/project/BDA-Project-Explore-Data.ipynb>
  - [21] S. M. Shiverick. 2017. Project Data Visualization. Github. (Dec. 2017). <https://github.com/bigdata-i523/hid335/blob/master/project/BDA-Project-Explore-Data.ipynb>
  - [22] S. M. Shiverick. 2017. Project Workflow Pipeline. Github. (Dec. 2017). <https://github.com/bigdata-i523/hid335/blob/master/project/readme.md>
  - [23] J. Swendsen. 2016. Contributions of mobile technologies to addiction research. *Dialogues Clinical Neuroscience* 18, 2 (June 2016), 213–221. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4969708/>
  - [24] Jake VanderPlas. 2017. *Python Data Science Handbook*. O'Reilly Media Inc., Sebastopol, CA. <https://jakevdp.github.io/PythonDataScienceHandbook/>
  - [25] Upkar Varshney. 2013. Smart medication management system and multiple interventions for medication adherence. *Decision Support Systems* 55, 5 (May 2013), 538–551. <https://doi.org/10.1016/j.dss.2012.10.011>
  - [26] Nora D. Volkow, Thomas R. Frieden, Pamela S. Hyde, and Stephen S. Cha. 2014. Medication-Assisted Therapies: Tackling the Opioid-Overdose Epidemic. *New England Journal of Medicine* 370, 22 (2014), 2063–2066. <https://doi.org/10.1056/NEJMmp1402780> arXiv:<http://dx.doi.org/10.1056/NEJMmp1402780> PMID: 24758595.

## A CODE REFERENCES

All code, notebooks, files, and folders for this project can be found in the i523/hid335/project github repository: <https://github.com/bigdata-i523/hid335/tree/master/project>. An outline of the workflow pipelines was included as a readme.md markdown file [22].

### A.1 Download and Extract Data

The get-data.py function was written to download the data, unzip the data files, extract the data, and write the NSDUH-2015 dataset to CSV file [4].

### A.2 Data Cleaning and Preparation

Data cleaning and preparation steps was conducted using an interactive python Jupyter Notebook [16] based on examples in Python for Data Analysis [9] and the Python Data Science Handbook [24].

### A.3 Exploratory Data Analysis

Exploratory Data Analysis of the NSDUH-2015 dataset was conducted using an interactive python notebook [20] based on examples from Python for Data Analysis [9], and the Python Data Science Handbook [24].

## A.4 Data Visualization

Several plots and graphs were constructed in a Data Visualization interactive python notebook [21] using Matplotlib and Seaborn python visualization packages [9, 24].

## A.5 Classification Algorithms

Machine learning classification models were constructed using SciKit Learn [10, 14] in two separate Jupyter Notebooks, one for classifier models of Heroin Use as the target variable [17], and another for classifier models of Prescription Opioid PRL Misuse as the target [18].

## B ISSUES

DONE:

Example of done item: Once you fix an item, change TODO to DONE

### B.1 Assignment Submission Issues

DONE:

Do not make changes to your paper during grading, when your repository should be frozen.

### B.2 Uncaught Bibliography Errors

DONE:

Missing bibliography file generated by JabRef

DONE:

Bibtex labels cannot have any spaces, \_ or & in it

DONE:

Citations in text showing as [?]: this means either your report.bib is not up-to-date or there is a spelling error in the label of the item you want to cite, either in report.bib or in report.tex

### B.3 Formatting

DONE:

Incorrect number of keywords or HID and i523 not included in the keywords

DONE:

Other formatting issues

### B.4 Writing Errors

DONE:

Errors in title, e.g. capitalization

DONE:

Spelling errors

DONE:

Are you using *a* and *the* properly?

DONE:

Do not use phrases such as *shown in the Figure below*. Instead, use *as shown in Figure 3*, when referring to the 3rd figure

DONE:

Do not use the word *I* instead use *we* even if you are the sole author

DONE:

Do not use the phrase *In this paper/report we show* instead use *We show*. It is not important if this is a paper or a report and does not need to be mentioned

DONE:

If you want to say *and* do not use & but use the word *and*

DONE:

Use a space after . , :

DONE:

When using a section command, the section title is not written in all-caps as format does this for you

\section{Introduction} and NOT \section{INTRODUCTION}

## B.5 Citation Issues and Plagiarism

DONE:

It is your responsibility to make sure no plagiarism occurs. The instructions and resources were given in the class

DONE:

Claims made without citations provided

DONE:

Need to paraphrase long quotations (whole sentences or longer)

DONE:

Need to quote directly cited material

## B.6 Character Errors

DONE:

Erroneous use of quotation marks, i.e. use “quotes”, instead of ” ”

DONE:

To emphasize a word, use *emphasize* and not “quote”

DONE:

When using the characters & # % \_ put a backslash before them so that they show up correctly

DONE:

Pasting and copying from the Web often results in non-ASCII characters to be used in your text, please remove them and replace accordingly. This is the case for quotes, dashes and all the other special characters.

DONE:

If you see a ffigure and not a figure in text you copied from a text that has the fi combined as a single character

## B.7 Structural Issues

DONE:

Acknowledgement section missing

DONE:

Incorrect README file

DONE:

In case of a class and if you do a multi-author paper, you need to add an appendix describing who did what in the paper

DONE:

The paper has less than 2 pages of text, i.e. excluding images, tables and figures

DONE:

The paper has more than 6 pages of text, i.e. excluding images, tables and figures

DONE:

Do not artificially inflate your paper if you are below the page limit

## B.8 Details about the Figures and Tables

DONE:

Capitalization errors in referring to captions, e.g. Figure 1, Table 2

DONE:

Do use *label* and *ref* to automatically create figure numbers

DONE:

Wrong placement of figure caption. They should be on the bottom of the figure

DONE:

Wrong placement of table caption. They should be on the top of the table

DONE:

Images submitted incorrectly. They should be in native format, e.g. .graffle, .pptx, .png, jpg

DONE:

Do not submit eps images. Instead, convert them to PDF

DONE:

The image files must be in a single directory named "images"

DONE:

In case there is a powerpoint in the submission, the image must be exported as PDF

DONE:

Make the figures large enough so we can read the details. If needed make the figure over two columns

DONE:

Do not worry about the figure placement if they are at a different location than you think. Figures are allowed to float. For this class, you should place all figures at the end of the report.

DONE:

In case you copied a figure from another paper you need to ask for copyright permission. In case of a class paper, you must include a reference to the original in the caption

DONE:

Remove any figure that is not referred to explicitly in the text (As shown in Figure ..)

DONE:

Do not use `textwidth` as a parameter for `includegraphics`

DONE:

Figures should be reasonably sized and often you just need to add `columnwidth`

e.g.

```
/includegraphics[width=\columnwidth]{images/myimage.pdf}  
re
```

## LIST OF FIGURES

1	Proportion of Individuals Who Reported Ever Misusing Prescription Opioid Pain Relievers and Proportion Who Reported Using Heroin	13
2	Plot of Opioid Pain Medication Misuse and Abuse and Heroin Use with Regression Slopes Weighted by Metropolitan Area Size	14
3	Pairplots of Mental Health, Prescription Opioid Misuse and Abuse, Heroin Use, and Size of City Metropolitan Area	15
4	Coefficients of Logistic Regression Classifier of Heroin Use (With L1 Penalty and Values of Regularization Parameter C)	16
5	Decision Tree Classification of Heroin Use (Partial View)	17
6	Feature Importance for Random Forests Classifier for Heroin Use	18
7	Feature Importance for Gradient Boosting Classifier for Heroin Use	19
8	Logistic Regression Classification of Prescription Opioid (PRL) Misuse with L2 Penalty	20
9	Decision Tree for Prescription Opioid (PRL) Misuse	21
10	Feature Importance for Random Forest Classifier of Prescription Opioid (PRL) Misuse	22
11	Feature Importance for Gradient Boosted Classifier Tree of Prescription Opioid (PRL) Misuse	23

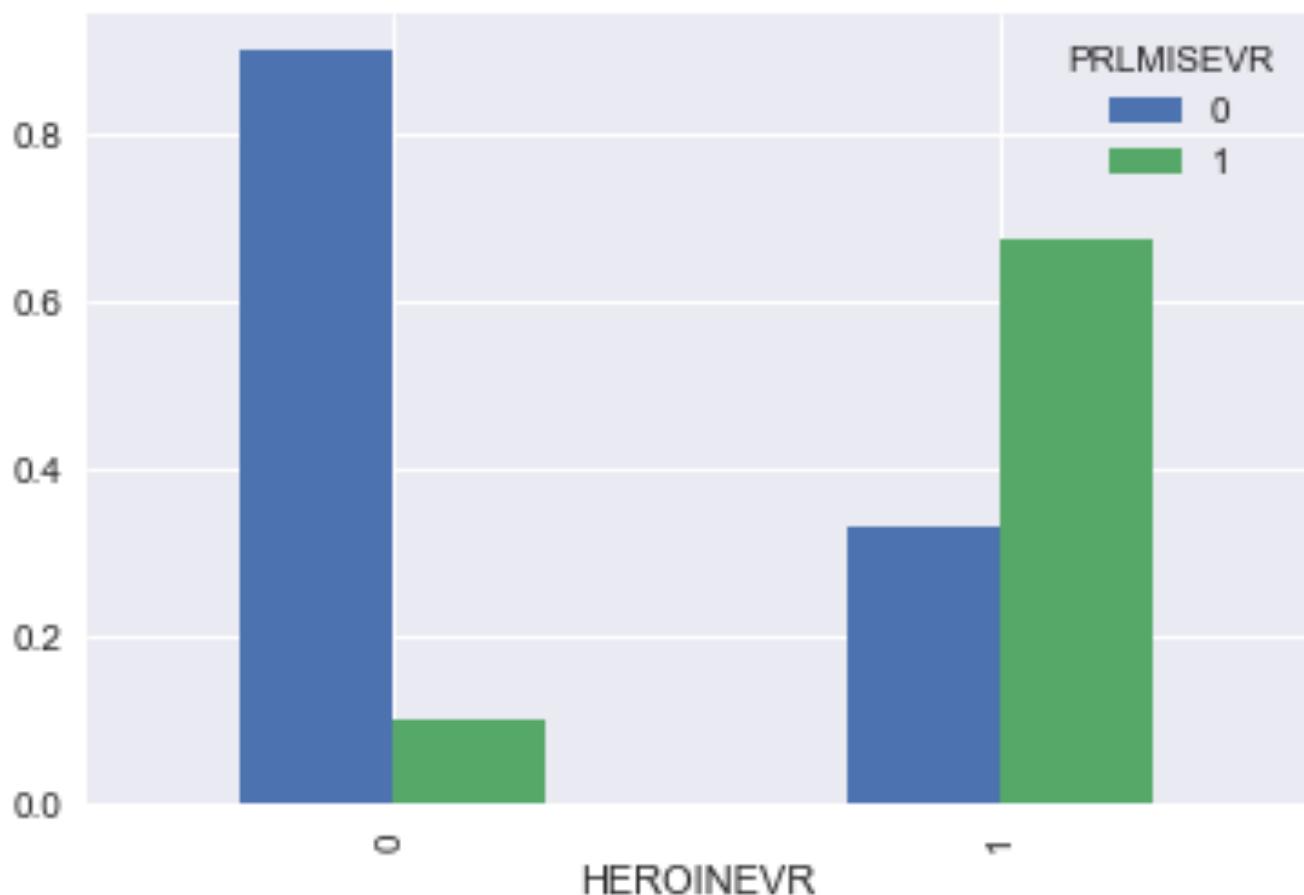


Figure 1: Proportion of Individuals Who Reported Ever Misusing Prescription Opioid Pain Relievers and Proportion Who Reported Using Heroin

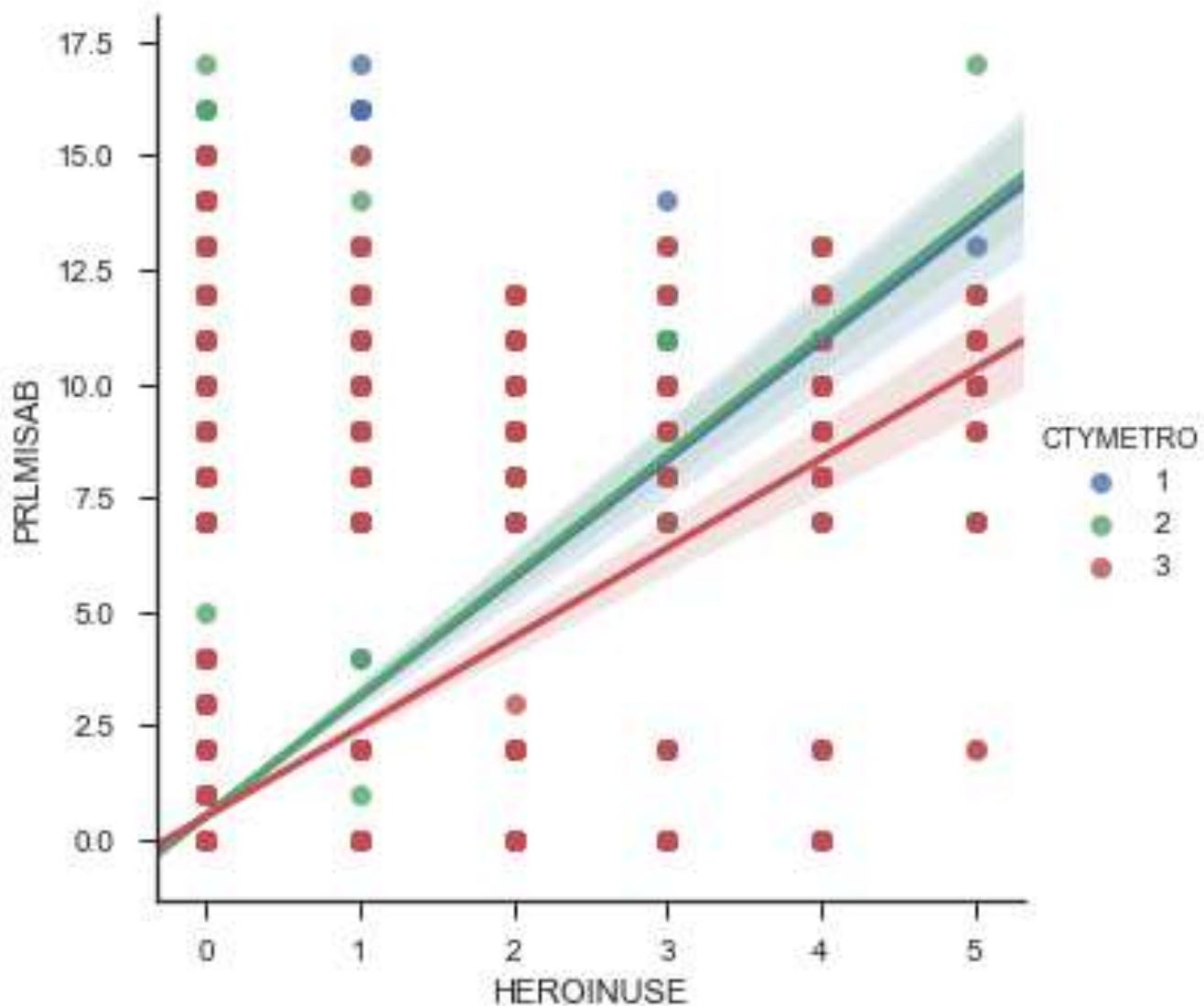


Figure 2: Plot of Opioid Pain Medication Misuse and Abuse and Heroin Use with Regression Slopes Weighted by Metropolitan Area Size

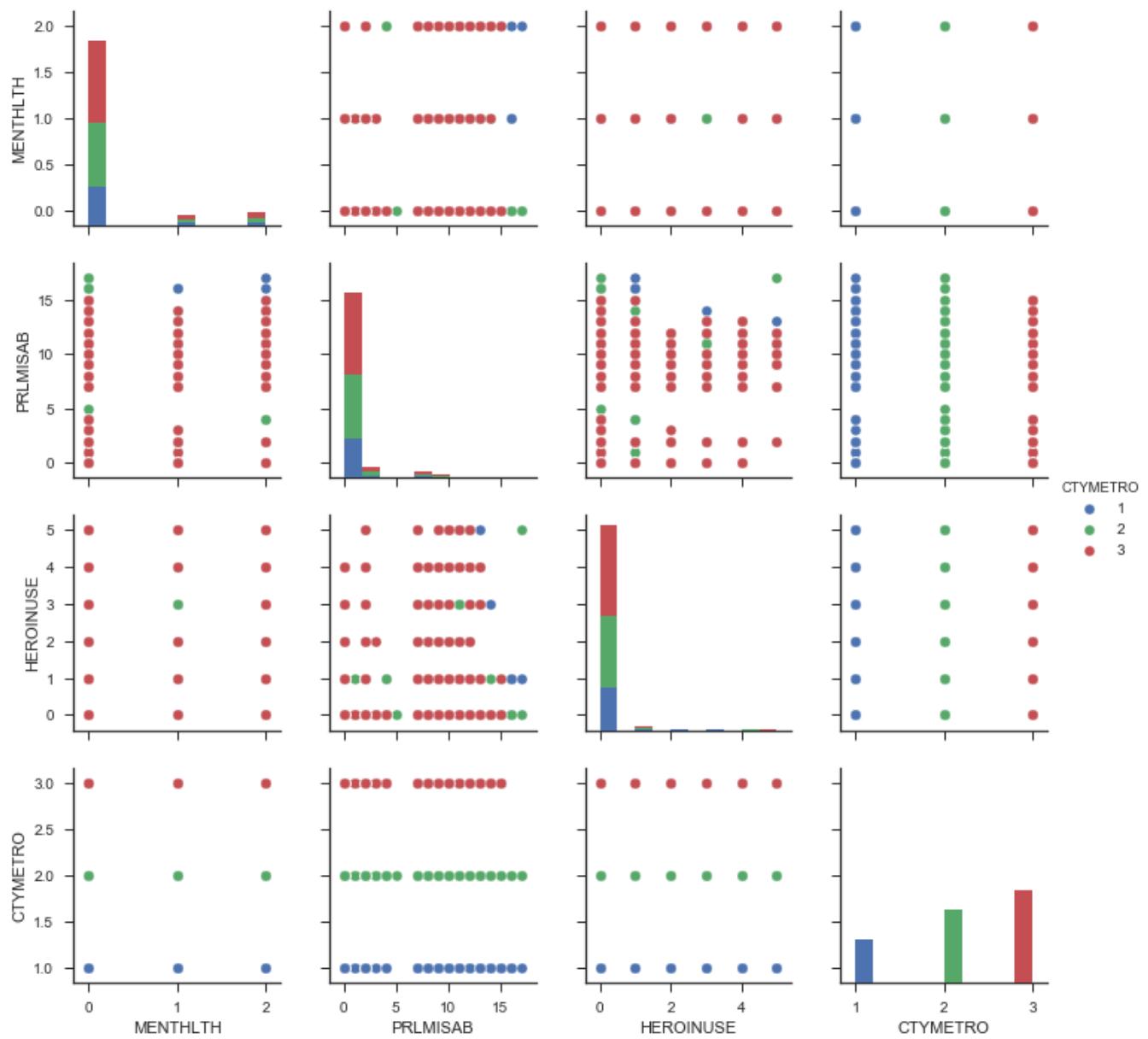


Figure 3: Pairplots of Mental Health, Prescription Opioid Misuse and Abuse, Heroin Use, and Size of City Metropolitan Area

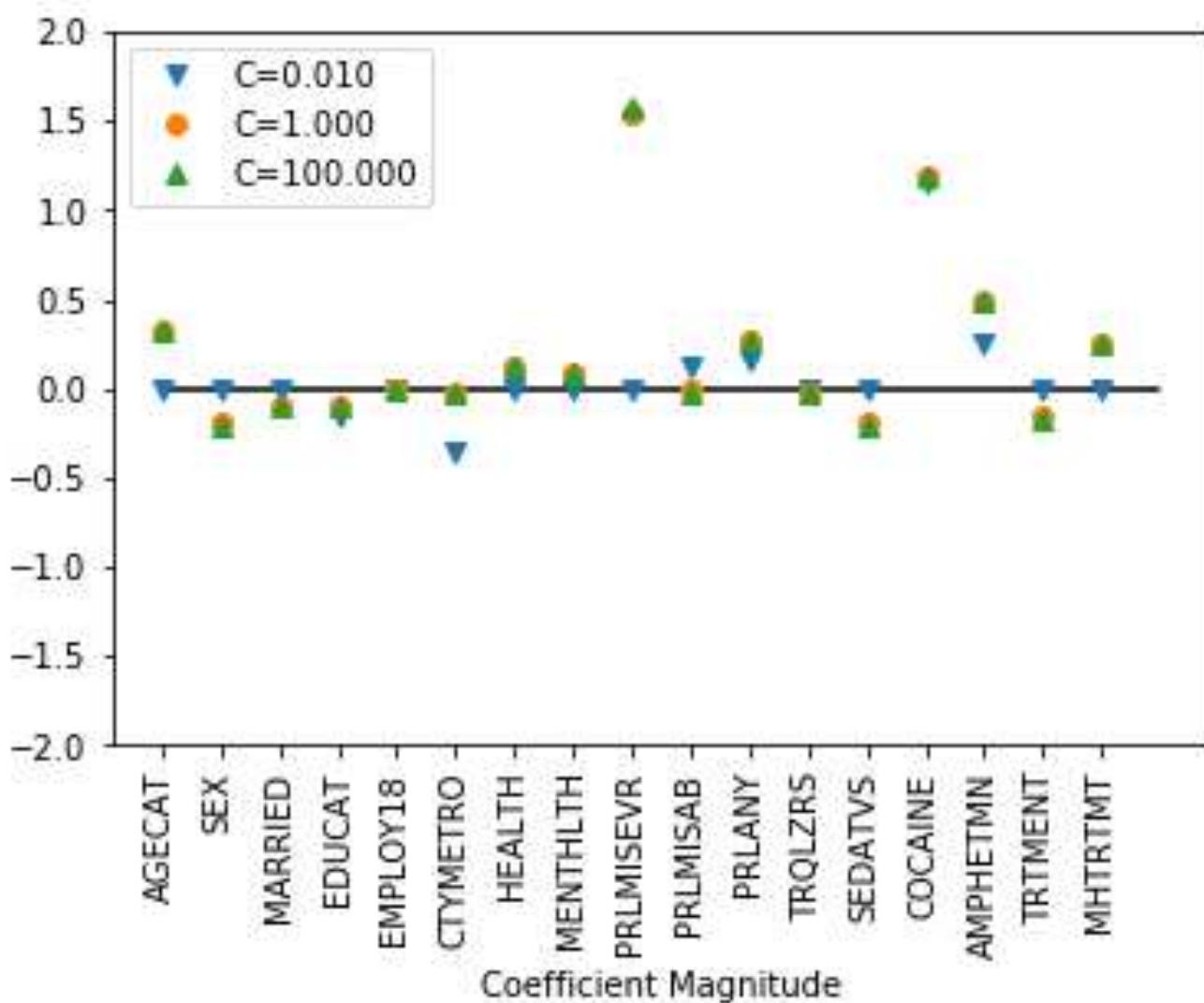


Figure 4: Coefficients of Logistic Regression Classifier of Heroin Use (With L1 Penalty and Values of Regularization Parameter C)

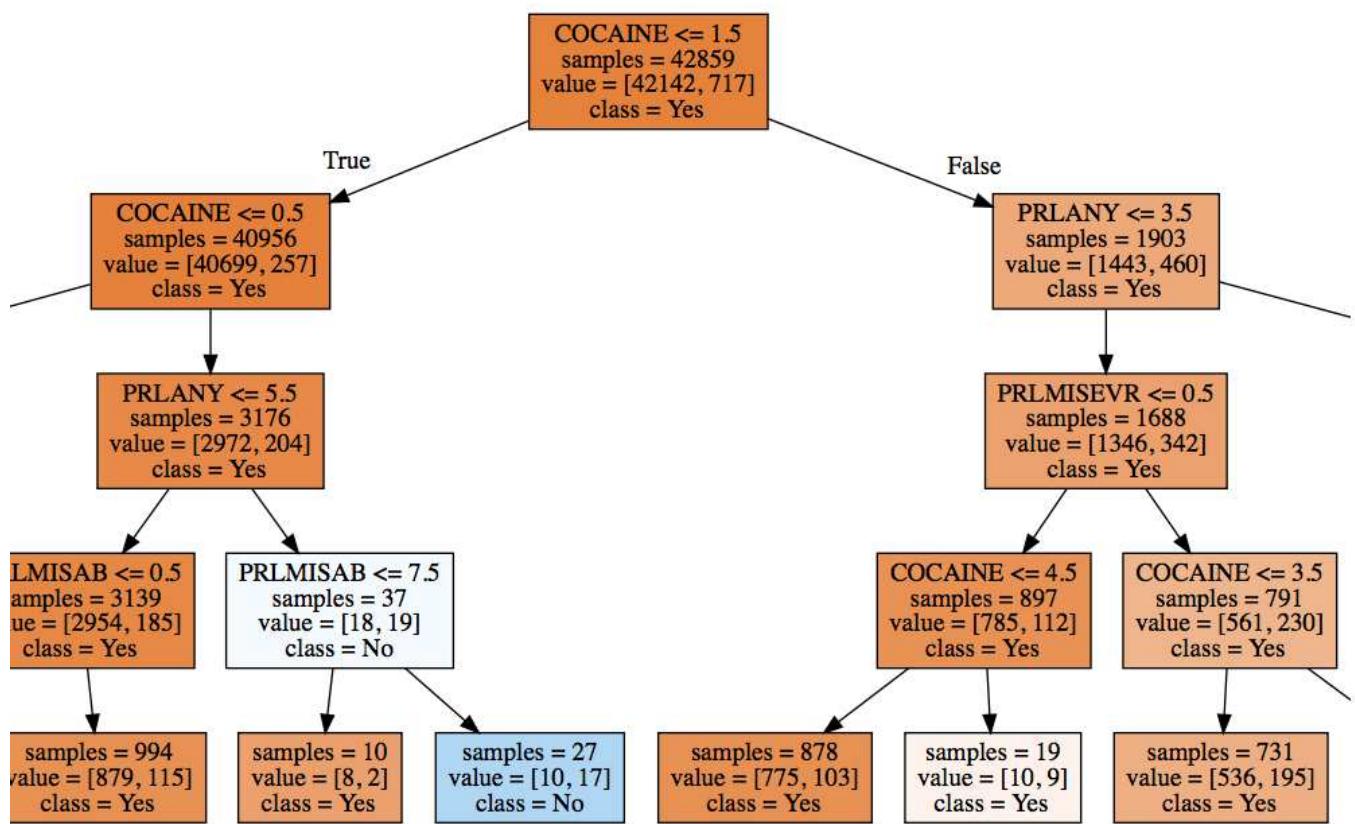


Figure 5: Decision Tree Classification of Heroin Use (Partial View)

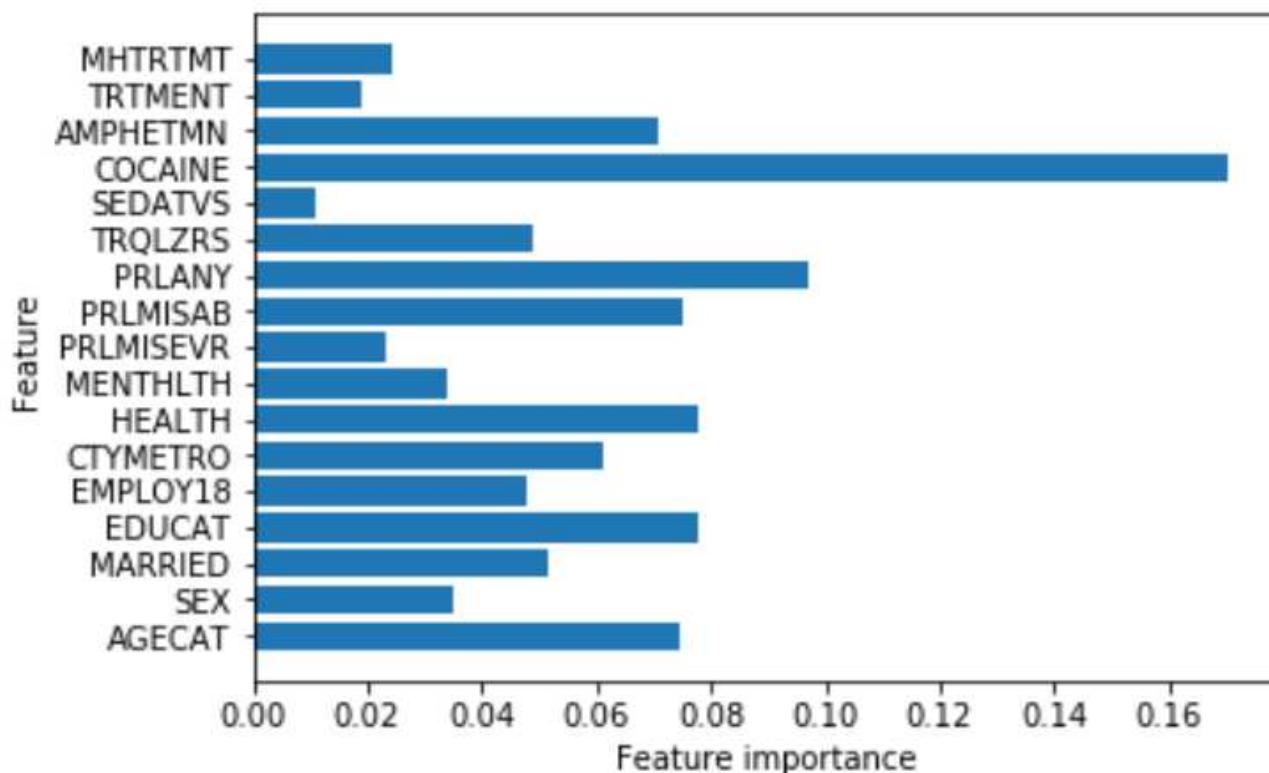


Figure 6: Feature Importance for Random Forests Classifier for Heroin Use

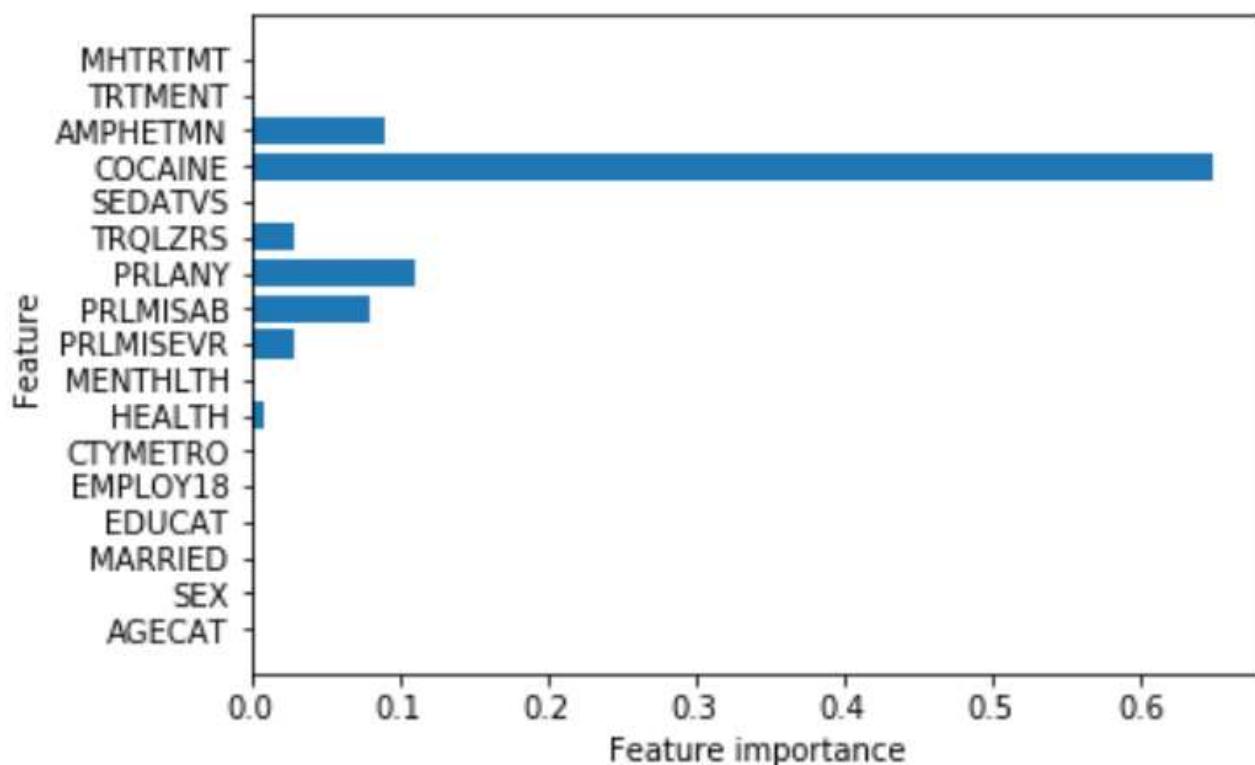


Figure 7: Feature Importance for Gradient Boosting Classifier for Heroin Use

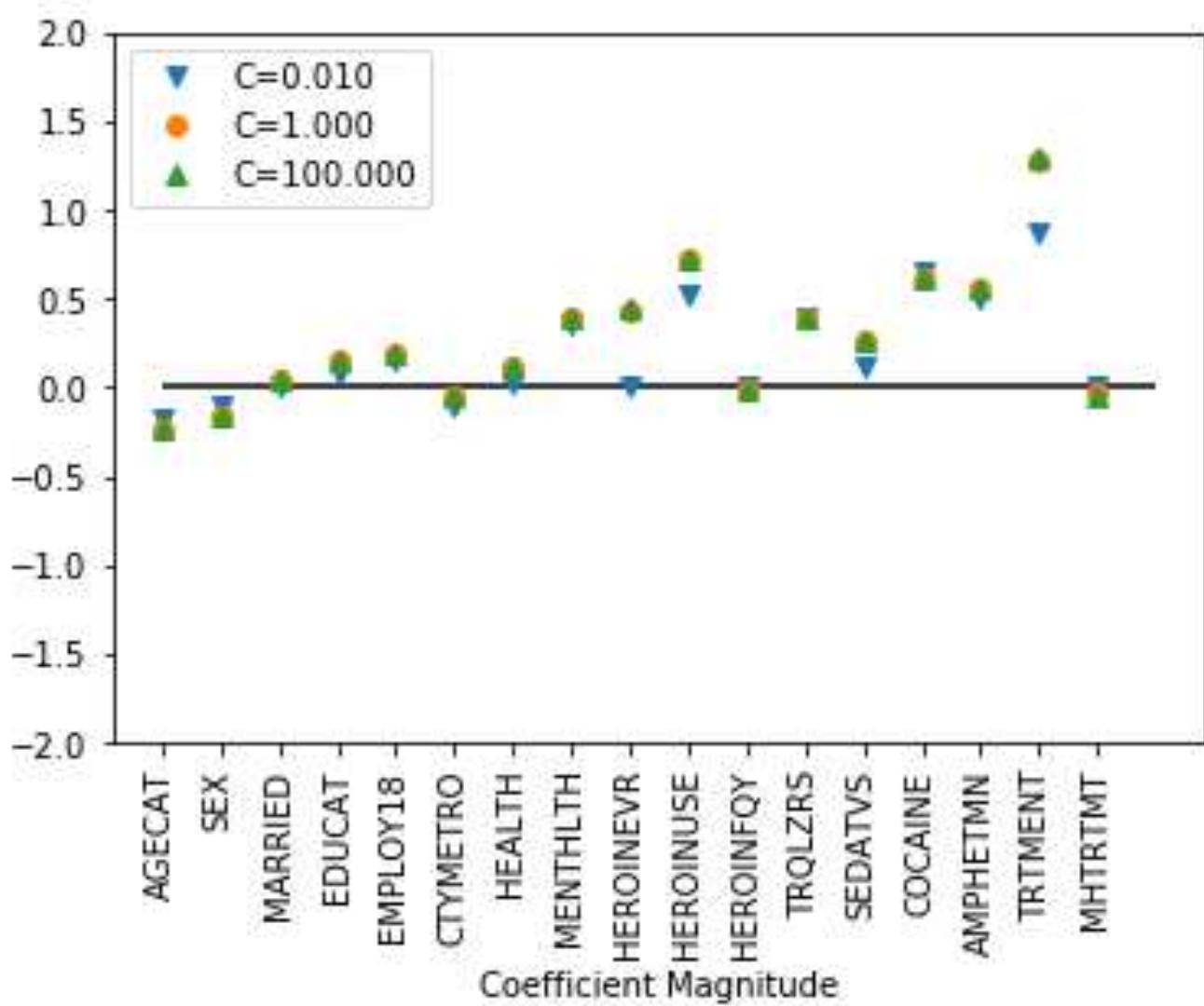


Figure 8: Logistic Regression Classification of Prescription Opioid (PRL) Misuse with L2 Penalty

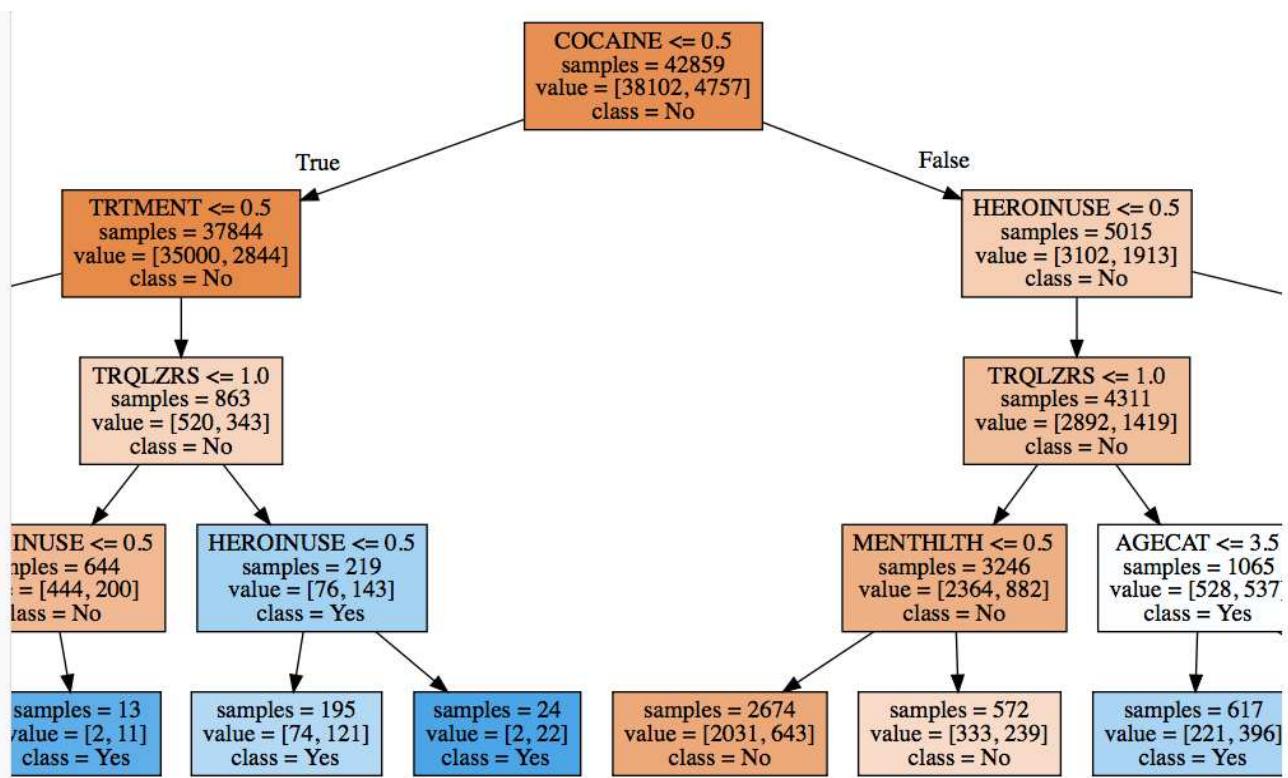


Figure 9: Decision Tree for Prescription Opioid (PRL) Misuse

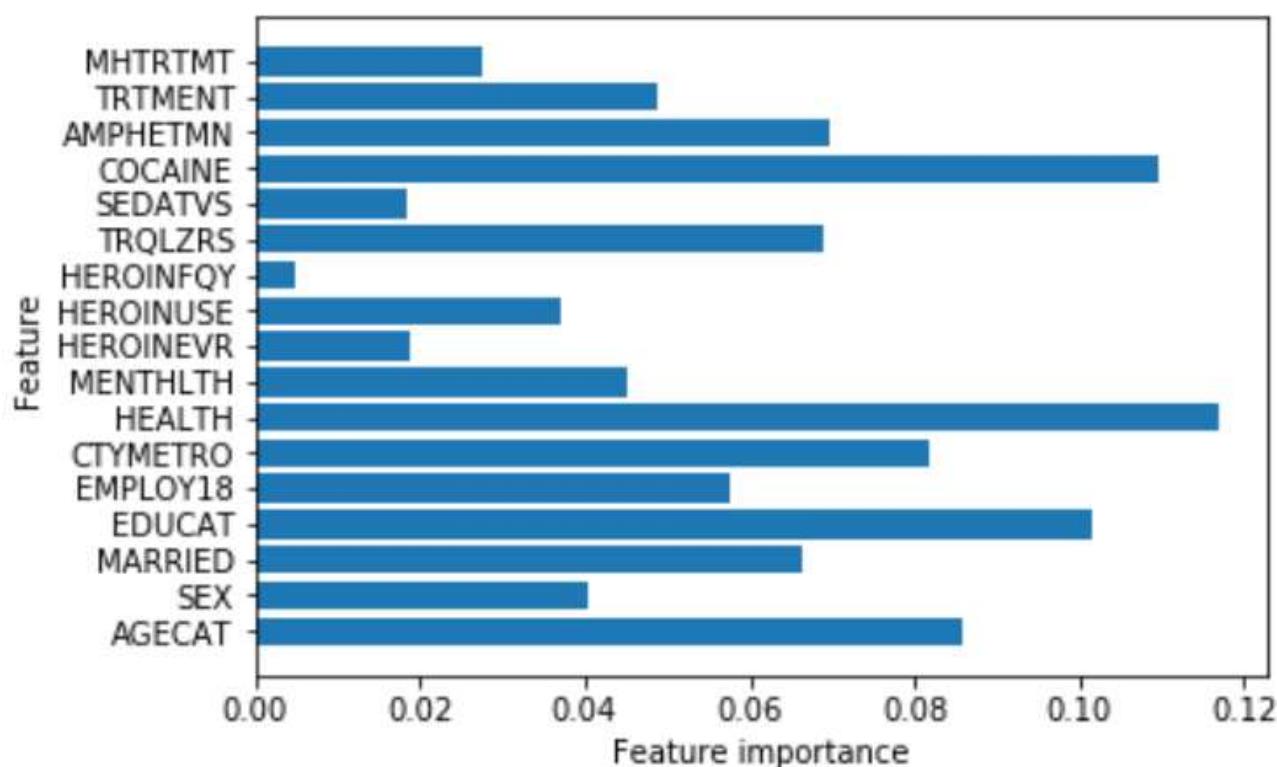


Figure 10: Feature Importance for Random Forest Classifier of Prescription Opioid (PRL) Misuse

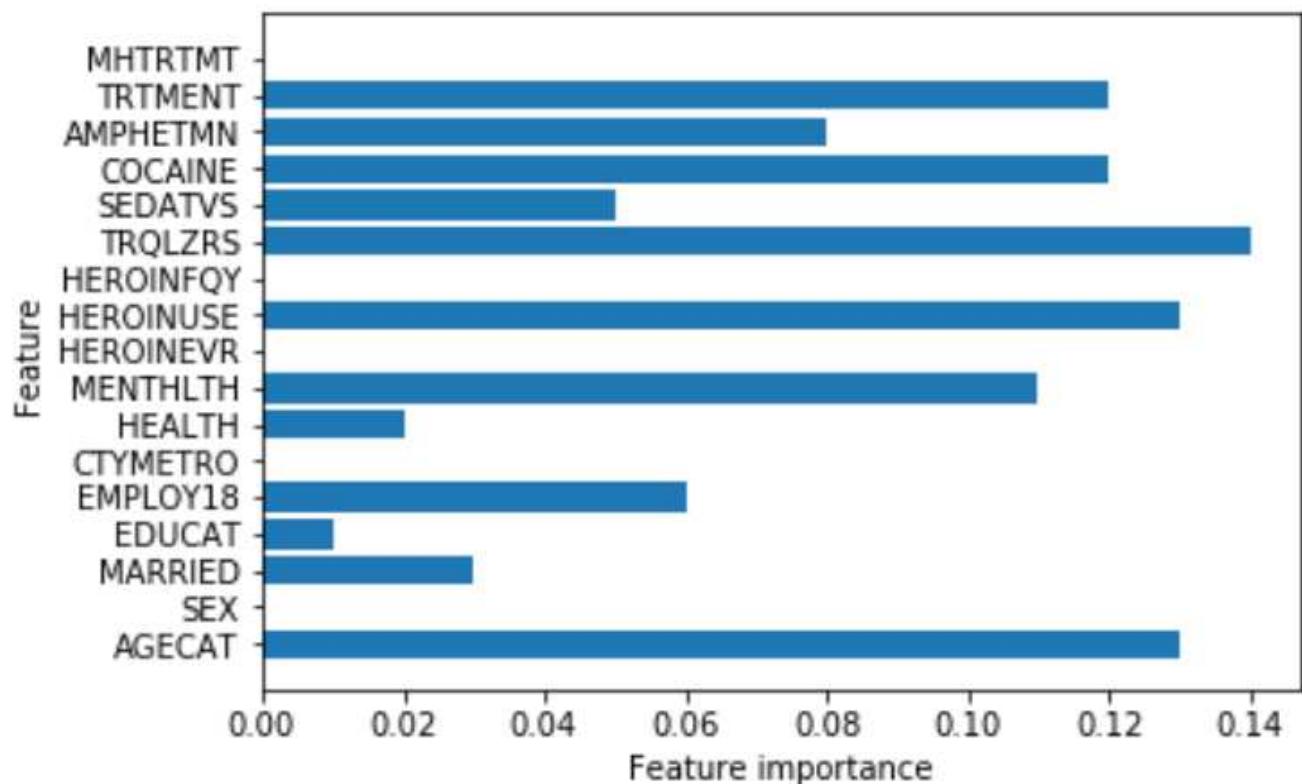


Figure 11: Feature Importance for Gradient Boosted Classifier Tree of Prescription Opioid (PRL) Misuse

LIST OF TABLES

1	Substance Use by Age Group Counts - NSDUH 2015 [1]	25
2	Frequency Table of Mental Health Issues and Treatment NSDUH 2015 [1]	25

**Table 1: Substance Use by Age Group Counts - NSDUH 2015 [1]**

Age Group	12-17	18-25	26-34	35-49	50+
Sample Size	13585	14553	9084	11169	8755
Oxycodone	545	1632	1132	1345	1044
Hydrocodone	831	2936	2233	2781	2103
Tramadol	241	753	654	829	734
Morphine	251	431	236	313	286
Fentanyl	28	97	81	96	86
Demerol	26	74	49	64	71
Buprenorphine	43	197	167	124	51
Oxymorphone	46	88	57	47	41
Hydromorphone	24	94	107	118	81
PRL Misuse Ever*	798	2127	1475	1343	600
Tranquilizers	405	1469	1064	1405	1153
Sedatives	204	242	157	256	226
Methadone Ever	32	83	96	71	46
Heroin Use Ever*	22	261	259	250	164
Cocaine Use Ever	109	1645	1626	1954	1406
Amphetamines Ever	932	1836	627	383	164
Methamphetamine	42	481	700	898	492
Hallucinogens	450	2660	2020	2127	1197
LSD Use Ever	190	1114	874	1442	907
Ecstasy (MDMA)	199	1867	1403	947	149

**Table 2: Frequency Table of Mental Health Issues and Treatment NSDUH 2015 [1]**

Age Group	12-17	18-25	26-34	35-49	50+
In Hospital Overnight	730	1149	821	890	1173
Adult Depression	0	2413	1395	1766	967
Mental Health Treatment					
Private Therapist	0	592	434	554	311
Treatment Gap*	469	931	321	239	90

```
bibtext report
```

```
=====
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
Warning--I didn't find a database entry for "granovetter73"
Warning--I didn't find a database entry for "watts98"
Warning--page numbers missing in both pages and numpages fields in herland14
Warning--no number and no volume in johnson11
Warning--page numbers missing in both pages and numpages fields in johnson11
(There were 5 warnings)
```

```
bibtext _ label error
```

```
=====
bibtext space label error
```

```
=====
bibtext comma label error
```

```
=====
latex report
```

```
[2017-12-11 13.31.04] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflatex)
Label `tab:freq' multiply defined.
p.8 L930 : [granovetter73] undefined
p.8 L934 : [watts98] undefined
Missing character: ""
There were undefined citations.
bookmark level for unknown defaults to 0.
The anchor of a bookmark and its parent's must not be the same. Added a new anchor.
```

The anchor of a bookmark and its parent's must not be the same. Added a new anchor.  
There were multiply-defined labels.  
Typesetting of "report.tex" completed in 1.4s.

---

## Compliance Report

---

```
name: Sean Shiverick
hid: 335
paper1: 10/25/17 100%
paper2: 100%
project: 100%
```

```
yamlcheck
```

---

## wordcount

---

```
25
wc 335 project 25 8446 report.tex
wc 335 project 25 9539 report.pdf
wc 335 project 25 1078 report.bib
```

```
find "
```

---

```
passed: True
```

```
find footnote
```

---

```
passed: True
```

```
find input{format/i523}
```

---

```
4: \input{format/i523}
```

```
passed: True
```

```
find input{format/final}
```

---

```
passed: False

floats
-----
353: \begin{table}
356: \label{tab:freq}
401: \begin{table}
404: \label{tab:freq}
433: \begin{figure}[!ht]
434: \centering\includegraphics[width=\columnwidth]{images/Figure1.pdf}
}
437: \label{f:Figure1}
461: \begin{figure}[!ht]
462: \centering\includegraphics[width=\columnwidth]{images/Figure2.pdf}
}
465: \label{f:Figure2}
500: \begin{figure}[!ht]
501: \centering\includegraphics[width=\columnwidth]{images/Figure3.pdf}
}
504: \label{f:Figure3}
550: \begin{figure}[!ht]
551: \centering\includegraphics[width=\columnwidth]{images/Figure4.pdf}
}
554: \label{f:Figure4}
599: \begin{figure}[!ht]
600: \centering\includegraphics[width=\columnwidth]{images/Figure5.pdf}
}
602: \label{f:Figure5}
637: \begin{figure}[!ht]
638: \centering\includegraphics[width=\columnwidth]{images/Figure6.pdf}
}
640: \label{f:Figure6}
674: \begin{figure}[!ht]
675: \centering\includegraphics[width=\columnwidth]{images/Figure7.pdf}
}
677: \label{f:Figure7}
714: \begin{figure}[!ht]
715: \centering\includegraphics[width=\columnwidth]{images/Figure8.pdf}
}
719: \label{f:Figure8}
745: \begin{figure}[!ht]
746: \centering\includegraphics[width=\columnwidth]{images/Figure9.pdf}
}
748: \label{f:Figure9}
```

```
766: \begin{figure} [!ht]
767: \centering\includegraphics[width=\columnwidth]{images/Figure10.pdf}
    f}
770: \label{f:Figure10}
795: \begin{figure} [!ht]
796: \centering\includegraphics[width=\columnwidth]{images/Figure11.pdf}
    f}
799: \label{f:Figure11}
```

figures 11

tables 2

includegraphics 11

labels 13

refs 0

floats 13

True : ref check passed: (refs >= figures + tables)

True : label check passed: (refs >= figures + tables)

True : include graphics passed: (figures >= includegraphics)

False : check if all figures are referred to: (refs >= labels)

Label/ref check

89: abusing prescribed opioid medication who also used heroin, shown  
in Figure 1.

333: 386 females). Table 1 shows the raw counts of individual  
substance use by age

392: Table 2 shows the frequency of individuals reporting that they  
had experienced

421: Figure 1 shows the proportion of individuals who reported  
misusing prescription

423: Figure 1 shows the majority of respondents (89 percent) stated  
they had never

426: Figure 1 shows that, of those individuals who reported using  
heroin, the

441: Figure 2 shows the aggregated measure of Opioid Pain Reliever  
misuse and abuse

469: Figure 3 shows the pairplots of demographic features including  
mental health

535: few features. Figure 4 plots the coefficients of logistic  
regression classifier

541: accuracy of 0.98 and test set accuracy of 0.98. Figure 4 shows  
that the

572: Figure 5 shows a partial view of the decision tree classifier of  
heroin use

625: feature importance for a single tree. Figure 6 shows the feature  
importance

665: or by reducing the learning rate. Figure 7 shows that the feature  
importance  
703: Figure 8 plots the coefficients of logistic regression classifier  
for  
705: Figure 8, the features with the highest coefficient values were  
Treatment  
728: of the pruned tree was 0.902 and test set accuracy was 0.902.  
Figure 9 shows  
731: notebook) \cite{classifyPRL}. As Figure 9 shows, the decision  
tree classifier  
756: 0.896, which suggests that the model overfit the data. Figure 10  
shows the  
756: 0.896, which suggests that the model overfit the data. Figure 10  
shows the  
758: PRL Misuse. As Figure 10 shows, several features were identified  
as important  
758: PRL Misuse. As Figure 10 shows, several features were identified  
as important  
784: prescription opioid abuse. Figure 11 shows the feature importance  
for the  
784: prescription opioid abuse. Figure 11 shows the feature importance  
for the  
785: gradient boosting classifier tree. As Figure 11 shows, several  
features were  
785: gradient boosting classifier tree. As Figure 11 shows, several  
features were  
818: used heroin than had not (see Figure 1), which partially supports  
the  
passed: False -> labels or refs used wrong

When using figures use columnwidth  
[width=1.0\columnwidth]  
do not change the number to a smaller fraction

find textwidth

---

passed: True

below\_check

---

WARNING: algorithm and below may be used improperly

126: classification algorithms are considered below.

bibtex

---

label errors

bibtex errors

```
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
Warning--I didn't find a database entry for "granovetter73"
Warning--I didn't find a database entry for "watts98"
Warning--page numbers missing in both pages and numpages fields in herland14
Warning--no number and no volume in johnson11
Warning--page numbers missing in both pages and numpages fields in johnson11
(There were 5 warnings)
```

bibtex\_empty\_fields

---

entries in general should not be empty in bibtex

find ""

---

passed: True

ascii

---

=====

The following tests are optional

=====

Tip: newlines can often be replaced just by an empty line

find newline

---

```
passed: True  
cites should have a space before \cite{} but not before the {
```

```
find cite {  
-----
```

```
passed: True
```

# IoT and Big Data Analytics for Equipment Predictive Health Management (PHM)

Ashok Reddy Singam

Indiana University

711 N Park Ave

Bloomington, Indiana 47408

asingam@iu.edu

Anil Ravi

Indiana University

711 N Park Ave

Bloomington, Indiana 47408

anilravi@iu.edu

## ABSTRACT

The predictive health management (PHM) is an enabling discipline consisting of technologies and methods to assess the reliability of a product in its actual life cycle conditions to determine the advent of failure and mitigate system risk. The PHM system will monitor environmental, operational, and performance related characteristics of the product and gathered data analyzed to assess product health and predict remaining life.

In this application, the industrial rotating equipment such as compressors, vacuum blowers, pumps, and valves etc. are considered to monitor and analyze their operational behavior. The product critical operational parameter data such as vibration, temperature, and load current will be collected from field sensors and analyzed to predict the failure using kNN machine learning classification algorithms. The data will be collected from the field using wireless sensors and stored on the cloud based AWS database server. The product data will be analyzed and made available to all stakeholders to take appropriate preventive actions via web/mobile applications.

## KEYWORDS

i523, HID333, HID337, KNN, IoT, Big Data, Analytics

## 1 INTRODUCTION

The PHM technology can be put within a broader business context by relating it to the Product-Service System (PSS) business model. PSS can be defined as an integrated combination of products and services where the emphasis is put on the ‘sale of use’ rather than the ‘sale of product’. Central to this new business model is a shift from selling a product, and its related spare parts as required, to selling a solution that supports customer needs in the form of a service delivering a fully maintained and useable product [2]. As shown in Figure (1), There are several wireless technologies such as 802.11, cellular, and short distance wireless protocols can be used to collect and send data to the centralized servers. Also, data can be stored in cloud based technologies such as AWS, Microsoft Azure, IBM and Google etc. for processing.

[Figure 1 about here.]

**Problem Statement:** in the manufacturing operations, automotive and other process industries rotating equipment such as pumps, valves, compressors, and blowers are commonly used equipment for various purposes. These equipment are severely suffered from wear and tear, bearing degradation, shaft misalignment, corrosion, and other mechanical breakdowns. Due to the limitations of wireless enabled sensors based data acquisition it was very difficult to collect this data in the past. Also, due to real-time nature of the data

acquisition, it was a huge challenge to store the data locally and process the information for applying machine learning algorithms. All these technological and infrastructure limitations caused industrial equipment health monitoring had become one of the sector businesses are losing the money due to operation shutdowns and unplanned maintenance etc.

**Solution Approach:** with the wireless sensors and cloud based server technologies, it has become possible to deploy hundreds of sensors in the manufacturing plant and collect the data and store with minimal costs. Once the data is stored on the servers with high computing power, machine learning algorithms can be used to process the sensor data to predict the equipment failures with reasonable accuracy. This approach has been named as predictive or prognostics health management of the equipment which is widely available in the recent times due to the availability of technological infrastructure.

The PHM generally combines sensing, collecting, storing and analyzing of environmental, operational, and performance related parameters to assess the health of a product and predict remaining useful life. Assessing the health of a product provides information that can be used to meet several critical goals [1]:

- Providing advance warning of failures
- Minimizing unscheduled maintenance, extending maintenance cycles, and maintaining effectiveness through timely repair actions
- Reducing the life cycle cost of equipment by decreasing inspection costs, downtime, and inventory
- Improving qualification and assisting in the design and logistical support of fielded and future systems

The PHM is not a new concept, however, with the advent of sensors, machine learning algorithms, and computing capacity of the servers it has become more prevalent in the recent days. In this application, an attempt has been made to prove the concept of simple PHM implementation and use in real world applications. The application can be re-architected to address more complex products/systems with considerations of scalability, performance, cost and reliability. The limitations of the current application are described in the end of this report.

The parameter monitoring and the analysis of acquired data using prognostic models are fundamental steps for the PHM methods. The sensors are the essential devices used to monitor parameters and obtain long-term accurate information to provide anomaly detection, fault isolation, and rapid failure prediction [1].

Firstly, PHM requires monitoring a large number of product parameters to evaluate the health of a product. Depending on the

complexity of the monitored product, it is possible to monitor thousands of parameters in the entire life cycle of the product to provide the information required by PHM. These parameters include operational and environmental loads as well as the performance conditions of the product, for example, temperature, vibration, shock, pressure, acoustic levels, strain, stress, voltage, current, humidity levels, contaminant concentration, usage frequency, usage severity, usage time, power, and heat dissipation. In each case, a variety of monitoring features such as magnitude, variation, peak level, and rate of change may be required in order to obtain characteristics of parameters.

In this application, commonly used equipment in industrial and automobile operations such as air compressors, vacuum blowers, and smart valves are considered for analysis. The critical operational parameters of these products will be collected using applicable sensors from the field and fed to a database at regular intervals.

In general design, the frequency of data collection and storage depends on the number of parameters to be analyzed, cost of the system and operational behavior of the equipment. For this application, since products with rotating parts are considered, the critical parameters that would define the health of the equipment are: input or load current, internal ambient temperature, and vibration of the equipment.

The PHM application design process is shown in Figure (2), which describes various steps of the processes involved. For the implementation of this project, the sensor generated data is simulated using SQL scripts due to development time constraints. However, a detailed step-by-step approach is provided if we need to plug-in the sensor modules in to the application.

[Figure 2 about here.]

**Data Acquisition Stage:** It is required to have a description of a machine behaving normally that can be used for early detection of anomalies. This calls for a proper characterization of machine health. As part of this process, various methods are identified to extract health information from vibration measurements and investigate strengths and weaknesses of these methods as health descriptors. This stage will be the core part of PHM application where vibration data were experimentally obtained from a compressor using triaxial accelerometer to collect transverse, longitudinal and vertical axes vibration signals. For the experimental data collection, ACC301A triaxial accelerometer and National Instruments data acquisition system was used. A total of 8 parameters

- (1) Input Current
- (2) Input Voltage
- (3) Internal ambient temperature
- (4) External ambient temperature
- (5) Transverse vibration
- (6) Longitudinal vibration
- (7) Vertical axis vibration
- (8) Acquisition time

were captured at one second rate, which generated about 65000 records. This data has been analyzed for identifying the feature classification.

**Pre Processing stage:** During this stage collected data will be filtered and processed for accuracy in order to adapt them to subsequent feature extraction stage. In this application, all the

pre-processing has been done manually to validate the accuracy of the data based on the system conditions. Since spectral analysis of vibration signals are not done ( one of the limitations for this application, captured in the end), the data generated from the compressor is considered as the primary frequency of the equipment (which is isolated from the rest of the attachments).

**Feature extraction and selection stage:** during this stage domain specific vibration spectral analysis has been performed but only considered time-domain behavior for various system operational conditions such as increased load, modified input voltage, and modified external ambient temperature etc. Based on the response of the machine vibration to various external conditions were noted down. This data is used to identify the following feature vectors.

- NORMAL OPERATION AT 30 DEG CENTIGRADE
- OVER CURRENT FAULT OPERATION
- OVER TEMPERATURE FAULT OPERATION
- INPUT OVER VOLTAGE FAULT OPERATION
- ABNORMAL OPERATION AT 30 DEG CENTIGRADE
- BEARING DEGRADATION OPERATION

**kNN classification stage:** this stage is core part of the PHM application, which will predict the unknown test data to be classified in to a known label based on the training data set using nearest neighbor algorithm.

**Classifier performance evaluation stage:** this stage will be used to evaluate the classifier accuracy of prediction. In this application, k-fold cross-validation method has been used to perform the evaluation.

The data is generated and made available in Oracle database on AWS cloud to perform analysis. The application developed in this project will consist of the following components:

- Sensor Data Generator
- Machine Learning Algorithm
- Big Data and IoT
- PHM Dashboard
- Decision Alerts
- Application Script

The following sections will describe the architectural and design aspects of the PHM system implementation in detail.

## 2 PROGNOSTICS MODEL EVALUATION

[6] The prediction is typically performed only after the *health* of the component or system deteriorates beyond a certain threshold. In this application, faults and failures are identified in the training data set. The faults identified are: Over current fault, over temperature fault and over voltage fault. If over current fault is occurred, the equipment will tend to draw higher current than nominal values which if continued further several times eventually leads to a permanent failure of the equipment. In this application, when motor bearing starts degrading, the first observation will be over current followed by over temperature conditions. Often times, that threshold is tripped because a fault occurs. A fault is a state of a component or system that deviates from the normal state such that the integrity of the component is outside of its required specification. A fault does not necessarily imply that the overall system does not operate anymore; however, the damage that characterizes the fault often grows under the influence of operations to a failure. The

latter is the state at which the component or system does not meet its desired function anymore. It is the task of prognostics to estimate the time that it takes from the current time to the failed state, conditional on anticipated future usage. This would give operators access to information that has significant implications on system safety or cost of operations. Where safety is impacted, the ability to predict failure allows operators to take action that preserves the assets either through rescue operation or through remedial action that avert failure altogether. Where minimizing cost of operations is the primary objective, predictive information allows operators to avert secondary damage, or to perform maintenance in the most cost-effective fashion. Often times, there is a mix of objectives that need to be optimized together, sometimes weighted by different preferences.

As emphasized above, predictive models evaluation needs to take domain specificities into account. Such specificities cover two aspects: capability of failure prediction and TTF estimation. From the point of view of TTF, it is desirable that a predictive model can generate alerts in a *targeted* time window prior to a failure. A model that predicts a failure too early leads to non-optimal component use [7] which will impact the reliability or availability of the system.

[Figure 3 about here.]

As shown in the Figure (3), the time to failure prediction will be estimated based on the classified result data set and alert the stakeholders to take relevant actions. The target alert zone will be identified based on the abnormal behavior of the equipment over the period.

### 3 APPLICATION DESIGN ANALYSIS

The PHM application in this project considered to use rotating equipment temperature, load current and vibration data for analyzing and predicting the future operational behavior. Vibration signals from rotating components are usually analyzed in the frequency domain, because significant peaks in the signal spectrum appear at frequencies that are related to the rotation frequency of the component. In this application, only time domain parameters with peak vibration magnitudes irrespective of the frequency component. The training data set consists of normal, abnormal, and fault conditions vibration patterns describes the system characteristics from which its status can be estimated. The PHM application for industrial equipment machine failure detection problem directly correlates to the pattern classification problem. From the vibration data collected, each accelerometer will output values of X, Y, and Z data then using a KNN we can similarly identify which vibration parameter(s) determines problems in our machines, or *likely to experience failure*. The typical defects or failures that can be detected are: machine imbalance, shaft misalignment, pumps cavitation, structural and rotating looseness, early stage bearing wear, gear teeth problems, and other high-frequency defects.

This application used *Sensor Data Gen* SQL script module to generate the sensor data and store in Oracle database on AWS. This is the critical module as we have not used the real data collection from the field. However, the sensor hardware and necessary environment to generate the data is identified and experimented to work with. A brief description about the hardware is provided in the end of this report.

The PHM application is designed such that the fundamental concepts can be verified to open a discussion on limitations, performance, scalability, ROI and reliability of the system.

The following sections describe the application design components with necessary implementation details:

#### 3.1 Sensor Data Generator

The SQL data generator script is designed to generate training data as well test data for this application with following eleven parameters: Acquisition time, equipment name, part number, serial number, internal ambient temperature, external ambient temperature, input voltage, input current, and vibration data for x, y, and z axes. The following database design architecture followed for Sensor Data Gen module:

- Sensor Data Generator PL SQL Objects
  - Tables
    - \* SENSOR TRAIN DATA for storing training data
    - \* SENSOR TEST DATA for storing testing data
  - Views
    - \* SENSOR TRAIN DATA VIEW: Created View on top of SENSOR TRAIN DATA with logic to translate string label data into numbers
    - \* SENSOR TEST DATA VIEW Created View on top of SENSOR TEST DATA with logic to translate string label data into numbers
  - Packages BIG DATA 503 PRJ PKG
    - \* Generate Test Set: Pl/Sql procedure to insert sensor test data into SENSOR TRAIN DATA table
    - \* Generate Train Set: Pl/Sql procedure to insert sensor train data into SENSOR TRAIN DATA table
    - \* Delete Data Set: Pl/Sql procedure to delete all training and test data.
    - \* Update Test Data Labels: Pl/Sql procedure to update SENSOR TRAIN DATA table with KNN algorithm predicted label values

#### 3.2 Machine Learning Algorithm

3.2.1 *Classifier evaluation.* Typical classifier evaluation methods include ROC Curves, Reject Curves, Precision-Recall Curves, and Statistical Tests. The statistical tests consists of following methods to perform evaluation:

- Estimating the error rate of a classifier
- Comparing two classifiers
- Estimating the error rate of a learning algorithm
- Comparing two algorithms

Out of the listed statistical tests, the error rate estimation method is used in this application to evaluate the performance. An experimental data is used to estimate the error rate or accuracy of various classifiers. Then a comparison has been made to choose the classifier to use in the application.

The following list of performance for various classifiers is observed during the accuracy calculation. The same set of training

data has been used for all the classifiers, which has resulted the following performance. All values are mentioned in percents between 0 to 1, 1 means 100 percent accuracy.

- LogisticRegression: 0.963636
- KNN: 0.981818
- DecisionTreeClassifier: 0.964394
- SVM: 0.972727

Based on the performance as shown in Figure (4), kNN has been selected to use for this application.

[Figure 4 about here.]

**3.2.2 *k Nearest Neighbor - kNN*.** [8] In this application, neighbors-based classification is chosen to classify the unknown instance to the known trained labels. Neighbors-based classification does not attempt to construct a general internal model, but simply stores instances of the training data.

Classification is computed from a simple majority vote of the nearest neighbors of each point: a query point is assigned the data class which has the most representatives within the nearest neighbors of the point.

KNN falls in the supervised learning family of algorithms. Informally, this means that we are given a labelled dataset consisting of training observations  $(x, y)$  and would like to capture the relationship between  $x$  and  $y$ . More formally, our goal is to learn a function

$$h : X - \rightarrow Y$$

so that given an unseen observations  $x$ ,  $h(x)$  can confidently predict the corresponding output  $y$ .

In the classification setting, the K-nearest neighbor algorithm essentially boils down to forming a majority vote between the  $K$  most similar instances to a given unseen observation. The number of neighbors for  $k$ -nearest neighbors ( $k$ ) can be any value less than the number of rows from dataset. Looking at only a few neighbors makes the algorithm perform better but the less similar the neighbors, the worse the prediction will be. Similarity is defined according to a distance metric between two data points. A popular choice is the Euclidean distance given by:

$$d(x, x') = \sqrt{(x_1 - x'_1)^2 + (x_2 - x'_2)^2 + \dots + (x_n - x'_n)^2}$$

But other measures can be more suitable for a given setting and include the Manhattan, Chebyshev and Hamming distance. An alternate way of understanding KNN is by thinking about it as calculating a decision boundary (i.e. boundaries for more than 2 classes) which is then used to classify new points.

Another characteristic of KNN is it is instance based learning algorithm. Means it doesn't explicitly learn a model. Instead, it chooses to memorize the training instances which are subsequently used as knowledge for the prediction phase. It is also means the algorithm does not build a model until the time that a prediction is required. It is also lazy learning because it only does work at the last second. This has the benefit of only including data relevant to the unseen data, called a localized model. A disadvantage with lazy model is it can be computationally expensive to repeat the same or similar searches over larger training datasets.

In the application design, sci-kit open source python libraries are used for implementing the kNN algorithms. Scikit is built on NumPy, SciPy, and matplotlib. The k-neighbors classification in KNeighborsClassifier is the more commonly used of the two techniques. The optimal choice of the value  $k$  is highly data-dependent: in general a larger  $k$  suppresses the effects of noise, but makes the classification boundaries less distinct.

The sklearn.neighbors.KNeighborsClassifier class has the following methods, which are used in the application design:

- fit: Fit the model using  $X$  as training data and  $y$  as target values.
- get params: Fit the model using  $X$  as training data and  $y$  as target values.
- kneighbors: Finds the  $K$ -neighbors of a point.
- kneighbors graph: Computes the (weighted) graph of  $k$ -Neighbors for points in  $X$ .
- predict: Predict the class labels for the provided data.
- predict\_proba: Return probability estimates for the test data  $X$ .
- score: Returns the mean accuracy on the given test data and labels.
- set params: Set the parameters of this estimator.

**3.2.3 *K-fold cross-validation*.** To estimate the test error in the model, a cross-validation approach followed in which a subset of the training set will be holding out from the fitting process. This subset, called the validation set, can be used to select the appropriate level of flexibility of our algorithm. There are different validation approaches that are used in practice, and we will be exploring one of the more popular ones called **k-fold cross validation**. The k-fold cross validation (the  $k$  is totally unrelated to  $K$ ) involves randomly dividing the training set into  $k$  groups, or folds, of approximately equal size. The first fold is treated as a validation set, and the method is fit on the remaining  $k - 1$  folds. The misclassification rate is then computed on the observations in the held-out fold. This procedure is repeated  $k$  times; each time, a different group of observations is treated as a validation set. This process results in  $k$  estimates of the test error which are then averaged out.

In this application, an average k-fold cross validation accuracy of 0.99 percent achieved, which is explained in the appendix section of the report. Figure(5) shows Classification and Confusion report output obtained from the KNN model we used for this project.

[Figure 5 about here.]

### 3.3 Big Data and IoT

In PHM systems big data is characterized by one or more 3Vs: volume, velocity and variety due to streaming of real-time IoT sensors. Most of the IoT systems present challenges in combinations of velocity and volume. The important feature of the IoT application is that by observing the behavior of "many things" it will be possible to gain important insights, optimize processes, etc. This requires storing all the events (velocity and volume challenge) to run analytical queries over the stored events and perform analytics (data mining and machine learning) over the data to gain insights. In general PHM applications, data will be collected through field sensors at specific rate which accounts for large amount of data per day in the order of multi-million records. This data will be stored

in any NOSQL or RDBMS based database for storage and processing. Since the big data infrastructure is much reliable and available widely from multiple vendors, it would help to build complex PHM systems with large number of feature vectors for classification.

In this application, for the demonstration of the concept, the real vibration data from the compressor equipment has been collected via accelerometer sensors. This vibration data has been analyzed in time domain and established the labels based on the compressor design performance parameters. Later, this data analysis is used to design a SQL script for generating training and test data sets. However, the real-time PHM system will have continuous streaming of data coming from hundreds of devices at faster rates (in the order of milliseconds to tens of seconds). This data needs to be captured by reliable and scalable platforms such as AWS IoT or similar and use the machine learning algorithms to classify the unknown data.

### 3.4 PHM Dashboard

Once all the test data set has been classified in to appropriate labels, the prediction of the failure can be performed based on the trending of the equipment behavior over the period. In order to understand the equipment performance insight, following queries will be used on the classified data:

- Faults Reported by Equipment Part Number
- Faults Reported by Serial Number
- Abnormal Behavior by Equipment Part Number
- Abnormal Behavior by Serial Number over the period range

There can be more application specific information obtained from classified data set to take various decisions. Figure(6), figure(7) and figure(8) show various PHM data analytics for this project. Figure(6) displays all the serial numbers of equipment 1 with bearing degradation problems. The X axis gives serial numbers while the Y axis gives number of occurrences of bearing degradation for that particular serial number. Similarly figure(7) and figure(8) give the details of over temperature and over current faults of various serial numbers.

As part of data visualization, result data file is queried based on the analytics metrics interested. The python matplotlib package has been used to draw the charts as needed for showing the analytics. In real world application, a more sophisticated business intelligence tools such as Tableau, Microsoft BI, and Amazon Quick Sight can be used to show the PHM dashboards. These dashboards are targeted for business users so that they will be able to customize the views, add filters and drill down in to specific information as needed.

Sample screen shots for the following scenarios are included from python code output:

- Bearing Degraded Serial numbers for Equipment Part Number1: Figure (6)
- Over Temperature Fault Serial numbers for Equipment Part Number1: Figure (7)
- Over Current Fault Serial numbers for Equipment Part Number1: Figure (8)

[Figure 6 about here.]

[Figure 7 about here.]

[Figure 8 about here.]

### 3.5 Decision Alerts

Once the results data set has been generated by the prediction algorithm, and then based on the analytics queries, PHM system can send out the alert messages to appropriate stakeholders. The typical messages include the following s minimum:

- SN 10002: Faulted X times on over temperature in last Y days, needs maintenance to clean the filter
- SN 10005: Consistently indicating bearing degradation from last X days, needs lubrication maintenance
- SN 10009: Consistently drawing over current from last X days, needs mechanical load maintenance

In this application all the unseen test data is classified and labeled in the result data file. However, in real world application, along with the dashboards a comprehensive alerting capabilities can be built. The application checks for out of range alert conditions on selected incoming report parameters, looking for warning or alarm conditions that are higher or lower than expected under normal operating conditions.

### 3.6 Application Code Development

Code required for this project is divided into two categories

- Python Coding: We used **Anacoda Navigator ver 1.6.9** [3] installation on windows7 laptop which includes Jupyter notebook application for python coding. **Anacoda Navigator** also supports multiple installation and management of python environments using gui interface. The location of the Jupyter notebook file we developer for this project is mentioned in appendix b.
- PL/SQL Coding: This application specific training/test sensor data has been created/generated on AWS cloud database using pl/sql coding which will be accessed during the run time by python notebook. We used **Orcale Sql Developer** [4] for pl/sql coding.

### 3.7 Python - Oracle Interface

For this project we used python library called **cxOracle** [5] to enable access to Oracle Database. It can be installed easily using **pip** and it supports both Python 2 and 3. This library supports:

- SQL and PL/SQL Execution. The underlying Oracle Client libraries have significant optimizations including compressed fetch, pre-fetching, client and server result set caching, and statement caching with auto-tuning.
- Extensive Oracle data type support, including large object support like CLOB and BLOB)
- Batch operations for efficient INSERT and UPDATEs

In the following scenarios we used **cxOracle** libraries:

- To read sensor training data from Oracle database
- To read sensor test data from Oracle database
- After classification of labels, to update test data with classified labels

## 4 APPLICATION LIMITATIONS

The PHM application developed in this project has several limitations. Typically, PHM applications suffer from the prediction accuracy rate which influences the ability to take decisions that

will have broader impact on the business operations and financial aspects. However, with the advanced machine learning classifiers and model evaluation methods this can be addressed to achieve reasonable confidence. Following are some of the limitations of this application, which can be addressed and improved in large real-time PHM systems.

**Data acquisition hardware:** in this application the data is not collected from real-time sensors for voltage, current, temperature and vibration data. There will be inherent accuracy in the raw data generated by SQL script. However, a sample vibration dataset has been collected from the field, which is used as basis to generate the simulated data set.

**Feature extraction analysis:** the equipment performance parameters of interest need to be down selected from large set of incoming parameter data.

When analyzing vibration data in the time domain only few parameters are available in quantifying the strength of a vibration profile: amplitude, peak-to-peak value, and RMS. The amplitude is valuable for shock events but it does not take into account the time duration and thus the energy in the event. The same is true for peak-to-peak with the added benefit of providing the maximum excursion of the wave, useful when looking at displacement information, specifically clearances. The RMS value is generally the most useful because it is directly related to the energy content of the vibration profile and thus the destructive capability of the vibration.

This requires in-depth domain specific analysis, in this case a detailed mathematical modeling of vibration spectral analysis to precisely select the features and corresponding behavior patterns. Such analytical data should be used for training data feature set. In this application, a primitive approach of time-domain analysis of vibration magnitudes used for determining features. However, in real application these features need to be mathematically analyzed to identify the features that represent the system behavior as close as possible.

**Model accuracy and scoring :** the kNN algorithm used in this application validated using k-cross fold cross-validation. There are several other model evaluation and scoring methods such as accuracy (or error rate), True Positive Rate (TPR), False Positive Rate (FPR), False Negative Rate (FNR), True Negative Rate (TNR), sensitivity etc. These metrics provide a simple and effective way to measure the performance of a classifier. This application can be further improved by applying more performance measurement methods to increase the effectiveness of the algorithm design.

**Scalability:** the application designed in this project is very primitive to understand the basic concepts of PHM and kNN classifier implementation. This application cannot be used for PHM application in business use. To implement real world PHM application, a more comprehensive design needed by considering modularity, service oriented architecture, large number of sensors integration, big data and analytics integration etc.

## 5 RECOMMENDATIONS

**Generality:** since each rotating equipment vibrates in a different manner, a monitoring method needs to be retrained for each machine. The training on several repeated measurements on several

similar equipment in several operating modes may allow for a more general monitoring method.

**Feature extraction and dimensionality:** In this application it has been assumed that a proper feature (selection) has been chosen, such that the feature dimensionality is not too high. If the data lies in a subspace, application of an initial dimensionality reduction may be a good idea. It is highly recommended to perform spectral analysis on vibration data and identify various fault frequencies and their sources. This would help to extract the optimized feature vectors for the given application followed by selecting the more relevant ones.

**Model evaluation:** classifier accuracy and effectiveness will be varied based on the test data set. It is highly recommended to evaluate multiple models with appropriate test data to choose the best classifier for the given application.

**Domain specific modeling:** It is highly recommended to perform more and more domain-oriented feature vector analysis to meet the needs of predictive model evaluation for PHM applications. Domain-oriented approaches helpful and useful in evaluating classifier for applications. Generic evaluation methods could help developers in investigating overall performance of a model from the statistical viewpoint at the initial stage of model development. Domain-oriented approaches should be further used to evaluate the usefulness and business value [7].

## 6 CONCLUSION

In this project, the problem statement around industrial rotating equipment maintenance is described and solution principle to address the same using PHM concept is defined, experimented and results are discussed. Since this application is developed to prove the only concept but not the complete solution a section with limitations and recommendations for real world system development is described. Overall, PHM application with kNN classifier algorithm and cross validation accuracy of 0.99 percent has been implemented, verified and results are analyzed for business decisions.

## ACKNOWLEDGMENTS

The authors would like to thank professor Gregor von Laszewski and his team for providing *LaTex* templates, assistance with the *JabRef* tool and in helping to fix compile issues with latex and yaml formats.

## REFERENCES

- [1] Shunfeng Cheng, Michael H. Azarian, and Michael G. Pecht. 2010. Sensor Systems for Prognostics and Health Management. (2010), 24 pages. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3247731/pdf/sensors-10-05774.pdf>
- [2] Tonci Grubic, Ian Jennions, and Tim Baines. 2009. The Interaction of PSS and PHM - a mutual benefit case. (2009), 10 pages. [https://www.phmsociety.org/sites/phmsociety.org/files/phm\\_submission/2009/phmc\\_09\\_49.pdf](https://www.phmsociety.org/sites/phmsociety.org/files/phm_submission/2009/phmc_09_49.pdf)
- [3] Anaconda Inc. 2017. Anaconda Python Data Science platform. (2017). <https://www.anaconda.com/what-is-anaconda/>
- [4] Oracle. 2017. Oracle SQL Developer. (2017). <http://www.oracle.com/technetwork/developer-tools/sql-developer/overview/index.html>
- [5] Oracle OTN. 2005. Using Python With Oracle Database 11g. (2005). <http://www.oracle.com/technetwork/articles/dsl/python-091105.html>
- [6] Abhinav Saxena, Jose Celya, Bhaskar Saha, Sankalita Saha, and Kai Goebel. 2009. Sensor Systems for Prognostics and Health Management. (2009), 16 pages. <https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20100023445.pdf>
- [7] Chunsheng Yang, Yanni Zou, Jie Liu, and Kyle R Mulligan. 2014. Predictive Model Evaluation for PHM. (2014), 11 pages. [https://www.phmsociety.org/sites/phmsociety.org/files/phm\\_submission/2014/ijphm.14.019.pdf](https://www.phmsociety.org/sites/phmsociety.org/files/phm_submission/2014/ijphm.14.019.pdf)

- [8] Kevin Zakka. 2016. A Complete Guide to K-Nearest-Neighbors with Applications in Python and R. (2016). <https://kevinzakka.github.io/2016/07/13/k-nearest-neighbor/>

## A WORK BREAKDOWN

### A.1 HID 333:Anil Ravi

- Identified Project topic.
- Created architecture of the application.
- Ran experimental test to collect vibration data
- Extracted and analyzed feature vectors
- Studied, designed and reviewed kNN algorithm
- Created draft project report
- Reviewed the draft project report.

### A.2 HID 337:Ashok Reddy Singam

- Implemented sensor data generation SQL script.
- Implemented kNN algorithm in Python
- Implemented k-fold cross validation design
- Created data analytics charts
- Reviewed the draft project report.

## B CODE REFERENCE

All code, notebooks and files for this project can be found in the github repository: <https://github.com/bigdata-i523/hid337/blob/master/project/jupyter>

#### LIST OF FIGURES

1	System Architecture	9
2	PHM Design Process	9
3	Time relation between alert time and failure time	10
4	Algorithm performance	10
5	KNN Classification and Confusion matrix report	11
6	Bearing Degradation	11
7	Over Temperature Fault	12
8	Over current Fault	13

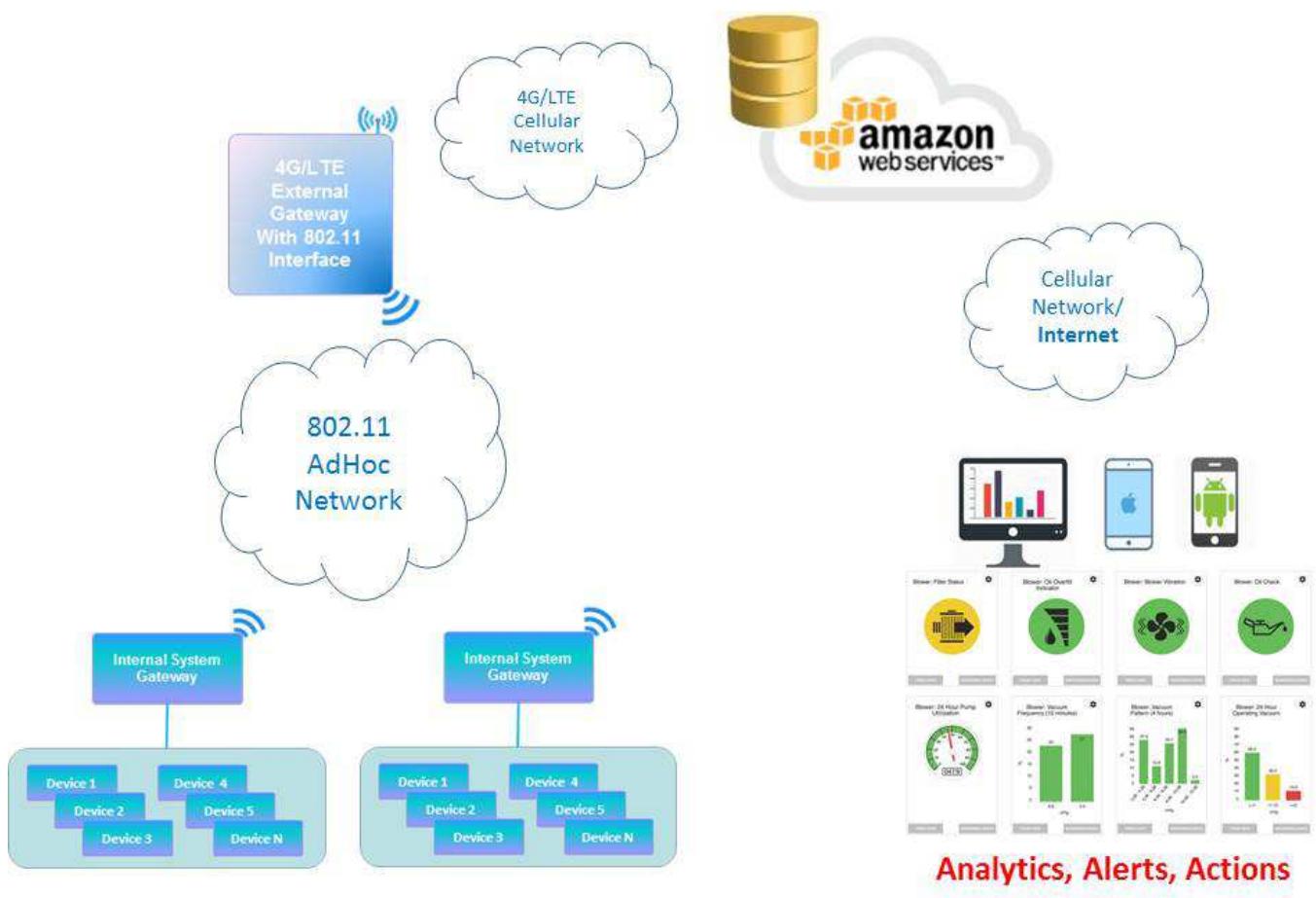


Figure 1: System Architecture

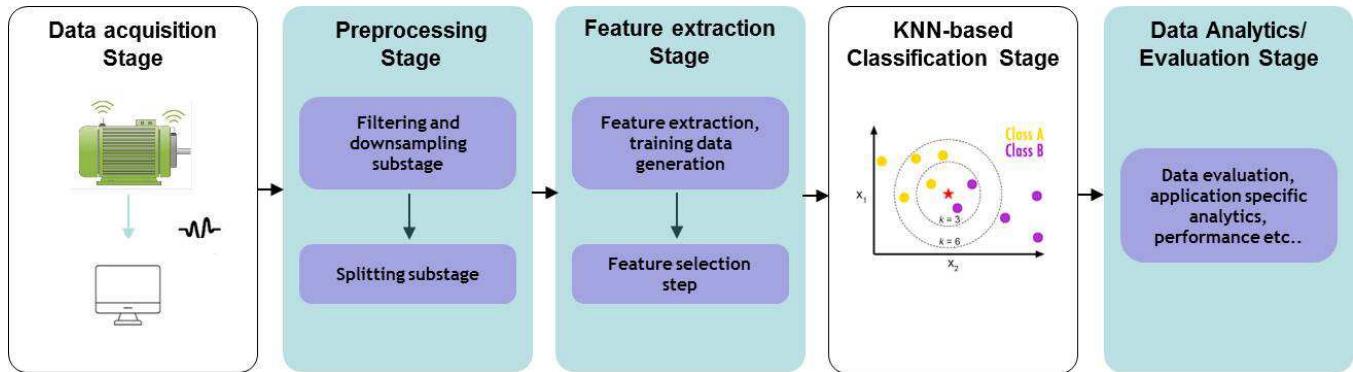


Figure 2: PHM Design Process

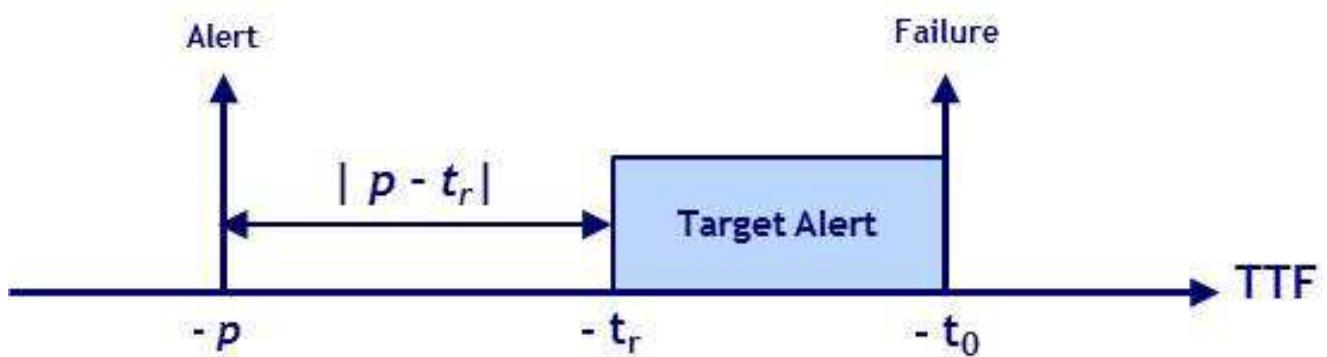


Figure 3: Time relation between alert time and failure time

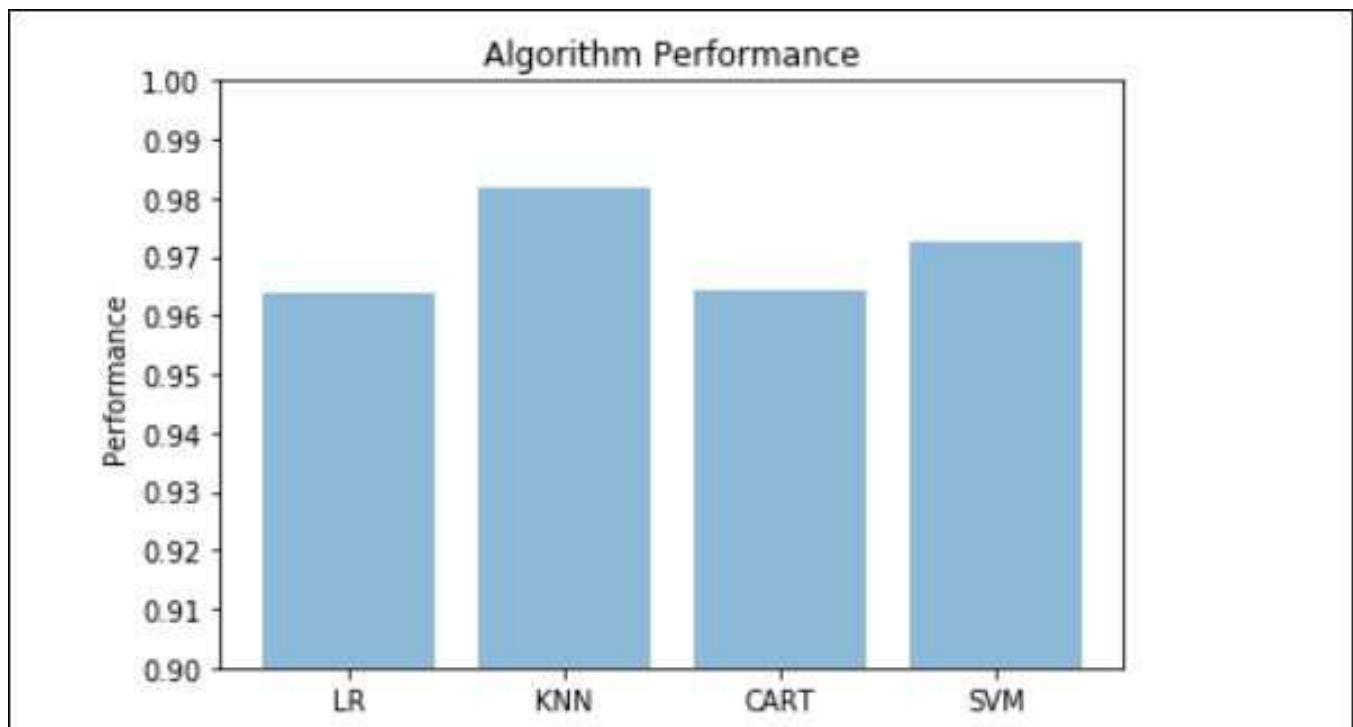


Figure 4: Algorithm performance

	precision	recall	f1-score	support
ABNORMAL_OP_30_DEG_C	1.00	1.00	1.00	997
BEARING_DEGRADE_OP	1.00	1.00	1.00	100
INPUT_OVER_VOLT_FAULT_OP	0.99	0.96	0.98	1059
NORMAL_OP_30_DEG_C	0.96	0.99	0.97	931
OVER_CURRENTFAULT_OP	1.00	1.00	1.00	1005
OVER_TEMP_FAULT_OP	1.00	1.00	1.00	1130
avg / total	0.99	0.99	0.99	5222
[[ 997 0 0 0 0]				
[ 0 100 0 0 0]				
[ 0 0 1018 41 0]				
[ 0 0 10 921 0]				
[ 0 0 0 0 1005]				
[ 0 0 0 0 1130]]				

Figure 5: KNN Classification and Confusion matrix report

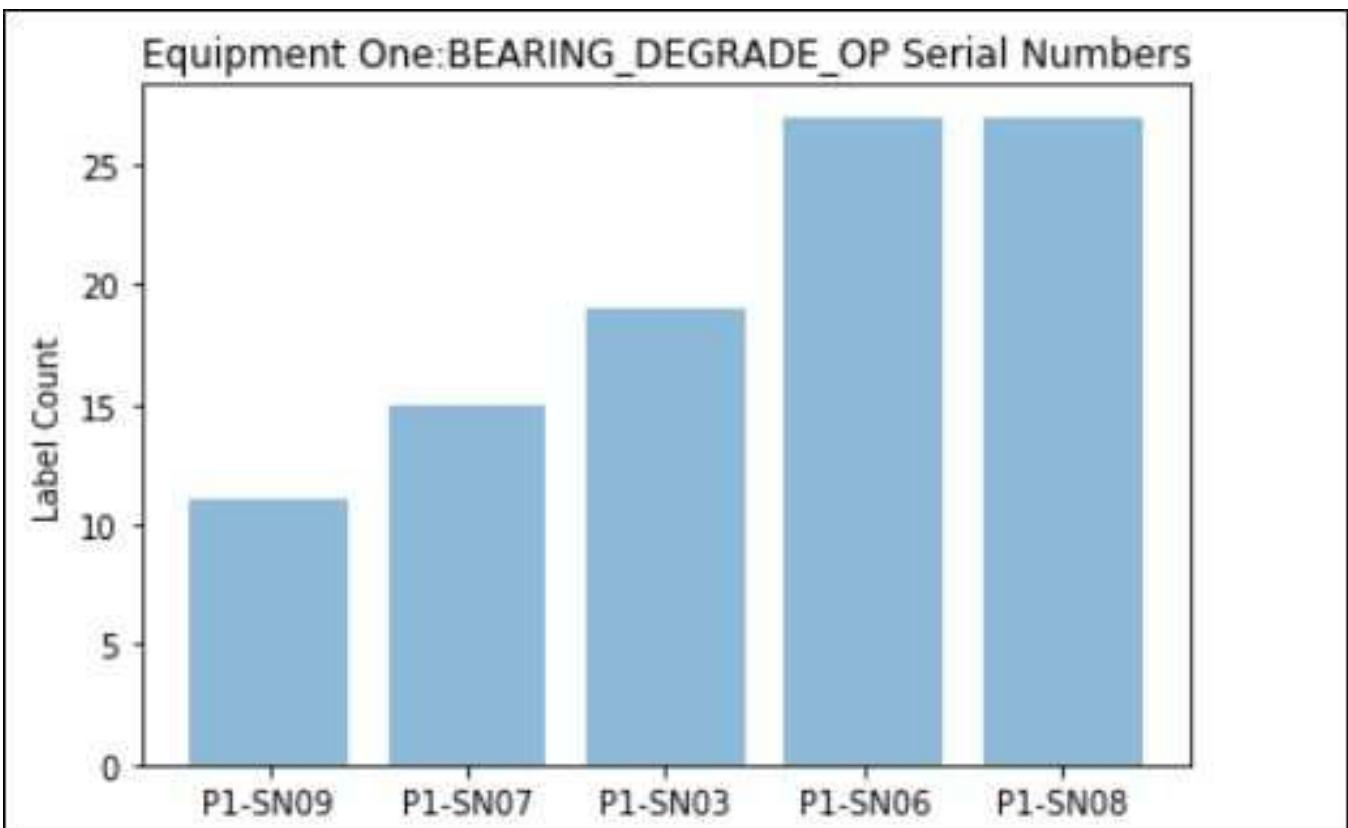


Figure 6: Bearing Degradation

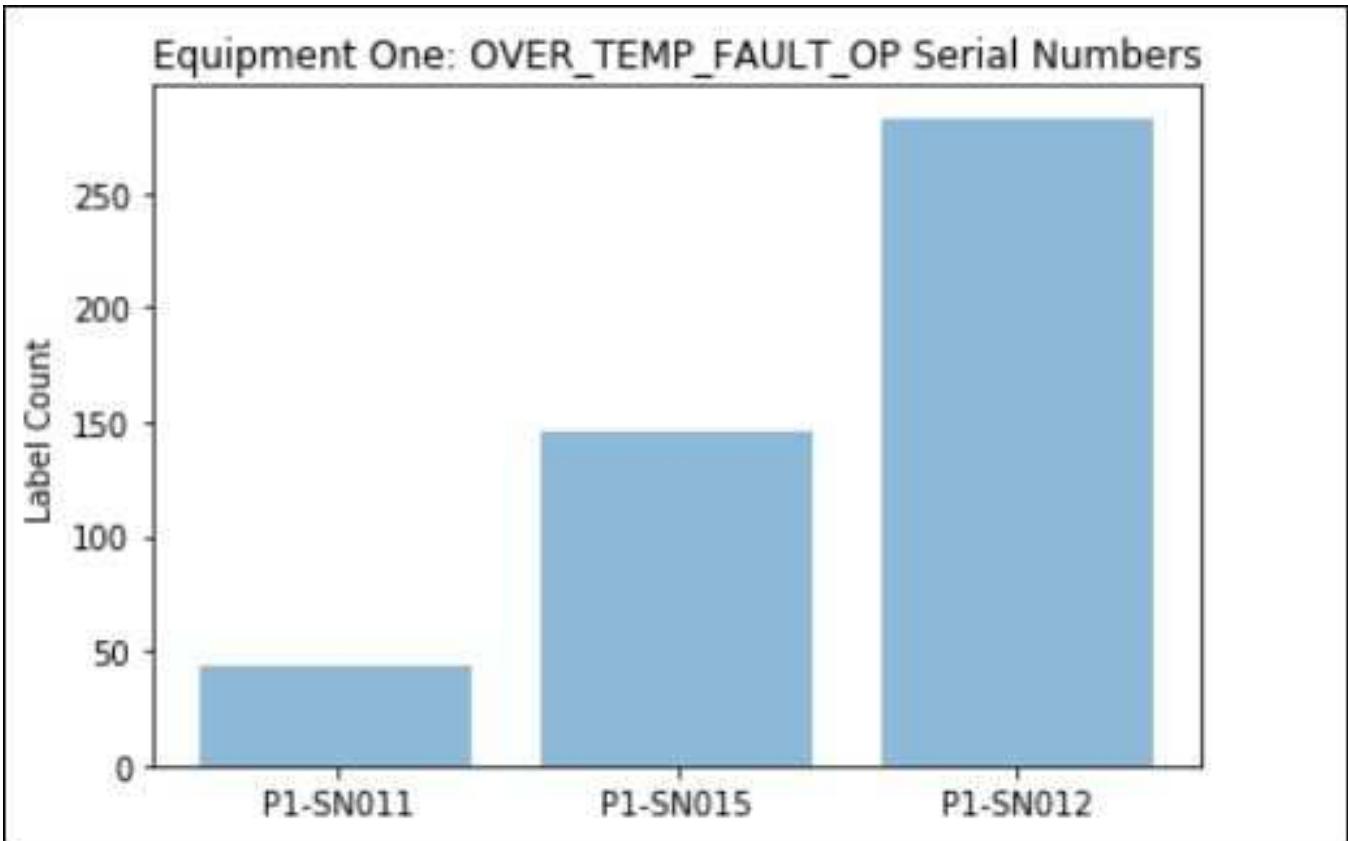


Figure 7: Over Temperature Fault

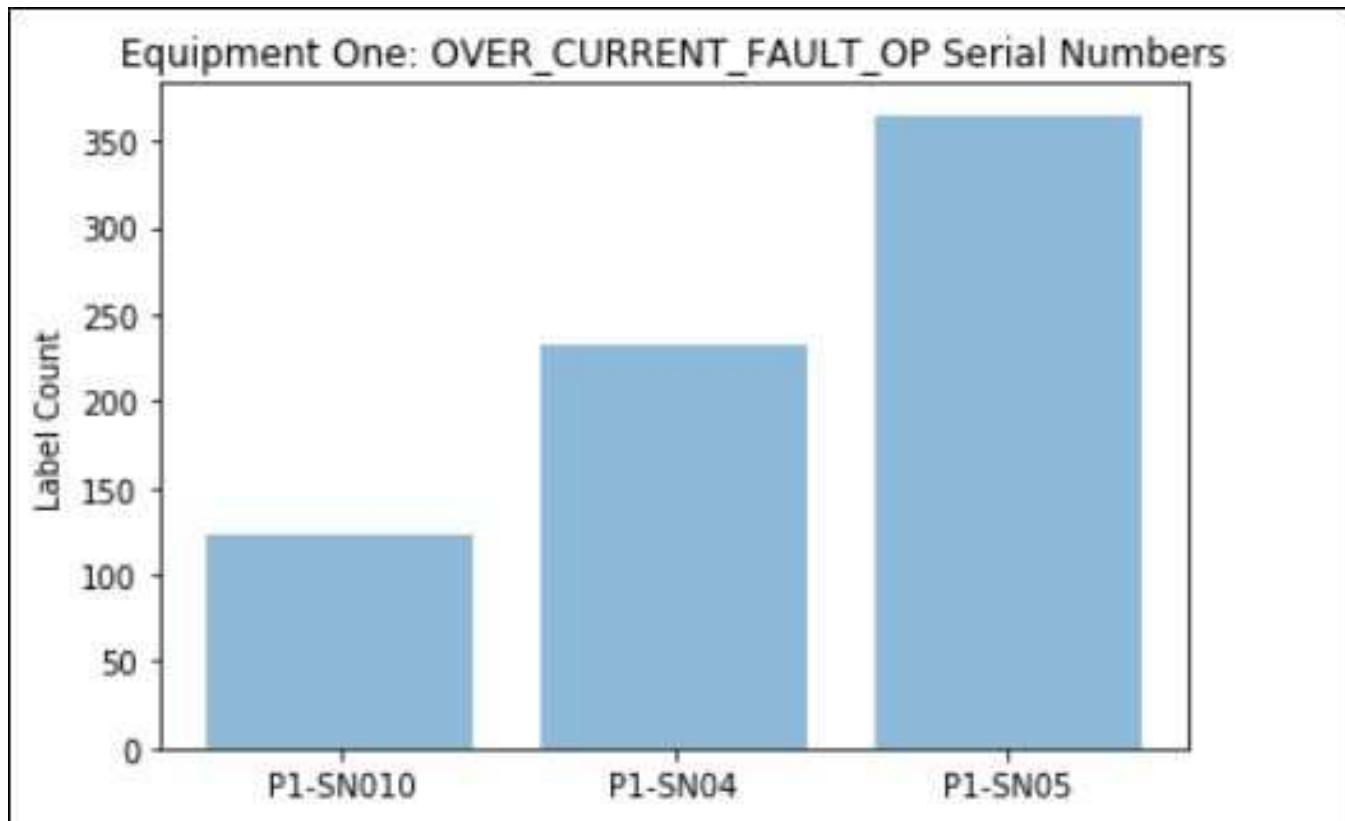


Figure 8: Over current Fault

```
bibtext report
```

```
=====
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
```

```
bibtext _ label error
```

```
=====
bibtext space label error
```

```
=====
bibtext comma label error
```

```
=====
latex report
```

```
=====
[2017-12-11 13.31.11] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflatex)
bookmark level for unknown defaults to 0.
The anchor of a bookmark and its parent's must not be the same. Added a new anchor.
Typesetting of "report.tex" completed in 1.0s.
```

```
=====
Compliance Report
```

```
=====
name: Ashok Reddy Singam
hid: 337
paper1: Nov 01 17 100%
paper2: Nov 06 17 100%
project: Dec 04 17 100%
```

```
=====
yamlcheck
```

```
=====
wordcount
```

```
-----  
13  
wc 337 project 13 5478 report.tex  
wc 337 project 13 5407 report.pdf  
wc 337 project 13 281 report.bib  
  
find "  
-----  
  
passed: True  
  
find footnote  
-----  
  
passed: True  
  
find input{format/i523}  
-----  
  
3: \input{format/i523}  
  
passed: True  
  
find input{format/final}  
-----  
  
passed: False  
  
floats  
-----  
  
34: The PHM technology can be put within a broader business context by relating it to the Product-Service System (PSS) business model. PSS can be defined as an integrated combination of products and services where the emphasis is put on the \lq sale of use \rq rather than the \lq sale of product \rq. Central to this new business model is a shift from selling a product, and its related spare parts as required, to selling a solution that supports customer needs in the form of a service delivering a fully maintained and useable product \cite{Tonci2009}. As shown in Figure (\ref{fig:Figure1}), There are several wireless technologies such as 802.11, cellular, and short distance wireless protocols can be used to collect and send data to the centralized servers. Also, data can be stored in cloud based technologies such as AWS, Microsoft Azure, IBM and Google etc. for processing.
```

```

37: \begin{figure}
38: \includegraphics[width=1.0\columnwidth]{images/system_architecture}
39: \caption{System Architecture} \label{fig:Figure1}
66: The PHM application design process is shown in Figure
    (\ref{fig:Figure2}), which describes various steps of the
    processes involved. For the implementation of this project, the
    sensor generated data is simulated using SQL scripts due to
    development time constraints. However, a detailed step-by-step
    approach is provided if we need to plug-in the sensor modules in
    to the application.
70: \begin{figure}
71: \includegraphics[width=1.0\columnwidth]{images/phm_process_1}
72: \caption{PHM Design Process} \label{fig:Figure2}
126: \begin{figure}
127: \includegraphics[width=1.0\columnwidth]{images/ttf_1}
128: \caption{Time relation between alert time and failure time}
    \label{fig:Figure3}
131: As shown in the Figure (\ref{fig:Figure3}), the time to failure
    prediction will be estimated based on the classified result data
    set and alert the stakeholders to take relevant actions. The
    target alert zone will be identified based on the abnormal
    behavior of the equipment over the period.
195: Based on the performance as shown in Figure (\ref{fig:Figure4}),
    kNN has been selected to use for this application.
197: \begin{figure}
198: \includegraphics[width=1.0\columnwidth]{images/algperformance}
199: \caption{Algorithm performance} \label{fig:Figure4}
245: In this application, an average k-fold cross validation accuracy
    of 0.99 percent achieved, which is explained in the appendix
    section of the report. Figure(\ref{fig:Figure8}) shows
    Classification and Confusion report output obtained from the KNN
    model we used for this project.
247: \begin{figure}
248: \includegraphics[width=1.0\columnwidth]{images/knnclassification}
249: \caption{KNN Classification and Confusion matrix report}
    \label{fig:Figure8}
265: There can be more application specific information obtained from
    classified data set to take various decisions.
    Figure(\ref{fig:Figure5}), figure(\ref{fig:Figure6}) and
    figure(\ref{fig:Figure7}) show various PHM data analytics for
    this project. Figure(\ref{fig:Figure5}) displays all the serial
    numbers of equipment 1 with bearing degradation problems. The X
    axis gives serial numbers while the Y axis gives number of
    occurrences of bearing degradation for that particular serial
    number. Similarly figure(\ref{fig:Figure6}) and

```

```
    figure(\ref{fig:Figure7}) give the details of over temperature  
    and over current faults of various serial numbers.  
272: \item Bearing Degraded Serial numbers for Equipment Part Number1:  
    Figure (\ref{fig:Figure5})  
273: \item Over Temperature Fault Serial numbers for Equipment Part  
    Number1: Figure (\ref{fig:Figure6})  
274: \item Over Current Fault Serial numbers for Equipment Part  
    Number1: Figure (\ref{fig:Figure7})  
277: \begin{figure}  
278: \includegraphics[width=1.0\columnwidth]{images/DEGRADE}  
279: \caption{Bearing Degradation} \label{fig:Figure5}  
282: \begin{figure}  
283: \includegraphics[width=1.0\columnwidth]{images/OVERTEMP}  
284: \caption{Over Temperature Fault} \label{fig:Figure6}  
287: \begin{figure}  
288: \includegraphics[width=1.0\columnwidth]{images/OVERCURR}  
289: \caption{Over current Fault} \label{fig:Figure7}
```

figures 8

tables 0

includegraphics 8

labels 8

refs 9

floats 8

False : ref check passed: (refs >= figures + tables)

True : label check passed: (refs >= figures + tables)

True : include graphics passed: (figures >= includegraphics)

True : check if all figures are referred to: (refs >= labels)

Label/ref check

passed: True

When using figures use columnwidth

[width=1.0\columnwidth]

do not change the number to a smaller fraction

find textwidth

---

passed: True

below\_check

---

bibtex

---

label errors

bibtex errors

```
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
```

bibtex\_empty\_fields

---

entries in general should not be empty in bibtex

find ""

---

passed: True

ascii

---

=====

The following tests are optional

=====

Tip: newlines can often be replaced just by an empty line

find newline

---

passed: True

cites should have a space before \cite{} but not before the {

find cite {

---

passed: True

# Diagnosis of Coronary Artery Disease Using Big Data Analysis

Hady Sylla

Indiana University Bloomington

Bloomington, Indiana 47401

hsylla@iu.edu

## ABSTRACT

This paper is about Big Data application in the diagnostic of coronary Artery disease .The paper explore contribution that big data had on the analysis of numerous data by researchers.

## KEYWORDS

i523, hid339, ???

### 0.1 Introduction

Cardiovascular illness is the primary source of dreariness and mortality in the Western world[1]. Early location of coronary corridor ailment (CAD) is of fundamental significance as favorable treatment may permanently lessen grimness and death. Albeit obtrusive coronary angiography remains the standard of reference for the assessment of CAD, multisector processed tomography coronary angiography (CTA) has as of late developed as a reliable imaging methodology for the non-intrusive evaluation of CAD. THE analysis of coronary-supply route infection on the premise of history and physical examination alone is regularly troublesome. Many complex tests have accordingly been produced to permit an early and more precise analysis. Albeit many tests are presently immovably settled in clinical practice, none is especially suited to wide-scale, savvy application,<sup>1</sup> in light of the fact that every ha constraints concerning affectability and specificity. Along these lines, when a positive test outcome happens in a patient with a low probability of illness, it is of restricted analytic importance.<sup>2 3 4</sup> A "positive" electrocardiographic anxiety test in an asymptomatic patient, for instance, has a prescient precision of just 30 for every penny for the nearness of angiographic coronary-conduit malady.

## 1 DATASET

For this project will use Z-Alizadeh Sani aataset for the data analuysis.

## 2 ANALYTIC

I will use Random Forest machine learning for predicting the feature to diagnose coronary artery disease.

## 3 RISK FACTORS

In spite of the fact that an extensive variety of hazard factors for coronary heart dis-ease have been distinguished from populace ponders, these measures, separately or in blend, are inadequately pow-erful to give a solid, noninvasive conclusion of the nearness of coronary illness. Here we demonstrate that pat-tern-acknowledgment systems connected to proton atomic mag-netic reverberation (1H-NMR) spectra of human serum can effectively analyze the nearness, as well as the seriousness, of coronary illness.

Utilization of super-vised halfway minimum squares-discriminant examination to orthogo-nal flag amended informational collections permits >90% of subjects with stenosis of every one of the three noteworthy coronary vessels to be distin-guished from subjects with angiographically ordinary coro-nary supply routes, with a specificity of >90%. Our examinations appear out of the blue a system fit for giving an accu-rate, noninvasive and fast conclusion of coronary heart dis-facilitate that can be utilized clinically, either in populace screening or to permit compelling focusing of medicines, for example, statins. Coronary illness (CHD) is a noteworthy reason for mortality and bleakness in created nations, influencing upwards of one of every three people previously the age of 70 years<sup>1</sup>. In the course of recent decades a scope of ecological and biochemical chance elements for the advancement of CHD have been iden-tified in cross-sectional studies<sup>2</sup>. For instance, tobacco smoking is related with a roughly two-overlap in-wrinkled danger of CHD<sup>3</sup>. Additionally, abnormal amounts of cholesterol in substantial, triglyceride-rich lipoprotein particles (mostly low-thickness lipoprotein (VLDL) and low-thickness lipoprotein (LDL)) and lower levels of cholesterol in high-thickness lipoprotein (HDL) particles are known to be related with expanded danger of CHD<sup>4</sup>. As of late, be that as it may, there have been specialized advances that have permitted to a great degree high-thickness informational collections to be de-veloped from people. Methods, for example, genomics, proteomics and metabonomics (a frameworks way to deal with ex-amining the adjustments in hundreds or thousands of low-mol-ecular-weight metabolites in an in place tissue or biofluid<sup>9</sup>) offer the possibility of effectively recognizing people with specific malady or lethal states. Of these methods, NMR-based metabonomics offers a few particular preferences in a clinical setting. To begin with, it can be completed on standard arrangements of serum, plasma or urine<sup>10,11</sup>, evad-ing the requirement for master arrangements of cell RNA and ace-tein required for genomics and proteomics, respectively<sup>12</sup>!!<sup>14</sup>. Second, huge numbers of the hazard factors effectively recognized, (for example, levels of different lipids) are little atom metabolites that will add to the metabonomic informational collection.

## 4 ANALYTIC

Figure 1 shows ....

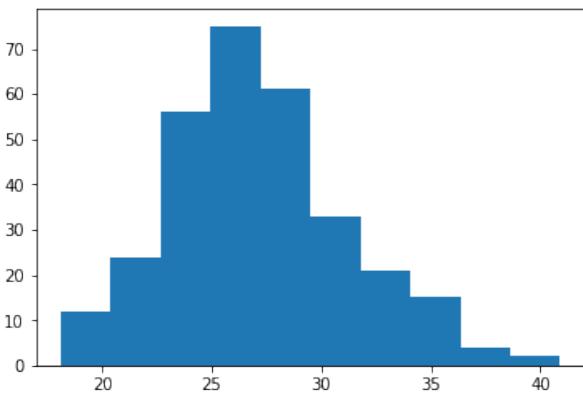
[Figure 1 about here.]

---

```
In [6]:def extract_col(df,col_name):
    return list(df[col_name])
```

```
col = extract_col(df, 'BMI')
plt.hist(col)
plt.show()
```

---



2.1 Taking `all` categorical features that have only 2 levels and label encoding them to get binary features

```
In [2]:cols = df.columns
num_cols = df._get_numeric_data().columns
cat_cols=list(set(cols) - set(num_cols))
##BBB, VHD have multiple levels rest are binary
###Cath is predictor var
cat_cols.remove('VHD')
cat_cols.remove('Cath')
cat_cols.remove('BBB')
cat_cols
```

```
Out[2]:['DLP',
'CRF',
'Obesity',
'Poor R Progression',
'Exertional CP',
'Airway disease',
'LowTH Ang',
'Sex',
'Nonanginal',
'Diastolic Murmur',
'Dyspnea',
'Thyroid Disease',
'Lung rales',
'CVA',
'CHF',
'Weak Peripheral Pulse',
'Atypical',
'LVH',
'Systolic Murmur']
```

```
In [3]:## Fitting our Encoder
In [4]:df[cat_cols]=df[cat_cols].apply(LabelEncoder()
.fit_transform)
```

One hot encoding our multiple level features: 'VHD' and 'BBB'

```
In [5]:from sklearn.feature_extraction import DictVectorizer
def encode_onehot(df, cols):
vec = DictVectorizer()
```

```
vec_data =
pd.DataFrame(vec.fit_transform(df[cols].to_dict('records'))).to
vec_data.columns = vec.get_feature_names()
vec_data.index = df.index

df = df.drop(cols, axis=1)
df = df.join(vec_data)
return df
X = encode_onehot(df, cols=['BBB'])
X1 = encode_onehot(X, cols=['VHD'])
X1.columns
```

4.0.1 *Taking all categorical features that have only 2 levels and label encoding them to get binary features.* Numerous variable choice strategies depend on the participation of variable significance for positioning and model estimation to create, assess and think about a group of models. Following Kohavi and John (1997) and Guyon and Eliseff (2003), it is common to recognize three sorts of variable determination techniques: "channel" for which the score of variable significance does not rely upon a given model outline strategy; "wrapper" which incorporate the expectation execution in the score figuring; lastly "implanted" which join all the more firmly factor choice and model estimation [2].

```
Out[5]:
Index(['Age', 'Weight', 'Length', 'Sex', 'BMI', 'DM',
'HTN', 'Current Smoker',
'EX-Smoker', 'FH', 'Obesity', 'CRF', 'CVA', 'Airway
disease',
'Thyroid Disease', 'CHF', 'DLP', 'BP', 'PR', 'Edema',
'Weak Peripheral Pulse', 'Lung rales', 'Systolic
Murmur',
'Diastolic Murmur', 'Typical Chest Pain', 'Dyspnea',
'Function Class',
'Atypical', 'Nonanginal', 'Exertional CP', 'LowTH
Ang', 'Q Wave',
'St Elevation', 'St Depression', 'Tinversion', 'LVH',
'Poor R Progression', 'FBS', 'CR', 'TG', 'LDL',
'HDL', 'BUN', 'ESR',
'HB', 'K', 'Na', 'WBC', 'Lymph', 'Neut', 'PLT',
'EF-TTE', 'Region RWMA',
'Cath', 'BBB=LBBB', 'BBB=N', 'BBB=RBBB',
'VHD=Moderate', 'VHD=N',
'VHD=Severe', 'VHD=mild'],
dtype='object')
```

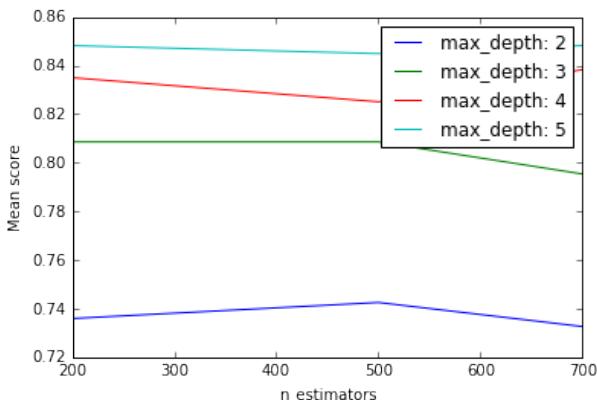
Getting our pred variable and removing from original df  
In [6]:y=df["Cath"].map({0:'Normal':1})  
X1.drop('Cath',1,inplace=True)

Splitting into train and test  
In [8]:from sklearn.model\_selection import train\_test\_split
X\_train, X\_test, y\_train, y\_test = train\_test\_split(X1,
y, test\_size=0.2, random\_state=3)

```
Hypertuning our featureset using grid search to get best
possible params
In [21]:from sklearn.model_selection import GridSearchCV
Depth=[2,3,4,5]
Trees=[200,500,700]
tuned_parameters = [{"n_estimators":
    Trees, 'max_depth':Depth}]
RFM = RandomForestClassifier()
clf = GridSearchCV(RFM, tuned_parameters,
    cv=5,scoring='accuracy')
clf.fit(X1,y)
print("Best parameters set found on development set:")
print()
print(clf.best_params_)
```

Best parameters set found on development set:

```
{'max_depth': 5, 'n_estimators': 200}
In [22]:scores = [x[1] for x in clf.grid_scores_]
scores = np.array(scores).reshape(len(Depth),
    len(Trees))
for ind, i in enumerate(Depth):
    plt.plot(Trees, scores[ind], label='max_depth: ' +
        str(i))
plt.legend()
plt.xlabel('n_estimators')
plt.ylabel('Mean score')
plt.show()
```



This is for plotting our feature importances with their standard deviations

```
In [23]:clfs = RandomForestClassifier(n_estimators=
    clf.best_params_['n_estimators'],
max_depth=clf.best_params_['max_depth'])
clfs.fit(X1, y.ravel())

importance = clfs.feature_importances_
importance = pd.DataFrame(importance, index=X1.columns,
columns=["Importance"])

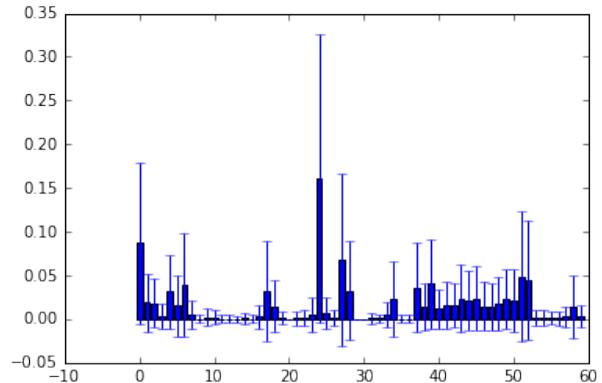
importance["Std"] = np.std([tree.feature_importances_
```

```
for tree in clfs.estimators_], axis=0)
```

```
x = range(importance.shape[0])
y = importance.ix[:, 0]
yerr = importance.ix[:, 1]

plt.bar(x, y, yerr=yerr, align="center")

plt.show()
```



In [24]:importance

Importance	Std
------------	-----

This table shows the importance of each feature in the diagnostic of CAD. Based on the analysis the age, HTN, DM, BP, Typical chest pain, Non-Anginal CP, T inversion, Q wave, ST elevation, PR, and ST depression are the features with the highest impact on CAD.

Variable significance is figured restrictively to a given acknowledgement notwithstanding for reproduced datasets. This decision which is criticizable if the goal is to achieve a decent estimation of a basic steady, is predictable with remaining as near as conceivable to the exploratory circumstance managing a given dataset[2].

Out[24]:

Age	0.086779	0.092552
Weight	0.018672	0.032956
Length	0.018183	0.027928
Sex	0.003414	0.013603
BMI	0.031609	0.042215
DM	0.015427	0.034442
HTN	0.038735	0.059296
Current Smoker	0.005287	0.015965
EX-Smoker	0.000587	0.004463
FH	0.002472	0.010216
Obesity	0.002333	0.008439
CRF	0.000534	0.004805
CVA	0.000469	0.004108
Airway disease	0.000474	0.003385
Thyroid Disease	0.001086	0.006045
CHF	0.000407	0.004046
DLP	0.003296	0.013382

BP	0.031988	0.056958
PR	0.014277	0.029568
Edema	0.001758	0.007496
Weak Peripheral Pulse	0.000000	0.000000
Lung rales	0.001332	0.007931
Systolic Murmur	0.001853	0.008916
Diastolic Murmur	0.004950	0.020264
Typical Chest Pain	0.161378	0.165277
Dyspnea	0.006237	0.017810
Function Class	0.001798	0.008412
Atypical	0.067681	0.098911
Nonanginal	0.032299	0.056285
Exertional CP	0.000000	0.000000
LowTH Ang	0.000000	0.000000
Q Wave	0.000976	0.005923
St Elevation	0.000906	0.004613
St Depression	0.005336	0.014798
Tinversion	0.022593	0.042821
LVH	0.000687	0.004452
Poor R Progression	0.000546	0.004018
FBS	0.036382	0.051744
CR	0.014378	0.027248
TG	0.040267	0.051599
LDL	0.011746	0.022602
HDL	0.015883	0.027084
BUN	0.015708	0.025022
ESR	0.023733	0.038713
HB	0.021479	0.033380
K	0.023375	0.036937
Na	0.014249	0.027810
WBC	0.014354	0.026839
Lymph	0.018147	0.030376
Neut	0.022351	0.034163
PLT	0.021242	0.035043
EF-TTE	0.049034	0.074234
Region RWMA	0.044669	0.068886
BBB=LBBB	0.001229	0.008498
BBB=N	0.002322	0.008860
BBB=RBBB	0.000986	0.006953
VHD=Moderate	0.001554	0.008084
VHD=N	0.002828	0.011718
VHD=Severe	0.014135	0.036448
VHD=mild	0.003587	0.013188

---

\large Final checking of our model's accuracy on test set

In [25]:

```

clf = RandomForestClassifier(n_estimators=
    clf.best_params_['n_estimators'],max_depth=clf.best_params_['max_depth'])
clf.fit(X_train, y_train)
y_pred=clf.predict(X_test)
print(accuracy_score(y_test,y_pred))
print()
print()
print('The accuracy score on test by Random
Forest:{}' .format(accuracy_score(y_test,y_pred)))
0.885245901639

```

---

The accuracy score on test by Random  
Forest:0.8852459016393442

## 4.1 Conclusion

A period of open data in social insurance is currently under way. We have effectively encountered a time of advance in digitizing therapeutic records, as pharmaceutical organizations and different associations total a long time of innovative work information in electronic databases. The national government and other open partners have additionally quickened the push toward straightforwardness by making many years of put away information usable, accessible, and significant by the medicinal services segment in general. Together, these increments in information liquidity have conveyed the business to the tipping point.

## 4.2 Acknowledgment

I would like to thank Dr. Gregor von Laszewski for his support and suggestions to write this paper.

## REFERENCES

- [1] ALIZADEHSANI, R., HABIBI, J., HOSSEINI, M. J., BOGHRTI, R., GHANDEHARIOUN, A., BAHADORIAN, B., AND SANI, Z. A. Diagnosis of coronary artery disease using data mining techniques based on symptoms and ecg features. *European Journal of Scientific Research* 82, 4 (2012), 542–553.
- [2] GENUER, R., POGGI, J.-M., AND TULEAU-MALOT, C. Variable selection using random forests. *Pattern Recognition Letters* 31, 14 (2010), 2225 – 2236.

LIST OF FIGURES

1 Read data

6

---

```
In [2]: import matplotlib.pyplot as plt
%matplotlib inline
from sklearn.metrics import accuracy_score

In [4]:from sklearn.ensemble import RandomForestClassifier
import pandas as pd
from sklearn.preprocessing import LabelEncoder
import numpy as np
df=pd.read_excel('Z-Alizadeh sani dataset.xlsx')
df.head()
```

---

Figure 1: Read data

```
bibtext report
```

```
=====
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: acm.bst
Database file #1: report.bib
```

```
bibtext _ label error
```

```
=====
bibtext space label error
```

```
=====
bibtext comma label error
```

```
=====
latex report
```

```
=====
[2017-12-11 13.31.17] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflatex)
Missing character: ""
bookmark level for unknown defaults to 0.
The anchor of a bookmark and its parent's must not be the same. Added a new anchor.
Typesetting of "report.tex" completed in 1.1s.
./README.yml
18:14     warning  truthy value is not quoted  (truthy)
31:14     warning  truthy value is not quoted  (truthy)
```

```
=====
Compliance Report
=====
```

```
name: Hady Sylla
hid: 339
paper1: 100% Oct 27 2017
paper2: Nov 5 2017 100%
project: Dec 2 2017 100%
```

```
yamlcheck
```

---

```
wordcount
```

---

```
6
```

```
wc 339 project 6 1755 report.tex  
wc 339 project 6 1700 report.pdf  
wc 339 project 6 122 report.bib
```

```
find "
```

---

76: THE analysis of coronary-supply route infection on the premise of history and physical examination alone is regularly troublesome. Many complex tests have accordingly been produced to permit an early and more precise analysis. Albeit many tests are presently immovably settled in clinical practice, none is especially suited to wide-scale, savvy application,<sup>1</sup> in light of the fact that every ha constraints concerning affectability and specificity. Along these lines, when a positive test outcome happens in a patient with a low probability of illness, it is of restricted analytic importance.<sup>2 3 4</sup> A "positive" electrocardiographic anxiety test in an asymptomatic patient, for instance, has a prescient precision of just 30 for every penny for the nearness of angiographic coronary-conduit malady.

182: Numerous variable choice strategies depend on the participation of variable significance for positioning and model estimation to create, assess and think about a group of models. Following Kohavi and John (1997) and Guyon and Eliseff (2003), it is common to recognize three sorts of variable determination techniques: "channel" for which the score of variable significance does not rely upon a given model outline strategy; "wrapper" which incorporate the expectation execution in the score figuring; lastly "implanted" which join all the more firmly factor choice and model estimation \cite{GENUER20102225}.

205: In [6]:y=df["Cath"].map({'Cad':0,'Normal':1})

222: print("Best parameters set found on development set:")

252: columns=["Importance"])

```
254: importance["Std"] = np.std([tree.feature_importances_
261: plt.bar(x, y, yerr=yerr, align="center")
passed: False
find footnote
-----
passed: True
find input{format/i523}
-----
4: \input{format/i523}
passed: True
find input{format/final}
-----
passed: False
floats
-----
92: Figure \ref{F:readdata} shows ....
94: \begin{figure}[htb]
108: \caption{Read data}\label{F:readdata}
119: \includegraphics[width=0.95\columnwidth]{images/output_2_0.png}
181: features}\label{taking-all-categorical-features-that-have-
    only-2-levels-and-label-encoding-them-to-get-binary-features}
184: \% \includegraphics[width=0.95\columnwidth]
241: \includegraphics[width=0.95\columnwidth]{images/output_15_1.png}
265: \includegraphics[width=0.95\columnwidth]{images/output_17_0.png}

figures 1
tables 0
includegraphics 4
labels 2
refs 1
floats 1

True : ref check passed: (refs >= figures + tables)
False : label check passed: (refs >= figures + tables)
False : include graphics passed: (figures >= includegraphics)
```

False : check if all figures are referred to: (refs >= labels)

Label/ref check  
passed: True

When using figures use columnwidth  
[width=1.0\columnwidth]  
do not change the number to a smaller fraction

find textwidth

---

passed: True

below\_check

---

bibtex

---

label errors

bibtex errors

This is BibTeX, Version 0.99d (TeX Live 2016)  
The top-level auxiliary file: report.aux  
The style file: acm.bst  
Database file #1: report.bib

bibtex\_empty\_fields

---

entries in general should not be empty in bibtex

find ""

---

passed: True

ascii

---

```
non ascii found 8211
```

---

```
=====  
The following tests are optional  
=====
```

```
Tip: newlines can often be replaced just by an empty line
```

```
find newline
```

---

```
passed: True  
cites should have a space before \cite{} but not before the {
```

```
find cite {
```

---

```
passed: True
```

# **Big Data Application in Precision Medicine and Pharmacogenomics**

Budhaditya Roy

Indiana University

School of Information and Computing

Bloomington, IN 47040

royb@indiana.edu

## **ABSTRACT**

This article focuses on the impending impact of big data analytics improving health, preventing and detecting illness at a preliminary stage of illness and personalize interferences. The complexity and diversity of biological data are pouring the need of big data analytics and how it is applied in biological field especially in Pharmacogenetics, personalized/precision medicine. Big data is particularly very useful in the healthcare industry as a whole for its data handling intensive nature. Over the past decade, electronic health records (EHR) have become an extensively accepted in hospitals and clinics worldwide and the amount of information is generally daily from a single patient is increasing day by day. Important clinical acquaintance and a deeper understanding of patient disease patterns can be deliberate from such data. It will help to improve patient care as well improve efficiency of patient care and disease prevention. There are few applications pointed out to be effective using big data such as Healthcare data solutions and big data in cancer therapy, continuous monitoring of patients symptoms, healthcare intelligence, fraud prevention and detection. Many people heard about the proposition of precision medicine in State of Union speech of President Obama in 2012. Since then the revolutionary process of precision medicine started to grow rapidly in healthcare industry. On January 30, on the same Precision based medical initiatives, the Obama administration exposed facts about the Precision Medicine tentative plans. Threw with a 215 million dollar investment in the US President's 2016 budget,, the Precision Medicine Initiative will product a new model of patient driven research that eventually support delivering the right treatment to the right patient at the right time[7]. On March 11, 2015, it is reported that China is planning to invest 60 billion Yuan almost 10 billion) in precision medicine (20 billion from the Central Government and the remaining 40 billion from local governments and companies) before 2030 [[7]]. There is a similar necessity of big data application to this latest emergence of biomedical domain. . Big data in precision medicine is the most widely used methods in precision and personalized medicine which is a life changing event in healthcare industry. Personalized medicine or called as Precision medicine is product and services that leverage the science of genomics and proteomics and take advantage of on the trends concerning wellness to enable preventive care. By using big data analytics, prevention and detection of diseases are in a new era of healthcare which essentially improving daily life of every patient. Personalize medicine is the in an era of new modern healthcare innovation. The role that big data analytics may have in interrogating the patient electronic health record headed for improved clinical decision support is discussed. In this paper we try to examine developments in pharmacogenetics

that have enflamed our appreciation of the reasons why patients respond inversely to chemotherapy in cancer treatment. We also try to measure the development of online health infrastructures and the way healthcare data may be capitalized in order to detect public health warnings and control or comprise epidemics. Finally, this paper talks about how a new generation of body sensors in form of implanted in human body may improve comfort, rationalize management of chronic diseases and progress the superiority of surgical implants which could be effectively used in near future. So, let's talk about what is precision medicine? How is it related to other dealings such as personalized medicine and omics technologies (especially in Pharmacogenomics and in Pharmacoproteomics)

## **KEYWORDS**

Keywords- HID348, Precision Medicine, Pharmacogenomics, Pharmacoproteomics. Big data Application, Data analytics, Big data infrastructure

## **1 INTRODUCTION**

The complexity, diversity, and rich context of data being generated in healthcare are driving the development of big data for health [3].The data captured at these portals can also help significantly reduce the cost of drug discovery as improved predictive analytics to determine which drugs work well and which are not as effective for certain conditions. Big data analytics may even allow for uploading the genomics of large populations that can be warehoused for researching new generations of drug remedies. Big data analytics is becoming increasingly popular in modern world with almost every domain. Big Data means lots of data used for analysis and get the insight from the data. Big data applications in general is applicable to any domain such as Retail, Healthcare, Finance and Supply Chain efficiently but in current world the application of Big Data has a major impact on healthcare sector where daily volume of data quadruple in every minute around the globe. Organizations are using Big Data to envisage the future with the goal of making them smarter and competitive in daily work. Applications from Big Data has become from retail industry where Big Data helps retailers gain insights into the customer needs and by monitoring customers' habits future can be effectively utilized, HealthCare and Hospitality[3]. Government agencies are progressively integrating Big Data analytics to control crime and sustain law by foresee the circumstances and by using social media, it is trying to achieve other benefits out of it. So, to get actionable data and perform analytics requires specialized tools which can handle this

massive amount of data as well as help in analysis of the information. There are thousands of Big Data tools available in the market right now which contribute significantly to healthcare analytics. There are open source tools like Hadoop, which is named as big data umbrella and in big data ecosystem. Today's healthcare data are beginning analyzed using aforesaid big data tools such as Pig, Cassandra, MongoDB and others. The quality of health care services in US and across the globe have been enhanced tremendously because of the advancement in health care services, advancements of technologies and Artificial intelligence process which improved the accuracy of healthcare as a service to next level. According to Google Trends analysis, the number of searches using the keyword "fibig data" started to increase dramatically in 2011 and reached the peak in 2017 [3]. Although the term "fibig data" resonates as if it is connected to the area of data science, big data and data science both play a significant role in healthcare research especially in precision and emergency medicine. Conventionally, scientists have adopted the traditional 4Vs criteria to describe big data: volume of data, velocity of data which is speed of incoming and outgoing data and variety which is range of data types [4]. However, from the perspective of medical science, this classification may not be real world sufficient as the 3Vs criteria are forceful and time-reliant. Big Data is a capacious collection of data that cannot be achieved by traditional database management systems (RDBMS). Big Data is an umbrella term used for the enormous amount of data produced from countless of sources such as mobile, web, sensor devices, enterprise applications and rigorous digital repositories. In big data umbrella, data can be structured as well as unstructured or semi-structured. The data varieties from terabytes to exabytes of data [4]. The relational database management systems (RDBMS) have proven inefficient to handle such huge volumes of data in form of patient records such as X-ray, Scan and routine checkup results. Another important factor which renders the conventional database systems inappropriate is that the majority of data being generated as unstructured, the RDBMS systems are only adept to handle structured relational data. Hereafter new tools and systems for data analysis and management have emerged. Volume, velocity, variety, veracity, variability, and value are the three must-haves of big data and these are condensed in the integral challenges of biomedical and health informatics. Effective ways of confronting these challenges would cover the way for more intellectual healthcare systems focused on early detection, prevention and personalized treatments. As Big data is characterized by the 4 Vs. We discussed about the 4Vs below which essentially contributed to the success of healthcare management [3]. 1. Volume- As data are increasing day by day, it is always in light of voluminous collection of data. The complete volume of data generated these days by real-time applications such as X-ray machines and MRI systems and other external data sources such as Facebook comments, tweets or even patients' data, it runs to petabytes and exabytes of data. Big Data technology empowers us to store this amount of data on dispersed systems [[4]]. 2. Velocity- is the proportion at which data is arrived. As an example, in whole gene sequencing process, one sequence generates huge volume of data and when the sequencing process is completed data arrives at a very higher speed having few time to store and analyze the data. 3. Veracity- When the volume increases so does the quality. Veracity refers to the quality of the data.

There are doubts of good quality of accurate data being generated in recent times. Big Data applications empower working with data which are large in volume, accurate and insightful [4]. 4. Value- Everything in the world has a value so does the data. There is an intrinsic value that the data holds and discovers for analysis. Value is the heart of Big data analytics and the way data generated value in healthcare is just enormous. Modern technologies have made it possible to find the insight from data. Since data is huge and storage capacity leads to an expensive turnaround. Apache Hadoop is the savior in these kinds of applications by processing gigabytes of data in a very short span of time and Hadoop ecosystem consisting of MapReduce a different language processing system or Hive and Drill an analytical SQL platform on Hadoop or Spark, in memory data flow system or HBase/MongoDB in memory database systems or HDFS, capable of storing petabytes of data or streaming systems such as Apache Storm and Kafka, overall these are highly capable tools can be profoundly effective in healthcare biomedical data analytics.

## 2 WHAT IS PRECISION MEDICINE

Precision medicine in broader terms is another name of preventive medicine. According to the Precision Medicine Initiative and American Healthcare Association, precision medicine is an emerging approach for disease treatment and prevention that takes into interpretation of individual heterogeneity in genes, environment and routine and lifestyle for each person. This approach will allow medical professionals and researchers to predict more accurately about the treatment and prevention approaches for any particular disease about its effectiveness. It is in divergence to a one-size-fits-all approach in which disease treatment and prevention strategies are developed for the average person with less deliberation for the genetics-based differences between each individual. Though the term precision medicine is relatively new in medical industry, the concept has been there and as a part of healthcare for many years. As an example, a patient who needs a blood transfusion is not given random blood from a blood bank storage rather it will be from a process of donor's blood type is matched to the recipient to decrease the risk of future problems [8]. Although illustrations can be found in numerous areas of medicine the role of precision medicine in day-to-day healthcare is relatively restricted. Medical researchers expect that this approach will increase in many areas of health and other healthcare domains in upcoming years. Precision medicine is being sought to transform how we as a whole improve health, treat and prevent disease. Today most of the medical treatments are intended for the average patient using the one-and-one approach. However, in many cases, this approach is not at all effective because treatments can be very successful for some patients but not for every patient [4]. As an example, if Patient A and Patient B both have stage 3 lung cancer, giving the same chemotherapy to both the patient helps one but not the other. In precision medicine with the help of big data technology, medicines are targeted to specific genomic sequences rather than a random selection. In advanced countries like USA, a rigorous process is already in place to target particular genes after finding the root cause of the disease. It is a big data application which enables to store the data and use analytical tools to get useful information out of it. Overall, Precision medicine is a field

of medicine that takes into interpretation individual differences in people's genes, environments, microbiomes, habitual effects, and family history to make diagnostic and beneficial strategies accurately personalized to individual patients. Precision medicine is a newer term referring to a similar ground compared to another word if personalized medicine. The term if precision medicine arrived the scientific dictionary in the year 2008 when business strategist Clayton Christensen, of Harvard Business School in Boston, invented the appearance to describe how molecular diagnostics allows physicians to unambiguously diagnose the cause of a disease without having to rely on perception [4]. The name precision medicine didn't gain enough attention until 2011 when a committee convened by the US National Research Council placed out a plan for modernizing the classification of disease on the foundation of molecular information such as causal genetic variants instead of a symptom based cataloguing system. The committee called the report Toward Precision Medicine [3]. There are many areas where precision medicine is vastly applicable and are very much beneficial such as, finding correct dose of prescription drugs, root cause analysis of a disease and so on. The field of pharmacogenomics aims to understand how genetic variations influence individual responses to medications. Genetic tests for supervisory treatment decisions are becoming increasingly available across miscellaneous areas of medical care. These kind of tests provide more effective drugs to patients earlier in their treatment and with fewer negative side effects and in less costly than previous tests. Precision medicine is also pertinent in Cancer detection, Genomics and cure process. Oncology is the target of some of the most auspicious precision medicine approaches available today. Cancer forms through the gradual accumulation of genetic DNA changes in genes that regulate cell growth. That is why, cancer is very much an illness of the genome. Depending on where in the body the cancer started and the types of genetic changes the cells grow, different types of cancer have very different genetic profiles which completely varies person to person and highly dependent on their family history. These genetic sequences can be used in a number of ways to help medical professionals choosing the best treatments for each individual patient. Growing tissues replacement is another way to apply precision medicine in pharmacogenomics [3].

## 2.1 Personalized Medicine

The concept of personalized medicine dates back many hundreds of years although the term seemingly similar meaning with precision medicine. Mere from 19th century, scientists started to measure the chemistry of root cause of any illness and the research improvements are granular over time. With the growth of the pharmaceutical industry and medical technology industries in recent times came the rise of genetics, data mining and imaging. Halfway over the period, comments of specific alterations in retort to drugs contributed growth to a body of study attentive on classifying crucial enzymes that play an important role in disparity in drug absorption and reaction and this is helped as the basis for pharmacogenetics. In recent times, sequencing of the human genome customary in motion the transformation of personalized medicine from an knowledge to a practice. Personalized diagnosed tools are now created with rapid developments in genomics along with advances high critical areas

such as computational biology, medical imaging, and regenerative medicine and treatment [4]. Personalized medicine first appeared in available mechanism in 1999 with the creation of some of the domain specific core concepts even dating back to 19th century [2]. So basically personalized medicine is referred to treatment depending on each individual's personal structure and history. Initially, personalized medicine is the idea that assortment of a treatment should be custom-made giving to the individual patient's specific physiognomies, including age, sex, gender, height, ethnicity, diet, and environmental factors against traditional clinical trials on group of people which has been happening since the invention of medicine happened [4]. Scientists got interested on personalized medicine when medical professional started understanding the essence of gene in human development. Several human genome projects have been conducted since then and the importance of personalized medicine started in limelight. With deceitful out in order the 3.2 billion units of our DNA, scientists flashed a blaze of detection and a detonation of genomic knowledge in medical science history [2]. Novel omics technologies including microarrays, whole genome single nucleotide polymorphism [SNP] chips, RNA interference high-throughput transmission, next generation sequencing are the few procedure which accompanied with this revolution. All the above launch a new epoch in personalized medicine which is called genomic revolution era which bids us limitless probable and countless promise containing the expansion of personalized medical products for each individual based on their sole genomic information [8]. Advancement of genomics science along with the developing of new omics technologies, personalized medicine is today frequently well-defined as a combination of molecular profiling (omics methods) and customary methods such as family history, lifestyle and environment, which create analytic and beneficial strategies precisely personalized to individual patients [2, 4].

## 3 BIG DATA IN PRECISION MEDICINE

Once again the term Big data is signifies in collection of large and complex data sets which are difficult or sometimes impossible to process using common database management tools or traditional data processing applications even with modern advancement of traditional data warehousing tools such as Amazon Redshift. In 2012, the Obama administration announced the Big Data Research and Development Initiative [6], which explored how big data could be utilized to address important problems faced by the overall healthcare system. Since then, Big Data becomes such a big term that people tend to claim any kind of data analysis to be if Big Data if characterization. The overall concept of big data can be explained in various ways. One way is, Big data is a comprehensive term for any collection of data sets are so voluminous that processing the data in the begging stage itself is very hard. With four if Vif characterization of big data m, complexity arises more to collect data and make meaningful information out of it. Omics data, mobile internet real-time data and electronic health record data are the top three areas for Big Data in medical research. Precision medicine will use all of these three Big Data. In fact, among the 215 million investment in the USA President's 2016 Budget, 130 million (over 60 percent) will be used for building a large US cohort for precision research [6]. In this regiment study, the scientists will use widespread omics

data, electronic health record data gathered from several hospital and private practices along with mobile internet data [\*]. Thus, omics and medical big data are one of the key pairs in the success of precision medicine in healthcare industry as a whole.

## 4 BIG DATA CHALLENGES IN HEALTHCARE

- Whenever anything benefits us, that comes with its own challenges and problems. The primary idea of big data to be applied in healthcare is to roll massive healthcare dataset with individual information. As the need of more data driven enterprise grows Besides general challenges inherent to the analysis of big data such as missing data, vague data, and varied data, employing big data in health care systems imposes new challenges which includes the lack of reliability and a solid data governance of some biomedical data, issues of privacy and security and confidentiality, insufficient data from random clinical trials including successful and failed trials, and overall low quality data. Challenges in machine learning and statistical applications also put the analytics in challenging situation where model development and execution are critical to success[2]. Healthcare providers who have hardly come to grasps with driving data into their electronic health records (EHR) are now being questioned to pull actionable insights out of them and apply those learnings to complex initiatives that straight impact their repayment rates. Organizations who can integrate this data driven technological innovation to their healthcare operations are in the most benefit[5]. Data assets and data insights can be achieved by using healthier patients, increased visibility in operational excellence, lower care costs and higher staff and consumer satisfaction rates are among the many benefits of turning data assets into data insights. The journey to evocative healthcare analytics is difficult challenge and problems by solving those will benefit the industry to the highest extent. The way overall big data analytics work, collecting, storing, analyzing the data require clear presentation to the staff members to understand the overall workflow process[4]. Analyzing genomic data is a computationally are some of the top challenges organizations typically aspect when striking up a big data analytics program and how can organizations overawed these issues to attain their data driven clinical and financial goals are the most important aspect of big data implementation. Understanding unstructured clinical nodes, storing unstructured patients health records are complex in nature and specialized training is required in implementing the analytics platform is essential. Some of the pitfall of big data application in precision medicine is discussed below.

### 4.1 Data Collection

This is the most crucial stage in any data driven technologies, capturing the patient's behavioral data through several sensing processes; with their numerous social interactions and communications. The data many come from many sources or in different format but not everywhere data governance is properly applied while collecting the data. Capturing data which is clean, comprehensive, correct, and well formatted for use in diverse systems is an ongoing combat for organizations, many of which are not on the endearing side of the battle. As an example, electronic health record capturing in right movement help physicians to access the accurate picture

of the patient's history. Oftentimes, delay in collecting this data create problems which eventually leads to unhealthy environment and future risks. Revolving Healthcare Big Data into Actionable Clinical Intelligence Providers can start to recover their data capture procedures by ranking valuable data categories for their specific plans, conscripting the data governance and honesty knowledge of health information management professionals and evolving clinical documentation improvement programs that tutor clinicians about how to confirm that data is valuable for downstream analytics [4]

### 4.2 Data Cleaning

Healthcare providers are well familiar with the importance of cleanliness in the clinic and the operating room but they are not aware on many things which could lead to a clear picture of the meaningful data. Data which is dirty and raw might have a potential impact on big data analytics projects and can screw up the true insight completely. Data cleaning also known as data scrubbing always ensures that data is not inconsistent, proper and useful in perspective and predictive analytics. Though when everything started, data cleaning was a manual process, but now with the help of big data quality tools, cleaning data has been easier than ever before. Since data cleaning is complex and tedious process in particular healthcare system, oftentimes big data analytical tools stand by the first door where data streamline occurs. Which eventually cleans the data with a global standard before it entered to main stream pipeline.

### 4.3 Data Storage

This is the most critical place where big data application play a key role. As the volume of healthcare data grows exponentially many healthcare providers are not able to manage the costs and effects of on premise data centers. Although many organizations are most happy with on premise data storing which also leads to security issues and data governance issues. With the help of cloud storage almost 90 percent of healthcare providers have chosen cloud based data storage centers which provides better flexibility and availability of data. Amazon web services, Microsoft Azure cloud and Salesforce cloud have put the data storage industry to the utmost point where no longer organizations need to worry about the cost and capacity of storing humongous amount of data. The cloud offers sprightly disaster recovery, lower set up and upfront costs and easier development, although organizations must be extremely careful about choosing partners that understand the significance of HIPAA law and other healthcare related compliance and security issues [3]. Many organizations finish up with an amalgam approach to their data storage agendas, which may be the furthestmost supple and workable approach for providers with variable data access and storage necessities. When creating hybrid substructure providers should be cautious to safeguard that dissimilar systems are able to interconnect and portion the data through extra segments of the organization when necessary [5].

### 4.4 Data Security

Data security is the number one priority for healthcare organizations, particularly in the wake of a hundreds of data breaches, hackings, and intrusion incidents. Data is so sensitive especially

in healthcare systems that a proper security measure has to be taken to protect the data. Healthcare privacy law such as HIPAA and others put the organizations in the front door where every healthcare providers must conform the law to protect the data. For precision based medicine era, this has become more important with each individual patient's data being captured and analyzed. Since genomic science is completely depending on data architecture, one data breach can push the healthcare provider in a tremendous reputation and financial loss. Due to this, security is one of the most talked topic in personalized medicine [2].

#### 4.5 Data Governance

Healthcare data, particularly on the clinical side has a long ledger life. In accumulation to existence required to keep patient data available for at least six years of time frame, providers might request to use de identified datasets for scientific projects, which makes continuing stewardship and curation an important concern [2]. Any data can be used for variety of other purposes as long as data masking is properly applied to the dataset. Understanding of the data when it is created by whom and for purpose can lead to positive results while in research.

#### 4.6 Data Querying

Vigorous metadata and robust stewardship procedures make organizations to comply with data querying very effectively. There are many tools in the market which can give access to query from databases to get the useful information. Azure datalakes, different programming based API, Hadoop Sqoop are the few tools which help in big data query language extraction. Many organizations use Structured Query Language (SQL) to dive into large datasets and relational databases, but this can only be true if end user can trust the data they are working on which can provide useful information to them [2]. Data Reporting: Reporting is the end to end product of any data collection and process. Big data reporting can help transform virtually all aspects of the enterprise. From quickly producing actionable intelligence to driving productivity to gain real time visibility into customers and markets, big data analysis and big data reporting promise to deliver a wealth of benefits for competitive advantage [\*]. Many companies including not for profit hospitals use reporting as their sole decision making procedure. Data reporting helps unveils the insight into charts and graphs and visualize the insight to the target audience. In healthcare organizations especially in precision medicines, reporting of the finding is must have to take informed decision. Data reporting has the potential to show about the result of an ongoing research study or data findings. [2].

#### 4.7 Data Visualization

In patient care, a clean and attractive data visualization can make it much easier for a clinical staff to understand the fundamental very easily and take decision based on it. Color visualizations are a popular data visualization technique that typically yields an immediate response as an example, red, black color divergence is. Organizations must also consider good data presentation practices such as charts, graphs, scatterplot. Common examples of data visualizations include heat maps, bar charts, pie charts, scatterplots,

and histograms, all of which have their own specific uses to prove concepts and material.

#### 4.8 Data Update

Healthcare data is non-static and almost all the elements requires an update in daily interval. Some datasets such as patient vital signs and symptoms, may require frequent update. But patient's demographic information may change once in a while. Since in genomic since changes are captured in every interval or procedure, there has to be constant update to the proper and existing dataset. The most critical phrase comes when incremental data is gathered and new data is added to existing dataset. In precision medicine, medical professional compares whole gene sequences in different timeframe of the disease and in different medicine stages. In these case, data has to be updated regularly in order to analyze the proper data [? ]. Organizations should also confirm that they are not making needless identical records when endeavoring an update to a single component which may make it problematic for clinical staff members to access needed information for patient decision making.

#### 4.9 Data Sharing

With the essence of electronic medical record, data sharing become easier but complicated in healthcare analytics. With large volume and the structure of the data, healthcare providers and researchers are immensely beholding data which may contribute finding to their scientific invention. Data exchange is a perpetual worry for organizations at any costs. With the increase data volume and nature of the data, it is getting more and more difficult for the organization to move data from one to place to another without losing information and change in pattern on data lineage can lead to significant mislead information. [5].

### 5 PRECISION MEDICINE AND OMICS

With the growth of big data, organizations move into NOSQL databases where security is a growing concern. Though we found there are severe security issues in most of the NOSQL databases which are used today in big data environment. Lack of security measures put extra sensitivity to the overall big data applications being NOSQL databases are heart of any big data project. Though not reached at pick, constant evaluation and research are in process to make NOSQL databases more secure in near future. The evolution of omics outlining technologies significantly benefited studies are conducted on diseases mechanism, molecular diagnosis and personalized treatment [4]. The study of omics is strongly related to the study of biology as a whole and precision medicine. There is a strong connection between Omics and Precision medicine and big data as a whole has become the core of precision medicine. The advancement of precision/personalized medicine depends heavily on the ability to acquire biological aces at omics interval though the training of precision medicine does not use sole omics data and omics knowledge [1]. This happens due to molecular characteristics found from omics data can categorize diseases and classify population of patients appropriate to assured common treatment more exactly [4]. Biology has become more data intensive and technological intensive subject .Following this trend, many of the emerging fields of large-scale data rich biology are designated by

adding the suffix fi-omicsfi to previously used definitions. Particularly, the word omics refers to a field of study in biology ending in the suffix fi?! omics and it is related addresses the objects of study of such a field[4][2]. Pharmacogenomics is the study of how a person's response to drugs is affected by his genetic makeup [3]. It combines pharmacology which is also called the science of drugs and genomics which is the study of gene and their functions to develop effective, proper medications that will be personalized to a person's genetic makeup. Pharmacoproteomics, essentially a sub discipline of functional pharmacogenomics which is a study of how the protein content of a cell or tissue changes qualitatively and quantitatively in response to treatment or disease, what the protein-protein and protein ligand interactions are in related to drug response, and how a person's protein variants in quality and quantity affect a person's response to a drug [8]. In modern days, the pharmaceutical industry has developed strong interest in Pharmacoproteomics with the anticipation that this technology will lead to the empathy and authentication of protein targets and eventually to the detection and growth of feasible drug candidates. Pharmacogenomics and Pharmacoproteomics will help the prescription of drug and related doses to a patients based on response to a drug which greatly indorsing the advance and practice of precision/personalized medicine [8]

## 6 CONCLUSION

Personal medicine, Omics technologies and pharmacogenomics are the evolutionary invention in medical industry, holding the hands of these medical concepts scientist can not only find the cancer cell in human body parts as well as start of cancer as a disease in a particular cell. These all is possible due to the essence of big data which not only helped organizations to tackle the voluminous data effectively but to use them in a way to get meaningful insight out of it. Massive parallel computing and clustering are now opened up new window in medical research where processing of huge amount data is better than ever before as well as build automated model on top of it. Whole gene sequencing is an example of how a big data can help strong millions of genetic information in a single storage system and take useful information out of it. With the help of big data in precision based medicines scientists are now able to predict the origin of the disease, track and cure it more effectively.

## ACKNOWLEDGMENTS

The author would like to thank Dr. Gregor von Laszewski and I523.

## REFERENCES

- [1] Shein-Chung Chow<sup>1</sup> and Fuyu Song<sup>2</sup>. 2016. *Some Thoughts on Precision Medicine*. Journal of Biometrics and Biostatistics, Chapter 1, 1.
- [2] Mohit Dayal<sup>1</sup> and Nanhay Singh. 2012. *Indian Health Care Analysis using Big Data Programming Tool*. Web, Chapter 1, 2. NA
- [3] Andy Futrel. 2012. *Building Genomic medicine capability* (1st. ed.). 4th, Vol. 2. MD Anderson Cancer Research center, Boston, MA, Chapter 1. [https://doi.org/10.1007/978-1-4614-3540-4\\_2](https://doi.org/10.1007/978-1-4614-3540-4_2)
- [4] Daniel Richard Leff and Guang-Zhong Yang\*. 2015. . 1, Vol. 1. Engineering.org, Chapter 1, 2. <https://engineering.org>
- [5] IEEE Chih-Wen Cheng Member IEEE Chanchala D. Kaddi Member IEEE Janani Venugopalan Member IEEE Ryan Hoffman Member IEEE Po-Yen Wu, Member and IEEE May D. Wang, Senior Member. 2016. *Omic and Electronic Health Record Big Data Analytics for Precision Medicine*. Chapter 1, 1. <https://doi.org/10.7339/9781509037876>
- [6] White House Press Release. 2015. *Precision Medicine Initiatives*. White House Press Conferences, NY, Chapter 1, 1. [https://doi.org/10.1007/978-1-4614-3540-4\\_2](https://doi.org/10.1007/978-1-4614-3540-4_2)
- [7] Whitehouse (Ed.). 2012. *Precision medicine Initiatives*. 1st, Vol. 1. Purdue University Press, Purdue University, Indiana. <https://doi.org/10.7339/9781509037876>
- [8] Xiaohua Douglas Zhang. 2015. *Pharmacogenomics & Pharmacoproteomics*. 1, Vol. 1. Merck Research Laboratories.

```
bibtext report
```

---

```
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
I was expecting a ',' or a '}'---line 61 of file report.bib
:
: chapter = "1",
(Error may have been on previous line)
I'm skipping whatever remains of this entry
I was expecting a ',' or a '}'---line 123 of file report.bib
:
: doi = "10.7/3-105.876",
(Error may have been on previous line)
I'm skipping whatever remains of this entry
I was expecting a ',' or a '}'---line 138 of file report.bib
:
: doi = "10.7/3-105.876",
(Error may have been on previous line)
I'm skipping whatever remains of this entry
Warning--I didn't find a database entry for "editor0"
Name 1 in "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Chanchala D. Kaddi, Me
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha
while executing---line 3085 of file ACM-Reference-Format.bst
```









```
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha
while executing---line 3229 of file ACM-Reference-Format bst
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha
while executing---line 3229 of file ACM-Reference-Format bst
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha
while executing---line 3229 of file ACM-Reference-Format bst
Warning--empty publisher in editor07
Warning--empty address in editor07
Warning--empty chapter and pages in editor01
Warning--unrecognized DOI value [NA]
Warning--empty address in editor04
Warning--empty chapter and pages in editor04
(There were 107 error messages)
make[2]: *** [bibtex] Error 2
```

latex report



```
Missing character: ""
There were undefined citations.
Typesetting of "report.tex" completed in 0.9s.
```

---

## Compliance Report

---

```
name: Budhaditya Roy
hid: 348
paper1: 100% Oct 25 17
paper2: 100%
project: 100%
```

```
yamlcheck
```

---

```
wordcount
```

---

```
6
wc 348 project 6 579 content.tex
wc 348 project 6 5477 report.pdf
wc 348 project 6 439 report.bib
```

```
find "
```

---

```
103: Do not use "these quotes" but use these ``these quotes''.
```

```
passed: False
```

```
find footnote
```

---

```
113: \footnote{do not use footnotes}.
```

```
passed: False
```

```
find input{format/i523}
```

---

```
4: \input{format/i523}
```

```
passed: True
```

```
find input{format/final}
```

---

```
passed: False
```

```
floats
```

---

```
62: In Figure \ref{f:fly} we show a fly. Please note that because we  
use
```

```
69: \begin{figure}[!ht]
```

```
70: \% \centering\includegraphics[width=\columnwidth]{images/fly.pdf}
```

```
71: \caption{Example caption}\label{f:fly}
```

```
86: or generate them by hand while using the provided template in  
Table\ref{t:mytable}. Not ethat
```

```
89: \begin{table}[htb]
```

```
92: \label{t:mytable}
```

```
figures 1
```

```
tables 1
```

```
includegraphics 1
```

```
labels 2
```

```
refs 2
```

```
floats 2
```

```
True : ref check passed: (refs >= figures + tables)
```

```
True : label check passed: (refs >= figures + tables)
```

```
True : include graphics passed: (figures >= includegraphics)
```

```
True : check if all figures are refered to: (refs >= labels)
```

```
Label/ref check
```

```
106: Do not use Figure 1 user the ref for the figure while using its  
label
```

```
passed: False -> labels or refs used wrong
```

```
When using figures use columnwidth  
[width=1.0\columnwidth]
```

```
do not change the number to a smaller fraction
```

```
find textwidth
```

---

```
passed: True
```

---

```
below_check
```

---

```
WARNING: figure and below may be used improperly
```

```
67: figure below.
```

```
bibtex
```

---

```
label errors
```

```
bibtex errors
```

```
This is BibTeX, Version 0.99d (TeX Live 2016)
```

```
The top-level auxiliary file: report.aux
```

```
The style file: ACM-Reference-Format.bst
```

```
Database file #1: report.bib
```

```
I was expecting a ',' or a '}'---line 61 of file report.bib
```

```
:
```

```
: chapter = "1",
```

```
(Error may have been on previous line)
```

```
I'm skipping whatever remains of this entry
```

```
I was expecting a ',' or a '}'---line 123 of file report.bib
```

```
:
```

```
: doi = "10.7/3-105.876",
```

```
(Error may have been on previous line)
```

```
I'm skipping whatever remains of this entry
```

```
I was expecting a ',' or a '}'---line 138 of file report.bib
```

```
:
```

```
: doi = "10.7/3-105.876",
```

```
(Error may have been on previous line)
```

```
I'm skipping whatever remains of this entry
```

```
Warning--I didn't find a database entry for "editor0"
```

```
Name 1 in "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Chanchala D. Kaddi, Me
```









```
while executing---line 3229 of file ACM-Reference-Format.bst
Name 1 in "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Chanchala D. Kaddi, Me
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha
while executing---line 3229 of file ACM-Reference-Format.bst
Warning--empty publisher in editor07
Warning--empty address in editor07
Warning--empty chapter and pages in editor01
Warning--unrecognized DOI value [NA]
Warning--empty address in editor04
Warning--empty chapter and pages in editor04
(There were 107 error messages)
```

bibtex\_empty\_fields

---

entries in general should not be empty in bibtex

find ""

---

4: author = "",

```
13: chapter =      "",  
14: page =      "",  
22: editor =      "",  
32: pages =      "",  
40: editor =      "",  
45: address =      "",  
55: editor =      "",  
64: address =      "",  
70: editor =      "",  
76: publisher =      "",  
84: editor =      "",  
109: address =      "",  
122: address =      "",  
137: address =      "",  
passed: False
```

```
ascii
```

---

```
=====
```

The following tests are optional

---

```
=====
```

Tip: newlines can often be replaced just by an empty line

```
find newline
```

---

```
=====
```

passed: True  
cites should have a space before \cite{} but not before the {

```
find cite {
```

---

```
-----  
passed: True
```

# Diversification of Big Data

Shiqi Shen

Indiana University Bloomington  
1575 S Ira St  
Bloomington, Indiana 47401  
shiqshen@indiana.edu

## ABSTRACT

There are some ideas around the conception of big data and how it is used in the market outside of a programmer's computer. What I mean to try, and state here is that there is an idea of how big data is used and functions in the world. As a programmer, we cannot just focus on what we contribute to the programs, meaning we cannot only look at our computer data and creation of such programs as a means for our own scientific endeavors and egos, we must understand how such creation are reflected into the world and how such programs create a very interesting market exposure. Therefore, we write this paper to analyze the enterprise of big data. We will use this paper to seek out the multiplicity of avenues in which big data is used by our technological world (mainly those that seek to use big data to create diversified consumer experiences).

## KEYWORDS

i423, hid109, Big Data; Social Media; Online, Shopping, Customers; Pricing; Dynamic, Internet, application

## 1 INTRODUCTION

In order to begin a type of observation around this topic, we need to first establish what types of enterprise we will look into in order to make our observations. We will look at Online Shopping, streaming services, and Social media. These forms of enterprise concentrate their use of big data to create a diversified user experience, one which looks to creating more possibilities and other forms of consumer products for consumers. We would also like to use such an observation to understand the use of big data in these types of enterprises and how such uses render experiences and technology of big data as a good form of consumer study.

## 2 OBSERVATION

Before we begin our in-depth observation, it is important to showcase the Big Data technologies, some components of big data that make it what it is. There is of course an array of technologies that facilitate and create big data. In the creations of big data, we can see many enterprises like the generations of web pages (in which individuals, corporations, government, and the like, produce these pages with data). We can also see digital imagers that facilitate data collection as well. These types of data can come from telescopes, MRI Machines, and Video Cameras. Another source of data production can come from biological and chemical sensors, things like microarrays and environmental monitors [21].

From production of data there must be a medium that collects this data, that is of course usually computers. These data can be collected in things like the internet and localized sensor networks.

Qiaoyi Liu

Indiana University of Bloomington  
3209 E 10th St  
Bloomington, Indiana 47408  
ql30@umail.iu.edu

These sources of course can both collect and analyze the data. From collection, there are also storage capacity for these data. In that, we find that there are many storage type disk (magnetic disk) that can hold tons of data. There are also cloud services with which data can be stored.

Big data is crucial to the developments of many businesses that drive on the interactions and recommenders of their clientele. When we are looking to understand big data and how it is used by big contenders, we must first look to the companies to see what uses they have for big data. Let us begin with looking at Amazon, an empire built on e-commerce, the selling and recommending of products to consumers at efficient and effective rates. When we look at how this company succeeds, we can see that the main component to its success comes from its unique and bold use of big data.

## 3 ALGORITHMS, PREDICTIONS, AND BIG DATA CRAZE WITH ONLINE SHOPPING

When a consumer is shopping on Amazon, they can click on the image of the product they would like to view. During this process, they will probably go back and look at other products to compare the product to other products that are similar. Therefore, what can a company do to ensure that they are helping their consumers with this process? By pooling this data into other similar consumer searches and sells. For Amazon to do this they have a very complex system of enterprise that is meant to survey this, their use of big data. But it also comes down to very simple standards that are used by consumers.

When a consumer is shopping on Amazon they can administer filters to help them narrow their searches of products that they like. But during this process of administering these filters, the consumer is also being surveyed of searches made in the past to better associate what the consumers wants and needs within the products that they are shopping for.

This type of use of big data allows for any company (mainly Amazon like companies) to focus their resources on the calculation of products that an individual consumer would want based on searches that they have made. But the best part about this is that the information is pooled on multiple similar searches and consumers to best devise the necessary searches to show a consumer what to buy [25].

But beyond this simple observation of the big data usage, we need to understand what forms of implementation must be taken into consideration when creating these types of programs. Now,

what really needs to be focused on here is the interpretation of the data that is collected by Amazon to provide consumers with their recommendations.

Within big data itself, the name gives away what it is, it is a cluster of data, a large, almost seemingly insurmountable amount. There must be some sort of way to interpret the data results as they come in. For this process, it needs to be understood that this data interpretation cannot happen within a void, rather what is done for these data to be interpreted and re-designated as recommendations back to a user would be through a process of examining detailed assumptions and rethinking the analysis [1].

This process of observing and interpreting big data itself can have issues, such as those of bug interference based on programs that are being used in order to interpret these large data pools, and data can become erroneous. However, a way in which these types of issues can be resolved comes in the form of predetermined data assumptions. Data assumptions that are made to help companies who use big data to narrow in on data pools to create a seamless connection of data gathered to products showcased based on data that is collected, interpreted, and re-designated to users. This is done to help devise the necessary sales goals or recommendation services that come from these companies use of big data.

Also, when we are looking at the process by which data is collected, there will be data that is of no importance to the necessary processes by which the data is collected for. The difficult task becomes that of filtering out the useless information without removing the information that is of importance. To do this, there have been advancements made within the scientific community to reduce the plausibility of such turn out to happen. The process seeks to monitor faultiness that can be caused by sensors to lower the chances of data that is useful to not be discarded alongside data of no importance. This is also where algorithms that are made to establish key components of data come into play as well.

These forms of interpretation of course come from algorithms that are used to detect the data in forms of patterns. A perfect example of this would be comparing both Amazon and Netflix's use of algorithms that recommend to users what to buy or watch next. Machine learning within the recommendation system is what enterprises this use of algorithms. The machine learning itself will compare the histories of purchase and views to establish a statistical model of a collective pool of millions of other users to generate the necessary continued recommendations to users. The use of algorithms and machine learning also helps to establish a base line of what it is that users like and are more attuned. The use of algorithms is to ensure that the data that is being collected is correctly sorted and re-designated to users in the forms of recommendations. This is because large quantities of data require these types of assumptions to calculate and extract knowledge from the data that is collected [21].

As a programmer, machine learning and the creation of these algorithms truly fascinate when they are being used in order to create a recommendation system. Since the collection of data is

pooled, the use of algorithms to reach a particular end, based on data analysis, makes for a very interesting usage. As the use of data algorithms not only seek to facilitate the necessary recommendations, but also filter through millions of data informatics to agitate the necessary products to ensure the transparency of recommendations created. This can also be seen in what was stated above regarding consumers being able to filter products with precise search options. These filters can act as filters for filters. As in the filters chosen by consumers help to facilitate precise sifters and allows for algorithms to pinpoint precise outlines of data that help to recommend even finer tuned recommendations for consumers.

Big data, from this perspective, can be seen as a general tool that can be created to make precise measurements that are used in order to facilitate the necessary recommendations for consumers. This is of course a process that is keyed out by the use of interpretations and algorithms that are necessary to the specifications made on big data that is collected. Therefore, big data is used precisely well as a tool to pool and narrow pattern like data to ensure that completed data of particular patterns can always correctly correlate to consumers as well as viewers who use services such as Amazon and Netflix (This paper will not observe Netflix in full as it did with Amazon since the use of big data analyzes and processes to creating recommendations for consumers are very similar).

Therefore, when we are looking at companies like Amazon and Netflix, we are looking at companies that use big data to create predictive analytics. Predictive analytics has many uses however and cannot just be subjected to Amazon and Netflix's use on commercial needs. The use is mainly attributed to uncovering patterns and highlighting relationships with the data that is being observed. Because of this very nature, big data that goes under predictive analytics are being done soon to search out these two main components. There is also the process of trying to find past data patterns outcome variables and trying to deduce them for the use of the future (observing patterns from a specific time-period to see if such trend continues again at another given future based on certain functions that existed in the data of the past and then comparing that to the future) [14].

Further delving into predictive analytics, which is the main use of big data for our commercial subjects, we see that there are also forms of linear regressions. The use of this is to find interdependencies within outcome variables and explanatory variables in order to use them in the process of making predictions or to right out make the prediction itself. This looks to focus the data that is collected into predictive measures that are precise to the sets of data that are collected and distributed through this type of analysis. Above, when looking at machine learning, this is categorized as a neural network. Neural networks are a collective entity, something like that of the human brain, if an artificial neural network can be defined as a computing system made up of number of simple highly interconnected processing elements which processes information by their dynamic state response to external inputs. Within the neural networks, machine learning works within the sphere of the networks to generate and learn from data collected to predict,

showcase patterns, and classifying input data [22].

To finish up the observation here on big data uses for predictive analytics and the basis of enterprise that is that of Amazon and Netflix's ability to use big data to create such systems, the paper needs to understand a little further into the principles of how this portrays use in consumer settings. What is meant to be clarified here is how big data effects the process by which consumers use services and are data mined. To affectively understand this component, there needs to be an understanding that big data and it cultivation are just as it is named, a collection of data on a mass scale meant to just be data. It is however, the enterprise of the scientific community as well as commercial bodies that induce a specific plethora of uses to assimilate the necessary components to use big data effectively. Such things make it to where consumers have an easier time with their collective use of these commercial branches (Amazon and Netflix, as well as unnamed companies). Consumers use of these sites creates data, a process by which these companies then use the process of data mining to achieve the best standards of prediction and analytics that help them to enforce their use of big data as their tool in apprehending consumers by prediction. Because of this process, consumers are the ones who are helping these companies further develop their own uses of big data by allowing these corporate bodies to data mine them, collect data, and analyze the data provided. But this process is crucial to the experience of the consumer, as this data helps with creating the necessary components of these corporate bodies, allowing them to further create advanced algorithms that help these corporate bodies devise the necessary recommendations and make the predictions to the habits of these consumers. That makes it easier for consumers to use these products offered by these corporate bodies.

In all, big data, with modifications from things such as data mining, data collection, machine learning, algorithms, and prediction analyst are all components which excel the use of big data (because of the necessary enterprise that it takes to stay updated with this matter) and insure the that consumers and users of big data can reach out to their own platforms easily.

#### 4 PRODUCT RECOMMENDER SYSTEM

In the recent past, Amazon has moved from operating as a pure e-commerce firm to a major player in the internet services industry, with focus on offering a wide variety of services to both individuals as well as companies. The firm started to shift its focus on big data and started the journey to transition from a typical online retailer into one a major force in the realm of big data. Around 2000, the company, along with other internet firms such as Google, Yahoo, and Twitter realized that they had voluminous data about their customers, which could be put used to improve their performance. Although the other firms did not initially concentrate majorly on big data, Amazon swiftly moved to take advantage of the invaluable database of individuals who used its e-commerce platforms around the world to shop. The team charged with the responsibility of recommending the products to the customers came up with innovative strategies that the firm could make use of the data collected

by the firm about their customers. The end result of the move was a huge success in big data, which revolutionized how the company did business.

As a major player in the e-commerce domain, the success of Amazon was always pegged on availing the right products to the customers. The efficacy of providing the right products for the customers in turn largely depended on a proper understanding of the needs of the consumers. A proper market research was necessary in order to understand the customer's needs and tastes. Since it was founded, Amazon has created a name for itself because of its superior product recommender system, which suggests products to consumers on the basis of their last purchase. The major driving force behind the recommender system is the data gathered from the customers.

The product recommender system is essential for the personalization of each customer's experience when they are shopping in the firm's online store [25]. The firm employs collaborative filtering and clustering algorithms to classify clients on the basis of preferences. Customers are grouped on the basis of same search as well as collaborative filtering between items. Content-based search employs the shopping history of customers and item ratings to establish a search query capable of finding other items that match the tastes of consumers. For instance, if a customer purchases a book, the product recommender systems will suggest books from the same author, publisher, or subject area. The product recommendations are not only used by the company in the online stores, but it also doubles up as a marketing tool useful in conducting email campaigns. There is a recommendation link that enables shoppers to filter products by several criteria depending on the items that they have in their shopping carts.

#### 5 BIG DATA FOR DYNAMIC PRICING

Dynamic pricing entails the use of big data such as clickstreams, purchase history, cookies, etc. to offer customized discounts to customers or to alter the prices of items being sold dynamically. The technology enables the real-time price customization for an item to suit a specific customer. This explains why it is sometimes possible for two different sets of customers to buy the same item at different prices from the same online store [23]. Despite the immense benefits of this technology, some customers may always feel discriminated against due to the price differences. Amazon has successfully used the power of big data to implement a price discrimination system. For example, there was an incident in which some Amazon customers were aggravated about price variations of a certain DVD. One of the customers noted that there was a difference of nearly two points five dollars in the price if the cookies were deleted from the computer. Price discrimination was also experienced in the sale of a product known as Diamond Rio MP3 Player.

Big data also enables price optimization. This enables the firm to manage the prices of commodities and grow its profits by twenty-five percent annually. Several factors are used to set the prices of commodities. Some of them are: activity of the customer on the

firm's shopping portal, availability of the product, competitor's prices, order history, item preferences, and the anticipated profit margin [23]. The prices are normally refreshed every ten minutes as big data become updated. Due to this, Amazon provides customers with discounts on best-selling commodities and accrue large profit margins on the items that are less popular with customers.

## 6 BIG DATA AND CUSTOMER SERVICE

Big data is also extensively being used for customer service at Amazon. The acquisition of Zappos has often been viewed as a major element in the same. Since it was founded, Zappos has enjoyed a good reputation for the excellence in customer service and was usually viewed as a world leader in this domain. Much of the success can be attributed to their advanced relationship management systems which extensively employed their own customer data. After the acquisition of the firm in 2009, the procedures were integrated together with those of Amazon. Today's business environment is changing at a rapid rate, and consumers are also using their voices faster. Within a few moments after undergoing a bad experience, customers can swiftly move into social media and spread the news about their negative experience [17]. The only strategy for an organization to survive under such conditions is to employ the power of analytic to streamline and shorten the response time, as well as fix the customer support issues. The customers of the present day are not only looking for a product that works, but also one that is personalized and able to recognize their interests and save them time.

## 7 ONE CLICK ORDERING

Amazon used big data to create one-click ordering. This feature is activated automatically when the customer places his first order, enters a shipping address as well as a method of payment. When using the one-click feature, the customer is given thirty minutes to change his mind about the particular purchase. This system was created on the premise that a simplified path to purchase would increase conversion rates. Since the introduction of the technology, the firm's revenues have increased year after year. The significance of this application pushed the company to patent it to prevent other companies from using it without authorization. Reorganizing the purchase process is currently one of the most significant differentiates in the current marketplace. The service enables users to make payments without having to exchange cards or money physically. Amazon has also greatly benefited from impulse buying, which is accelerated by one-click buying. Research has shown that the largest percentage of people normally purchase things they don't require or did not plan to purchase in the first place [5].

## 8 USING BIG DATA TO SUPPORT OTHER COMPANIES

Amazon also uses its big data platform to support and help other companies improve their operations. Organizations can employ AWS toolkit provided by Amazon to create scalable big data applications that have the capacity to improve business performance [25]. Besides, they would be able to secure these applications easily without the need to spend on expensive infrastructure and hardware. The big data applications including data warehousing, clickstream

analytic, fraud detection, internet of things, and several others are delivered via cloud computing. Hence, there is no need for an organization to incur additional costs in setting up a data center. The Amazon web services can enable companies to analyze spending habits, customer demographics, and other related information to enable them effectively cross-sell some of the firm's products in patterns similar to Amazon. That is to say that the retailers will also be able to stalk their customers, recommend products to them, and improve their customer experience.

## 9 BIG DATA TECHNOLOGIES

**Amazon EMR:** This technology offers a managed Hadoop framework that simplifies and hastens the processing of huge amounts of data across scalable Amazon EC2 instances. Amazon EMR also supports other common distributed frameworks including HBase, Apache Spark, Flink, and Presto [3]. Besides, it reliably and safely handles a wide range of big data use cases, such as web indexing, log analysis, financial analysis, machine learning, and bioinformatics.

**Amazon Athena:** It denotes an interactive query service that simplifies data analysis in Amazon S3 via standard SQL. Since it is service less, one only pays for the queries they run and there is no infrastructure to be managed [3]. The technology is quite straightforward and delivers results within the shortest time possible. Moreover, it does not require complex ETL jobs to prepare data for analysis.

**Amazon Kinesis Firehouse:** This is one of the simplest methods to import streaming data into Amazon Web Services. The technology can be used to gather, transform, and import streaming data into Amazon S3, Amazon Kinesis analytic, and Amazon Redshift, to permit instant analytic with the current BI tools and dashboards currently being used. It is a comprehensively managed service that can expand automatically with the increase in data throughput.

## 10 UNSTRUCTURED DATA AND AI COMPONENTS OF SOCIAL MEDIA

Next, we will look to observe big data and its uses within social media. First and foremost, it is important to understand that social media is an outlet that is massive. The many posts, tweets, likes, shares, and other social media actions all develop unstructured forms of data that are then considered by corporations to understand the market of users. It is within the creation of this unstructured data that creates such an importance to talking about social media and its perfect relationship to big data. Because of the importance of social media to businesses (due to trend and the fast-paced living environment we live in, if a business is behind on trend it becomes behind on sales and other forms of innovations), there is a large component of dependence towards retrieving big data from social media to calculate and predict trend. But beyond just trend, businesses are also trying to get their hands on the enormous amount of big data that exists within the social media sphere. There is recognition of the value of unstructured data that is sourced within social media. The value comes from consumers using social media to broadcast what they are thinking, want, and are doing. These

types of information, one might think it private, but the internet is a very transparent source of data. In so, businesses value the perspective of consumers and create ways in which they can follow through with interacting on a very contingent basis, they data mine and from that, they advertise based on the collected big data from social media [12].

This brings to focus the use of advertisement and the necessity for big data to be used. Within digital advertising, the one who collects and analyzes big data effectively and efficiently with accurate uses is king. To be successful in this method of advertising, businesses need to be prodigious at their collection of data, integration of that data and analysis of that data. The reason being is if these three things are managed well, the use of the data is much more successful. What is challenging about this type of work however is that a majority of the data that is collected from social media is unstructured, it is in a word, messy. These forms of unstructured data are in our own use of social media, usually within posts, videos, tweets, photo post to Instagram, mass use of Snapchat, etc. Because these forms of data are so much more unstructured and more difficult to analyze using traditional analysis methods, businesses needed to enterprise methods for them to collect these data forms and have the necessary big analytics platforms to analyze the data. Keep in mind the data has the most information about us, therefore the use of these unstructured data is key to devising targeted and precise marketing executions [20].

There is also the collection of real-time data. Because of the advances within the technology, the process by which real-time data can be analyzed has sped up exponentially compared to the past where this process would be impossible and yield no results. With the ability to now look real-time data and have the capacity to analyze it, businesses (those that are marketers) have the possibility of taking action instantly to provide consumers on social media with their own personalized ad. The use of personalized ads of course comes from the massive amounts of data that consumers put out on social media, allowing marketers to collect these data (things we like, talk about, and do daily). And because they have these data, they can target specific ads to consumers without missing a beat. But what happens when we begin to incorporate other technologies that allow us to always be able to access our social medias? Mobile devices provide the quintessential provisions needed for data to have a constant flow to advertisers. Big data then is a facet within the life of social media. It must be done effectively in order to continue its uses of data collected and then expounded on to get customers to buy products or even interact with specific businesses. With the development of the mobile device, location also becomes part of big data, as it allows for your location to be collected, you leave not only a digital footprint, but also a physical one which can be collected as data and used [20].

Because of such enterprise, big data can gather almost every facet of information that is readily available to the internet. Such a mass look on data also comes back to the revision on algorithms that are meant to specify what is being observed and analyzed in the data. Of course, for these algorithms to work, there has to be a steady stream of data in order for the algorithms to process and do

its job. Because of the accessibility of data through the means of mass social media and how quickly consumers use and are exposed to social media (due to mobile phone), businesses are more creative in their approach to their algorithms. These algorithms can now pop up wherever consumers are on the social sphere. In doing this, big data itself changed the advertising platform. Advertiser now must create enticing messages from big data to continue their reach to consumers. The change is incredibly eye opening. As the use of data itself can create a massive alteration in how a marketer begins to try and craft their ads to highlight what it is that individuals want. Advertising becomes more individualized and work closely around the sphere of social media to allocate their ads effectively [9].

Within the use of big data, there is also the use of AI to help with the process of analyzing big data. AI creates a much more effective measure when it comes to analyzing hundreds and thousands of data in detail. Because AI can do that, it allows for businesses to have a better idea around what perspective they themselves must take when advertising based on detail production from AI technology. AI technology becomes a very important component to being able to go analyze big data in order to provide the necessary specific information that would help with creating the necessary ads [12].

We are then looking to see how this affects the user, or better how this type of big data in social media affects consumer experiences. The use of big data comes back to the work of the consumer and the business. The consumer creates big data from their usage of social media, social media in turn collects and responds to the data that is being created by the consumer (user). Through the process of facilitating the data, analyzing it, interpreting it, pooling it, and filtering it through algorithms and AI technology, consumers using social media get access to tailored advertisements. The consumer experiences a personalized social media experience and personalized advertisement trail based on the collected data that is reinforced by big data that is pooled together in order to do this. Does big data have any other uses in social media then? Yes, it does.

Social media can also use big data in order to create studies and infiltrate certain components of a consumer's private life. Facebook for example, a top social media company prides itself in its technological advances that use big data. Facebook uses is considered a top user of digital advertisement, so it begs to question what else does Facebook do with the mass big data that it collects? Facebook has also used its big data resources to try and act on social media experiments. During a time when there were the I Voted stickers sprawled on Facebook for users to share and gimmick that they had voted, Facebook was using this in order to incite and boost voter turnout. The method was to first isolate and use the stickers with particular groups of people, small groups at first in order to test the role. After having enough data collected on the presence of the stickers and what they meant for users, Facebook began to mass data span by incorporating the stickers in a much more massive turnout. With midterms of 2010, there is studied on behalf of Facebook scientist, that say 340,000 more people voted in the 2010 midterms [18].

From these the observation of these big data uses, we can showcase the how and what makes big data so important to the creation of diversified consumer experiences but also showcase the necessary components to how big data is used by the new technologies we have. It is however imperative that we understand these different phenomena that come from the use of big data.

## 11 SOCIAL MEDIA IS SIGNIFICANT FOR COMPANIES AND INDIVIVUALS

Although big data is said to come from several different sources, the largest proportion of it is said to originate from unstructured sources. As it can be imagined, social media makes up the largest source of unstructured content for big data. All the activities that users perform on social media such as views, retweets, comments, favorites, likes, etc. can be gathered and explored by interested individuals.

In the current digital world, social media plays a vital role in many companies. Having a presence on various social media platforms such as Instagram, Facebook, and Twitter is imperative since it enables individuals to interact with an organization on an ostensibly personal level and at the same time helps businesses across several domains get in touch with their customers. Currently, Facebook alone has over two billion users on their platform; this is roughly twenty-six percent of the world population [12]. It is therefore important to consider the fact that big data, from the social media platforms, can reach any people in different forms. Besides that, social media interactions have continued to play a big role and will continue to play a big role in business decisions. For example, some insurance companies have declined to offer life insurance policies to individuals solely based on their social media posts. If you frequently post, on any of these platforms, about how you are drinking or going to drink, insurance companies would be reluctant to offer you a life insurance policy as this is a risk to them.

It will not be long before organizations discover new and better strategies for making sense of big data. But, at the moment, the concept of big data is still new and rapidly evolving. Nevertheless, some businesses have found ways of interacting and using this data, which is just but the beginning, but still a good way to begin. To elaborate, a marketing company whose interest is promoting a new product could employ machine learning algorithms that enable it to gather data from individuals who meet certain attributes [12]. Consequently, by employing artificial intelligence technology, they will also be capable of drawing insights from millions of users and create campaigns. This will increase their levels of precision and focus, a technique usually referred to as targeted marketing, and present an excellent opportunity for finding the perfect audience and satisfy its preferences.

## 12 BIG DATA IN SOCIAL MEDIA ADVERTISING

Fundamentally, advertising revolves around communication since it is all about sensitizing consumers on products and services that an organization is selling. However, different consumers will always want to hear varied messages, which is a vital fact to consider when

new clients are being recruited into the internet bandwagon due to the growing popularity of smart phones. Big data has the capacity to customize these messages, project what consumers would like to hear, and establish new perceptions on what customers like or prefer [8]. The above steps are all revolutionary and are expected to have a significant impact on how marketers in various organizations advertise.

Furthermore, there are some occurrences which several people do not view as advertising but are still interactions between big data and marketing like product recommendation. An obvious example is Netflix [9]. Although the company does not have a concrete advertisement plan, it employs a lot of algorithms to recommend various movies and shows to its customers. The approach saves the organization a lot of money by reducing the rate of customer exit and ensures that the right shows are marketed to the right individuals. The company's strategy is to target consumers with shows specifically tailored for them. Apart from them, other firms such as Amazon, YouTube, etc. also do the same by using product recommendation to target their customers [9]. In order to stay up to date, the algorithms need constant flow of data to help it work more efficiently. With the growth of the internet, users leave huge volumes of data not only on social media platforms but also on other places they visit online in the form of a digital footprint. This provides advertisers with new avenues to tailor their messages to meet their customer demands.

The digital footprints left by online advertisers provides new insights to marketers on what a consumer really needs, and this sometimes may be more accurate than what the customer actually says on social media. However, marketers are worried about how to safeguard the privacy and security of their consumers and therefore companies that are careless in handling data collected from consumers usually ignite a backlash which greatly impact their business. Even though targeted advertising has been in existence for quite a while [9] the more the data that is collected by advertisers, the more personalized and effective marketing is expected to be. Organizations will strive not just to gather as much data as they can, but also to gather information which typically represents the individual consumer's needs in order to enable them to market to their personalized tastes.

## 13 ANALYZING LINKS

Big data collected from social media can lead to the discovery of new information regarding each individual customer that can help in creating a customized appeal to that specific customer. However, with the new insights, marketers can enhance how advertising is approached as they create new strategies. The new growth in content marketing is usually perceived as a primary beneficiary of big data, although the concept of content marketing could be older than the internet itself.

Another essential point is that big data enables digital marketers to target users effectively with more personalized advertisements which they might prefer to see. Facebook and Google are among the biggest players in this domain of digital advertising. They have

discovered excellent ways of creating and delivering more appealing advertisements in ways that do not intrude on the rights and preferences of the consumers [10]. Most of their advertisements feature services and goods that consumers would like most to enhance their lives and almost all of these advertisements are reliant on huge amounts of personal data that users usually provide from what they are up to, what they share and like things online.

Experts contend that it is possible to accurately make predictions on an array of individual attributes that are more sensitive merely through an analysis of an individual's Facebook or Twitter likes [20]. For example, the likes on these social media websites are critical in predicting one's religion, sexual orientation, emotional stability, life satisfaction, age, relationship status, and many other attributes. Companies like Facebook successfully linked political activity with user commitment when they created a sticker enabling most of their users to declare on their profiles that they had voted. The initiative was conducted during the 2010 midterm polls and was very effective as more people turned up to vote as compared to the 2006 midterm elections [18]. Individuals who saw the feature had high chances of voting and actively engaged in a conversation about the same after seeing their friends and peers participate in the activity. Later on, during the 2016 polls, Facebook escalated their role into the voting process by providing users with not only constant reminders but also with directions about their polling stations [19]. Apart from that, they also enabled users to easily get access to registration information, news, voting guides and other tools that would have made them more equipped to go through the election process.

## 14 USER RATING AND POP UP ADS

Depending on the user preferences and the content that they often access on social media, pop-up advertisements can be created to target users every time they are online. For example, an ad can be created on the Facebook Messenger app to open inside that particular app every time the user hits the CTA button. When clicked, such ads would redirect the user to a page where they would be required to answer some question, claim a reward or send some feedback regarding a product or service. Before creating such ads, it is imperative to establish a custom audience of the individuals who would be targeted with that particular pop-up ads. For instance, individuals who have previously liked the company's products on their Facebook page or other social media sites can be included on the list of target audience to receive the ad [4]. Another strategy that can be employed is to rate users by tracking their cookies. In most cases, user activities are usually tracked across the internet using cookies whenever a user logs into one of the social media sites and is concurrently browsing other sites. Whenever this happens the other sites that the user is visiting can be easily tracked and the data used accordingly.

## 15 RELEVANCY OF BIG DATA ANALYTICS IN GROCERIES STORES

### 15.1 Increases the customer shopping experience

As per a current SHSFoodThink white paper "Are We Chain Obsessed?" 64% of customers said that the previous shopping experience is what makes them keep coming back! not the items themselves [24]. By utilizing bits of knowledge received from the information transaction database, online networking, promotional activity, customers purchasing behavior, and client movement patterns, grocery stores can find a way to guarantee they are engaged with their customers that matter most.

For instance, they can investigate customers shopping movement to enhance the layout of their store, or recognize attrition risks for clients who have not as of late bought staple things, similar to milk. In like manner, chains can construct item varieties demonstrated with the customer needs and purchase patterns in certain regions [2, 13, 24]. Regardless of whether it is through reconsidering store layout or furnishing store attended with mobile apps to better serve clients, analytics can enable grocers to change consumer's expectations.

### 15.2 RESTRUCTURE THE SUPPLY CHAIN

Grocery stores can likewise utilize analytic to investigate the production of their products, monitor production processes, and quality control, and improve straightforwardness with buyers about their sustenance production practices of foods [16]. Suppliers remain to profit from the evaluation also, with access to secure, customized content of information identified with performance sales of the product, stock, margins, and marketing effectiveness. Giving supplier an opportune profitable business knowledge that supports joint ventures, drives performance, and decreases waste products

### 15.3 BUILD SUPERIOR MARKETING PROGRAMS

Loyalty programs furnish grocery merchants with an abundance of data to enable them to distinguish client segments and precisely characterize item preferences. By joining this information with different data sources! like healthful patterns, favored technique for accepting marketing promotion, customer movement patterns, and weather-related event! grocery merchants can concentrate on enhancing, and derive income from, the general shopping experience [24]. For instance, grocery retailers can utilize analytics to customize the advancements they offer to clients given what they are well on the way to buy. They can likewise time advancements fittingly, and offer codes to customers who often as possible buy certain things.

### 15.4 IMPROVES HR STRATEGIES

Supermarket stores utilize analytics to manage work-related decisions. Information freely accessible through online networking accounts and different means can be examined in conjunction with a grocer's internal information to direct decision identified with selection and recruitment, employee termination, and performance management and advancements [11]. For example, an investigation

of late action on LinkedIn can reveal insight into which representatives are destined to leave an organization.

Grocery merchants can likewise break down information to control the advancement approaches that will build workforce performance. For example, they could explore different avenues regarding organizing a social gathering for representatives at a subset of their stores, and analyze information on profitability, morale, and turnover in the preceding months [13]. They may find that the gathering information prompted a more positive workplace where workers feel more noteworthy engagement at work, and soon after that, they could roll the strategy out to different stores.

## 15.5 USING BIG DATA FOR COMPETITIVE ADVANTAGE AND ATTRACTING CUSTOMERS

Numerous grocery stores have been utilizing transaction and client information for a considerable length of time, despite the fact that many still have not completely used all that can be proficient with these types of information. For Small to Medium Sized grocery merchants, many have swung to subcontracted point solutions because of an absence of available analytics assets and potential framework investment required [11, 24]. The issue with point solutions recently is that if? they independently work out for a particular business section and the evaluation is cookie cutter. In this way, the 'information' is not coordinated and hard if not difficult to give an all-encompassing picture of client conduct overall touch focuses for instance. Nor are the investigations offering a cross-functional observation that is pertinent to all business partners as far as driving differentiation in the commercial center in promoting, advertising, store operations and supply chain.

As far as utilizing 'new' data sources, for example, mobile, social and text, the industry is particularly occupied with a discovery phase of investigation with an assortment of center sections, testing and figuring out how to extricate an incentive from these rich new sources of information. There are two common paths grocery merchants takes with little respect of the 'size' of the organization: to start with is Strategic Commitment, in which there is C-level (hierarchical) commitment making the venture in the assets to get the majority of the in-house data and evaluated it [13].

Presently like never before, information, analytics, and IP are seen as vital resources and competitive discriminators. The other is Business Discovery; in which grocery merchants outsource to an Analytics as a Service firm to use internal and external information. Performing analytics speeds the construction of business advantages creating new users case and helps catch 'quick wins' before making resource commitment to technological innovation and human capital in advance [11]. In view of progress, and a wit, trusted stakeholder willing to share the techniques and explanatory models, can assist grocery merchants to proceed with an outsourced administrations supplier or relocate the data, analytics in addition to IP in-house.

## 16 RECOMMENDATIONS

### 16.1 Real-time insight on product demand

Nowadays, retailers can get to information on item demand levels instantly on a chain of stores. Nevertheless, numerous merchants are still in the earliest stages in regards to evaluating and monetizing the huge amount accessible data [2]. This prompts stocking deficits, for example, evaluating item demanded based exclusively on past historical information. It can likewise convey about wrong promoting endeavors: If a customer purchased ketchup on Saturday, an email coupon for it on Sunday is not well planned and make little sense to the shopper.

This is the place data from store loyalty programs in addition to credit card sales can prove to be useful. Its data can be utilized to define needs of the customers in future. For example, grocery merchants can use data analytics to decide how regularly customers purchase sugar, flavors, or different items, and after that send every family unit coupons given their propensity to buy [24].

### 16.2 Enhancing in-store stock management

Perishable basic supplies, for example, dairy, meat, and fish call for precise stock administration, regularly on an hourly premise. Client analytics and prediction tools can enable grocery merchants to calibrate their inventory levels by assessing buyer purchasing behavior and requested products from various viewpoints and situations [24].

For example, grocery retailers might need to screen cycles like when customers go for particular nourishment, purchasing patterns amid sales deals when storing activity peaks or seasonally inspired buys. As indicated by a report from Manthan, this methodology worked for U.K. food grocery merchant Waitrose: a deeper understanding of buyer purchasing behavior and demand outlines using cutting edge client analytics and predicting tools helped the store [11]. Concurrently, retailers can utilize these systems to all the more deftly change their stock levels and amplify high-buy products.

### 16.3 Leveraging Predictive Analytics

Amazon spearheaded item proposal engine: the "if you purchased that, you may like this" invention. This strategic changing web-based shopping feature mirrors the retailer's profound assessment of buyers' shopping basket. Proposal engine is intended to enable customers to find items they were not sorting out but rather would be interested in purchasing [13]. Today, general grocery merchants are progressively tapping the global innovation behind proposal engine: predictive analytics. This kind of assessment measures future patterns in light of present and past information, and it can enable stores to improve business. Information is driven, all-encompassing assessment of "purchasing triggers, for example, regularity, weather, stock, and advancements, is progressively informing grocery stores' product blend, marketing plans, and sales forecast [2]. Furnished with these information-driven tools, stores can better distinguish what items customers need today and what they will be demanding in future, and this learning will enable them to stay competitive for a considerable length of time to come.

## **17 INTRODUCTION**

Digitization set apart by an increasing number social media and mobile devices is shifting the business landscape in every sector insurance included. The opportunity presented by this aspect for insurance companies are immense. Communities and social networks enable insurers to interface with their clients better, which to their advantage improves branding, customer retention, and acquisition [24]. Insurance companies additionally get a plenty of contributions from computerized data as feedbacks, which likewise can be utilized to develop unique products and aggressive valuing. Digitization of big data analytics offers numerous opportunities that Insurances Company can harness to detect fraud among their customers. Dealing with fraud manually has dependably been expensive for insurance firms regardless of the possibility that maybe a couple of minor fraud went undetected [6]. What's more, the trends in big data (the evolution in unstructured information) are prone to numerous fraud, which can go without notice if analysis is performed correctly. In the proceeding section, the article will examine important of big data in insurance fraud detection and its relevancy.

## **18 IMPORTANCE BIG DATA AND INSURANCE FRAUD DETECTION**

Conventionally, insurance firms utilize statistical models to recognize fraudulent cases. These models have their limitation [15]. To start with, they employ sampling techniques to assess information, which prompts at least one fraud going unnoticed. There is a punishment for not performing a proper assessment of the data provided. Subsequently, this strategy depends on the cases analyzed before. Therefore, every time different fraud takes place, insurance firms need to manage the impact for the first time. Lastly, the conventional strategy works in silos and is not correctly equipped for taking care of the natural developing wellsprings of data from various diverts and diverse capacities in an integrated way. Analytics tends to be difficult and assumes an exceptionally pivotal part in fraudulent recognition for insurance firms. In the proceeding section, the significant benefits of utilizing big analytics in fraud detection assessed.

### **18.1 Identification of low incidence events:**

Utilizing sampling methods accompanies its particular arrangement of acknowledged mistakes. By using analytics, insurance can manufacture frameworks that go through every fundamental datum. This like this distinguishes events with low frequency (0.001%) [7]. Methods such as predictive modeling can be utilized to altogether break down processes of fraud, channel clear cases, and allude low-rate fraud cases for facilitating analytics.

### **18.2 Enterprise-wide solution:**

Analytics help in building a global point of view of the anti-fraud endeavors all through the undertaking. Such a point of view regularly prompts dominant fraud location by connecting related data inside the association. Fraud can happen at various source focuses premium, claims or surrender, application, employee-related or outsider fraud. In the meantime, insurance channel broadening is

adding to the breakdown of identifiable information. Insurance-related exercises should be possible using cell phones separated from the conventional face-to-face and online Insurance [15, 25]. This can be seen as an expansion to data storehouses in the Insurance business. Given more prominent channel enhancement and the development of ranges where fraud can happen, it is vital for insurers to have reachable enterprise-level data about their business and clients.

### **18.3 Data Integration:**

Analytics assumes a vital part in incorporating information. Viable fraud recognition abilities can be worked by joining information from different sources. Analytics additionally help in integrating inside information with outsider information that may have predictive significance, for example, public records. Information sources with derogatory properties are on the whole public documents that can be incorporated into a model. Cases include liquidations, liens, criminal records, judgment, abandonment, or even deliver change speed to show transient conduct. Different sorts of outsider information can be useful in upgrading effectiveness, for example, audit evaluating data to decide whether harms coordinate portrayal or misfortune or injury being guaranteed [6]. A standout amongst the most under-used information sources is doctor's visit expense audit information. This information, if utilized as a part of a model legitimately, is a gold dig for organizations researching medical fraud. Revealing peculiarities, in charging and adding these to the next scoring motors or interpersonal organization analytics will diminish the measure of time an agent or expert spends endeavoring to pull the majority of the pieces together to recognize deceitful action.

### **18.4 Harnessing Unstructured Data:**

Analytics is useful for getting the best incentive from unstructured information. Fraud can be delicate or hard. This depends on whether it comprises of a policyholder's misrepresented cases, or on the off chance that it contains of a policyholder arranging or creating a misfortune. At an abnormal state, fraud can happen amid commission discounting, because of false documentation, an arrangement between parties or from is offering [24]. Albeit bunches of organized data is put away in an information distribution center as a component of numerous applications, a significant portion of the vital data about a fraud is in unstructured information, for example, outsider reports, which are not assessed. In most insurance firms, data accessible in online networking is not suitably stored. An uncommon investigative-unit specialist will concur that unstructured information is vital for fraud examination. Since textual information is not straightforwardly utilized for reporting, it does not discover a place in most information stockrooms [7]. This is the place content examination can assume a crucial part in checking on this unstructured information and giving some valuable experiences in fraud discovery.

## **19 RELEVANCE OF BIG DATA IN INSURANCE FRAUD DETECTION**

Big data analytics is a reality for the insurance company because of its capability to enhance various conventional technologies and

be used to detect fraudulent acts. In the proceeding section, the relevance of big data and insurance fraud detection will be examined.

### 19.1 Text analysis

In numerous Insurance fraud recognition ventures, from 33% to oneportion of factors utilized as a part of the fraud location model originate from unstructured content data. This is particularly helpful for long-tail claims, for example, damage claims, because the best information frequently is found in claim notes [15]. Content mining is something beyond keyword sorting. Excellent content analytics apparatuses translate the importance of the words to establish context. Innovation that is adroit at preparing common dialect can help remove factors from the unstructured content that can be utilized for assist fraud modeling.

### 19.2 Data Management

Regardless of where your information is stored from legacy frameworks to the valid information stockpiling structure, Hadoop an information administration framework can enable insurers to make reusable information rules. They give a standard, repeatable strategy for enhancing and incorporating information [7]. Preferably, you need a framework that interfaces with different information sources. It ought to have streamlined information league, relocation, synchronization, organization, and visual assessment.

### 19.3 Event Stream Processing

This enables insurers to investigate and processes in movement (i.e., process streams). Rather than putting away information and running questions against data, you store the inquiries and stream the data through them [24]. This is foundational to both ongoing fraud identification (invigorating fraud scoring) and successful utilization of great high-speed information sources similar to vehicle telematics.

### 19.4 Hadoop

A free programming structure that assesses and prepares of tremendous collected information in a distributed environment of computing. It offers gigantic details stockpiling and super-quick processing at around 5 percent of the cost of convection less-adaptable databases. Hadoop's mark quality is the capacity to deal with organized and unstructured information (counting sound, text, and visual), and in expansive volumes. Insurers either can employ Hadoop specialists to exploit the structure or purchase items that scaffold to existing databases and information distribution centers[6, 7]. This foundational innovation for making predictive analytics models stays one-step in front of fraudsters and spillage of paid-out cases cash. The exchange observing advancement innovation used to battle regularly complicated illegal tax avoidance utilizes Hadoop as a center stockpiling and sorting out innovation. Complex organized crack rings and therapeutic factories, for instance, are conveying progressively modern techniques for laundering cash stolen from auto insurers.

### 19.5 In memory

In-memory analytics is a processing style in which all information utilized by an application is put away inside the principal memory

of the computing condition. Instead of being available on a disc, the data stays suspended in the mind of useful sets of PCs. Different clients can share this information with numerous applications in a quick, secure, and simultaneous way. In-memory analytics likewise exploits multi-threading and distributed registry [6, 24]. This implies clients can disseminate the information (and complex workloads that process the data) over different machines in a group or inside a single server condition. In-memory analytics manages questions and information analytics, yet also is utilized with morecomplex procedures, for example, predictive analytics, machine learning, and analytics. The sorts of neural-network analytics that assist insurer in discovering association among suspects sustaining claim and premium fraud depending on the kind of processes

### 19.6 Software as a Service (SaaS)

Predictive modeling and different analytics were accessible to large insurance net providers willing to introduce the innovation on location as of not long ago. Software as a service has advanced to even where genuinely little insurers can exploit Big Data analytics [6]. Insurance providers subscribe to a service keeps running by a seller as opposed to paying for the vast buy, establishment, and support of in-house frameworks. SaaS likewise is named "on-demand software."

## 20 DISCUSSION

Form such progress we can see the diverse reach that big data has and how it affects the users in their experiences with big data. Not only do we see big data creating advertising products to consumers, we are also seeing social media sites using big data to influence other functions of consumeris daily lives. To further that observation, there is necessity behind seeing the claims that Facebook makes on its ability to influence its consumers to influence those within the social media sphere. If big data has the access to arrange itself around the sphere of the consumer to have the consumer act on certain task, what does that mean the uses of big data are to social media sites? We can assume from our knowledge that social media sites like Facebook could be interjecting into the private lives of their users by instigating on the data that is collected in processed to pinpoint user habits. It can also be seen that big data facilitates the necessary components of information to allow social media sites to specify their own approaches to their consumer in ways that can be seen as going over the line when it comes to their connection with their consumers. It is however, still a very enterprise avenue of using big data. As it allows for the social media to influence social, economical, and political landscapes. But that in itself is also a very dangerous power to have. As those who use the big data and direct their resources into specific marketing strategies can alter nearly whatever they fid like to in front of the irrational consumer.

As we have observed of big data above, we also learn of the prediction value and how big data escalates the ability for businesses to predict and recommend to consumers different products. The use of pooling data and sifting it through algorithms in order to precisely choose methods of spawning products before consumers

is a fundamental use of big data. Big data becomes a tool that assimilates the data that is created to businesses to create more big data. The process is unending and constantly provides businesses with unlimited amounts of data that can be used to spearhead their campaigns. Does big data then become a commodity that is used like currency to businesses? Well, it is very possible. As the use of big data is how businesses maneuver their strategies to get consumer to consume. If these algorithms meant to increase sales were used for something else, say medical awareness to the issues that exists within smoking cigarettes, how will the big data be used and what forms of algorithms would be used? The use of prediction analyst suits sales, but based on the observation made above, it is probably even more effective in helping to create a knowledgeable public. By facilitating the big data and being able to sort out the necessary public images that have control over social sphere through mass social medias, there can be an exchange of data between consumers. Because of the interplay between data and how consumers absorb them and create more data that is spawned for more information, big data can in turn control knowledgeable outcomes in public opinions. The use of big data is vast then when it comes to understanding the many components that make up what big data is.

Our observation above also places the consumer experience as an important facilitator for big data to exist. The habits and practices of consumers as well as the opinions and locations of the consumer can truly inhibit how big data is filtered to facilitate the necessary components of data to market and process information. This process can of course be seen using AI technology in order to expedite the data that is coming in. It does this so that the data is filtered and able to be used to immediately influence commercial markets, social media spheres, and consumer habits. That in turn regulates and begins to push out even more big data from the interactions that consumer have with the new platforms that are created from old big data that was used in order to create their new purchases or opinions. AI technology then becomes a fundamental component to the access, collection, analysis, and interpretation of big data. Its use of manipulating and translating data in order to be used to create enterprise is crucial to the development of more big data. From the observation above, it can also be said that AI technology will has also positioned itself in a way where it has become fundamental to the big data analysis and because of that, AI technology is part of big data.

The paper also sought to observe the nature of consumers having the capacity to control the big data that flows into collection. The use comes from filter settings like those included on Amazon in order to help narrow searches of items or on Netflix in order to create the right kind of streaming that the consumer requires or likes. If that is the case, the consumer actually holds a lot of power when it comes to the collection, analysis, and interpretation process of big data. Within how the consumer chooses to reside over these social medias and commercial businesses determines how the social media sites and businesses get their data. Beyond that, the consumer has no knowledge of how to analyze or interpret big data, yet holds the key to the very idea of big data. Because of this notion, the discussion here seeks to try and highlight the importance of

businesses maintaining and using data collected responsibly.

Big data is used in order to interact with consumers in order to sell or sway. The use of data however is also created by consumers. For this symbiotic relationship to exist and stay peaceful, businesses must be sure that they are not over stepping privacy issues when it comes to the use of big data. By enterprise methods to help consumers with choices and options through their recommendation systems and early predictive measures of trend by their collected big data, that is fine. But when business pressure consumers with the use of big data, the business will most likely end up losing new data to collect. As if no one is using their sources to create data, their lack of data causes them to have slow flow, and that leads to isolation of data.

Take for instance the process of data mining. Data mining is used in order to receive particular forms of information about consumers. This information is in the form of data, this data is put through special filters to narrow in on what it is that businesses want to know about particular groups of consumers to achieve the best methods of interacting with the consumers in order to highlight necessary products to the consumer. What happens if the algorithms for this particular data mine was off? This would mean that the data that was supposed to continue in the line of procession ends up lost. Because missing the mark with data mining and interpretation means that businesses loses their edge with their consumers.

Big data is a very complex topic to talk about. It is however, a very interesting topic to look at. As when we are observing what forms of big data are used in order to create experiences for consumers and business practices for businesses we can see the importance of having a very strong handle on the idea of big data. It is not just a process by which you collect massive amounts of information and then through it back out into a market. Big data must be molded around using algorithms, AI technology, studies done to mine particular forms of data, and even understanding the complex notions of unstructured data. Because of these reasons, the study of big data is still relatively incomplete. The use of big data however, should be understood as a relationship between consumers and those who seek to use big data to facilitate their individual means.

## 21 CONCLUSION

The paper sought out to examine the complexities of big data, but to be more precise, this paper seeks out to the multiplicity of avenues in which big data is used by our technological world in regards to Online shopping, Streaming Services, and Social Medias. The conclusion is that the multiple complex systems which make up the forefront and system of enterprise around big data falls under the very distinctive relationship that exists around the users of these services, and the sources the users use in order to create more big data.

Such complex multiplicity of diversified uses alter the understanding of big data by showcasing that big data in itself is easily

manipulated and altered. This becomes the case because of the multiple layers of data that exists in any given moment. The use of these data are incorporated in a way that there are many organization that are still trying to spearhead further in the endeavors of big data. The papers observation of the multiple forms of big data conversions, analytics, prediction standards, and even experimental uses of data reinforces the concept that big data is as it is called, massive. Because of such presence, to truly be able to look into the multiplicity of big data would mean a massive overhaul of research meant to showcase the existence. This paper however does not do that, but would also rather seek to showcase that.

Online shopping and streaming sites uses a multiplicity of tools alongside big data in order to function with consumers and creates a diversified experience for consumers. The tools that are used by these shopping and steaming businesses alter big data into sustainable forms of information that are then used in order to predict and recommend to consumers what products to purchase and recommendations. These are all done through algorithms used to analyze the data. The paper observes and concludes that the standing made by these businesses are diversified and are meant to showcase the suitable substance of the big data to consumers. The use their procedures not only diversify the consumer experience but also diversifies the way that big data is collected and used. Big data within these forms of principles are collected and retained under algorithmic databases that are then filtered out when it is being generated by consumers. The process of big data filtering is not only done by businesses, they are also given as options to consumers. Through the process of filtering consumers can do the exact same thing.

Social media sites use big data just at online shopping and streaming services do, but social media also has another power. One which allows them to facilitate their studies into experiences around the consumer. By doing this, social media can use big data in order to influence and manipulate consumers into specific acts of studies that are being done by the social media sites. This form of big data usage not only diversifies but also includes possibility for growth in collection of information. As when social media analysis these complex forms of data known as unstructured data, they are having a deeper perspective into consumer habits, wants, and additional personal information that expands their usage of big data.

Big data is a very diversified entity. Even though it can be narrowed down to certain institutions or entities, it in itself is able to expand largely in those narrowed views. The diversification of big data is very crucial to the survival and usage of big data. Without multiple sources to collect necessary data, there would be no big data. Therefore, the userfis experiences around big data needs to be one that is flexible and seeks to incorporate the right amounts of AI and Algorithms in order to maintain steady flow of data which encapsulates the very idea of diversified big data.

In summary, the multiplicity of avenues that exists around big data creates the very core study that it takes in order to understand the practices and exhibits which are induced by the use of big data itself. By observing real world applications of big data we can see

that the diversification of big data does not need to be mellowed or shallowed out from the perspective of a programmer, as it seems that the market capitalizes on big data and therefore creates the very enterprise of multiplicity within its diversification.

## ACKNOWLEDGMENTS

My group work very hard to facilitate a study around diversification and observation. Their hard work in reading through these sources multiple times to see the rights of the observation is very much appreciated. We would also like to thank our professor for the chance to take on such a free ranging topic, as it has allowed us to further appreciate big data and its importance in our future endeavors.

## REFERENCES

- [1] Bertino, Davidson S. Dayal U. Franklin M. Agrawal D., Bernstein P. and Colleagues. 2012. Challenges and Opportunities with Big Data: A White Paper Prepared for the Computing Community Consortium committee of the Computing Research Association. (2012). <http://cra.org/ccc/resources/ccc-led-whitepapers/>
- [2] J. Aloysius, H. Hoehle, S. Goodarzi, and V. Venkatesh. 2016. Big data initiatives in retail environments: Linking service process perceptions to shopping outcomes. (2016). [http://www.venkatesh.com/wp-content/uploads/dlm\\_uploads/2016/07/2016-AOR-Aloysius-et-al.pdf](http://www.venkatesh.com/wp-content/uploads/dlm_uploads/2016/07/2016-AOR-Aloysius-et-al.pdf)
- [3] Amazon. 2017. Big Data on AWS. (2017). <https://aws.amazon.com/big-data/>
- [4] John Aycock. 2010. Springer Sciences & Business Media. *Spyware and Adware* 50 (2010), 71–109.
- [5] Roy F Baumeister. 2002. Yielding to temptation: Self-control failure, impulsive purchasing, and consumer behavior. *Journal of consumer Research* 52, 4 (2002), 670–676.
- [6] Chui M. Brown, B. and J Manyika. 2011. Are you ready for the era of fibig datafi? *McKinsey Quarterly* 4, 1 (2011), 24–35.
- [7] A. A. Crdenas, P. K. Manadhata, and S. P. Rajan. 2013. Big Data Analytics for Security. *IEEE Security Privacy* 11, 6 (2013), 74–76.
- [8] Kyle Hensel & Michael H. Deis. 2010. Using Soical Media To Increase Advertising And Improve Marketing. *The Entrepreneurial Executive* 15 (2010), 87.
- [9] Gary Eastwood. 2017. Big Data, Algorithms and the Future of Advertising. (2017). <https://www.networkworld.com/article/3194585/big-data/big-data-algorithms-and-the-future-of-advertising.html>
- [10] W. Glynn Mangold & David J. Faulds. 2009. Social media: The new hybrid element of the promotion mix. *Business Horizons* 52 (2009), 357–365.
- [11] Ban G-Y. 2014. Business analytics in the age of big data. *Business Strategy Review* 25, 3 (2014), 8–9.
- [12] David Geer. 2017. Will Big Data Change how we use Social Media? (2017). [https://thenextweb.com/contributors/2017/07/06/will-big-data-change-use-social-media/#.tnw\\_DPcEKg97](https://thenextweb.com/contributors/2017/07/06/will-big-data-change-use-social-media/#.tnw_DPcEKg97)
- [13] M. Ghemsoune, M. Lebbah, and H. Azzag. 2016. State-of-the-art on clustering data streams. *Big Data Analytics* 1, 1 (2016), 134–145.
- [14] Amir Gandomi & Murtaza Haider. 2015. Beyond the Hype: Big Data Concepts, Methods, and Analytics. *International Journal of Information Management* 35, 2 (2015), 137–144.
- [15] Shaun Hipgrave. 2013. Smarter fraud investigations with big data analytics. *Network Security* 13, 12 (2013), 7–9.
- [16] A. Hussain and A. Roy. 2016. The emerging era of Big Data Analytics. *Big Data Analytics* 1, 1 (2016), 249.
- [17] Randal E. Bryant & Randy H. Katz & Edward D. Lazowska. 2008. Big-Data Computing: Creating revolutionary breakthroughs in commerce, science, and society. (2008). <https://cra.org/ccc/wp-content/uploads/sites/2/2015/05/Big>Data.pdf>
- [18] Dara Lind. 2014. Facebookfis fil Votedfi Sticker was a secret experiment on its users. (2014). <https://www.vox.com/2014/11/4/7154641/midterm-elections-2014-voted-facebook-friends-vote-polls>
- [19] Sarah Perez. 2016. Facebook gives its Election 2016 hub top billing by pinning it to your Favorites. (2016). <https://www.qubole.com/blog/big-data-advertising-case-study/>
- [20] Nate Philip. 2014. The Impact of Big Data on the Digital Advertising Industry. (2014). <https://www.qubole.com/blog/big-data-advertising-case-study/>
- [21] Randy H. Katz Randal E. Bryant and Edward D. Lazowska. 2008. Big-Data Computing: Creating Revolutionary Breakthroughs in Commerce, Science, and Society. (2008). <https://cra.org/ccc/wp-content/uploads/sites/2/2015/05/Big>Data.pdf>
- [22] Chetan Sharma. 2014. Big Data Analytics Using Neural Networks. (2014). <http://scholarworks.sjsu.edu/etd.projects/368>

- [23] Benjamin Reed Shiller. 2014. First-Degree Price Discrimination Using Big Data. (2014). [http://benjaminshiller.com/images/First\\_Degree\\_PD\\_Using\\_Big\\_Data.Jan.18.\\_2014.pdf](http://benjaminshiller.com/images/First_Degree_PD_Using_Big_Data.Jan.18._2014.pdf)
- [24] Eric Siegel. 2013. *Predictive analytics: the power to predict who will click, buy, lie, or die*. Vol. 51. Wiley, New York.
- [25] Hsinchun Chen & Roger H L Chiang & Veda C. Storey. 2012. Business intelligence and analytics: From big data to big impact. *MIS Quarterly: Management Information Systems* 36, 4 (2012), 1165–1188.

bibtext report

=====

This is BibTeX, Version 0.99d (TeX Live 2016)  
The top-level auxiliary file: report.aux  
The style file: ACM-Reference-Format.bst  
Database file #1: report.bib  
Name 1 in "Randal E. Bryant, Randy H. Katz, and Edward D. Lazowska" has a comma at the e  
while executing---line 3085 of file ACM-Reference-Format.bst  
Name 1 in "Randal E. Bryant, Randy H. Katz, and Edward D. Lazowska" has a comma at the e  
while executing---line 3085 of file ACM-Reference-Format.bst  
Name 1 in "Randal E. Bryant, Randy H. Katz, and Edward D. Lazowska" has a comma at the e  
while executing---line 3085 of file ACM-Reference-Format.bst  
Name 1 in "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., Franklin M., and C  
while executing---line 3085 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U.,  
while executing---line 3085 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U.,  
while executing---line 3085 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U.,  
while executing---line 3085 of file ACM-Reference-Format.bst  
Name 1 in "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., Franklin M., and C  
while executing---line 3085 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U.,  
while executing---line 3085 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U.,  
while executing---line 3085 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U.,  
while executing---line 3085 of file ACM-Reference-Format.bst  
Name 1 in "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., Franklin M., and C  
while executing---line 3085 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U.,  
while executing---line 3085 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U.,  
while executing---line 3085 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U.,  
while executing---line 3085 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Brown, B., Chui, M. and Manyika, J" for entry 16  
while executing---line 3085 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Brown, B., Chui, M. and Manyika, J" for entry 16  
while executing---line 3085 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Brown, B., Chui, M. and Manyika, J" for entry 16  
while executing---line 3085 of file ACM-Reference-Format.bst  
Name 1 in "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., Franklin M., and C  
while executing---line 3131 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., while executing---line 3131 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., while executing---line 3131 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., while executing---line 3131 of file ACM-Reference-Format.bst

Name 1 in "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., Franklin M., and C while executing---line 3131 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., while executing---line 3131 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., while executing---line 3131 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., while executing---line 3131 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Brown, B., Chui, M. and Manyika, J" for entry 16 while executing---line 3131 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Brown, B., Chui, M. and Manyika, J" for entry 16 while executing---line 3131 of file ACM-Reference-Format.bst

Name 1 in "Randal E. Bryant, Randy H. Katz, and Edward D. Lazowska" has a comma at the e while executing---line 3131 of file ACM-Reference-Format.bst

Name 1 in "Randal E. Bryant, Randy H. Katz, and Edward D. Lazowska" has a comma at the e while executing---line 3131 of file ACM-Reference-Format.bst

Name 1 in "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., Franklin M., and C while executing---line 3229 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., while executing---line 3229 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., while executing---line 3229 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., while executing---line 3229 of file ACM-Reference-Format.bst

Name 1 in "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., Franklin M., and C while executing---line 3229 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., while executing---line 3229 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., while executing---line 3229 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., while executing---line 3229 of file ACM-Reference-Format.bst

Name 1 in "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., Franklin M., and C while executing---line 3229 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., while executing---line 3229 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., while executing---line 3229 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., while executing---line 3229 of file ACM-Reference-Format.bst

```
Too many commas in name 1 of "Brown, B., Chui, M. and Manyika, J" for entry 16
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Brown, B., Chui, M. and Manyika, J" for entry 16
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Brown, B., Chui, M. and Manyika, J" for entry 16
while executing---line 3229 of file ACM-Reference-Format.bst
Name 1 in "Randal E. Bryant, Randy H. Katz, and Edward D. Lazowska" has a comma at the e
while executing---line 3229 of file ACM-Reference-Format.bst
Name 1 in "Randal E. Bryant, Randy H. Katz, and Edward D. Lazowska" has a comma at the e
while executing---line 3229 of file ACM-Reference-Format.bst
Name 1 in "Randal E. Bryant, Randy H. Katz, and Edward D. Lazowska" has a comma at the e
while executing---line 3229 of file ACM-Reference-Format.bst
(There were 48 error messages)
make[2]: *** [bibtex] Error 2
```

latex report

[2017-12-11 13:24:58] pdflatex report.tex



```
Missing character: ""

Typesetting of "report.tex" completed in 1.2s.
./README.yml
 56:72   error    trailing spaces  (trailing-spaces)
 57:75   error    trailing spaces  (trailing-spaces)
 58:75   error    trailing spaces  (trailing-spaces)
 59:77   error    trailing spaces  (trailing-spaces)
 60:77   error    trailing spaces  (trailing-spaces)
 61:75   error    trailing spaces  (trailing-spaces)
 62:77   error    trailing spaces  (trailing-spaces)
 63:78   error    trailing spaces  (trailing-spaces)
 64:77   error    trailing spaces  (trailing-spaces)
 65:77   error    trailing spaces  (trailing-spaces)
 66:78   error    trailing spaces  (trailing-spaces)
 67:76   error    trailing spaces  (trailing-spaces)
 68:51   error    trailing spaces  (trailing-spaces)
```

---

## Compliance Report

---

```
name: Shiqi Shen
hid: 109
paper1: complete 100% Oct 27th
paper2: complete 100% Nov 4th
project: 100% Dec 4th
```

```
yamlcheck
```

---

```
wordcount
```

---

```
13
wc 109 project 13 11810 report.tex
wc 109 project 13 11844 report.pdf
wc 109 project 13 891 report.bib
```

find "

---

- 159: As per a current SHSFoodThink white paper "Are We Chain Obsessed?" 64{\%} of customers said that the previous shopping experience is what makes them keep coming backnot the items themselves \cite{12}. By utilizing bits of knowledge received from the information transaction database, online networking, promotional activity, customers purchasing behavior, and client movement patterns, grocery stores can find a way to guarantee they are engaged with their customers that matter most.
- 207: Amazon spearheaded item proposal engine: the "if you purchased that, you may like this" invention. This strategic changing web-based shopping feature mirrors the retailer's profound assessment of buyers' shopping basket. Proposal engine is intended to enable customers to find items they were not sorting out but rather would be interested in purchasing \cite{10}. Today, general grocery merchants are progressively tapping the global innovation behind proposal engine: predictive analytics. This kind of assessment measures future patterns in light of present and past information, and it can enable stores to improve business. Information is driven, all-encompassing assessment of "purchasing triggers, for example, regularity, weather, stock, and advancements, is progressively informing grocery stores' product blend, marketing plans, and sales forecast \cite{14}. Furnished with these information-driven tools, stores can better distinguish what items customers need today and what they will be demanding in future, and this learning will enable them to stay competitive for a considerable length of time to come.
- 259: Predictive modeling and different analytics were accessible to large insurance net providers willing to introduce the innovation on location as of not long ago. Software as a service has advanced to even where genuinely little insurers can exploit Big Data analytics \cite{16}. Insurance providers subscribe to a service keeps running by a seller as opposed to paying for the vast buy, establishment, and support of in-house frameworks. SaaS likewise is named "on-demand software."

passed: False

find footnote

---

passed: True

```
find input{format/i523}
```

---

```
4: \input{format/i523}
```

```
passed: True
```

```
find input{format/final}
```

---

```
passed: False
```

```
floats
```

---

```
figures 0  
tables 0  
includegraphics 0  
labels 0  
refs 0  
floats 0
```

```
True : ref check passed: (refs >= figures + tables)
```

```
True : label check passed: (refs >= figures + tables)
```

```
True : include graphics passed: (figures >= includegraphics)
```

```
True : check if all figures are refered to: (refs >= labels)
```

```
Label/ref check
```

```
passed: True
```

```
When using figures use columnwidth
```

```
[width=1.0\columnwidth]
```

```
do not change the number to a smaller fraction
```

```
find textwidth
```

---

```
passed: True
```

```
below_check
```

---

WARNING: algorithm and above may be used improperly

69: As a programmer, machine learning and the creation of these algorithms truly fascinate when they are being used in order to create a recommendation system. Since the collection of data is pooled, the use of algorithms to reach a particular end, based on data analyzation, makes for a very interesting usage. As the use of data algorithms not only seek to facilitate the necessary recommendations, but also filter through millions of data informatics to agitate the necessary products to ensure the transparency of recommendations created. This can also be seen in what was stated above regarding consumers being able to filter products with precise search options. These filters can act as filters for filters. As in the filters chosen by consumers help to facilitate precise sifters and allows for algorithms to pinpoint precise outlines of data that help to recommend even finer tuned recommendations for consumers. \\

WARNING: algorithm and above may be used improperly

265: As we have observed of big data above, we also learn of the prediction value and how big data escalates the ability for businesses to predict and recommend to consumers different products. The use of pooling data and sifting it through algorithms in order to precisely choose methods of spawning products before consumers is a fundamental use of big data. Big data becomes a tool that assimilates the data that is created to businesses to create more big data. The process is unending and constantly provides businesses with unlimited amounts of data that can be used to spearhead their campaigns. Does big data then become a commodity that is used like currency to businesses? Well, it is very possible. As the use of big data is how businesses maneuver their strategies to get consumer to consume. If these algorithms meant to increase sales were used for something else, say medical awareness to the issues that exists within smoking cigarettes, how will the big data be used and what forms of algorithms would be used? The use of prediction analyst suits sales, but based on the observation made above, it is probably even more effective in helping to create a knowledgeable public. By facilitating the big data and being able to sort out the necessary public images that have control over social sphere through mass social medias, there can be an exchange of data between consumers. Because of the interplay between data and how consumers absorb them and create more data that is spawned for more information, big data can in turn control knowledgeable outcomes in public opinions. The use of big data is vast then

when it comes to understanding the many components that make up what big data is.\\

bibtex

---

label errors

bibtex errors

This is BibTeX, Version 0.99d (TeX Live 2016)

The top-level auxiliary file: report.aux

The style file: ACM-Reference-Format.bst

Database file #1: report.bib

Name 1 in "Randal E. Bryant, Randy H. Katz, and Edward D. Lazowska" has a comma at the end while executing---line 3085 of file ACM-Reference-Format.bst

Name 1 in "Randal E. Bryant, Randy H. Katz, and Edward D. Lazowska" has a comma at the end while executing---line 3085 of file ACM-Reference-Format.bst

Name 1 in "Randal E. Bryant, Randy H. Katz, and Edward D. Lazowska" has a comma at the end while executing---line 3085 of file ACM-Reference-Format.bst

Name 1 in "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., Franklin M., and O'Neil C." has a comma at the end while executing---line 3085 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., Franklin M., and O'Neil C." while executing---line 3085 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., Franklin M., and O'Neil C." while executing---line 3085 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., Franklin M., and O'Neil C." while executing---line 3085 of file ACM-Reference-Format.bst

Name 1 in "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., Franklin M., and O'Neil C." while executing---line 3085 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., Franklin M., and O'Neil C." while executing---line 3085 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., Franklin M., and O'Neil C." while executing---line 3085 of file ACM-Reference-Format.bst

Name 1 in "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., Franklin M., and O'Neil C." while executing---line 3085 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., Franklin M., and O'Neil C." while executing---line 3085 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., Franklin M., and O'Neil C." while executing---line 3085 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., Franklin M., and O'Neil C." while executing---line 3085 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., Franklin M., and O'Neil C." while executing---line 3085 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., Franklin M., and O'Neil C." while executing---line 3085 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Brown, B., Chui, M. and Manyika, J" for entry 16  
while executing---line 3085 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Brown, B., Chui, M. and Manyika, J" for entry 16  
while executing---line 3085 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Brown, B., Chui, M. and Manyika, J" for entry 16  
while executing---line 3085 of file ACM-Reference-Format.bst

Name 1 in "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., Franklin M., and C  
while executing---line 3131 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U.,  
while executing---line 3131 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U.,  
while executing---line 3131 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U.,  
while executing---line 3131 of file ACM-Reference-Format.bst

Name 1 in "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., Franklin M., and C  
while executing---line 3131 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U.,  
while executing---line 3131 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U.,  
while executing---line 3131 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U.,  
while executing---line 3131 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Brown, B., Chui, M. and Manyika, J" for entry 16  
while executing---line 3131 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Brown, B., Chui, M. and Manyika, J" for entry 16  
while executing---line 3131 of file ACM-Reference-Format.bst

Name 1 in "Randal E. Bryant, Randy H. Katz, and Edward D. Lazowska" has a comma at the e  
while executing---line 3131 of file ACM-Reference-Format.bst

Name 1 in "Randal E. Bryant, Randy H. Katz, and Edward D. Lazowska" has a comma at the e  
while executing---line 3131 of file ACM-Reference-Format.bst

Name 1 in "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., Franklin M., and C  
while executing---line 3229 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U.,  
while executing---line 3229 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U.,  
while executing---line 3229 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U.,  
while executing---line 3229 of file ACM-Reference-Format.bst

Name 1 in "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., Franklin M., and C  
while executing---line 3229 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U.,  
while executing---line 3229 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U.,  
while executing---line 3229 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U.,  
while executing---line 3229 of file ACM-Reference-Format.bst

Name 1 in "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., Franklin M., and C while executing---line 3229 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., while executing---line 3229 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., while executing---line 3229 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., while executing---line 3229 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Brown, B., Chui, M. and Manyika, J" for entry 16 while executing---line 3229 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Brown, B., Chui, M. and Manyika, J" for entry 16 while executing---line 3229 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Brown, B., Chui, M. and Manyika, J" for entry 16 while executing---line 3229 of file ACM-Reference-Format.bst  
Name 1 in "Randal E. Bryant, Randy H. Katz, and Edward D. Lazowska" has a comma at the e while executing---line 3229 of file ACM-Reference-Format.bst  
Name 1 in "Randal E. Bryant, Randy H. Katz, and Edward D. Lazowska" has a comma at the e while executing---line 3229 of file ACM-Reference-Format.bst  
Name 1 in "Randal E. Bryant, Randy H. Katz, and Edward D. Lazowska" has a comma at the e while executing---line 3229 of file ACM-Reference-Format.bst  
(There were 48 error messages)

bibtex\_empty\_fields

---

entries in general should not be empty in bibtex

find ""

---

passed: True

---

ascii

---

non ascii found 8217  
non ascii found 8217  
non ascii found 8217  
non ascii found 8220  
non ascii found 8221  
non ascii found 8217  
non ascii found 8220  
non ascii found 8221  
non ascii found 8217

```
non ascii found 8217
```

```
=====
The following tests are optional
=====
```

```
Tip: newlines can often be replaced just by an empty line
```

```
find newline
-----
```

```
passed: True
cites should have a space before \cite{} but not before the {
```

```
find cite {
-----
```

```
passed: True
```

# Big Data Analytics on Food Products Around the World

Karthik Vegi

Indiana University Bloomington  
College Mall Apartments  
Bloomington, Indiana 47401  
kvegi@iu.com

Nisha Chandwani

Indiana University Bloomington  
Park Doral Apartments  
Bloomington, Indiana 47408  
nchandwa@iu.edu

## ABSTRACT

Food is one of the basic necessities of human-being. It helps us gain energy to recharge our body to do the daily activities of moving, playing, and thinking. From being a cave man to producing a wide variety of foods, we have come a long way. The civilizations shaped the food habits of the world and there is a lot of variance in the food habits across countries. We analyze the *Open Food Facts* database that gathers information on food products from around the world to unearth some food habits of the world and we predict the food grade based on the nutrition facts of the food products.

## KEYWORDS

i523, hid231, hid203, big data, food habits, food products, nutrition

## 1 INTRODUCTION

*Open Food Facts* is a non-profit initiative started by Stephane Gigaandet and run by thousands of volunteers around the world. Any person around the world can contribute to the database by simply scanning a product using a mobile app which is made available to IOS and Android. This massive database of food products opens up a lot of opportunities to analyze the food products around the world and understand the food habits. We are particularly interested in the consumption of nutrients that come along with the food items across the world, the composition of different fat content, and the prediction of nutrition grade based on the nutrients.

## 2 FOOD ANALYSIS: IMPORTANCE AND RELATED WORK

In recent times, more and more companies try to market their food as low-fat or low-calories in order to fool consumers into buying their products. The increasing concern of public health has led to a significant interest in detecting the health-related properties of food products [2]. Thus, there is no question about the importance of analysis of the nutrition grade and food safety in today's world. The analysis of food requires more robust and efficient methodologies in order to ensure the quality and safety of the food products [2]. Previous methods based on the so-called wet-chemistry have now evolved into more powerful techniques which are used in the food laboratories. These methods provide a massive improvement in analytical accuracy thus expanding the limits of food applications [2]. The traditional methods of food analysis can be classified based on the underlying principle. Some of these categories are spectroscopic, biological, electrochemical, supercritical fluid chromatography [2]. All these techniques provide information about the sample under study and this information is derived from a specific physical-chemical interaction [2]. A different approach to analyzing and detecting the food quality is by using machine

learning techniques. We will discuss one of these modern methods of food analysis which can be widely used across countries.

## 3 ANALYSIS OF NUTRIENTS IN FOOD

Fat is definitely a nutrient that the body needs and is an essential nutrient that aids in cell growth, helps with energy generation, maintaining body temperature, protect organs, help absorb other essential nutrients that aid in producing energy, improve blood cholesterol level, help reduce inflammation in case of injury, and help in storing energy that can be used for survival when you go without food for few days [1]. But we do need to keep a track of the consumption because anything that is remotely excess leads to a variety of serious health issues [1].

### 3.1 Dietary Fats

There are different types of fat if? some are good and some are bad and some needs to be taken within a certain limit [1].

*3.1.1 Saturated Fat.* More intake of saturated fats results in the cholesterol levels in the blood which increases the risk of heart-related diseases [1]. The American Heart Association suggests around 5 percent of daily calories from foods containing saturated fat [1]. Meat, cheese, and milk are some of the sources of saturated fat [1].

*3.1.2 Trans Fat.* Any type of trans fat whether it is natural or artificial is not good [1]. The reason why food manufacturers use trans-fat is that they are less expensive, can be produced artificially, easy to use with other ingredients, last for a long time and also aid in improving the taste of the food [1]. Trans fats raise the bad fat levels and decrease the good fat levels [1]. The American Heart suggests to completely cut off trans-fat from the diet [1].

*3.1.3 Monounsaturated Fat.* Monounsaturated fats have a good effect on the body when taken within limit [1]. They help reduce the bad cholesterol levels in the blood and thereby decrease the risk of heart diseases [1]. They also help in gaining vitamin E which is a good nutrient that acts as antioxidant [1]. Olive oil, avocados, and sesame oil are some of the sources of monounsaturated fats [1].

*3.1.4 Polyunsaturated Fat.* Polyunsaturated fats have a good effect on the body when taken within limit [1]. They help reduce the bad cholesterol levels in the blood and thereby decrease the risk of heart diseases [1]. They also provide some nutrients that are essential for the body [1]. Soybean oil and sunflower oil are some of the sources of polyunsaturated fats [1].

### 3.2 Data Cleaning and Transformation

To make the analysis more interesting, the top 20 countries with most value counts for the attributes have been considered. The countries with names combined with other countries were also cleaned in the process. The data was analyzed for missing values and the attributes with more than 60 percent missing values were removed from the analysis to add consistency. Only the columns that are meaningful in the analysis were retained and the rest were removed from further analysis.

We then display the top 5 countries as a pie-chart and the 5 countries are namely United States, France, Switzerland, Germany, and Spain as shown in Figure 1.

[Figure 1 about here.]

We then impute all the null values with zeroes and we then check the dietary fat content in the foods and check the top countries with fat content using a histogram. The analysis with respect to the fat countries is as follows

### 3.3 Fat Content

The top 5 countries with most fat content in the food items are Serbia, United States, Switzerland, Germany, and Sweden as shown in Figure 2.

[Figure 2 about here.]

The top 5 countries with most saturated fat content in the food items are Serbia, United States, Germany, France and Switzerland as shown in Figure 3

[Figure 3 about here.]

The top 5 countries with most trans-fat content in the food items are United States, Brazil, Canada, Australia, Russia, and Serbia as shown in Figure 4.

[Figure 4 about here.]

The top 5 countries with most cholesterol content in the food items are United States, Canada, Portugal, Brazil, France, and Italy as shown in Figure 5.

[Figure 5 about here.]

### 3.4 Sugar and Salt Content

Although the body needs sugar, high intake of artificial and processed sugar is bad for health as it does not add any nutrients but only adds calories [5]. It is always better to rely on the natural sugar that comes with fruits and milk [5]. Artificial sugars tooth decay and diabetes [5]. Just as fat, sodium which is the main source of iodine is essential for health although its intake should be within limit [5]. Increase in intake of salt leads to blood pressure and has an effect on the heart [5].

The top 5 countries with most sugar content in the food items are United States, Serbia, Switzerland, France, and Sweden as shown in Figure 6.

[Figure 6 about here.]

The top 5 countries with most sodium content in the food items are United States, Hungary, Serbia, Sweden, and France as shown in Figure 7.

[Figure 7 about here.]

## 4 NUTRITION GRADE LABELLING SYSTEM

France recently took a decision to implement a nutri-score system which will use a color coding mechanism to label the food products that will help consumers know the nutrition grade of the product [3]. The World Health Organization regional office for Europe as a part of its 5-year action plan from 2015-2020 recommends a labeling mechanism for the consumers to know about the quality of the food products at a first glance [3]. This will not only make it easier for the consumers to pick healthier options but it will also regulate food manufacturers to resort to healthier ingredients instead of going for low cost artificial or less healthy ingredients [3].

France after the United Kingdom became the second country to implement this system to indicate the main ingredients like fat, salt and sugar content in the food items [3]. France made use of an evidence-based system to study different labeling systems to arrive at the best one [3]. By implementing this system, the World Health Organization will keep a check on the growing number of diet-related diseases in the Europe region [3]. Europe being the largest consumer of cheese wants to regulate the ingredients that go into the manufacturing process so that people are well informed about their food choices [3].

### 4.1 Nutrition Grade Prediction as a Big Data Problem

We build a predictive classification model to predict the food nutrition grade based on the ingredients of the food. The goal is to apply various machine learning algorithms to the problem at hand, measure the prediction accuracy to compare and contrast the different algorithms and arrive at the best algorithm that suits the given data and the problem. This problem can be solved using Big Data and Machine Learning techniques given the size and the complexity of the data.

## 5 MACHINE LEARNING

Machine Learning is a field in which we train computers in a way that they can learn from the input data [6]. The ideology is that computers use the training data that is made available to them, learn from it, build a model and use this experience to build knowledge that can be applied on new unseen data [6]. A wonderful example to demonstrate machine learning is the application to detect spam emails where the machine builds knowledge from previously seen emails which are marked as spam, checks new emails to see if they match the historic spam emails and label them as spam or non-spam [6].

## 5.1 Types of Machine Learning Algorithms

There are primarily two types of machine learning algorithms, descriptive models and predictive models [6]. A *Descriptive Model* is described as the analysis done and insights gained from slicing and dicing the data in new and interesting ways [6]. One example of a descriptive model is pattern discovery that is often used in market basket analysis where transnational purchase details are analyzed [6]. A *Predictive Model* on the other hand involves predicting one value using one or more variables [6]. The learning algorithms tried to build a model that captures the relationship between a response variable and the independent variables [8].

## 5.2 Types of Learning

*Unsupervised Learning* is the process where there is no explicit training data to learn from, so there is simply no mechanism where the machine can learn from previously available data [6]. The same email example can be looked at in a different way where we now want to do anomaly detection in emails [6]. Here the main goal is to detect unusual messages from the bunch of messages and we do not have experience of previous data [6].

*Supervised Learning* in contrast is the process of gaining knowledge or expertise from the training data which can be applied to future unseen data [6]. Here the model is first trained by using a bulk of training examples and this model is applied to testing data to measure the accuracy [6]. The variable that we need to predict is identified which is called the response variable and the variables that are used to predict the response variables, called the predictor variables are identified [6]. If the existing variables are not sufficiently giving the accuracy that is expected, a method called feature engineering is done where new variables are derived by combining existing variables [6].

## 6 PREDICTION ANALYSIS

Prediction analysis is the process of working on a large dataset using a combination of statistical, data mining and machine learning algorithms to predict the outcome based on past data [6]. There are primarily two types of prediction analysis in machine learning, namely regression and classification [8]. In regression, we try to predict a continuous variable from the predictor variables [8]. A good example of regression is to predict the housing prices from different parameters like the year of construction, location, amenities, number of bedrooms etc [8]. Here the response variable is continuous and it is not predefined [8]. Classification, on the other hand, tries to predict a categorical variable in which we assign each record with a predefined label or a class [8].

Classification is the task of assigning each data record to a predefined class [8]. In machine learning, classification is categorized as a supervised learning technique [8]. This problem has applications in various fields like spam detection, medical applications, astronomy, and banking to identify fraudulent transactions from genuine transactions [8]. It is the task of coming up with a model which is essentially a function that maps every data record to a class label [8].

The task at hand is a classification problem since we are trying to predict the food nutrition grade of the products based on the ingredients that go into the product. For this problem, we are considering only the data for the country France, since the nutrition grade is available for most food products from the country. Another reason is that France is the first country in the region to come up with the idea of adding a color-coded label to the food products mentioning the nutrition grade. In the subsequent sections, we discuss the machine learning techniques used to solve this problem.

## 6.1 K Nearest Neighbors

**6.1.1 Overview.** Some of the classification algorithms in machine learning work on the principle of eager learning that involves a two-step process where first a model is built from the training data and the model is applied on testing data [8]. In contrast, K nearest neighbors is a lazy learning algorithm where the process of modeling the training data is not done until the test examples are classified [8]. *Rote Classifier* is a good example of lazy learning algorithm which memorizes the entire training data to perform classification but has the drawback of not being able to map every test example against the training example [8]. K nearest neighbors algorithm overcomes this drawback by finding all the records that are closest or nearest to the training records [8].

The nearest neighbor puts each attribute list as a data point in the n-dimensional space, given n the number of attributes [8]. Once we have the training examples, we take each test example and compute its distance to the training example classes and assign a class label [8]. Any of the popular distance measures among Euclidean distance, Manhattan distance, Minkowski distance and Mahalanobis distance can be used [8]. The k denotes the k closest points to the test example [8]. Figure 8 shows the algorithm [8].

[Figure 8 about here.]

**6.1.2 Support in Python.** KNeighborsClassifier is available in the scikit learn python library.

## 6.2 Logistic Regression

Logistic regression or logit regression is a special type of regression analysis where the response variable that we need to predict is a categorical variable [8]. Typically, logistic regression models the response variable to take two values, 1 or 0, pass or fail, win or lose [8]. Logistic regression that takes more than two values for the response variable is called multinomial logistic regression [8]. Here the probability of the response variable to take a categorical value is modeled as a function of the predictor variables [8].

Like a lot of machine learning algorithms, logistic regression works by making a lot of assumptions which should be taken care as a part of the data cleaning and transformation process [6]. It does not assume a linear relationship between the response variables and the predictor variables [6]. Since it applies a log transformation on the predicted probabilities, it can handle a variety of relationship between the predictor variables [6]. If the predictor variables are multivariate normal, the algorithm achieves the best result although it works even if they are not [6]. The stepwise method must be used in the logistic regression to ensure that we are neither overfitting

nor underfitting the data [6]. A very important assumption to be noted in logistic regression is that each attribute list must be independent, in the sense, the data records must not be derived from a before-after setup experiment [6]. It also requires a decently large sample size to work on [6].

**6.2.1 Support for Python.** LogisticRegression is available in the scikit learn python library.

### 6.3 Random Forest Classifier

Random forest is an ensemble classification algorithm which is very powerful [8]. Ensemble method is a special process to improve the accuracy of the prediction [8]. The classification algorithms we have seen so far predict the response variable using a single classifier on the test data but ensemble methods use multiple classifiers in tandem and aggregate the predictions to boost the accuracy by a huge margin [8]. Using a combination method, the ensemble method derives a set of base classifiers from the training data and on each iteration takes a vote of all the base classifiers to arrive at a result [8].

Random forest is an ensemble method which works very well for classification problems [8]. It combines the predictions made by multiple classifiers where each classifier independently works on the training data and casts its vote [8]. Unlike methods like AdaBoost which generates values based on independent random vectors using a varied probability distribution, random forest generates values based on fixed probability distribution [8].

**6.3.1 Rationale for Random Forest.** Consider an example, where we have 25 base classifiers and each base classifier has an error rate of 0.35 [8]. As discussed, the random forest takes the majority vote given by the base classifiers [8]. The model makes a wrong prediction if half or more base classifiers predict inaccurately. The accuracy is improved with an error rate of 0.06 which is far better than using just a single classifier [8].

**6.3.2 Support for Python.** RandomForestClassifier is available in the scikit learn python library.

## 7 EXPERIMENTS AND RESULTS

In this section, we will introduce the algorithm along with the details of experiments and methodology for predicting the nutrition grade of food products in France.

### 7.1 Algorithm

The problem at hand is to correctly identify the nutrition grade of the food item. The possible labels are, *a* to *e*, with *a* being the best and *e* being the worst grade for a food item. For this task, we have used machine learning techniques that help in predicting the label of each food item. Before getting into the details of each step of the method, we first present a concise version of the algorithm used for this task:

- (1) Select all the records for the country, France. Drop records where nutrition grade is not populated.
- (2) Separate the predictors from the response variable in order to perform data cleaning and data transformation steps.

- (3) Check for missing values in the predictors obtained in the step above. Drop columns with more than 60% missing values.
- (4) Impute the missing values with 0 for remaining columns.
- (5) After imputing the missing values, standardize all the numerical predictors using the standard scaler.
- (6) Check for the correlation between different numerical predictors. Drop one predictor from each pair of predictors that show high correlation.
- (7) Combine the pre-processed predictors and the response variable in a single data frame.
- (8) Divide the data obtained in step above into training and test data using stratified sampling.
- (9) Train different classifiers on the training data and check the performance of each classifier on the test data.

### 7.2 Data set

For the classification problem, we selected the records for country France.

Number of examples: 123,961

Number of variables: 12

Response variables: *Nutrition Grade*

Predictor variables: *Energy per 100g*, *Fat per 100g*, *Saturated Fat per 100g*, *Carbohydrates per 100g*, *Sugars per 100g*, *Fiber per 100g*, *Proteins per 100g*, *Salt per 100g*, *Trans-fat per 100g*, *Sodium per 100g*

### 7.3 Python Packages Used

The following Python packages were used to solve the classification problem:

- Pandas: Provides high-performance data structures for data analysis and data munging
- Matplotlib: Plotting library that helps to embed plots into applications using GUI
- Seaborn: Visualization package based on matplotlib used for drawing high-level statistical graphics
- Scikit-learn: Toolbox with solid implementation of machine learning and other algorithms
- Scipy: Package that supports scientific computing with modules for linear algebra and integration

### 7.4 Data Cleaning

**7.4.1 Step 1: Data Sparsity.** Data sparsity refers to the situation where a lot of attributes have missing values which is an advantage in some cases because you only need to store and analyze the data that is available to you and save on computation time and storage [8]. We first check the data value counts for each country. United States, France, Switzerland, Germany, and Spain come as the top 5 countries with most data. Since the food nutrition grade was implemented in France, it has most products for which nutrition grade is labeled. So for this classification problem, we use the food data from France for analysis.

**7.4.2 Step 2: Handling Missing Values.** Missing values is a common scenario and they can be handled in different ways. You could

choose to eliminate the data objects with missing values but at the expense of missing some critical analysis [8]. Estimating the missing values is also a good way to handle them, especially when the data comes from time series etc, where you could possibly interpolate the missing values from the ones that are closer to it [8]. Ignoring the missing values is another technique which can be applied to tasks like clustering where the similarity can be calculated using the attributes other than the missing ones [8].

The data set was first analyzed to check the missing values in all the columns. The threshold limit has been set at 60 percent. All the columns with missing values more than 60 percent were removed from the analysis to make the result more consistent. Once the columns were removed, the data set has to be re-indexed to maintain the order. Only the columns that are important for the prediction task have been retained from the original dataset. In this case, all the ingredients which are primarily the predictor variables were included. The missing values in the response variable also need to be taken care of. Removing the records with missing values for the response variable proved to be the best option for trying out various things.

Imputation was used to handle the null values in the predictor variables. Imputation can be done in a variety of ways, for example, replacing the missing values with zero or imputing the missing values for numerical columns with the mean and the categorical columns with the mode. Since all the predictor variables have numeric values, all the null values have been replaced with zero. To ensure that the imputation process has been done correctly, the sum of missing values is calculated since post-imputation, this sum should be zero.

**7.4.3 Step 3: Outlier Treatment.** Outliers are data objects with quite distinct characteristics from the other data records [8]. There is a considerable difference between anomalies and outliers, where anomalies refer to data records that have bad data, which is noise and need to be ignored, anomalies often contain interesting aspects and can lead to some good analysis [8]. In applications like *Fraud Detection*, anomalies could be of utmost importance [8]. The outliers in the data have been looked at by using box plots and have been handled as a part of the data cleaning process.

## 7.5 Exploratory Data Analysis

For exploratory data analysis, we used the Seaborn package along with Matplotlib for visualizations. The measure of spread, that is the range and variance of the values, is a good way to understand the different aspects of the predictor variables. Box-plots are a method of visualization to look at the distribution of values for a numerical attribute [8]. The box plots show the percentiles where the lower and upper ends of the box indicate 25<sup>th</sup> and 75<sup>th</sup> percentile, the line inside the box indicates the 50<sup>th</sup> percentile, the tails indicate the 10<sup>th</sup> and 90<sup>th</sup> percentile respectively [8].

**7.5.1 Bi-variate box-plots.** Bi-variate box-plots go beyond univariate box plots by showing the relationship between the predictor variable and the response variable [8]. We look at the bi-variate

box-plots for each of the important predictor variables namely, saturated fat, polyunsaturated fat, sugars and salt and the response variable, nutrition grade. Figure 9 shows the bi-variate box plots.

[Figure 9 about here.]

By looking at the box plots, we can understand some important aspects of how the response variable is related to the predictor variables. We see that as the average saturated fat content increases, the food grade decreases and as the average polyunsaturated fat content increases the nutrition grade is better. When the sugar levels increase, the health quotient of the food comes down. The energy levels behave in an interesting manner where the energy for the nutrition grade A is higher whereas in general, the average energy level slightly increases with the decreasing nutrition grade. While increase in energy does not necessarily imply that the nutrition quality is high, as there are a lot of instant energy foods that have a lot of additives, but they are often rated low when it comes to health.

**7.5.2 Correlation.** Correlation between data objects is the measure of the linear relationship between the attributes of the object that are continuous variables [8]. Correlation analysis is the process of finding of the correlations between the different predictor variables and identify high collinearity problem [6]. The relationship could be either linear or non-linear based on the given data [8]. The correlation coefficient can range anywhere between -1 and 1, where 1 indicates a very high positive correlation and -1 indicates a very high negative correlation [6]. Correlation plot visually shows the correlation coefficient between the variables in a nicely laid out plot. Figure 10 shows the correlation plot.

[Figure 10 about here.]

By looking at the correlation plot, we can see that sugars, fat, energy are positively correlated with the nutrition grade. This indicates that these variables will play an important role in the prediction algorithm. However, sodium and salt are highly correlated with each other and this may lead to collinearity problem if not handled. Collinearity is the state where the independent variables are highly correlated with each other which can add a lot of noise to the data [7]. Some of the problems because of collinearity are that the regression coefficients may not be estimated correctly. Also, collinearity makes it very difficult to explain the response variables using the predictor variables [7]. So we remove sodium from the predictor variables and proceed to the next step.

**7.5.3 Data Transformation.** Data transformation refers to the transformation that is applied to the variables [8]. For each data object, we apply a transformation function to all the attributes of the object to ensure that the attributes do not have a lot of variance in the data [8]. This process is also called standardization since we are applying a standard function to make sure all the attributes fall within a given range [8]. There are different methods that can be applied to achieve scaling namely log transformation, absolute value, square root transformation [8].

We use the method called normalization where all the values fall in the range, 0 to 1. To achieve this, we use the prepossessing package from sklearn which provides utility functions and transformer classes to change raw data into a standard representation. A lot of machine learning algorithms work well on standardized data. If some of the variables have extreme values, they might dominate the model function and might disturb the estimation parameter. Thus, for such extreme values, standardization helps achieve better results.

On scaling the data, there was a massive improvement in the prediction accuracy of the algorithms, implemented for this task. Thus, this proves the importance of data standardization with respect to machine learning algorithms.

## 7.6 Data Sampling

In a supervised machine learning approach, the model is trained on one sample of the data and later tested on a different sample of the data. Thus, in order to test the performance of the nutrition grade classifier, the data for the country France was divided into two samples, training and testing. There are various ways to achieve this split or sampling of the data. Some of these sampling methods are:

- Simple Random Sampling: This is one of the simplest sampling techniques. In this technique, every data point has an equal chance of being selected. In other words, it works similar to a lottery system where every outcome has an equal probability. The biggest advantage of this technique is the ease of implementation and its unbiased nature while generating the sample. However, random sampling might not always result in a sample that can represent the true population. It generally works well when we have huge data to sample from.
- Stratified Sampling: This technique is a more sophisticated method of sampling data. Stratified sampling generates a sample such that the proportion of each class in the sample is same as that in the true population. In this technique, the entire population is divided into groups or strata. The next step is to randomly select data points from each stratum such that the final sample has the same proportion for each stratum as that present in the true population. Thus, the sample generated by this technique is a good representative of the true population. Stratified sampling is a very useful technique when the classes in the data are highly imbalanced.

For our classifier, we chose to divide the data for France into training and test samples using stratified sampling technique. The strata or groups were created based on the response variable, i.e., food grade. This ensured that the training and test data had the same proportion of each food grade.

## 7.7 Data Modeling

Once the data was divided into training and test data, the next step was to train different classifiers and tune their respective parameters for better accuracy. We implemented three different models for

classifying the food grade. Each of these models along with their parameters is:

- K Nearest Neighbors (kNN): For kNN, the grade of a food item in test data is classified by first finding the  $k$  most similar food items in the training data. It then takes the vote (food grade label) from each of these neighbors and based on the majority vote, the food item from the test data is assigned a food grade. Thus, one of the most important parameter for kNN is  $k$ , i.e., the number of neighbors to consider from the training data. We tried different  $k$  values and found that  $k = 3$  gives the best accuracy.
- Logistic Regression: For logistic regression, one of the important parameters is the penalty. This parameter specifies the kind of regularization to be applied. This parameter can take two possible values,  $l_1$  regularization and  $l_2$  regularization. Both these values penalize high magnitude of the coefficients of the predictors in order to prevent the model from over-fitting. For our model, we have used  $l_2$  regularization as it works well even in the presence of highly correlated features.
- Random Forest: For the random forest, there are many parameters, such as the number of trees in the forest, the maximum depth of the trees, maximum number of features to consider at each split, the minimum number of samples required in a sub-tree to qualify for a further split, the minimum number of samples required to qualify as a leaf node, etc. For our data, we have kept most of the parameters at their default values, except for, the number of estimators or trees in the forest. We have set this value to 100, as the classifier resulted in very high accuracy with 100 trees in the forest.

## 7.8 Evaluation Metrics and Results

There are various evaluation metrics for assessing the performance of classifiers. Some of these evaluation metrics are [4]:

- Accuracy: This metric gives the proportion of the total number of correctly classified instances
- Precision: This gives the proportion of the true positive instances from the total instances classified as positive
- Recall: This gives the proportion of the positive instances that are correctly classified
- F-Measure: This gives the harmonic mean between precision and the recall values
- Confusion Matrix: This is a useful way of checking the accuracy of the classifier. It clearly shows the number of instances correctly classified for each label. Thus, if we know that the classes in the data are not well-balanced, it's always a good idea to check the confusion matrix along with accuracy. Consider a case where 95% of the instances belong to class A and only 5% of the instances belong to class B. If a classifier is trained on a dataset with such imbalance, there is a high chance that the classifier would return label A for each test instance. The classifier would still be able to correctly classify 95% of the test instances resulting in 95% accuracy. This is a case where accuracy can be misleading and thus a quick look at the confusion

matrix can help understand the problem with the classifier. For such a case, the confusion matrix will clearly show that all the instances of the minority class, B, have been misclassified.

For our model, we used accuracy as well as confusion matrix for evaluating the results. The confusion matrix did not show any serious issues for any of the classifiers. The accuracy for each of the three classifiers was:

- (1) Logistic Regression: With  $l_2$  penalty, the accuracy of logistic regression was 78.9%. Figure 11 shows the confusion matrix.

[Figure 11 about here.]

- (2) K Nearest Neighbors: With k as 3, the accuracy of kNN was 95.74%. Figure 12 shows the confusion matrix.

[Figure 12 about here.]

- (3) Random Forest: With a number of trees as 100, the accuracy of random forest classifier was 99.68%. Figure 13 shows the confusion matrix.

[Figure 13 about here.]

Thus, we obtained the best results with Random Forest classifier.

## 8 CONCLUSION

Analysis of food content is very important in today's world as most of the companies try to fool consumers by labeling their product as low-fat. It's important for the consumers to know the true nutrition grade while purchasing any food item. Thus, we analyzed the nutrition grade based on the composition of various components of the food items. We developed a model that labels a food item purely on the basis of its nutrients, thus eliminating any bias, such as, the production company or the brand name. For accurate labeling, we applied different data cleaning and data transformation techniques. With this transformed data, we tried various machine learning models. We got the best results using random forest classifier which was able to accurately label 99% of the food products. Since the model is trained only for France, as part of future work, we can try and scale our model for different countries. However, to achieve similar results for other countries, we need to collect more data. The current data has many missing values for countries other than France. Once we collect enough data for these countries, we can also try and implement more sophisticated models like neural networks in future.

## ACKNOWLEDGMENTS

This project was undertaken as a part of the course objective for I523: Big Data Applications and Analytics at Indiana University, Bloomington. We would like to thank Dr. Gregor von Laszewski and all the TAs for their help, support, and suggestions.

## A WORK BREAKDOWN

**Dataset identification:** Karthik Vegi, Nisha Chandwani: work equally split between.

**Requirement Gathering:** Karthik Vegi, Nisha Chandwani: work equally split between.

**Learning Machine Learning Concepts:** Karthik Vegi, Nisha Chandwani: work equally split between.

**Data analysis and implementation of the Logistic Regression:** Karthik Vegi.

**K nearest neighbors and Random Forest algorithms:** Nisha Chandwani

**Writing the project report:** Karthik Vegi, Nisha Chandwani: work equally split between.

## REFERENCES

- [1] American Heart Association. 2017. Dietary Fats. Webpage. (March 2017). <https://healthyforgood.heart.org/eat-smart/articles/dietary-fats>
- [2] Alejandro Cifuentes. 2012. Food analysis: present, future, and foodomics. *ISRN Analytical Chemistry* 2012 (2012), 16.
- [3] World Health Organization Europe. 2017. Labelling systems to guide consumers to healthier options. Webpage. (March 2017). <http://www.euro.who.int/en/countries/france/news/news/2017/03/france-becomes-one-of-the-first-countries-in-region-to-recommend-colour-coded-front-of-pack>
- [4] M Hossin and MN Sulaiman. 2015. A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process* 5, 2 (2015), 1.
- [5] Healthy Eating SFGate. 2017. Recommended Daily Allowances of Fats, Sugars, Sodium for Adults. Webpage. (2017). <http://healthyeating.sfgate.com/recommended-daily-allowances-fats-sugars-sodium-adults-2976.html>
- [6] Shai Shalev-Shwartz and Shai Ben-David. 2014. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, New York, USA.
- [7] Statistics Solutions. 2017. Multicollinearity. Webpage. (March 2017). <http://www.statisticssolutions.com/multicollinearity/>
- [8] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. 2006. *Introduction to Data Mining*. Pearson, Boston, USA.
- [9] Karthik Vegi and Nisha Chandwani. 2017. Code base - Analysis on food products around the world. github. (Dec. 2017). <https://github.com/bigdata-i523/hid231/tree/master/project/code>

#### LIST OF FIGURES

1	Top 5 countries [9]	9
2	Top 5 countries with most fat content [9]	10
3	Top 5 countries with most saturated fat content [9]	11
4	Top 5 countries with most trans-fat content [9]	12
5	Top 5 countries with most cholesterol content [9]	13
6	Top 5 countries with most sugar content [9]	14
7	Top 5 countries with most sugar content [9]	15
8	K nearest neighbors algorithm[8]	15
9	Bi-variate box plots [9]	16
10	Correlation Plot [9]	17
11	Confusion matrix for Logistic Regression [9]	18
12	Confusion matrix for K Nearest Neighbors [9]	19
13	Confusion matrix for Random Forest [9]	20

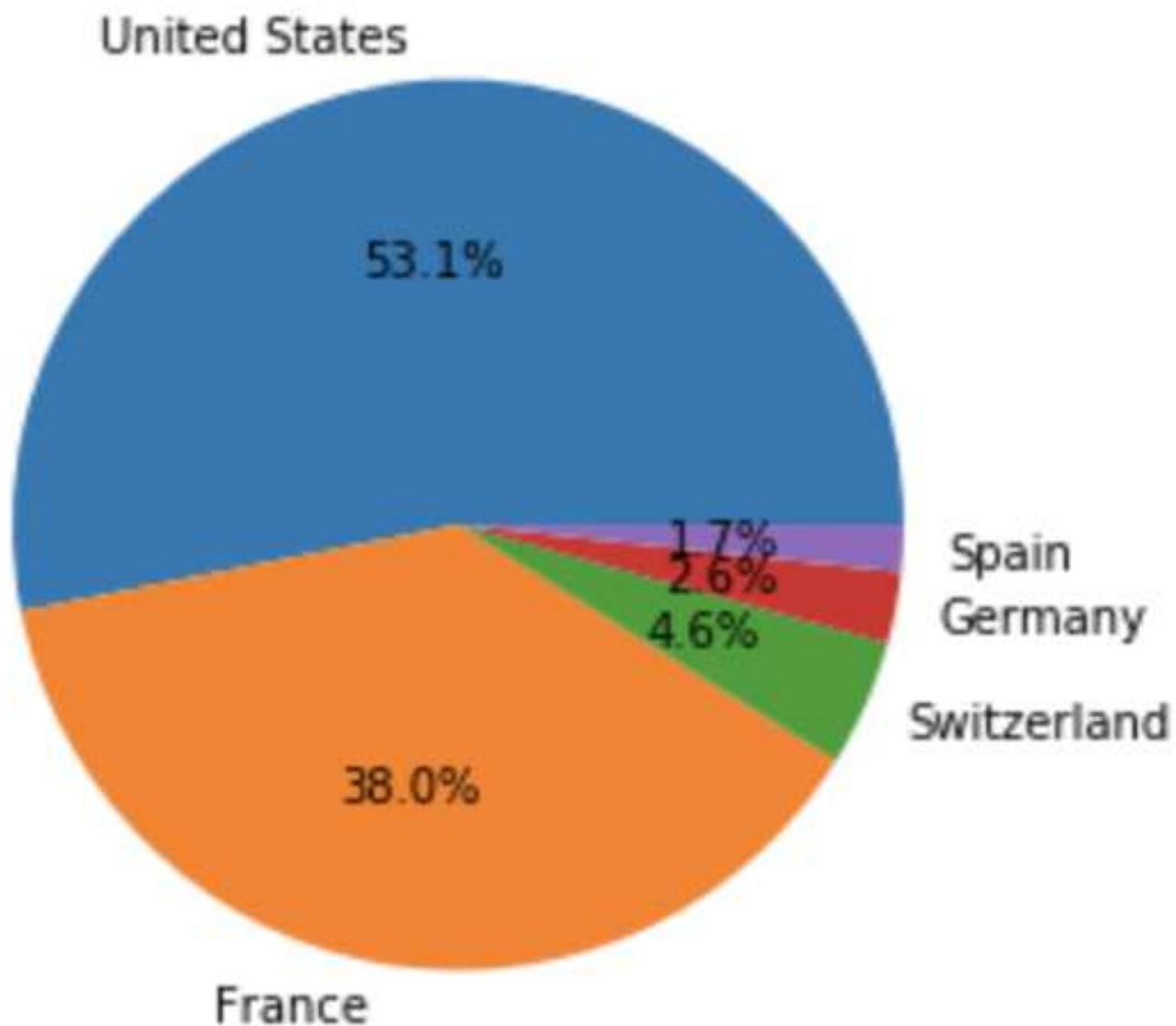


Figure 1: Top 5 countries [9]

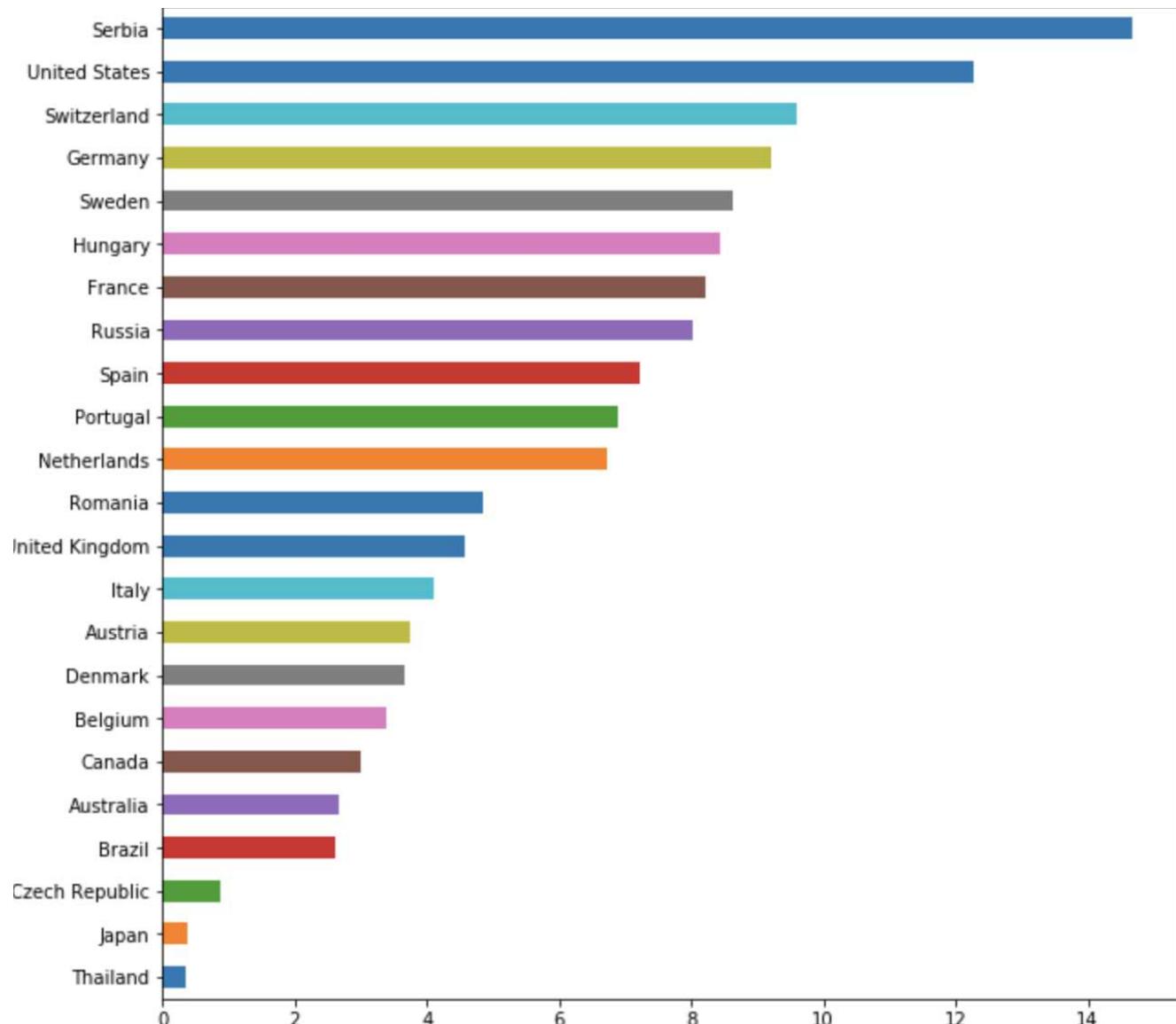


Figure 2: Top 5 countries with most fat content [9]

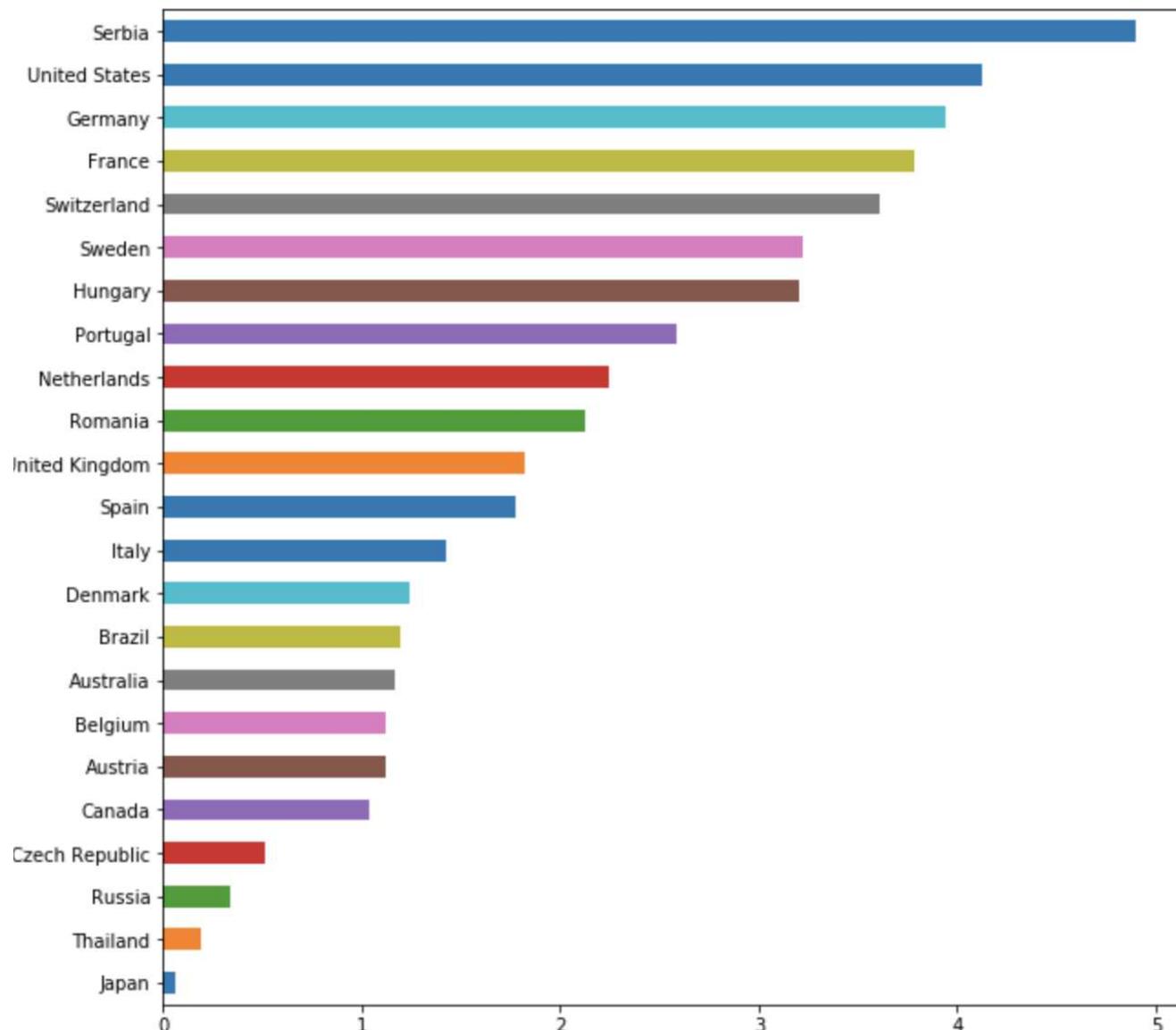


Figure 3: Top 5 countries with most saturated fat content [9]

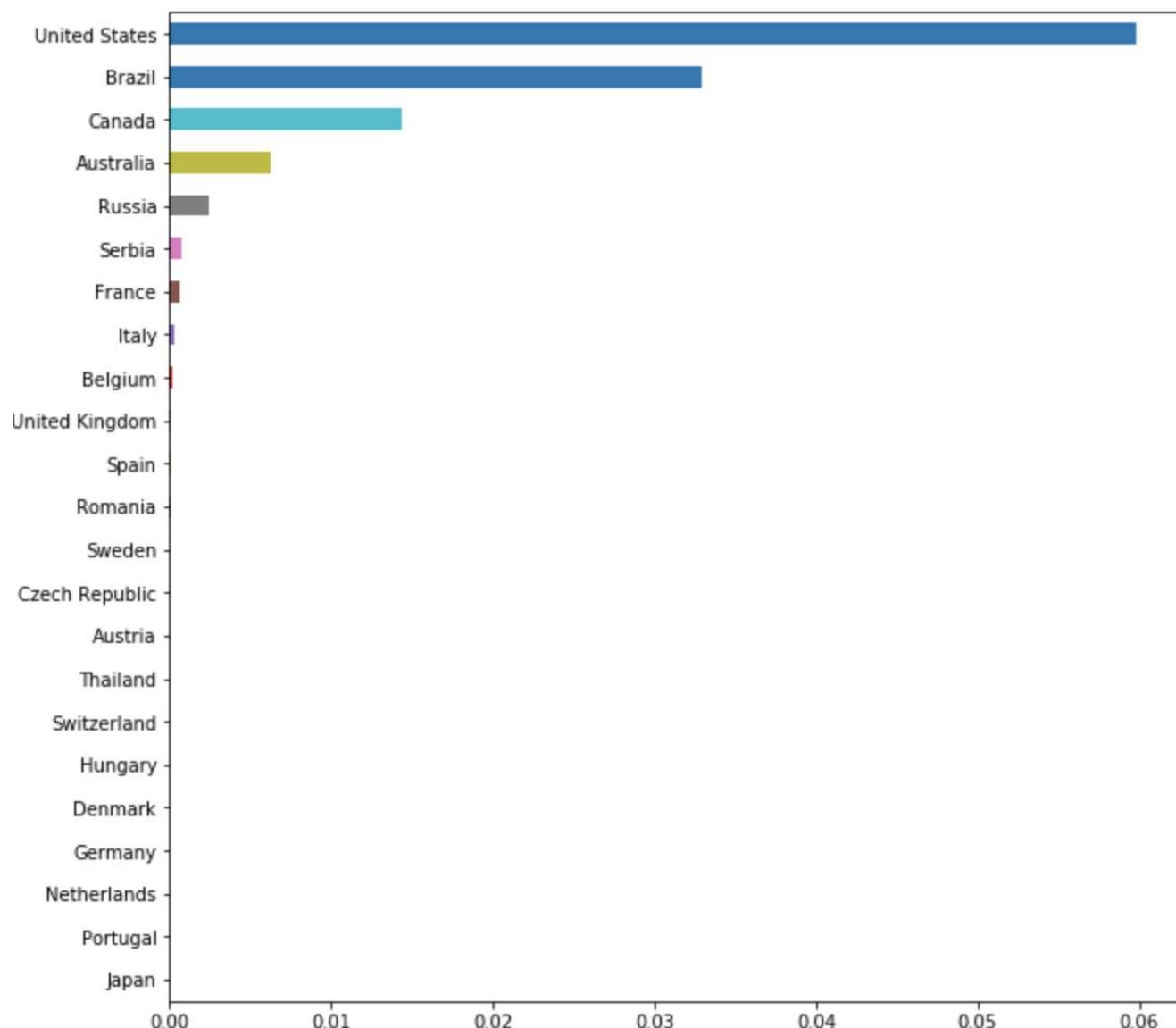


Figure 4: Top 5 countries with most trans-fat content [9]

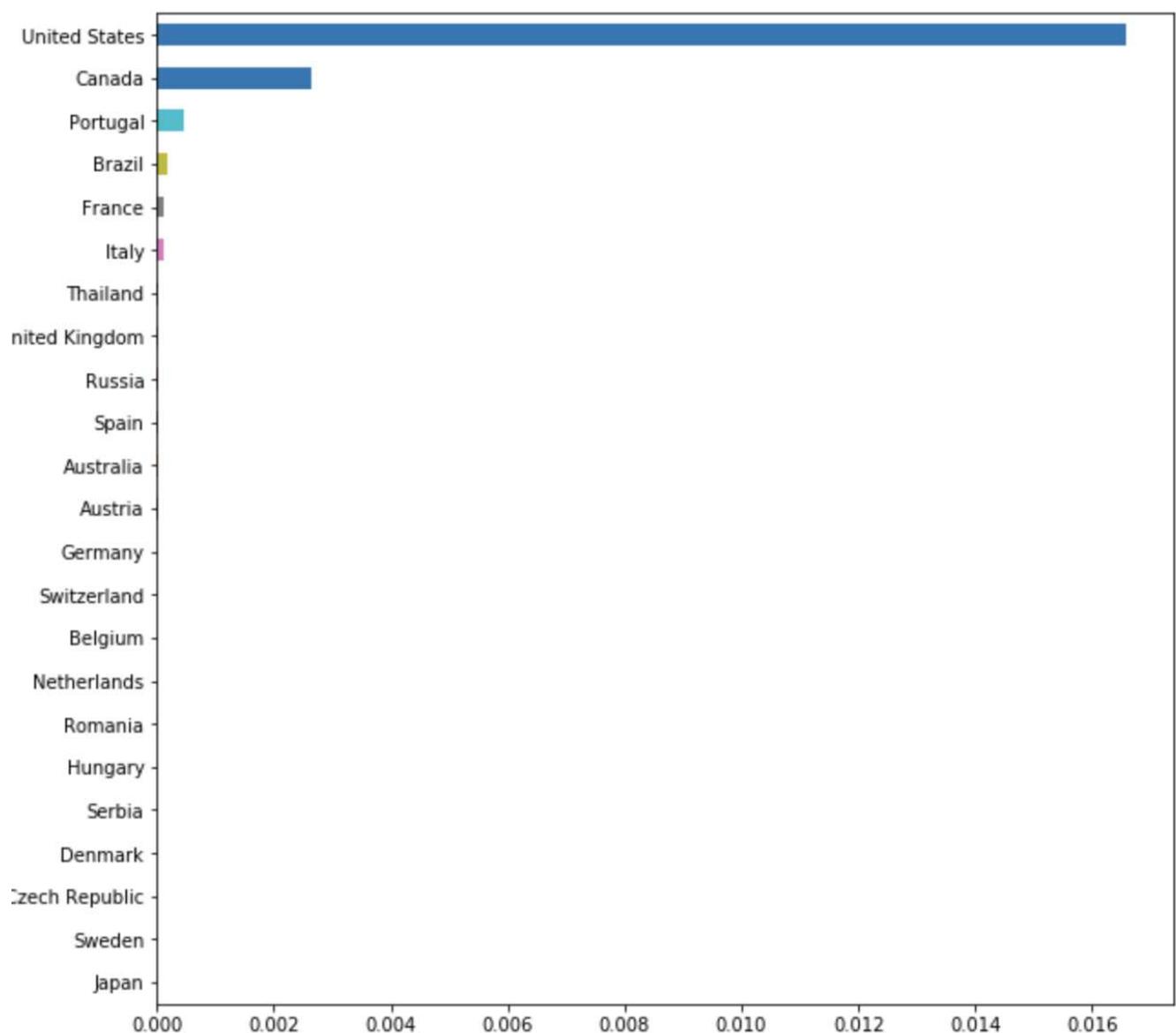


Figure 5: Top 5 countries with most cholesterol content [9]

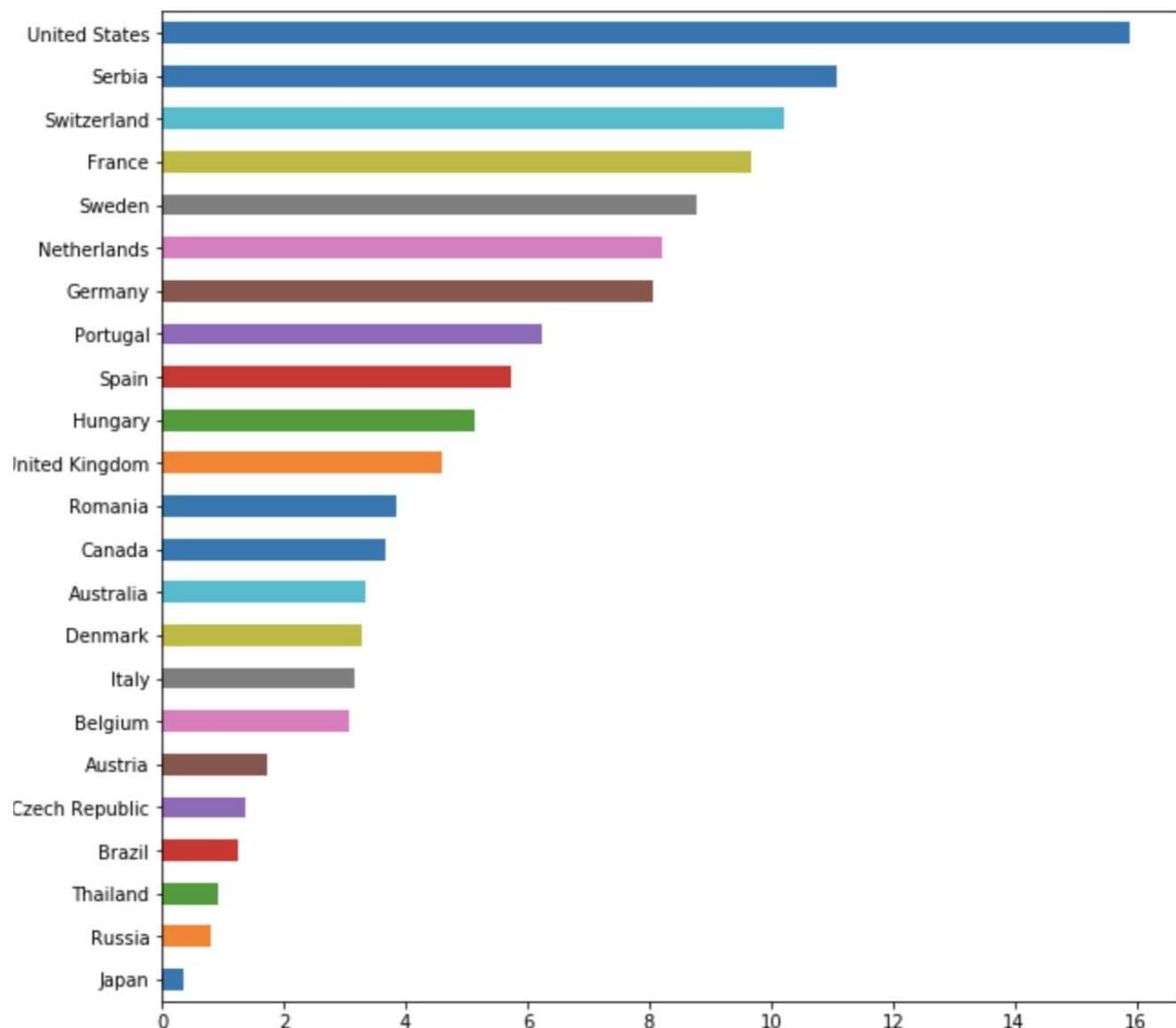


Figure 6: Top 5 countries with most sugar content [9]

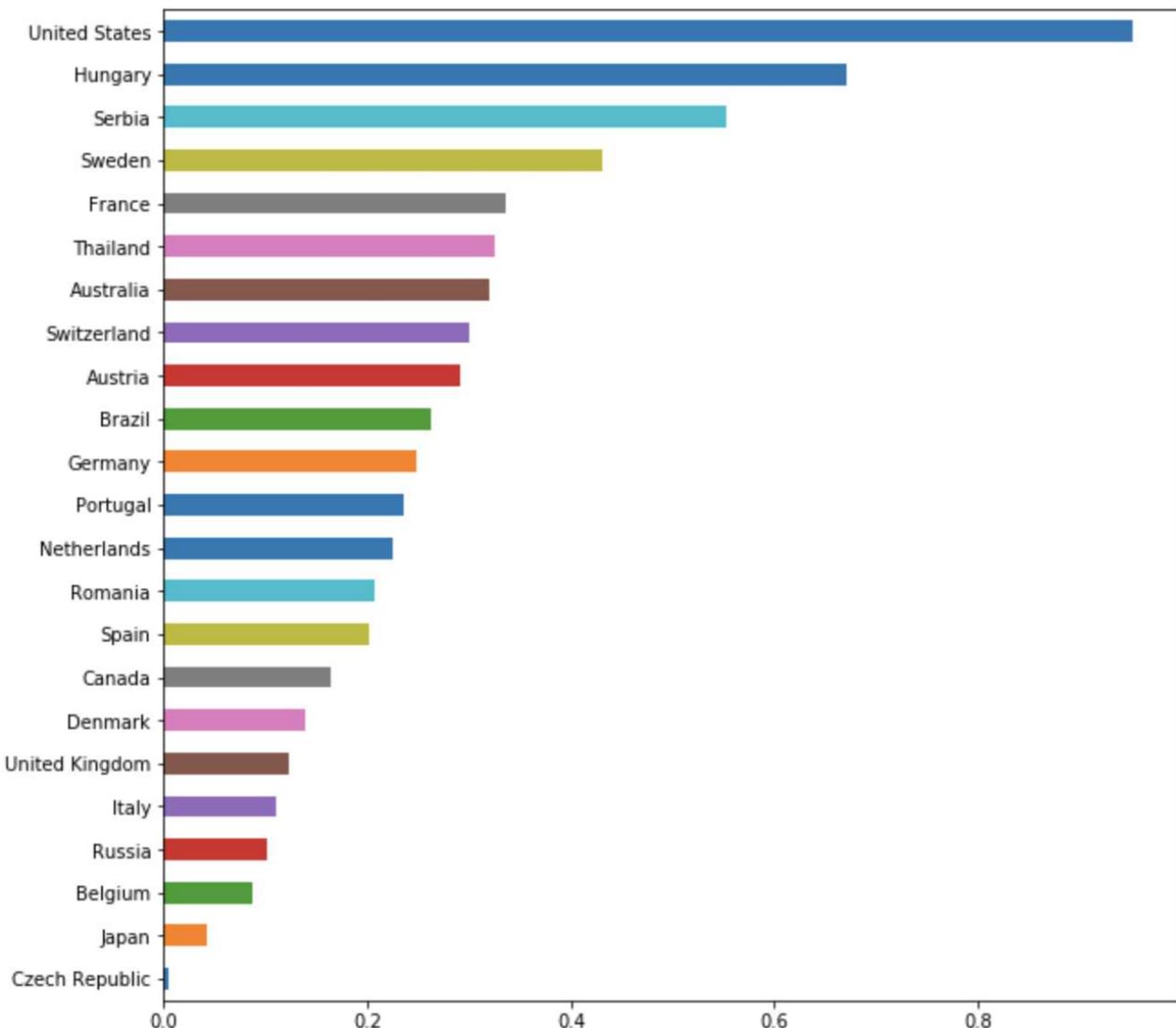


Figure 7: Top 5 countries with most sugar content [9]

---

### Algorithm 5.2 The $k$ -nearest neighbor classification algorithm.

---

- 1: Let  $k$  be the number of nearest neighbors and  $D$  be the set of training examples.
  - 2: **for** each test example  $z = (\mathbf{x}', y')$  **do**
  - 3:   Compute  $d(\mathbf{x}', \mathbf{x})$ , the distance between  $z$  and every example,  $(\mathbf{x}, y) \in D$ .
  - 4:   Select  $D_z \subseteq D$ , the set of  $k$  closest training examples to  $z$ .
  - 5:    $y' = \operatorname{argmax}_v \sum_{(\mathbf{x}_i, y_i) \in D_z} I(v = y_i)$
  - 6: **end for**
- 

Figure 8: K nearest neighbors algorithm[8]

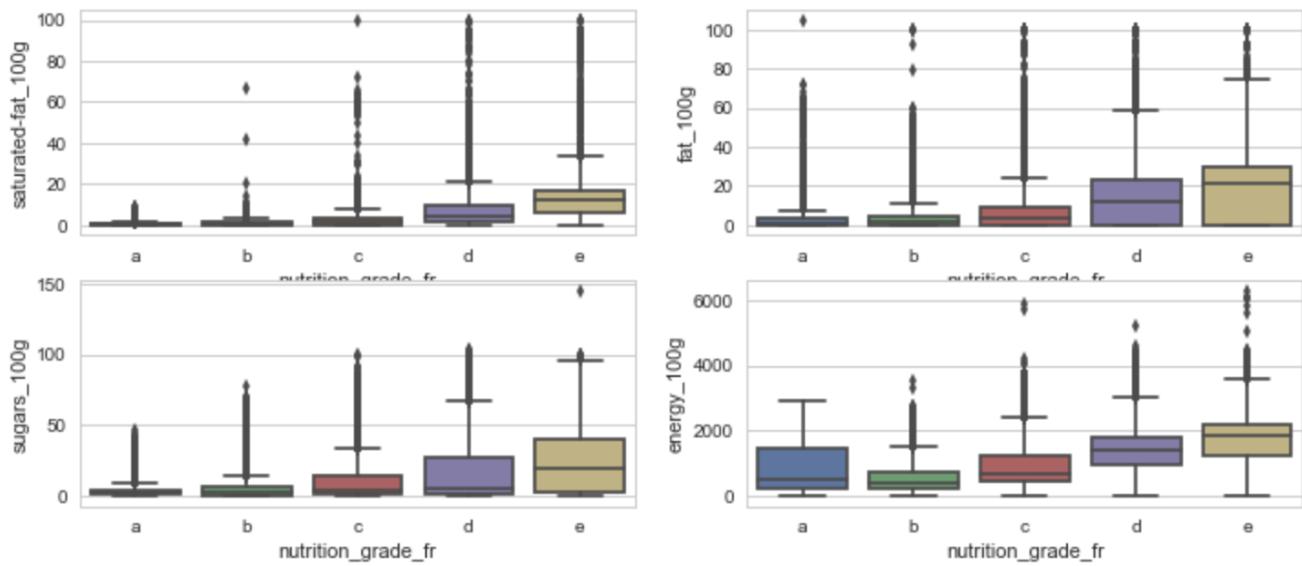


Figure 9: Bi-variate box plots [9]

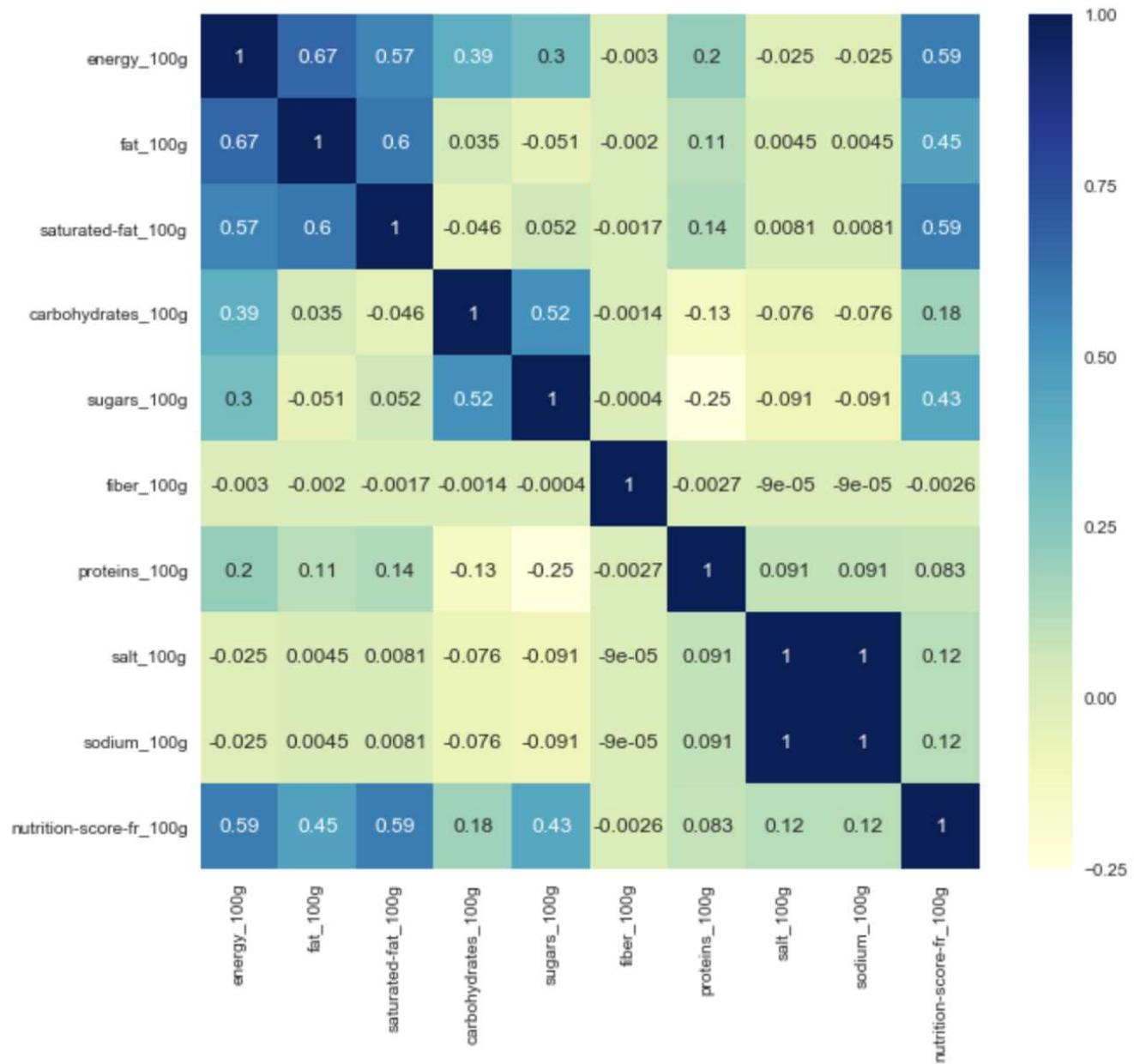


Figure 10: Correlation Plot [9]

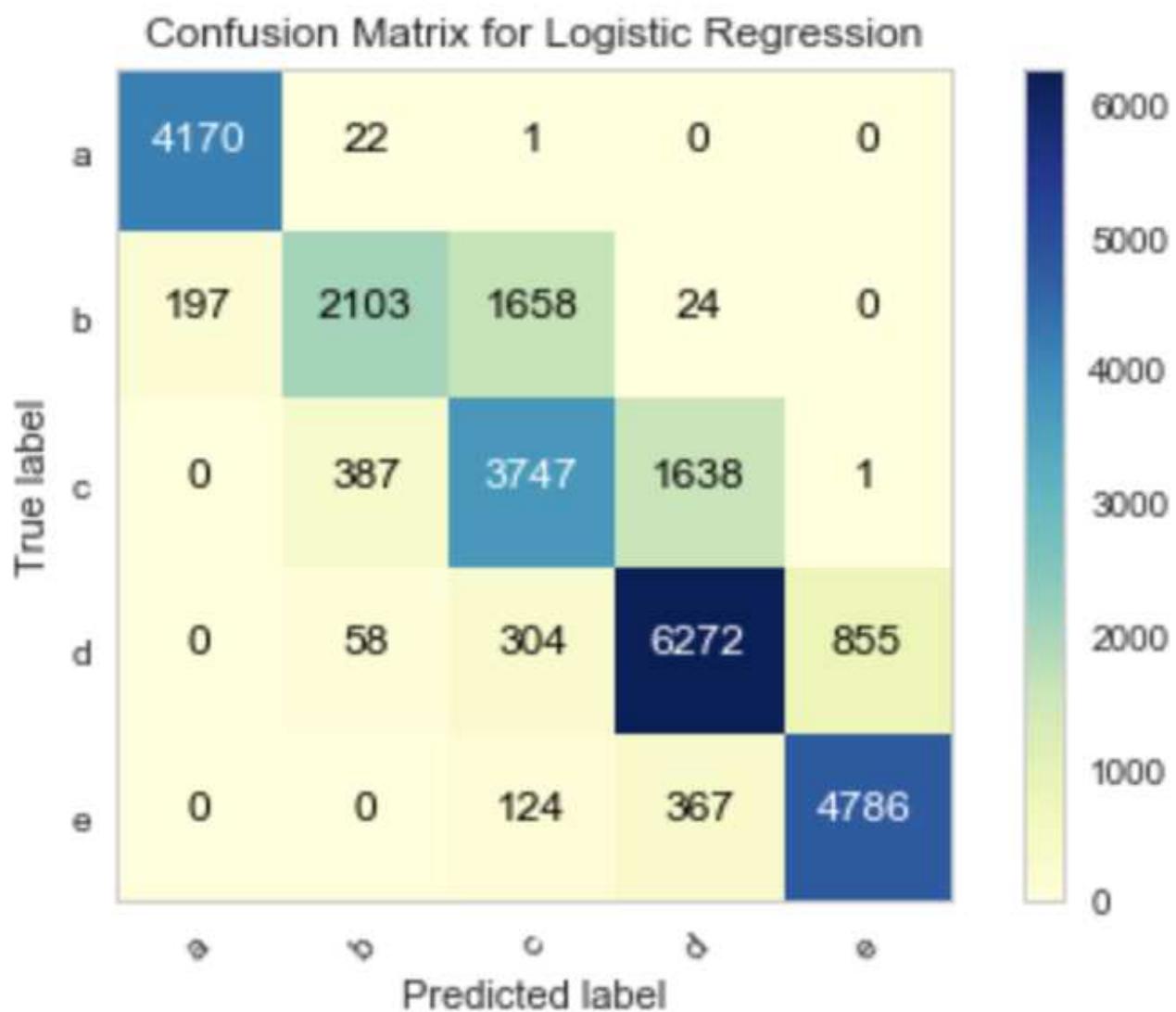


Figure 11: Confusion matrix for Logistic Regression [9]

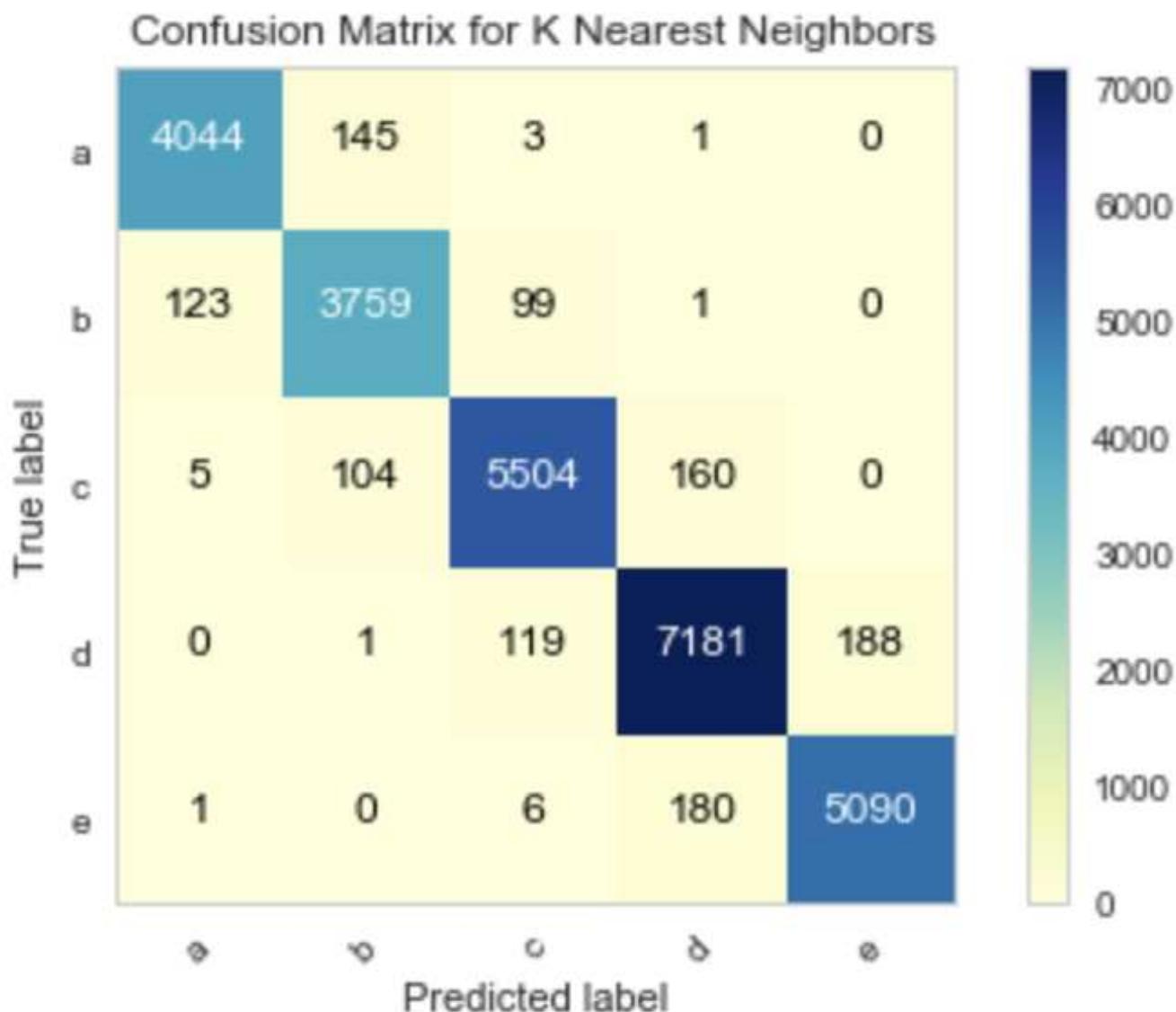


Figure 12: Confusion matrix for K Nearest Neighbors [9]

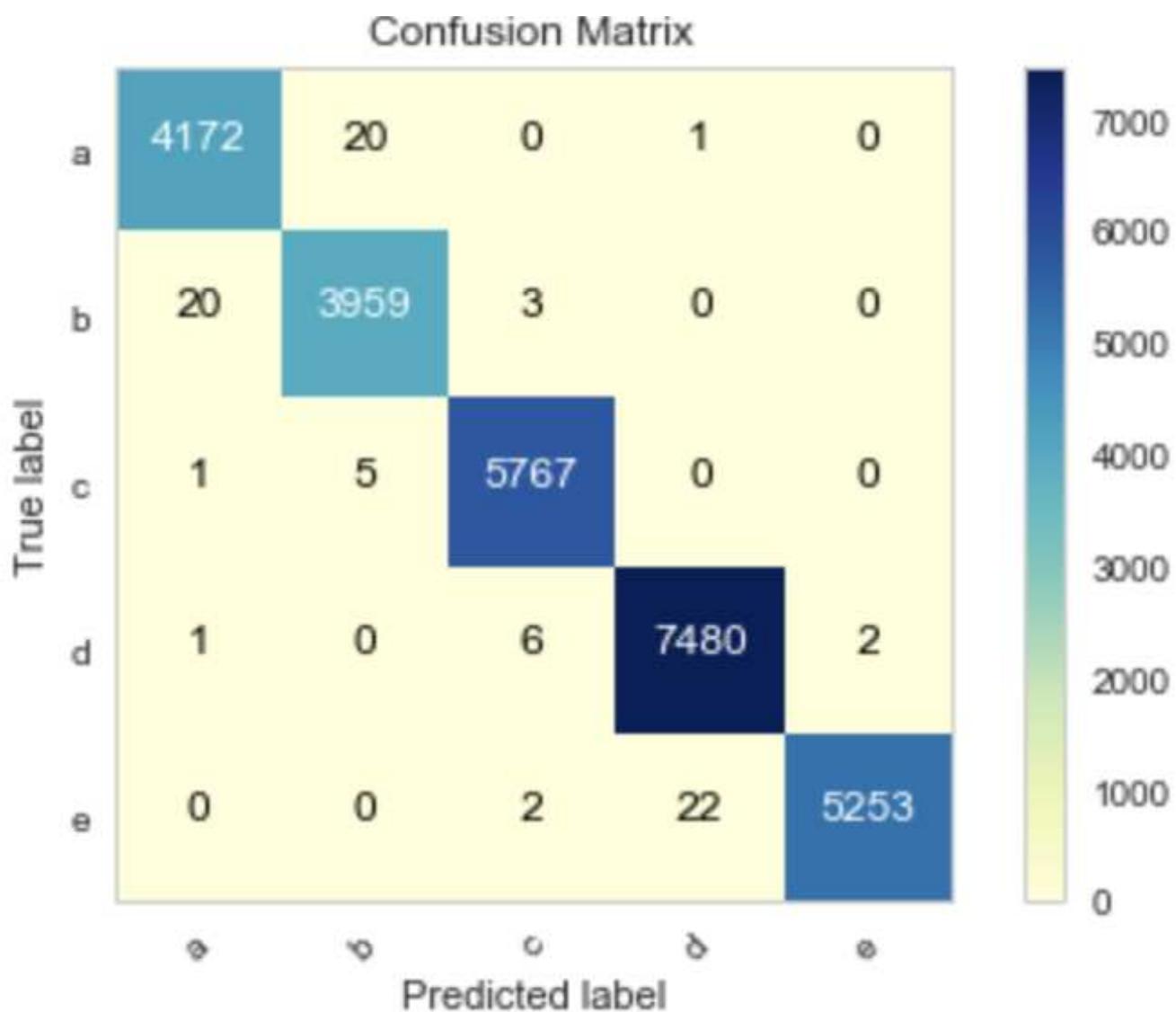


Figure 13: Confusion matrix for Random Forest [9]

```
bibtext report
```

```
=====
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
```

```
=====
bibtext _ label error
```

```
=====
bibtext space label error
```

```
=====
bibtext comma label error
```

```
=====
latex report
```

```
=====
[2017-12-11 13.27.41] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflatex)
Missing character: ""
bookmark level for unknown defaults to 0.
The anchor of a bookmark and its parent's must not be the same. Added a new anchor.
Typesetting of "report.tex" completed in 2.5s.
./README.yml
36:70     error    trailing spaces (trailing-spaces)
37:81     error    line too long (85 > 80 characters) (line-length)
```

```
=====
Compliance Report
```

```
=====
name: Vegi, Karthik
hid: 231
paper1: Oct 29 17 100%
paper2: 100%
project: 100% Dec 4, 2017
```

```
yamlcheck
```

---

```
wordcount
```

---

```
(null)
wc 231 project (null) 6225 report.tex
wc 231 project (null) 6232 report.pdf
wc 231 project (null) 250 report.bib
```

```
find "
```

---

```
passed: True
```

```
find footnote
```

---

```
passed: True
```

```
find input{format/i523}
```

---

```
4: \input{format/i523}
```

```
passed: True
```

```
find input{format/final}
```

---

```
passed: False
```

```
floats
```

---

```
69: We then display the top 5 countries as a pie-chart and the 5
countries are namely United States, France, Switzerland, Germany,
and Spain as shown in Figure \ref{fig:Fig7}.
```

```
71: \begin{figure}
```

```
72: \includegraphics[width=1.0\columnwidth]{images/fig7.png}
```

```
74: \label{fig:Fig7}
```

```
80: The top 5 countries with most fat content in the food items are
Serbia, United States, Switzerland, Germany, and Sweden as shown
in Figure \ref{fig:Fig8}.
```

```

82: \begin{figure}
83: \includegraphics[width=1.0\columnwidth]{images/fig8.png}
85: \label{fig:Fig8}
88: The top 5 countries with most saturated fat content in the food
     items are Serbia, United States, Germany, France and Switzerland
     as shown in Figure \ref{fig:Fig9}
90: \begin{figure}
91: \includegraphics[width=1.0\columnwidth]{images/fig9.png}
93: \label{fig:Fig9}
97: The top 5 countries with most trans-fat content in the food items
     are United States, Brazil, Canada, Australia, Russia, and Serbia
     as shown in Figure \ref{fig:Fig10}. \\
99: \begin{figure}
100: \includegraphics[width=1.0\columnwidth]{images/fig10.png}
102: \label{fig:Fig10}
105: The top 5 countries with most cholesterol content in the food
     items are United States, Canada, Portugal, Brazil, France, and
     Italy as shown in Figure \ref{fig:Fig11}. \\
107: \begin{figure}
108: \includegraphics[width=1.0\columnwidth]{images/fig11.png}
110: \label{fig:Fig11}
116: The top 5 countries with most sugar content in the food items are
     United States, Serbia, Switzerland, France, and Sweden as shown
     in Figure \ref{fig:Fig12}. \\
118: \begin{figure}
119: \includegraphics[width=1.0\columnwidth]{images/fig12.png}
121: \label{fig:Fig12}
124: The top 5 countries with most sodium content in the food items
     are United States, Hungary, Serbia, Sweden, and France as shown
     in Figure \ref{fig:Fig13}. \\
126: \begin{figure}
127: \includegraphics[width=1.0\columnwidth]{images/fig13.png}
129: \label{fig:Fig13}
163: The nearest neighbor puts each attribute list as a data point in
     the n-dimensional space, given n the number of attributes
     \cite{book-tan}. Once we have the training examples, we take each
     test example and compute its distance to the training example
     classes and assign a class label \cite{book-tan}. Any of the
     popular distance measures among Euclidean distance, Manhattan
     distance, Minkowski distance and Mahalanobis distance can be used
     \cite{book-tan}. The k denotes the k closest points to the test
     example \cite{book-tan}. Figure \ref{fig:Fig1} shows the
     algorithm \cite{book-tan}.
165: \begin{figure}
166: \includegraphics[width=1.0\columnwidth]{images/fig1.png}
168: \label{fig:Fig1}

```

```

243: \subsubsection{Bi-variate box-plots} Bi-variate box-plots go
beyond uni-variate box plots by showing the relationship between
the predictor variable and the response variable \cite{book-tan}.
We look at the bi-variate box-plots for each of the important
predictor variables namely, saturated fat, polyunsaturated fat,
sugars and salt and the response variable, nutrition grade.
Figure \ref{fig:Fig2} shows the bi-variate box plots. \\
245: \begin{figure}
246: \includegraphics[width=1.0\columnwidth]{images/fig2.png}
248: \label{fig:Fig2}
254: Correlation between data objects is the measure of the linear
relationship between the attributes of the object that are
continuous variables \cite{book-tan}. Correlation analysis is the
process of finding of the correlations between the different
predictor variables and identify high collinearity problem
\cite{book-shai}. The relationship could be either linear or non-
linear based on the given data \cite{book-tan}. The correlation
coefficient can range anywhere between -1 and 1, where 1
indicates a very high positive correlation and -1 indicates a
very high negative correlation \cite{book-shai}. Correlation plot
visually shows the correlation coefficient between the variables
in a nicely laid out plot. Figure \ref{fig:Fig3} shows the
correlation plot. \\
256: \begin{figure}
257: \includegraphics[width=1.0\columnwidth]{images/fig3.png}
259: \label{fig:Fig3}
304: \item Logistic Regression: With $l_2$ penalty, the accuracy of
logistic regression was 78.9\%. Figure \ref{fig:Fig4} shows the
confusion matrix. \\
306: \begin{figure}
307: \includegraphics[width=1.0\columnwidth]{images/fig4.png}
309: \label{fig:Fig4}
312: \item K Nearest Neighbors: With k as 3, the accuracy of kNN was
95.74\%. Figure \ref{fig:Fig5} shows the confusion matrix. \\
314: \begin{figure}
315: \includegraphics[width=1.0\columnwidth]{images/fig5.png}
317: \label{fig:Fig5}
320: \item Random Forest: With a number of trees as 100, the accuracy
of random forest classifier was 99.68\%. Figure \ref{fig:Fig6}
shows the confusion matrix. \\
322: \begin{figure}
323: \includegraphics[width=1.0\columnwidth]{images/fig6.png}
325: \label{fig:Fig6}

```

figures 13  
tables 0

```
includegraphics 13
labels 13
refs 13
floats 13
```

```
True : ref check passed: (refs >= figures + tables)
True : label check passed: (refs >= figures + tables)
True : include graphics passed: (figures >= includegraphics)
True : check if all figures are refered to: (refs >= labels)
```

```
Label/ref check
passed: True
```

```
When using figures use columnwidth
[width=1.0\columnwidth]
do not change the number to a smaller fraction
```

```
find textwidth
```

---

```
passed: True
```

```
below_check
```

---

```
bibtex
```

---

```
label errors
```

```
bibtex errors
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
```

```
bibtex_empty_fields
```

---

```
entries in general should not be empty in bibtex
```

```
find ""
```

---

```
passed: True
```

```
ascii
```

---

```
non ascii found 8211
```

```
=====
The following tests are optional
=====
```

```
Tip: newlines can often be replaced just by an empty line
```

```
find newline
```

---

```
passed: True
```

```
cites should have a space before \cite{} but not before the {
```

```
find cite {
```

---

```
passed: True
```

# Recipe Ingredient Analysis

Sushant Athaley  
Indiana University  
sathaley@iu.edu

## ABSTRACT

Food is the unavoidable part of day to day of human life. Ingredients play a major role or are the basic requirement in preparation of any kind of food. We can find the humongous list of ingredients getting used across globally along with other details which constitute to big data. We explore ingredients getting used in various recipes across the globe to understand most used ingredient, key ingredients of various cuisine and the relationship between the ingredients to find out closely related ingredients which can always provide great dish if used together.

## KEYWORDS

i523, hid302, big data, ingredient, recipe, analysis, python, gephi

## 1 INTRODUCTION

Ingredients are vital for human existence as well as for food or restaurant industry. We use it every day for cooking and food industry uses it to produce consumable for their customers. Ingredient inspires chefs to come up with new culinary artistry. So what do we know about this essential element of the life and what data tell us? Ingredients come in different size, color, shape, flavor, nutrition, taste, texture, grows in specific weather conditions and this provides a great opportunity for various analysis which can be useful for the human being as well as business industries. So main focus of this study is on the ingredients used in various recipes across the cuisines. This study evaluates recipe ingredient dataset from Kaggle [8] to analyze most used ingredients, key cuisine ingredients and ingredient relationship.

This study is organized as follows, section *Ingredient* defines ingredient and its various characteristics. section *Ingredieint Analytics* describes various analytics which can be performed on the ingredient with some examples. Section *Project* describes the aim of this study. Section *technologies* provides information on the tools and technologies used for this project. Section *Methodology* covers overall process carried out in this project. Section *Dataset* describes data structure used along with loading process and data findings. Section *Analysis and Findings* describes various analysis carried out on the data and the visual representation of the analysis. Section *Shortcomings* captures shortcomings of the project. Section *Limitations* talks about limitations and what else can be done with this dataset which is not covered in the current scope of the project. Section *Conclusion* concludes the study.

## 2 INGREDIENT

Food is defined as “Edible or potable substance (usually of animal or plant origin), consisting of nourishing and nutritive components such as carbohydrates, fats, proteins, essential mineral and vitamins, which (when ingested and assimilated through digestion) sustains life, generates energy, and provides growth, maintenance,

and health of the body” [2]. Thus food is the basic necessity for human for the sustainability. Food can be eaten raw, cooked or processed. As human race evolved over the period of time, the way we eat food is also evolved. Food cooking is just not the basic necessity but its an art and science in today’s era. Food preparation consists of various cooking techniques, tools, and ingredients to make it palatable or edible by humans. The ingredient is by far the most important part of any food or recipe preparation. The recipe consists of the list of ingredients and the set of instruction to cook particular food dish [6]. An ingredient is defined as “Any of the foods or substances that are combined to make a particular dish” [10]. Ingredients impart various flavors, aroma, texture, and color to the cooking dish. Ingredients are mostly derived from vegetables, fruits, nuts, grains, living organisms, herbs, flowers, and spices. It comes in both solid and liquid forms. Another characteristic of ingredients is the nutritional value they provide which is essential for the human body.

## 3 INGREDIENT ANALYTICS

Ingredients characteristics and the combination of other related data provides various opportunities to analyze ingredient in different ways. Analysis of the flavors present in ingredient can provide us with the categorization of the different ingredient by the flavor profile which can be helpful in deciding substitute ingredient if a certain ingredient is not present or pairing ingredient from different flavor categories to construct the dish as per the taste required [1]. This analysis also helps to understand which ingredients cannot be used together. A similar analysis is carried out to correlate ingredient across recipes to come up with top 50 combinations of ingredients which can be used together [7]. Flavourspace application provides functionality to search recipe based on the ingredients, suggests alternate ingredient if not present, adjust the recipe as per the taste which is a good example of big data analytics in food industry [11]. Foodpairing application takes another approach to form the connection between unfamiliar ingredients and provides information on how to use such ingredient to make a dish, this is very helpful in terms of sustainability as we can use ingredient which is ample available but not in use due to the absence of information on using such ingredients [9].

Recipe recommendation system uses users recipe browsing history or rating history to suggest the recipe. It also relay on the ingredient present in the recipe and look for the overlap or key ingredients while matching other recipes. Another approach is to recommend recipe based on the nutritional values or healthy food choice which is dependent on the ingredient used in the recipe. Models are made to recommend recipe based on the available ingredients and personal nutrition needs. Chen-Yuen et al. [4] derived network of complimenting and substituting ingredients. They also demonstrated that network can be used to predict which recipe

would be successful. To understand the complimenting ingredient they constructed network based on pointwise mutual information (PMI) defined on pairs of ingredients. This PMI provides the probability of those two ingredients occurring together. Their study found out 2 main cluster as savory and sweet dishes along with the a satellite cluster of mixed drink ingredients. This study also finds out ingredient adjustment and substitution based on the comments on the recipe. Recipe comments provides insight into which ingredients quantity is increased or decreased to get more flavors or which ingredient is used instead of some ingredient in the recipe since ingredient mentioned in recipe is not present or to get different taste. The words like add, omit, instead, adding, using, more etc in comments provides this insight. Ingredient which are considered as unhealthy like sugar, fats are often reduced and ingredient which adds flavors like soy sauce, lemon juice, cinnamon are added more in quantity. Chicken can be substituted by turkey, beef, sausage, chicken breast, bacon and olive oil by butter, apple sauce, oil, banana, margarine, and Tilapia by cod, catfish, flounder, halibut, orange roughy to name few.

Another study conducted on most used ingredient provides insight that sugar, oil, pepper, and salt are most commonly occurring ingredient, among spices clove, in vegetable onion, garlic , and tomatoes, butter in milk product, eggs followed by chicken in the animal product are the most used ingredient in the categories [3]. This information can help in better planning and sourcing of such ingredients which are in high demand.

Ingredient nutrition analysis can help find out nutrition of the food prepared by those ingredients. This would be helpful in menu planning where nutrition information is the key factor such as school, hospitals or any other dietary program [5].

Recipe cost is calculated by including the cost of the ingredient used in that recipe. Ingredient cost as per the quantity used in recipe provides base information to calculate the price of any recipe. This ingredient cost analysis provides an avenue to reduce the cost of the recipe by using substitute ingredient of lesser cost. This can also help in household budget to keep in check as well as make restaurant industry profitable.

Ingredient used in recipe can provide insight into type of weather received by that cuisine as ingredient can grow in certain weather condition. This can help chef locally source the ingredient and maintain local agriculture sustainability.

## 4 PROJECT

This project study is conducted to analyze ingredients getting used in various recipes across the cuisines to find out

- Most used ingredients across cuisines or globally
- Key ingredients used by cuisines
- Ingredient relationship or connection to understand the related ingredients

### 4.1 Technologies

Technologies and tools used in this projects are

- Python version 3.6 is used for data load and processing
- Gephi 0.9.2 for visualization
- Spyder 3.0 as a Python IDE

### 4.2 Code Organization

Code is checked-in in Github at location  
<https://github.com/bigdata-i523/hid302/tree/master/project/code>  
 Code is organized as described in Figure 1

[Figure 1 about here.]

#### Python Scripts

- *ingredientAnalysis.py* - This python script loads dataset from datafile train.json present in *data* directory and process dataset to find out recipe distribution across cuisines, top 20 ingredient used across cuisines and top 10 key ingredients for every cuisine. The graph generated during analysis is stored in *images* folder.
- *ingCluster.py* - loads dataset from datafile train.json present in *data* directory and process dataset to create relationship file required by Gephi in excel format. It establishes ingredients relationship by relating ingredient in recipe with each other to generate *nodes.xlsx* and *edges.xlsx* files and stores in *data* directory. These generated files are then imported into the Gephi to create the visualization.
- *geph\_ing\_big\_data.gephi*
  - This is project file from Gephi which can be re-opened in Gephi software to view or re-run the analysis.

### 4.3 Methodology

The first step was to source the data. We were interested in the dataset which provides recipe information along with the ingredient used in the recipe. Since we wanted to analyze distribution across cuisines, data should also contain cuisine tagging. This dataset can be generated by pulling recipe data from various online applications or pick from publicly available datasets. We finalized publicly available dataset at Kaggle application satisfying need for this project.

Figure 2 shows methodology used for this project to analyze ingredient data.

[Figure 2 about here.]

DataSet is loaded through Python script and further processed to clean the data. This cleaned data then processed to analyze ingredient distribution across cuisine and per cuisine. Gephi software is used to analyze the relationship and to find out the ingredient modularity. The Python script is used to create the network files required by the Gephi tool. Gephi requires Nodes and relationship in terms of Edges between the nodes for the analysis. The Python script is used to create Node and Edges file in excel format so that it can be imported into Gephi. Distinct ingredients used in recipe becomes the nodes. Edges or relationship between ingredients is derived by relating ingredients appearing in the same recipe. All ingredient in the same recipe is considered related to each other. Network files created by Python are imported in Gephi to produce the graph for the visualization. Gephi tools data laboratory is used to clean up the data and filters are applied to provide usable network visualization.

### 4.4 Dataset

The dataset for this study is sourced from Kaggle application [8]. This dataset is publicly available and featured in *What's Cooking?*

competition. This dataset is in JSON format and of 12MB size. This dataset contains recipe id, cuisine and list of ingredients as described in Figure 3.

[Figure 3 about here.]

This dataset contains total 39774 recipes across various cuisines. We used two different methods to load this data. Cuisine and ingredient analysis is done by loading data into *pandas dataframe* and to analyze ingredient relationship data has been loaded into *json* object. Figure 4 shows the code for data loading used in this project.

[Figure 4 about here.]

Ingredient extraction from the data structure and processing was challenging as ingredients are listed comma separated for each recipe. Also, ingredient list can vary by recipe and there is no proper structure. Another issue with the ingredient list is ingredient appears in various forms but it's the same ingredient which gives duplicate data. For example, salt appears as salt, kosher salt, Morton Salt, sea salt, table salt, Himalayan salt, fine sea salt, low sodium salt, fine salt. This is the same ingredient but come across in recipe as a different ingredient and getting counted as a separate ingredient in the analysis. Some ingredients are listed along with measures like (10 oz.) frozen chopped spinach, (10 oz.) frozen chopped spinach, thawed and squeezed dry, (14.5 oz.) diced tomatoes and getting counted as a separate ingredient. Some ingredients are listed along with the brand name like KRAFT Reduced Fat Shredded Mozzarella Cheese, Johnsonville Smoked Sausage, Johnsonville Mild Italian Sausage Links etc and also constitutes to the ingredient list. This variation makes difficult to get the proper ingredient list for the analysis. Extensive work is needed to clean and correct the noisy data so that proper analysis can be carried out. This correction process is not carried out as part of this project.

Certain ingredients like salt or water etc should be avoided from the analysis as those are not the ingredient we are looking for the analysis. We tried to clean such elements during ingredient relationship analysis but we had little success as those ingredients are present in the dataset in various forms.

## 4.5 Analysis and Findings

**4.5.1 Recipe Distribution By Cuisine.** We first analyze entire dataset to understand the total number of recipes and their distribution across various cuisines. We use Pythons Panda library to get the total recipe count as 39774 and plot the distribution. Figure 5 shows number of recipes per cuisine. Dataset is heavily dominated by Italian cuisine followed by Mexican cuisine and with very fewer recipes from Russian and Brazilian cuisines. This also highlights another shortcoming of the dataset that it doesn't have equal representation of all cuisines which might give us biased analysis.

[Figure 5 about here.]

Table1 describes recipe count for every cuisine.

[Table 1 about here.]

**4.5.2 Most Used Ingredients All Cuisines.** The second analysis is carried out to understand top 20 ingredients getting used across cuisine or globally. Ingredient *Salt* is obvious topper followed by *Oil* and *Onions*. This also proves our craving for salty and fatty food. Top 20 ingredient also contain duplicate ingredient like garlic

and garlic clove, salt and kosher salt, eggs and large eggs which shows shortcoming of the dataset. Also ingredient like salt, oil and water could be avoided to get analysis of real ingredients as these are commonly use ingredient and doesn't contribute much to the study. Figure 6 shows top 20 ingredient across cuisines.

[Figure 6 about here.]

**4.5.3 Ingredients Distribution By Cuisines.** The third analysis is carried out to understand key ingredient for each cuisine. These key ingredients define those cuisines and provide unique test characterized by that cuisine. We limited ingredient list to top 10 to get the key ingredients for each cuisine. Study shows *Italian* cuisine is characterized by olive oil, garlic, cheese, black pepper, onion and butter, *Mexican* by onion, cumin, garlic, chili powder, jalapeno chilies, sour cream, tortillas and avocado, *Southern US* by butter, all-purpose flour, sugar, eggs, baking powder, milk and butter milk, *Indian* by onion, garam masala, turmeric, garlic, cumin and oil, *Chinese* by soy sauce, sesame oil, corn starch, sugar, garlic, green onions and scallions. Similarly it is applicable for all other cuisines present in the dataset and it is very close representation of all cuisines. Figure 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26 shows top 10 key ingredient used in the corresponding cuisines.

[Figure 7 about here.]

[Figure 8 about here.]

[Figure 9 about here.]

[Figure 10 about here.]

[Figure 11 about here.]

[Figure 12 about here.]

[Figure 13 about here.]

[Figure 14 about here.]

[Figure 15 about here.]

[Figure 16 about here.]

[Figure 17 about here.]

[Figure 18 about here.]

[Figure 19 about here.]

[Figure 20 about here.]

[Figure 21 about here.]

[Figure 22 about here.]

[Figure 23 about here.]

[Figure 24 about here.]

[Figure 25 about here.]

[Figure 26 about here.]

**4.5.4 Ingredients Relationship.** Forth analysis is carried out to understand the relationship between the ingredient to find out ingredient clusters. This analysis helps us understand the ingredient combinations which can be used together to provide great dish every time. This model can be used to predict ingredients for certain recipe based on the cluster. We used Gephi tool to analyze and produce the graph for this analysis. Gephi accepts network structure in terms of Node and Edge relationship. We created this network using python by relating all ingredients present in the recipe with each other. Ingredients become the node and source and target nodes become the edges. These network files generated

in excel spreadsheet and converted to CSV format and imported into the Gephi tool. Import created 5405 Nodes and 290828 edges for processing and analysis. Force Atlas 2 layout present in Gephi has been applied to the network which brings nodes with higher weights and shared connections closer to each other. We also used Gephi Data Laboratory to clean up duplicate or unwanted nodes. Filtering based on Degree Range and Edge Weight has been applied to data to reduce node and edges to get the graph which can be used for analysis and avoid crashing Gephi due to large data. Modularity statistic uncovered 5 ingredient clusters which can be identified by different colors in the graph. This cluster can approximately relate to the cuisines present in our dataset and confirms our earlier analysis of ingredient by cuisine.

- Orange - Mexican
- Brown - Indian
- Blue - Chinese
- Green - Italian
- Gray - Southern US

Figure 27 shows ingredient cluster of more than 1000 nodes. This graph is nice to look at but difficult to read due to lot many nodes and edges in the graph.

[Figure 27 about here.]

Figure 28 shows ingredient cluster of around 100 nodes. We generated this graph by reducing nodes and edges to make it more readable. This graph provides us with our top 5 cuisine clusters.

[Figure 28 about here.]

## 4.6 Shortcomings

Improper documentation of ingredient names in the dataset reduces the correctness of this analysis. In absence of proper ingredient name and duplication of ingredient name prevents getting exact ingredient weight into the analysis. A dataset with uniform ingredient name can help this analysis to achieve its best. If we don't find proper ingredient name then this analysis needs to include extensive data cleaning process which can be considered an improvement to this project.

Network file creation algorithm can be enhanced further by considering the number of recipes for the ingredient to provide additional weight to the relationship which can provide the stronger bond between the ingredients.

## 4.7 Limitations

This dataset can be analyzed to find out ingredient overlap between various cuisine and can provide insight into the influence of one cuisine on another which is not covered as part of this study. Usually, geographically neighboring cuisines are influenced by each other as they share common ingredients.

## 5 CONCLUSION

This project shows most used ingredient, ingredient distribution by cuisine and predictive ingredient relationship model as per the goal of the project. We also show various opportunities present with ingredient data analysis and role of big data analytics. We prove human craving for salty and fatty food as salt and oil are most used ingredient across cuisines as per the analysis. We understand now

based on our analysis key ingredient of any cuisine. Ingredient cluster shows why those ingredients are the base of certain cuisine and recipe of those ingredients always turn out delicious. We also crave for the good data so that we can provide more accurate analysis of the ingredients. Ingredient analysis has potential not only to help restaurant and food industry but it can help with our social responsibility of sustainability and understanding different cuisines and culture. As food industries interest grows in big data analytics, we will continue to see more evaluations of the ingredients.

## ACKNOWLEDGMENTS

The author would like to thank Dr. Gregor von Laszewski for his support and suggestions in this project. The author would also like to acknowledge Kaggle application for hosting ingredient dataset which is used in this project and various application users contributing in data analysis. We also acknowledge various online resources which helped understand Python and Gephi.

## REFERENCES

- [1] Bagrow James P Ahn Yong-Yeol, Ahnert Sebastian E. 2011. Flavor network and the principles of food pairing. (2011). <https://www.nature.com/articles/srep00196#supplementary-information>
- [2] businessdictionary. 2017. Food. web. (2017). <http://www.businessdictionary.com/definition/food.html>
- [3] Usashi Chatterjee, Vinit Kumar, and Devika P. Madalli. 2016. Formalizing Food Ingredients for Data Analysis and Knowledge Organization. *COLLNET Journal of Scientometrics and Information Management* 10 (07 2016), 289–309. [https://www.researchgate.net/publication/311337510\\_Formalizing\\_Food\\_Ingredients\\_for\\_Data\\_Analysis\\_and\\_Knowledge\\_Organization](https://www.researchgate.net/publication/311337510_Formalizing_Food_Ingredients_for_Data_Analysis_and_Knowledge_Organization)
- [4] Lada A. Adamic Chun-Yuen Teng, Yu-Ru Lin. 2011. Recipe recommendation using ingredient networks. web. (2011). <https://arxiv.org/pdf/1111.3919.pdf>
- [5] S. M. Church. 2015. The importance of food composition data in recipe analysis. web. (2015). <http://onlinelibrary.wiley.com/doi/10.1111/nbu.12125/abstract>
- [6] collinsdictionary. 2017. Recipe. web. (2017). <https://www.collinsdictionary.com/us/dictionary/english/recipe>
- [7] inkhorn82. 2014. A Delicious Analysis. web. (2014). <https://www.r-bloggers.com/a-delicious-analysis-aka-topic-modelling-using-recipes/>
- [8] kaggle. 2015. What's Cooking? web. (2015). <https://www.kaggle.com/c/whats-cooking/data>
- [9] Bernard Lahousse. 2016. Using Big Data to Transform Unfamiliar Ingredients Into Tasty Recipes. web. (2016). <https://foodtechconnect.com/2016/04/20/big-food-data-recipes-from-unfamiliar-ingredients/>
- [10] oxforddictionaries. 2017. Ingredient. web. (2017). <https://en.oxforddictionaries.com/definition/ingredient>
- [11] Matthew Robinson. 2015. Big Data Analytics and Food Come Together At Flavourspace. web. (2015). <http://www.theculinaryexchange.com/food-innovation/big-data-analytics-and-food-come-together-at-flavourspace/>

## LIST OF FIGURES

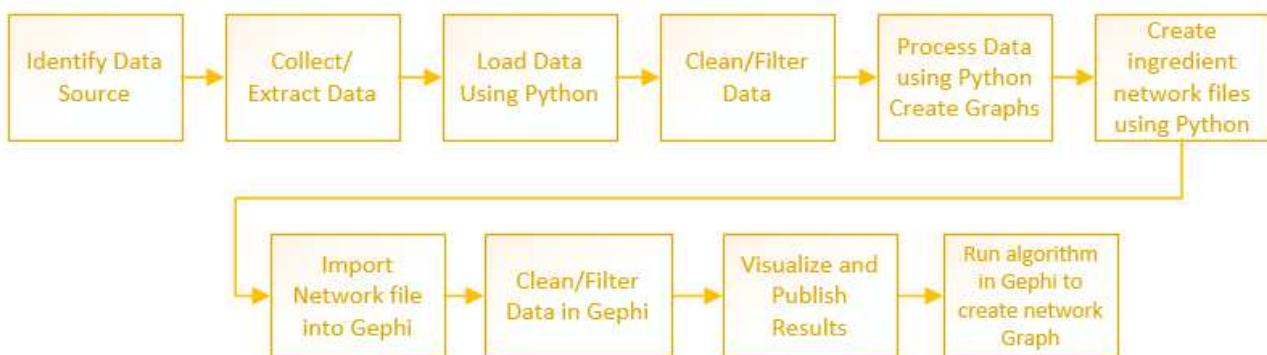
1	Code Structure	6
2	Flowchart of the Methodology to Analyze Ingredients	6
3	Ingredient Data Structure	6
4	Data Loading	7
5	Recipe Distribution By Cuisine	7
6	Top 20 Ingredients	8
7	Top 10 Ingredients	9
8	Top 10 Ingredients	10
9	Top 10 Ingredients	11
10	Top 10 Ingredients	12
11	Top 10 Ingredients	13
12	Top 10 Ingredients	14
13	Top 10 Ingredients	15
14	Top 10 Ingredients	16
15	Top 10 Ingredients	17
16	Top 10 Ingredients	18
17	Top 10 Ingredients	19
18	Top 10 Ingredients	20
19	Top 10 Ingredients	21
20	Top 10 Ingredients	22
21	Top 10 Ingredients	23
22	Top 10 Ingredients	24
23	Top 10 Ingredients	25
24	Top 10 Ingredients	26
25	Top 10 Ingredients	27
26	Top 10 Ingredients	28
27	Ingredient Cluster	29
28	ingredient Cluster 100 Nodes	30

```

code
- ingredientAnalysis.py
- ingredientAnalysis.py
- data
  - train.json
  - nodes.xlsx
  - edges.xlsx
- images
- gephi
  - geph Ing big data.gephi

```

**Figure 1: Code Structure**



**Figure 2: Flowchart of the Methodology to Analyze Ingredients**

```

{
  "id": 24717,
  "cuisine": "indian",
  "ingredients": [
    "tumeric",
    "vegetable stock",
    "tomatoes",
    "garam masala",
    "naan",
    "red lentils",
    "red chili peppers",
    "onions",
    "spinach",
    "sweet potatoes"
  ]
},

```

**Figure 3: Ingredient Data Structure**

```

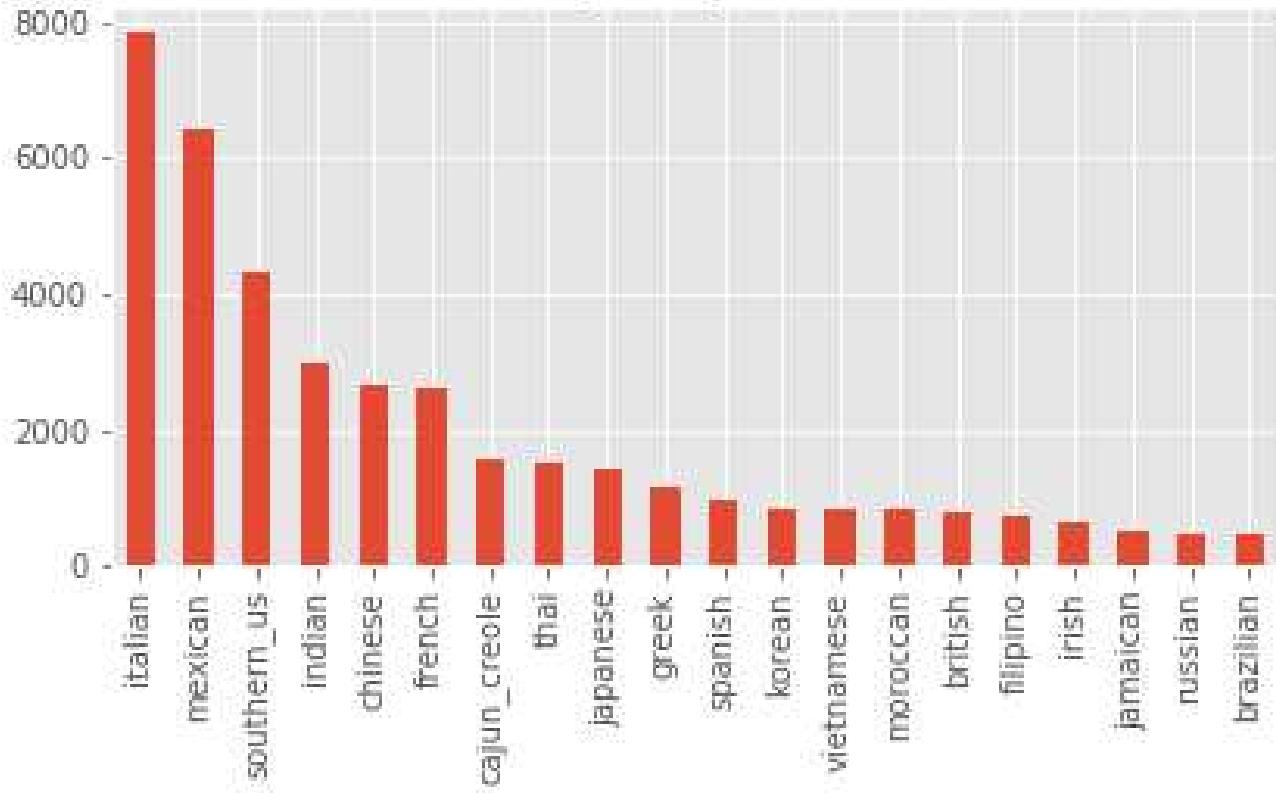
#read the ingredient data using pandas
dfTrain = pd.read_json('./data/train.json')

#load data using json
dataFilePath='./data/train.json'
with open(dataFilePath) as data_file:
    data = json.load(data_file)

```

**Figure 4: Data Loading**

## Recipies By Cuisine



**Figure 5: Recipe Distribution By Cuisine**

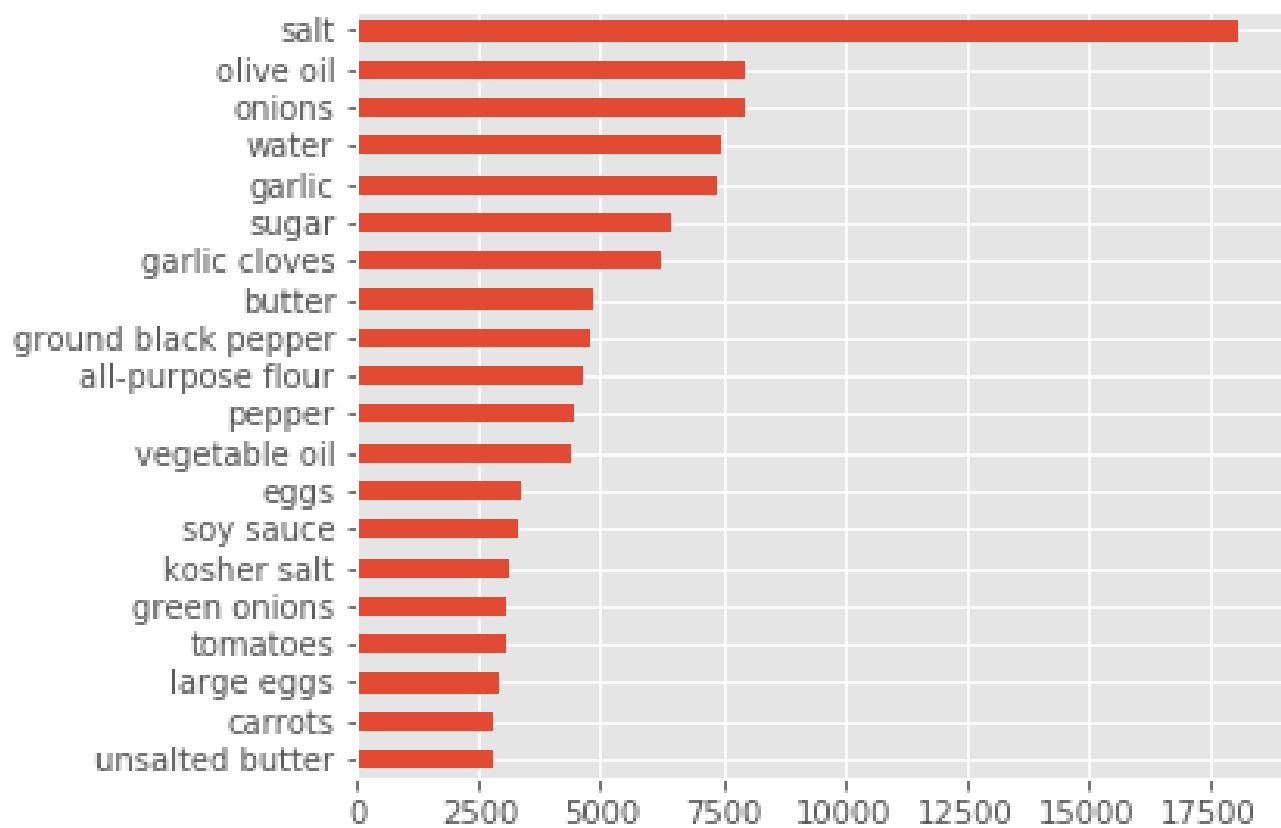


Figure 6: Top 20 Ingredients

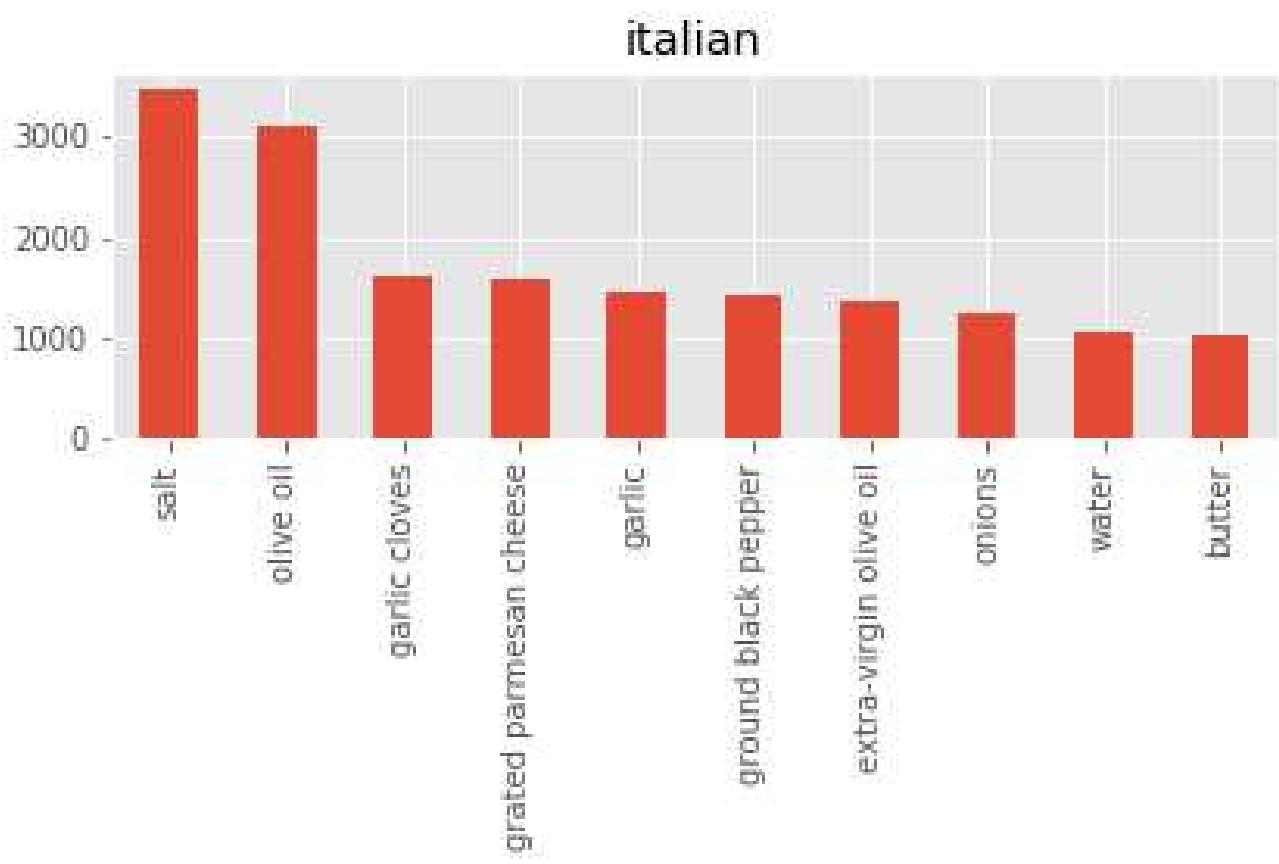


Figure 7: Top 10 Ingredients

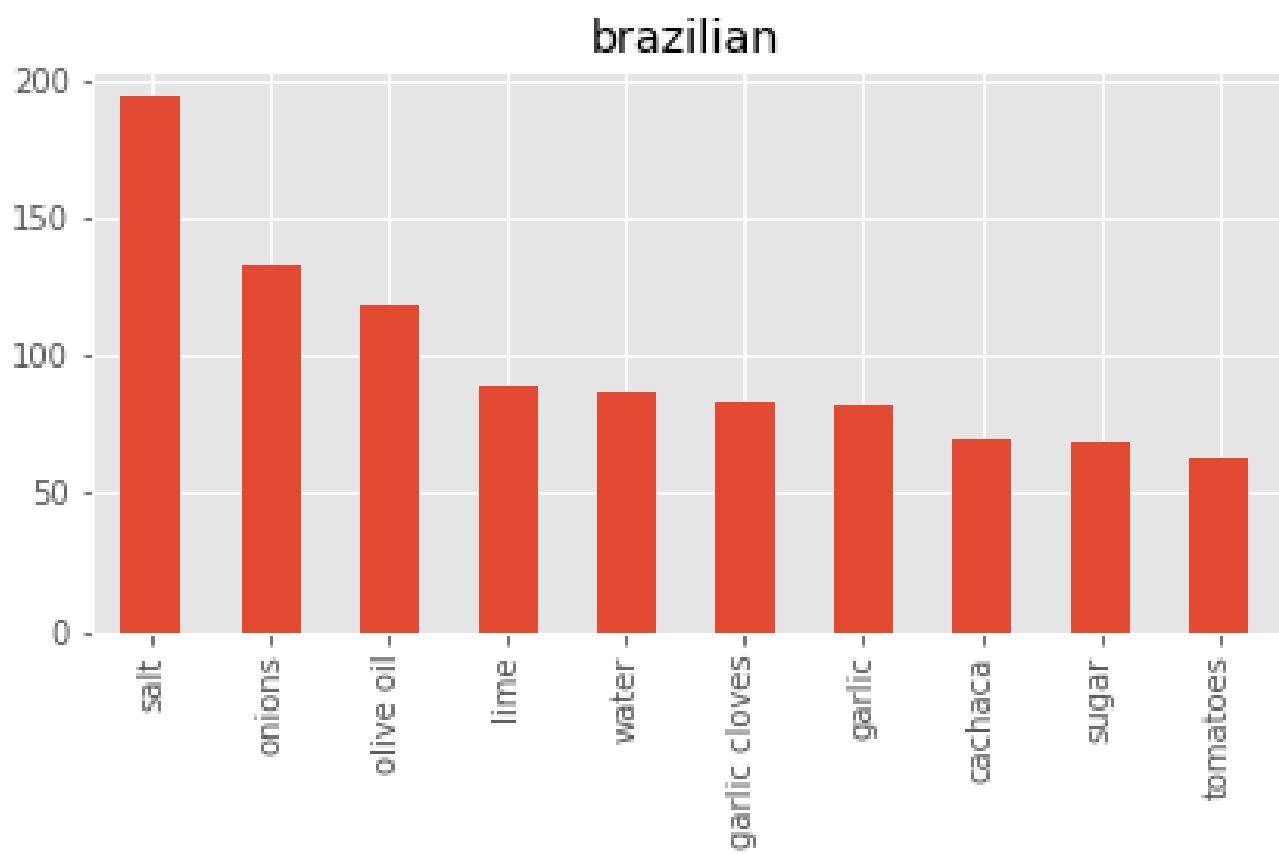
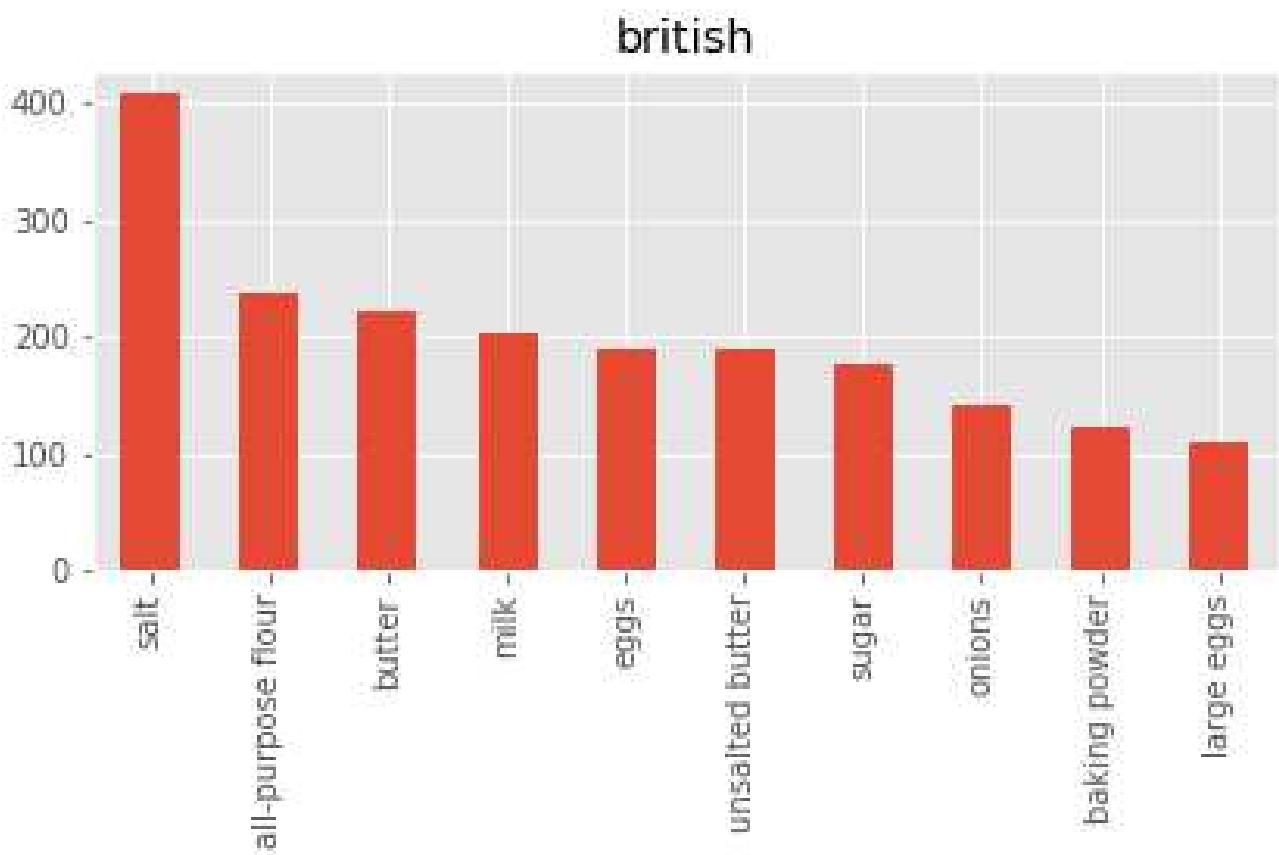


Figure 8: Top 10 Ingredients



**Figure 9: Top 10 Ingredients**

### cajun\_creole

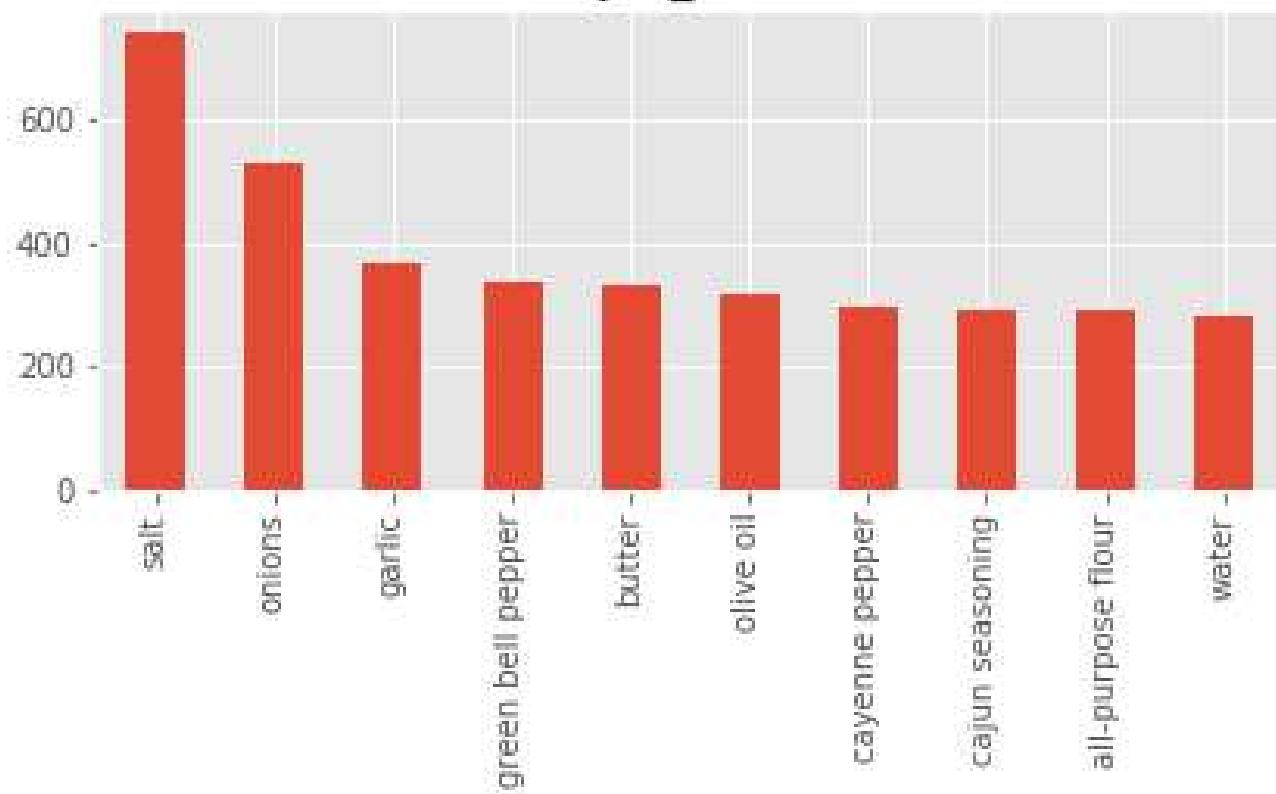


Figure 10: Top 10 Ingredients

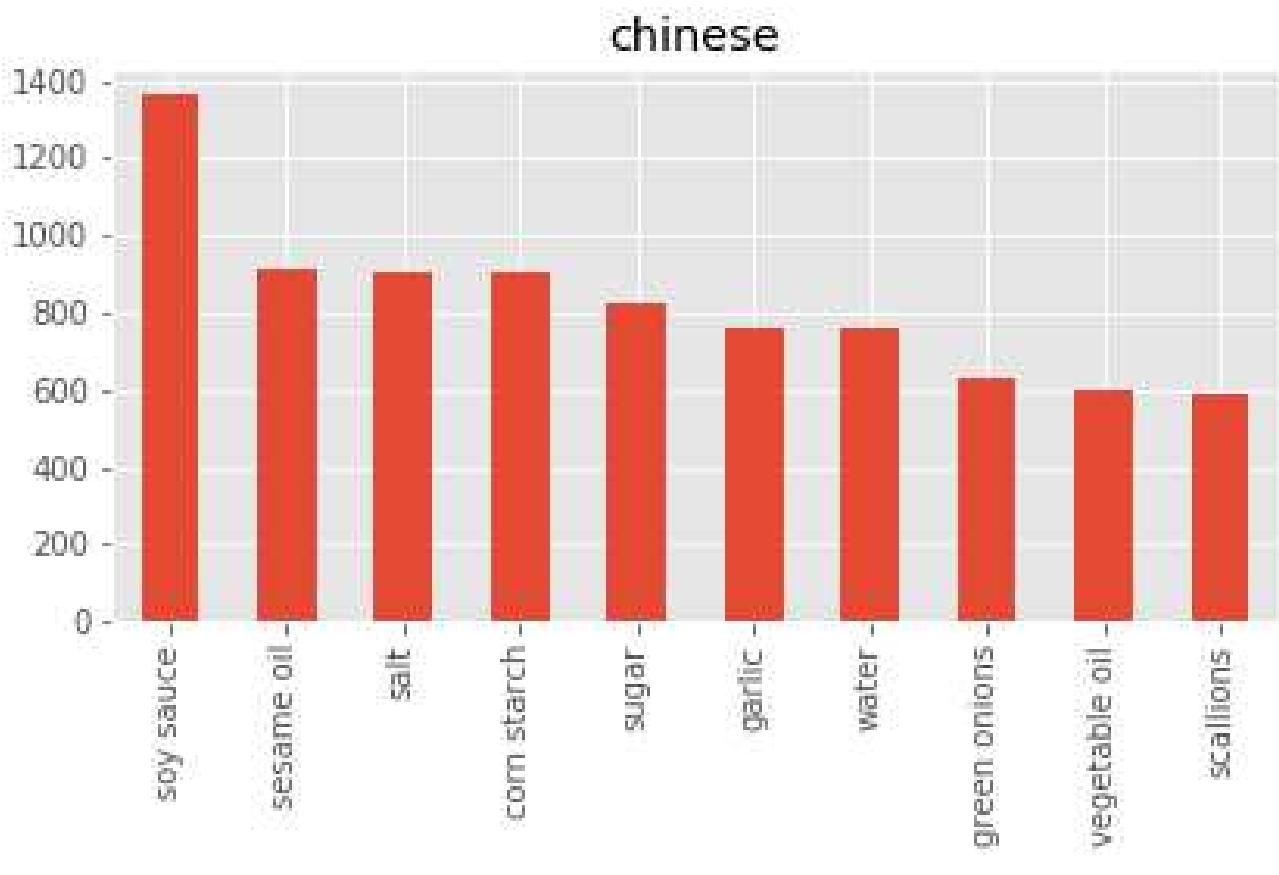


Figure 11: Top 10 Ingredients

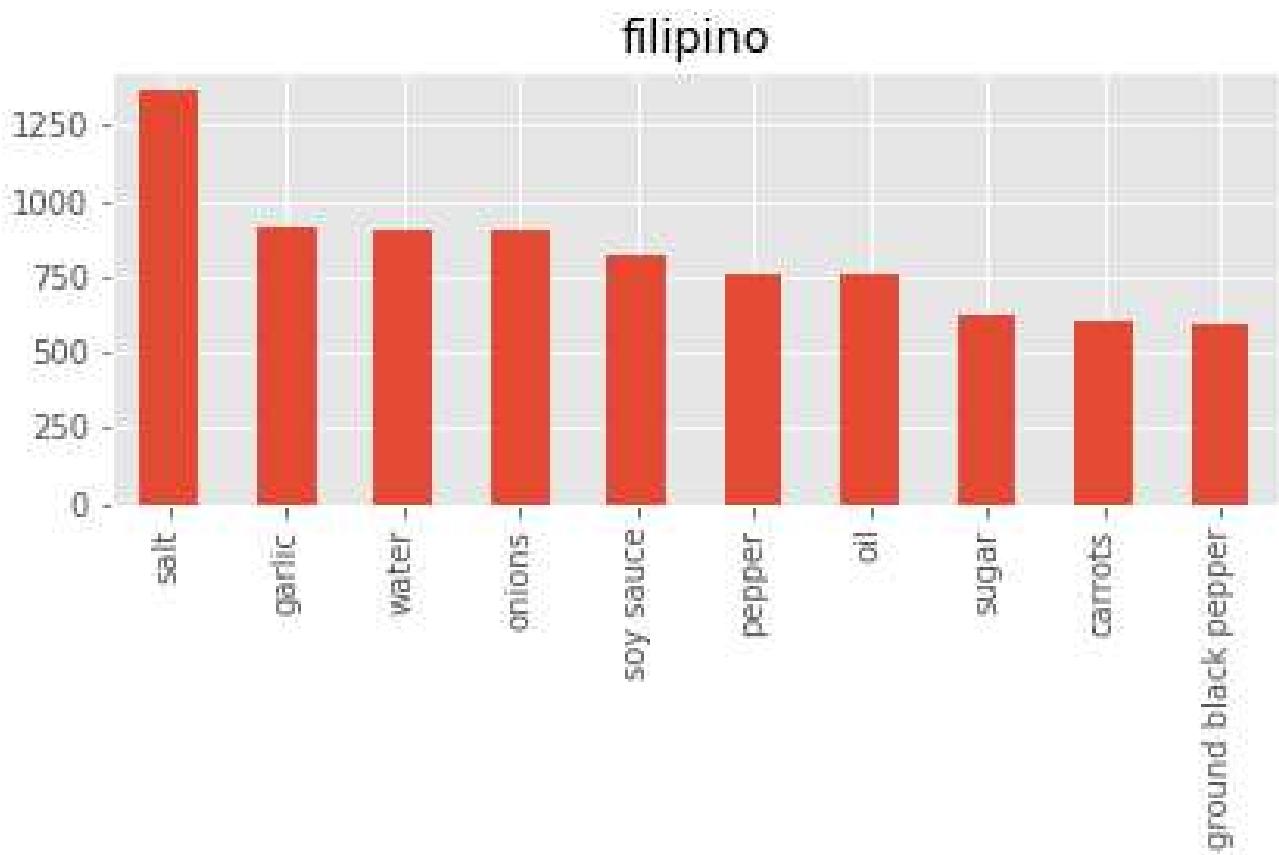


Figure 12: Top 10 Ingredients

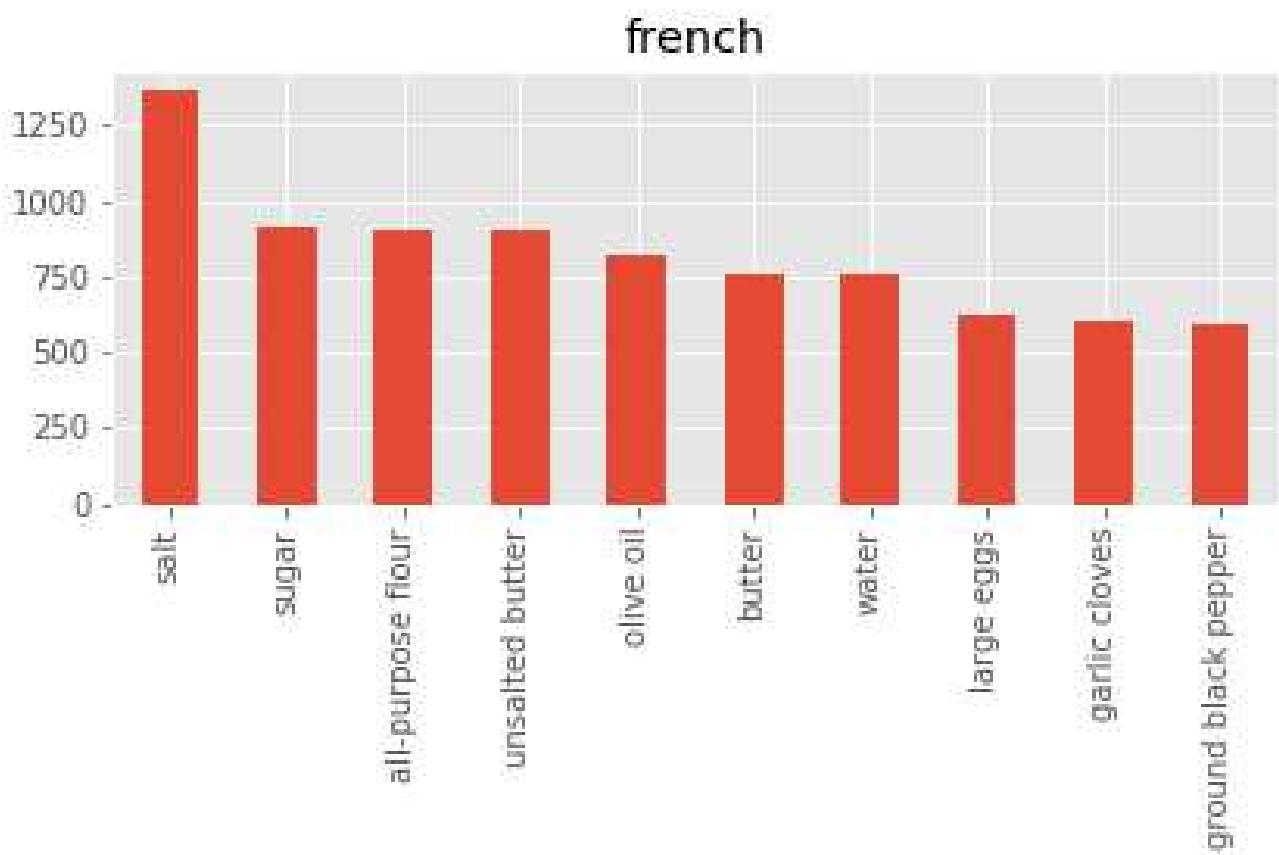


Figure 13: Top 10 Ingredients

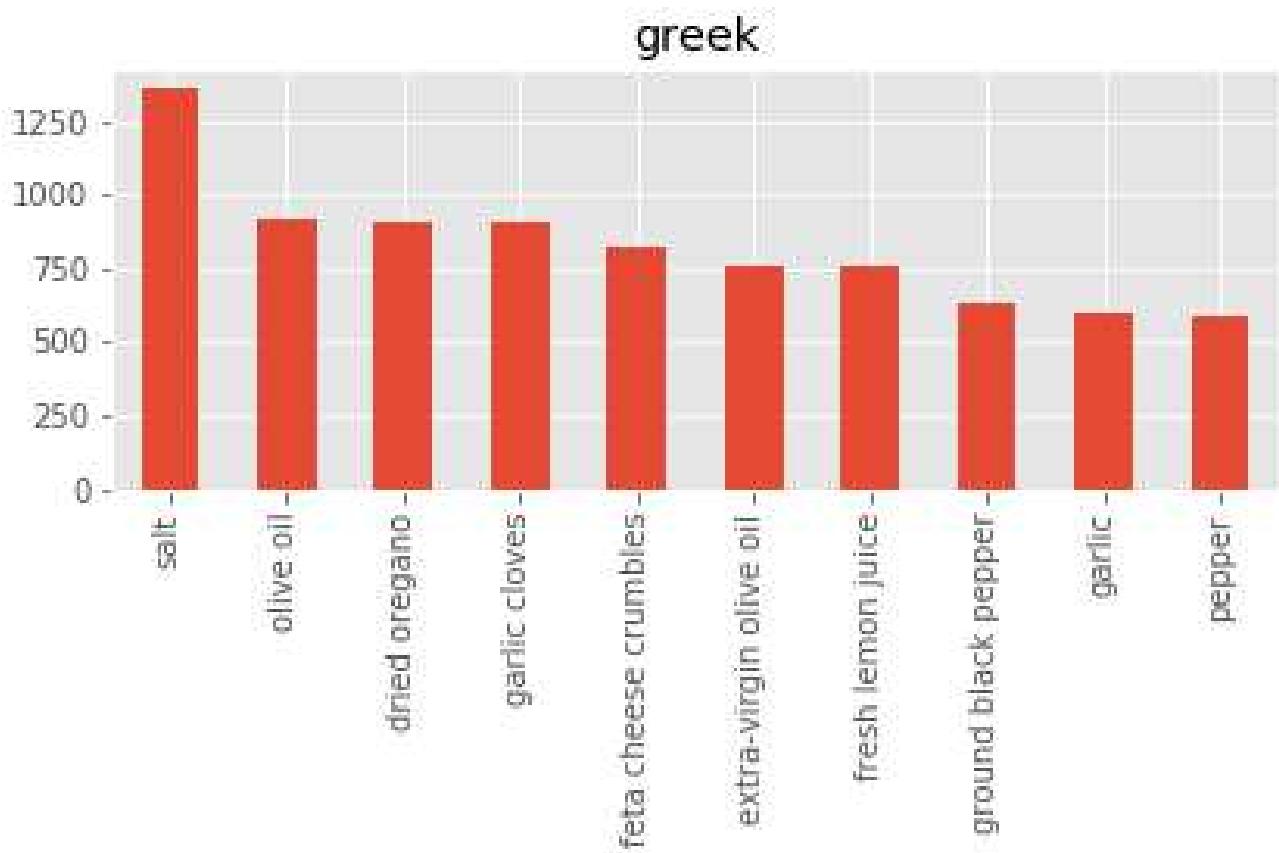
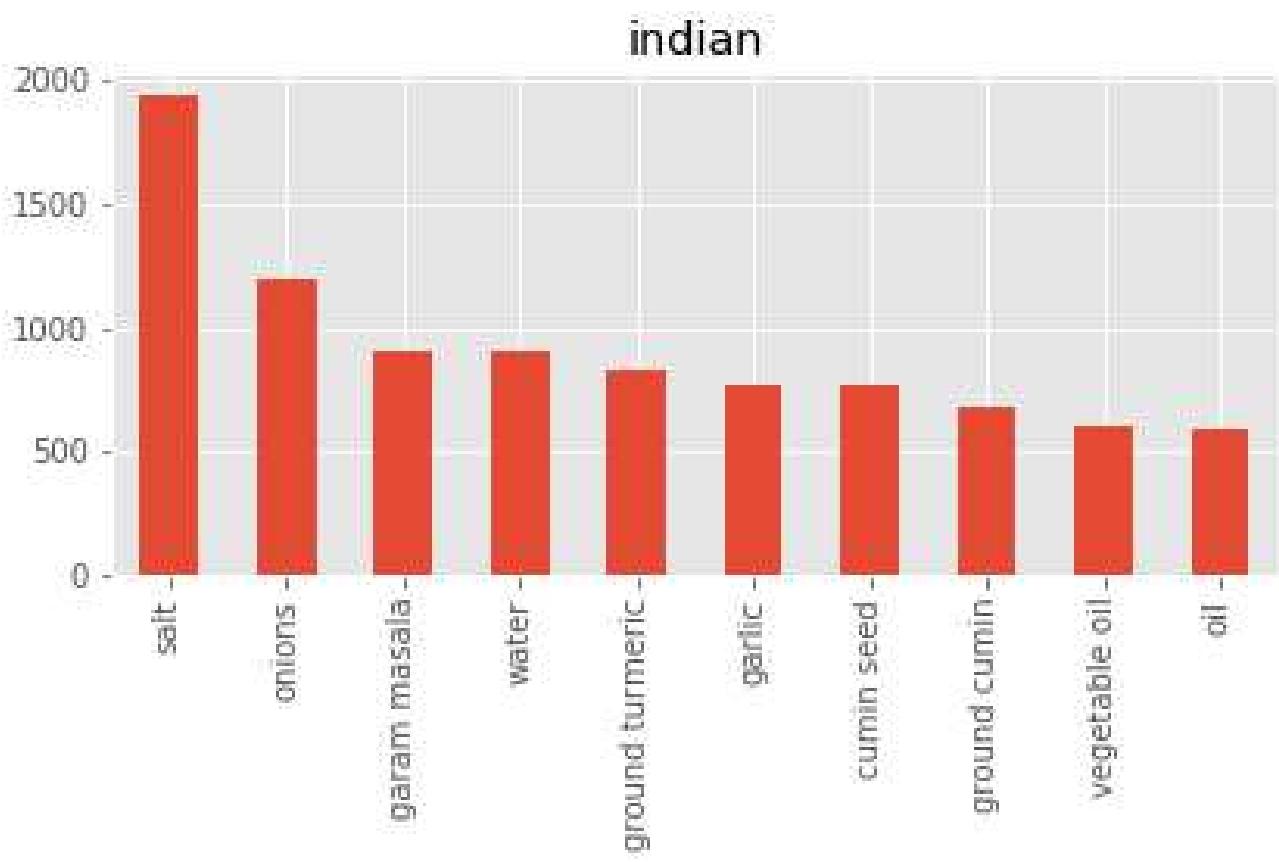


Figure 14: Top 10 Ingredients



**Figure 15: Top 10 Ingredients**

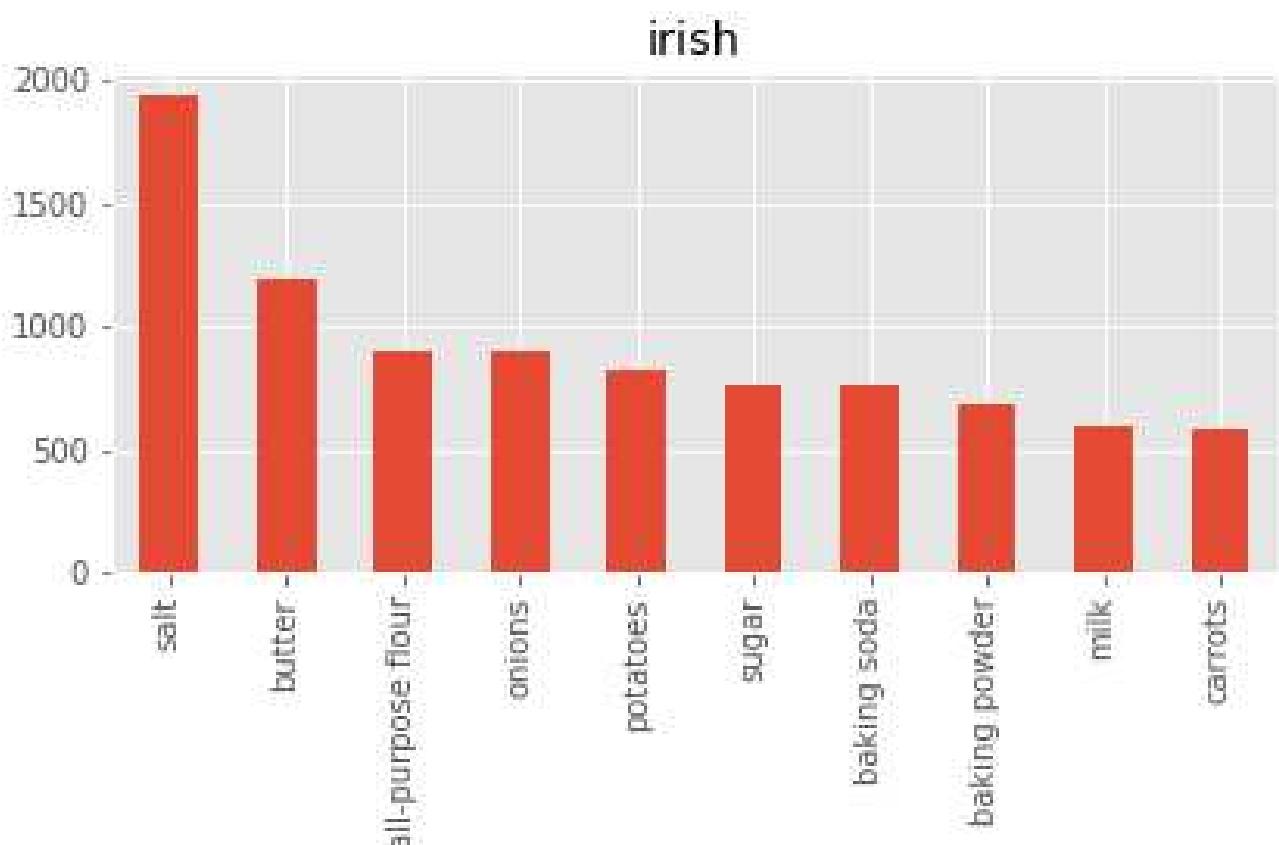


Figure 16: Top 10 Ingredients

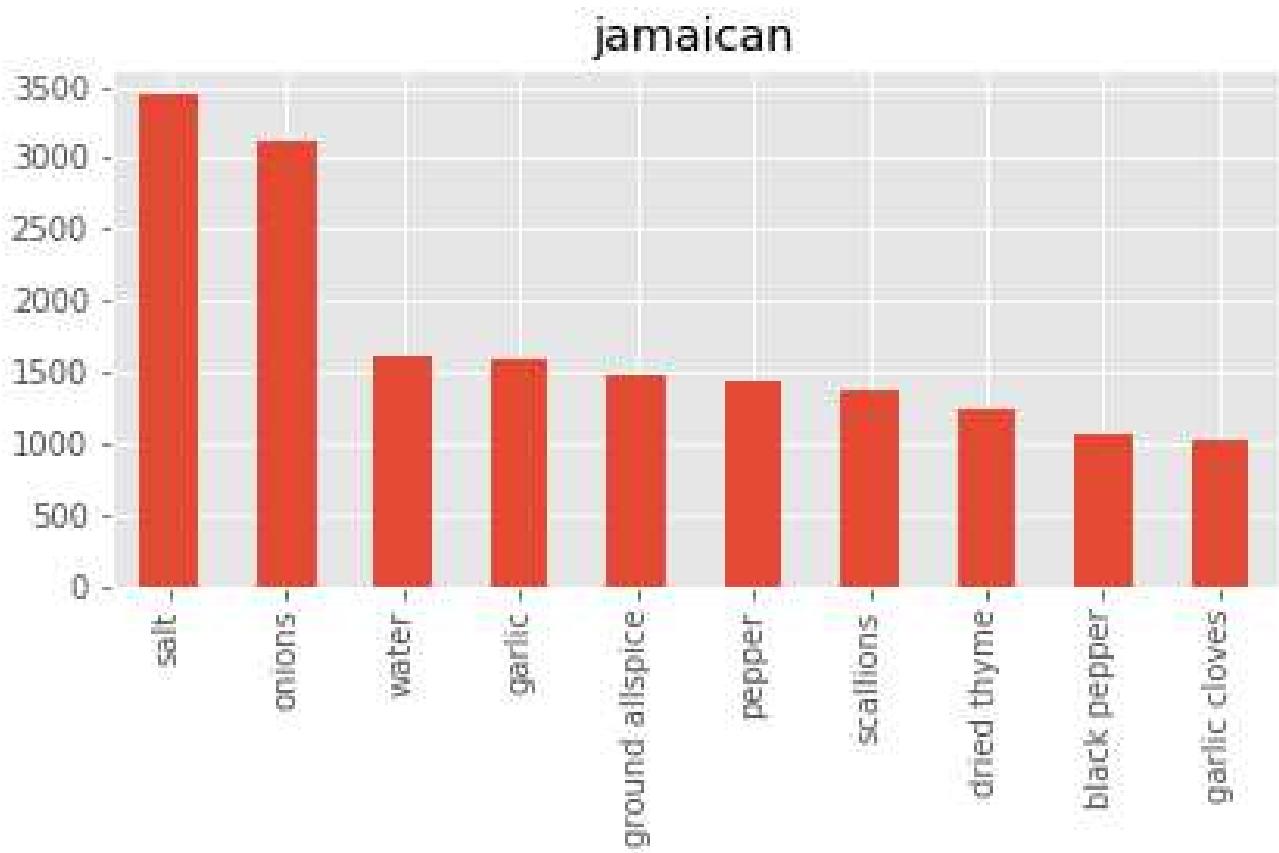


Figure 17: Top 10 Ingredients

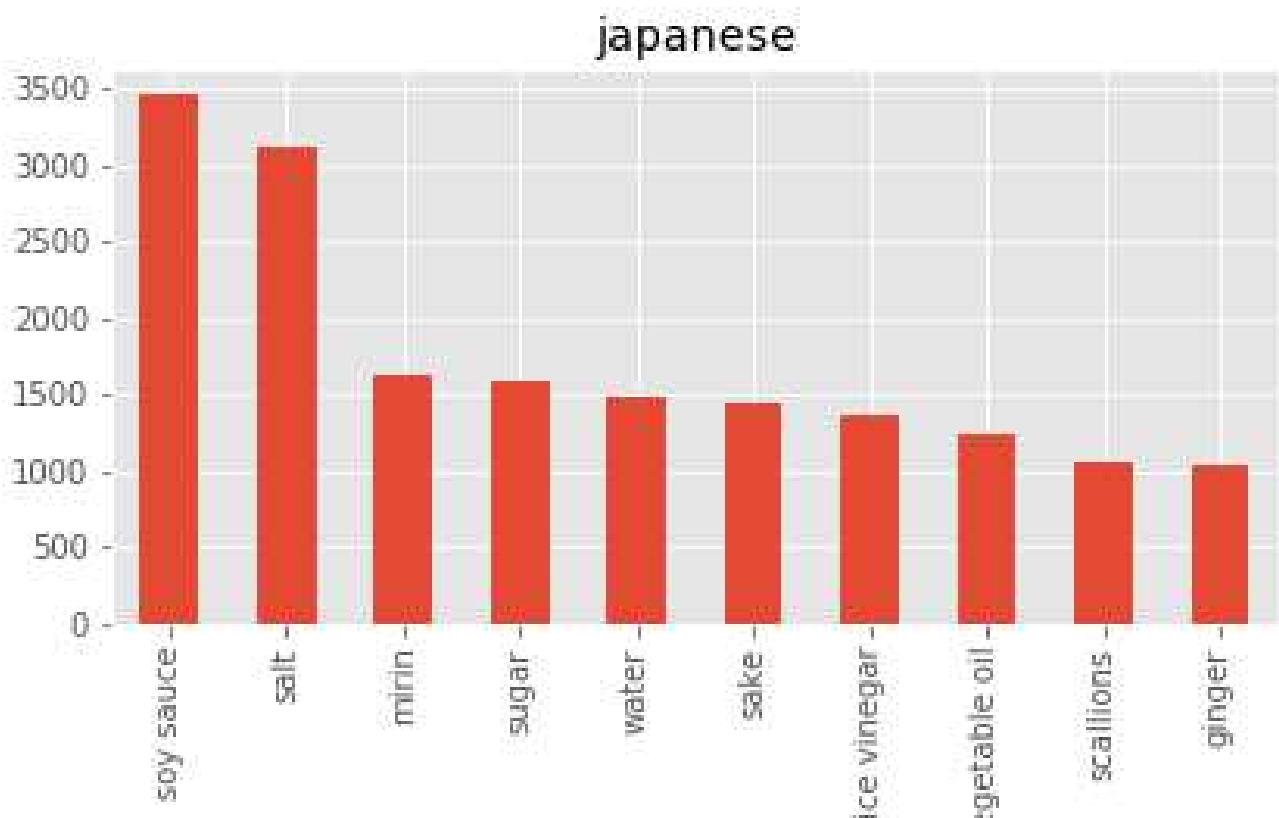


Figure 18: Top 10 Ingredients

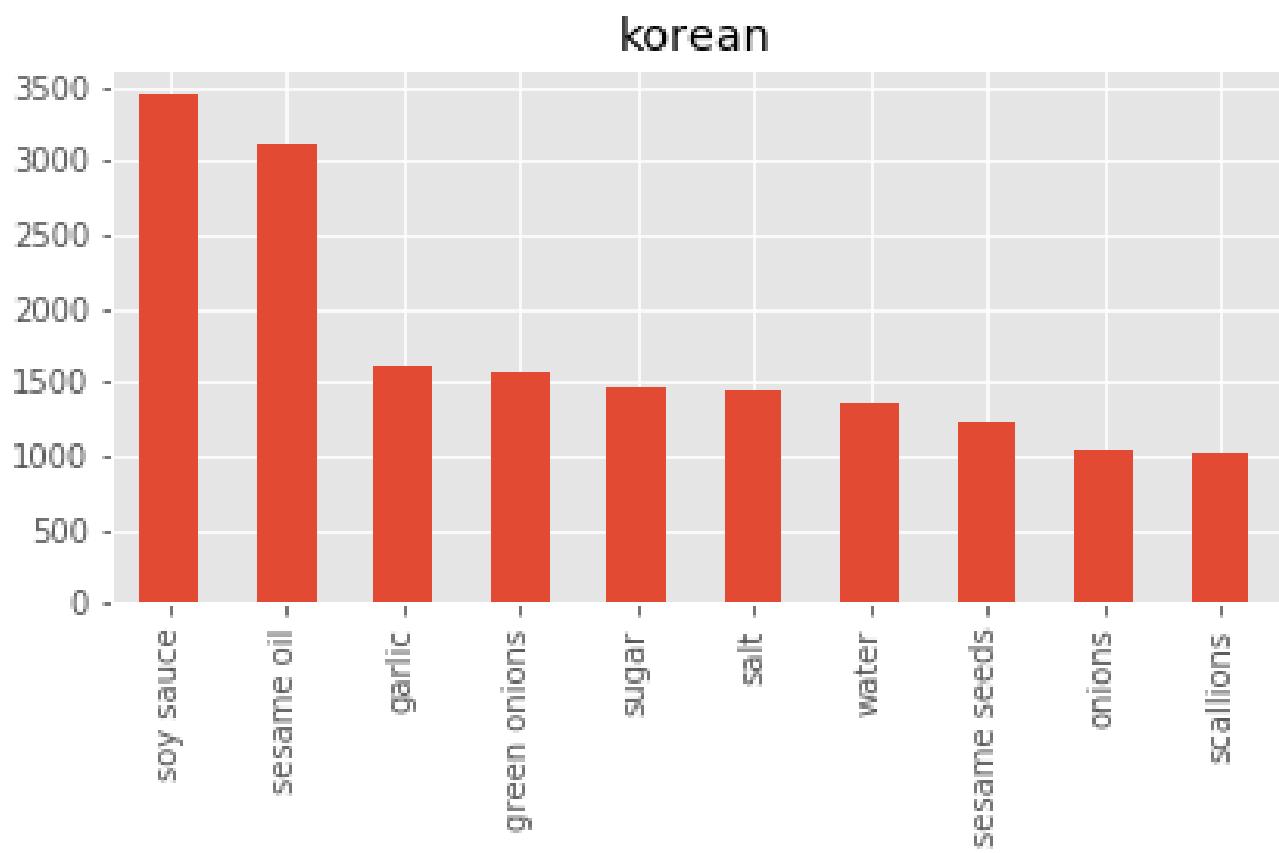


Figure 19: Top 10 Ingredients

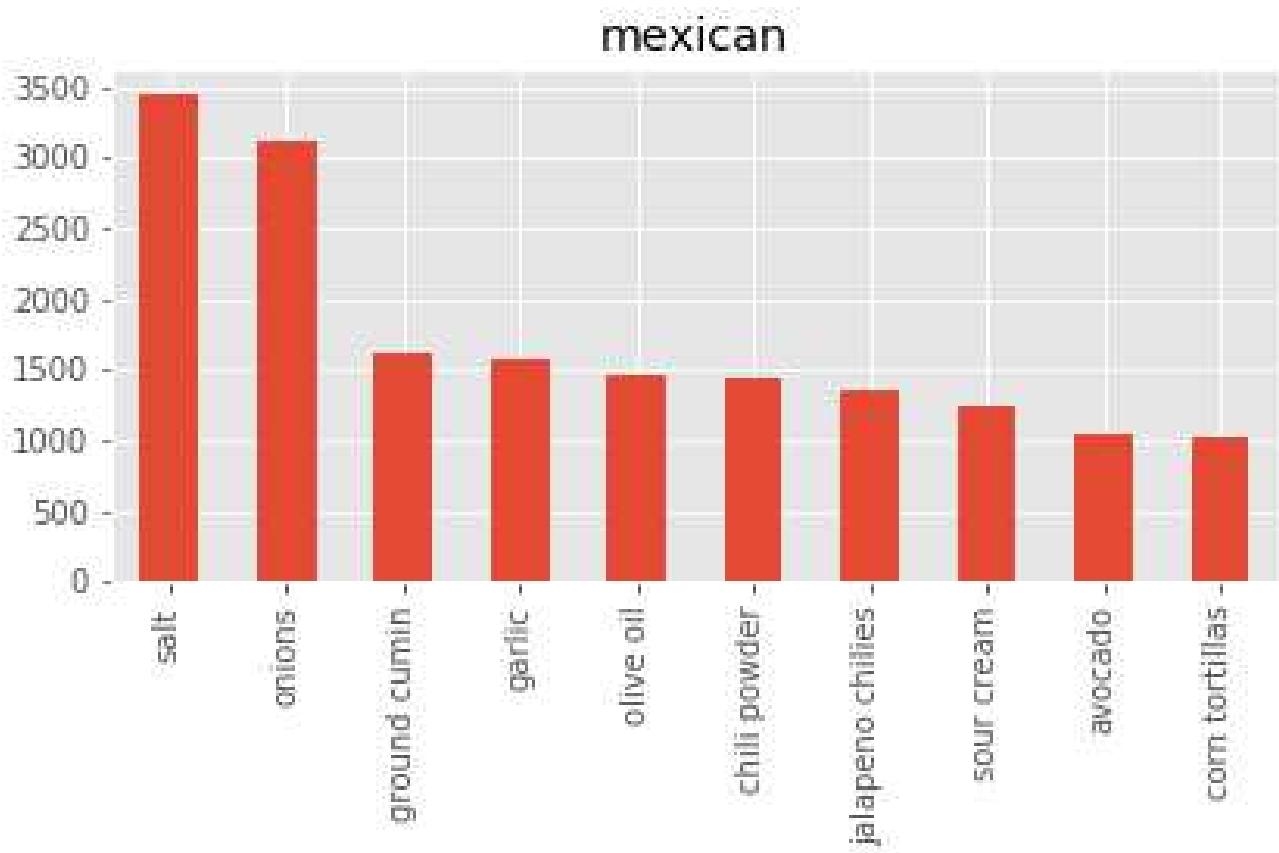


Figure 20: Top 10 Ingredients

## moroccan

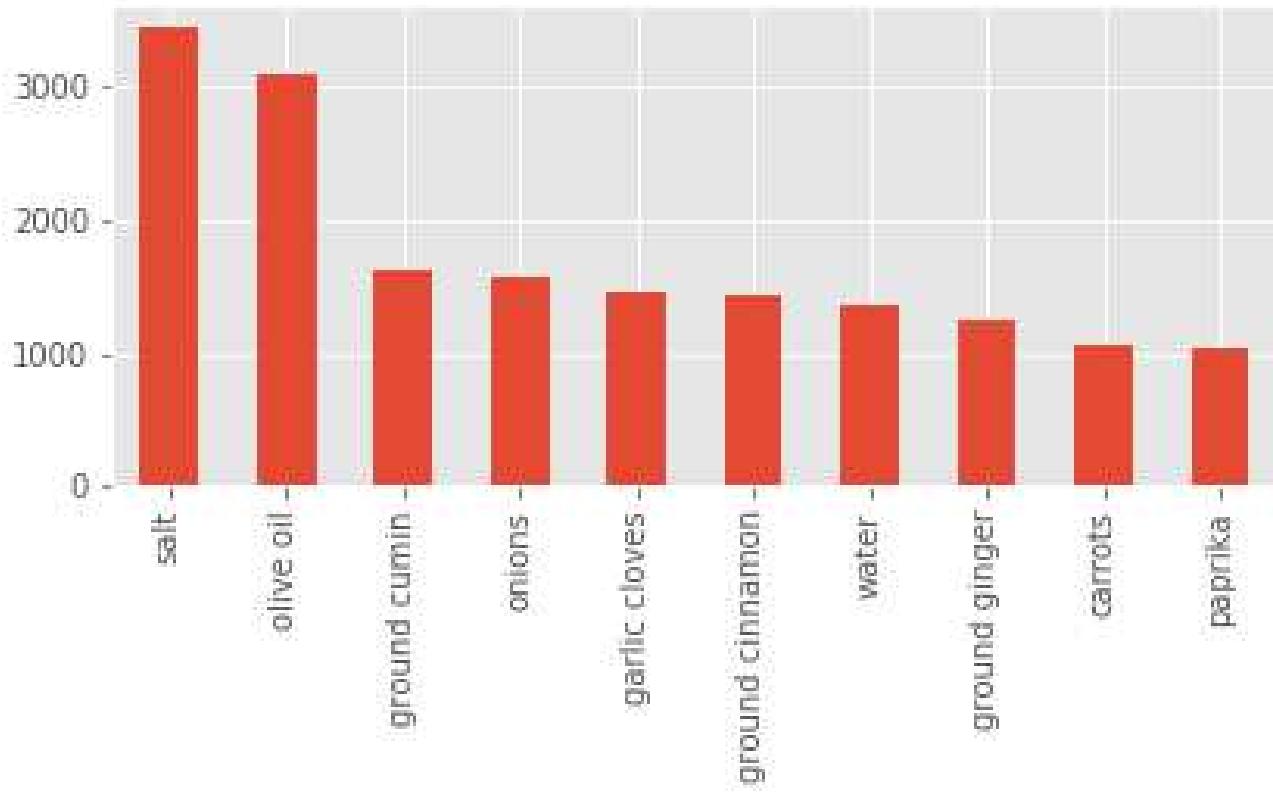


Figure 21: Top 10 Ingredients

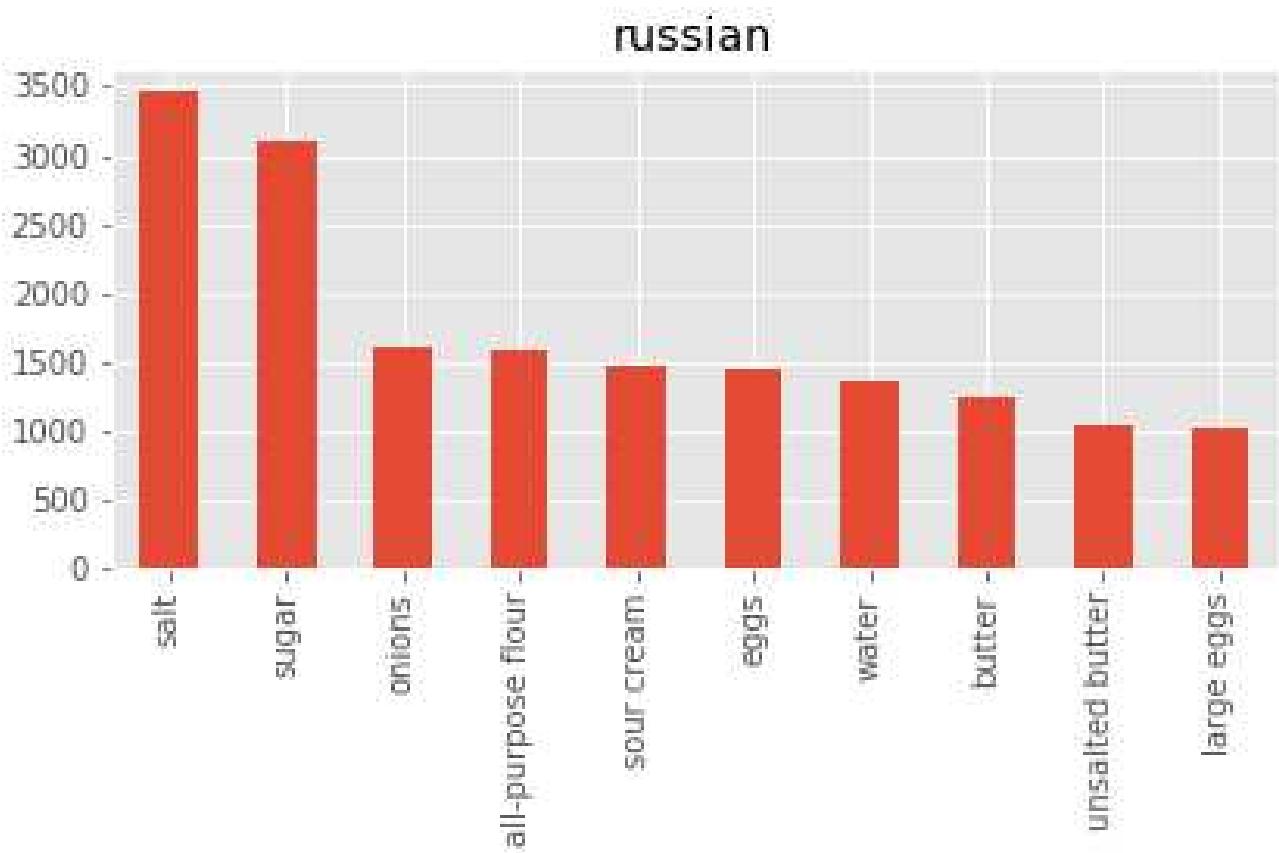


Figure 22: Top 10 Ingredients

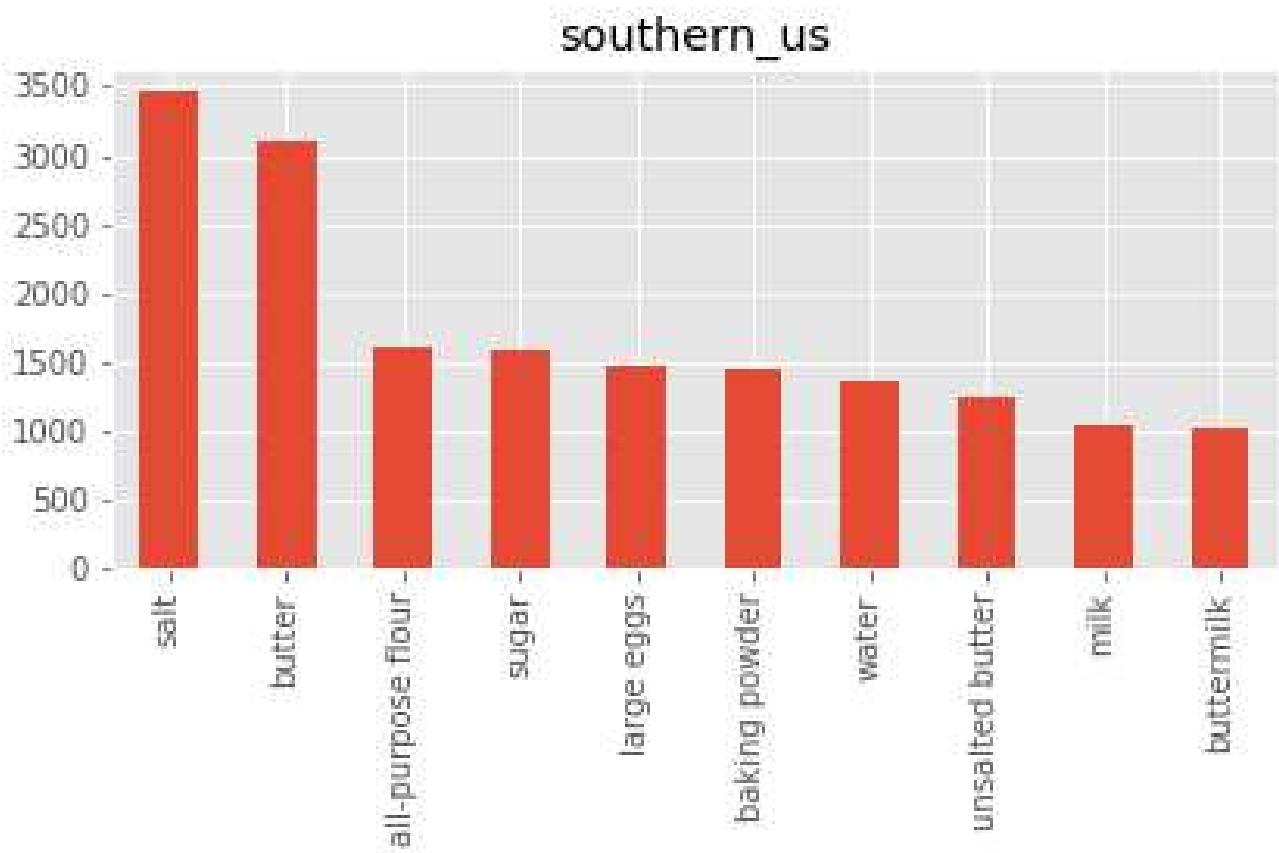


Figure 23: Top 10 Ingredients

spanish

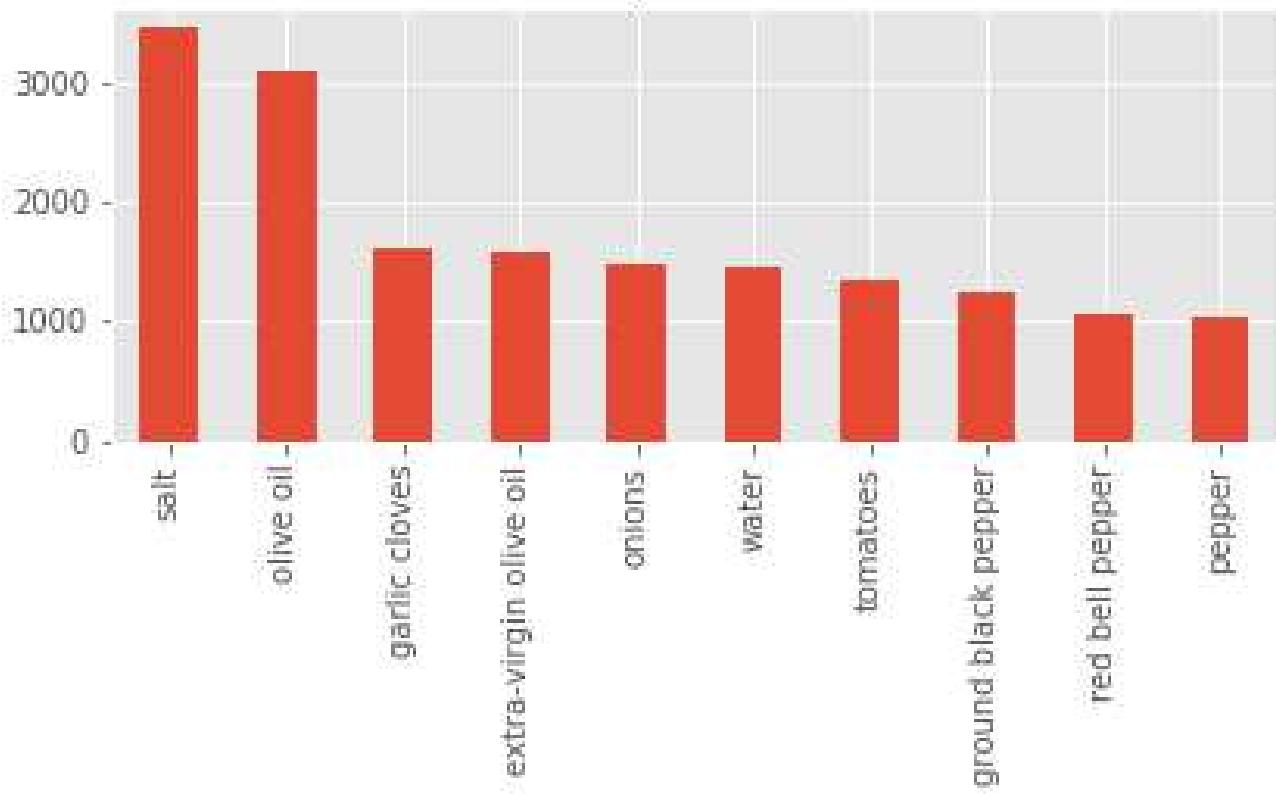


Figure 24: Top 10 Ingredients

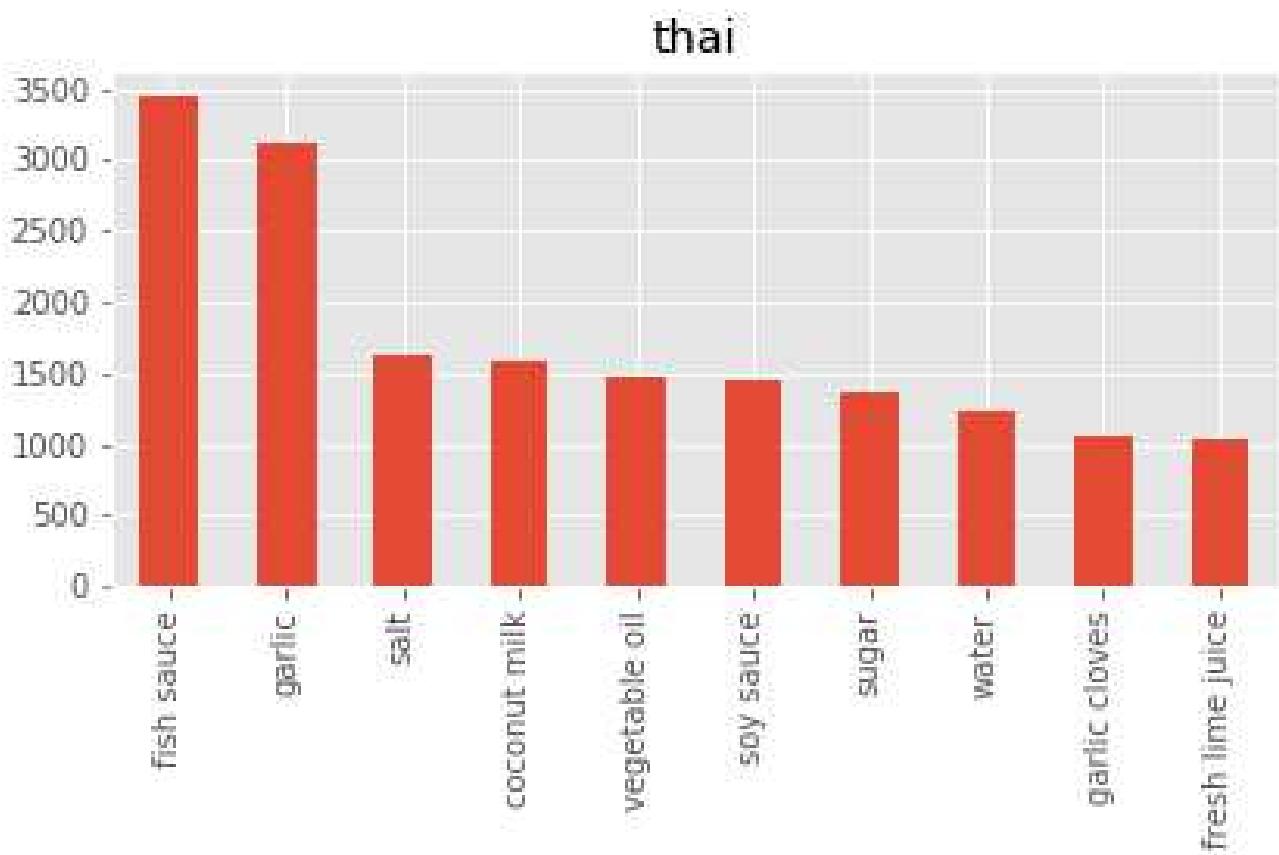


Figure 25: Top 10 Ingredients

## vietnamese

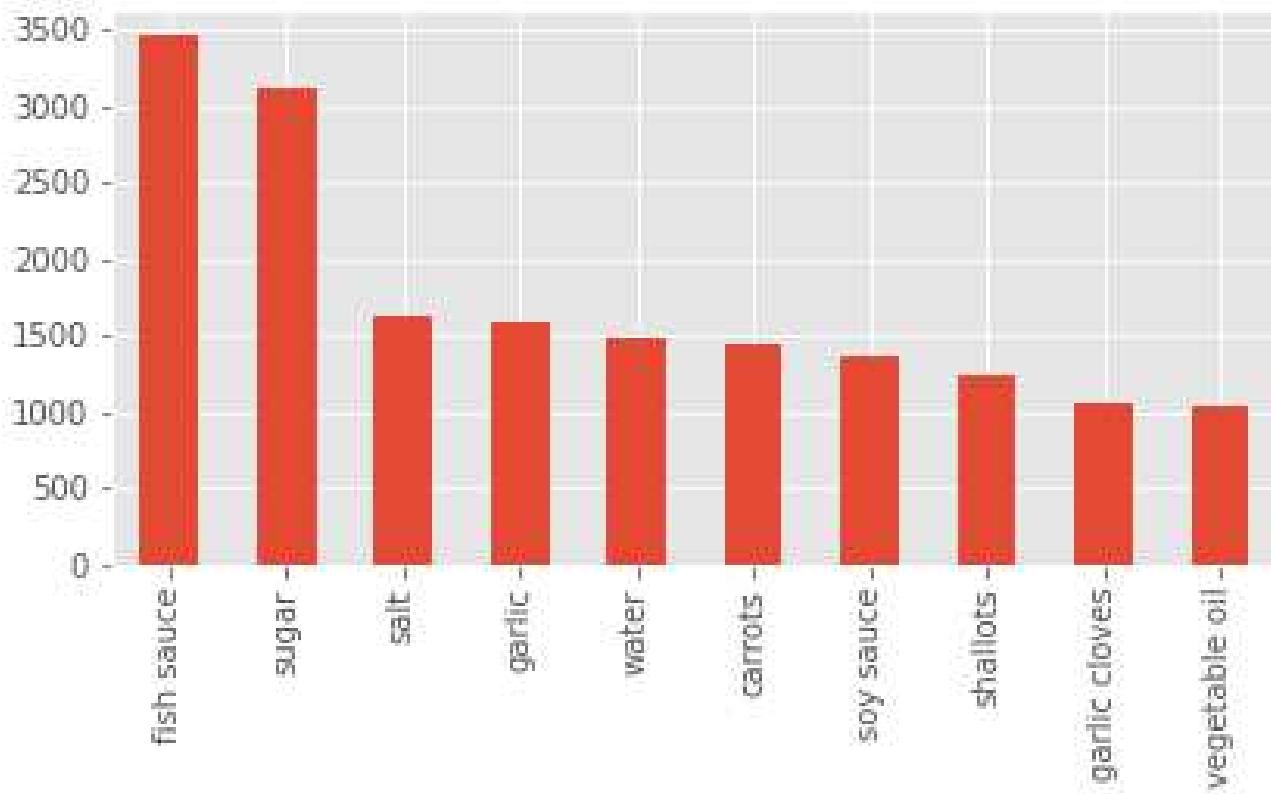
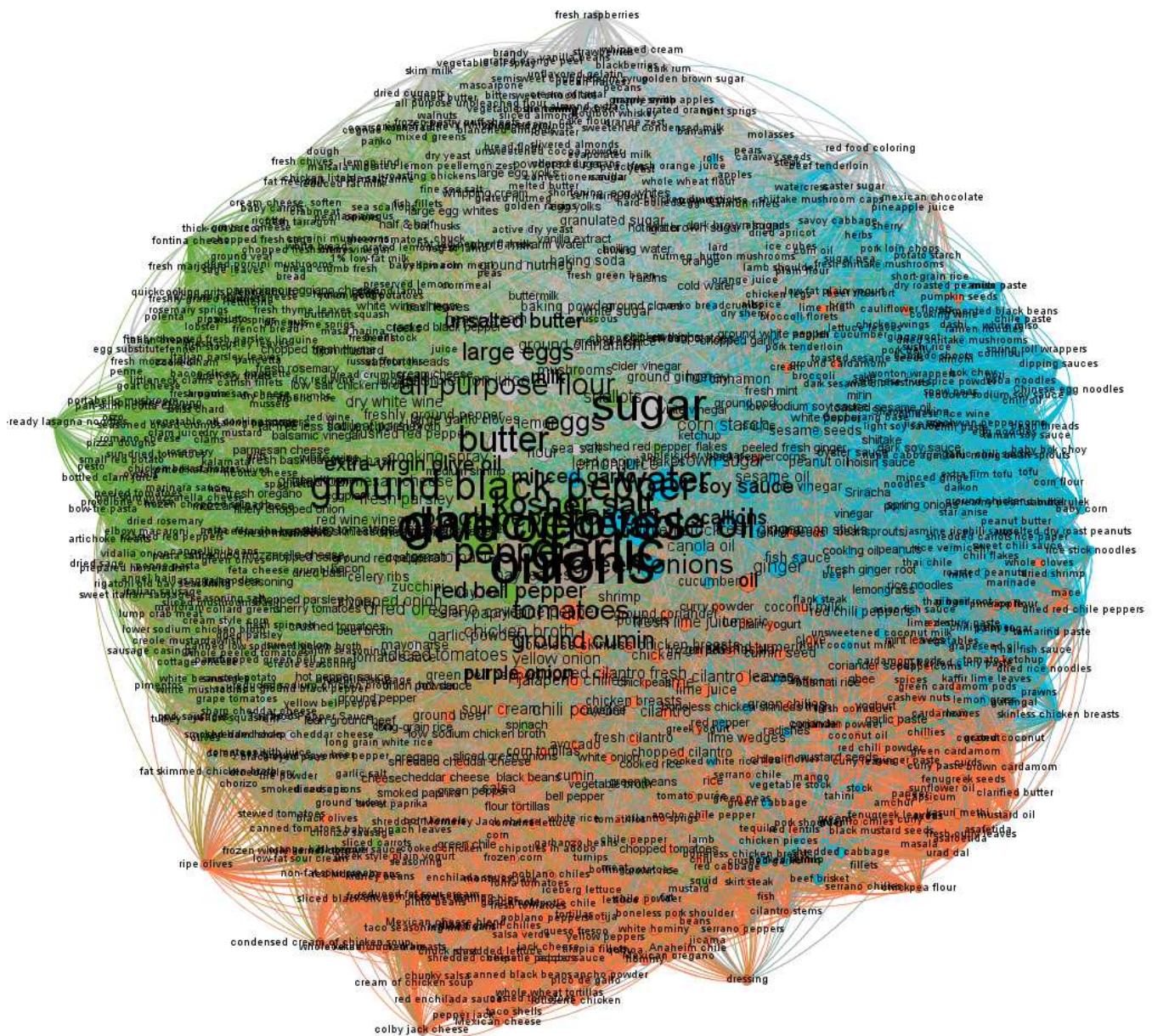
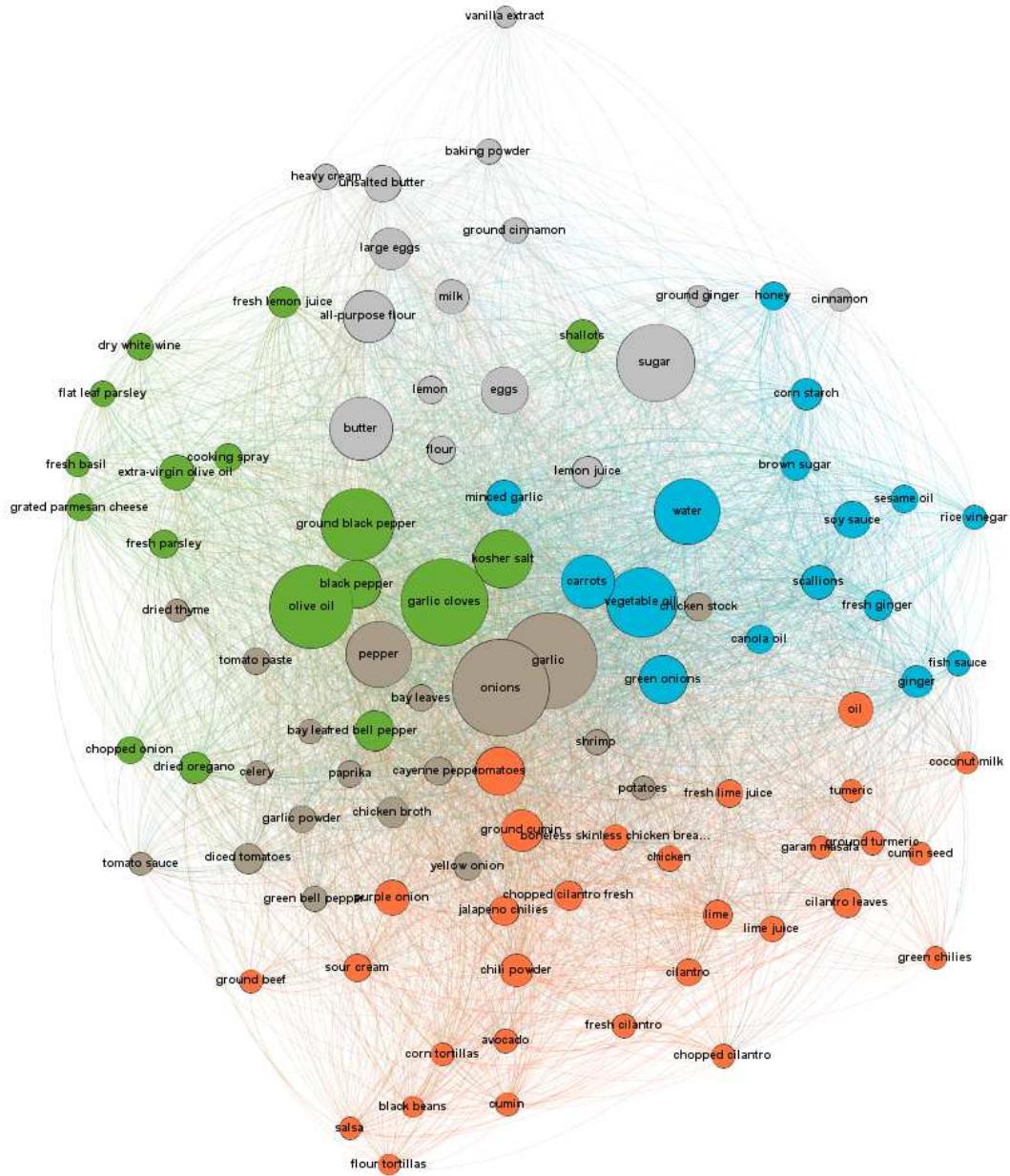


Figure 26: Top 10 Ingredients



**Figure 27: Ingredient Cluster**



**Figure 28: ingredient Cluster 100 Nodes**

LIST OF TABLES

1 Recipe Count By Cuisine

32

**Table 1: Recipe Count By Cuisine**

Cuisine	Recipe Count
brazilian	467
british	804
cajun creole	1546
chinese	2673
filipino	755
french	2646
greek	1175
indian	3003
irish	667
italian	7838
jamaican	526
japanese	1423
korean	830
mexican	6438
moroccan	821
russian	489
southern us	4320
spanish	989
thai	1539
vietnamese	825

```
bibtext report
```

```
=====
```

```
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
```

```
bibtext _ label error
```

```
=====
```

```
bibtext space label error
```

```
=====
```

```
bibtext comma label error
```

```
=====
```

```
latex report
```

```
=====
```

```
[2017-12-11 13.28.41] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflatex)
bookmark level for unknown defaults to 0.
The anchor of a bookmark and its parent's must not be the same. Added a new anchor.
The anchor of a bookmark and its parent's must not be the same. Added a new anchor.
Typesetting of "report.tex" completed in 2.5s.
```

```
=====
```

```
Compliance Report
```

```
=====
```

```
name: Sushant Athaley
hid: 302
paper1: Nov 3 2017 100%
paper2: 100%
project: 100%
```

```
yamlcheck
```

```
-----
```

```
wordcount
```

---

```
32
```

```
wc 302 project 32 3612 report.tex  
wc 302 project 32 3815 report.pdf  
wc 302 project 32 358 report.bib
```

```
find "
```

---

```
109: "id": 24717,
```

```
110: "cuisine": "indian",
```

```
111: "ingredients": [
```

```
112: "tumeric",
```

```
113: "vegetable stock",
```

```
114: "tomatoes",
```

```
115: "garam masala",
```

```
116: "naan",
```

```
117: "red lentils",
```

```
118: "red chili peppers",
```

```
119: "onions",
```

```
120: "spinach",
```

```
121: "sweet potatoes"
```

```
135: dataFilePath=". ./data/train.json"
```

```
passed: False
```

```
find footnote
```

---

```
passed: True
```

```
find input{format/i523}
-----
passed: False

find input{format/final}
-----
passed: False

floats
-----
68: Code is organized as described in Figure \ref{c:code-structure}
69: \begin{figure}[htb]
82: \caption{Code Structure}\label{c:code-structure}
95: Figure \ref{f:methodology} shows methodology used for this project
   to analyze ingredient data.
96: \begin{figure}![ht]
97: \centering\includegraphics[width=\columnwidth]{images/methodology.
   PNG}
98: \caption{Flowchart of the Methodology to Analyze Ingredients
   }\label{f:methodology}
105: The dataset for this study is sourced from Kaggle application
   \cite{www-kaggle}. This dataset is publicly available and
   featured in \emph{What's Cooking?} competition. This dataset is
   in JSON format and of 12MB size. This dataset contains recipe id,
   cuisine and list of ingredients as described in Figure
   \ref{c:data-structure}.
106: \begin{figure}[htb]
125: \caption{Ingredient Data Structure}\label{c:data-structure}
127: This dataset contains total 39774 recipes across various
   cuisines. We used two different methods to load this data.
   Cuisine and ingredient analysis is done by loading data into
   \emph{pandas dataframe} and to analyze ingredient relationship
   data has been loaded into \emph{json} object. Figure \ref{c:loading}
   shows the code for data loading used in this project.
128: \begin{figure}[htb]
139: \caption{Data Loading}\label{c:loading}
149: We first analyze entire dataset to understand the total number of
   recipes and their distribution across various cuisines. We use
   Pythons Panda library to get the total recipe count as 39774 and
   plot the distribution. Figure
   \ref{f:Number_of_recipes_by_cuisine} shows number of recipes per
   cuisine. Dataset is heavily dominated by Italian cuisine followed
   by Mexican cuisine and with very fewer recipes from Russian and
```

Brazilian cuisines. This also highlights another shortcoming of the dataset that it doesn't have equal representation of all cuisines which might give us biased analysis.

```
150: \begin{figure} [!ht]
151: \centering\includegraphics[width=\columnwidth]{images/Number_of_recipes_by_cuisine.png}
152: \caption{Recipe Distribution By Cuisine
153: }\label{f:Number_of_recipes_by_cuisine}
155: Table\ref{t:recipecount} describes recipe count for every
156: cuisine.
156: \begin{table}[htb]
159: \label{t:recipecount}
188: The second analysis is carried out to understand top 20
189: ingredients getting used across cuisine or globally. Ingredient
190: \emph{Salt} is obvious topper followed by \emph{Oil} and
191: \emph{Onions}. This also proves our craving for salty and fatty
192: food. Top 20 ingredient also contain duplicate ingredient like
193: garlic and garlic clove, salt and kosher salt, eggs and large
194: eggs which shows shortcoming of the dataset. Also ingredient like
195: salt, oil and water could be avoided to get analysis of real
196: ingredients as these are commonly use ingredient and doesn't
197: contribute much to the study. Figure
198: \ref{f:Ingredient_Distribution} shows top 20 ingredient across
199: cuisines.
189: \begin{figure} [!ht]
190: \centering\includegraphics[width=\columnwidth]{images/Ingredient_Distribution.png}
191: \caption{Top 20 Ingredients }\label{f:Ingredient_Distribution}
195: The third analysis is carried out to understand key ingredient
for each cuisine. These key ingredients define those cuisines and
provide unique test characterized by that cuisine. We limited
ingredient list to top 10 to get the key ingredients for each
cuisine. Study shows \emph{Italian} cuisine is characterized by
olive oil, garlic, cheese, black pepper, onion and butter,
\emph{Mexican} by onion, cumin, garlic, chili powder, jalapeno
chilies, sour cream, tortillas and avocado, \emph{Southern US} by
butter, all-purpose flour, sugar, eggs, baking powder, milk and
butter milk, \emph{Indian} by onion, garam masala, turmeric,
garlic, cumin and oil, \emph{Chinese} by soy sauce, sesame oil,
corn starch, sugar, garlic, green onions and scallions. Similarly
it is applicable for all other cuisines present in the dataset
and it is very close representation of all cuisines. Figure
\ref{f:italian_10_most_used_ingredients},
\ref{f:brazilian_10_most_used_ingredients},
\ref{f:british_10_most_used_ingredients},
\ref{f:cajun_creole_10_most_used_ingredients},
```

```

\ref{f:chinese_10_most_used_ingredients},
\ref{f:filipino_10_most_used_ingredients},
\ref{f:french_10_most_used_ingredients},
\ref{f:greek_10_most_used_ingredients},
\ref{f:indian_10_most_used_ingredients},
\ref{f:irish_10_most_used_ingredients},
\ref{f:jamaican_10_most_used_ingredients},
\ref{f:japanese_10_most_used_ingredients},
\ref{f:korean_10_most_used_ingredients},
\ref{f:mexican_10_most_used_ingredients},
\ref{f:moroccan_10_most_used_ingredients},
\ref{f:russian_10_most_used_ingredients},
\ref{f:southern_us_10_most_used_ingredients},
\ref{f:spanish_10_most_used_ingredients},
\ref{f:thai_10_most_used_ingredients},
\ref{f:vietnamese_10_most_used_ingredients} shows top 10 key
ingredient used in the corresponding cuisines.

196: \begin{figure}[!ht]
197: \centering\includegraphics[width=\columnwidth]{images/italian_10_
most_used_ingredients.png}
198: \caption{Top 10 Ingredients
}\label{f:italian_10_most_used_ingredients}
201: \begin{figure}[!ht]
202: \centering\includegraphics[width=\columnwidth]{images/brazilian_1
0_most_used_ingredients.png}
203: \caption{Top 10 Ingredients
}\label{f:brazilian_10_most_used_ingredients}
206: \begin{figure}[!ht]
207: \centering\includegraphics[width=\columnwidth]{images/british_10_
most_used_ingredients.png}
208: \caption{Top 10 Ingredients
}\label{f:british_10_most_used_ingredients}
211: \begin{figure}[!ht]
212: \centering\includegraphics[width=\columnwidth]{images/cajun_creol
e_10_most_used_ingredients.png}
213: \caption{Top 10 Ingredients
}\label{f:cajun_creole_10_most_used_ingredients}
216: \begin{figure}[!ht]
217: \centering\includegraphics[width=\columnwidth]{images/chinese_10_
most_used_ingredients.png}
218: \caption{Top 10 Ingredients
}\label{f:chinese_10_most_used_ingredients}
221: \begin{figure}[!ht]
222: \centering\includegraphics[width=\columnwidth]{images/filipino_10
_most_used_ingredients.png}
223: \caption{Top 10 Ingredients
}

```

```

} \label{f:fili10}
226: \begin{figure} [!ht]
227: \centering \includegraphics [width=\columnwidth] {images/french_10_m
ost_used_ingredients.png}
228: \caption{Top 10 Ingredients
} \label{f:fili10}
231: \begin{figure} [!ht]
232: \centering \includegraphics [width=\columnwidth] {images/greek_10_mo
st_used_ingredients.png}
233: \caption{Top 10 Ingredients
} \label{f:greek10}
236: \begin{figure} [!ht]
237: \centering \includegraphics [width=\columnwidth] {images/indian_10_m
ost_used_ingredients.png}
238: \caption{Top 10 Ingredients
} \label{f:indian10}
241: \begin{figure} [!ht]
242: \centering \includegraphics [width=\columnwidth] {images/irish_10_mo
st_used_ingredients.png}
243: \caption{Top 10 Ingredients
} \label{f:irish10}
246: \begin{figure} [!ht]
247: \centering \includegraphics [width=\columnwidth] {images/jamaican_10
_most_used_ingredients.png}
248: \caption{Top 10 Ingredients
} \label{f:jamaican10}
251: \begin{figure} [!ht]
252: \centering \includegraphics [width=\columnwidth] {images/japanese_10
_most_used_ingredients.png}
253: \caption{Top 10 Ingredients
} \label{f:japanes10}
256: \begin{figure} [!ht]
257: \centering \includegraphics [width=\columnwidth] {images/korean_10_m
ost_used_ingredients.png}
258: \caption{Top 10 Ingredients
} \label{f:korean10}
261: \begin{figure} [!ht]
262: \centering \includegraphics [width=\columnwidth] {images/mexican_10_
most_used_ingredients.png}
263: \caption{Top 10 Ingredients
} \label{f:mexican10}
266: \begin{figure} [!ht]
267: \centering \includegraphics [width=\columnwidth] {images/moroccan_10
_most_used_ingredients.png}
268: \caption{Top 10 Ingredients
} \label{f:moroccan10}

```

```

271: \begin{figure} [!ht]
272: \centering\includegraphics[width=\columnwidth]{images/russian_10_
    most_used_ingredients.png}
273: \caption{Top 10 Ingredients
    }\label{f:russian_10_most_used_ingredients}
276: \begin{figure} [!ht]
277: \centering\includegraphics[width=\columnwidth]{images/southern_us
    _10_most_used_ingredients.png}
278: \caption{Top 10 Ingredients
    }\label{f:southern_us_10_most_used_ingredients}
281: \begin{figure} [!ht]
282: \centering\includegraphics[width=\columnwidth]{images/spanish_10_
    most_used_ingredients.png}
283: \caption{Top 10 Ingredients
    }\label{f:spanish_10_most_used_ingredients}
286: \begin{figure} [!ht]
287: \centering\includegraphics[width=\columnwidth]{images/thai_10_mos
    t_used_ingredients.png}
288: \caption{Top 10 Ingredients
    }\label{f:thai_10_most_used_ingredients}
291: \begin{figure} [!ht]
292: \centering\includegraphics[width=\columnwidth]{images/vietnamese_
    10_most_used_ingredients.png}
293: \caption{Top 10 Ingredients
    }\label{f:vietnamese_10_most_used_ingredients}
307: Figure \ref{f:ingredient_modularity} shows ingredient cluster of
    more than 1000 nodes. This graph is nice to look at but difficult
    to read due to lot many nodes and edges in the graph.
308: \begin{figure} [!ht]
309: \centering\includegraphics[width=\columnwidth]{images/ingredient_
    modularity.png}
310: \caption{Ingredient Cluster }\label{f:ingredient_modularity}
313: Figure \ref{f:ingredient_modularity100} shows ingredient cluster
    of around 100 nodes. We generated this graph by reducing nodes
    and edges to make it more readable. This graph provides us with
    our top 5 cuisine clusters.
314: \begin{figure} [!ht]
315: \centering\includegraphics[width=\columnwidth]{images/ingredient_
    modularity100.png}
316: \caption{ingredient Cluster 100 Nodes
    }\label{f:ingredient_modularity100}

```

figures 28  
 tables 1  
 includegraphics 25  
 labels 29

```
refs 10
floats 29
```

```
True : ref check passed: (refs >= figures + tables)
True : label check passed: (refs >= figures + tables)
True : include graphics passed: (figures >= includegraphics)
False : check if all figures are referred to: (refs >= labels)
```

```
Label/ref check
passed: True
```

```
When using figures use columnwidth
[width=1.0\columnwidth]
do not change the number to a smaller fraction
```

```
find textwidth
```

---

```
passed: True
```

```
below_check
```

---

```
bibtex
```

---

```
label errors
```

```
bibtex errors
```

```
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
```

```
bibtex_empty_fields
```

---

```
entries in general should not be empty in bibtex
```

```
find ""
```

---

```
passed: True
```

```
ascii
```

---

```
=====
The following tests are optional
=====
```

```
Tip: newlines can often be replaced just by an empty line
```

```
find newline
```

---

```
passed: True
```

```
cites should have a space before \cite{} but not before the {
```

```
find cite {
```

---

```
passed: True
```

# Comparison between different classification algorithms in Digit Recognizer

Junjie Lu

Indiana University Bloomington  
3322 John Hinkle Place  
Bloomington, Indiana 47408  
junjlu@iu.edu

Yuchen Liu

Indiana University Bloomington  
1750 N Range Rd  
Bloomington, Indiana 47408  
liu477@iu.edu

Wenxuan Han

Indiana University Bloomington  
1150 S Clarizz Blvd  
Bloomington, Indiana 47401-4294  
wenxhan@iu.edu

## ABSTRACT

Digit Recognizer is becoming more and more important in many different areas, such as zip code recognizer, banking receipt and balance sheet. Many technology companies are trying to use Big Data to develop more efficient and accurate algorithm for Digit Recognizer. This project uses Digit Recognizer data set from Kaggle.com. There are more than 42000 samples in the data set. Each sample contains 784 features which contain pixel information from a  $28 \times 28$  graph. Each pixel has a value between 0 to 255. We use binary classification technique for data cleaning and PCA for feature extraction. For the classification model, we choose five most commonly used classification algorithms, which include Decision Tree (DT), Naive Bayes (NB), Logistic Regression (LR), Random Forest (RF) and Support Vector Machine (SVM). From the result, SVM classifier on PCA data produces the highest accuracy with 0.9813. The time spend is 127 seconds. Naive Bayes classifier on PCA data spends the least amount of time to finish the classification task. It takes less one second and reaches a 0.8651 accuracy.

## KEYWORDS

I523, HID213, HID214, HID209, Big Data, Digit Recognition, Cross Validation, Decision Tree, Naive Bayes, Logistic Regression, Random Forest, Support Vector Machine

## 1 INTRODUCTION

People have made a great improvement in digital recognition in recent years. And it plays significant roles in many different areas. Zip code recognizer can scan zip code for post office automatically. Recognizer in banks can help managing user account by scanning their account number. They help people a lot in increasing working efficiency. And many new productions use digital recognition to authenticate password. In this situation, the accuracy and efficiency of recognition become more and more essential and methods in order to increase the accuracy and efficiency are also required.

Fortunately, people have already developed many different types of techniques to avoid faults and decrease running time in recent few years. Several algorithms will be mentioned here. Logistic regression, the most frequently used algorithm in the field of machine learning, also has a good performance in digital recognition. Decision tree is commonly used in decision analysis. It can identify strategies to get a result, in this case, it can also play an important role in digital recognition. Naive Bayes classifier would also be used. Random forest is also a widely used technique in the field of classification and regression. Its special structure with the multitude of decision trees would help it get a fantastic result. Support

vector machine can efficiently perform non-linear classification hence it also be considered frequently. These algorithms have different structures so that they have different performance. We can also observe running time and accuracy of different algorithms with different kind of data. In this paper, we are going to talk about this and make the comparison between algorithms in accuracy and efficiency.

## 2 EXPERIMENT PREPARATION

In this paper, we choose the data of Digit Recognizer from Kaggle.com in order to test different classification algorithms [5]. The goal of this experiment is to correctly identify digits from a data set of tens of thousands of handwritten images. Thus, we could compare the pros and cons of each technique through the recognition accuracy and time-consuming.

### 2.1 Data Set Description

In train.csv data file, it contains 42000 gray-scale images of hand-drawn digits, from zero through nine. Each image is a  $28 \times 28$  pixels matrix with a total of 784 pixels [5]. Each pixel has a single pixel-value which is an integer from 0 to 255 associated with it, indicating the lightness or darkness of that pixel (higher numbers meaning lighter). In this experiment, we have plotted the graph in order to see the appearance of these digits easily. Figure 1 shows the first 70 samples.

The training data set has 785 columns. The first column called “label”, is the digit that was drawn by the user. The rest of the columns contain the pixel-values of the associated image. Each pixel column in the training set has a name like  $pixelx$ , where  $x$  is an integer between 0 and 783. To locate a pixel on the image, suppose that we have decomposed  $x$  as  $x = i * 28 + j$ , where  $i$  and  $j$  are integers between 0 and 27. Then  $pixelx$  is located in row  $i$  and column  $j$  of this matrix [5]. Visually, if we omit the “pixel” prefix, the pixels make up the image like the following form:

000	001	002	003	...	026	027
028	029	030	031	...	054	055
056	057	058	059	...	082	083
:	:	:	:	:	:	:
728	729	730	731	...	754	755
756	757	758	759	...	782	783

### 2.2 Data Cleaning

As we mentioned above, it can be seen from both the figure and the pixel-value that the value varies from 0 to 255, which means each

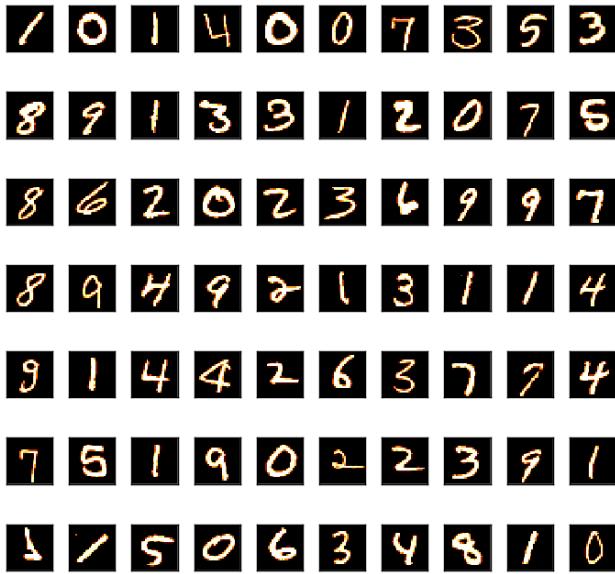


Figure 1: 70 samples of hand-drawn digits in this data set.

feature is a continuous value. Thus, it is possible that such continuous values might affect our later feature selection. Our observation shows that the values are not very high at the boundaries of 0 and  $> 0$ . So here exist three ways to handle it [23]:

- (1) Not do any processing on image;
- (2) Binarize the image. That is, for the values which are 0, keep them as 0; for the values which are greater than 0, change to 1;
- (3) Binarize the image by setting a threshold. That is, for the values which are greater than this threshold, change to 1; otherwise, change to 0.

Obviously, method (2) and (3) will cause the loss of the original information. However, this information may not as important as our expected during the execution of classification algorithms, it could play a positive role in increasing the performance without reducing the accuracy.

In our experiment, we selected method (2) to clean the raw data. Figure 2 shows this operation.

### 2.3 Feature Extraction

Dimension reduction in the field of machine learning refers to using a mapping method to map the data points in the original high-dimensional space into the low-dimensional space. The essence of dimension reduction is to learn a mapping function  $f : x \rightarrow y$ , where  $x$  is the expression of the original data point,  $y$  is the low-dimensional vector representation after the data point mapping [9].

The reason why we use data after dimension reduction is that the redundant information and noise information are contained in the original high-dimensional space, which reduces the accuracy of our model. By dimension reduction, we hope to reduce the error

---

```

from numpy import *
# The data is from 0-255 for each cell.
# Normalize data by set all value > 0 to 1
def data_clean(data):
    m, n = shape(data)
    new_data = zeros((m, n))
    for i in range(m):
        for j in range(n):
            if data[i, j] > 0:
                new_data[i, j] = 1
            else:
                new_data[i, j] = 0
    print("Data clean completed.")
    return new_data

```

---

Figure 2: Bla bla

caused by redundant information and improve the accuracy of identification. We also hope to find the intrinsic structure of the data structure through the dimension reduction algorithm. Also, in this example, there are 784 features in our data. Space, time and computation complexity are all unacceptable. There are many different dimension reduction algorithms for us to choose. In this project, we choose to use Principle Component Analysis (PCA).

#### 2.3.1 PCA

Principal Component Analysis (PCA) is the most commonly used method of supervised linear dimension reduction. Its goal is to map high-dimensional data to a low-dimensional representation of space by some kind of linear projection. The variance of the data is expected to be maximized in the projected dimension. By keep the variance of data as high as possible, PCA can reduce the dimension of data and keep the loss of information of the data as a minimum [3].

A common understanding is that if all the points are mapped together, almost all information (such as the distance between points) is lost. If the post-mapping variance is as large as possible, the data points are spread apart to preserve more information. It can be proved that PCA is a linear dimension reduction method that loses the original least data information.

One of the questions we faced while we are using PCA is that: how many components should we choose for the model after dimension reduction. In order to solve this problem, we use Explained Variance as our threshold standard. Explained Variance is an important indicator of PCA dimension reduction. The Explained Variance shows the amount of variance explained by each of the selected components. The first column of the PCA model always explains the most variance and the variance explained will keep decrease as the number of column increase. Generally, a dimension with a cumulative contribution rate of about 90% is selected as a reference dimension for PCA dimensionality reduction. In this project, in order to get a more accurate result, we choose 95% as our threshold.

---

```
from sklearn.decomposition import PCA
```

---

```

def feature_selection(data):
    pca = PCA()
    pca.fit(data)
    ev = pca.explained_variance_
    ev_ratio = []
    for i in range(len(ev)):
        ev_ratio.append(ev[i] / ev[0])

    # select number of component which have a higher ratio
    # than 0.05 with the first components
    n = 0
    for i in range(len(ev_ratio)):
        if ev_ratio[i] < 0.05:
            n = i
            break

    # Then, PCA the model by the number of components
    pca = PCA(n_components=n, whiten=True)
    return pca.fit_transform(data)

```

---

After calculating the explained variance for each component, we decide to choose 30 components for our model. Which shows that there will be 30 features in our model.

### 3 EXPERIMENT ALGORITHMS

We aim to select five most commonly used classification algorithms which include Decision Tree, Naive Bayes (NB), Logistic Regression (LR), Random Forest (RF) and Support Vector Machine (SVM). This section offers a broad overview of these algorithms before applying them to the digit recognizer problem to compare their characteristics. Then, the result of the different algorithm on different data will show on a table.

For each algorithm, we use:

- (1) PCA data - data after using PCA to reduce the dimension on raw data
- (2) Clean data - data after our data cleaning process, which set all values greater than 0 to 1 in our data
- (3) PCA Clean daata - data after using PCA to reduce the dimension on clean data after data cleaning process

#### 3.1 Cross-Validation

When we build the model, it is normal to follow the principle of simplification since the simpler model we built, the better performance we will get. However, for some complicated problems, our model will also become more complex which might cause the overfitting problem. In order to solve this problem, we introduce the cross-validation technique. Cross-validation is a technique to evaluate predictive models by partitioning the original sample into a training set to train the model, and a test set to evaluate it [13].

The purpose of cross-validation is to select the model with the optimal parameters. After the model is set up, tuning the parameters is a very time-consuming process. Through cross-validation, we can get the model with the optimal parameters much easier. Here are some steps about cross-validation procedure:

- (1) Prepare the candidate models,  $M_1, M_2, M_3, \dots$  (the model framework is consistent, only different on the parameters);

- (2) For each model, use cross-validation to return the accuracy and error rate information of the model, the result should be the average of cross-validation;
- (3) Select the best model by comparing the accuracy or error of the different models.

There are some types of cross-validation which are common to use: K-fold cross-validation and Leave-one-out cross-validation.

- K-fold:

This method is to divide the data set into  $k$  subsets. Each time, select one of the  $k$  subsets as the test set and the other  $k - 1$  subsets become a training set. Then the average accuracy or error across all  $k$  trials is computed [13]. In general, we choose 10 as the value of  $k$ .

- Leave-one-out (LOO):

This method is K-fold cross-validation taken to its logical extreme, with  $k = n$  ( $n > k$ ), the number of data points in the set [13]. That is, it randomly select  $n$  samples as a training set and the rest as a test set. Since the time complexity of this cross-validation is factorial, it is not an appropriate method for big data set.

In this project, we use K-fold cross validation technique to reduce over-fitting of our model and increase the accuracy in each model. We use the function `cross_val_score` from the `sklearn` package. It have several important parameters to set [7].

- (1) CV: int, cross-validation generator or an iterable, optional  
This parameter determines the cross-validation splitting strategy, which determined the number of fold we need to use. In our project, we use the default 3-fold cross-validation. Because 3-fold provide us the result in a reasonable time and accuracy.
- (2) Scoring: string, callable or None, optional, default: None  
The scoring parameter determines what to return after we call the function. We just this parameter to ‘accuracy’, which will return the accuracy between 0 to 1 for each model.

After we receive the result for each validation, we generate the mean of each result and use the result as the accuracy of the model.

---

```

from datetime import datetime
from sklearn.cross_validation import cross_val_score

def model_acc(data, label, model):
    start = datetime.now()
    acc = cross_val_score(model, data, label, cv=5, scoring="accuracy").mean()
    end = datetime.now()
    time_use = (end - start).seconds

    print("Time use: ", time_use)
    print("Accuracy by cross validation: ", acc)

```

---

#### 3.2 Decision Tree

##### 3.2.1 Introduction

Decision tree builds classification or regression models in the form of a tree structure (either binary or non-binary) [17]. Each of its non-leaf nodes represents a test on the characteristic attributes,

and each leaf node stores a category. The process of decision making using decision tree has the following steps [19]:

- (1) Start at the root node;
- (2) Test the corresponding characteristic attribute of the items that need to be classified;
- (3) Select the branch based on the value until the leaf node is reached;
- (4) The category of stored in the leaf node is the result.

The decision tree construction process rely on attribute selection metrics in order to choose the attribute which has the capability to divide tuples into different classes best. The key step in constructing a decision tree is split attributes which means to construct different branches according to the different partition of a certain characteristic attribute at a node. The goal of this step is to make each split subset as “pure” as possible. Split attributes are divided into three different situations:

- (1) Attributes are discrete values and do not require to generate a binary decision tree. This time, each partition of an attribute becomes a branch;
- (2) Attributes are discrete values and require to generate a binary decision tree. This time, a subset of attribute partitions is used for testing, broken down two branches according to “subordinate to this subset” and “not subordinate to this subset”;
- (3) Attributes are continuous values. This time, determine a value as a *split\_point* and generate two branches according to  $> \text{split\_point}$  and  $\leq \text{split\_point}$ .

There are many attribute selection metric algorithms (e.g. ID3, C4.5, CART, etc.), generally using top-down recursive method with non-backtrack greedy strategy. In our experiment, we applied optimized version of the Classification And Regression Trees (CART) algorithm from scikit-learn library.

The CART algorithm uses a binary recursive segmentation technique [1]: the current sample set is divided into two sub-sample sets, so that each non-leaf node have two branches. Therefore, the decision tree generated by the CART algorithm is a concise binary tree with the root node represents a single input variable ( $x$ ) and a split point on that variable and the leaf nodes contain an output variable ( $y$ ) which has the capability to make a prediction [1].

The first key step of CART algorithm is creating the tree model, it examines each variable and all possible partitions of this variable to observe the best partitions. For discrete values such as  $U = \{x, y, z\}$ , there are three cases of partitions [6]:

$$\{\{x, y\}, \{z\}\}, \{\{x, z\}, \{y\}\}, \{\{y, z\}, \{x\}\}$$

except  $\emptyset$  and  $U$ ; for continuous values, it introduces the idea of “split point”. Suppose one attribute of a sample has  $n$  continuous values, it then has  $n - 1$  splitting points where each of them is the average of two consecutive values  $(a[i] + a[i + 1])/2$ . Partitions of each attribute are sorted by the amount of impurities that they can reduce. The reduction of impurities could use the most popular method of impurity metric which is: Gini index. If we use  $k$  ( $k = 1, 2, 3, \dots, C$ ) to represent the class, where  $C$  is the dependent variable number of the category set. Thus, the Gini impurity of a Node  $A$  could be

defined as [6]:

$$Gini(A) = 1 - \sum_{k=1}^C p_k^2$$

Where  $p_k$  denotes the probability of observation points which belong to class  $k$ . When  $Gini(A) = 0$ , all samples belong to the same class. When  $Gini(A)$  is the maximum, which is  $\frac{(C-1)C}{2}$ , all classes occur with the same probability in nodes.

The second key idea in the CART process is to prune the trees of the training set with independent validation data sets. Analyzing the recursive tree construction of classification and regression tree, it is easy to find that there exists a data over-fitting problem [1]. In the construction of decision tree, many branches reflect the abnormality in training data due to the noises or outliers inside. Using such decision tree to classify the data with unknown categories, the accuracy of classification is not high. So it is essential to detect and subtract these branches. Generally, tree pruning method uses statistical metrics, subtract the least reliable branches, which results in faster classification and improves the ability to separate correctly from the training data. The CART algorithm often adopts the post-pruning method, which is implemented by pruning the branches in a fully grown tree. By deleting the branch of the node to cut tree nodes, the bottom non-pruned node becomes a leaf.

The following part of codes shows how we called CART algorithm in our experiment.

---

```
# Import Library
from sklearn import tree

def dt_classifier(data, label, data_type):
    dt_model = tree.DecisionTreeRegressor()
    dt_model.fit(data, label)
    print("Test " + data_type + " using DT: ")

    # Train the model using the training sets and check score
    model_acc(data, label, dt_model)
```

---

### 3.2.2 Advantage and Disadvantage

Decision Tree has advantages as follow [4]:

- (1) Decision trees are easy to understand and implement, and people have the ability to understand what the decision tree means by explaining it.
- (2) Data preparation is often simple or unnecessary for decision trees, and other techniques often require first generalizing data, such as removing redundant or blank attributes.
- (3) Feasible and effective results for large data sources in a relatively short period of time.
- (4) Not sensitive to missing values
- (5) Can handle irrelevant feature data
- (6) High efficiency. Decision tree only needs to build once. The maximum number of calculations for each prediction does not exceed the depth of the decision tree

Decision Tree also has disadvantages as follow [4]:

- (1) Hard to predict features with continues value
- (2) Need to do a lot of data reprocessing work for time-series data

- (3) When the category is too large, the error rate may increase.
- (4) It does not look good when dealing with data that has a strong correlation between each feature.

### 3.2.3 Result

From table 1, we can find that the Decision Tree algorithm has a highest accuracy 0.8378 when we using Clean data. That's because the Clean data contains all 784 features in the data set. It has the minimum information loss among all three data set. Clean data also have the longest running time, which is 20 seconds.

PCA Clean Data have the second highest accuracy with the lowest running time. By using the PCA to reduce the dimension of the clean data, the running time reduced a lot. The accuracy only decreases by 0.01, which shows that the process of PCA did not lose a lot of information.

When we use decision tree algorithm, PCA data have the lowest accuracy. That's may because the raw data have may noise and redundant information. After we remove this information from our data pre-processing step, our accuracy increased.

## 3.3 Naive Bayes

### 3.3.1 Introduction

Naive Bayes algorithm is a classification technique based on Baye's Theorem with an assumption of independence among predictors [15]. That is to say, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. For example, we may guess a fruit is an orange if it is yellow, round and about 3 inches in diameter. Even if these features depend on each other, all properties independently contribute to the probability that this fruit is an orange, which explain the term 'Naive' [15].

The Baye's Theorem is particularly useful and not complicated. It solves many problems encountered in our life. The purpose of this theorem is that given a conditional probability of a certain condition, obtain the probability of exchanging two conditions. That is, to get  $P(B|A)$  while given  $P(A|B)$ .  $P(A|B)$  is the posterior probability which is also the conditional probability (likelihood) and  $P(A)$  or  $P(B)$  is called a prior probability. We use the following equation to express this theorem:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

The idea of Baye's Theorem is very simple and directly: For the given item which need to be classified, compute the probability of each category under this item. We consider this item belongs to the category with the largest value. The work process of Naive Bayes classification is as follows [20]:

- (1) Let  $D$  be the set of training tuples associated with their class labels. Each tuple is represented by an n-dimensional attribute vector  $X = x_1, x_2, \dots, x_n$ ;
- (2) Suppose there are  $m$  classes  $C_1, C_2, \dots, C_m$ . For the given tuple  $X$ , the classification algorithm will predict that  $X$  belongs to the class with the highest posterior probability. That is, Naive Bayes classification predicts that  $X$  belongs to class  $C_i$  if and only if  $P(C_i|X) > P(C_j|X), 1 \leq j \leq m, j \neq i$ . Thus, the class  $C_1$  with the largest  $P(C_i|X)$  is called the

maximum posterior probability according to the Baye's Theorem:  $P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}$ ;

- (3) Since  $P(X)$  is a constant for all classes, we only require the maximum of  $P(C_i|X)P(C_i)$ . If the prior probability of a class is unknown, then generally assume these classes are equiprobable (i.e.  $P(C_1) = P(C_2) = \dots = P(C_m)$ ) and maximize  $P(C_i|X)$  based on this assumption. Otherwise, maximize  $P(C_i|X)P(C_i)$ ;
- (4) Given a data set with multiple attributes, the computational cost of  $P(C_i|X)$  is very large. In order to reduce this cost, we could make the naive assumption about conditional independent of the class. For the label of a given tuple class, assuming the attribute values are conditionally independent. Therefore, we have

$$P(X|C_i) = \prod_{k=1}^n P(x_k|C_i)$$

To examine whether the attribute is classified or continuous value, we need to consider the following two cases:

- (a) If  $A_k$  is a classified attribute, then  $P(x_k|C_i)$  is the number of tuples of class  $C_i$  whose value is  $x_k$  for attribute  $A_k$  in  $D$  divided by the number of tuples of class  $C_i$  in  $D$  ( $|C_i, D|$ );
- (b) If  $A_k$  is a continuous value attribute, then assume the attribute obeys a Gaussian distribute with the mean  $\eta$  and standard deviation  $\sigma$ , as defined by:

$$g(x, \eta, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\eta)^2}{2\sigma^2}}$$

Thus,  $P(x_k|C_i) = g(x_k, \eta_{C_i}, \sigma_{C_i})$ .

- (5) To predict the label of class  $X$ , calculate  $P(C_i|X)P(C_i)$  for each class  $C_i$ .

The whole Naive Bayes classification could be divided into three stages:

- (1) Preparation stage. The task of this stage is to make the necessary preparation for the Naive Bayes classification. The main work is to determine the characteristic attributes according to the specific situations and make the appropriate partition for each characteristic attribute, and then manually classified some of the items to constitute a training sample set. The input of this stage is all data that need to be classified and the output is the characteristic attribute and the training sample.
- (2) Classifier training stage, the task of this stage is to generate a classifier. The main work is to compute the occurrence frequency of each class in training sample and the conditional probability of all partitions in each category, and then record the results. The input is characteristic attributes and a training sample, the output is a classifier. This stage could be completed automatically by a program.
- (3) Application stage. The task of this stage is to classify items using classifier. The input is classifier and items, and the output is the mapping between items and categories. This stage could also be completed by a program.

	PCA Data	PCA Clean Data	Clean Data
Time	12	9	20
Accuracy	0.8012	0.8234	0.8378

**Table 1: Result For Decision Tree**

The following part of codes shows how we called Naive Bayes algorithm in our experiment.

---

```
# Import Library
from sklearn.naive_bayes import GaussianNB

def nb_classifier(data, label, data_type):
    nb_model = GaussianNB()
    nb_model.fit(data, label)
    print("Test " + data_type + " using NB: ")

    # Train the model using the training sets and check score
    model_acc(data, label, nb_model)
```

---

### 3.3.2 Advantage and Disadvantage

Naive Bayes has advantages as follow [10]:

- (1) Naive Bayesian model originated in classical mathematical theory, which is stable.
- (2) Have a good performance on small-scale data,
- (3) Can handle multi-category tasks.
- (4) For incremental training, especially when the amount of data exceeds memory, we can use batch training to save training time.

Naive Bayes also has disadvantages as follow [10]:

- (1) In theory, the naive Bayes model has the smallest error rate compared to other classification methods. However, this is not always the case. This is because the naive Bayesian model assumes that the features are independent of each other. This assumption often does not hold in practice. When the number of attributes is large or the correlation between attributes is large, the error rate will be huge.
- (2) Need to know the prior probability, and the probability of prior probability depends on the assumption. There are many kinds of hypothetical models, so the prediction results will be poor at some time due to the choice of hypothetical model.
- (3) Because we determine the posterior probability by priority and data to determine the classification, there is a certain error rate in the classification decision.
- (4) Sensitive to the type of raw data.

### 3.3.3 Result

From table 2, we can find that Clean Data have a really low accuracy with the highest time spent. That's because the raw data set did not match the assumption of Naive Bayes. The features are not conditionally independent of each other. The pixels are continues. For example, if pixel1 and pixel3 are both greater than 0, pixel2 will have a more probability to have a value greater than 0.

After we use the dimension reduction technique to reduce the dimension of the data, each component of the data becomes a linear

combination of the original data. The new data fits the assumption of Naive Bayes more. Therefore, the PCA Data and PCA Clean Data have a much better performance than Clean Data. They also have the lowest running time compare to any other algorithms.

The PCA Clean Data have the highest accuracy of 0.8710 which higher than the PCA Data. That's may because of the noise and redundant in the original data.

## 3.4 Logistic Regression

### 3.4.1 Introduction

Logistic regression is a static regression model with a category of the dependent variable. It uses a binary logistic model to estimate binary response probability on predictor variables. In this case, we can know which specific factor makes influence in the presence of risk increasing odds when getting outcomes. We use logistic regression to find the best fitting model to conclude the relationship between variables. Logistic regression generates the coefficients (and its standard errors and significance levels) of a formula to predict a logit transformation of the probability of the presence of the characteristic of interest [18]:

$$\text{logit}(p) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \dots + b_kX_k$$

$p$  is the probability of the presence of the characteristic of interest and odds is logical transformation.

$$\text{adds} = \frac{p}{1-p} = \frac{p(\text{presence of characteristic})}{p(\text{absence of characteristic})}$$

$$\text{logit}(p) = \ln \frac{p}{1-p}$$

There are four ways to input independent variables into the model:

- (1) Enter: enter all variables at the same time
- (2) Forward: enter essential variables one by one
- (3) Backward: enter all variables first and delete non-essential variables one by one
- (4) Stepwise: enter essential variables one by one and check the importance of each variable, delete non-essential ones.

It still has other options:

- (1) Remove variable. Variables would be removed from the model if its significant level is greater than P-value.
- (2) Classification table cutoff value: a value between 0 and 1 which will be used as a cutoff value for a classification table. The classification table is a method to evaluate the logistic regression model. In this table the observed values for the dependent outcome and the predicted values (at the selected cut-off value) are cross-classified [18].
- (3) Categorical: Identify variables in the category.

The following part of codes shows how we called Logistic Regression algorithm in our experiment.

	PCA Data	PCA Clean Data	Clean Data
Time	0	0	20
Accuracy	0.8651	0.8710	0.5397

Table 2: Result For Navie Bayes

---

```
# Import Library
from sklearn.linear_model import LogisticRegression

def lr_classifier(data, label, data_type):
    lr_model = LogisticRegression()
    lr_model.fit(data, label)
    print("Test " + data_type + " using LR: ")

# Train the model using the training sets and check score
model_acc(data, label, lr_model)
```

---

### 3.4.2 Advantage and Disadvantage

Logistic Regression has advantages as follow [11]:

- (1) Very simple to implement and use, widely used in industrial issues
- (2) The amount of computation is very small when classified. Therefore the running time is low and the requirement for the storage space is also low.
- (3) The sigmoid score for each sample is easy to observe. The threshold can be easily determined by user.
- (4) For logistic regression, multicollinearity is not a problem, it can be solved in conjunction with L2 regularization;

Logistic Regression also has disadvantages as follow [11]:

- (1) When the feature space is large, the performance of logistic regression is not very good.
- (2) May have the under-fitting problem, the general accuracy is not high.
- (3) Can only deal with the binary classification problem (based on this, softmax can be used for multi-classification), and must be linearly separable.
- (4) For non-linear features, normalization is required.

### 3.4.3 Result

The result of logistic regression is pretty impressive. This is a 10-categorical classification problem, and logistic regression did a good job on this task.

When we get this result, we are thinking if we having an over-fitting result. Therefore, we add a regularization parameter to penalize the features. We use l2 regularization as our parameter when we create our logistic classifier. We also use cross-validation skill to increase our sample size. The results show that the accuracy is still around 90%. Therefore, we are not having an over-fitting problem.

The running time of logistic regression is relatively high. For Clean Data, it received the accuracy of 0.9064 with 218 seconds. PCA Data and PCA Clean Data have a lower accuracy with a much lower time spend. Also, we noticed that the PCA Data accuracy is a little bit higher than the PCA Clean Data. That's may because the clean data make some of the information loss in the raw data.

## 3.5 Random Forest

### 3.5.1 Introduction

Random forest uses a random way to build a forest within many decision trees. There is no correlation between each tree in a random forest [21]. After getting the forest, when a new input sample comes in, each decision tree required to make a judgment separately in order to see which class the sample belongs to (for the classification algorithm), and predict the sample for the category which has most selected.

Random forest is mainly used for regression and classification. It is somewhat similar to the bagging which utilizes decision trees as a basic classifier. Bagging could generate a decision tree after replay a sample in each bootstrap and do not make more intervention while generating these trees. Random forest is also sampling with bootstrap, but the difference is that when constructing each tree, every node variable is generated only in a small number of randomly selected variables. Therefore, not only the samples are random, but also the generation of each node's features. Since the combination classifier is more effective than the single classifier, random forest could classify the data and give the importance evaluation of each variable.

The basic principle of random forest is to get a new training sample set by selecting  $k$  samples from the original training sample set  $N$ , and then make up a random forest according to  $k$  classification trees. The classification result of the new data depends on the score of the tree votes [14]. In essence, it is an improvement on the decision tree algorithm: it combines multiple decision trees, each tree established depends on an independently sample and has the same distribution. The classification error relies on each the classification ability of a tree and the correlation between them. Feature selection uses a random method to split each node, and then compare the error generated in different situations. The inherent estimation error, classification ability and relevance determine the number of features [14].

Since there are many decision trees in the forest, once a new input sample comes in, each decision tree make a decision to check what the class the sample belongs to, and which one is chosen most to the prediction. There are two selection metrics for decision trees to split attributes [19]:

- (1) Information gain
  - (a)  $I(s_1, s_2, \dots, s_m) = -\sum_{i=1}^m p_i \log_2(p_i)$ , where  $S$  is the data set,  $m$  is the number of categories,  $p_i \approx \frac{|S_i|}{|S|}$  is the probability for any sample belongs to  $C_i$ ,  $C_i$  is a class label and  $s_i$  is the number of samples on  $C_i$ ;
  - (b) The smaller  $I(s_1, s_2, \dots, s_m)$ , the more ordered of the sample and the better the classification effect;
  - (c) Entropy of the subsets partitioned by attribute  $A$ :  $A$  has  $V$  different values,  $S$  is partitioned by  $A$  into  $V$  subsets  $s_1, s_2, \dots, s_V$ , where  $s_{ij}$  is the number of samples

	PCA Data	PCA Clean Data	Clean Data
Time	27	21	218
Accuracy	0.8891	0.8862	0.9064

**Table 3: Result For Logistic Regression**

of  $C_i$  in subset  $s_j$ . Then, we have

$$E(A) = \sum_{j=1}^V \frac{(s_{1j} + \dots + s_{mj})}{s * I(s_{1j}, \dots, s_{mj})}$$

- (d)  $G = I(s_1, s_2, \dots, s_m)E(A)$ ;
  - (e) Select the attribute with the maximum information gain as the split attribute.
- (2) Gini index
- (a) Set  $S$  contains  $N$  categories of records, then its Gini index is the frequency of the occurrence of  $p_j$ ;
  - (b) If set  $S$  is partitioned into  $m$  parts  $s_1, s_2, \dots, s_m$ , this segmentation is the Gini split;
  - (c) Select the attribute with the smallest Gini split as a split attribute.

In order to implement random forest, we should follow these steps:

- (1) The input original training set is  $N$ , use bootstrap to extract  $k$  samples randomly and build  $k$  decision trees;
- (2) Suppose there are  $m_A$  variables, then randomly extract  $m_T$  variables from each node of each tree to find one of the variables with the highest classification ability in  $m_T$  variables. The threshold of the variable classification is determined by checking each classification point;
- (3) Maximize the growth of each tree without any pruning;
- (4) Constitute the random forest with these decision trees. Use random forest to determine and classify the new data, and the results are based on votes amount of the tree classifier.

The following part of codes shows how we called Random Forest algorithm in our experiment.

---

```
# Import Library
from sklearn.ensemble import RandomForestClassifier

def rf_classifier(data, label, flag):
    rf_model = RandomForestClassifier(n_estimators=100)
    rf_model.fit(data, label)
    print("Test " + flag + " using RF: ")

# Train the model using the training sets and check score
model_acc(data, label, rf_model)
```

---

### 3.5.2 Advantage and Disadvantage

Random Forest has advantages as follow [2]:

- (1) It can handle very high-dimensional data, and do not have to do feature selection, feature subset is randomly selected
- (2) It can provide which feature is more important after training.

- (3) When creating a random forest, the use of generalization error is an unbiased estimation, which shows that this model has a high generalization ability.
- (4) Easy to make a parallel method, training tree and tree are independent of each other.
- (5) In the training process, the algorithm is able to detect the interaction between the features.
- (6) For unbalanced data sets, it can balance the model automatically.
- (7) If a large part of the features is lost, the model can still maintain the accuracy.

Random Forest also has disadvantages as follow [2]:

- (1) There may be many similar decision trees that mask the real results.
- (2) Small data or low dimensional data may not produce the best classification.
- (3) Much slower than single decision tree algorithm.
- (4) Random forests can be over-fitting on some noisy classifications or regression problems
- (5) For feature with different value range, the more value-separated features will have a greater impact on random forests

### 3.5.3 Result

From table 4, we can find that Clean Data performed perfectly in this case. It takes the shortest time and reached a 0.9647 accuracy.

The result shows an interesting phenomenon: Clean Data cost less time than PCA Data and PCA Clean Data. In order to explain this phenomenon, we have to check what parameter we choose when we build our random forest classifier. From sklearn API document, we can find that the first default parameter is the number of trees in the forest. For all the data, we set the number of trees to the default number, which is 10. However, in Clean Data, many features are correlated to each other, which means that there may many similar decision trees. For PCA Data and PCA Clean Data, most of the features are independent of each other. Therefore, the running time for Clean Data is higher than PCA Data and PCA clean Data.

Also, we know that when there are similar decision trees in the random forest, the real results may be masked. Therefore, although the Clean Data have a really high accuracy, it may still not as good as the PCA Data and PCA Clean Data result. When we running the classifier on an untested data set, the classifier made by PCA Clean Data may have the best performance among the three.

## 3.6 Support Vector Machine

### 3.6.1 Introduction

Support vector machines are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis [22]. It is mostly used in classification.

	PCA Data	PCA Clean Data	Clean Data
Time	126	107	56
Accuracy	0.9483	0.9497	0.9647

**Table 4: Result For Random Forest**

People can plot each data as a point in an n-dimensional space and give each feature a value. Finding the hyperplane which can differentiate two classes very well can complete classification. As for hyperplane, we must know the notation used to define a hyperplane [12]:

$$f(x) = \beta_0 + \beta^T x$$

$\beta$  is weight and  $\beta_0$  is bias. The optimal hyperplane can be represented in an infinite number of different ways by scaling of  $\beta$  and  $\beta_0$ . The one we choose is [12]:

$$|\beta_0 + \beta^T x| = 1$$

$x$  is the training sample who is the most closest to hyperplane. It is known as canonical hyperplane. Distance between point and hyperplane is [12]:

$$\text{distance} = \frac{|\beta_0 + \beta^T x|}{\|\beta\|}$$

$$\text{distance}_{\text{support vector}} = \frac{|\beta_0 + \beta^T x|}{\|\beta\|} = \frac{1}{\|\beta\|}$$

$$M = 2 * \text{distance}_{\text{support vector}} == \frac{2}{\|\beta\|}$$

$$\min L(\beta) = \frac{1}{2} \|\beta\|^2 \text{ subject to } y_i(\beta^T + \beta_0) \geq 1, \forall i$$

In Python, scikit-learn is a widely used library for implementing machine learning algorithms, SVM is also available in the scikit-learn library and follows the same structure (Import library, object creation, fitting model and prediction). Let's look at the below code [16]:

---

```
# Import Library
from sklearn import svm

# Assumed you have, X (predictor) and Y (target) for training data set and validation or testing data set

# Create SVM classification object
model = svm.SVC(kernel='rbf', C=10)

# there are various option associated with it, like changing kernel, gamma and C value
# Train the model using the training sets and check score
model.fit(X, y)
model.score(X, y)

# Predict Output
predicted = model.predict(x_test)
```

---

The e1071 package in R is used to create Support Vector Machines with ease. It has helper functions as well as code for the Naive Bayes Classifier. The creation of a support vector machine in R and Python follow similar approaches, let's take a look now at the following code [16]:

---

```
# Import Library
require(e1071) #Contains the SVM

Train <- read.csv(file.choose())
Test <- read.csv(file.choose())

# there are various options associated with SVM training; like changing kernel, gamma and C value

# create model
model <- svm(Target~Predictor1+Predictor2+Predictor3, data=Train,
kernel='linear', gamma=0.2, cost=100)

# Predict Output
preds <- predict(model, Test)
table(preds)
```

---

### 3.6.2 Advantage and Disadvantage

Support vector machine has advantages as follow:

- (1) More efficient in high dimensional space.
- (2) Effective when the number of samples is smaller than the number of dimensions.
- (3) Can memorize efficiently by using a subset of training sample in decision function.
- (4) Flexible by changing Kernel functions for different customers.

And it also has disadvantages as follow:

- (1) It would over-fitting in choosing Kernel functions when the number of samples is much smaller than the number of features.
- (2) Must pay more attention to regularization term.
- (3) It can only get probability by an expensive five-fold cross-validation instead of calculating directly.

### 3.6.3 Result

By using SVM to build our classifier, we received a really great accuracy score. The PCA Data received a 0.9814 accuracy of 127 seconds. The running complexity of SVM is  $O(N^3 + LN^2 + d * L * N)$ , which  $N$  is the number of support vector choose,  $L$  is the number of samples and  $d$  is the number of features of the data set. Therefore, SVM algorithm will run really slow on the large data set. Therefore, when we use the Clean Data, which include more than 42000 samples and 784 features, it takes 1029 seconds to finish the job. We also try to use SVM direct on our raw data. It takes forever to get a result.

SVM can get much better results than other algorithms in the small sample training set. SVM has become one of the most commonly used and effective classifiers. By using the concept of margin, a structured description of the data distribution is obtained, thereby reducing the need for data size and data distribution.

	PCA Data	PCA Clean Data	Clean Data
Time	127	90	1029
Accuracy	0.9814	0.9785	0.9575

**Table 5: Result For Support Vector Machine**

SVM model has three very important parameters kernel, C and gamma[8].

- (1) Kernel: string, optional. This parameter specifies the kernel type to be used in the algorithm. There are many different kernels that can be used in SVM. For example, linear, polynomial, sigmoid, Radial basis function (RBF) and pre-computed. In this project, choose to use RBF. Because:
  - (a) The RBF kernel function can map a sample to a higher dimensional space, and the linear kernel function is a special case of RBF. That is to say, if RBF is considered, then it is unnecessary to consider the linear kernel function.
  - (b) Compared with polynomial kernel function, RBF needs to determine fewer parameters, the number of kernel function parameters directly affect the complexity of the function. In addition, when the order of the polynomial is relatively high, the elemental values of the kernel matrix will tend to positive infinity or negative infinity, while the RBF will reduce the numerical calculation difficulties.
  - (c) RBF and sigmoid have similar performance for some parameters.
- (2) C is the penalty coefficient, which shows the tolerance of the bias. If your C is small, it will give you a great distance, but as a trade-off, we have to ignore some misclassified samples; on the other hand, if you have a large C, you will try to correctly classify all the samples, but the price is the margin space will be small. In our example, we choose c equals to 10, which is a relatively large c value, which brings us a more accurate classifier.
- (3) Gamma defines how much influence a single training example has. It determines the distribution of the data after mapping to a new feature space. The larger the gamma is, the less the support vector it will be. The smaller the gamma value is, the more the support vector it will be. The number of support vectors affects the speed of training and prediction. Also, if we set gamma large, it will have the over-fitting problem. Therefore, in this project, we decided to use the default gamma value, which is

$$\text{gamma} = \frac{1}{\text{number of features}}$$

In this task, SVM have a really great performance, the running time is also acceptable.

## 4 CONCLUSION

From table 6, we can easily find that when we use SVM classifier on PCA Data, we will receive the highest accuracy among all 5 different algorithms. The highest accuracy we reached for this project is 0.9813, which shows that our classifier predicts 98.13%

of the sample correct by using our SVM classifier. The time of training the model takes 127 seconds. The time spent is acceptable. The accuracy of SVM on PCA Clean Data has the second highest accuracy, which is 0.9785. The difference between first and second highest accuracy is about 0.0028, which is really small. However, the time spent saved 41.1%. Therefore, SVM on PCA Clean Data is also a reasonable choice for the Digit Recognition task.

Random Forest can be explained as a combination of many decision trees. Decision tree can be explained as a special case of Random Forest, which set the number of trees in the Random Forest to 1. Therefore, Random Forest has a much better performance than decision tree in all three data set. As a trade-off, the time spent for Random Forest is much higher than Decision Tree.

Compare to other four Classifiers, Naive Bayes has the fastest training speed. For PCA Data and PCA Clean Data, Naive Bayes Classifier takes less than one second to train the classifier. And for Clean data, which contains all 784 features, it takes only 6 seconds to train the classifier. The reason why Naive Bayes is fast is that:

- (1) The algorithm does not need to iterate to get the result. The running time is approximately linear.
- (2) It makes an assumption of independence between its features, so that parameter estimates can be calculated independently and thus possibly very quickly.
- (3) The prior probability values do not change. Therefore, the prior probability can be calculated and stored in memory in the first place.

However, we have to be very careful about the assumption made by Naive Bayes, or we will get a very low accuracy.

Logistic Regression received an average performance among the 5 algorithms. It achieves a 0.8891 accuracy in 27 seconds on PCA data. However, when we use logistic regression, we have to pay a lot of attention to over-fitting problem. We should use regularization and cross-validation to reduce the probability of over-fitting problems.

To conclude, we decide to use SVM classifier for Digit Recognition Task. We should definitely use feature extraction on the data because of the running time and over-fitting problem. The Binary Data cleaning method is optional. If we want to have higher accuracy, we should not use Binary Data cleaning. As a trade-off, if we want to have faster training speed, we should use Binary Data cleaning.

## 5 LIMITATIONS

Our analysis is far from perfect. There are several points that we want to point out as discussion and also opportunities for future improvement.

- (1) We can try several more classification algorithms. For example,  $K^{th}$  Nearest Neighbour (KNN) and Neural Network.

	Decision Tree	Naive Bayes	Logistic Regression	Random Forest	Support Vector Machine
PCA Time	12	<b>0</b>	27	126	127
PCA Accuracy	0.8012	0.8651	0.8891	0.9483	<b>0.9813</b>
PCA Clean Time	9	<b>0</b>	21	107	90
PCA Clean Accuracy	0.8234	0.8710	0.8862	0.9497	0.9785
Clean Time	20	6	218	56	1029
Clean Accuracy	0.8378	0.5397	0.9064	0.9647	0.9575

**Table 6: Result For Different Algorithm with Different Data Cleaning & Feature Extraction method**

We can use some more complex algorithms too, such as Convolution Neural Network (CNN).

- (2) We can focus more on tune parameter. For example, we can use the Grid Search on SVM to get a better parameter combination.
- (3) We can choose a different Data Cleaning Method. For example, we can set a threshold on data. Any value greater than 50 will be set to 1.
- (4) We can choose a different Feature Extraction or Feature Selection method. For example, LDA. Unlike PCA, LDA is an unsupervised dimension reduction method.

## ACKNOWLEDGMENTS

The authors would like to thank Professor Gregor von Laszewski and all TAs for providing the resource, tutorials and other related materials to write this paper.

## REFERENCES

- [1] Jason Brownlee. 2016. Classification And Regression Trees for Machine Learning. (April 2016). <https://machinelearningmastery.com/classification-and-regression-trees-for-machine-learning/>
- [2] Daniel S. Chapman, Aletta Bonn, William E. Kunin, and Stephen J. Cornell. 2009. Random Forest characterization of upland vegetation and management burning from aerial imagery. *Journal of Biogeography* 37, 1 (2009), 37–46. <https://doi.org/10.1111/j.1365-2699.2009.02186.x>
- [3] Kazuhiro Hotta. 2008. Non-linear feature extraction by linear PCA using local kernel. *2008 19th International Conference on Pattern Recognition* (2008). <https://doi.org/10.1109/icpr.2008.4761721>
- [4] Hemant Ishwaran and J. Sunil Rao. 2009. Decision Tree: Introduction. *Encyclopedia of Medical Decision Making* (2009). <https://doi.org/10.4135/9781412971980.n97>
- [5] Kaggle. 2015. Data Description. (2015). <https://www.kaggle.com/c/digit-recognizer/data>
- [6] Scikit Learn. 2007. Decision Trees. (2007). <http://scikit-learn.org/stable/modules/tree.html>
- [7] Scikit Learn. 2007. sklearn model selection cross val score. (2007). [http://scikit-learn.org/stable/modules/generated/sklearn.model\\_selection.cross\\_val\\_score.html](http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.cross_val_score.html)
- [8] Scikit Learn. 2007. sklearn svm SVC. (2007). <http://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html>
- [9] Sheng-Jie Liang, Zhi-Hua Zhang, and Li-Lin Cui. 2010. Feature extraction method Based PCA and KICA. *2010 Second International Conference on Computational Intelligence and Natural Computing* (2010). <https://doi.org/10.1109/cinc.2010.5643821>
- [10] J. Luengo and Rafael Rumi. 2015. Naive Bayes Classifier with Mixtures of Polynomials. *Proceedings of the International Conference on Pattern Recognition Applications and Methods* (2015). <https://doi.org/10.5220/0005166000140024>
- [11] Scott Menard. 2010. Introduction: Linear Regression and Logistic Regression. *Logistic Regression: From Introductory to Advanced Concepts and Applications* (2010), 1–18. <https://doi.org/10.4135/9781483348964.n1>
- [12] OpenCV. 2017. Introduction to Support Vector Machines. (December 2017). [https://docs.opencv.org/2.4/doc/tutorials/ml/introduction\\_to\\_svm/introduction\\_to\\_svm.html](https://docs.opencv.org/2.4/doc/tutorials/ml/introduction_to_svm/introduction_to_svm.html)
- [13] OpenML. 2013. 10-fold Crossvalidation. (2013). <https://www.openml.org/a/estimation-procedures/1>
- [14] Savan Patel. 2017. Chapter 5: Random Forest Classifier. (May 2017). <https://medium.com/machine-learning-101/>
- [15] Sunil Ray. 2015. 6 Easy Steps to Learn Naive Bayes Algorithm (with codes in Python and R). (September 2015). <https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/>
- [16] Sunil Ray. 2015. Understanding Support Vector Machine algorithm from examples (along with code). (October 2015). <https://www.analyticsvidhya.com/blog/2017/09/understanding-support-vector-machine-example-code/>
- [17] Dr. Saed Sayad. 2010. Decision Tree - Classification. (2010). [http://www.saedsayad.com/decision\\_tree.htm](http://www.saedsayad.com/decision_tree.htm)
- [18] MedCalc Software. 2017. Logistic regression. (February 2017). [https://www.medcalc.org/manual/logistic\\_regression.php](https://www.medcalc.org/manual/logistic_regression.php)
- [19] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. 2006. *Introduction to Data Mining*. Addison Wesley.
- [20] Wikipedia. 2017. Naive Bayes classifier. (December 2017). [https://en.wikipedia.org/wiki/Naive\\_Bayes\\_classifier](https://en.wikipedia.org/wiki/Naive_Bayes_classifier)
- [21] Wikipedia. 2017. Random forest. (November 2017). [https://en.wikipedia.org/wiki/Random\\_forest](https://en.wikipedia.org/wiki/Random_forest)
- [22] Wikipedia. 2017. Support vector machine. (December 2017). [https://en.wikipedia.org/wiki/Support\\_vector\\_machine](https://en.wikipedia.org/wiki/Support_vector_machine)
- [23] Hui Xiong, Gaurav Pandey, Michael Steinbach, and Vipin Kumar. 2005. Enhancing Data Analysis with Noise Removal. (2005). <https://doi.org/10.21236/ada439494>

## A CODE ATTACHMENT

---

```

##Author: Yuchen Liu HID213, Wenxuan Han HID209, Junjie Lu HID214
##Data: 2017.12.01
##Reference: http://blog.csdn.net/tinkle181129/article/details/55261251

from datetime import datetime
import matplotlib.pyplot as plt
import pandas as pd
from numpy import *
from sklearn import svm
from sklearn import tree
from sklearn.cross_validation import cross_val_score
from sklearn.decomposition import PCA
from sklearn.ensemble import RandomForestClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.naive_bayes import GaussianNB

# 1. read data from csv
def read_data():
    data_set = pd.read_csv("train.csv")
    data = data_set.values[0:, 1:]
    label = data_set.values[0:, 0]
    print("Data load completed.")
    return data, label

# plot 70 samples
def show_pic(data):
    print(shape(data))

```

```

plt.figure(figsize=(7, 7))
for digit_num in range(0, 70):
    plt.subplot(7, 10, digit_num + 1)
    grid_data = data[digit_num].reshape(28, 28)
    plt.imshow(grid_data, interpolation="none", cmap="afmhot")
    plt.xticks([])
    plt.yticks([])
plt.tight_layout()
plt.savefig("data_samples.png")

# 2. Data Cleaning
# The data is from 0-255 for each cell.
# Normalize data by set all value > 0 to 1
def data_clean(data):
    m, n = shape(data)
    new_data = zeros((m, n))
    for i in range(m):
        for j in range(n):
            if data[i, j] > 0:
                new_data[i, j] = 1
            else:
                new_data[i, j] = 0
    print("Data clean completed.")
    return new_data

# 3. Feature Selection by PCA
def feature_selection(data):
    # First, use explained_variance to get recommended number of component
    pca = PCA()
    # pca_parameter = pca.fit(data)
    pca.fit(data)
    ev = pca.explained_variance_
    ev_ratio = []
    for i in range(len(ev)):
        ev_ratio.append(ev[i] / ev[0])

    # select number of component which have a higher ratio
    # than 0.05 with the first components
    n = 0
    for i in range(len(ev_ratio)):
        if ev_ratio[i] < 0.05:
            n = i
            # print(n)
            break

    # Then, PCA the model by the number of components
    # pca = PCA(n_components=n, whiten=True)
    pca = PCA(n_components=n, whiten=True)
    print("Feature selection completed.")
    return pca.fit_transform(data)

# 4. Model Selection
def model_acc(data, label, model):
    start = datetime.now()
    acc = cross_val_score(model, data, label, scoring="accuracy").mean()
    end = datetime.now()
    time_use = (end - start).seconds
    print("Time use: ", time_use)
    print("Accuracy by cross validation: ", acc)

def dt_classifier(data, label, data_type):
    dt_model = tree.DecisionTreeRegressor()
    dt_model.fit(data, label)
    print("Test " + data_type + " using DT: ")
    model_acc(data, label, dt_model)

def nb_classifier(data, label, data_type):
    nb_model = GaussianNB()
    nb_model.fit(data, label)
    print("Test " + data_type + " using NB: ")
    model_acc(data, label, nb_model)

def lr_classifier(data, label, data_type):
    lr_model = LogisticRegression()
    lr_model.fit(data, label)
    print("Test " + data_type + " using LR: ")
    model_acc(data, label, lr_model)

def rf_classifier(data, label, flag):
    rf_model = RandomForestClassifier(n_estimators=100)
    rf_model.fit(data, label)
    print("Test " + flag + " using RF: ")
    model_acc(data, label, rf_model)

def svm_classifier(data, label, flag):
    svm_model = svm.SVC(kernel="rbf", C=10)
    svm_model.fit(data, label)
    # svc_clf = NuSVC(nu=0.1, kernel='rbf', verbose=True)
    print("Test " + flag + " using SVM: ")
    model_acc(data, label, svm_model)

def main():
    data, label = read_data()
    # show_pic(data)
    clean_data = data_clean(data)

    test_type = 3
    for i in range(1, 3):
        print("In %d test" % i)

        if test_type == 0:
            input_data = data
            str = "raw data"
        elif test_type == 1:
            input_data = clean_data
            str = "clean data"
        elif test_type == 2:
            input_data = feature_selection(data)
            str = "pca data"
        elif test_type == 3:
            input_data = feature_selection(clean_data)
            str = "pca clean data"

        dt_classifier(input_data, label, str)
        nb_classifier(input_data, label, str)

```

```
lr_classifier(input_data, label, str)
rf_classifier(input_data, label, str)
svm_classifier(input_data, label, str)
```

---

```
main()
```

bibtext report

=====

This is BibTeX, Version 0.99d (TeX Live 2016)  
The top-level auxiliary file: report.aux  
The style file: ACM-Reference-Format.bst  
Database file #1: report.bib  
Warning--entry type for "cart" isn't style-file defined  
--line 5 of file report.bib  
Warning--entry type for "sklearn.cv" isn't style-file defined  
--line 57 of file report.bib  
Warning--entry type for "sklearn.dt" isn't style-file defined  
--line 68 of file report.bib  
Warning--entry type for "svm.form" isn't style-file defined  
--line 117 of file report.bib  
Warning--entry type for "10f.cv" isn't style-file defined  
--line 129 of file report.bib  
Warning--entry type for "sp.rfc" isn't style-file defined  
--line 139 of file report.bib  
Warning--entry type for "nb.steps" isn't style-file defined  
--line 150 of file report.bib  
Warning--entry type for "svm.code" isn't style-file defined  
--line 162 of file report.bib  
Warning--entry type for "ss.dt" isn't style-file defined  
--line 174 of file report.bib  
Warning--entry type for "lr.form" isn't style-file defined  
--line 185 of file report.bib  
Warning--entry type for "wiki.nb" isn't style-file defined  
--line 208 of file report.bib  
Warning--entry type for "wiki.rf" isn't style-file defined  
--line 219 of file report.bib  
Warning--entry type for "wiki.svm" isn't style-file defined  
--line 231 of file report.bib  
Warning--no number and no volume in PCA  
Warning--page numbers missing in both pages and numpages fields in PCA  
Warning--no number and no volume in DT  
Warning--page numbers missing in both pages and numpages fields in DT  
Warning--no number and no volume in feature\_extra  
Warning--page numbers missing in both pages and numpages fields in feature\_extra  
Warning--no number and no volume in NB  
Warning--page numbers missing in both pages and numpages fields in NB  
Warning--no number and no volume in LR  
Warning--empty address in intro.dm  
Warning--no journal in data\_clean  
Warning--no number and no volume in data\_clean

```
Warning--page numbers missing in both pages and numpages fields in data_clean  
(There were 26 warnings)
```

```
bibtext _ label error
```

```
=====
```

```
report.bib:243:@Article{data_clean,  
report.bib:89:@Article{feature_extra,
```

```
=====
```

```
bibtext space label error
```

```
=====
```

```
bibtext comma label error
```

```
=====
```

```
=====
```

```
latex report
```

```
=====
```

```
[2017-12-11 13.25.28] pdflatex report.tex
```

```
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflatex)
```

```
Missing character: ``"
```

```
Missing character: ``"
```

```
Typesetting of "report.tex" completed in 1.5s.
```

```
./README.yml
```

```
24:81    error    line too long (82 > 80 characters) (line-length)  
45:81    error    line too long (85 > 80 characters) (line-length)  
47:79    error    trailing spaces (trailing-spaces)  
48:81    error    line too long (82 > 80 characters) (line-length)  
48:82    error    trailing spaces (trailing-spaces)  
49:79    error    trailing spaces (trailing-spaces)  
50:81    error    line too long (81 > 80 characters) (line-length)  
50:81    error    trailing spaces (trailing-spaces)  
51:81    error    line too long (89 > 80 characters) (line-length)  
51:89    error    trailing spaces (trailing-spaces)  
52:81    error    line too long (81 > 80 characters) (line-length)  
52:81    error    trailing spaces (trailing-spaces)  
53:81    error    line too long (81 > 80 characters) (line-length)  
53:81    error    trailing spaces (trailing-spaces)  
54:81    error    line too long (86 > 80 characters) (line-length)  
54:86    error    trailing spaces (trailing-spaces)
```

```
=====
```

```
Compliance Report
```

```
=====
name: Han, Wenxuan
hid: 209
paper1: Oct 29 2017 100%
paper2: 100%
project: Dec 04 17 100%
```

```
yamlcheck
```

---

```
wordcount
```

---

```
13
wc 209 project 13 8951 report.tex
wc 209 project 13 8934 report.pdf
wc 209 project 13 719 report.bib
```

```
find "
```

---

```
157: print("Data clean completed.")

250: acc = cross_val_score(model, data, label, cv=5,
                           scoring="accuracy").mean()

254: print("Time use: ", time_use)

255: print("Accuracy by cross validation: ", acc)

302: print("Test " + data_type + " using DT: ")

395: print("Test " + data_type + " using NB: ")

476: print("Test " + data_type + " using LR: ")

565: print("Test " + flag + " using RF: ")

801: data_set = pd.read_csv("train.csv")

804: print("Data load completed.")

815: plt.imshow(grid_data, interpolation="none", cmap="afmhot")
```

```
819: plt.savefig("data_samples.png")

834: print("Data clean completed.")

861: print("Feature selection completed.")

868: acc = cross_val_score(model, data, label,
    scoring="accuracy").mean()

871: print("Time use: ", time_use)

872: print("Accuracy by cross validation: ", acc)

878: print("Test " + data_type + " using DT: ")

885: print("Test " + data_type + " using NB: ")

892: print("Test " + data_type + " using LR: ")

899: print("Test " + flag + " using RF: ")

904: svm_model = svm.SVC(kernel="rbf", C=10)

907: print("Test " + flag + " using SVM: ")

918: print("In %d test" % i)

922: str = "raw data"

925: str = "clean data"

928: str = "pca data"

931: str = "pca clean data"

passed: False

find footnote
-----
passed: True

find input{format/i523}
-----
```

```
passed: False

find input{format/final}
-----
4: \input{format/final}

passed: True

floats
-----
109: \begin{figure} [!ht]
111: \includegraphics[width=\columnwidth]{images/data_samples}
139: data. Figure \ref{A:xyz} shows this operation.
141: \begin{figure} [htb]
160: \caption{Bla bla}\label{A:xyz}

figures 2
tables 0
includegraphics 1
labels 1
refs 1
floats 2

True : ref check passed: (refs >= figures + tables)
True : label check passed: (refs >= figures + tables)
True : include graphics passed: (figures >= includegraphics)
True : check if all figures are referred to: (refs >= labels)

Label/ref check
106: In train.csv data file, it contains $42000$ gray-scale images of hand-drawn digits, from zero through nine. Each image is a $28 \times 28$ pixels matrix with a total of 784 pixels \cite{kaggle}. Each pixel has a single pixel-value which is an integer from 0 to 255 associated with it, indicating the lightness or darkness of that pixel (higher numbers meaning lighter). In this experiment, we have plotted the graph in order to see the appearance of these digits easily. Figure 1 shows the first 70 samples.
339: From table 1, we can find that the Decision Tree algorithm has a highest accuracy 0.8378 when we use Clean data. That's because the Clean data contains all 784 features in the data set. It has the minimum information loss among all three data sets. Clean data also have the longest running time, which is 20 seconds.
429: From table 2, we can find that Clean Data have a really low
```

accuracy with the highest time spent. That's because the raw data set did not match the assumption of Naive Bayes. The features are not conditionally independent of each other. The pixels are continuous. For example, if pixel1 and pixel3 are both greater than 0, pixel2 will have a more probability to have a value greater than 0.

passed: False -> labels or refs used wrong

When using figures use columnwidth  
[width=1.0\columnwidth]  
do not change the number to a smaller fraction

find textwidth

---

passed: True

---

below\_check

---

WARNING: figure and above may be used improperly

129: As we mentioned above, it can be seen from both the figure and the pixel-value that the value varies from 0 to 255, which means each feature is a continuous value. Thus, it is possible that such continuous values might affect our later feature selection. Our observation shows that the values are not very high at the boundaries of 0 and \$>0\$. So here exist three ways to handle it \cite{data\_clean}:

WARNING: code and below may be used improperly

633: In Python, scikit-learn is a widely used library for implementing machine learning algorithms, SVM is also available in the scikit-learn library and follows the same structure (Import library, object creation, fitting model and prediction). Let's look at the below code \cite{svm.code}:

WARNING: algorithm and below may be used improperly

633: In Python, scikit-learn is a widely used library for implementing machine learning algorithms, SVM is also available in the scikit-learn library and follows the same structure (Import library, object creation, fitting model and prediction). Let's look at the

below code \cite{svm.code}:

bibtex

---

label errors

89: feature\_extra: do not use underscore in labels:  
243: data\_clean: do not use underscore in labels:

bibtex errors

This is BibTeX, Version 0.99d (TeX Live 2016)  
The top-level auxiliary file: report.aux  
The style file: ACM-Reference-Format.bst  
Database file #1: report.bib  
Warning--entry type for "cart" isn't style-file defined  
--line 5 of file report.bib  
Warning--entry type for "sklearn.cv" isn't style-file defined  
--line 57 of file report.bib  
Warning--entry type for "sklearn.dt" isn't style-file defined  
--line 68 of file report.bib  
Warning--entry type for "svm.form" isn't style-file defined  
--line 117 of file report.bib  
Warning--entry type for "10f.cv" isn't style-file defined  
--line 129 of file report.bib  
Warning--entry type for "sp.rfc" isn't style-file defined  
--line 139 of file report.bib  
Warning--entry type for "nb.steps" isn't style-file defined  
--line 150 of file report.bib  
Warning--entry type for "svm.code" isn't style-file defined  
--line 162 of file report.bib  
Warning--entry type for "ss.dt" isn't style-file defined  
--line 174 of file report.bib  
Warning--entry type for "lr.form" isn't style-file defined  
--line 185 of file report.bib  
Warning--entry type for "wiki.nb" isn't style-file defined  
--line 208 of file report.bib  
Warning--entry type for "wiki.rf" isn't style-file defined  
--line 219 of file report.bib  
Warning--entry type for "wiki.svm" isn't style-file defined  
--line 231 of file report.bib  
Warning--no number and no volume in PCA  
Warning--page numbers missing in both pages and numpages fields in PCA  
Warning--no number and no volume in DT

```
Warning--page numbers missing in both pages and numpages fields in DT
Warning--no number and no volume in feature_extra
Warning--page numbers missing in both pages and numpages fields in feature_extra
Warning--no number and no volume in NB
Warning--page numbers missing in both pages and numpages fields in NB
Warning--no number and no volume in LR
Warning--empty address in intro.dm
Warning--no journal in data_clean
Warning--no number and no volume in data_clean
Warning--page numbers missing in both pages and numpages fields in data_clean
(There were 26 warnings)
```

bibtex\_empty\_fields

---

entries in general should not be empty in bibtex

find ""

---

passed: True

---

ascii

---

non ascii found 65292

---

=====

The following tests are optional

=====

Tip: newlines can often be replaced just by an empty line

find newline

---

passed: True

cites should have a space before \cite{} but not before the {

find cite {

---

passed: True

# Income Prediction based on Machine Learning Techniques

Borga Edionse Usifo

Indiana University

Bloomington, Indiana 47408

busifo@iu.edu

## ABSTRACT

This project takes a closer look to some of the most used supervised learning algorithms in machine learning. We start with the description of the each of the algorithms then we move it to analytics and findings by using that particular algorithm in our data-set. We also provide advantages and disadvantages of each supervised machine learning algorithm for future reference. We mainly focus on our prediction of the income level of individuals by looking at their age, gender, education, location, and other features given by our data-set. We will try each algorithm and try to pick the best features from our data-set to have an optimal prediction.

## KEYWORDS

i523, HID343, Machine Learning, Income Prediction, Logistic Regression, Ensemble methods

## 1 INTRODUCTION

In this project, we try to showcase the performance of the machine learning algorithms on data which we gather from UCI machine learning repository [22]. This data used by Kohavi R. and Becker B. for their research in improving the in Naive Bayes Classifier's accuracy [21].

Data consists of 15 variables, and we try to predict the income of the individuals. To do this prediction task, we first started with data preparation because the data we receive from UCI machine learning repository [22] not fully prepared for any machine learning algorithm. Our first task was the clean the data while applying some statistical techniques to get insights from the dataset. We also used data transformation methods like One-Hot-Encoding[45] to apply logarithmic functions for improving the machine learning algorithms performance before training the data.

Machine Learning algorithms that we discuss in this paper are Gaussian Naive Bayes [46], K Nearest Neighbors [29], Ensemble Methods (Boosting) [8], Support Vector Machines [6], Logistic Regression [34], and Decision Trees [49]. We try to show their weakness, advantages, and their time consumption while training each of them in machine learning algorithms section.

After providing a brief introduction of each of the supervised machine learning algorithms, we will discuss our findings for of each of the algorithms by comparing their accuracy score, F-1 score, recall, and lastly time comparison.

## 2 IMPORTANCE OF BIG DATA ANALYTICS FOR PREDICTIVE CLASSIFICATION

Importance of big data analytics is getting higher every day since the algorithms become more powerful to predict, classify and cluster any given data set. Importance of our case is any company can be used to predict individuals income to refer them goods in their

income range or governments can provide additional support for the areas that have lower income range. There can be many possible things that can do with this kind of classification predictions.

## 3 DATA PREPARATION

We first used the pandas [28] to help to load the data in data frame format. This gave us a unique advantage, and faster processing of comma separated values for putting into data frame [48]. Our data consist of 15 variables. Some of these variables are continuous, and some of them are categorical variables, and our target variable was "income" attribute. After putting the data into data frames, we first got a statistical snapshot of continuous variables ( age, education, capital gain, capital loss, hours worked) by using the pandas [27] functions as shown in Table 1.

[Table 1 about here.]

### 3.1 Data Cleaning

After getting a snapshot from income data frame, we recognized that there is a column which has no meaning. The first task was to remove this entire column from our dataset we used pandas drop function for doing this task. After removing this column, we had more concise dataframe to analyze.

Moreover, removing the column we have encountered some missing values which labeled as "question marks" in data frame. In order to remove this values we first changed all the "question mark" values to "NaN" values by using pandas "replace" function [26]. After replacing all the question marks with "NaN" values, we used pandas missing value dropping function to remove all the "NaN" values from our dataset.

Furthermore, we start investigating the types of the variables, and in our case, we found two types of variable one of them labeled as "int64" which stands for integer values, other one labeled as object type of variable. From our previous example especially in "scikit-learn" it is better to use float object rather than "int64" for training the machine learning algorithms. Because their numerical output most of the time is "float64" object. We transferred all the "int64" objects to "float64" objects. This was the last step of the cleaning process.

Our last process is changing the string values to numerical values on our target data which consist of string values ("\$ 50K") for machine learning algorithms to understand this target data we need to transfer it to numerical values. Since we have only two categories, we will assign 1 and 0 as numerical values as shown in Table 2.

[Table 2 about here.]

Our shape of the data will also receive impact from changing to numerical. Our number of futures will go from 14 to 103. This is because we implemented one-hot-encode to our dataset. It is called

one hot encoded because we transform the categorical variables into a more acceptable shape for the machine learning algorithms to perform well [45]. In other words “we implement binarization of the category to include as a future to train model [45]”. As we can see in Table 3 and Table 4.

[Table 3 about here.]

[Table 4 about here.]

## 4 DATA EXPLORATION

After cleaning the data, we started our data exploration to learn little bit more from our data and make necessary changes if needed before putting into our machine learning algorithms. The first step in this process is getting the total count of the individuals as well as the count of the individuals who are making more than \$50K and less than \$50K which can be seen in below Table 5.

[Table 5 about here.]

Moreover, we also look at the statistical values of each of the continuous variable we have. Those values given in Table 6. As we can see we have individuals who’re age ranging from 17 to 90 years old with a mean of 38.58. If we look at the capital gains and capital losses, we have a standard deviation of 7385 and 402 respectively this is also another indication of skew in these variables.

[Table 6 about here.]

We used scatter matrix plot and applied the correlation function to see if we have any reliable correlation between any of the variables. As we can see from the correlation matrix Table 7 and correlation numbers Figure 1 we do not have the high correlation between any variables. Correlation values range between -1 to 1. The correlation value of 1 is an indication of perfect positive correlation and correlation number -1 indicates a negative correlation between variables [15]. Because of lower correlation values, it will be tough to determine the classification by just looking at the correlations; this indicates we have sophisticated algorithms to determine the relationship between variables to classify individuals incomes.

[Table 7 about here.]

[Figure 1 about here.]

Furthermore, we also explore the capital gains, capital losses, and hours per week variables which we used a histogram to plot the data into distribution form so we can see how all these attributes distributed. The reason we do the histogram is we want to see any skewness in our data. As shown in the histogram graphs in Figure 2 and Figure 3 in capital gains and capital loss we have highly skewed data which can cause issues later on in our algorithms. We apply a logarithmic function to do highly skewed data to less skewed [24]. Using logarithmic functions adds more value to data from the interpretable standpoint and “it helps to meet the assumptions of inferential statistics [24]”.

[Figure 2 about here.]

[Figure 3 about here.]

Moreover, applying logarithmic function had an impact on distribution. We can see the changes on skew data in Figure 4 after applying logarithmic function.

[Figure 4 about here.]

## 5 MACHINE LEARNING ALGORITHMS TO CONSIDER

We have multiple algorithms to consider when we are doing the supervised learning. Each algorithm has its benefits and drawbacks. We will consider several supervised machine learning algorithms for our predictions. The application we will use to implement these algorithms will be Python Scikit-Learn library. We will briefly explain each parameter included in these algorithms in Scikit-Learn.

First we’ll look at the Scikit-Learn in Python framework we will go through the advantages in Scikit-Learn how we can implement any machine learning in just couple of simple line of codes in Scikit-Learn.

### 5.1 Why Scikit-Learn?

Scikit-learn developed by David Cournapeau in 2007. The development came from while he was working on summer code project for Google. After recognized and published by INRIA in 2010 project start the get more attention among worldwide. There are more than 30 active contributors and has secured several sponsorships from big technology companies[17]. “It also has a goal of providing common algorithms to Python users through consistent interface[2]”. Scikit-Learn consists of several elements to make analytical predictions. These elements are shown below[23]:

**Supervised Learning Algorithms:** One of the most fundamental reason that Scikit-Learn’s popularity comes from highly available supervised learning algorithms. These algorithms vary from regression models to decision trees and many more[23].

**Cross Validation:** Scikit-Learn includes various techniques to check the accuracy or any statistical measure between training and unseen testing set[23].

**Unsupervised Learning Algorithms:** Scikit-Learn had also various algorithms to support many unsupervised algorithms some of these include clustering, factor analysis, and neural network analysis[23].

**Various example data-sets:** Scikit-Learn comes with different data sets included in its package so users can start learning Scikit-Learn without the need of any data-sets[23].

**Feature extraction:** It has rich feature for extracting images or text from data-sets[23].

Algorithms that we will investigate shown below; we will go more deep analysis on each of these algorithms.

- Gaussian Naive Bayes
- Logistic Regression
- K-Nearest Neighbors (KNN)
- Stochastic Gradient Descent Classifier
- Support Vector Machines
- Decision Trees

### 5.2 Gaussian Naive Bayes

Naive Bayes bring many beneficial features; it is widely popular among machine learning applications[41]. The popularity of Naive Bayes comes from being able to handle large projects and data-sets faster than most algorithms[41]. It also can handle complex data-sets with categorical and non-categorical inputs [41]. Naive Bayes based on probabilistic classifier of Bayesian theory. It is also a favorite way of doing text categorization [46].

Term naive comes from it is the method of use probability among categories which assumes of independence among given class of attributes as shown in Figure 5. In other words, if we try to classify individuals from their email communications it will not take the order of words into account. Whereas in the English language we can tell the difference between sentence makes sense or not if we randomly re-order our words in the sentences. So it does not understand the text, it only looks at word frequencies as a way to do the classification. This is why it is called “Naive”.

[Figure 5 about here.]

As we state above Naive Bayes derives from Bayesian Theory where the dimensionality of inputs is relatively high. Bayesian Theorem is stated below [16].

$$P(C | X) = \frac{P(X | C) \times P(C)}{P(X)} \quad (1)$$

Naive Bayes Classifier works as follows [16]:

#### **Advantages of Naive Bayes [16]:**

- Faster classification time for training data-set.
- Because of independent classification it improves classification performance.
- Performance is relatively good.

#### **Disadvantages of Naive Bayes[16]:**

- Often it requires a large number of data-sets to give adequate results.
- On some occasions which are relative to data-sets, it can give less accuracy.

### **5.3 Logistic Regression**

Logistic Regression widely used for predicting “probability of failure in a given system, product, and process [34]”. Logistic Regression also used in natural language analysis, it is an extension of conditional random fields [34]. It works as a classifier which learns the features from the input given and classifies them by multiplying the input value with the weight value [14].

$$P(C | X) = \sum_{i=1}^N W_i \times f_i \quad (2)$$

Main reason that Logistic Regression differs from Linear Regression is output variable for Logistic Regression is binary whereas output variable in Linear Regression is discrete(continuous) [12].

#### **Advantages of Logistic Regression:**

- It does not have any assumptions over distribution of classes [18].
- It is fast to train [18].
- Logistic Regression has fast classifying method of unknown data [18].
- We can easily extend to other regression for multiple classes like multinomial regression [18].

#### **Disadvantages of Logistic Regression:**

- One of the disadvantages of linear regression is it is not providing flexibility in some instances. What we mean by the “lack of flexibility is the linear dependency, and linear decision boundary in the instance space is not valid [42]”.

This disadvantage can be improved changing from Logistic Regression to Choquistic Regression[42].

- Logistic regression can provide poor results when there are more complex relationships in data [9].
- Logistic models also have over-fitting problems which come from a result of sampling bias [31].
- Because of Logistic Regression’s predictions comes from the independent variable if the researcher includes wrong independent variables then model’s prediction will have no value [31].
- Because it is predictions based on 1 and 0 model will have poor performance when predicting continuous variables [31].

### **5.4 K-Nearest Neighbors (KNN)**

K Nearest neighbor has been primarily studied, and this popularity comes from it has been applied to many applications some of these applications are “spatial databases, pattern recognition, geographic information, image retrieval, computer game, and many other applications [29]”. Due to an increase of mobile devices and people tends to use of applications like navigation K-nearest neighbor found itself another widely used area of location-based services due to an ability to found a target location [29].

Intuition behind the K Nearest Neighbor can be described as follows: “ for a set P of n objects and a querying point q, return the k objects in P that are closest to q [29].“

#### **Advantages of K Nearest Neighbors:**

- K Nearest Neighbor is a basic and simple approach to implement [35].
- K Nearest Neighbor can perform well and efficiently with the large amount of data [43].
- K nearest Neighbor also does effectively well with noisy data sets (“if the inverse square of weighted distance used as the distance [43]”). In other words, it is flexible to feature and distance choices [35].

#### **Disadvantages of K Nearest Neighbors:**

- K Nearest Neighbor typically require large dataset to perform well [35].
- Time complexity could be high due to computing distance of each query to all training data points [43]. This time might be improved with some indexing (K-D Tree) [43].
- Determining the value of K can be time-consuming [43].
- It can be unclear to know which type of distance to use, as well as which variability to use to get the optimal results [43].
- Switching the different K values can result in the predicted class labels [30].

Many of these disadvantages are improving with the help of parallel distributed computing. Recent improvements in MapReduce framework allows users to run KNN algorithms in the cluster which had a significant effect on reducing the computation time [19].

Another area of improvements on KNN, is to implement different mapping functions such as kernel KNN, kernel difference weighted KNN, adaptive quasi-conformal kernel nearest neighbor, angular

similarity, local linear discriminant analysis, and Dempster-Shafer [10].

## 5.5 Decision Trees

Decision Tree is another widely used algorithm model for classification and regression. Decision Trees uses a recursive split model where each recursive split is identified by each data point; this is an example of non-parametric hierarchical model [13].

Representation of decision trees is as follows; we sort the instances from root to leaf nodes, this sorting gives insights about the classification of the instance, every outcome descending from the root node corresponds to possible values for that variable [33]. We can classify an instance by starting from the root node and checking the attributes labeled on that node and moving down from that node based on attribute given attribute values [33] as shown in Figure 6.

[Figure 6 about here.]

### Advantages of Decision Trees:

- Decision Tree applications are easy to interpret and understand [32]. This ease comes from their schematic representation [32]. Interpretation between alternatives can be expressed with single numerical number which is the expected value (EV) [32].
- Decision Trees can handle noisy or incomplete data-sets [32]. In other words it requires little effort of data preparation because of its flexibility [7].
- It can handle both nominal and numerical variables [32].
- It can be modified easily whenever the new information is available [32].
- 

### Disadvantages of Decision Trees:

- Because of its use of divide and conquer method they can demonstrate good performance if there are few attributes exist when the attributes level goes into large number decision tree become more complex which will result in poor performance [32].
- Decision Trees are also susceptible to training set which can give a result of over-fitting [32]. In other words, it can believe the training set completely which will give an abysmal performance on testing set.
- ID3 and C4.5 decision tree algorithms require discrete values as input data.

## 5.6 Stochastic Gradient Descent Classifier (SGD)

Stochastic Gradient Descent recently got became more popular because of its large-scale learning ability in machine learning problems [11]. It is a useful and straightforward way approach of linear classifiers under convex problems which is Support Vector Machines or Conditional Random Fields [3]. The originality of SGD derives from “Stochastic Approximation” which is a work from Robinson and Monroe [5].

### Advantages of Stochastic Gradient Descent:

- One of the advantages of stochastic gradient descent is, it is easy to implement [38].

- Stochastic Gradient Descent is also efficient because of each step only relies on a single derivative which makes the computational cost  $1/n$  than normal gradient descent [37].

### Disadvantages of Stochastic Gradient Descent:

- Stochastic Gradient Descent can be required to have many iterations, and it also requires some hyper-parameters [38].
- Feature scaling is a practice which is used in the standardization of range of independent variables [47]. SGD also used this feature scaling technique and it can be sensitive to feature scaling [38].
- Another drawback of Stochastic Gradient Descent is while using GPU they are hard to parallelize or distributing them using computer clusters [25].

## 5.7 Support Vector Machines

Support Vector Machines is fallen under the classification methods in machine learning [6]. It is also a robust classification method that has been widely found itself an area ranging from pattern recognition to text analysis [6].

Fitting a boundary between data points is the principle of the support vector machines. This boundary divides the data points between classes, and each similar data point puts under the same class classification [6]. After training the support vector machines with training data-set, we only need to check whether the test data lies under the boundaries for testing set. Another thing to consider is after it creates the boundaries of the data remaining training data becomes obsolete because we only need the core set of points which supports the boundaries to classify the new data set. This core data points called “support vectors”. It is called vector because of each data point contains a row of observed data values for attributes [6].

[Figure 7 about here.]

Traditionally boundaries are called “hyperplanes” and it is used to describe boundaries in more than three dimensions because they are hard or sometimes impossible to visualize [7]. Figure 7. Optimality of hyperplane expressed as a linear function which requires maximum distance between the identified classes. It only considers a small number of training examples to build this hyperplane. SVM hyperplanes based on “separation of positive (+1) and negative (-1) with the largest margin [39]“.

One of the main characteristic of the machine learning is to generalize. In other words, we want to give a general idea that tends to fit any of our testing datasets optimally. Support vector machines are a perfect regarding generalizations because once the training data fitted by the support vector machines other than support vector data inside the training data becomes redundant which means that even with the small changes inside the data will not have a significant effect on general boundaries [6].

### Advantages of Support Vector Machines:

- Generalizes the data well with the help of boundaries. Which reduces the overfitting [6].
- Classification accuracy in basic support vector machine will yield a 95 percent accuracy with a default settings [6].
- SVM can deliver a unique solution, because of optimality solution is convex. This will give an advantage over Neural

Networks which has multiple solutions in local minima [1].

#### **Disadvantages of Support Vector Machines:**

- One common disadvantage of SVM, is the lack of transparency because of its non-parametric techniques [1].
- Another biggest disadvantage of SVM is it requires high algorithmic complexity and high level of memory for the large-scale implementations [39].
- According to Burges, biggest limitation of the SVM is in the choice of kernel [4].

## **5.8 Ensemble Methods**

Ensemble methods goes into classification algorithm category, they are learning algorithms which uses weighted vote for it is prediction methods, in other words, it is learning rules over a small subset of data then we combine these rules which we learn from the small subset of data to make predictions and/or classification on the testing data [8]. The originality of the Ensemble method comes from Bayesian averaging, but with the recent algorithms include “Bagging, error-correcting, and boosting [8]“.

Bagging refers to simply the looking at data-sets and dividing the data-set to it is small subsets then learning the rules of that particular small subset. Next step is combining each learned rule from subsets to apply to more significant data set. Combining method mostly done with averaging the learned rules. Bagging also does better on testing set than standard Linear Regression analysis and linear regression does better on training set especially in third order polynomial [8].

#### **Stacking**

Boosting is another method used in Ensemble Methods. The difference from bagging is in boosting we need to pick subsets or examples that we are not good at in other words hardest examples. Then we combine these learned rules with the weighted mean instead mean used in bagging method.

Boosting is little different then bagging.

#### **Advantages of Ensemble Methods:**

- Prediction of the ensemble methods is better than most of the algorithms because of the combining methods intuition makes the model less noisy [36].
- They are more stable than other algorithms. [36]

#### **Disadvantages of Ensemble Methods:**

- Over-fitting may cause some disadvantages for ensemble learning but bagging operation will reduce this overfitting [36].

## **6 FITTING DATA INTO MACHINE LEARNING ALGORITHMS**

In this section, we will show the techniques we used on the execution of the prepared data into machine learning algorithms. Before fitting the data into the machine learning algorithms, we split the data into two sets. These sets are the training set and the testing set. We do splitting because of gaining an access of the future data will most likely be hard before future occurs, and because of this fact, it is a good idea to test our model with a dataset which our model has not seen it [40].

We used scikit-learn for splitting data into train and test we saved 20% of data for testing purposes as shown in Table 8 .

[Table 8 about here.]

Furthermore, after splitting the data we put all of our training data into to each of the machine learning algorithm to get their prediction results. We also provided code at the beginning and the end of each algorithm to calculate their running time.

Before we move further we need to discuss critical characteristics of a machine learning algorithm. These are;

- Confusion Matrix
- Accuracy
- Recall
- F-1 Score
- Precision

**6.0.1 Confusion Matrix:** Confusion matrix develops from 4 key elements. These elements are true positive, true negative, false negative, and false positive. As shown in Figure 8 about the constructing a confusion matrix. If we want to build a confusion matrix by targeting individuals who are making more than \$50K our true positive, true negative, false positive, and false negative explained below.

[Figure 8 about here.]

**True Positive (TP):** We can explain true positive as if the individuals make more than \$50K and our model correctly classifies them as individuals who makes more than \$50K, then this individual is in higher income range, in this case, we call it a true positive [20].

**True Negative (TN):** Intuition of true negative is if an individual makes less than \$50K and our model correctly classifies them as individuals who makes less than \$50K, then this individual is in lower income range. We call this true negative [20].

**False Negative (FN):** When an individual makes less than \$50K and our model incorrectly classifies them in higher income range by making a mistake causes a false negative to happen [20].

**False Positive (FP):** When an individual is making more than \$50K and our model classifies them in lower income range by mistake. This is called false positive [20].

**6.0.2 Accuracy:** Accuracy answers the question of how good is the model is. In our case this question will be out of all the individuals, how many did the models classify the individuals correctly. The mathematical expression of the accuracy is the ratio between the number of correctly classified points and the number of total points. We can think that if we have high accuracy, our model is excellent, but this is only where we have identical false positive and false negative values in our dataset [20].

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \quad (3)$$

**6.0.3 Precision.** Precision answers the questions of out of all the points predicted to be positive how many of them were actually positive? If we translate this question into our case, we will have out of all the individuals that we are classified as lower income how many were actually have lower income. Higher precision indicates that we have low false positive rate [20]. Mathematical expression of precision is;

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

**6.0.4 Recall (Sensitivity).** Recall answers the question of “out of the points that are labeled positive how many of them were correctly predicted is positive ? ”. If we translate this to our case, we will have “out of the points that are labeled higher income how many of them correctly predicted is in higher income range ? ”. Mathematical expression of the recall is;

$$Precision = \frac{TP}{TP + FN} \quad (5)$$

**6.0.5 F-1 Score.** The F-1 score is the idea of giving a decision by looking at only one score which will include precision, and recall scores. We cannot just take the average of precision and recall because if either of them is very low. We need a number to be low, even if the other one is not. This will lead us to look at the harmonic mean, and it works as follow. Let’s say we have two numbers X and Y. X is smaller than Y, and we have the arithmetic mean, and it always lies between X and Y. It is a mathematical fact that the harmonic mean is always less than the arithmetic mean which is closer to the smaller number than to the higher number. Mathematical expression of F-1 score is;

$$F1Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (6)$$

## 6.1 Results

Now we can look at the results from each of the machine learning algorithm. Results also showed in Table 9 with the visualization of Figure 10. We can also see the running time of the each of the algorithm in Figure 9. Support Vector Machines is the winner for the highest running time for training the algorithm.

[Figure 9 about here.]

[Table 9 about here.]

**6.1.1 Naive Bayes.** As shown in the Figure 10 we have a comparison of several supervised machine learning algorithms on our dataset. We can see that from the accuracy standpoint Naive Bayes algorithms have the lowest score which means that it did not do a good job for labeling true positives regards to all data but it did a good job in precision standpoint while doing a bad classification from recall standpoint. Two key element for us in this situation is accuracy and f1 score(which consist of precision and recall).

**6.1.2 Support Vector Machine.** Support Vector Machine is the second best algorithm in our case. This algorithm did very well job on classification it has the second highest accuracy and f1 score.

**6.1.3 AdaBoost.** As we stated before ensemble algorithms learn from the small portion of the data and combine these learning to do the predictive task. As shown in Figure 10 adaboosting has the highest accuracy score among all the other algorithms. This algorithm should be our first choice to do predictive modeling. We believe that there is still an improvements on accuracy

**6.1.4 K-Nearest Neighbors.** K-Nearest Neighbor algorithm in our project we set the k value to 5. K Nearest Neighbor algorithm also did a good job by placing itself third in accuracy score.

**6.1.5 Decision Tree.** Decision Tree is gave a good accuracy but fall behind on f1 score as shown in Figure 10.

[Figure 10 about here.]

## 7 CONCLUSION

We presented the importance of analytical approach with machine learning algorithms and how they can be used to predict or classify the individuals with many different attributes like age, education, income, etc. We also presented weaknesses and strengths of these algorithms along with their precision, accuracy, recall, and F-1 scores by presenting with the visualizations. We also demonstrated the running time for each algorithm while using big data sets. The source code of this project can found Github website which presented in reference section [44].

## ACKNOWLEDGMENTS

The author would like to thank Dr. Gregor von Laszewski for his support and suggestions to write this paper.

## REFERENCES

- [1] L. Auria and A. R. Moro. 2008. Support Vector Machines (SVM) as a Technique for Solvency Analysis. Online. [http://www.diw.de/english/products/publications/discussion\\_papers/27539.html](http://www.diw.de/english/products/publications/discussion_papers/27539.html)
- [2] L. Ben. 2015. Six Reasons why I recommend scikit-learn. Online. (Oct. 2015). <https://www.oreilly.com/ideas/six-reasons-why-i-recommend-scikit-learn>
- [3] L. Bottou. 2010. Stochastic Gradient Descent. Online. (2010). <http://leon.bottou.org/projects/sgd>
- [4] C. J. C. Burges. 1998. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery* 2, 2 (01 Jun 1998), 121–167. <https://doi.org/10.1023/A:1009715923555>
- [5] N. Deanna, S. Nathan, and W. Rachel. 2016. Stochastic gradient descent, weighted sampling, and the randomized Kaczmarz algorithm. *Mathematical Programming* 155, 1 (01 Jan 2016), 549–573. <https://doi.org/10.1007/s10107-015-0864-7>
- [6] B. Deshpande. 2013. When do support vector machines trump other classification methods. Online. (Jan. 2013). <http://www.simafore.com/blog/bid/112816/When-do-support-vector-machines-trump-other-classification-methods>
- [7] B. Desphande. 2011. 4 key advantages of using decision trees for predictive analytics. Online. (July 2011). <http://www.simafore.com/blog/bid/62333/4-key-advantages-of-using-decision-trees-for-predictive-analytics>
- [8] G. T. Dietterich. n.d. Ensemble Methods in Machine Learning. (n.d.). <http://web.engr.oregonstate.edu/~tgd/publications/mcs-ensembles.pdf>
- [9] EliteDataScience. 2016. Modern Machine Learning Algorithms: Strengths and Weaknesses. Online. (May 2016). <https://elitedatascience.com/machine-learning-algorithms>
- [10] O. F. Ertugrul and M. E. Tagluk. 2017. A novel version of k nearest neighbor: Dependent nearest neighbor. *Applied Soft Computing* 55, Supplement C (2017), 480 – 490. <https://doi.org/10.1016/j.asoc.2017.02.020>
- [11] M. Fan. n.d. How and Why to Use Stochastic Gradient Descent? (n.d.). <http://anson.ucdavis.edu/~minjay/SGD.pdf>
- [12] J. Fang. 2013. Why Logistic Regression Analyses Are More Reliable Than Multiple Regression Analyses. *Journal of Business and Economics* 4, 7 (July 2013), 620–633. <http://www.academicstar.us/UploadFile/Picture/2014-6/201461494819669.pdf>
- [13] M. A. Hassan, A. Khalil, S. Kaseb, and M. A. Kassem. 2017. Potential of four different machine-learning algorithms in modeling daily global solar radiation. *Renewable Energy* 111, Supplement C (2017), 52 – 62. <https://doi.org/10.1016/j.renene.2017.03.083>
- [14] S. T. Indra, L. Wikarsa, and R. Turang. 2016. Using logistic regression method to classify tweets into the selected topics. *2016 International Conference on Advanced Computer Science and Information Systems (ICACSIS), Advanced Computer Science and Information Systems (ICACSIS), 2016 International Conference on*, 385–389 (2016), 385. <http://proxyiub.uits.iu.edu/login.aspx?direct=true&db=edssee&AN=edssee.7872727&site=eds-live&scope=site>
- [15] Investopedia. n.d. Correlation Coefficient. Online. (n.d.). <https://www.investopedia.com/terms/c/correlationcoefficient.asp>
- [16] D. S. Jadhav and H. P. Channe. 2014. Comparative Study of K-NN, Naive Bayes and Decision Tree Classification Techniques. *International Journal of Science and Research (IJSR)* 5, 1 (Jan. 2014), 1842–1845. <https://www.ijsr.net/archive/v5i1/NOV153131.pdf>

- [17] B. Jason. 2014. A gentle introduction to Scikit-Learn: Python Machine Learning Library. Online. (April 2014). <https://machinelearningmastery.com/a-gentle-introduction-to-scikit-learn-a-python-machine-learning-library/>
- [18] H. Jeff. 2012. Introduction to Machine Learning. Online. (Jan. 2012). [http://courses.washington.edu/css490/2012/Winter/lecture\\_slides/05b\\_logistic\\_regression.pdf](http://courses.washington.edu/css490/2012/Winter/lecture_slides/05b_logistic_regression.pdf)
- [19] J. Jiaqi and Y. Chung. 2017. Research on K nearest neighbor join for big data. In *2017 IEEE International Conference on Information and Automation (ICIA)*. IEEE, Department of Computer Engineering Wonkwang University Iksan 54538, Korean, 1077–1081. <https://doi.org/10.1109/ICInFA.2017.8079062>
- [20] R. Joshi. 2016. Accuracy, Precision, Recall, and F1 Score: Interpretation of Performance Measures. Online. (Sept. 2016). <http://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures>
- [21] R. Kohavi. 1996. Improving the Accuracy of Naive-Bayes Classifiers: A Decision-tree Hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96)*. AAAI Press, Silicon Graphics, Inc, 202–207. <http://dl.acm.org/citation.cfm?id=3001460.3001502>
- [22] R. Kohavi and B. Becker. n.d. Predicting whether income exceeds \$50K/yr based on census data. Online. (n.d.). <https://archive.ics.uci.edu/ml/datasets/Census+Income>
- [23] J. Kunal. 2015. Scikit-Learn in python - The most important Machine Learnig Tool I learnt last year. Online. (Jan. 2015). <https://www.analyticsvidhya.com/blog/2015/01/scikit-learn-python-machine-learning-tool/>
- [24] M. D. Lane. n.d. Log Transformations. Online. (n.d.). <http://onlinestatbook.com/2/transformations/log.html>
- [25] V. Q. Le, J. Ngiam, A. Coates, A. Lahiri, B. Prochnow, and A. Y. Ng. 2011. On optimization methods for deep learning. In *International Conference of Machine Learning*. Stanford University, International Conferenfe of Machine Learning, Stanford University, NA. <https://cs.stanford.edu/~acoates/papers/LeNgiCoaLahProNg11.pdf>
- [26] Pandas Library. n.d.. Dataframe replace. Online. (n.d.). <https://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.replace.html>
- [27] Pandas Library. n.d.. Pandas Dateframe describe. Online. (n.d.). <https://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.describe.html>
- [28] Pandas Py Data Library. n.d.. Pandas for Python. Online. (n.d.). <https://pandas.pydata.org/>
- [29] L. J. Moon. 2017. Fast k-Nearest Neighbor Searching in Static Objects. *Wireless Personal Communications* 93, 1 (01 Mar 2017), 147–160. <https://doi.org/10.1007/s11277-016-3524-1>
- [30] G. Nick. 2014. KNN. Online. (April 2014). <http://www.nickgillian.com/wiki/pmwiki.php/GRT/KNN>
- [31] R. Nick. NA. The Disadvantages of Logistic Regression. Online. (NA). <http://classroom.synonym.com/disadvantages-logistic-regression-8574447.html>
- [32] C. Petri. 2010. Decison Trees. Online. (2010). <http://www.cs.ubbcluj.ro/~gabisa/DocDiplome/DT/DecisionTrees.pdf>
- [33] U. Princeton. NA. Decision Tree Learning. Online. (NA). <http://www.cs.princeton.edu/courses/archive/spr07/cos424/papers/mitchell-decrees.pdf>
- [34] S. A. Raj, L. J. Fernando, and S. Raj. 2017. Predictive Analytics On Political Data. Congress. *World Congress on Computing and Communication Technologies* 10, 1109 (2017), 93–96.
- [35] M. Ray. 2012. Nearest Neighbours: Pros and Cons. Online. (April 2012). <http://www2.cs.man.ac.uk/~raym8/comp37212/main/node264.html>
- [36] S. Ray. 2015. 5 Easy Questions on Ensemble Modeling Everyone Should Know. Online. (Jan. 2015). <https://www.analyticsvidhya.com/blog/2015/09/questions-ensemble-modeling/>
- [37] J. Rie and Z. Tong. 2013. Accelerating Stochastic Gradient Descent using Predictive Variance Reduction. In *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Eds.). Curran Associates, Inc., Rutgers University, New Jersey, USA, 315–323. <http://papers.nips.cc/paper/4937-accelerating-stochastic-gradient-descent-using-predictive-variance-reduction.pdf>
- [38] Scikitlearn. n.d.. Stochastic Gradient Descent. Online. (n.d.).
- [39] K. N. Shrivastava, P. Saurabh, and B. Verma. 2011. An Efficient Approach Parallel Support Vector Machine for Classification of Diabetes Dataset. *International Journal of Computer Applications in Technology* 36, 6 (Dec. 2011), 19–24. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.259.3757&rep=rep1&type=pdf>
- [40] D. Steinberg. 2014. Why Data Scientist Split Data into Train and Test. Online. (March 2014). <https://info.salford-systems.com/blog/bid/337783/Why-Data-Scientists-Split-Data-into-Train-and-Test>
- [41] K. B. Tapan. 2015. Naive Bayes vs Logistic Regression: Theory, Implementation and Experimental Validation. *Inteligencia Artificial, Vol 18, Iss 56, Pp 14-30 (2015)* 1, 56 (2015), 14. <http://proxyiub.uits.iu.edu/login?url=https://search-ebscohost-com.proxyiub.uits.iu.edu/login.aspx?direct=true&db=edsdoj&AN=edsdoj.0e372b34c5d48bc72cd437eede1fd1&site=eds-live&scope=site>
- [42] A. F. Tehrani, W. Cheng, and E. Hullermeier. 2011. Choquistic Regression: Generalizing Logistic Regression Using the Choquet Integral. Online. (July 2011). <https://www-old.cs.uni-paderborn.de/fileadmin/Informatik/eim-i-is/PDFs/Talk.EUSFLAT.11.pdf>
- [43] K. Teknomo. 2017. K-Nearest Neighbor Tutorial. Online. (2017). <http://people.revoledu.com/kardi/tutorial/KNN/Strength%20and%20Weakness.htm>
- [44] E. B. Usifo. 2017. Income Prediction. Github. (Dec. 2017). <https://github.com/bigdata-i523/hid343/tree/master/project>
- [45] R. Vasudev. n.d.. What is One Hot Encoding? do you have to use it ? Online. (Aug. n.d.). <https://hackernoon.com/what-is-one-hot-encoding-why-and-when-do-you-have-to-use-it-e3c6186d008f>
- [46] Wikipedia. 2017. Naive Bayes. Online. (Nov. 2017). [https://en.wikipedia.org/wiki/Naive\\_Bayes\\_classifier](https://en.wikipedia.org/wiki/Naive_Bayes_classifier)
- [47] Wikipedia. NA. Feauture Scaling. Online. (NA). [https://en.wikipedia.org/wiki/Feature\\_scaling](https://en.wikipedia.org/wiki/Feature_scaling)
- [48] Wikipedia. n.d.. Comma Separated Values. Online. (n.d.). [https://en.wikipedia.org/wiki/Comma-separated\\_values](https://en.wikipedia.org/wiki/Comma-separated_values)
- [49] Wikipedia. n.d.. Decision Trees. Online. (n.d.). [https://en.wikipedia.org/wiki/Decision\\_tree](https://en.wikipedia.org/wiki/Decision_tree)
- [50] H. Zhang. 2004. *The Optimality of Naive Bayes*. resreport, University of New Brunswick. <http://www.cs.unb.ca/~hzhang/publications/FLAIRS04ZhangH.pdf>

#### LIST OF FIGURES

1	Scatter Matrix Plot [44].	9
2	Histogram of Capital Gain [44].	10
3	Histogram of Capital Loss [44].	10
4	After Logarithmic Function Applied Histogram of Capital Gain [44].	11
5	Example of Naive Bayes [50].	12
6	Example of Decision Tree Construction[33].	12
7	Example of Shows the Hyperplanes [6].	13
8	Example of Confusion Matrix Construction [20].	13
9	Supervised Learning Algorithm Running Time Results [44].	14
10	Supervised Learning Algorithm Results [44].	15

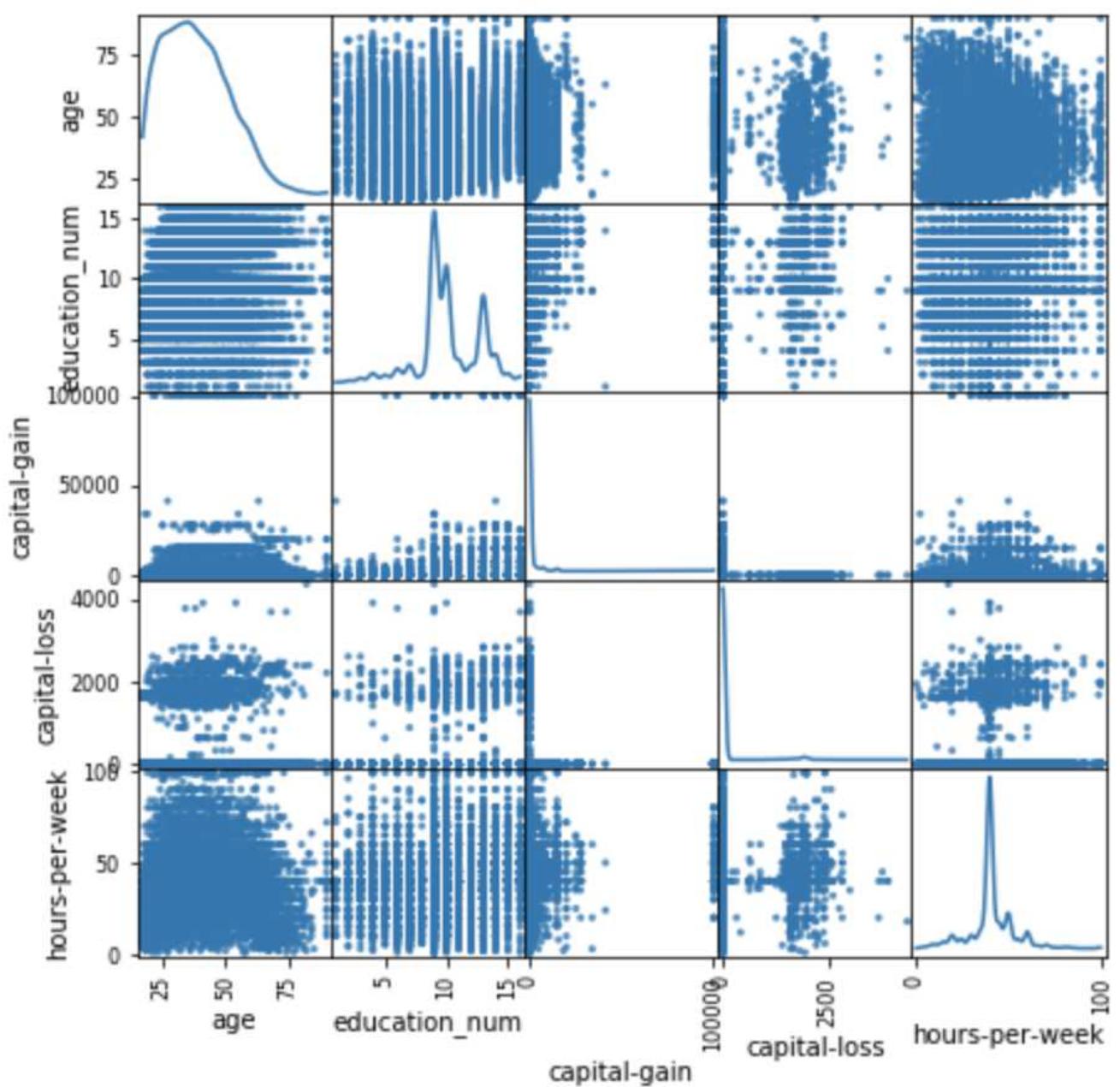


Figure 1: Scatter Matrix Plot [44].

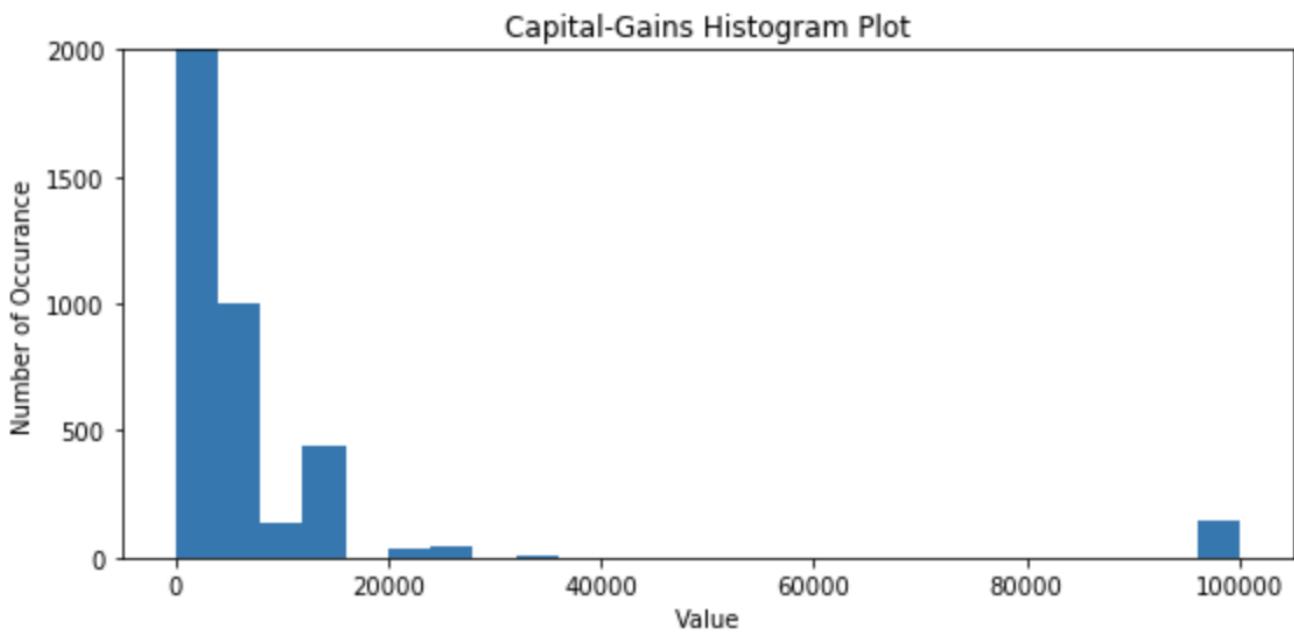


Figure 2: Histogram of Capital Gain [44].

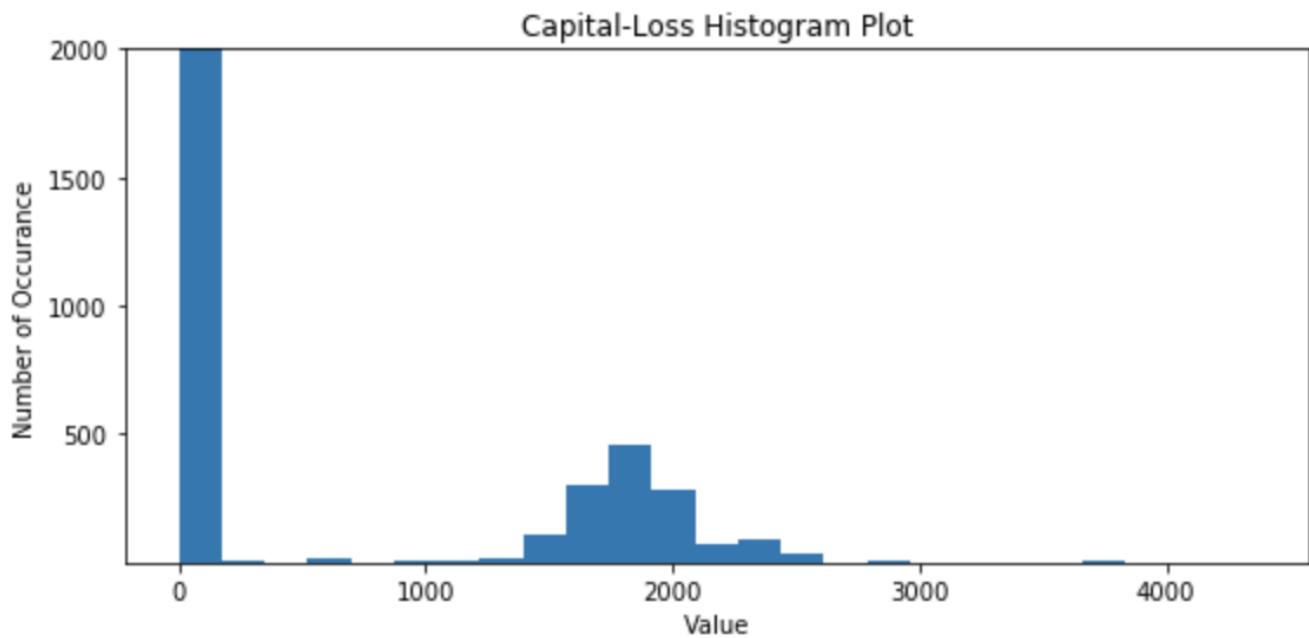


Figure 3: Histogram of Capital Loss [44].

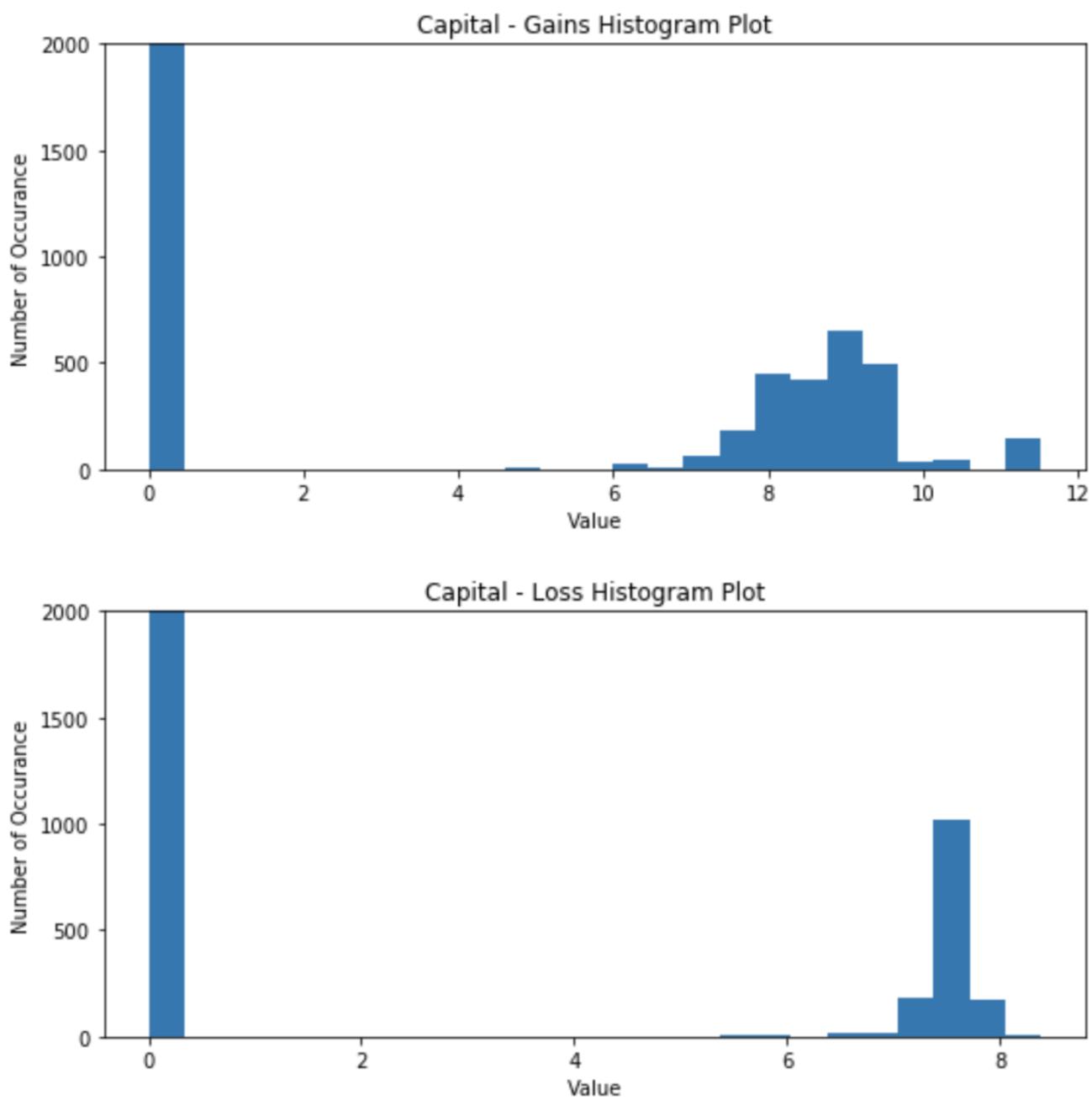


Figure 4: After Logarithmic Function Applied Histogram of Capital Gain [44].

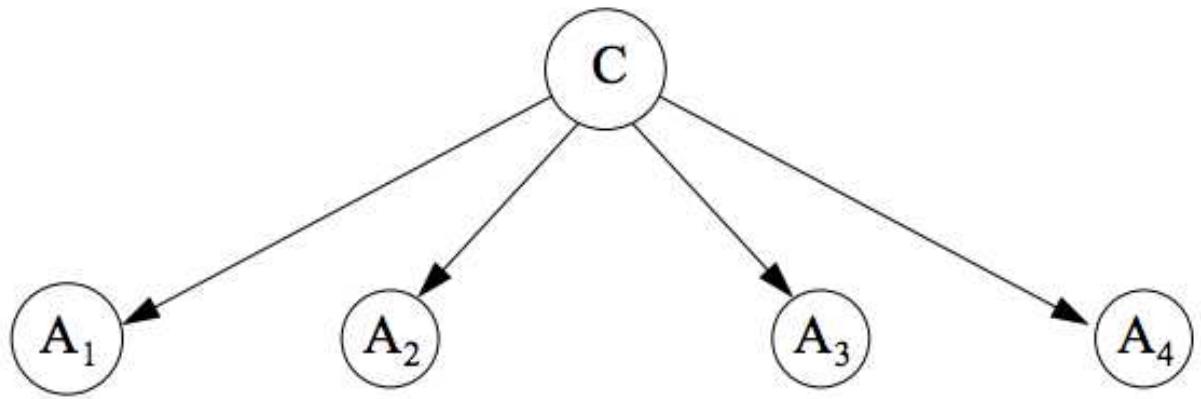


Figure 5: Example of Naive Bayes [50].

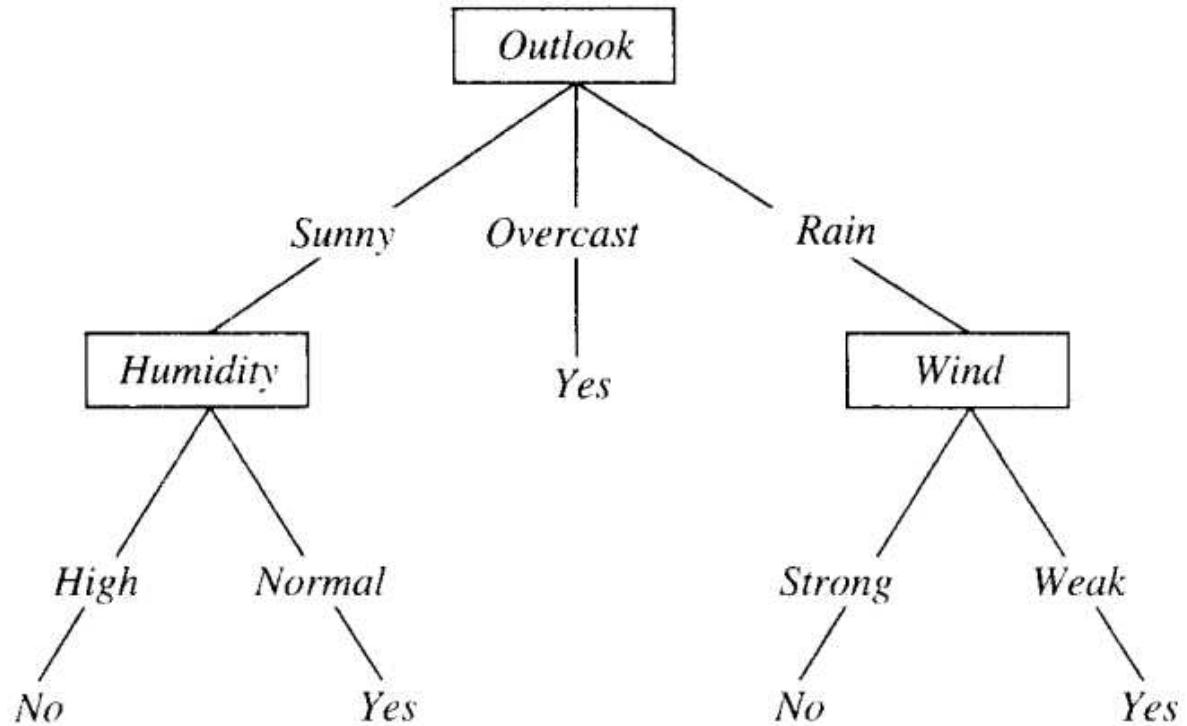


Figure 6: Example of Decision Tree Construction[33].

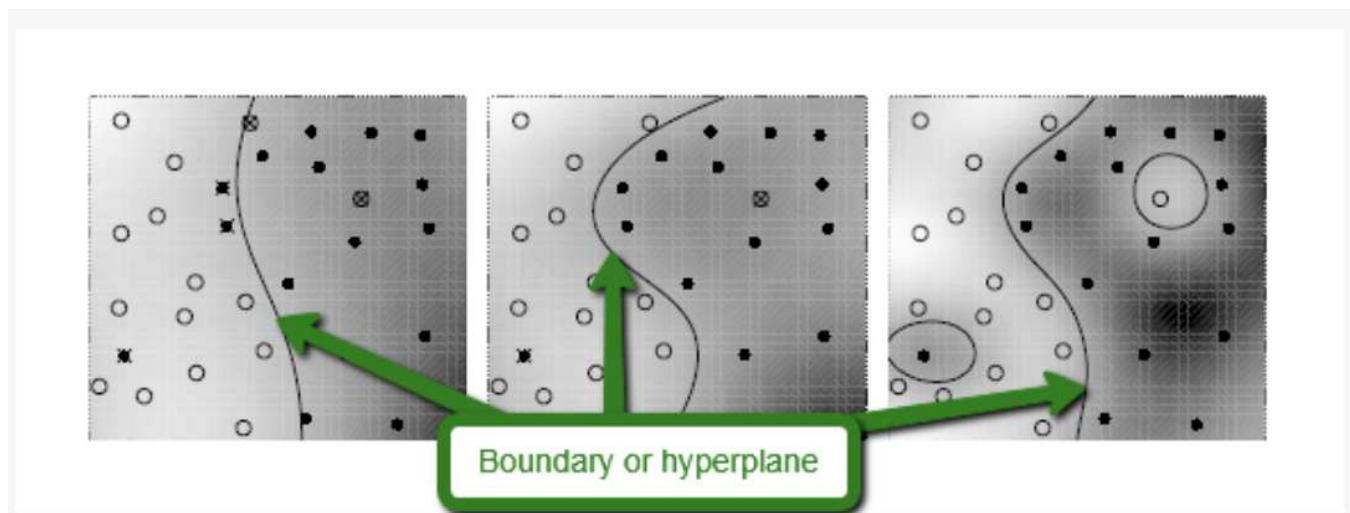
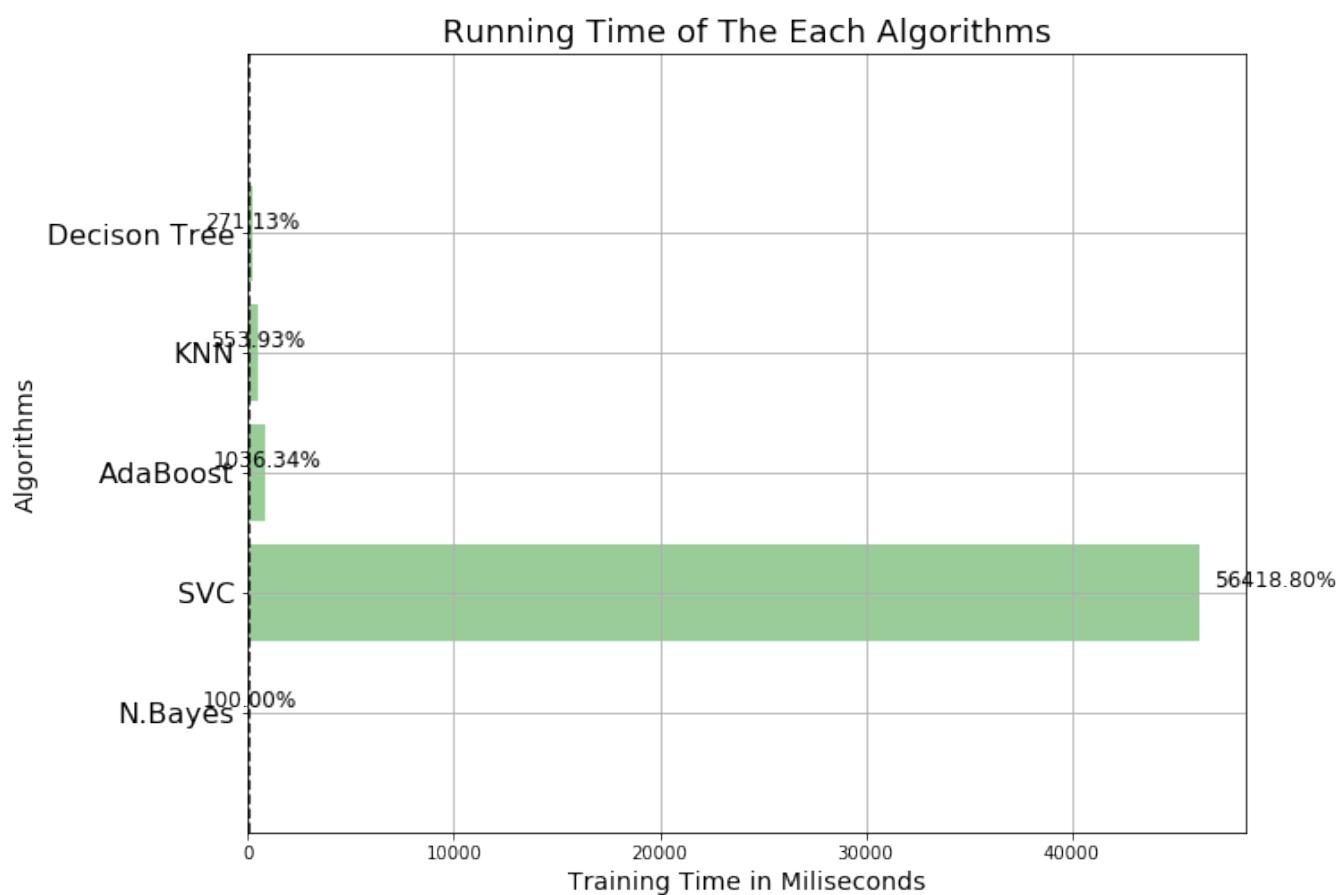


Figure 7: Example of Shows the Hyperplanes [6].

	Predicted class		
Actual Class		Class = Yes	Class = No
	Class = Yes	True Positive	False Negative
	Class = No	False Positive	True Negative

Figure 8: Example of Confusion Matrix Construction [20].



**Figure 9: Supervised Learning Algorithm Running Time Results [44].**

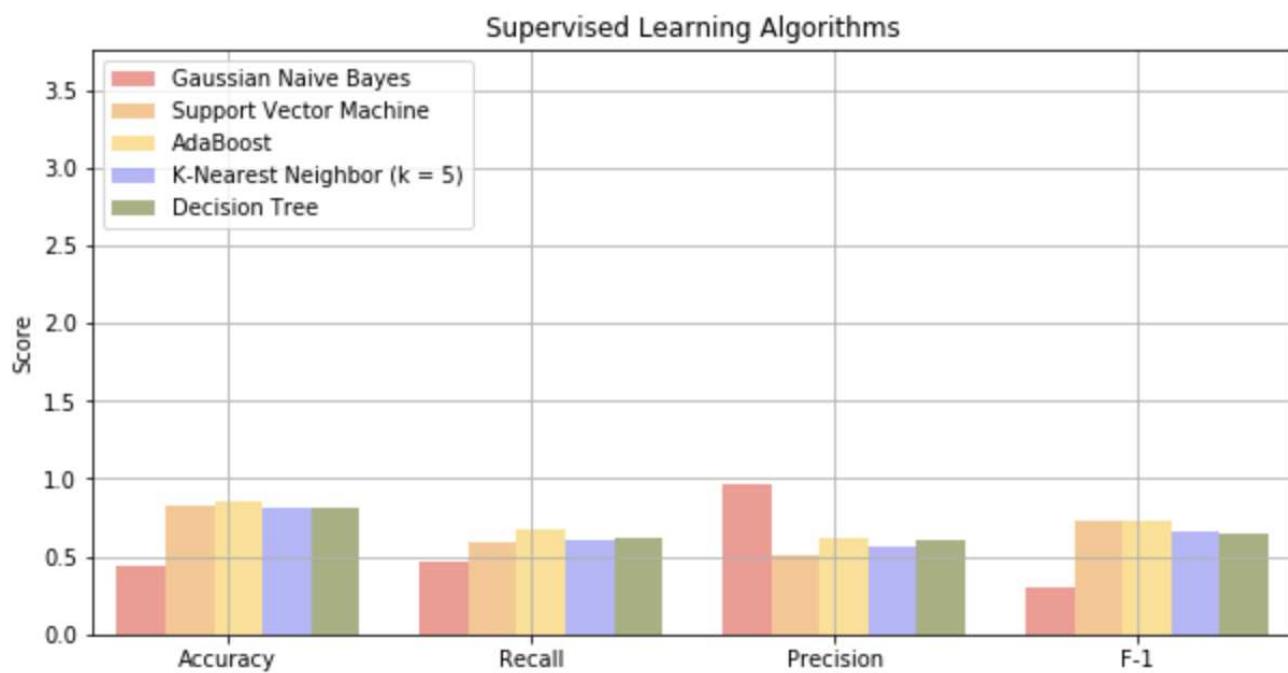


Figure 10: Supervised Learning Algorithm Results [44].

## LIST OF TABLES

1	Statistical Summary of The Continuous Variables	17
2	Description of the Binary Values	17
3	Example of One Hot Encoding Before [45].	17
4	Example of One Hot Encoding After [45].	17
5	Count of Income Variable Regarding to Individuals	17
6	Statistical Summary of Continuous Variables [44].	18
7	Correlation Matrix [44].	18
8	Train-Test-Split [44].	18
9	Results of the Algorithms [44].	18

	<b>age</b>	<b>education</b>	<b>cap gain</b>	<b>cap loss</b>	<b>hours</b>
<b>count</b>	32561	32561	32561	32561	32561
<b>mean</b>	38.581	10.08	1077.64	87.303	40.437
<b>std.</b>	13.640	2.572	7385.292	402.960	12.347
<b>min.</b>	17.0	1.0	0	0	1.0
<b>25%</b>	28.0	9.0	0	0	40.0
<b>50%</b>	37.0	10.0	0	0	40.0
<b>75%</b>	48.0	12.0	0	0	45.0
<b>max</b>	90.0	16.0	0	4356.0	99.0

**Table 1: Statistical Summary of The Continuous Variables**

Description	Assigned Value
Individuals who makes more than \$50K	1
Individuals who makes at or less than \$50K	0

**Table 2: Description of the Binary Values**

<b>Company Name</b>	<b>Categorical Variable</b>	<b>Price</b>
VW	1	2000
Acura	2	10011
Honda	3	50000
Honda	3	10000

**Table 3: Example of One Hot Encoding Before [45].**

VW	Acura	Honda	Price
1	0	0	20,000
0	1	0	10,011
0	0	1	50,000
0	0	1	10,000

**Table 4: Example of One Hot Encoding After [45].**

<b>Description</b>	<b>Count</b>
Total Number of Individuals	30162
Individuals who makes more than \$50K	7508
Individuals who makes at or less than \$50K	22654

**Table 5: Count of Income Variable Regarding to Individuals**

	<b>Age</b>	<b>Gain</b>	<b>Loss</b>	<b>Hours</b>
<b>Number of Instances</b>	32,561	32,561	32,561	32,561
<b>Mean</b>	38.58	1077.64	87.303	40.437
<b>Standard Deviation</b>	13.640	7385.292	402.960	12.347
<b>Minimum Value</b>	17	0	0	1
<b>25th percentile</b>	28	0	0	40
<b>50th percentile</b>	37	0	0	40
<b>75th percentile</b>	48	0	0	45
<b>Maximum Values</b>	90	99999	4356	99

Table 6: Statistical Summary of Continuous Variables [44].

	<b>Age</b>	<b>Education</b>	<b>Capital Gain</b>	<b>Capital Loss</b>	<b>Hours Per Week</b>
<b>Age</b>	1.0	0.043	0.080	0.060	0.101
<b>Education</b>	0.043	1.0	0.124	0.079	0.152
<b>Capital Gain</b>	0.080	0.124	1.0	-0.032	0.080
<b>Capital Loss</b>	0.060	0.796	-0.032	1.0	0.052
<b>Hours Per Week</b>	0.101	0.152	0.080	0.052	1.0

Table 7: Correlation Matrix [44].

<b>Splitting the Data</b>	<b>Sample Size</b>
Training	24129
Testing	6033

Table 8: Train-Test-Split [44].

Name	Accuracy	Recall	Precision	F1 Score
Naive Bayes	0.4442	0.4642	0.9680	0.3053
SVC	0.8301	0.5969	0.5056	0.7284
AdaBoost	0.8499	0.6724	0.6189	0.7361
KNN	0.8184	0.6090	0.5682	0.6561
Decision Tree	0.8161	0.6231	0.6109	0.6459

Table 9: Results of the Algorithms [44].

```
bibtext report
```

```
=====
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
```

```
=====
bibtext _ label error
```

```
=====
report.bib:518:@incollection{NIPS2013_4937,
```

```
=====
bibtext space label error
```

```
=====
report.bib:172:@INPROCEEDINGS{knn-chung,
```

```
=====
bibtext comma label error
```

```
=====
latex report
```

```
=====
[2017-12-11 13.31.29] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflatex)
bookmark level for unknown defaults to 0.
The anchor of a bookmark and its parent's must not be the same. Added a new anchor.
The anchor of a bookmark and its parent's must not be the same. Added a new anchor.
Typesetting of "report.tex" completed in 1.9s.
```

```
./README.yml
```

```
8:81      error    line too long (188 > 80 characters)  (line-length)
8:188     error    trailing spaces  (trailing-spaces)
21:81     error    line too long (534 > 80 characters)  (line-length)
34:81     error    line too long (666 > 80 characters)  (line-length)
34:666    error    trailing spaces  (trailing-spaces)
35:12     error    trailing spaces  (trailing-spaces)
37:30     error    trailing spaces  (trailing-spaces)
42:5      error    duplication of key "type" in mapping  (key-duplicates)
```

```
=====
Compliance Report
```

```
name: Usifo, Borga
hid: 343
paper1: 100 %
paper2: 100 %
project: 100 %
```

```
yamlcheck
```

---

```
wordcount
```

---

```
18
wc 343 project 18 5794 report.tex
wc 343 project 18 6335 report.pdf
wc 343 project 18 3252 report.bib
```

```
find "
```

---

```
passed: True
```

```
find footnote
```

---

```
passed: True
```

```
find input{format/i523}
```

---

```
4: \input{format/i523}
```

```
passed: True
```

```
find input{format/final}
```

---

```
passed: False
```

```
floats
```

---

```
47: We first used the pandas \cite{www-pandas} to help to load the
```

data in data frame format. This gave us a unique advantage, and faster processing of comma separated values for putting into data frame \cite{www-commasep}. Our data consist of 15 variables. Some of these variables are continuous, and some of them are categorical variables, and our target variable was “income” attribute. After putting the data into data frames, we first got a statistical snapshot of continuous variables ( age, education, capital gain, capital loss, hours worked) by using the pandas \cite{www-pandas.describe} functions as shown in Table \ref{stats-table}.

```

49: \begin{table}[!ht]
64: \label{stats-table}
76: \par Our last process is changing the string values to numerical values on our target data which consist of string values (''\$ 50K'') for machine learning algorithms to understand this target data we need to transfer it to numerical values. Since we have only two categories, we will assign 1 and 0 as numerical values as shown in Table \ref{assign-values}.
78: \begin{table}[!ht]
87: \label{assign-values}
90: \par Our shape of the data will also receive impact from changing to numerical. Our number of futures will go from 14 to 103. This is because we implemented one-hot-encode to our dataset. It is called one hot encoded because we transform the categorical variables into a more acceptable shape for the machine learning algorithms to perform well \cite{www-hackernoon}. In other words ‘‘we implement binarization of the category to include as a future to train model \cite{www-hackernoon}’’. As we can see in Table \ref{one-hot-before} and Table \ref{one-hot-after}.
93: \begin{table}[!ht]
104: \label{one-hot-before}
108: \begin{table}[!ht]
119: \label{one-hot-after}
125: After cleaning the data, we started our data exploration to learn little bit more from our data and make necessary changes if needed before putting into our machine learning algorithms. The first step in this process is getting the total count of the individuals as well as the count of the individuals who are making more than \$50K and less than \$50K which can be seen in below Table \ref{my-label-2}.
127: \begin{table}[!ht]
137: \label{my-label-2}
141: \par Moreover, we also look at the statistical values of each of the continuous variable we have. Those values given in Table \ref{my-label}. As we can see we have individuals who’re age ranging from 17 to 90 years old with a mean of 38.58. If we look
  
```

at the capital gains and capital losses, we have a standard deviation of 7385 and 402 respectively this is also another indication of skew in these variables.

```
143: \begin{table}![ht]
158: \label{my-label}
161: \par We used scatter matrix plot and applied the correlation function to see if we have any reliable correlation between any of the variables. As we can see from the correlation matrix Table \ref{scatter-matrix} and correlation numbers Figure \ref{fig:scatter} we do not have the high correlation between any variables. Correlation values range between -1 to 1. The correlation value of 1 is an indication of perfect positive correlation and correlation number -1 indicates a negative correlation between variables \cite{www-investopedia}. Because of lower correlation values, it will be tough to determine the classification by just looking at the correlations; this indicates we have sophisticated algorithms to determine the relationship between variables to classify individuals incomes.
163: \begin{table}![ht]
176: \label{scatter-matrix}
180: \begin{figure}![ht]
182: \includegraphics[width=\columnwidth]{images/scatter-matrix.png}
183: \caption{Scatter Matrix Plot
\cite{Borga2017}.}\label{fig:scatter}
186: \par Furthermore, we also explore the capital gains, capital losses, and hours per week variables which we used a histogram to plot the data into distribution form so we can see how all these attributes distributed. The reason we do the histogram is we want to see any skewness in our data. As shown in the histogram graphs in Figure \ref{fig:Hist-capital} and Figure \ref{fig:loss-capital} in capital gains and capital loss we have highly skewed data which can cause issues later on in our algorithms. We apply a logarithmic function to do highly skewed data to less skewed \cite{www-onlinestat}. Using logarithmic functions adds more value to data from the interpretable standpoint and ‘‘it helps to meet the assumptions of inferential statistics \cite{www-onlinestat}’’.
188: \begin{figure}![ht]
190: \includegraphics[width=\columnwidth]{images/capital-gain.png}
191: \caption{Histogram of Capital Gain
\cite{Borga2017}.}\label{fig:Hist-capital}
194: \begin{figure}![ht]
196: \includegraphics[width=\columnwidth]{images/capital-loss.png}
197: \caption{Histogram of Capital Loss
\cite{Borga2017}.}\label{fig:loss-capital}
200: \par Moreover, applying logarithmic function had an impact on
```

distribution. We can see the changes on skew data in Figure \ref{fig:Hist-capital-log} after applying logarithmic function.

202: \begin{figure}[!ht]

204: \includegraphics[width=\columnwidth]{images/logarithmic-applied.png}

205: \caption{After Logarithmic Function Applied Histogram of Capital Gain \cite{Borga2017}.}\label{fig:Hist-capital-log}

242: \par Term naive comes from it is the method of use probability among categories which assumes of independence among given class of attributes as shown in Figure \ref{fig:Naive Bayes}. In other words, if we try to classify individuals from their email communications it will not take the order of words into account. Whereas in the English language we can tell the difference between sentence makes sense or not if we randomly re-order our words in the sentences. So it does not understand the text, it only looks at word frequencies as a way to do the classification. This is why it is called ‘‘Naive’’.

244: \begin{figure}[!ht]

247: \includegraphics[width=\columnwidth]{Naive-bayes}

248: \caption{Example of Naive Bayes \cite{Zhang}.}\label{fig:Naive Bayes}

333: \par Representation of decision trees is as follows; we sort the instances from root to leaf nodes, this sorting gives insights about the classification of the instance, every outcome descending from the root node corresponds to possible values for that variable \cite{www-cs.princeton}. We can classify an instance by starting from the root node and checking the attributes labeled on that node and moving down from that node based on attribute given attribute values \cite{www-cs.princeton} as shown in Figure \ref{fig:Decision Tree}.

335: \begin{figure}[!ht]

337: \includegraphics[width=\columnwidth]{images/decison\_tree.png}

338: \caption{Example of Decision Tree Construction\cite{www-cs.princeton}.}\label{fig:Decision Tree}

386: \begin{figure}[!ht]

388: \includegraphics[width=\columnwidth]{images/hyperplane-boundary.png}

389: \caption{Example of Shows the Hyperplanes \cite{www-simafore-svm}.}\label{fig:Hyperplane}

392: \par Traditionally boundaries are called ‘‘hyperplanes’’ and it is used to describe boundaries in more than three dimensions because they are hard or sometimes impossible to visualize.\cite{www-simafore}. Figure \ref{fig:Hyperplane}. Optimality of hyperplane expressed as a linear function which requires maximum distance between the identified classes. It only considers a small number of training example to build this

hyperplane. SVM hyperplanes based on “ separation of positive (+1) and negative (-1) with the largest margin \cite{verma-ssv}”.

437: \par We used scikit-learn for splitting data into train and test we saved 20\% of data for testing purposes as shown in Table \ref{split} .

439: \begin{table}[!ht]

448: \label{split}

464: Confusion matrix develops from 4 key elements. These elements are true positive, true negative, false negative, and false positive. As shown in Figure \ref{fig:confusion-matrix} about the constructing a confusion matrix. If we want to build a confusion matrix by targeting individuals who are making more than \\$50K our true positive, true negative, false positive, and false negative explained below.

466: \begin{figure}[!ht]

468: \includegraphics[width=\columnwidth]{images/confusion-matrix.png}

469: \caption{Example of Confusion Matrix Construction \cite{www-exsilio}.}\label{fig:confusion-matrix}

511: Now we can look at the results from each of the machine learning algorithm. Results also showed in Table \ref{result-table} with the visualization of Figure \ref{fig:result-algo}. We can also see the running time of the each of the algorithm in Figure \ref{fig:result-time}. Support Vector Machines is the winner for the highest running time for training the algorithm.

513: \begin{figure}[!ht]

515: \includegraphics[width=\columnwidth]{images/running-time.png}

516: \caption{Supervised Learning Algorithm Running Time Results \cite{Borga2017}.}\label{fig:result-time}

519: \begin{table}[!ht]

531: \label{result-table}

535: As shown in the Figure \ref{fig:result-algo} we have a comparison of several supervised machine learning algorithms on our dataset. We can see that from the accuracy standpoint Naive Bayes algorithms have the lowest score which means that it did not do a good job for labeling true positives regards to all data but it did a good job in precision standpoint while doing a bad classification from recall standpoint. Two key element for us in this situation is accuracy and f1 score(which consist of precision and recall).

539: As we stated before ensemble algorithms learn from the small portion of the data and combine these learning to do the predictive task. As shown in Figure \ref{fig:result-algo} adaboosting has the highest accuracy score among all the other algorithms. This algorithm should be our first choice to do predictive modeling. We believe that there is still an

```
improvements on accuracy
544: Decision Tree is gave a good accuracy but fall behind on f1 score
      as shown in Figure \ref{fig:result-algo}.
551: \begin{figure}[!ht]
553: \includegraphics[width=\columnwidth]{images/result-score.png}
554: \caption{Supervised Learning Algorithm Results
      \cite{Borga2017}.}\label{fig:result-algo}
```

```
figures 10
tables 9
\includegraphics 10
labels 19
refs 17
floats 19
```

```
True : ref check passed: (refs >= figures + tables)
True : label check passed: (refs >= figures + tables)
True : include graphics passed: (figures >= \includegraphics)
False : check if all figures are referred to: (refs >= labels)
```

```
Label/ref check
passed: True
```

```
When using figures use columnwidth
[width=1.0\columnwidth]
do not change the number to a smaller fraction
```

```
find textwidth
```

---

```
passed: True
```

---

```
below_check
```

```
WARNING: algorithm and below may be used improperly
```

```
125: After cleaning the data, we started our data exploration to learn
      little bit more from our data and make necessary changes if
      needed before putting into our machine learning algorithms. The
      first step in this process is getting the total count of the
      individuals as well as the count of the individuals who are
      making more than \$50K and less than \$50K which can be seen in
      below Table \ref{my-label-2}.
```

WARNING: code and below may be used improperly

218: Scikit-learn developed by David Cournapeau in 2007. The development came from while he was working on summer code project for Google. After recognized and published by INRIA in 2010 project start the get more attention among worldwide. There are more than 30 active contributors and has secured several sponsorships from big technology companies\cite{www-machinelearningmystery}. ‘‘It also has a goal of providing common algorithms to Python users through consistent interface\cite{www-oreily}’’. Scikit-Learn consists of several elements to make analytical predictions. These elements are shown below\cite{www-analyticvidhya}:

WARNING: algorithm and below may be used improperly

218: Scikit-learn developed by David Cournapeau in 2007. The development came from while he was working on summer code project for Google. After recognized and published by INRIA in 2010 project start the get more attention among worldwide. There are more than 30 active contributors and has secured several sponsorships from big technology companies\cite{www-machinelearningmystery}. ‘‘It also has a goal of providing common algorithms to Python users through consistent interface\cite{www-oreily}’’. Scikit-Learn consists of several elements to make analytical predictions. These elements are shown below\cite{www-analyticvidhya}:

WARNING: algorithm and below may be used improperly

228: \par Algorithms that we will investigate shown below; we will go more deep analysis on each of these algorithms.

bibtex

---

label errors

518: NIPS2013\_4937: do not use underscore in labels:

bibtex errors

This is BibTeX, Version 0.99d (TeX Live 2016)  
The top-level auxiliary file: report.aux

The style file: ACM-Reference-Format.bst  
Database file #1: report.bib

bibtex\_empty\_fields

---

entries in general should not be empty in bibtex

find ""

---

passed: True

---

ascii

---

=====  
The following tests are optional  
=====

Tip: newlines can often be replaced just by an empty line

find newline

---

passed: True

cites should have a space before \cite{} but not before the {

find cite {

---

passed: True

# New Approaches to Managing Metadata at Scale in Research Libraries

Timothy A. Thompson

Indiana University Bloomington

School of Informatics, Computing, and Engineering

Bloomington, Indiana 47408

timathom@indiana.edu

## ABSTRACT

The analysis of big data often relies on distributed storage and computation; however, access to big data—and to the platforms capable of managing and processing it—continues to be largely centralized. Centralization is particularly evident in the case of the metadata produced, managed, and disseminated by academic and research libraries. Libraries typically create and share their catalog records by uploading them to a central data warehouse, which can then be searched by other libraries for records that can be copied and added to an institution’s local catalog. Centralization has the advantage of scalability and availability, but it comes at the cost of a loss of autonomy. Existing metadata workflows can be optimized through the adoption of entity resolution and machine learning algorithms, including approaches based on neural network models. Although innovation is possible within the current paradigm, it can only reach its full potential in the context of peer-to-peer platforms that would allow libraries to share their data directly.

## KEYWORDS

i523, HID340, Research Libraries, Library Catalogs, Entity Resolution, Neural Networks

## 1 INTRODUCTION

The problem of entity resolution (also known as record linkage or data matching [? ]) is one that has a direct impact on the work of information professionals in research libraries. In library units responsible for catalog management, many workflows center on a procedure known as copy cataloging, which aims to expedite the processing of new acquisitions. Copy cataloging involves searching a shared database for records created by another cataloging agency, but that describe identical publications that have been acquired locally [? ]. In the current environment, a single company, the Online Computer Library Center (OCLC—<http://www.oclc.org>), is the only viable platform for global cooperative cataloging [? ]. OCLC provides data aggregation and warehousing services that allow libraries to effectively share their data, but its business model does not encourage peer-to-peer interaction and innovation among individual libraries. This centralized model, which operates on the basis of membership fees, has the advantage of scalability and availability, but it comes at the cost of a loss of control over the data itself, and it entails the acceptance of a business model that, in effect, charges libraries for serving their own data back to them.

## 2 NEW APPROACHES TO METADATA MANAGEMENT

Libraries have a tradition of experience with record matching and automation [? ], but now stand to benefit from the increasingly mainstream availability of algorithms and routines developed within the context of data science and machine learning. Sophisticated algorithms for string comparison and probabilistic data record linkage have long been available, but are not widely used by libraries, with the exception of large-scale projects such as the Social Networks and Archival Context Project (SNAC) (<http://snaccooperative.org/>) and the Virtual International Authority File (VIAF) (<http://viaf.org/>). The former has employed methods based on Naïve Bayes classification algorithms to aggregate and disambiguate data from across a wide range of libraries and archives. The reported accuracy of this approach fell with the range of 80-90 percent.

### 2.1 Neural Networks for Data Matching

## 3 CONCLUSION

The blockchain-based database BigchainDB (written in Python) provides an alternative, and to benefit from features such as data immutability and an asset-based transactional model. A working prototype installation of a BigchainDB node has the potential to provide an example of how libraries can abandon centralized models for managing their data at scale.

## ACKNOWLEDGMENTS

The author would like to thank Dr. Gregor von Laszewski and the i523 teaching assistants for their support and suggestions in writing this report.

## REFERENCES

```
bibtext report
```

```
=====
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
Warning--I didn't find a database entry for "pC12"
Warning--I didn't find a database entry for "cD17"
Warning--I didn't find a database entry for "aT10"
Warning--I didn't find a database entry for "jM92"
(There were 4 warnings)
```

```
bibtext _ label error
```

```
=====
bibtext space label error
```

```
=====
bibtext comma label error
```

```
=====
latex report
```

```
=====
[2017-12-11 13.31.22] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflatex)
p.1 L31 : [pC12] undefined
p.1 L31 : [cD17] undefined
p.1 L31 : [aT10] undefined
p.1 L34 : [jM92] undefined
Empty 'thebibliography' environment.
There were undefined citations.
Typesetting of "report.tex" completed in 0.8s.
```

```
=====
Compliance Report
=====
```

```
name: Tim Thompson
hid: 340
```

```
paper1: Oct 25 17 100%
paper2: 100% Nov 8 17
project: Dec 7 17 66%
```

```
yamlcheck
```

---

```
wordcount
```

---

```
1
wc 340 project 1 607 report.tex
wc 340 project 1 573 report.pdf
wc 340 project 1 730 report.bib
```

```
find "
```

---

35: Libraries have a tradition of experience with record matching and automation \cite{jM92}, but now stand to benefit from the increasingly mainstream availability of algorithms and routines developed within the context of data science and machine learning. Sophisticated algorithms for string comparison and probabilistic data record linkage have long been available, but are not widely used by libraries, with the exception of large-scale projects such as the Social Networks and Archival Context Project (SNAC) (\url{http://snaccooperative.org/}) and the Virtual International Authority File (VIAF) (\url{http://viaf.org/}). The former has employed methods based on Naive Bayes classification algorithms to aggregate and disambiguate data from across a wide range of libraries and archives. The reported accuracy of this approach fell with the range of 80-90 percent.

```
passed: False
```

```
find footnote
```

---

```
passed: True
```

```
find input{format/i523}
```

---

```
7: \input{format/i523}
```

```
passed: True
```

```
find input{format/final}
```

---

```
passed: False
```

```
floats
```

---

```
figures 0
tables 0
includegraphics 0
labels 0
refs 0
floats 0
```

```
True : ref check passed: (refs >= figures + tables)
```

```
True : label check passed: (refs >= figures + tables)
```

```
True : include graphics passed: (figures >= includegraphics)
```

```
True : check if all figures are refered to: (refs >= labels)
```

```
Label/ref check
```

```
passed: True
```

```
When using figures use columnwidth
```

```
[width=1.0\columnwidth]
```

```
do not change the number to a smaller fraction
```

```
find textwidth
```

---

```
passed: True
```

```
below_check
```

---

```
bibtex
```

---

```
label errors
```

bibtex errors

```
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
Warning--I didn't find a database entry for "pC12"
Warning--I didn't find a database entry for "cD17"
Warning--I didn't find a database entry for "aT10"
Warning--I didn't find a database entry for "jM92"
(There were 4 warnings)
```

bibtex\_empty\_fields

---

entries in general should not be empty in bibtex

find ""

---

passed: True

---

ascii

---

=====

The following tests are optional

=====

Tip: newlines can often be replaced just by an empty line

find newline

---

passed: True

cites should have a space before \cite{} but not before the {

find cite {

---

passed: True

# Big Data Analytics in Detection of DDoS (Distributed Denial-of-Service) attacks

Neha Rawat  
Indiana University  
Bloomington, Indiana  
nrawat@iu.edu

## ABSTRACT

With the increase in internet traffic, threats on the network have also increased. Denial-of-service attacks are cyber attacks wherein a perpetrator, due to any kind of malicious intent, tries to make a resource on the network unavailable to its intended users and carries it out by swamping the system or resource with excess requests in order to overload it and prevent users from accessing it. A much more dangerous variety of such an attack is if it is distributed i.e. coming from various sources. Big Data analytics, however, can be used to detect such attacks by having the ability to store the voluminous logs of such attacks and using the data and machine learning techniques to design an anomaly detection system (using a classification model) to detect and prevent these attacks. This project will aim to explore such classification models, design and train the most optimum model and display its effects using a DDoS network traffic logs dataset.

## KEYWORDS

i523, HID224, Denial-of-Service, Intrusion Detection, KDD Cup'99 dataset, Machine Learning, Apache Spark

## 1 INTRODUCTION

The Internet allows us several comforts and functionalities in our day-to-day lives. With the increasing flexibility and accessibility provided by technology, the Internet has become an indispensable part of our life. However, this same accessibility often provides openings for malicious attackers to enter. Security over the Internet is an interdependent factor, with the security of one user depending on rest of the global network [1]. Denial-of-Service attacks are attacks by such malicious users in order to disrupt the accessibility of other legitimate users to a Web Service or application [7]. The objectives of such attacks are mainly malicious, driven out of revenge or for some material gain. The attacks seriously hinder the productivity of the victim, as the resources available are not sufficient to handle the oncoming flood of requests. This attack increases in complexity when there are multiple sources of attacks, resulting in a Distributed Denial-of-Service attack. “In the case of a Distributed Denial-of-Service (DDoS) attack, an attacker uses multiple sources - which may be compromised or controlled by a group of collaborators - to orchestrate an attack against a target” [7]. A small batch of requests sent by an attacker may be enough to generate a large amount of unwanted traffic. The earliest of these attacks was when a DDoS tool called Trinoo, deployed in at least 227 systems, flooded a University of Minnesota computer, which was subsequently rendered useless for more than two days [1]. Figure 1 shows how a Distributed Denial-of-Service attack occurs.

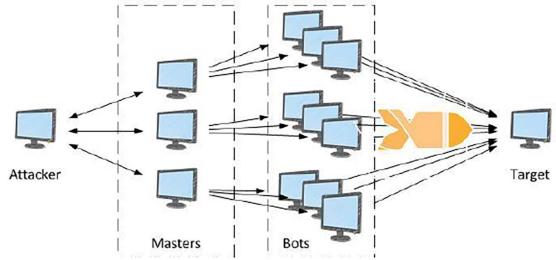


Figure 1: Distributed Denial-of-Service Attack [7]

As the connectivity increases in our everyday lives, so have the risks for DDoS attacks. The Internet of Things (IoT) for example, has opened up a whole new avenue for Denial-of-Service attackers. Earlier, limited to attacks over the Internet which mostly affected a user’s computer, with the advent of IoT, the scope of attacks on other smart devices has increased considerably. These devices could be used as pawns in a DDoS attack network and could even be the intended targets for such an attack. Some of the largest DDoS attacks till date are as given: In March of 2013, the DDoS attack on Spamhaus saw 120 Gbps of traffic on their network, in August of 2013, a “part of the Chinese internet went down in one of the largest DDoS attacks”, in the Spring of 2015, UK-based phone carrier Carphone Warehouse got attacked and hackers stole millions of customers’ data and in January of 2016, some HSBC customers were inhibited from accessing their online banking accounts, which caused a great upheaval as it was “two days before the tax payment deadline in the United Kingdom” [11]. We can see that these attacks, if allowed to happen, have great damage potential. Hence, DDoS mitigation service providers like Imperva Incapsula Enterprise, Arbor Cloud, Verisign, DOSarrest and CloudFlare, have their work cut out for them to detect and prevent such attacks, which are increasing in their reach and complexity [10].

## 2 DDoS ATTACK TYPES AND ARCHITECTURE

In order to prevent a DDoS attack, it is important to know the points in a network where the attack is expected to occur and the type of attack that can occur. Referring to an Open Systems Interconnection(OSI) model, we can usually narrow down the layers which could be affected by a potential attack to the Network, Transport, Presentation and Application layers [7]. Figure 2 shows an Open Systems Interconnection Model with the layers highlighted where DDoS attacks are most common.

#	Layer	Unit	Description	Vector Examples
7	Application	Data	Network process to application	HTTP floods, DNS query floods
6	Presentation	Data	Data representation and encryption	SSL abuse
5	Session	Data	Interhost communication	N/A
4	Transport	Segments	End-to-end connections and reliability	SYN floods
3	Network	Packets	Path determination and logical addressing	UDP reflection attacks
2	Data Link	Frames	Physical addressing	N/A
1	Physical	Bits	Media, signal, and binary transmission	N/A

Figure 2: Open Systems Interconnection Model [7]

Apart from this, the DDoS attacks generally have a specific architecture and follow certain strategies. Knowledge of the pathway which a Denial-of-Service attack follows is essential to detecting and mitigating it.

## 2.1 DDoS Attack Types

The DDoS attacks in the Network and Transport layers are generally of the User Datagram Protocol (UDP) reflection and synchronize (SYN) flood types [7]. The UDP protocol can allow the attacker to fake the source of a request sent to a server and generate a larger response. The amplification factor of a protocol (request to response size) will result in an overwhelming response to a comparatively smaller request. “For example, the amplification factor for DNS can be in the 28 to 54 range - which means an attacker can send a request payload of 64 bytes to a DNS server and generate over 3400 bytes of unwanted traffic” [7]. A SYN flood attack is based on employing all the resources of a system and exhausting them by leaving connections half-open. For example, when an user connects to a TCP service, the client will send a SYN packet and the server will return a SYN-ACK, expecting the client to return an ACK and completing the handshake. In a SYN flood attack, the ACK is not returned and so the server is stuck in this state which prevents other users from connecting to it [7].

In the Presentation and Application layers, the DDoS attacks are slightly different. The most common of such attacks are “HTTP floods, cache-busting attacks, and WordPress XML-RPC floods” [7]. In an HTTP flood attack, the attacker sends HTTP requests under the guise of a real user or web service. These attacks target a resource or try to emulate human behavior. Cache-busting attacks are a specialized version of HTTP flood attacks that use “variations in the query string to circumvent content delivery network (CDN) caching which results in origin fetches, causing additional strain on the origin web server” [7]. A WordPress XML-RPC flood (WordPress pingback flood) is used by an attacker to misuse the XML-RPC API function of a website hosted on WordPress software to generate HTTP flood requests. This type of attack has *WordPress* present in the HTTP request header and so is clearly recognizable [7].

## 2.2 DDoS Attack Architecture

“DDoS attack networks follow two types of architectures: the Agent-Handler architecture and the Internet Relay Chat (IRC)-based architecture” [1]. The components of an Agent-Handler architecture are clients, handlers, and agents. In this type of architecture, the attacker connects with the rest of the attack system at the client point. The handlers are generally software packages available over

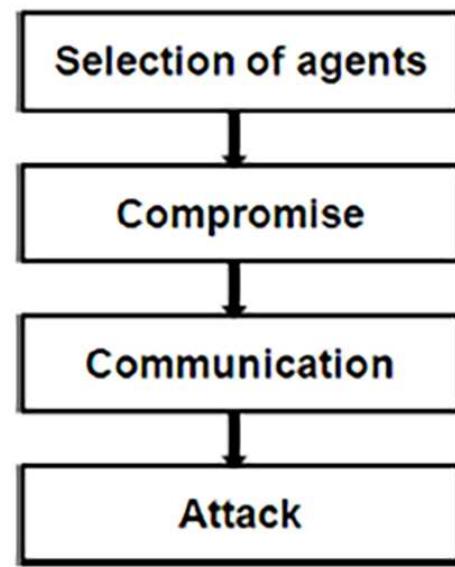
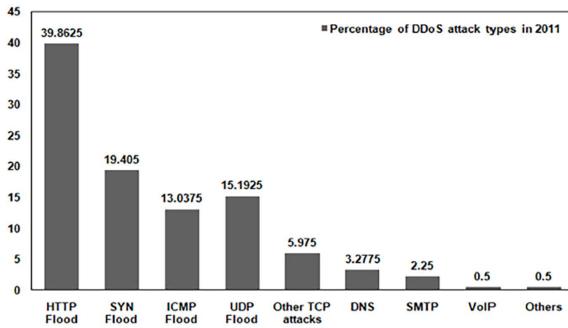


Figure 3: Steps of a Denial-of-Service attack [1]

the Internet which are used by the client to connect to the agents. The agent softwares are placed in the vulnerable systems that are finally used to implement the attack. Often, the users of the agent systems are not aware of the attack being carried out [1]. In the IRC-based architecture, “an IRC communication channel is used to connect the client(s) to the agents” [1]. IRC ports are employed to send commands to the agents, making the DDoS command packets harder to trace (as these channels have a lot of traffic) [1]. When launching a DDoS attack, the attacker goes through some steps common to both types of architectures [1]. First, the attacker tries to identify vulnerable systems that can be used as agents. The resources of these systems are used to generate a powerful attack stream. Next, the attacker plants the handler software code in the compromised system and ensures steps to prevent the code from being detected. These compromised systems are often referred to as *zombies*. Sometimes, the attacker creates several intermediate layers between the *zombies* and the victim to hinder traceability. Thirdly, the attacker communicates with the handler codes placed via protocols like TCP or UDP, and decides the scheduling of the attacks. Post the complete setup, the attacker launches the attack on the victim’s machine or server and renders it unusable [1]. In an IRC-based architecture, most of the above steps remain same, but an IRC-channel is used for communication purposes. This helps the attacker as even if one *zombie* or *bot* is discovered, the identities of the others is still hidden, as IRC-channels are difficult to detect [1]. Figure 3 shows the steps of a Denial-of-Service attack execution.



**Figure 4: Different Denial-of-Service attack type statistics [1]**

### 3 DDoS ATTACK DEFENSE METHODOLOGIES

In the previous section, we explored the common types of DDoS attacks and the general architecture that they follow. These different types of attacks are used with variation by attackers in their attempts to obstruct utilization of resources. Figure 4 shows the percentage of different Denial-of-Service attacks in 2011 by type. The different types of DDoS attacks and their improvement throughout time has also invoked different defense mechanisms against these attacks. DDoS defense mechanisms are usually employed at three points in the attack network : Victim-end, Source-end and Intermediate-Network [1]. Victim-end detection approaches are generally incorporated in the routers of victim networks. A detection system is used to detect intrusion based on different techniques. Detecting DDoS attacks at this point is relatively easy and the most practically applicable, but has the disadvantage of detection only after the attack has reached the victim and legitimate users have already been denied services [1]. Source-end detection system works similarly to the victim-end detection system apart from “a throttling component”, which is added to force a rate limit on outgoing connections. The detection system then compares both incoming and outgoing network traffic with normal traffic benchmarks to detect an attack. This is probably the ideal defense mechanism, but faces challenges in the deployment of a detection system at the source and difficulty in identification in case of multiple sources [1]. The intermediate-network defense mechanism acts like a middle-ground between the victim-end and source-end systems. It acts like a collaborative model which depends upon communication and sharing of information between all routers on the network. Hence, this too suffers from the problem of deployability, as even one router missing on the network could hinder the traceback process [1].

From the above defense mechanism schemes, we can garner that detection of these attacks forms a major part of the preventive process. The most commonly used detection methodologies for defense against DDoS are as follows: Statistical Methods, Soft-Computing and Machine Learning Methods and Knowledge-Based Methods [1].

### 3.1 Statistical Methods

Statistical Methods follow the statistical properties of the distribution of incoming and outgoing network traffic for detection of DDoS attacks. The distributions (or statistical estimates generated using it) are compared with those for a normal traffic signature. An example of the same is the use of cumulative deviation from normal to detect DDoS attacks. Similarly, a periodic deviation analysis from the normal pattern can be used to detect intrusions [1]. Another example, is the use of a two-sample t-test to detect DDoS signatures by comparing the SYN arrival rate distribution with the distribution of a normal SYN arrival rate (after confirming a gaussian distribution for it). If the difference is considered significant according to the t-test, the traffic is marked as potentially containing attack packets [1]. A prediction method designed by Zhang et al. [12] uses an Auto Regressive Integrated Auto Regressive (ARIMA) model for their detection system.

### 3.2 Soft-Computing and Machine Learning Methods

The voluminous network traffic data generated can be leveraged by a soft-computing system like a neural network or a data mining/machine learning model to design a classifier that differentiates between normal traffic and intrusions. An example is the use of statistical preprocessing for extraction of relevant features from the traffic followed by an unsupervised neural net to classify traffic signatures as either a DDoS attack or normal [1]. Another case is the use of a Radial Basis Function (RBF) neural network to analyze attack packets and classify them as normal or harmful [1]. Machine learning algorithms like K-Nearest Neighbors and Support Vector Machines can be used as excellent classifiers for incoming network traffic. Fuzzy networks can also be used in the decision-making process while separating normal traffic packets from potentially harmful ones [1].

### 3.3 Knowledge-Based Methods

In knowledge-based methods, network traffic features are compared with predefined patterns of attack. Some examples of knowledge-based methodologies include “expert systems, signature analysis, self organizing maps, and state transition analysis” [1]. Heuristics can be used to analyze traffic characteristics and classify them as DDoS or otherwise. An excellent example is that of a DDoS detection system which used a “gossip based communication mechanism” to exchange information about network attacks among independent detection nodes in order to use the aggregate data to identify network attacks [1]. Another model, used temporal-correlation based method to extract features and spatial-correlation for detection to correctly identify DDoS attacks [1].

## 4 DDoS ATTACK DETECTION MODEL

For this project, we have worked on the design and implementation of an optimal DDoS detection model (based on Soft-Computing and Machine Learning algorithms) by training and implementing several potential models and creating an ensemble model from the best ones. We have also explored the traffic logs dataset to identify patterns via unsupervised means.

## 4.1 Data Description

The KDD Cup'99 dataset [6] has been used for our data analysis. This dataset has been derived from the 1998 DARPA Intrusion Detection Evaluation Program dataset [8] which was prepared and managed by MIT Lincoln Labs. The data was simulated to evaluate study in intrusion detection. It comprises of a “wide variety of intrusions simulated in a military network environment” [6]. The original data comprised of around five million records. Hence, we use a 10 percent subset of the original train and test datasets for our analysis purposes.

## 4.2 Data Exploration and Processing

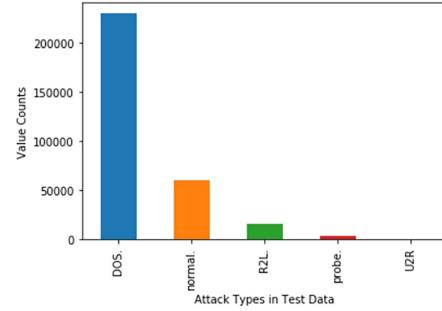
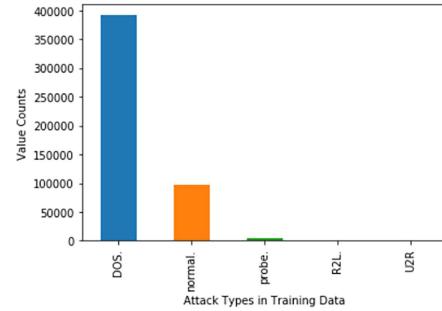
The data exploration and analysis for this project has been implemented using Python on *Jupyter Notebook*. The *Jupyter Notebook* provides us with “an open-source web application that allows us to create and share documents that contain live code, equations, visualizations and narrative text” [5].

For data loading, we use the *Pandas* library in python, which is one of the largest and most flexible data managing libraries and offers a wide variety of options for data handling and manipulation using data frames. After loading the datasets, we explore some of the features of the dataset. From the documentation on the KDD Cup'99 dataset, we know that the data consists of a wide variety of network attacks, but the five main classes of network traffic are as follows: normal (normal network traffic), DoS/DDoS (Denial-of-Service network traffic), R2L (unauthorized access from a remote machine traffic), U2R (unauthorized access to local superuser privileges traffic) and probing [6]. Also, the test dataset consists of an additional 14 attack types which are not present in the training data. However, these new attack types are also a part of the above five categories and the purpose behind their addition in the dataset was to prove that new variants can also be detected using signatures of the preexisting types of attacks [6].

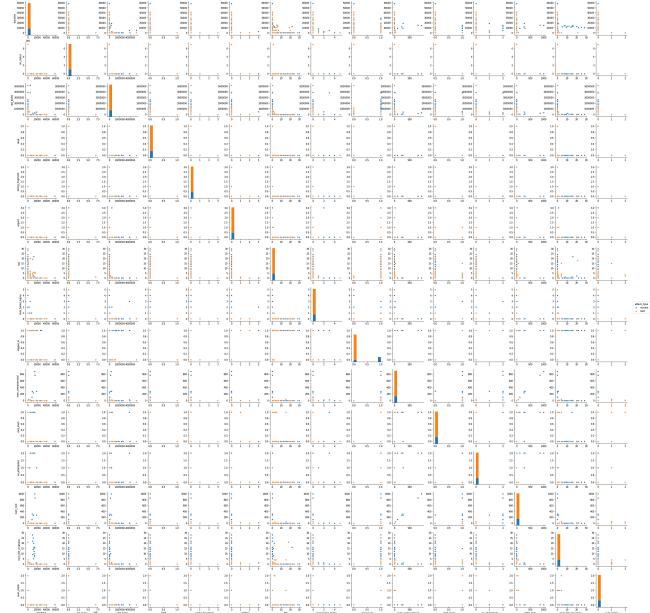
For plotting and visualization purposes, we use *Matplotlib* and *Seaborn* - two excellent visualization libraries offered by Python. First, we check for nulls in the train and test dataset, but find none. Secondly, we check the three categorical columns in the data, to ensure same levels in both the training and test dataset. We find that the training dataset has an additional level in the *service* column. For simplicity, we remove the categorical columns from our analysis dataset and continue our work on only the numerical columns. We now explore the target label column which specifies the *attack type* or the network traffic class. We map the labels to five core categories discussed previously and compare them for the training and testing set. Figure 5 shows the Attack Type distribution in the training and test datasets

We can observe that DoS attacks form the majority of all the attack types (98.67 percent out of all attacks in training set; 91.78 percent out of all attacks in test set). Hence, we broadly classify the target labels as *normal* and *bad* for intrusion detection. We also include the individual labels for the multi-label classification part.

Post this, we create pair plots for the first few variables in order to view individual distributions as well as correlations. Figure 6 shows the pair plot between the first 15 variables in the training dataset. We observe that the data seems to be skewed, indicating the need for standardizing the features. Also, there do not seem to



**Figure 5: Attack Type Distributions**



**Figure 6: Pair plot for Training Features**

be a lot of correlated variables in the dataset.

We proceed with separating the binary variables (mentioned in the documentation) from the continuous variables and scaling the continuous variables using mean normalization in the training dataset. We then apply the same transformations to the test dataset. Post

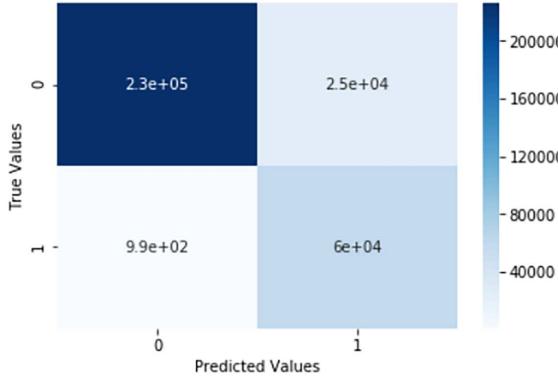


Figure 7: Logistic Regression Confusion Matrix - 2-class classification

this, we consolidate all our features and get the final processed datasets for training and testing.

### 4.3 Data Analysis

Once we are ready with our final datasets, we design the required detection models by training them on the training data and testing their performance on the test data. For the design of the models, we use the *scikit-learn* or *sklearn* package in python, which contains a plethora of resources for statistical and machine learning methodologies. For most models, we also employ the *n\_jobs* parameter present in the models for parallelization purposes [9]. For performance tests, we calculate the accuracy, precision, recall and F1 score for the model (for both 2-label and multi-label classification). The confusion matrix generated in each case displays the classes as follows: 2-class classification (0 - bad, 1 - normal) ; Multi-class classification (0 - Dos/DDoS, 1 - R2L, 2 - U2R, 3 - normal, 4 - probe).

**4.3.1 Logistic Regression.** Logistic Regression is a machine learning algorithm based on the regression model which is used to fit a model to describe the relationship between a dependent (categorical target) and one or more independent variables. Used mainly for classification purposes, the target variable in a logistic regression model is mainly binary, although the method can be used for multi-class classification too. The basis of logistic regression is a *logistic function* (usually a sigmoid function) which keeps the output values bounded between 0 and 1. This function is fit using a *maximum likelihood* methodology which attempts to estimate the coefficients of the regression equation such that the probability outputs match as closely as possible to the true output values [4].

We train two logistic regression models - one for the 2-class classification and one for the multi-class classification. The model for the 2-class classification is fit as per the default parameters, with the regularization parameter as 0.01 for stronger regularization. For the multi-class classification (since this is not the default type for a logistic regression model), we use a specific solver method known as *Stochastic Average Gradient Descent Solver* [9]. Figure 7 shows the 2-class confusion matrix for logistic regression. Figure 8 shows the multi-class confusion matrix for logistic regression.

The overall accuracy, recall, precision and F1 score for the 2-class

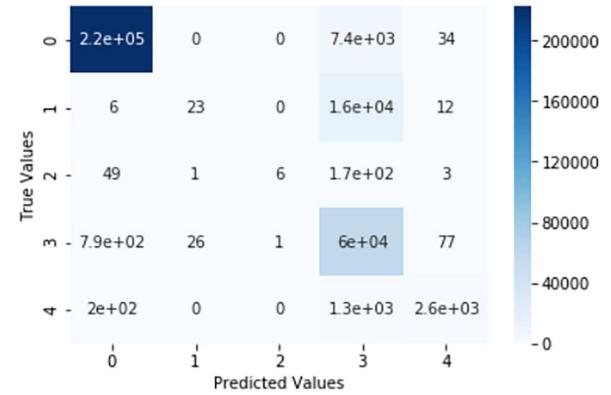


Figure 8: Logistic Regression Confusion Matrix - Multi-class classification

classification are as follows: 91.7, 94.2, 85.0 and 88.3 percent. The same for the multi-class classification are as follows: 91.5, 52.2, 79.4 and 52.3. We can observe that the accuracy of the model seems to be good for the 2-class classification but the recall and F1 scores decrease for the multi-class classification (due to the decrease in recall for the U2R and R2L classes, which have a higher proportion in test as compared to train data).

**4.3.2 K-Nearest Neighbors.** The K-Nearest Neighbors algorithm selects the  $k$  nearest points to the test data point (depending upon a predefined distance metric), present in the training data point, and assign it the class label depending on the majority class label present among the  $k$  training data points. Being a non-parametric method, KNN does not assume any initial distribution or form of data [4].

We train two KNN ( $k=5$ ) models - one for the 2-class classification and one for the multi-class classification. For both the classification models, instead of taking the *brute force* or traditional approach, we use an optimized approach known as *Ball Tree Method* [9], which is a tree based method that endeavors to reduce the number of distance computations by encoding the distance information more efficiently. It recursively divides the data according to a hyper-sphere determined by a particular centroid and radius, and reduces the participants for a neighbor search using triangle inequality [9]. This method works better for data in high dimensions (similar to the dataset for our analysis). Also, we take the distance metric as Manhattan Distance instead of the commonly used Euclidean Distance metric, due to better properties of Manhattan distance in higher dimensions.

Figure 9 shows the 2-class confusion matrix for KNN. Figure 10 shows the multi-class confusion matrix for KNN. The overall accuracy, recall, precision and F1 score for the 2-class classification are as follows: 92.35, 94.64, 85.95 and 89.20 percent. The same for the multi-class classification are as follows: 92.08, 55.90, 80.16 and 55.17. We can observe that the accuracy of the model increases as compared to a simple logistic regression model for the 2-class classification. The recall and F1 scores too increase for the multi-class classification case.

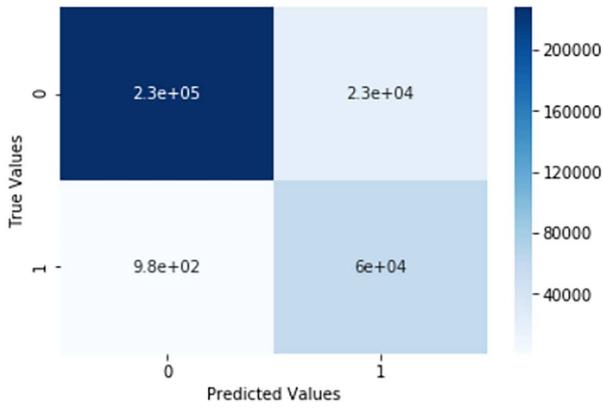


Figure 9: KNN Confusion Matrix - 2-class classification

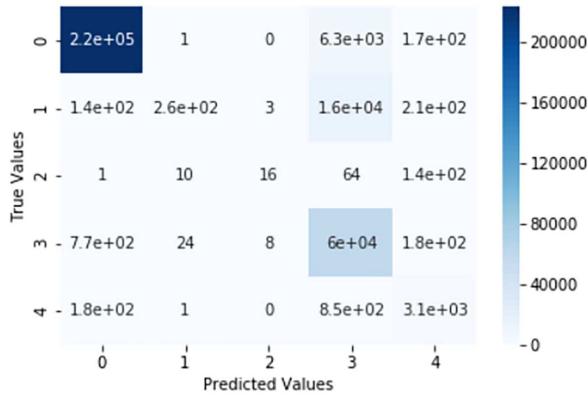


Figure 10: KNN Confusion Matrix - Multi-class classification

**4.3.3 Support Vector Machine - Linear.** A Support Vector Machine is a model based on the maximal margin classifier i.e. classification based on an optimal separating hyperplane. The support vector machine extends this concept further and to non-linear decision boundaries as well. It uses a function referred to as the *kernel* which acts as quantification of the similarity between observations. Therefore, for non-linear cases a variety of kernels such as radial or polynomial can be employed for classification purposes [4].

We train two linear SVM models - one for the 2-class classification and one for the multi-class classification. We implement this classifier using a *Bagging Classifier* which uses the base SVM classifier on different subsets of data drawn with replacement (also referred to as bootstrapping) and aggregates the results to given the final output [9]. Figure 11 shows the 2-class confusion matrix for linear SVM. Figure 12 shows the multi-class confusion matrix for linear SVM.

The overall accuracy, recall, precision and F1 score for the 2-class classification are as follows: 92.23, 94.55, 85.78 and 89.04 percent. The same for the multi-class classification are as follows: 89.14, 54.91, 83.44 and 55.81. We can observe that the accuracy of the model increases as compared to a simple logistic regression model

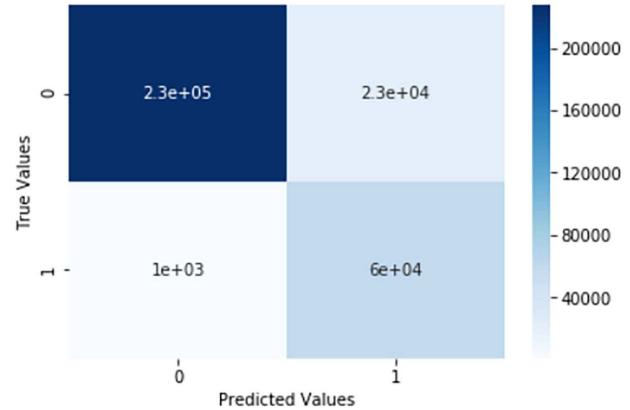


Figure 11: Linear SVM Confusion Matrix - 2-class classification

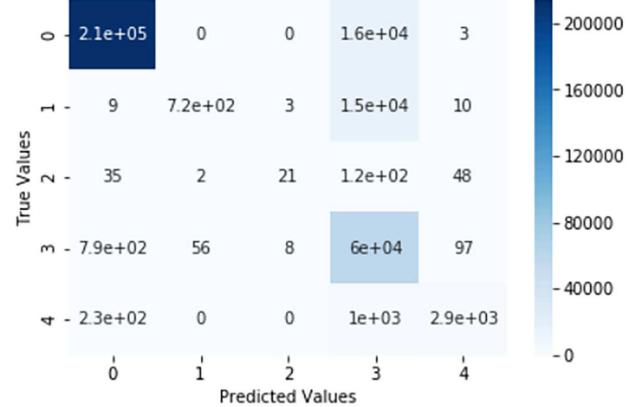
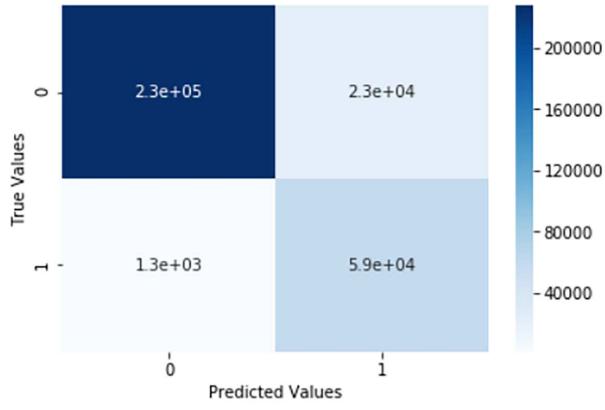


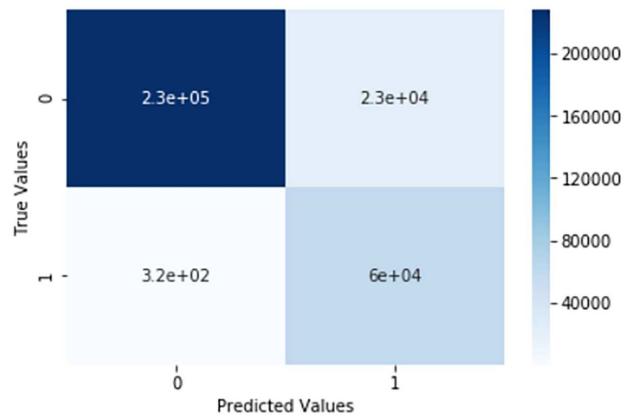
Figure 12: Linear SVM Confusion Matrix - Multi-class classification

but is lower than the KNN model for the 2-class classification. The recall and F1 scores too increase compared to logistic regression but are similar to that of KNN for the multi-class classification case.

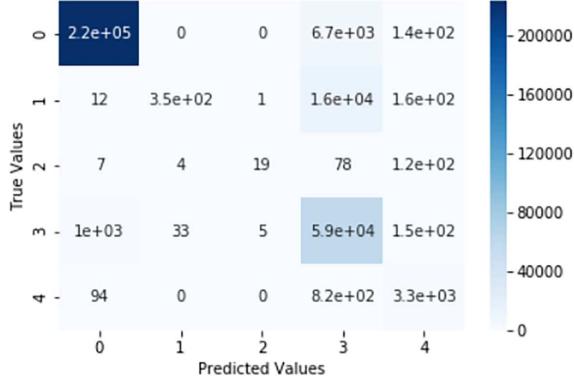
**4.3.4 Support Vector Machine - Polynomial.** Here, we train two SVM models (with polynomial kernels of degree=3) and using a *Bagging Classifier* - one for the 2-class classification and one for the multi-class classification. Figure 13 shows the 2-class confusion matrix for polynomial SVM. Figure 14 shows the multi-class confusion matrix for polynomial SVM. The overall accuracy, recall, precision and F1 score for the 2-class classification are as follows: 92.28, 94.41, 85.87 and 89.08 percent. The same for the multi-class classification are as follows: 91.95, 56.74, 84.60 and 56.38. We can observe that the accuracy of this model too is lower than the KNN model for the 2-class classification. However, the recall and F1 scores are higher than KNN too (correctly classifies more DoS/DDoS and probe attacks than linear



**Figure 13: Polynomial SVM Confusion Matrix - 2-class classification**



**Figure 15: Random Forest Confusion Matrix - 2-class classification**



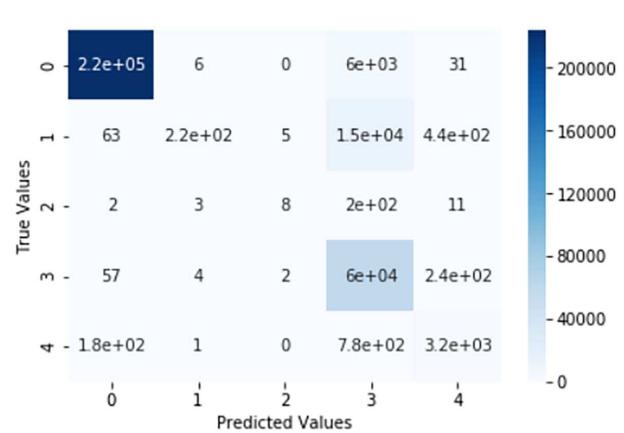
**Figure 14: Polynomial SVM Confusion Matrix - Multi-class classification**

SVM and more R2L and probe attacks than KNN) for the multi-class classification case. Overall, the performance is similar to KNN.

**4.3.5 Random Forest.** A random forest model works as an improvement over individual decision trees through building a number of decision trees on bootstrapped samples along with decorrelating the individual trees by choosing only a random subset of predictors out of the total predictors while constructing trees. At each split, a fresh subset of predictors is used which implements the decorrelation of features [4].

We train two random forest models - one for the 2-class classification and one for the multi-class classification. The selection of the subset of features is taken as the default parameter i.e. square root of the total number of features [9]. Figure 15 shows the 2-class confusion matrix for a Random Forest. Figure 16 shows the multi-class confusion matrix for a Random Forest.

The overall accuracy, recall, precision and F1 score for the 2-class classification are as follows: 92.64, 95.22, 86.31, 89.63 percent. The same for the multi-class classification are as follows: 92.44, 55.74, 80.39 and 54.26. We can observe that the accuracy of this model



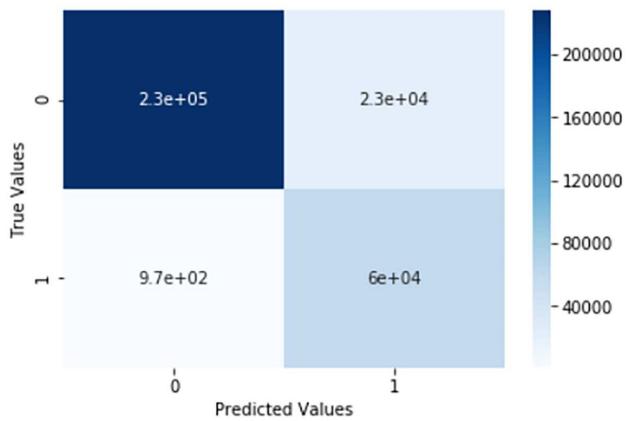
**Figure 16: Random Forest Confusion Matrix - Multi-class classification**

higher than all the previous models for the 2-class classification. The recall and F1 score for multi-class classification is comparable to the SVM models.

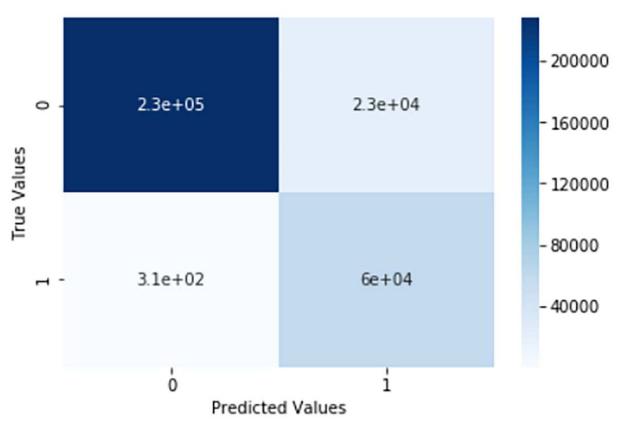
**4.3.6 Neural Networks : Multi-Layer Perceptron.** Neural Networks are soft-computing techniques that attempt to replicate information processing in biological systems, and thus have excellent learning capabilities. When used for pattern recognition or classification purposes, the most useful Neural Network is that of Multi-Layer Perceptron which basically acts as multiple layers of logistic regression models [2].

We train two MLP models (with a hyperbolic tan activation function as it has better convergence properties than a logistic or sigmoid function) - one for the 2-class classification and one for the multi-class classification. Figure 15 shows the 2-class confusion matrix for a Multi-Layer Perceptron. Figure 18 shows the multi-class confusion matrix for a Multi-Layer Perceptron.

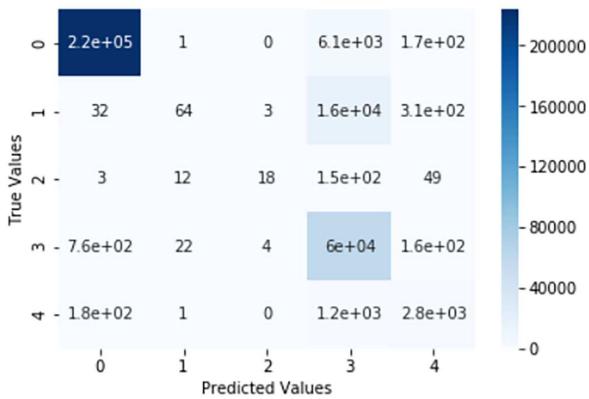
The overall accuracy, recall, precision and F1 score for the 2-class



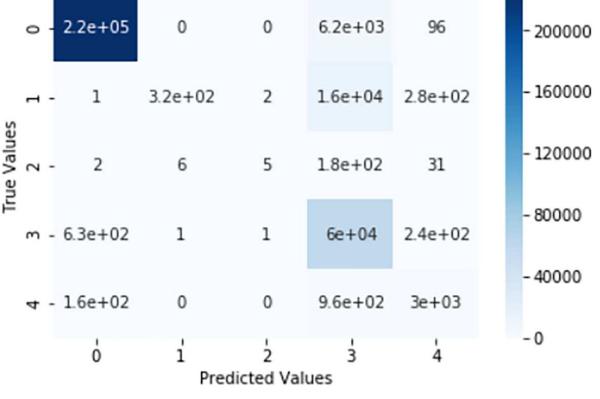
**Figure 17: Multi-Layer Perceptron Confusion Matrix - 2-class classification**



**Figure 19: Ensemble Model Confusion Matrix - 2-class classification**



**Figure 18: Multi-Layer Perceptron Confusion Matrix - Multi-class classification**



**Figure 20: Ensemble Model Confusion Matrix - Multi-class classification**

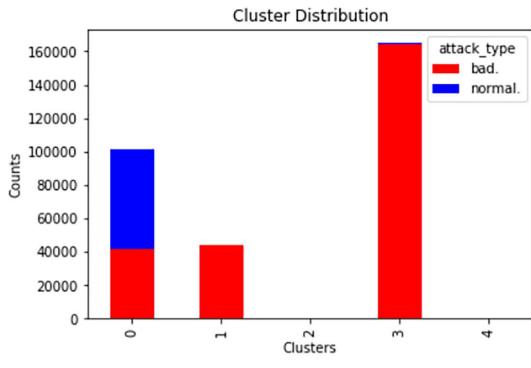
classification are as follows: 92.40, 94.68, 86.02 and 89.27 percent. The same for the multi-class classification are as follows: 91.98, 54.18, 77.53 and 53.90. We can observe that the accuracy of this model is similar to that of a random forest model for the 2-class classification. The recall and F1 score for multi-class classification is comparable to the random forest model.

**4.3.7 Ensemble Modeling.** Ensemble modeling deals with the combination of two or more machine learning models to generate a model with better accuracy. We have already observed that Random Forests have the highest accuracy for the 2-label classification whereas a polynomial SVM has better recall for the multi-label classification. Therefore, we try to get the best of both worlds by creating an ensemble of two Random Forest (with different rules for selection of the feature subset - square root of features and log of features) and one polynomial SVM model.

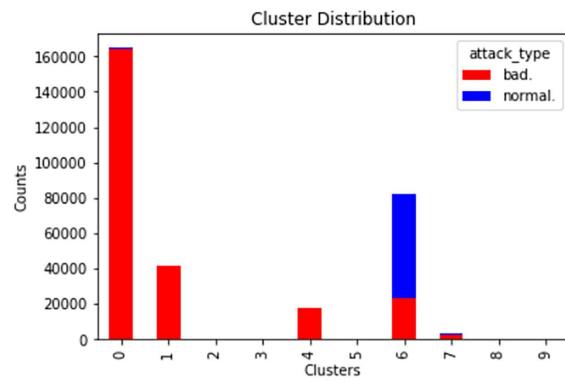
We train two ensemble models - one for the 2-class classification and one for the multi-class classification. Figure 19 shows the 2-class confusion matrix for an ensemble model. Figure 20 shows the

multi-class confusion matrix for an ensemble model. The overall accuracy, recall, precision and F1 score for the 2-class classification are as follows: 92.66, 95.25, 86.33 and 89.65 percent. The same for the multi-class classification are as follows: 92.26, 56.85, 84.77 and 55.41. We can observe that the accuracy and F1 score of this model is higher than all individual models for the 2-class classification. The recall and F1 score for multi-class classification is balanced between that of the random forest and the polynomial SVM but is higher than most individual models.

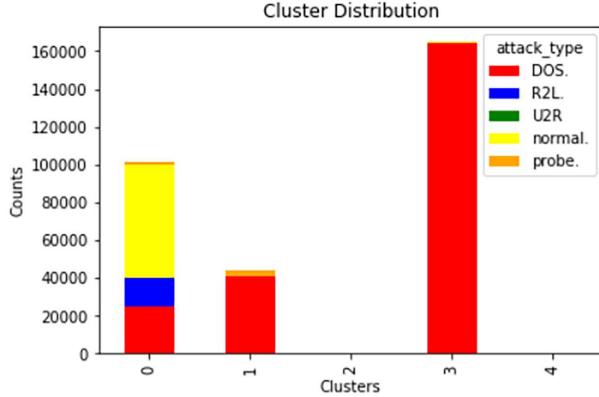
**4.3.8 Unsupervised Learning - Clustering.** Up till here, we observed and evaluated a variety of supervised learning models. As a result, we came to the conclusion that an ensemble of two good models often results in a better and more balanced result than individual models. In this section, we will examine how exploring the test data by means of a clustering algorithm (with no support from the training data) helps provide a good idea of the patterns within the data. The clustering algorithm we will use for this purpose is K-Means which is used to partition data into a given number of



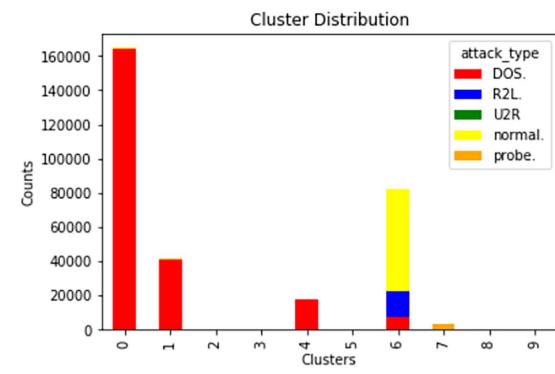
**Figure 21: Chart for k-means clustering (clusters=5) - 2-class classification**



**Figure 23: Chart for k-means clustering (clusters=10) - 2-class classification**



**Figure 22: Chart for k-means clustering (clusters=5) - multi-class classification**



**Figure 24: Chart for k-means clustering (clusters=10) - multi-class classification**

non-overlapping clusters based on a distance metric [4]. We train two k-means clustering models for both 2-class and multi-class classification - one for clusters=5 and the other for clusters=10 (for greater granularity).

Figure 21 shows the 2-class chart for k-means clustering with clusters=5 (some clusters not visible due to small size). We see that most of the clusters show one of the classes as a dominant proportion of the cluster. We can validate the same by comparing with the multi-class labels as well.

Figure 22 shows the multi-class chart for k-means clustering with clusters=5 (some clusters not visible due to small size).

We also run the analysis for clusters=10, for greater granularity. Figure 23 shows the 2-class chart for k-means clustering with clusters=10 (some clusters not visible due to small size).

Figure 24 shows the multi-class chart for k-means clustering with clusters=10 (some clusters not visible due to small size).

We can observe the same trend here as well.

#### 4.4 Results

Among the supervised models, we observe that on comparison, some models perform better in terms of accuracy whereas some

perform better in terms of recall. We also observe that most models find it easier to perform a 2-class classification (due to the high volume of attack labels in both the datasets as compared to normal labels), but face difficulties in identifying the individual classes (especially R2L and U2R which have a higher proportion in the test data compared to the training data). Overall, for the purpose of Dos/DDoS and intrusion detection, we see that most machine learning models give good results (KNN for example), and an ensemble of a random forest and polynomial SVM model gives the best accuracy among all.

When we venture into unsupervised learning we observe that clustering algorithms too can work well on network traffic data by creating clusters of traffic logs through pattern recognition. Though clustering does not provide us with exact labels, it can be useful in cases where we do not have any training or benchmark data, by giving us a fair idea of the direction in which to proceed.

### 5 APACHE SPARK - USING PYSPARK

The volume of network traffic data generated is generally quite huge, and thus requires Big Data technologies to deal with it. Our demonstration was for a smaller subset of the actual dataset (which in itself consists of five million records). However, this larger dataset

too consists of logs only for seven weeks of monitoring. We can therefore imagine how voluminous the datasets would begin to get with constant monitoring of systems. In such cases, Big Data cloud technologies can come to the aid of analytics, and help create a sustainable system for such intrusion detection purposes.

Our analysis was carried out using Python on an individual system. But often for larger datasets, we need additional resources. The PySpark API, from Apache Spark (an open-source processing engine), can help us gain “access to the extremely high-performance data processing enabled by Spark’s Scala architecture - without the need to learn any Scala” [3]. The smallest building blocks of Spark are referred to as RDDs (Resilient Distributed Datasets) and these along with Spark’s DataFrame can act as useful alternatives to the *Pandas* data frames, in case of large datasets, where the distributed processing power of Spark can come into play [3].

We can install PySpark on a Windows machine using GOW (incorporates Linux commands in Windows like gzip, curl and tar) and Anaconda (an open-scale distribution containing Jupyter Notebook and other resources for Python). The package can be installed from the Apache Spark website, following which we perform *gzip* and *tar* operations on it. After adding the windows binary for Hadoop and modifying a few environment variables, you can launch Spark locally from Command Prompt. We have not used Spark for our analyses further as Python was able to handle the 10 percent datasets locally. However, PySpark can prove to be a great tool for analyzing data and creating models for larger datasets using a familiar and flexible language like Python. The presence of libraries like *mllib* in PySpark can offer us a wide variety of learning algorithms (similar to the *sklearn* library in Python).

## 6 CONCLUSION

The detection and prevention of DDoS attacks is a crucial problem for the safety and stability of networks. With the increasing use and dependence on technology and connectivity, this affects a huge cohort of people today. The data generated from day-to-day network traffic is huge and largely unstructured, but it can be captured and modified into an understandable structure, to be analyzed and used to generate efficient solutions. Through our analysis, we affirm the efficiency of machine learning technologies as tools for Big Data analytics and the use of open-source distributed processing systems as supports towards utilization of these tools. We observe that not only do supervised learning methods work well towards this objective, but unsupervised learning techniques such as clustering also provide us with helpful insights on pattern detection in the data. Therefore, Big Data technologies along with intelligent analytic solutions can help create new and improve existing defense systems to ensure security from such malicious attacks and intrusions.

## REFERENCES

- [1] Monowar H. Bhuyan, H. J. Kashyap, D. K. Bhattacharyya, and J. K. Kalita. 2014. Detecting distributed denial of service attacks: methods, tools and future directions. *Comput. J.* 57 (2014), 537–556. <https://doi.org/10.1093/comjnl/bxt031>
- [2] Christopher M. Bishop. 2006. *Pattern Recognition and Machine Learning*. Springer, New York.
- [3] IBM. 2016. *PySpark High-performance data processing without learning Scala*. IBM.
- [4] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning with Applications in R*. Springer, New York.
- <https://doi.org/10.1007/978-1-4614-7138-7>
- [5] Jupyter. 2017. The Jupyter Notebook. (2017).
- [6] KDDCup99. 1999. KDD Cup 1999 Data. (1999).
- [7] Andrew Kiggins and Jeffrey Lyons. 2016. *AWS Best Practices for DDoS Resiliency*. Amazon Web Services.
- [8] MITLincolnLaboratory. 1998. DARPA Intrusion Detection Evaluation. (1998).
- [9] scikit learn. 2017. scikit-learn - Machine Learning in Python. (2017).
- [10] Jessica Stone. 2017. The Best DDoS Protection Services. (July 2017).
- [11] Lea Toms. 2016. Closed for Business - the Impact of Denial of Service Attacks in the IoT. (Feb 2016).
- [12] Guoxing Zhang, Shengming Jiang, and Gang Wei. 2009. A prediction-based detection algorithm against distributed denial-of-service attacks. In *Proceedings of the International Conference on Wireless Communications and Mobile Computing: Connecting the World Wirelessly*, Vol. 1. Leipzig, Germany, 106fi?!110.

```
bibtext report
```

```
=====
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
Warning--empty publisher in zhang05
(There was 1 warning)
```

```
bibtext _ label error
```

```
=====
bibtext space label error
```

```
=====
bibtext comma label error
```

```
=====
latex report
```

```
=====
[2017-12-11 13.26.59] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflatex)
Missing character: ""
Missing character: ""
Missing character: ""
Typesetting of "report.tex" completed in 18.3s.
```

```
=====
Compliance Report
=====
```

```
name: Rawat, Neha
hid: 224
paper1: Nov 3 17 100%
paper2: Nov 6 17 100%
project: Dec 04 17 100%
```

```
yamlcheck
```

```
wordcount
```

---

```
10  
wc 224 project 10 5836 content.tex  
wc 224 project 10 5726 report.pdf  
wc 224 project 10 348 report.bib
```

```
find "
```

---

```
passed: True
```

```
find footnote
```

---

```
passed: True
```

```
find input{format/i523}
```

---

```
passed: False
```

```
find input{format/final}
```

---

```
4: \input{format/final}
```

```
passed: True
```

```
floats
```

---

```
29: \begin{figure}  
30: \includegraphics[width=1.0\columnwidth]{images/DDoS.PNG}  
32: \label{F:ddos}  
34: Figure \ref{F:ddos} shows how a Distributed Denial-of-Service  
    attack occurs.\\  
40: \begin{figure}  
41: \includegraphics[width=1.0\columnwidth]{images/OSI.PNG}  
43: \label{F:osi}  
45: Figure \ref{F:osi} shows an Open Systems Interconnection Model  
    with the layers highlighted where DDoS attacks are most common.\\  
53: \begin{figure}
```

```

54: \includegraphics[width=1.0\columnwidth]{images/dossteps.PNG}
55: \label{F:doss}
56: Figure \ref{F:doss} shows the steps of a Denial-of-Service attack
   execution.\\
57: \begin{figure}
58: \includegraphics[width=1.0\columnwidth]{images/dosattackstat.PNG}
59: \label{F:dosstat}
60: Figure \ref{F:dosstat} shows the percentage of different Denial-
   of-Service attacks in 2011 by type.\\
61: \begin{figure}
62: \includegraphics[width=1.0\columnwidth]{images/attack_type.PNG}
63: \label{F:att}
64: Figure \ref{F:att} shows the Attack Type distribution in the
   training and test datasets\\
65: \begin{figure}
66: \includegraphics[width=1.0\columnwidth]{images/pairplot.png}
67: \label{F:pair}
68: Figure \ref{F:pair} shows the pair plot between the first 15
   variables in the training dataset. We observe that the data seems
   to be skewed, indicating the need for standardizing the features.
   Also, there do not seem to be a lot of correlated variables in
   the dataset.\\
69: \begin{figure}
70: \includegraphics[width=1.0\columnwidth]{images/logreg2.PNG}
71: \label{F:logreg2}
72: Figure \ref{F:logreg2} shows the 2-class confusion matrix for
   logistic regression.
73: \begin{figure}
74: \includegraphics[width=1.0\columnwidth]{images/logregall.PNG}
75: \label{F:logregall}
76: Figure \ref{F:logregall} shows the multi-class confusion matrix
   for logistic regression.\\
77: \begin{figure}
78: \includegraphics[width=1.0\columnwidth]{images/knn2.PNG}
79: \label{F:knn2}
80: Figure \ref{F:knn2} shows the 2-class confusion matrix for KNN.
81: \begin{figure}
82: \includegraphics[width=1.0\columnwidth]{images/knnall.PNG}
83: \label{F:knnall}
84: Figure \ref{F:knnall} shows the multi-class confusion matrix for
   KNN.\\
85: \begin{figure}
86: \includegraphics[width=1.0\columnwidth]{images/svm2.PNG}
87: \label{F:linsvm2}
88: Figure \ref{F:linsvm2} shows the 2-class confusion matrix for
   linear SVM.

```

```

158: \begin{figure}
159: \includegraphics[width=1.0\columnwidth]{images/svmall.PNG}
161: \label{F:linsvmall}
163: Figure \ref{F:linsvmall} shows the multi-class confusion matrix
   for linear SVM.\\
168: \begin{figure}
169: \includegraphics[width=1.0\columnwidth]{images/svmpoly2.PNG}
171: \label{F:polysvm2}
173: Figure \ref{F:polysvm2} shows the 2-class confusion matrix for
   polynomial SVM.
174: \begin{figure}
175: \includegraphics[width=1.0\columnwidth]{images/svmpolyall.PNG}
177: \label{F:polysvmall}
179: Figure \ref{F:polysvmall} shows the multi-class confusion matrix
   for polynomial SVM.\\
186: \begin{figure}
187: \includegraphics[width=1.0\columnwidth]{images/rf2.PNG}
189: \label{F:rf2}
191: Figure \ref{F:rf2} shows the 2-class confusion matrix for a
   Random Forest.
192: \begin{figure}
193: \includegraphics[width=1.0\columnwidth]{images/rfall.PNG}
195: \label{F:rfall}
197: Figure \ref{F:rfall} shows the multi-class confusion matrix for a
   Random Forest.\\
204: \begin{figure}
205: \includegraphics[width=1.0\columnwidth]{images/nn2.PNG}
207: \label{F:nn2}
209: Figure \ref{F:nn2} shows the 2-class confusion matrix for a
   Multi-Layer Perceptron.
210: \begin{figure}
211: \includegraphics[width=1.0\columnwidth]{images/nnall.PNG}
213: \label{F:nnall}
215: Figure \ref{F:nnall} shows the multi-class confusion matrix for a
   Multi-Layer Perceptron.\\
222: \begin{figure}
223: \includegraphics[width=1.0\columnwidth]{images/ensemble2.PNG}
225: \label{F:en2}
227: Figure \ref{F:en2} shows the 2-class confusion matrix for an
   ensemble model.
228: \begin{figure}
229: \includegraphics[width=1.0\columnwidth]{images/ensembleall.PNG}
231: \label{F:enall}
233: Figure \ref{F:enall} shows the multi-class confusion matrix for
   an ensemble model.\\
240: \begin{figure}

```

```

241: \includegraphics[width=1.0\columnwidth]{images/cluster52graph.PNG
}
243: \label{F:cg52}
245: Figure \ref{F:cg52} shows the 2-class chart for k-means
    clustering with clusters=5 (some clusters not visible due to
    small size).\\
247: \begin{figure}
248: \includegraphics[width=1.0\columnwidth]{images/cluster5allgraph.P
    NG}
250: \label{F:cg5all}
252: Figure \ref{F:cg5all} shows the multi-class chart for k-means
    clustering with clusters=5 (some clusters not visible due to
    small size).\\
254: \begin{figure}
255: \includegraphics[width=1.0\columnwidth]{images/cluster102graph.PN
    G}
257: \label{F:cg102}
259: Figure \ref{F:cg102} shows the 2-class chart for k-means
    clustering with clusters=10 (some clusters not visible due to
    small size).\\
260: \begin{figure}
261: \includegraphics[width=1.0\columnwidth]{images/cluster10allgraph.
    PNG}
263: \label{F:cg10all}
265: Figure \ref{F:cg10all} shows the multi-class chart for k-means
    clustering with clusters=10 (some clusters not visible due to
    small size).\\

```

figures 24

tables 0

includegraphics 24

labels 24

refs 24

floats 24

True : ref check passed: (refs >= figures + tables)

True : label check passed: (refs >= figures + tables)

True : include graphics passed: (figures >= includegraphics)

True : check if all figures are referred to: (refs >= labels)

Label/ref check

passed: True

When using figures use columnwidth

[width=1.0\columnwidth]

do not change the number to a smaller fraction

find textwidth

---

passed: True

---

below\_check

---

WARNING: code and above may be used improperly

52: When launching a DDoS attack, the attacker goes through some steps common to both types of architectures \cite{monowar01}. First, the attacker tries to identify vulnerable systems that can be used as agents. The resources of these systems are used to generate a powerful attack stream. Next, the attacker plants the handler software code in the compromised system and ensures steps to prevent the code from being detected. These compromised systems are often referred to as {\em zombies}. Sometimes, the attacker creates several intermediate layers between the {\em zombies} and the victim to hinder traceability. Thirdly, the attacker communicates with the handler codes placed via protocols like TCP or UDP, and decides the scheduling of the attacks. Post the complete setup, the attacker launches the attack on the victim's machine or server and renders it unusable \cite{monowar01}. In an IRC-based architecture, most of the above steps remain same, but an IRC-channel is used for communication purposes. This helps the attacker as even if one {\em zombie} or {\em bot} is discovered, the identities of the others is still hidden, as IRC-channels are difficult to detect \cite{monowar01}.

bibtex

---

label errors

bibtex errors

This is BibTeX, Version 0.99d (TeX Live 2016)  
The top-level auxiliary file: report.aux  
The style file: ACM-Reference-Format.bst  
Database file #1: report.bib

```
Warning--empty publisher in zhang05
(There was 1 warning)
```

```
bibtex_empty_fields
```

---

```
entries in general should not be empty in bibtex
```

```
find ""
```

---

```
passed: True
```

```
ascii
```

---

```
non ascii found 8217
```

```
=====
The following tests are optional
=====
```

```
Tip: newlines can often be replaced just by an empty line
```

```
find newline
```

---

```
passed: True
```

```
cites should have a space before \cite{} but not before the {
```

```
find cite {
```

---

```
passed: True
```

# Big Data Analytics in Indian Premier League

Swargam, Prashanth  
Indiana University Bloomington  
107 S Indiana Ave  
Bloomington, Indiana 47408  
pswargam@iu.edu

## ABSTRACT

Cricket is one of the most admired sports across the globe. Indian Premier League is one of the professional cricket leagues conducted by Board of Cricket Control India in the months of April and May. This league is famous for its diversity of players and breath-taking cricket match endings. The factors of winning change for each moment as the game progresses. As there are many players and franchises involved in the game, these factors for winning changes for each team. Data related to each player is required to analyse his performance and predict his future scope in team. Data related to factors of winning is crucial and can be analysed for predicting the results of the game. This analysis would help the team management, league administration to wisely chose the players and modify rules according to the impact of each decision. Data related some of the important factors which plays major role in deciding the match winner are analysed. Their impact is predicted and compared with the actual results. Impact of these factors are studied for each individual team and individual season of this cricket tournament. Impact of each factor is plotted and its impact in next season is predicted.

## KEYWORDS

Big Data, Cricket, Indian Premier League, i523

## 1 INTRODUCTION

Fast paced games are gaining more importance in near future. This because there are many factors which contribute to the result of the game. These factors are minor but could change the results of the game dramatically. Indian Premier League is one such type of cricket league where there are a lot factors which have their influence on the results of the game. These factors are though minor or major, will have bigger part in deciding the results of the game. These factors from the previous games can be utilised wisely to predict their influence in the upcoming matches. These factors can be quantitatively represented in the form numbers, graphs or Booleans. This quantitative representation of data related the factors can be analysed using various analytic techniques to predict their impact on the game.

However, Analytics is a good way to go about this prediction, but there are several problems which should be addressed. Considering the role of batsman, it will be having parameters like balls faced, dot balls, number of boundaries, strike rate etc. Considering the role of bowler, there are various parameters like matches played, overs bowled, economy rate. There are many similar kinds of roles in the game and above-mentioned parameters are specific to one player playing only one role in the team. According to, around 500 players play for each season of cricket. These 500 players will be

filtered on various factor from the pool of nearly 5,31,253 cricket players across the globe. These players can play any of the role or play multiple to roles to contribute to the result of the match. These players and cricket matches produces large amount of data which when analysed to produce structured data and analytics. Hence, there is good scope of analytics ad big data in this sport.

The data produced by the matches happening in Indian Premier League can be used to fit in mathematical models. These mathematical models are then used to study the nature and trends of the factors which influence the results of the game. Extending this model to the known values of the input factors can produce the predicted values of the impacts of these factors. Models like Linear regression, polynomial regression, radial-basis approach can be applied to do these kinds of predictive analysis.

## 2 PROBLEM STATEMENT

There are various factors influencing the results of the game. As part of this analytics, data related to four of the most influencing factors is gathered and modelled for analysis. This data was available in raw formats which requires some amount of modelling for predictive analysis. The modelled data is used for building a mathematical model which would fit closely to the trend of these factors in matches played in all the past seasons. A part of data is assumed to be unknown. This unknown part of data is predicted by using the fitted mathematical models. Results obtained by these predictions are compared with the actual results from the data source. Impact of these factors are calculated to the ratio of one. These data is analysed for each independent team and each independent season.

The report is in regard to the predictive analysis conducted on only five of the most influencing factors in the game. There are other factors in the game which might influence the result of the game. This predictive analysis will only be considered reliable only if the predicted values of the results will have high accuracy with respect to the actual results of the available data. The data is divided into two parts. All the available data is sorted with respect to date. The latest match comes later in the dataset and the earliest first. The first part of data is used to train the mathematical model. The parameters in the later part of the dataset are used to predict the result of the match. These predicted results are then compared with the actual results in the datasets to determine the accuracy of the predictive model which is used to build the analytics. This analysis produce valuable insights on the influence of these factors and the mathematical model.

## 3 SCOPE

The scope of the analysis is:

1)The analytics uses the data for only five factors. These five factors are namely toss, Batting position, Range of score, portion of runs in boundaries.

2)The data is collected for all the seasons completed for this tournament. However, this data is sorted with respect to date and partitioned into training and testing data sets for calculating the accuracy of the model.

3) The values of these factors are represented in usable data formats like range, Boolean, integers for analysis.

## 4 FACTORS IN CONSIDERATION

### 4.1 Batting Sequence

Order of batting is considered as one of the factors in consideration. Batting order is one crucial parameter which depends on various other factors of the game. Some of these parameters are the status of the pitch for the game, climate conditions of the game, previous statistics of the game and the history of the team in similar situations. The toss winner will have the privilege to decide the order of the batting. As this factor is conglomeration of various other factors stated above, batting order is considered for the analytics. This data can be represented in the form of Boolean. Where true Boolean indicates that the team referring to the statistics have batted first in the game. False indicates that the team referring to the statistics have batted second in the game. This Boolean value depends on the values in for toss winner and toss decision.

### 4.2 Total Score

Score indicates the total number of runs scored by the team in any match. Score of the team depends on various other parameters of the game like team statistics and composition, impact of the opponents, and situation of the match. This parameter can be calculated from other values of extra runs, scored runs. This categorized into four categories. This first category of the innings scored not more than 100 runs. The second category of the innings scored more than 100 and less than 150 runs. The third category of the innings scored more than 150 and less than 200 hundred runs. All the other innings which scored more than 200 are categorised into fourth category. This categorization is done in accordance to the range of scores. The least scored innings were given least category value. The highest scored innings were given highest category value.

### 4.3 Score Composition

Composition of scored runs. The runs are majorly scored in the form of boundaries, players individual running, and the extra runs given by then bowling team. As IPL is a T20 game which is played for short duration of time, scoring runs quickly at right time is crucial factor. Boundaries contribute to runs scored in the form of fours and sixes in the game. This is the easiest way to quickly score the runs. Team scoring high majority of runs in the form of boundaries have higher chance of imposing a higher target to the opponents or chasing down the target imposed by the opponents. Hence, this parameter is considered for analysis. This value for this parameter is a Boolean. This value is set to true, if most of the runs scored by any team in any innings are from boundaries and vice versa.

### 4.4 Toss

The batting sequence is decided by the winner of the toss. Winner of the toss will have the initial upper hand in the game to decide the sequence of the game. The winner's decision will vary on various other factors of the game like, duration of the match, pitch behaviour throughout the game, statistics of the game. These various factors play an important role in deciding the toss winner's decision. The value for this parameter is Boolean. The true value of the parameter indicates the team has won the toss and the false value indicates the team has not won the toss.

## 5 DATA MANIPULATION

### 5.1 Team data

There are thirteen teams participated in IPL which was held for nine seasons. Each team was having its team name and teamId. These details are taken from the data source team.csv. Python's csv module is used to read the data from this csv file. As this data represents a key value pair, csv module's dictreader method is used to read the row in these files. Using this method, a dictionary was created which consisted of the teamId as keys and team name as value objects.

### 5.2 Match data

Details pertaining to a specific match are published to the Match.csv file. These details include host team name, guest team name, toss winner id, match winner id, decision of the toss, win type. This data was used and modified for calculating the impact of each factors stated in the factors description. Python's pandas dataframes were used to read the data from these csv files. All the missing values in the dataframes are replaced with 0 to ease the complexities that arise with null values. For each value in the team dictionary, the teamId is matched with the opponent team id column and team name id column in the match.csv. Python's operator module is used to obtain the or condition between these values. Dataframe is modified with the given conditions. This dataframe is converted to list. This way, list of matches played by a team is defined and stored into a list.

From the dataframe which contains the list of matches played by the team, column matchwinnerid is used to define if the match winner. A new dataframe is created with a condition if the matchwinnerid's value in the column is equals to the id of the current team. If the above-mentioned condition is true, then this value is added to the dataframe, else the value is removed from the dataframe. This way, list of matches won by a team is determined.

A new dataframe is created to store the Booleans of the toss decision. From the teamdata dataframe, the column toss winner id is used to determine the toss winner for each match. If the id value in this column is same as the team id of the current team, Boolean true is appended to the toss list. Else, Boolean false is appended to the toss list. This list contains only Booleans.

### 5.3 Ball-by-ball data

Ball by ball analyses will be required to calculate the scores for each ball. These details will include the number of runs scored in the ball, extra runs scored, bowler details, batsman details, over details. This

data is extracted from the ballbyball.csv file. This file contains ball by ball analysis of all the matches. This data is sorted according to the match id and is extracted for further analysis. Pandas readcsv method is used to read the csv file. A new data dataframe called balldata is created.

Balldata csv is used to for calculating the runs scored by a team in a specific match. The balldata dataframe is filtered for useless columns and null values. All the null values are replaced with 0 to ease the complexities which comes with usage of null values. Balldatadfis team batting id and match id are used to calculate the runs scored by a team in a specific match. This data frame is modified such that, the values in the match id column is equal to the current match id and the values in the team batting id is equal to the current team id. Python operator module is used to achieve the and condition in the above case. The modified ball data data frame is used to calculate the score. The sum method on the dataframe is used from the pandas library. This score is categorised into four categories. The first category included the score which are less than hundred. The second category included the scores which are between hundred and one hundred fifty. The third category included the scores which are between one hundred fifty and two hundred. The fourth category of scores contain scores which are above two hundred.

A new list is created which stored the category of the scores. If the score falls in first category, then an integer 1 is appended to the list. If the score falls in the second category, integer 2 is appended to the list. If the score falls in the category falls in the third category, an integer 3 is appended. If the score falls in the category four, an integer 4 is appended to the list. The other factor which was stated in the factors in consideration teamfis score composition. It is highly probable that a team scores a high total or chase down the target imposed by the opposition team quickly is the majority of runs are scored in the form of boundaries. The ball data dataframe which is created in the previous case is utilised with other conditions to calculate the contribution of boundaries to the total score. This dataframe is filtered with the match id and team id condition to obtain the current match and current batting team. This is again filtered for all the scores that are having values either four or six. The minified frame is filtered for all values of four and summed. The same is repeated with the sum of six values. Adding together these two sums will give the total amount of runs scored in form of boundaries.

A new list is created for storing the Booleans related to the contribution of boundaries to the score. This list is appended with value true, if the contribution of boundaries is more than other forms of runs, false if, the contribution of boundaries is less than the contribution of other forms of runs.

The lists which are created for each factor are used to define factorsfi impact on the gamefis result. These lists contain the values in the form of integers and Booleans which are obtained by using the existing parameters. These lists are created for each value in the team.csv file. Thus, they are independent to each team. A new DataFrame is created with these lists as columns in the frame. For each team, these frames are stored into another csv file using the pandas write csv function with the file name same as the value in the team dictionary.

## 6 PREDICTIVE ANALYSIS

The factors and their values are written in to a csv file which contains the factor names as the columns headings and their respective values in the rows. Each team has its independent statistics files mentioned in the above statement. These files will be used for conducting the predictive analysis. For each value in the team dictionary, these csv files which are specific to the team are read using the pandas csv reader function. There are two types of columns in these statistics file. The first type is predictors and the other type is targets. Columns related to the factors which impact the results are considered as the predictor columns, Columns related to the result of the match are considered as the targets column. Columns battingfirst, boundaries majority, wontoss, scorerange are predictors. Wonmatch column is target in the given scenario.

The data available in the statistics files is split into test and train sets. This is done by the function train test split in the sklearn library in the python. The data is split in the ratio of 3 is to 2. Sixty percent of the data is used as training set. Forty percent of the data is used as testing set. The predictors and target of the training set are used to build the mathematical model. The predictors of the testing set are used to predict the values of the targets. These predicted values are compared with the actual target values. This comparision is used to determine the accuracy of the prediction.

### 6.1 Implementation

Decision trees and Random forest are used in making the predictions. Decision trees are tree like structures which are built based on the values of the parameters. These trees are useful in defining the probability of the target value being attained. Trees are build with various stems which are drawn from conditional statements. Decision trees has three kinds of elements. The first element i.e., are the decision elements which refer to the block which checks for the condition or logic of the tree. Chance elements are the elements which occur depending on the condition or logic of the function. End elements are the results or outcome of the decision tree. These are basically leaf nodes of the tree. These nodes represent the result. Random Forests are conglomeration of decisions from various decision trees. In random forest approach, a dataset is divided into various subsets which will have some or all the input parameters as the decision makers and some or all the data which are the values for the parameters. Each subset is used to build a tree by using principles of decision tree. The predictions from these prediction trees are used in determining the final value. When a given set of input parameters are giving, they are predicted with all the decision trees developed by the forest. The outcome from all the decision trees are noted. The majority of these outputs is decided as result. This way, the errors which might arise in using only one decision tree can be eradicated. An error from one model of decision tree will be dominated by the results from all the other decision tree.

Random forest classifier from the module sklearn is used to build the various decision trees and predict these values. Classifier type object is instantiated in the code/cite. This object is assigned with the RandomForestclassifier and an attribute called n estimator. The variable n estimator will define the number of decision trees to be build for the analysis. Fit function from the sklearn module is used to develop the model for the random forest algorithm. Fit function

will take the training parameters and training targets as inputs. These are divided into fifty subsets in this scenario. Predict function is used to predict the value of the target given test predictor variable as inputs.

## 6.2 Accuracy

Generation of only one decision tree as the model for the given training data would produce erroneous results. Using the random forests, fifty decision trees are built to predict the correct value of the target. This way, errors produced by one of the decision tree will be corrected by the predictions from the other trees.

In the graph 3, the accuracy of using various number of decision trees are plotted against the number of decision trees. It can be observed that using less than five decision trees, the accuracy for the prediction for all the teams is around sixty percentage. As the number of decision trees increased, the accuracy of all the prediction is increased by at least ten percent. The hundred percent accuracy is because, those teams have played less number of games.

## 7 IMPACT OF FACTORS

This indicates the contribution of each factor for predicting the result of the game. These values will determine the probability of the value of result on any given values for the input variables. In the first step, the distinct values of the target value are noted. For all the distinct values of an input parameter from the set of values of one of the input parameter, the probability of different kind of results are calculated. This calculation is repeated for the distinct values in the set of input parameter and summed at the end. This produces the importance of the feature. The above procedure is repeated for all the decision tree and all the value are averaged for getting the overall value of the importance feature for that feature. This procedure is repeated for all the input parameters. This will give the contribution of each parameter to the result values.

### 7.1 Batting First

The first importance feature for all the teams are plotted in the graph 4 . It is clear from the graph that the importance feature batting first is highest for Rising Pune super giants. This indicates that this team have won most of their previous matches with while batting first in the game. While the Chennai Super kings and Kochi Tuskers are also high, but they are less than half. This indicates that batting first in the match will be a favourable condition for the above mentioned teams. For all the other teams except Kings xi Punjab, Pune warriors and Sunrisers Hyderabad, the batting first importance factor is nearly around 0.1. This implies, out of all that matches which were present in the training set, these team batted first for only ten percent of the games.

However, the total number matches should also be considered while validating this condition. Though the value for Kochi Tuskers is high, this might also be because they have played less number of matches and they have batted first in all the winning matches.

### 7.2 Score Composition

Score Composition composition is the combination of different ways of getting runs. For any high scoring and successful game, a team will have to score runs at faster rate. Hence, scoring runs in

form of boundaries will help a team a lot in turning the match in their favour. In the graph 5 , Contribution of this factor against each team is plotted. This graph summarises the contribution towards of the score composition factor towards the factor. This factor is very high for Kochi Tuskers team, because they have played considerably less number of games.

A factor around 0.2 to 0.3 seems around the average value for all the teams. It is clearly seen from the graph that this factor is high for Rising Pune Supergiants. That means, out of all the matches this team have won in all the season, in most of the matches, this team have scored most of their runs in form of boundaries. Then, Rajasthan royals and Delhi Daredevils are having next higher values. A lower value of this factor means, that out of all the matches which this specific team have won, they have scored most of them in form of another form. Sunrisers Hyderabad and Royal challengers are one such team.

This factor is calculated independent of the team composition. This factor can be normalised with respect to the team composition.

### 7.3 Score Range

Score is the total number of runs scored by a team in a specific innings of the match. This factor is categorised into four categories. Each category has a range of value for runs. Based, on this runs, the team is classified into categories. The first category of team will have scored less than hundred runs. The second category of team have scored less than one hundred fifty. The third category of team have scored less than two hundred runs. The remaining teams comes under the fourth category.

This categorisation is used as one factor in determining the results. From the graph, it can be seen that this value is very high for some of the teams and considerably less for other teams. A higher value of this factor means that out of all the matches the specific team has won, most of the matches they have scored the runs in the higher categories ,or out of all the matches they lost, they have scores in the lower categories of the score.

From the graph 7 ,It is clear that, teams Kolkatta Knight Risers and Sun Risers Hyderabad have a value around 0.6. This implies that the wonmatch of column for this teams was mostly decided by the score range category column. For the teams like Gujarat Lions and Rising Pune Super giants, this value is low. It can be inferred that the wonmatch column for this table is mostly decided by other columns in the team statistics table.

### 7.4 Toss

Toss is another important factor considered for this analysis. The winner of the toss has the power to decide the sequence of the match. This decision of the winning captain will effect the results of the match. Given this opportunity, any captian would take decision in favour their side. Hence, toss is one important factor in deciding the results of the match.

A higher value of this factor implies, that out of all the matches the specific team has won, they have also won the toss in most of the matches. It also implies that the column wontoss have contributed a large amount to the target column. A lower value of this factor implies that out of all the matches a specific team has lost, most of the matches they have lost the toss.

From the graph 6 it is clear that the this value is higher for teams Delhi Daredevils, Gujarat Lions and Royal challengers bangalore. This implies that out of the matches they have won, most of the matches they have won the toss. This implies toss is one of the important factor for this team to win a specific match. It is clear from the graph that this value is lower for teams like Kolkatta Knight risers. This implies that out of all the matches won by this team, they have won toss in less number of matches. This implies that toss is not one of the important factors for this team to win.

For the other teams, this value is around 0.15 which is considered as average contribution. For these teams, toss decision have a fair impact on the decision of the match. The wontoss column in the team statistics have contributed fair amount to the target column.

## 8 STATISTICAL ANALYSIS

This analysis will give the statistics of each factors and their importance in the previous matches for each team. This is done by gathering the data from the team statistics which is prepared as part of the first step in the predictive analysis.

A dictionary of team names is prepared using the teams.csv file from the source. This file is read using the read csv method in the csv library. An empty list is created for the factors batting order, score range, toss. As there is only one kind of value for all these factors, they are grouped and studied together. The score range is studied in a different program.

These empty lists are used for storing the percentage contribution of each factor towards the results of analysis. In this case, the data is not divided into test and training sets. All the data in the data sets are taken into consideration. The values used in the analysis are the actuals value and no predictions are made in defining these values.

For every element in the team dictionary, we iterate over the specific team statistics csv file created in the first step of this predictive analysis. From these csv files , we read the columns battingsfirst, majorityscore, wontoss columns. For each columns in the csv file, we create a corresponding dataframe in the python program using the pandas dataframe. This dataframes are constructed on based on two values. The wonmatch column value and the value of the factor being studied. We append the corresponding row to the dataframe only if both the conditions are satisfied. This kind of conditional statements can be achieved by python or operator.

For every factor, now independent dataframes are defined after both the conditions are satisfied. Total matches is the length of the csv file. From the above calculate the percentage of the dataframe which we have captured with respect to the total length of the csv file. This percentage value determines the percentage contribution of the factor towards determining the winning chances of the team.

These percentages are calculated for all the teams in the team dictionary. These values are stored in the respective factors list. This list is used for plotting the bar graphs using the pythons matplotlib.

### 8.1 Analysis of individual team

From the the lists of percentage contributions from the above analysis. Analysis of each factor and their contribution towards the

individual team has been plotted in the graph. These plots are plotted against the team name and the percentage bars which show the percentage of each factor. Graph 2 has been plotted with above lists.

For the team Kolkatta knight riders, out of all the matches they have won, in forty percentage of those matches, they have score majority of their score in form of boundaries. Out of all the matches they have won, they also won toss in thirty percentage of the matches and batted first in twenty percentage of the matches. This implies that the probability of winning a match for kolkatta Knight riders team is high if they score moajority of their score in form of boundaries. Then decision of toss and sequence of matches have considerable effect in the matchesf decision.

Team Royal Challengers bangalore have followed the same trend as the kolkatta team in the analysis. Out of all the matches they have won, they also scored majority of the runs in the form of boundaries. While ,out of these matches, they won only twenty percent of the tosses, they batted first in nearly twenty five percentage of the matches. Runs have been major contributing factor in this team also.

Chennai Super Kings have majority of their contribution from the toss factor and batting sequence factor together. That implies that, from the all the matches they have won, they either won the toss or batted first in the match. This implies that the team Chennai Super kings will have to win toss or bat first to win the match.

Teams punjab and Kochi Tuskers have followed trend similar to that of Kolkatta Knight risers and royal challengers bangalore. They scored majority of their score in the form of boundaries in nearly forty percent of the matches played by them. The other factors toss and batting sequence have contributed to nearly twenty percentage of their winnings. This implies that scoring majority of runs in form of boundaries will favour these teams.

Mumbai Indians team have the second highest percentage contribution from the factors toss and batting sequence compared to other teams. This team also have third highes contribution from the score composition factor. That implies, this team have higher chance of winning given any factor. Though the value from the any one factor is against them, the other factor will over ride the effect of the previous factor.

Gujarat Lions have the highest contribution from the factor batting order. That implies, out of all the matches won by this team, they have scored majority of runs in the form of boundaries. The other two factors are considerably low. This implies that Gujarat team will have to score most of the runs in the form of boundaries to win the match.

It can be inferred from the graph that rising pune super giants have not won a match while batting first in the game. Out of all the matches won by them, they have either batted second in the game or score majority of runs in the form of boundaries.

Team Deccan chargers have almost same amount of contribution from all the three factors. This team is consistently performing in any values of the factors. They have good balance of the conditions in the previous games.

Team Pune warriors has the lowest values for the factors toss and sequence of the match. They are also have less contribution from the factor composition of score. This implies that this team is under performing in any given condition.

## 8.2 Analysis of Range of Scores

The scores were divided into categories. This analysis will contribute the percentage contribution of each range of the score to the result of the matches played by the specific team. This analysis can be used as to determine the safe score range for every team.

Team dictionary is taken from the team.csv file. The team statistics csv file which is created in the predictive analysis will be utilised for the value for range of scores. The values from the columns score range and wonmatch will be utilised. The data in this csv file is not partitioned into training and testing sets. This analysis is performed on complete data. No data is predicted in this analysis also. This is analysis on all the available data.

For every value in the team dictionary, we locate the team statistics csv file which is generated as part of the predictive analysis. Four empty lists are generated to store the number of matches won by each team with score in the given score range. Total number matches can be calculated by using the length of team statistics csv file that is being studied upon. For every value in the team dictionary, we divide the csv file into four different data frames. Each data frame corresponds to each range of scores. This partition can be achieved by using pandas data frame.

This dataframes are extracted using the two conditions. They are the value of score range must be equal to the range we are fetching data for and the other condition is the wonmatch column must be true.

After corresponding data frames are extracted from the team staistics csv, we calculate the percentage of each range data frame length against the total length of the team statistics csv. This percentage value will give us the percentage contribution of each range towards the result of the match.

## 8.3 Graphical Analysis on scores

From the graph 1, it is clear that most of the wining scores fall in the second and third category. That implies for any team to win the match, it is most likely that the team must score atleast hundred runs in the match and come under second category or score atleast one hundred fifty runs and come under third category.

Teams Gujarat lions and pune warriors have never won match scoring less than hundred runs or more than two hundred runs. This implies that the range from hundred to two hundred is the safe score range for these teams. Gujarat have won most number of matches scoring in the third category that is atleast scoring one hundred fifty runs and atmost two hundred runs. This is more safe zone for them. While, for pune warriors team they have almost similar winning percentage in both the categories. Teams rising pune supergiants and kochi have never won a game scoring more than two hundred runs. This might be bacuse, they have not scored more than two hundred runs in any given match or they have not won in the matches in which they scored more than two hundred runs. That implies that they are having a safe scoring in the range of one hundred fifty and two hundred.

Sunrisers Hyderabad team has more contribution from the the second and third categories of scoring in the game compared to the other two teams. This indicates that they are consistent in this category of scoring than other teams. They do not have major difference in contribution from these categories. This indicates that

they are consistently scoring around one hundred fifty mark score which makes the assumption that this team have good batting side. From the statistics on this team it is clear that they have also won matches scoring less than hundred. It is clear that they are also having a good bowling side as well.

Team Royal Challengers Bangalore has highest contribution from the fourth category of scoring when compared to other teams. That implies, the team royal challengers bangalore have highest probability of winning the matches , if they are scoring more than two hundred runs in the match. Then their next contribution comes from the score range of third category. That implies this team has a good batting side. Because they are having high scoring percentages from the third and fourth categories.

Following the royal challenger banglore team is Chennai super kings team. This team have contributions from all the four categories of the scoring range. This implies this team is fairly consistent in both the categories of the game.

Team Mumbai Indians have the highest contribution from the third category of the scoring range after Gujarat lions. They are also having contributions from all the categories of range of scores. This will clearly show that Mumbai Indians team has an inconsistent batting team. If these totals are from the second innings, it can also be assumed that whenever their batsman failed to score many runs, bowler won the match.

Kolkatta team has the highest contribution from the first category of scoring after Kochi team and Pune super giants. This implies that the above three teams have a good batting side. Scoring less than hundred and winning match is sight of team having a good bowling side. Upon having good contributions from category one, kolkatta knight riders also has good contributions from the other three categories as well. This implies, that this team not only has good bowling side,it also ahs a good batting side as well.

## 9 CONCLUSION

Predictive Data analytics have provided promising solutions to various problems in wide variety of fields. As part of this study, predictive and statistical data analytics on data related to a cricket League, Indian Premier league is conducted. As part of predictive analytics, the available data is split into training and testing data sets. The values from the model are used to predict the target value. These values are compared to the original values for accuracy. Method of improving the accuracy of model is studied. This study would be useful to determine the impact of a factor on the result of the game. This analysis would help in predicting the results of the matches acuurately with the model developed from the training data.

Statistical analysis on the same data is conducted to get the details related to the impact of a factor quantitatively on a specific team. This analysis pertains to the available data and no predictions are done. This kind of analysis would be useful in determining the strengths of individual team. This kind of analysis can be conducted to the nature and strengths of each team.

## ACKNOWLEDGMENTS

The author would like to thank Dr. Gregor von Laszewski for his support and suggestions to write this report.

The author would like to thank assistant instructors for their support in completing this project.

The author would like to acknowledge that the base data for the analysis is provided by [1].

## REFERENCES

- [1] HarshaVardhan. 2016. Indian Premier League. (2016). <https://www.kaggle.com/harsha547/indian-premier-league-csv-dataset/data>

[Figure 1 about here.]

#### LIST OF FIGURES

1	Score Staistics	10
2	Other Factors Staistics	10
3	Tree number vs Accuracy	10
4	BattingFirst	10
5	Majority of runs in boundaries	10
6	Toss	10
7	Range of Scores	10



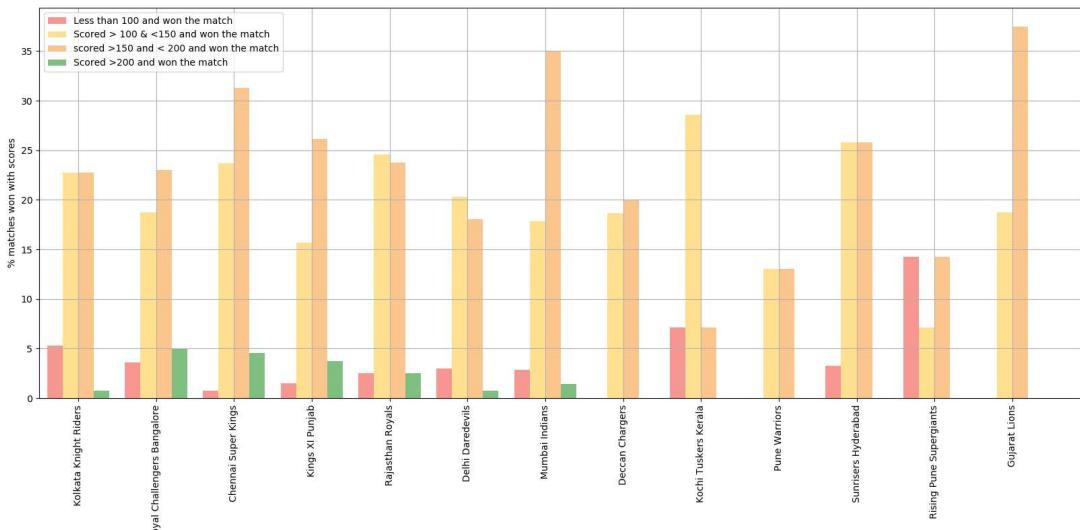


Figure 1: Score Staistics

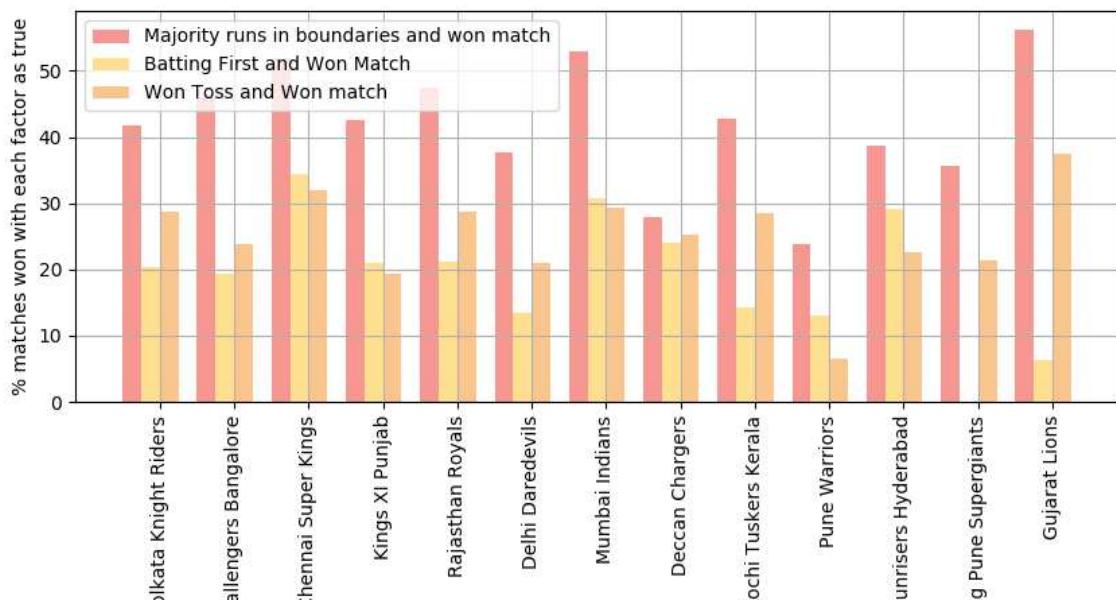
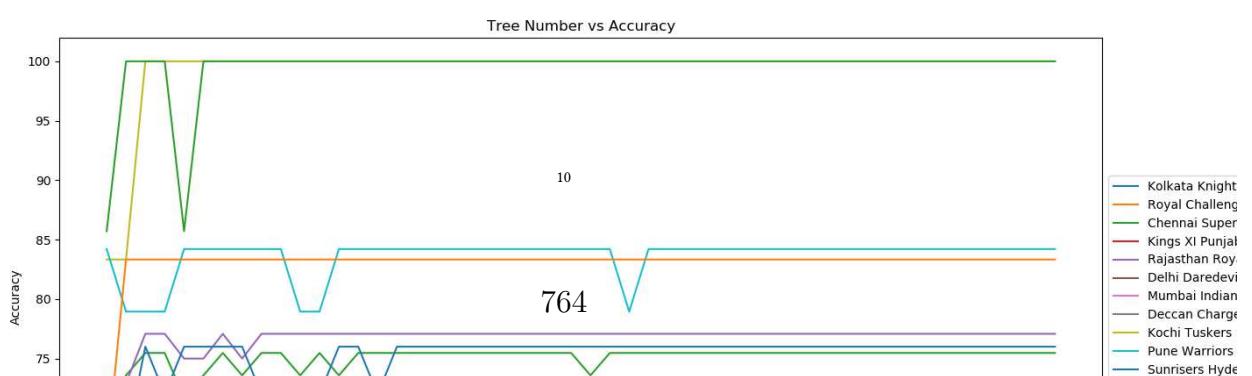


Figure 2: Other Factors Staistics



## bibtex report

```
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
```

bibtext \_ label error

bibtext space label error

bibtext comma label error

# latex report

[2017-12-11 13.27.25] pdflatex report.tex

```

Missing character: ""

bookmark level for unknown defaults to 0.
The anchor of a bookmark and its parent's must not be the same. Added a new anchor.
Float too large for page by 1818.09471pt.
Typesetting of "report.tex" completed in 1.7s.
./README.yml
 33:75   error   trailing spaces (trailing-spaces)
 36:81   error   line too long (82 > 80 characters) (line-length)
 36:82   error   trailing spaces (trailing-spaces)
 40:81   error   line too long (89 > 80 characters) (line-length)
 53:81   error   line too long (92 > 80 characters) (line-length)
 54:81   error   line too long (95 > 80 characters) (line-length)
 54:95   error   trailing spaces (trailing-spaces)
 55:81   error   line too long (91 > 80 characters) (line-length)
 56:81   error   line too long (91 > 80 characters) (line-length)
 56:91   error   trailing spaces (trailing-spaces)
 57:81   error   line too long (94 > 80 characters) (line-length)
 58:81   error   line too long (96 > 80 characters) (line-length)
 59:81   error   line too long (99 > 80 characters) (line-length)
 60:81   error   line too long (100 > 80 characters) (line-length)
 61:81   error   line too long (100 > 80 characters) (line-length)
 62:81   error   line too long (99 > 80 characters) (line-length)
 62:99   error   trailing spaces (trailing-spaces)
 63:81   error   line too long (99 > 80 characters) (line-length)
 64:81   error   line too long (101 > 80 characters) (line-length)
 65:81   error   line too long (103 > 80 characters) (line-length)

```

---

### Compliance Report

---

```

name: Swargam, Prashanth
hid: 228
paper1: Oct 20 17 100%
paper2: Nov 06 17 100%
project: Dec 04 17 100%

```

yamlcheck

---

```
wordcount
```

---

```
(null)
wc 228 project (null) 6688 report.tex
wc 228 project (null) 6475 report.pdf
wc 228 project (null) 24 report.bib
```

```
find "
```

---

```
passed: True
```

```
find footnote
```

---

```
passed: True
```

```
find input{format/i523}
```

---

```
4: \input{format/i523}
```

```
passed: True
```

```
find input{format/final}
```

---

```
passed: False
```

```
floats
```

---

134: In the graph `\ref{f:treenumvsaccuracy}`, the accuracy of using various number of decision trees are plotted against the number of decision trees. It can be observed that using less than five decision trees, the accuracy for the prediction for all the teams is around sixty percentage. As the number of decision trees increased, the accuracy of all the prediction is increased by at least ten percent. The hundred percent accuracy is because, those teams have played less number of games.

142: The first importance feature for all the teams are plotted in the graph `\ref{f:BattingFirst}` . It is clear from the graph that the importance feature batting first is highest for Rising Pune super

giants. This indicates that this team have won most of their previous matches with while batting first in the game. While the Chennai Super kings and Kochi Tuskers are also high, but they are less than half. This indicates that batting first in the match will be a favourable condition for the above mentioned teams. For all the other teams except Kings xi Punjab, Pune warriors and Sunrisers Hyderabad, the batting first importance factor is nearly around 0.1. This implies, out of all that matches which were present in the training set, these team batted first for only ten percent of the games.

- 148: Score Composition composition is the combination of different ways of getting runs. For any high scoring and successful game, a team will have to score runs at faster rate. Hence, scoring runs in form of boundaries will help a team a lot in turning the match in their favour. In the graph \ref{f:BoundariesMajority} , Contribution of this factor against each team is plotted. This graph summarises the contribution towards of the score composition factor towards the factor. This factor is very high for Kochi Tuskers team, because they have played considerably less number of games.
- 162: From the graph \ref{f:scorerange} ,It is clear that, teams Kolkatta Knight Risers and Sun Risers Hyderabad have a value around 0.6. This implies that the wonmatch of column for this teams was mostly decided by the score range category column. For the teams like Gujarat Lions and Rising Pune Super giants, this value is low. It can be inferred that the wonmatch column for this table is mostly decided by other columns in the team statistics table.
- 170: From the graph \ref{f:Wontoss} it is clear that the this value is higher for teams Delhi Daredevils, Gujarat Lions and Royal challengers bangalore. This implies that out of the matches they have won, most of the matches they have won the toss. This implies toss is one of the important factor for this team to win a specific match. It is clear from the graph that this value is lower for teams like Kolkatta Knight risers. This implies that out of all the matches won by this team, they have won toss in less number of matches. This implies that toss is not one of the important factors for this team to win.
- 189: From the the lists of percentage contributions from the above anaylsis. Analysis of each factor and their contribution towards the individual team has been plotted in the graph. These plots are plotted against the team name and the percentage bars which show the percentage of each factor.Graph \ref{f:otherstats} has been plotted with above lists.
- 229: From the graph \ref{f:scorestats}, it is clear that most of the wining scores fall in the second and third category. That implies

for any team to win the match, it is most likely that the team must score atleast hundred runs in the match and come under second category or score atleast one hundred fifty runs and come under third category.

```

270: \begin{figure} [!ht]
271: \centering\includegraphics[width=\columnwidth]{images/scorestatics.png}
272: \caption{Score Staistics}\label{f:scorestats}
274: \centering\includegraphics[width=\columnwidth]{images/otherstastics.png}
275: \caption{Other Factors Staistics}\label{f:otherstats}
277: \centering\includegraphics[width=\columnwidth]{images/treenumvsaccuracy.png}
278: \caption{Tree number vs Accuracy}\label{f:treenumvsaccuracy}
280: \centering\includegraphics[width=\columnwidth]{images/BattingFirst.png}
281: \caption{BattingFirst}\label{f:BattingFirst}
283: \centering\includegraphics[width=\columnwidth]{images/BoundariesMajority.png}
284: \caption{Majority of runs in boundaries}\label{f:BoundariesMajority}
286: \centering\includegraphics[width=\columnwidth]{images/Wontoss.png}
287: \caption{Toss}\label{f:Wontoss}
289: \centering\includegraphics[width=\columnwidth]{images/scorerange.png}
290: \caption{Range of Scores}\label{f:scorerange}

```

```

figures 1
tables 0
includegraphics 7
labels 7
refs 7
floats 1

```

```

False : ref check passed: (refs >= figures + tables)
False : label check passed: (refs >= figures + tables)
False : include graphics passed: (figures >= includegraphics)
True : check if all figures are refered to: (refs >= labels)

```

```

Label/ref check
passed: True

```

```

When using figures use columnwidth
[width=1.0\columnwidth]
do not change the number to a smaller fraction

```

```
find textwidth
```

---

```
passed: True
```

---

```
below_check
```

---

```
bibtex
```

---

```
label errors
```

```
bibtex errors
```

```
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
```

```
bibtex_empty_fields
```

---

```
entries in general should not be empty in bibtex
```

```
find ""
```

---

```
passed: True
```

---

```
ascii
```

---

```
non ascii found 8217
```

```
non ascii found 8217
```

```
=====
The following tests are optional
=====
```

```
Tip: newlines can often be replaced just by an empty line
```

```
find newline
=====
```

```
passed: True
cites should have a space before \cite{} but not before the {
```

```
find cite {
=====
```

```
passed: True
```

# Big Data in Job Recommendation Systems

Huiyi Chen

Indiana University Bloomington  
Bloomington, Indiana 47408  
huiyichen@indiana.edu

Yuanming Huang

Indiana University of Bloomington  
Bloomington, Indiana 47408  
huang226@indiana.edu

## ABSTRACT

In recent years, there are more and more recommendation systems merged as time goes on. It brought us a lot of convenience and improved the efficiency of finding resources by the development of technology. However, with the rapid growth of information, people feel hard to find the exact information they need in job searching sometimes. Thus, the birth of Job Recommendation satisfied the demand of finding exact job information as soon as possible and make this process more and more automatically. So, in this project, we will explore the system of Job recommendation to help people improve the efficiency in finding job information. *LATEX*

## KEYWORDS

Big Data, Job Recommendation Systems, Collaborative filtering, content-based filtering, Java, HID101, HID230

## 1 INTRODUCTION

As we browse on the internet nowadays, we realized that many websites push side advertisements that are related to what we have just browsed seconds ago. We were wondering how the internet has gotten so smart, yet it was a trick that many technology companies have made use of to better sort through the data they have about users and to generate profits more efficiently. These companies use data to filter out things that users like to read, which enhance users' browsing experience and therefore, use the website more often. The so-called recommendation system has been more and more popular and is now an integral part of many e-commerce sites such as Netflix, Amazon.com, Google, and etc. While these big commercial companies have been using the recommendation system for so long, we also want to analyze how some job websites can utilize this recommendation system to enhance user experience.

According to Wikipedia's definition, a recommendation system is a subclass of the information filtering system that seeks to forecast and predict the "preference" or "rating" that users would give to an item. In terms of a job recommendation system, it is the criteria of "interesting and useful" and "individualized" that separate the recommendation system from traditional search engine and information retrieval systems. It is more personalized for users that they could easily find the information they need without future clicks and research.

In regards of job recommendation systems, we want the system to recommend related job ads to job seekers based on their click history and basic information that they inserted in into the website such as majors, work history, extracurricular activities, skill sets, and etc. In other words, a job recommendation system has no difference than an automated form of a job search agency that not only it will show you a list of jobs that might fit you based on your information, but it will also utilized the relative jobs that

related users have gotten or interested in who share the similar backgrounds as you do. The system is well trained in up-selling and cross selling, just like an integrated job search agency. The fact that the job recommendation system could recommend personalized content based on past experience and related users' information bring users back to the website and keep using it.

## 2 IMPORTANCE OF JOB RECOMMENDATION SYSTEMS

With the rapid development and progression of the Internet in the world, people's lifestyle has undergone tremendous changes. In the past, people may need to look for employment information by reading newspapers and reading magazines, but nowadays, people's lifestyles are changed. They are more and more inseparable from the Internet and more and more employment information is directly put online. People start to search for jobs online as a more direct and convenient way. According to data from China Internet Network Information Center, as of 2014, the number of Internet users in China has reached 632 million. As a result, optimizing network information has become an increasingly important demand all around the world. Because of the development of the network, the amount of information increased more rapidly. However, people's ability to choose information can not keep up with the explosive growth of information, which leads to a huge conflict between information growth and information selection ability of people. Consequently, how to make this process more efficient by designing a system is becoming more and more important. On the other hand, the rapid development of the Internet brings us into a new era of big data. On the Internet, there are more and more different types of jobs, different types of requirements, and different types of recruitment processes. Job seekers often spend a lot of time to do research and browse for jobs information. However, they may still need to do something that is not their ideal position. Therefore, how to help job seekers intelligently to find the information they want in a short time has significant meaning at the time.

Besides helping the job seekers, we also need to focus on the company side. With the growing number internet companies, the competition is strong, and the companies need to compete against a big number of competitors. In order to differentiate themselves from others, using big data is a great method. Big data helps the company to load their websites more personalized and intelligent for the users. It gives the users what they want instead of giving everyone the homologous information. The fact that the users can user fewer clicks and less research time to find the information that fit their needs will increase the use and favor for certain websites. With that, internet companies can quickly differentiate themselves and increase profit. On the other hand, since the use of big data and recommendation systems becomes more and more popular, the ignorance of the technology might result in getting behind from the

competitors. As a result, recommendation systems are gaining high popularity, and companies need to learn how to properly utilize them.

### 3 RECOMMENDATION TECHNIQUES

According to Resnick, P. and Varian, H.'s *Recommender Systems*. Communications of the ACM, they mentioned that recommendation techniques have many possible classifications. It is not about the types of interfaces and the properties of users' interaction with the job recommendation systems, but it is more about the sources of data that the system is based on and also the use to which the data is put. More specifically, a job recommendation system needs to contain (i) background data, the information which the system retains before the job recommendation process starts, (ii) input data, the information which user must communicate to the system in order to generate a recommendation, for example, majors, work history, extracurricular activities, skill sets, and etc, and (iii) an algorithm that combines input and background data to arrive at its suggestions.

#### 3.1 Collaborative Filtering Algorithms

In order to understand how a job recommendation system works, we need to understand different recommendation systems approaches in order to pick the one that fits our needs the most. Collaborative filtering methods are based on a large section of collecting and analyzing information on users' activities, preferences and forecasting what job seekers will like based on similar job seekers who share similar background. One of the important advantages of the collaborative filtering approach is that it does not rely on machine analyzable content and therefore, it is capable of accurately recommending complex items such as a data science job without requiring and "understanding" of the job itself. Many algorithms are used in measuring item similarity and user similarity in recommendation systems.

Collaborative filtering algorithms are based on the assumptions that most of the people who agreed in the past will later agree in the future, and that they will tend to like similar types of items as they liked in the past.

When we are building a model from users' behaviors, we need to make a distinction between implicit and explicit forms of data collection.

Examples of the collection of implicit data could be the following:

- Observing the jobs that users view in the past.
- Analyzing job/user viewing times.
- Keeping a record of the jobs that users apply online.
- Obtaining a list of jobs that users have read to or researched on their computers.
- Analyzing the users' social network and discovering similar likes and dislikes.

Examples of the collection of explicit data could be the following:

- Asking users to rate a job on a number scale.
- Asking users to search.
- Asking users to make a rank of a collection of jobs from favorite to least favorite.

- Presenting two jobs to users and asking them to choose one of the better between the two.
- Asking users to come up with a list of jobs that they like.

The job recommendation system will compare the implicit and explicit data to similar and dissimilar data collected from outside resources and calculates a list of recommended jobs for the users.

Collaborative filtering methods often have these three drawbacks which are cold start, sparsity, and scalability.

- Cold start: These recommendation systems often require a great amount of pre-existing data on users in order to make some accurate recommendations. That means, if the data of one user is not comprehensive enough, there is a great possibility that the recommendation by the system does not align with the user's interest.
- Sparsity: The number of jobs posted on major job search sites is extremely large. The most involved and active users will only rate a small subset of the entire database. With that being said, even the most popular jobs will only have very few ratings.
- Scalability: In many of the system environments in which these recommendation systems make recommendations, there are over millions of users and jobs, which means a large amount of computation power is definitely required and necessary to calculate all of the recommendations for all the users.

Collaborative filtering approaches are classified as model-based and memory-based collaborative filtering. A well-known example of memory-based approaches is user-based algorithm and that of model-based approaches is Kernel-Mapping Recommender.

#### 3.2 Content-based Filtering Algorithms

Another popular approach that data scientists like to use to design job recommendation systems is content-based filtering. Content-based filtering approaches are based on profiles of the users' preferences and a description of the job item. In a content-based recommendation system, keywords are implemented to describe the jobs and users' profiles are built to indicate the types of jobs these users like. To put in a different way, this algorithm tries to recommend jobs that are similar to those that users viewed and liked in the past. In particular, various potential jobs are put together to compare with jobs previously rated by the users and the best-matching jobs are recommended. This particular approach has its origins in information filtering research and information retrieval.

To epitomize the features of the jobs in the recommendation system, an job presentation algorithm would be applied. A widely popular used algorithm is the  $tf\cdot tf \cdot idf$  representation.

To create a user profile and save it in the database, the recommendation system mostly wants to focus on two kinds of information:

- (1) A model of the users' preferences.
- (2) A history of the users' interactions with the recommendation system.

Generally, these methods use an job profile characterizing the job within the recommendation system. The system will create content-based profiles of users based on the item features. There might be different weights to every item in terms of users' preference. The weights that are assigned to each of the features depending

on the users' preference can be computed and calculated from individually rated content using a series of techniques. Simple approach uses the average rates of the item vector while other sophisticated approaches use machine learning techniques such as cluster analysis, Bayesian Classifiers, artificial neural networks and decision trees to calculate or estimate the potential probability that the users are going to like the job.

The feedback the system got directly from a user, usually in the form of a like or dislike button, can be used to allocate lower or higher weights on the importance of specific attributes.

An significant drawback with content-based filtering is that whether the recommendation system is actually capable to learn user preferences from users' historical actions regarding content sources and use the sources across all the other content types. When the recommendation system is limited to recommending jobs of the same types that the user is currently using, the retrieved value from the system is significantly less than that when another content type from other retrieving services could be recommended. For instance, recommending current news articles based on historical browsing of news is definitely useful, but it would be more useful when products, music, videos, discussions and etc. from different retrieving services could be recommended based on the news browsing.

A great example of content-based filtering approach being used in the real world is Pandora Radio. It plays music based on the user's initial feed to the recommendation system and deliver recommended music with similar characteristics. Besides that, most of the movie, music, and book recommendation systems are based on content-based filtering algorithm since this particular one works the best for personalization based on historical data.

### 3.3 Hybrid Recommendation Systems

Recent researches have demonstrated that hybrid approaches, combining both content-based filtering and collaborative filtering could be much more effective in many cases. Hybrid approaches could be implemented in the following ways: by adding collaborative-based approach to a content-based capabilities , or vice versa; by collaborative-based making and content-based predictions separately and later combining them; or by unifying both approaches into one integrated model. Recent studies empirically compare the performance of the hybrid recommendation systems with the pure collaborative and content-based methods and thus demonstrate that the hybrid approaches can provide much more accurate recommendations than the pure approaches. These approaches can also be used to overcome many of the common problems that happen in pure approach recommendation systems such as sparsity and the scalability problem.

One of the examples of the use of hybrid recommendation systems is Netflix. Although many movie and music recommendation systems use pure content-based approaches, Netflix chooses the strong hybrid approach that makes recommendations by using collaborative filtering, comparing the searching and watching behaviors of similar users as well as by content-based filtering, offering shows and movies which share similar characteristics with content that users have highly rated.

A variety numbers of techniques have been proposed as the fundamentals for recommendation systems: content-based, collaborative, demographic and knowledge-based techniques. Every single one of these techniques has some sort of drawbacks, such as the well-known scalability problem for collaborative approach and biased rating system for content-based systems. A hybrid recommendation system is one that integrates multiple techniques together in order to achieve some synergy between all of the approaches, minimizing the influences of those drawback from a specific method.

- Content-based: The system generates recommendations from particularly two sources: the ratings that a user has given them and the features associated with products. Content-based recommendation system treats recommendations as user-specific classification problems and learns a classifier for the users' likes and dislikes based on item features, which in our case, the job features.
- Collaborative: The system provides recommendations using exclusively information about the rating profiles from different users or jobs. Collaborative system locates similar users / jobs with rating histories similar to the incumbent user or job and generate recommendations using the neighborhood. The users based and the items based nearest neighbor algorithms can be integrated to deal with the cold start problem and thus, improve recommendation results.
- Knowledge-based: A knowledge-based recommendation system suggests jobs based on inferences about users' preferences and needs. This knowledge will contain explicit functional information about how certain job features would meet user needs.
- Demographic: A demographic recommendation system generates recommendations based on a demographic profile of the users. Recommended jobs can be retrieved for different demographic niches, combining the ratings of users in those niches.

The term hybrid recommendation system is used to describe any recommendation system that would combine multiple recommendation techniques as above together to produce the output. We can also combine several different techniques of the same type, for instance, two different content-based recommendation systems could function together, and many of the projects have investigated the type of hybrid: NewsDude, which has both kNN classifiers and naive Bayes in its news recommendations, is a good example.

There are seven hybridization techniques that are popular:

- Weighted techniques: The scores of different recommendation elements are combined numerically.
- Switching techniques: The recommendation system chooses among all the recommendation components and applies only the selected one.
- Mixed techniques: Recommendations from different recommendation systems are presented all together.
- Feature Combination techniques: Features that are derived from different knowledge resources are combined all together and delivered to a sole recommendation algorithm.
- Feature Augmentation techniques: Only one recommendation technique is used to calculate a set of features or a

feature, which is part of the input to the following technique.

- Cascade techniques: Recommendation systems are given strict pre-existed priority, with the lower priority ones breaking ties in the scoring of the higher ones.
- Meta-level techniques: Only one recommendation technique is applied at a time and produces a draft model, which is later the input that is used by the next technique.

## 4 HOW TO IMPLEMENT RECOMMENDATION SYSTEM

Recent researches has shown that there are increasing demands of Information Systems technologies for human resource management in the recruiting process in particular.

- (1) User Information Acquisition and Modeling: Because users have different interests and different industry preferences, we need to deal with the log files, find out users' explicit and implicit requirements, and then analyze and build the user mold.
- (2) Model Design and Implementation: At this stage, the main contents include the combination of feature variables, similarity calculation, positive and negative samples of mobile phones and internet devices, weight value calculation and knowledge classification logistic regression.
- (3) System Design and Implementation: The user model and big data platform combine to meet the needs of the company's job recommendation system.
- (4) System Verification and Comparison: The group calculates the conversion rate off-line, determines the characteristic variable combination and the similar algorithm of the recommended model and uses the filtering recommendation algorithms to compare and verify to get the optimal combination recommendation system.
- (5) The Application and Research of the System: It is necessary to establish the application framework of the recommendation system in other application fields to study how to integrate with other business systems of the enterprise and to realize the diversification of the recommendation system.

In order for the system to work, we need to utilize Hadoop as the platform for the recommendation system to be running on.

### 4.1 Based on Hadoop

The realization of our project is based on the Hadoop platform, so here we need to do a detailed study of Hadoop about the source and role of this platform. Hadoop is an open source framework for writing and running distributed applications for large-scale data, designed for offline and large-scale data analysis. It is not suitable for online transactions that randomly read and write to several records Processing mode. Just like HDFS (file system, data storage technology related) + Mapreduce (data processing), Hadoop data sources can be any kind of form and have better performance in dealing with semi-structured and unstructured data than relational databases , With more flexible processing power, no matter what form of data will eventually be converted to key / value. Key / value is the basic data unit. MapReduce can be used to replace SQL, the

standard query language, and MapReduce uses scripts and code, while Hadoop, which is used to relational databases and custom SQL, has an open source hive instead. Thus, we can understand that Hadoop is a distributed computing solution.

Hadoop features: Hadoop good log analysis, facebook to use Hive for log analysis, in 2009 Facebook had non-programmers, 30 percent of people use HiveQL for data analysis; China's Taobao search custom filtering is also used Hive; With Pig you can also do advanced data processing, including Twitter, LinkedIn to discover people you may know, and he can also implement recommendations similar to Amazon.com's collaborative filtering. And China's Taobao's product recommendation is such a process. At Yahoo, 40 percent of Hadoop jobs are run on pigs, including spam identification and filtering, and user feature modeling.

#### (1) Data integration

Data consolidation is called "enterprise data center" or "data lake." When users have different data sources, want to analyze their data. Such projects include getting data sources (real-time or batch) from all sources and storing them in hadoop. Sometimes this is the first step to becoming a "data-driven company"; sometimes you may only need a beautiful report. Enterprise Data Centers typically consist of HDFS file systems and tables in HIVE or IMPALA. In the future, HBase and Phoenix should have bigger development in big data integration and create a new situation to create a brand new world of beautiful data.

Often, salespeople love to say "read patterns," but in fact, to be successful, you have to know exactly what your own use cases will look like (Hive patterns do not look like you did in your enterprise data warehouse). The real reason is that a data lake has more horizontal scalability and much lower costs than Teradata and Netezza. Many people use Tableu and Excel when doing front-end analysis. Many sophisticated companies use "data scientists" as front ends with Zeppelin or IPython notebooks.

#### (2) professional analysis

Many data integration projects actually start with the analysis of specific needs and a data set system. These are often incredibly specific areas such as liquidity risk / Monte Carlo simulation in the banking sector. In the past, this professional analysis has relied on outdated, proprietary software packages, which often suffer from a limited feature set due to the inability to scale the data (largely because software vendors can not understand as much as professional organizations do).

In the world of Hadoop and Spark, take a look at these roughly the same data consolidation systems, but often have more, if not unique, HBase, custom non-SQL code, and fewer sources of data. More and more based on Spark.

#### (3) Hadoop as a service

Any large organization in a "professional analytics" project will inevitably start to feel "happy" (ie, ache) managing several differently configured Hadoop clusters, sometimes from different vendors. Next, they will say, "Maybe we should integrate these resource pools," rather than leaving most of the nodes idle most of the time. They should

make up cloud computing, but many companies often can not or do not because of security reasons. This usually means a lot of Docker container packages.

(4) Flow analysis

In general, flow analysis is a real-time version of an organization's batching. With anti-money laundering and fraud detection: why not on a transactional basis, take hold of it and not end it in a cycle? The same inventory management or anything else.

In some cases, this is a new type of trading system that analyzes the bits of a data bit because you are parallelizing it into an analysis system. These systems prove themselves as popular data stores such as Spark or Storm and Hbase. But flow analysis does not replace all forms of analysis, and for things you have never considered, people often want to analyze historical trends or look at past data.

(5) Complex event handling

Here we are talking about sub-second real-time event processing. While there is not yet fast enough for ultra-low latency (picosecond or nanosecond) applications like high-end trading systems, millisecond response times can be expected. Examples include real-time evaluation of call data records processed by internet telecommunication carriers for a thing or event. Sometimes you see that such systems use Spark and HBase, but in the end it usually has to be converted to Storm based on the interference mode developed by LMAX Exchange.

In the past, such systems have been based on custom messages or high performance from shelves, client-server messaging products - but today's data is overloaded. I have not used it yet, but the Apex project looks promising, claiming to be faster than Storm.

(6) ETL flow

Sometimes when we want to capture the stream data and store them up. These items usually coincide with No. 1 or No. 2, but add to their scope and characteristics. These are almost Kafka and Storm projects. Spark is also used, but there is no reason, because no memory analysis is required.

(7) Change or add SAS

We do not need to buy storage for your data scientists and analysts and you can "play" the data. In addition, you can do a few different things besides SAS can do or produce beautiful graphical analysis. This is your "data lake." Here is the IPython notebook (now) and Zeppelin (later). We use SAS to store the results.

These are all normal when I see other different types of Hadoop, Spark, or Storm projects every day. If you use Hadoop, you probably know about them. Some years ago I had implemented some of these projects, using other technologies. Although more and more things change, but the essence remains unchanged. We will find a lot of similarities, things you used to deploy and trendy technology are around the Hadoop Sphere rotation.

## 5 MAP AND REDUCE

Suppose the user wants to count a huge text file stored on a similar HDFS, want to know the frequency of occurrence of each word in this text. So start a MapReduce program. In the Map stage, hundreds of machines read all parts of this file at the same time and separately counted the frequencies of the parts they read separately, resulting in pairs like (hello, 1100), (world, 1214) These hundreds of machines each produced the same set, and then hundreds of other machines started Reduce processing. Reducer Machine A will receive all statistical results starting with Mapper Machine A, and Machine B will receive the vocabulary statistics beginning with B (but in fact it does not begin with a letter, but uses the function to generate a Hash value to avoid data Stringing. Since words beginning with X are certainly far fewer than the others, and we expect the workload of the data processing machines to differ too much). These reducers will then summarize again, (hello, 1100) + (hello, 1311) + (hello, 35881) = (hello, 38291). Each Reducer will do this, and eventually we get the word frequency result for the entire document.

This is a seemingly simple model, but many algorithms can be described using this model. The simple model of Map + Reduce, though easy to use, is also very cumbersome. The second generation of Tez and Spark In addition to new features such as memory caching, essentially the Map / Reduce model is more generic, blurring the boundaries between Map and Reduce, making data exchange more flexible and with less disk reads Write in order to more complex description of complex algorithms to achieve higher throughput.

With MapReduce, Tez, and Spark, programmers find it really troublesome to write MapReduce programs. So we want to simplify this process. We would like to have a higher level and more abstract language layer to describe the algorithm and data processing flow. So there is Pig and Hive. Pig is close to the script to describe MapReduce, Hive is using SQL. They translate scripts and SQL into MapReduce programs and throw them into computational engines for computation, and we're freed from cumbersome MapReduce programs and written in simpler and more intuitive languages.

With Hive, people found that SQL has a huge advantage over Java. One is that it is too easy to write. Just word frequency things, described in SQL on only one or two lines, MapReduce write about dozens of hundreds of rows. More importantly, users with non-computer backgrounds write SQL, so data analysts are finally freed from the dilemma of begging engineers and engineers are freed from the weird one-off handlers. This result makes the whole process more efficient. Hive has evolved into a big data warehouse core components. Even many company pipeline assembly is entirely SQL described, because easy to write and easy to maintain.

Since data analysts began to analyze data with Hive, they found that Hive was running too slow on MapReduce. But data analysis, people always want to run faster. For a huge site of massive data, this process may take dozens of minutes or even hours. And this analysis may be only a small part of the need to analyze how many people browse the electronic products, analysis of how many people read the Rachmaninoff CD, and so on, and then come to our proportion of the type of user. Due to the high demand for speed, the new Impala, Presto, Drill was born (and of course innumerable

non-famous interactive SQL engines). The core idea of fiQufiQuthe three systems is that the MapReduce engine is too slow because it is too generic, too strong, too conservative, and we SQL needs a lighter and faster access to resources, more specialized SQL optimization, and less Fault Tolerant Assurance (because of a system error, we can restart the task, if the entire processing time is shorter, such as a few minutes). These systems allow users to more quickly handle SQL tasks, sacrificing features such as general stability.

But in fact these systems, has not reached the desired level of popularity. Because at this time two new ones were made. They are Hive on Tez / Spark and SparkSQL. Their design philosophy is because MapReduce is slow, but if I run SQL with a new generation of general purpose computing engines like Tez or Spark, then I can run faster. And users do not need to maintain two systems. The above introduction, the basic structure is a data warehouse. The underlying HDFS runs above MapReduce / Tez / Spark and runs Hive, Pig on it. Or run Impala, Drill, Presto directly on HDFS. This solves the low-speed data processing requirements.

## 6 DESIGN SYSTEM MODEL

As a kind of data mining, recommendation system is one of the more special data mining systems. He embodies the system and user interaction and real-time. Recommend interest-based objects to users based on their hobbies or browsing behaviors, and further correct and optimize the recommendation results based on the feedback results of user interaction. In this professional recommendation system, there are mainly three parts, data collection, offline data processing and real-time online recommendation.

### 6.1 Data Collection

In the process of data collection. Because there are many ways for users to provide their preference information to the system, they can be divided into two kinds of explicit and implicit information. This information forms the basis of user behavior analysis. In this project, the main sources and channels of data are information about job-seekers registering, browsing jobs, and web-logging for job postings. User behavior categories: registration, browsing, residence time, job application. Their respective types of information are: explicit, implicit, implicit, implicit. The following is an explanation of the characteristics and actions of the four user behaviors.

Registration: job seekers registered behavior, including the basic characteristics of job seekers, registration information we can get job preferences, Through job seekers' preferences, we can get more precise career preferences.

Browse: job seekers on the job browsing information, through the frequency of the frequency of statistics, job seekers get the preference. This process can to some extent reflect their concern about job postings and the likelihood that they will be interested in positions. Thereby enhancing the accuracy of the analysis

Dwell time: The user's dwell time information analysis, you can know whether the user is interested in the content of the visit and the degree of concern, so as to get their preference information. The longer you stay on a page, the more likely they are to be interested in the content of the page, as well as the level of attention. However, there are occasional noise data that is difficult to use based on this standard.

Job Application: Boolean preferences, the value is 0 and 1. This information can be used to determine whether the user is interested in this position.

### 6.2 Offline Data Processing

Generally, the historical data of job seekers will be very large. Therefore, if the system wants to analyze massive data online and recommend it in real time, it is unrealistic. Therefore, if offline processing of data can make the data processing more Efficient and easy to implement. When we have collected enough user behavior data, we can pre-process the data off-line, such as noise reduction, and then analyze the user's behavior log and train user profiles through recommendation strategies. Finally, offline calculation of the user's character data and get the initial recommendation seen, the next step, the recommended results provided to the online implementation of the recommended use.

### 6.3 Online Real-Time Recommendation

The online real-time recommendation is to analyze the user's real-time behavior in a very short period of time and give the recommended result. Therefore, the online recommendation system and offline processing are two different processes and concepts. The online real-time recommendation can not process the user's historical behavior log and can not handle too complicated data. The online real-time recommendation usually deals with simple data, for example, querying the basic information of job seekers, job seekers applying. And then combine the result of the analysis with the user characteristic data that has been processed offline to get the final recommendation result to the job seekers through the multi-dimensional analysis of filtering, screening and recommendation ranking.

## 7 TECHNICAL MODEL

In order to build a job recommendation system, we need to implement a technical model. There are several things that we need to pay attention to when building the model.

### 7.1 Similarity retrieval technology

Content similarity retrieval technology refers to comparing the text feature information in a resource with the text feature value of a user's interest, comparing the user's historical preference information with the content feature of a resource, and calculating the similarity , Filtered out to meet the user's search expectations. An example of content similarity retrieval technology. First, you need to model the content and attributes of your position, for example, by industry, function, job type, and place of employment. And then through the characteristics of each position data to find the similarity between positions. If the professions, functions, types of jobs and workplaces are the same, we can think of these two posts as similar.

### 7.2 Demographic Collaborative Recommendation Technology

Demographic information can be viewed as a kind of user knowledge information that can be used to determine similar equivalence across networks, so that demographic information can be

considered as a synergistic approach. Collaborative demographic recommendation process: First, establish a data model for job seekers. Then according to job seekers model to calculate the similarity between job seekers. Find job seekers with the highest degree of similarity. Finally, recommend jobs to current job candidates based on their preferences. As a result, demographic data can initiate a referral system even when job seekers do not have a feedback evaluation of the position.

### 7.3 Collaborative Filtering Recommendation Techniques

Collaborative Filtering Recommendation Techniques is a technique that is widely used to predict user interest preferences. Its basic principle is based on the user's preference for the object, found the relevance of the object or user, and then recommend based on these correlations. The recommendation of collaborative filtering consists of three components: they are, Item-based recommendation, user-based recommendation and model-based recommendation. In this project, we mainly study this work recommendation system based on the idea of collaborative recommendation. That is to say, we need to train the recommendation model based on the sample data of job seekers' preferences and then make predictions based on the real-time information of job seekers Calculate recommended.

### 7.4 Big data processing technology

In this professional recommendation system, we assume that the big data processing framework that we need to use is the log acquisition system Flume, the big data platform Hadoop, the streaming computing framework Storm, and the message cache system Kafka.

Below we will introduce these different frameworks to understand the function and features of them.

**7.4.1 Flume.** Flume is a distributed, reliable, and highly available mass log collection, aggregation and delivery system. In this project, we anticipate that all the real-time input data needed will be realized through this technology platform. An Agent is the basic component of a Flume stream. An Agent contains Source, Channel, Sinks, and other components that use these components to pass an Event from one node to the next or for the final purpose. The data is finally encapsulated into an Event for transmission. The Source is used to accept an external source. For example, the Apache server delivers an Event to the source. The channel is used to temporarily store the Event. Sink then outputs the data. Other components can add pretreatment and classification capabilities.

**7.4.2 Hadoop.** At present, the large number of user data are obtained based on the site log records, job site for very large number of users, a simple single machine is difficult to complete offline data processing, so this project will be presumably we have introduced a distributed Framework to deal with, Hadoop is one of the most popular distributed framework today, which includes the distributed file system HDFS and distributed computing framework MapReduce and so on.

**7.4.3 Storm.** Storm is Twitter's open source distributed real-time computing flow data processing system, the calculation model is Topology as a unit, and a Topology is a series of Spout and Bolt formed by the graphic structure. Events Stream will flow between

Spout and Bolt, Spout will generate Event, and Bolt will logically process the received Event and get the result of the calculation.

**7.4.4 Kafka.** Kafka is a high-throughput, distributed messaging subscription system that is organized as a topic and can be used as an active data and offline processing system for real-time processing of caches between systems for offline and online data users Provide data pipes to handle activity data from different sources.

### 7.5 Mixed Recommendation Mechanism

In the popular web site, generally in order to achieve a better recommendation, we tend to mix a variety of recommended methods, rather than simply using a particular recommendation mechanism. This mechanism for mixing multiple recommended methods and strategies is called, Hybrid Recommendation Mechanism.

## 8 OVERALL GOAL OF THIS SYSTEM

In today's social conditions, job seekers want to find a suitable job is not easy. therefore. We need a home-based recommendation system for personal job seekers' information intelligence that uses big data mining and analytics technologies to change the traditional hiring process. With these ideas, we know that applying personalized recommendation techniques to job hunting, Work becomes easier, more efficient and smarter. In order to complete this system, we need to accomplish the following goals:

- (1) The recommendation system needs to intelligently collect explicit user needs and invisible user expectations information based on the confidentiality of user privacy information, and provide an effective data foundation for offline data processing and online real-time recommendation.
- (2) The recommended system should be able to handle a large amount of behavioral log data without affecting its normal use. Learning offline interests and training feature models through big data technology to provide data support for online real-time recommendation.
- (3) Recommended system must be based on user behavior online, rapid response to analyze user needs, in a timely manner to the user to make the list of their interest.

## 9 RECOMMENDATION SYSTEM FUNCTION DESCRIPTION

In this project, the user's occupation model, user interest in job hunting, job similarity, collection of positive and negative samples, machine learning and other algorithms related to the calculation of the model are based on the offline data processing module. However, its research is done under massive log file processing. Therefore, the system's offline data processing module is deployed on the Hadoop platform. As mentioned by the big data processing technology can know, Hadoop is one of the most popular distributed framework, including the distributed computing framework MapReduce and distributed file system HDFS.

Offline data processing results is to achieve career recommendation system online real-time recommended services important data support. The module externally requests the data from the data collection module by reading the request and reading the data

from the data collection module. The internal data processing request is executed through the general control interface. Various implementation algorithm models are run on the Hadoop platform and offline by MapReduce and HDFS. , The final results will be stored in the HBase database. Offline data processing module is mainly composed of the Hadoop platform running algorithms and algorithms to achieve the algorithm model, and the algorithm code, and by a calculation of the total control interface implementation of the call.

The essence of running algorithm is distributed framework Hadoop platform, which is mainly composed of MapReduce distributed computing framework and distributed file system HDFS. Job log files of job seekers are stored on HDFS, the log file data is cleaned by Pig Latin, formatted and stored in data warehouse Hive, and then processed by Hadoop Streaming stream or Hadoop Java API to calculate the job similarity, the job seeker's interest, etc. Series of MapReduce programs, model results processed offline will eventually be stored in the HBase database to provide data support for real-time analysis.

The implementation algorithm mainly includes the realization of the user-job model and the realization of the weighting coefficient. In this project, it takes a large amount of computation and time-consuming tasks to be completed offline. One of the most prominent features that need to be done off-line, such as user-job model calculation and machine learning, is that computing requires the user's historical behavior log, so implementing this aspect requires a computing platform that runs algorithms.

## 10 PERFORMANCE MEASURES

After the implementation of the job recommendation system, evaluation is always necessary and important in evaluating and assessing the effectiveness of the recommendation algorithms that we implemented. The most commonly used evaluation metrics are the root mean squared error and mean squared error, the former having been used in the Netflix Prize, one of the key events that energized research in recommendation systems. Such information retrieval metrics as recall and precision or DCG are very useful to assess the quality and performance of recommendation approaches. Recently, diversity, coverage, and novelty are also considered as important aspects in evaluating the recommendation systems. Results of so-called offline evaluations often do not correlate with actually assessed user-satisfaction.

To evaluate recommendation systems, we can use the popular concept of precision-recall. We need to be familiar with this in terms of the idea and classification is very similar.

- Recall:
  - What ratio of jobs that users like were actually recommended.
  - If a user likes 10 jobs and the recommendation decided to show 6 of them, then the recall is 0.6
- Precision
  - Out of all of the recommended jobs, how many does the user actually like?
  - If 10 items were recommended to the user out of which he or she liked 8 of them, then precision is 0.8

An ideal job recommendation system is the one that only recommends the jobs which a user likes. So in this case, precision=recall=1. This is an optimal recommendation, and we should try and get as close as possible.

## 11 CONCLUSIONS

In this paper we wrote, we used a comprehensive literature analysis of many fundamental elements, problems faced, and technical model related to the job recommendation systems. We also analyzed other popular recommendation systems that are widely used nowadays, and tried to implement the advantages of those systems into the job recommendation system. The job recommendation system technologies will accomplish significant success in broad ranges of applications and will be potentially a powerful searching and recommending techniques. Consequently, there is a great opportunity for applying these technologies in recruitment environment to improve the matching quality. Additionally, in order to have readers understand job recommendation problem, we detailed a series of recommendation techniques, how to implement the job recommendation system, and performance measurement methods. Finally, we plan as a continuation of this work to present a designed algorithm of job recommendation approaches that have been proposed to produce the best fit with design order and technical model.

## ACKNOWLEDGMENTS

The authors would like to thank Dr. Gregor von Laszewski for his support to write this paper as well as TAs' helpful suggestions on this paper.

## REFERENCES

```
bibtext report
```

---

```
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
I found no \citation commands---while reading file report.aux
Database file #1: report.bib
(There was 1 error message)
make[2]: *** [bibtex] Error 2
```

```
latex report
```

---

```
[2017-12-11 13.24.33] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflatex)
Missing character: ""
Empty 'thebibliography' environment.
Typesetting of "report.tex" completed in 1.0s.
./README.yml
17:14      warning  truthy value is not quoted  (truthy)
```

---

```
Compliance Report
```

---

```
name: Huiyi Chen
hid: 101
paper1: Oct 27th 100%
paper2: not yet started
project: Dec 09 17 100%
```

```
yamlcheck
```

---

wordcount

---

8

```
wc 101 project 8 7517 report.tex  
wc 101 project 8 7078 report.pdf  
wc 101 project 8 201 report.bib
```

find "

---

58: \par According to Wikipedia's definition, a recommendation system is a subclass of the information filtering system that seeks to forecast and predict the "preference" or "rating" that users would give to an item. In terms of a job recommendation system, it is the criteria of "interesting and useful" and "individualized" that separate the recommendation system from traditional search engine and information retrieval systems. It is more personalized for users that they could easily find the information they need without future clicks and research.

75: In order to understand how a job recommendation system work, we need to understand different recommendation systems approaches in order to pick the one that fit our need the most. Collaborative filtering methods are based on a large section of collecting and analyzing information on users' activities, preferences and forecasting what job seekers will like based on similar job seekers who share similar background. One of an important advantages of the collaborative filtering approach is that it does not rely on machine analyzable content and therefore, it is capable of accurately recommending complex items such as a data science job without requiring and "understanding" of the job itself. Many algorithms are used in measuring item similarity and user similarity in recommendation systems.

185: \par Data consolidation is called "enterprise data center" or "data lake." When users have different data sources, want to analyze their data. Such projects include getting data sources (real-time or batch) from all sources and storing them in hadoop. Sometimes this is the first step to becoming a "data-driven company"; sometimes you may only need a beautiful report. Enterprise Data Centers typically consist of HDFS file systems and tables in HIVE or IMPALA. In the future, HBase and Phoenix should have bigger development in big data integration and create a new situation to create a brand new world of beautiful data.

186: \par Often, salespeople love to say "read patterns," but in fact, to be successful, you have to know exactly what your own use cases will look like (Hive patterns do not look like you did in your enterprise data warehouse). The real reason is that a data lake has more horizontal scalability and much lower costs than Teradata and Netezza. Many people use Tableau and Excel when doing front-end analysis. Many sophisticated companies use "data scientists" as front ends with Zeppelin or IPython notebooks.

194: \par Any large organization in a "professional analytics" project will inevitably start to feel "happy" (ie, ache) managing several differently configured Hadoop clusters, sometimes from different vendors. Next, they will say, "Maybe we should integrate these resource pools," rather than leaving most of the nodes idle most of the time. They should make up cloud computing, but many companies often can not or do not because of security reasons. This usually means a lot of Docker container packages.

208: \par We do not need to buy storage for your data scientists and analysts and you can "play" the data. In addition, you can do a few different things besides SAS can do or produce beautiful graphical analysis. This is your "data lake." Here is the IPython notebook (now) and Zeppelin (later). We use SAS to store the results.

passed: False

find footnote

---

12: \renewcommand\footnotetextcopyrightpermission[1]{} % removes footnote with conference information in first column

passed: False

find input{format/i523}

---

passed: False

find input{format/final}

---

passed: False

floats

---

```
figures 0
tables 0
includegraphics 0
labels 0
refs 0
floats 0
```

```
True : ref check passed: (refs >= figures + tables)
True : label check passed: (refs >= figures + tables)
True : include graphics passed: (figures >= includegraphics)
True : check if all figures are refered to: (refs >= labels)
```

Label/ref check  
passed: True

When using figures use columnwidth  
[width=1.0\columnwidth]  
do not change the number to a smaller fraction

find textwidth

---

passed: True

below\_check

---

bibtex

---

label errors

17: las\_gergor00: do not use underscore in labels:

bibtex errors

```
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
```

```
I found no \citation commands---while reading file report.aux
Database file #1: report.bib
(There was 1 error message)
```

```
bibtex_empty_fields
```

---

```
entries in general should not be empty in bibtex
```

```
find ""
```

---

```
passed: True
```

```
ascii
```

---

```
non ascii found 8216
non ascii found 8217
non ascii found 8217
non ascii found 8211
non ascii found 8203
non ascii found 8203
```

```
=====
The following tests are optional
=====
```

```
Tip: newlines can often be replaced just by an empty line
```

```
find newline
```

---

```
passed: True
```

```
cites should have a space before \cite{} but not before the {
```

```
find cite {
```

---

```
passed: True
```

# **Big Data Analytics in Support Filtering Wrong Informations On Social Networking Sites**

Juan Ni

Bloomington, Indiana 47401

nijuaniu.edu

## **ABSTRACT**

In an era of information, people are more likely to get information from the ciber world. Due to the conflict of interest, many organizations hire “Spammers” to post a mass of wrong comments under some famous person’s post on Social Networking Sites for control the trend of public opinion[2]. So when user want to see the public opinion under some famous person’s post, they usually get the wrong information which doesn’t represent the real public opinion. Big Data analytic can provide information filtering to screening the fake comment based on data mining technique, and let the user be able to see the true of public opinion on social networking sites.

## **KEYWORDS**

523, HID 107, project, big data, weibo, spammers, data visualizationThChina

## **1 INTRODUCTION**

In highly informatization modern society, internet especially for social networking website carried mass information data. Along with the growth in users at social networking websites, social networking website become a platform which content infinite potential business opportunities and interests. Famous social networking website like Facebook and twitter already became leader which can drive the public opinion direction. I think the influence by social networking websites is completely different than traditional social media, social networking websites are more emphasize on audience’s acceptance and follow suit. For example, some politician announce some idea on news paper and TV news, the influence from it only work when it can make arouse sympathy to the audience, which mean it only work when the audience think it make sense for them. But social networking websites are difference, Seiter mention that“ Comments are a powerful emotional driver. Make the most of them by engaging often with your Facebook community and replying to fans’ comments to keep the conversation going.”[? ]. Social netorking website can make the user believed their opinion by drive their user follow the crowd because people love to view the comments and post comment, “A previous study showed that 45% of users on a social networking site readily click on links posted by their fifriendfi accounts, even if they do not know that person in real life” [? ]. For example, the President of United states Trump really like post his opinion via twitter, we can see lot of people post comments under his tweets, and also many people retweet his tweets. According to Seiter’s statistic, “to let others know what I believe in and who I really am (37%)” [2] is place on the fourth position at social website seeking primarily ranking. This draw a conclusion that if people saw retweets or comments from some Twitter who they believe that twitter is believable, people

will believe the retweets and comments from that twitter is making sense for them. Furthermore, the power of comments under the hot tweets is really powerful because making comment make user no longer be a spectator, they are actually involve into the event and be part of the society. Then, the comments with most retweets will gain people’s trust, and making people think that comment is represent the main strain. So this is how social networking websites impact the main social opinion trends.

## **2 THE ADVERSE EFFECT FROM SPAMMERS**

The social opinion trend at social networking community will drive personal and even company decision, then some trends might harm someone’s benefit because the power of social opinion trend is so powerful. Then people hired spammers to spread wrong information that lead the trend become advantage for them, but the users are become victim because they will make wrong decision because of the trend is control by someone on purpose. “Brown showed how it would be possible for spammers to craft targeted spam by leveraging the information available in online social networks.” [5], every spammer post must for some reason that beneficial for their employer, the most famous case for spammers the shampoo case at 2010. ”BaWang shampoo” is the most famous shampoo at China which advertised by super star Jackie Chan, ”Next Magazine” post a fake news claimed that using ”BaWang shampoo” could cause cancer [3]. I clear remember at that time, almost all social websites post new claim ”BaWang shampoo” is harmful at the same time without any authority judgment, and they put this new at the headline position to abstract user’s eye-ball. Even the authority department proof this new is unreliable, the business reputation of ”BaWang shampoo” had been damaged, lot of people around me stop using this shampoo any more. This case seems have no spammers involved, but actually the spammers for this case is social networking websites themselves instead of single person. The reason why they post slander is because they can get benefit from other shampoo companies in china, other shampoo companies can have more sales because the market-share of ”BaWang shampoo” will be decrease at this case.

The other reason why spammers getting so popular at social networking website is the operating cost of spammers is supper low. Try searching ”buy Facebook like” at Google, and there are over hundred million results come up. And the price of buying like and followers from that website is pretty low, so spammers can get an account with 1000 followers which look like a real account for only 5 dollar [6]. So spammer can create thousands this kind of fake account for posting the wrong information at social networking website, and the detected system is hard to find out those kind of account is real account or Zombie account those accounts have actually followers and like, even feeds. Then spammers can

use script to post batch of wrong information via those account. Furthermore, the cost for post batch of wrong information is unbelievable low; according to the internal information which provided from my friend who working at a IT company, there two ways to post wrong information at social networking website, first one is money reward system, second one is posting AI. For the reward system, a professional spammer company usually have about 10 teams, each team has 500 people; They use reward instead of constant salary, each feed related to the order topic is worth 0.5 Chinese dollar(equal to 5 cent in dollar), and each comment on the target feed is worth 0.2 Chinese dollar( equal to 3 cent in dollar), and the price of long text post is negotiated. The spammers accounts are provide from the company, the price for those account is also low; Most social networking websites only require email address for registered, then they buy email accounts from the retail like 100 Chinese dollar(equal to 15 dollar) for ten thousand accounts, and using script to register social websites accounts and making follow with each fake accounts. This is how they operate the spammer, now most social networking websites register require phone number verify, then they working with local sim card retail which infinity phone number on hand, but the price for each fake account is increase a lot like 3 Chinese dollar(like 50cent) per one, but still pretty cheap compare with other advertise way. Cheap labor force and the development of script technology rising the spammer company, but the lower the user experience at social websites because lot of trash information full of the social websites, people are hard to see the true at social websites any more. According to the network sites worldwide ranking[7], we can see WeChat has almost twice active user than Weibo, and WeChat is the most popular social networking app in China because spammer can not do any at that app. In WeChat, they post feed at the module which call "Friends circle", the feed formatting is pretty similar as weibo, but the app is semi-closed which mean it is complete private, user can only see their friend's post and the comment, no retweet allowed, and if someone who not in the user friend's list comment at user's feeds, user can not see that guys comments. Plenty of users quit traditional social networking website, and the semi-closed social networking app is getting popular, so the adverse effect of spammer is not only for the spam target but also for the platform. If we let spammer keep development, and don't have any way to filter their information, the traditional social websites will die soon.

### 3 BACKGROUND

According to the worldwide statistics data, "Sina Weibo" has 368 millions active users which more than 328 millions of twitter active user [? ], so I would like to using "Sina Weibo" as my investigate target instead of using Twitter. The other reason why I choose "Sina Weibo" as my investigate target is because I'm familiar with Chinese culture and I have been using "Sina Weibo" for more than 8 years. I think my knowledge about "Sina Weibo" will help me a lot at this project and better understand how spammers works at "weibo". The page frame at "weibo" is pretty similar to Twitter [figure1].

[Figure 1 about here.]

The four buttons under each feed are "collect, retweet, reply, like", and the capability for each button is same as twitter. When

user click into the "rely" button, user can see all the comments related to the current feed, and sort them by the amount of "like" that comments get from other user. The only things differen at "weibo" is user not only can see the comments but also can see the retweet information, twitter only allow user to see who retweet the feed. Then user can see the retweet's comments and sort the list by the retweet's times of the retweet feed. So people would love to check the retweet list to see which famous person retweet the feed, and what comment they put into the retweet. Spammer control the public opinion trends by putting wrong information that doesn't represent real public opinion into the comment for some hot feed, they utilize user's habit to reach their goal.

### 4 RELATED WORK

There are many researcher done previous research about how to distinguish the authenticity of information that post on social networking websites. Kr point out the user's social networking structure and the user's feed can represent the credible of the user, and kr using different order algorithm to rank the credible order based on the user's social networking structure and user's feed, and use it to judge the user whether is spammer or not [7]. According to Liu's idea, personal information source is really important for judge the source is reliable or not, like the user register time, the user operating frequency, and the relationship between the user and comment target will be three factors for supervise fake account [17]. In lou's article, he mention that we need to analysis the content and feed for judge the reliable level of information, he also point out the if only investigated the comment, retweet for detect spammer, it is hard to reach automatically fast and accuracy result, which mean it still require operator to control the analysis application [15]. Xu has really unique investigate area, she investigate the spammer in online business platform, and her idea can be work on social networking website [14]. In her article, she focusing on the speciality of spammer's behavior in online business websites, she collect sixty thousand comments and thirteen thousand product information that related to those comments at Amazon, she use those data to analysis the characteristic of user behaviour and set up a classifier for different characteristic of user behaviour; she also use the relationship between different spammers to improve the level of accuracy for detecting spammer.

### 5 METHOD

The method for Filtering spammer's Information at social networking websites can be divided into two part, first part is collecting data and the second part is produce data. Collecting data is the main part at this project because any analysis must base on the data, if the application can not collect the target data from third party platform, then is no way to start analysis.

#### 5.1 Data Collection

Using python 2.7 to collecting data from Weibo, and using the official SDK as my accessed method. First, setting a feed as the investigated target, I using [https://weibo.com/5305999252/Fy0sio7nQ?from=page\\_1005055305999252\\_profile&wvr=6&mod=weibotime&type=comment](https://weibo.com/5305999252/Fy0sio7nQ?from=page_1005055305999252_profile&wvr=6&mod=weibotime&type=comment) this feed to investigated the comment content. The person send this feed is my favorite gaming live streaming player, his name

is LuBen Wei. He is the most popular gaming live streaming in China, there are over four million audiences what his playing game every night. Moreover, this is his second account, so he always post some feeds that can not be post at his official account at Weibo, but there are still thirty thousand comments under this feed , the number of comments at this feed even more than the comments under every signaler feed from Donald J. Trump's account. And the feed's content is he complain about the cheating case, he announced that he never cheating at "PlayerUnknown's Battlegrounds", he claim that the rumor about his cheating is come from the spammers. After he post this feed, this feed became the top 1 hot feed at the feed ranking at Webo, and most comments under this feed are abuse him cheating. So I though there are must be spammer working under this feed, the comments at a feed from a gaming live streaming player's second account is more than the comments from United states's president's feed which is so ridiculous. Therefor I think there must be spammer involved into this feed, that is how we pick up the feed which involved spammer in social networking website. If a feed has unusual comments and likes amount compare with other feed post from the owner, that feed have huge possibility that involved spammer work.

First of all, Weibo require we use Weibo API with authentication, so we need to create a personal application first at the weibo application apply page [13]. Then the weibo official suggest us to use SDK to access the the API, so I came to the sdk websites [8] to get the Weibo SDK package. Normally can just type "pip install sinaweibopy" to install this sdk package to python, and also can download the sdk package, and put the webo.py with the py files I using to collect data into the same fiddler to use this sdk. I using the second method because I have issue pop this sdk. User can get the direction of how to use sdk via the weibo sdk wiki page [10], they provide many tutorial about how to use sdk on different environment not only for pyhton. For using Offical sdk, we need to use the "app\_key" and "app\_secret", we can find those code from the application page which I create the app apply before using my account. Those two codes are represent the user identity of who using the API, so weibo will ban the user's weibo account if they do something bad via weibo API because those two codes are directly link to user's weibo account. For getting the autorized for using the weibo API, I using Thinkgamer\_gyt's idea to get the authorized code [9], Weibo using OAuth 2 to check the user identity for using API. After the authorized page pump up, enter code which from the page url link which look like <https://api.weibo.com/oauth2/default.html?code=2024222384d5dc88316d21675259d73a>, and the code we need to enter is the string that after "code=" at the url link. ; then weibo will return an the access token for the API, then we using "Client" to activate the API, so we can get our target information via "Client".

After finishing open the API, the next step is to allocate the target feed page. For target the specific feed, we need to find out the id for each feed. Weibo is pretty tricky, they hide the real id and replace it as some codes at the feed URL, so we need to decode the Url to get the real feed id. [https://weibo.com/5305999252/Fy0sio7nQ?from=page\\_1005055305999252\\_profile&wvr=6&mod=weibotime&type=comment](https://weibo.com/5305999252/Fy0sio7nQ?from=page_1005055305999252_profile&wvr=6&mod=weibotime&type=comment) this link is my target feed link, take a look on the link, we can find out the the code before the question mark is pretty much look like the encryption feed id. Xuebuyuan find out the encryption rule for

feed id [16], based on his idea, each four characters from back to front is a group as sixty binary, and switch those sixty binary to ten binary and then link them together. I using his code to decode the feed id at the ipython notebook, so user want to change their focus feed, they can enter the different codes from their focus feed's url, then they can get their feed id.

Once we get the feed's id, we can use this id to allocate the target feed at API. The next step is to get the target information we want to analysis from the feed. We are looking for the comments information from this page, so we need to know the code about the API port for our target information. Weibo provide a API port instructions to guide the user how to access different information via API [12], the access port we are looking for is comment. The code of comment port can not directly use at python, then use dot instead of slash for the comment for fit the python coding rule. Then following the access port guide, setting the parameter like feed's id, the number of page we want to have for the comment, and the number of comment we want to have for each comment page. For here, I only set up return 200 comments from page one because set limited usage of API for each account, so each account only get 2000 comments via API if the account is using free API connection, I don't want to use all the attempts chances at once. Then we use the access token which we got in the previous section to open the API and active the data port we set up for the target information. Finally set a variable to storage the data which return from the comment port.

The last step for collecting data is to storage the target data as txt file. All the data that related to comment are saved into the variable now, so if we want to get our target return object from this data set, we need to use the special code for different kind of category inside this data set; Weibo also provide a specific instruction for the code of return object [11]. The data we need for wrong information filtration are the content of comments and the user information for each comments, the numbers of follower and the number of friends for the user who post the comments, also we need to get the number of feed that post from the user who comments the feed. Then use special code "follower\_count", "FRIENDS\_count", and "statuses\_count" to get those information, and save them into the txt file for next step data vualization. The reason why I save my target date into txt file because of the coding knowledge shortage, I don't know how to using the data analysis model at 2.7 python version, so I decide to using Python 2.7 to collecting data and using python 3.5 to do the data analysis.

## 5.2 Data visualization

The data visualization in this paper will be simple and straight forward, because of even I have some idea to analysis the data, but I can not represent it due to coding knowledge shortage. The models I decide to use for this data visualization are matplotlib, nltk, wordcloud, pandas, numpy, jieba and codecs. First, open the content.txt file at python and using readlines and appends to create a list that content all the comments. I intend to use worldcloud to visualize the words that has most frequency on the comment list, and I find out the worldcloud don't support chineeses really well, so I use the module "jieba" to reproduce the comments list. Jieba is the best module to support Word Segmentation, wen can using

this module to pick up the words which most frequently appear at the comment list[4]. I using FontTian's formatting as my main structure of jieba code [4], also we can add some new word into our word list and use it to make the jieba module can be able to indentify the new world, and we use "stopwords\_path" to filter the common word like "hello", and the we are using outside txt source which is Chinese vocabulary words out file as our stop word dictionary [1]. And during using this stop word file inside the jieba code, we have to encoding the file to "utf8" formatting, otherwise it will have some error that the jieba module can not distinguish the content inside the stop word file; also we need to set up the right font for the wordcloud by using the ttc file from the fonts document on the computer, if the font use on wordcloud doesn't support chinese, the final result will be a retangle for each word instand of actually Chinese.

The first outcome I got from the word cloud is look like this:

[Figure 2 about here.]

So now people can really quick have a pretty idea that what is the main trend of the comments, and what is all the comment talking about. But we can still find some noise inside the plt, like "ffh" which mean replyTh it doesn't contain any meaning, so we can add this word to our stop words list, then we try to create world cloud again, the new plot is look like this:

[Figure 3 about here.]

It seems more meaniful than before, and contain more information than before. So we can see that their lot of words like "ffh" inside the comments data is useless, then we can add those word into our stop words list to rip it up. Also we can use this method to filter the spammer's information, so the user can see the true information from the world cloud.

Then for detecting the spammer, analysis the user who post the comments is very important. To visualize the user data, using readlines to open the each txt file, and using solit and strip to reproduce the data formatting. Then putting all the data into the dataframe via pandas modules. When I look at the data type in dataframe, I found out that the data type inside the dataframe for each catory is not number, then I use astype to change the data type to number for calculator. Here is the three histograms I plot out for the the Feeds, Friends, and followers data:

[Figure 4 about here.]

[Figure 5 about here.]

[Figure 6 about here.]

The result is pretty interesting and surprise for me. The data represent those three histograms are pretty obviously, even the number of my sample is only 200 comments because of the limitation of weibo API. We can see that all the histograms are right-skewed distribution, according to the definition of histogram, the mean at right skewed distribution is the peak of the right side [1]. It doesn't look like normal because of the curve is not bell shape, if this data is from the real commets which mean post by the actual user not spammer, the curve of this three histograms will looks like normal distribution. The peal of those three histograms show us the majority of those three elements, for the number of friends, the majority is between 0 to 250; for the majority of feeds is between 0 to 1800, and the majority for followers is between 0 to 200. we

can see most majority are fall into the first interval, which mean it represent the friends, followers, and feed from the accounts I pick are pretty much same type of fake account. So it is pretty luck that I can dig out so much spammer account with small number of simple, so there no doubt that their are lot of spammers involve into this feed's comment because the comments post by the majority accounts are pretty much same type of account.

## 6 FUTURE WORK

The above sections bring out the idea about how to detect the spammer, and the method is pretty sample because of it just to use for proof my idea is feasible. To rise the filter wrong information to the big data level, we can using database to storage our data instead of storage the data into a txt files. We can make a connection between API and mysql, and the programming will automatically storage the data inside each table for different data category. Also the authorized code can be automatically get from the authorized page, so user no need to enter it by hand. All the analysis will be integration into one application, so user only need to copy and paste the feed link that they want to see the true for that feed, and the application will decode the feed id and collecting all kind of data into the mysql database. For collecting huge size of data like over ten thousand comments, we can using different virtual machine to get data from the API, so we don't need to worry about the daily API usage any more. Furthermore, using machine learning to train a model that can recognize the most frequency word that spammer use to post the wrong information, and use it the find out the spammer on the comment list. Finally, according to the comment list data analysis, save the user name which are define as spammer in database, and then remove the comments that related to those user in the comments content list, use this new comments content list to create wordcloud to represent the true trend and focus point for user's target feed. Moreover, the spammer data on the database will be cumulative, so the application analysis more fee, the spammer list will be increase, then each time the application can remove the comments which post by the spammer on the spammer list before analysis the spammer on the comments list which can improve the precision ratio of eliminate spammer information a lot. I think big data is based on the data precipitation, so most big data application won't have good performance at the begin because lack of data, when the application produce and save data until certain level, the performance the application will be increase.

## 7 CONCLUSIONS

The power of social opinion not only effect the ciber world, but also have great impact on real world. Therefore, it is really important to let the user in ciber world getting right information for the content that they are interesting in; the best way to achieve this gold is using application that base on big data analysis to filtering the wrong information that post by the spammer. There lot of ways to filtering the wrong information, but the collecting related data are always same, because no matter using which way to analysis the data, getting data is top priority than any things. I believe current technology can support big data storage really well, when the data storage reach certain amount, we can use it to decontaminate the ciber world, and maintain the ciber world envirnment that allow

people gain real information and create real social relationship on it. Therefore, improve the accuracy and adaptation for spammer will be meaningful to investigated.

## 8 ACKNOWLEDGEMENT

I would like to take this chance to thanks to my tutor Miao, in process on reviewing my paper, he gave me many useful comments and advises. At the same time, I would like to thanks my instructor laszewsk, give me useful knowledge about how to write a report on Latex format. Finally, I would love to thanks my friends who working at IT company give me many idea about how spammer work.

## REFERENCES

- [1] ASQ. 2017. Typical Histogram Shapes and What They Mean. (2017). <http://asq.org/learn-about-quality/data-collection-analysis-tools/overview/histogram2.html> [Online; accessed 1-Dec-2017].
- [2] Christina Hills. 2017. DIFFERENCE BETWEEN SPAMMERS AND HACKERS. (2017). <https://websitecreationworkshop.com/blog/wordpress-tips/difference-spammers-hackers/> [Online; accessed 1-Dec-2017].
- [3] Eddie Lee. 2016. Next Magazine to pay BaWang shampoo makers HK\$3 million compensation for defamation. (2016). <http://www.scmp.com/news/hong-kong/law-crime/article/1951576/next-magazine-pay-bawang-shampoo-makers-hk3-million> [Online; accessed 1-Dec-2017].
- [4] fxsjy. 2017. jieba. (2017). <https://github.com/fxsjy/jieba> [Online; accessed 1-Dec-2017].
- [5] Garrett Brown and Travis Howe and Micheal Ihbe and Atul Prakash and Kevin Borders. 2017. Social Networks and Context-Aware Spam. (2017). [http://web.eecs.umich.edu/~aprakash/papers/cscw08\\_socialnetworkspam.pdf](http://web.eecs.umich.edu/~aprakash/papers/cscw08_socialnetworkspam.pdf) [Online; accessed 1-Dec-2017].
- [6] isocialfame. 2017. Buy Real Facebook Page Likes+Followers (Business Pages). (2017). <https://isocialfame.com/collections/facebook-marketing/products/buy-facebook-fan-page-likes?variant=509391732763> [Online; accessed 1-Dec-2017].
- [7] KR Canini and B Suh and PL Pirolli. 2017. Finding Credible Information Sources in Social Networks Based on Content and Social Structure. (2017). <http://www.parc.com/content/attachments/finding-credible-information-preprint.pdf> [Online; accessed 1-Dec-2017].
- [8] michaelliao. 2017. sdk. (2017). <http://github.liaoxuefeng.com/sinaweibopy/> [Online; accessed 1-Dec-2017].
- [9] Thinkgamer. 2017. Weibo API using guide. (2017). <http://blog.csdn.net/gamer-gyt/article/details/51839159> [Online; accessed 1-Dec-2017].
- [10] Weibo. 2017. SDK. (2017). [http://open.weibo.com/wiki/SDK#Python\\_SDK](http://open.weibo.com/wiki/SDK#Python_SDK) [Online; accessed 1-Dec-2017].
- [11] weibo. 2017. weibo API return object code. (2017). <http://open.weibo.com/wiki/> [Online; accessed 1-Dec-2017].
- [12] weibo. 2017. weibo API wiki. (2017). <http://open.weibo.com/wiki/weiboAPI> [Online; accessed 1-Dec-2017].
- [13] weibo. 2017. Weibo application. (2017). <http://open.weibo.com/apps> [Online; accessed 1-Dec-2017].
- [14] Xu Chang. 2013. Detecting collusive spammers in online review communities. (2013). <http://www.ixueshu.com/document/43b579eeddbe46b2318947a18e7f9386.html> [Online; accessed 1-Dec-2017].
- [15] xudong lou and pin liu. 2011. analysis the spread of spammer. (2011). <http://www.ixueshu.com/document/43b579eeddbe46b2318947a18e7f9386.html> [Online; accessed 1-Dec-2017].
- [16] xuebuyuan. 2007. get mid. (2007). <http://www.xuebuyuan.com/1874313.html> [Online; accessed 1-Dec-2017].
- [17] zhibin liu and lanhua deng. 2017. analysis the credible in network information. (2017). <http://www.ixueshu.com/document/1453923d337e1742318947a18e7f9386.html> [Online; accessed 1-Dec-2017].

LIST OF FIGURES



联合国 V

12月3日 22:37 来自 iPhone 8 Plus

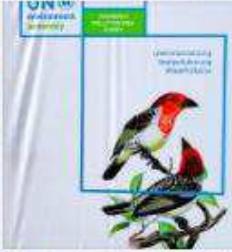


v

和@微公益一起近距离关注 2017 #联合国环境大会# @联合国环境规划署 @李晨

@微公益 V

#联合国环境大会# 即将开幕！微公益带你一起走进肯尼亚，迈向#零污染地球#！@联合国环境规划署 @熊猫守护者 @微环保



12月3日 17:16 来自 荣耀9 美得有声有色

126 | 15 | 76

☆ 收藏

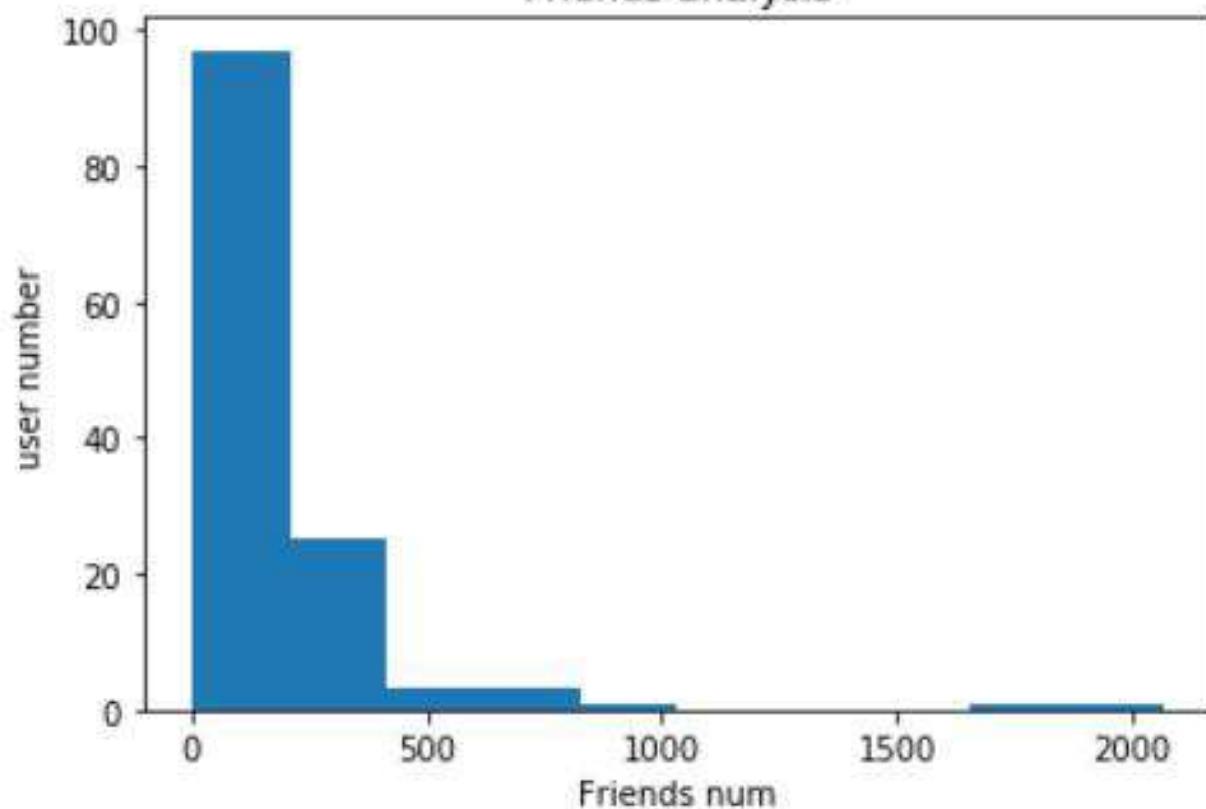
74

48

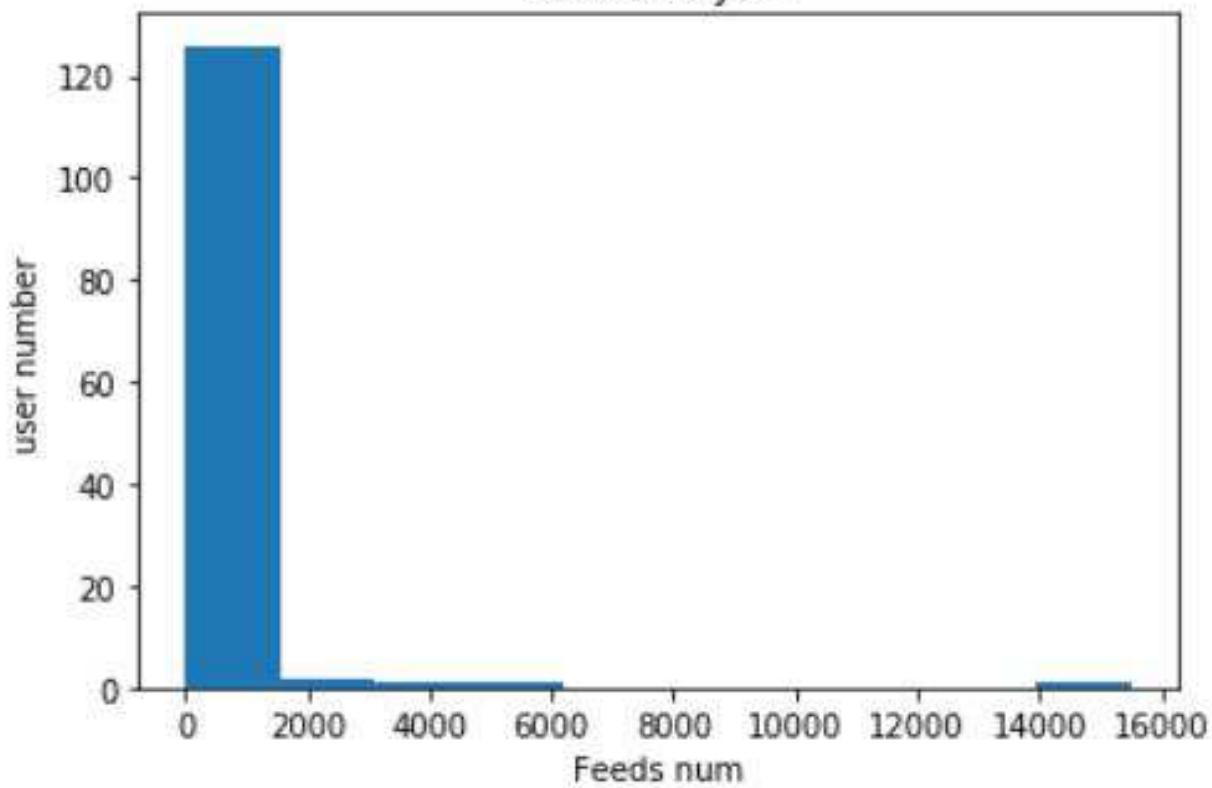
349



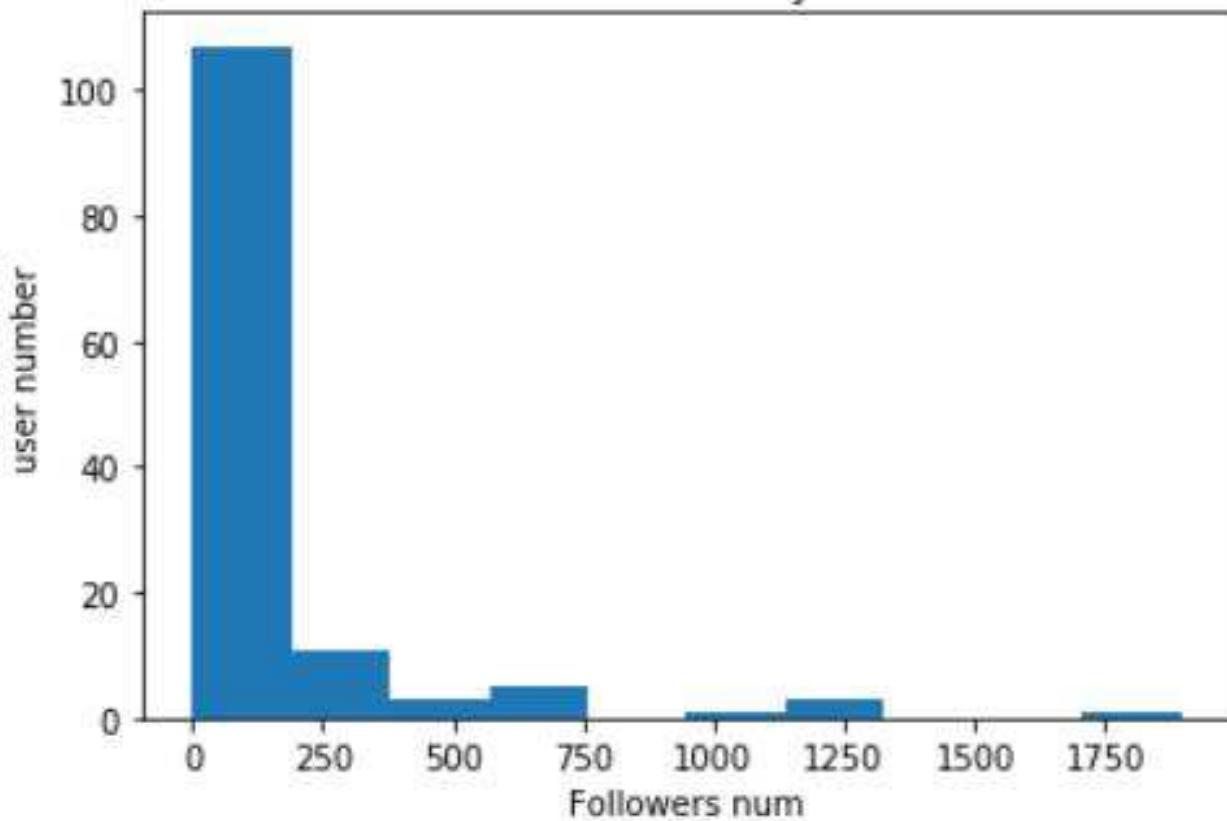
Friends analysis



Feed analysis



### Followers analysis



```
bibtext report
```

```
=====
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
Warning--I didn't find a database entry for "intro01"
Warning--I didn't find a database entry for "intro02"
Warning--I didn't find a database entry for "target01"
(There were 3 warnings)
```

```
bibtext _ label error
```

```
=====
bibtext space label error
```

```
=====
bibtext comma label error
```

```
=====
latex report
```

```
=====
[2017-12-11 13.24.52] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflatex)
Missing character: ""
Missing character: ""
Citation 'intro01' on page 1 undefined on input line
Missing character: ""
Citation 'intro02' on page 1 undefined on input line
Citation 'target01' on page 2 undefined on input lin
Missing character: ""
```

```
Missing character: ""
There were undefined citations.
bookmark level for unknown defaults to 0.
The anchor of a bookmark and its parent's must not be the same. Added a new anchor.
'h' float specifier changed to 'ht'.
Typesetting of "report.tex" completed in 1.0s.
./README.yml
8:81      error    line too long (84 > 80 characters) (line-length)
9:81      error    line too long (85 > 80 characters) (line-length)
10:81     error    line too long (85 > 80 characters) (line-length)
11:81     error    line too long (82 > 80 characters) (line-length)
22:1      error    trailing spaces (trailing-spaces)
25:15     error    too many spaces after colon (colons)
25:81     error    line too long (691 > 80 characters) (line-length)
41:81     error    line too long (96 > 80 characters) (line-length)
42:81     error    line too long (688 > 80 characters) (line-length)
```

---

## Compliance Report

---

```
name: Ni, Juan
hid: 107
paper1: Oct 22 1800 100%
paper2: 100%
project: Dec 08 0600 100%
```

```
yamlcheck
```

---

```
wordcount
```

---

```
11
wc 107 project 11 4636 report.tex
```

wc 107 project 11 4721 report.pdf  
wc 107 project 11 2025 report.bib

find "

---

36: The social opinion trend at social networking community will drive personal and even company decision, then some trends might harm someone's benefit because the power of social opinion trend is so powerful. Then people hired spammers to spread wrong information that lead the trend become advantage for them, but the users are become victim because they will make wrong decision because of the trend is control by someone on purpose. "Brown showed how it would be possible for spammers to craft targeted spam by leveraging the information available in online social networks." \cite{adv:01}, every spammer post must for some reason that beneficial for their employer, the most famous case for spammers the shampoo case at 2010. "BaWang shampoo" is the most famous shampoo at China which advertised by super star Jackie Chan, "Next Magazine" post a fake news claimed that using "BaWang shampoo" could cause cancer \cite{bawang:01}. I clear remember at that time, almost all social websites post new claim "BaWang shampoo" is harmful at the same time without any authority judgment, and they put this new at the headline position to abstract user's eye-ball. Even the authority department proof this new is unreliable, the business reputation of "BaWang shampoo" had been damaged, lot of people around me stop using this shampoo any more. This case seems have no spammers involved, but actually the spammers for this case is social networking websites themselves instead of single person. The reason why they post slander is because they can get benefit from other shampoo companies in china, other shampoo companies can have more sales because the market-share of "BaWang shampoo" will be decrease at this case.

38: The other reason why spammers getting so popular at social networking website is the operating cost of spammers is supper low. Try searching "buy Facebook like" at Google, and there are over hundred million results come up. And the price of buying like and followers from that website is pretty low, so spammers can get an account with 1000 followers which look like a real account for only 5 dollar \cite{buy:01}. So spammer can create thousands this kind of fake account for posting the wrong information at social networking website, and the detected system is hard to find out those kind of account is real account or Zombie account those accounts have actually followers and like, even feeds. Then spammers can use script to post batch of wrong information via

those account. Furthermore, the cost for post batch of wrong information is unbelievable low; according to the internal information which provided from my friend who working at a IT company, there two ways to post wrong information at social networking website, first one is money reward system, second one is posting AI. For the reward system, a professional spammer company usually have about 10 teams, each team has 500 people; They use reward instead of constant salary, each feed related to the order topic is worth 0.5 Chinese dollar(equal to 5 cent in dollar), and each comment on the target feed is worth 0.2 Chinese dollar( equal to 3 cent in dollar), and the price of long text post is negotiated. The spammers accounts are provide from the company, the price for those account is also low; Most social networking websites only require email address for registered, then they buy email accounts from the retail like 100 Chinese dollar(equal to 15 dollar) for ten thousand accounts, and using script to register social websites accounts and making follow with each fake accounts. This is how they operate the spammer, now most social networking websites register require phone number verify, then they working with local sim card retail which infinity phone number on hand, but the price for each fake account is increase a lot like 3 Chinese dollar(like 50cent) per one, but still pretty cheap compare with other advertise way. Cheap labor force and the development of script technology rising the spammer company, but the lower the user experience at social websites because lot of trash information full of the social websites, people are hard to see the true at social websites any more. According to the network sites worldwide ranking[7], we can see WeChat has almost twice active user than Weibo, and WeChat is the most popular social networking app in China because spammer can not do any at that app. In WeChat, they post feed at the module which call "Friends circle", the feed formatting is pretty similar as weibo, but the app is semi-closed which mean it is complete private, user can only see their friend's post and the comment, no retweet allowed, and if someone who not in the user friend's list comment at user's feeds, user can not see that guys comments. Plenty of users quit traditional social networking website, and the semi-closed social networking app is getting popular, so the adverse eff of spammer is not only for the spam target but also for the platform. If we let spammer keep development, and don't have any way to filter their information, the traditional social websites will die soon.

- 42: According to the worldwide statistics data, "Sina Weibo" has 368 millions active users which more than 328 millions of twitter active user \cite{target01} , so I would like to using "Sina Weibo" as my investigate target instead of using Twitter. The

other reason why I choose "Sina Weibo" as my investigate target is because I'm familiar with Chinese culture and I have been using "Sina Weibo" for more than 8 years. I think my knowledge about "Sina Weibo" will help me a lot at this project and better understand how spammers works at "weibo". The page frame at "weibo" is pretty similar to Twitter [figure1].

- 49: The four buttons under each feed are "collect, retweet, reply, like", and the capability for each button is same as twitter. When user click into the "rely" button, user can see all the comments related to the current feed, and sort them by the amount of "like" that comments get from other user. The only things differen at "weibo" is user not only can see the comments but also can see the retweet information, twitter only allow user to see who retweet the feed. Then user can see the retweet's comments and sort the list by the retweet's times of the retweet feed. So people would love to check the retweet list to see which famous person retweet the feed, and what comment they put into the retweet. Spammer control the public opinion trends by putting wrong information that doesn't represent real public opinion into the comment for some hot feed, they utilize user's habit to reach their goal.
- 59: from=page\\_1005055305999252\\_profile&wvr=6&mod=weibotime&type=comment} this feed to investigated the comment content. The person send this feed is my favorite gaming live streaming player, his name is LuBen Wei. He is the most popular gaming live streaming in China, there are over four million audiences what his playing game every night. Moreover, this is his second account, so he always post some feeds that can not be post at his official account at Weibo, but there are still thirty thousand comments under this feed , the number of comments at this feed even more than the comments under every signaler feed from Donald J. Trump's account. And the feed's content is he complain about the cheating case, he announced that he never cheating at "PlayerUnknown's Battlegrounds", he claim that the rumor about his cheating is come from the spammers. After he post this feed, this feed became the top 1 hot feed at the feed ranking at Webo, and most comments under this feed are abuse him cheating. So I though there are must be spammer working under this feed, the comments at a feed from a gaming live streaming player's second account is more than the comments from United states's president's feed which is so ridiculous. Therefor I think there must be spammer involved into this feed, that is how we pick up the feed which involved spammer in social networking website. If a feed has unusual comments and likes amount compare with other feed post from the owner, that feed have huge possibility that involved spammer work.

- 61: First of all, Weibo require we use Weibo API with authentication, so we need to create a personal application first at the weibo application apply page \cite{method:01}. Then the weibo official suggest us to use SDK to access the the API, so I came to the sdk websites \cite{method:03} to get the Weibo SDK package. Normally can just type "pip install sinaweibopy" to install this sdk package to python, and also can download the sdk package, and put the webo.py with the py files I using to collect data into the same fidder to use this sdk. I using the second method becuase I have issue pop this sdk. User can get the dirction of how to use sdk via the weibo sdk wiki page \cite{method:04}, they provide many tutorial about how to use sdk on different environment not only for pyhton. For using Offical sdk, we need to use the "app\\_key" and "app\\_secret", we can find those code from the application page which I create the app apply before using my account. Those two codes are represent the user identity of who using the API, so weibo will ban the user's weibo account if they do something bad via weibo API because those two codes are directly link to user's weibo account. For getting the autorized for using the weibo API, I using Thinkgamer\\_ggt's idea to get the authorized code \cite{method:05}, Weibo using OAuth 2 to check the user identity for using API. After the authorized page pump up, enter code which from the page url link which look like
- 62: \url{https://api.weibo.com/oauth2/default.html?code=2024222384d5dc88316d21675259d73a}, and the code we need to enter is the string that after "code=" at the url link.
- 63: ; then weibo will return an the access token for the API, then we using "Client" to activate the API, so we can get our target information via "Client".
- 71: The last step for collecting data is to storage the target data as txt file. All the data that related to comment are saved into the variable now, so if we want to get our target return object from this data set, we need to use the special code for different kind of category inside this data set; Weibo also provide a specific instruction for the code of return object \cite{method:07}. The data we need for wrong information filtration are the content of comments and the user information for each comments, the numbers of follower and the number of friends for the user who post the comments, also we need to get the number of feed that post from the user who comments the feed. Then use special code "follower\\_count", "FRIENDS\\_count", and "statuses\\_count" to get those information, and save them into the txt file for next step

data viualization. The reason why I save my target date into txt file because of the coding knowledge shortage, I don't know how to using the data analysis model at 2.7 python version, so I decide to using Python 2.7 to collecting data and using python 3.5 to do the data analysis.

74: The data visualization in this paper will be simple and straight forward, because of even I have some idea to analysis the data, but I can not represent it due to coding knowledge shortage. The models I decide to use for this data visualization are matplotlib, nltk, wordcloud, pandas, numpy, jieba and codecs. First, open the content.txt file at python and using readlines and appends to create a list that content all the comments. I intend to use wordcloud to visualize the words that has most frequency on the comment list, and I find out the wordcloud don't support chineses really well, so I use the module "jieba" to reproduce the comments list. Jieba is the best module to support Word Segmentation, wen can using this module to pick up the words which most frequently appear at the comment list\cite{method:08}. I using FontTian's formatting as my main structure of jieba code \cite{method:08}, also we can add some new word into our word list and use it to make the jieb module can be able to indentify the new world, and we use "stopwords\\_path" to filter the common word like "hello", and the we are using outside txt source which is Chinese vocabulary words out file as our stop word dictionary \cite{method:10}. And during using this stop word file inside the jieba code, we have to encoding the file to "utf8" formatting, otherwise it will have some error that the jieba module can not distinguish the content inside the stop word file; also we need to set up the right font for the wordcloud by using the ttc file from the fonts document on the computer, if the font use on wordcloud doesn't support chinese, the final result will be a retangle for each word instand of actually Chinese.

passed: False

find footnote

---

passed: True

find input{format/i523}

---

4: \input{format/i523}

```
passed: True
```

```
find input{format/final}
```

---

```
passed: False
```

```
floats
```

---

```
43: \begin{figure}[h]
44: \includegraphics[width=\columnwidth]{1.jpg}
77: \begin{figure}[h]
78: \includegraphics[width=\columnwidth]{2.jpg}
81: \begin{figure}[h]
82: \includegraphics[width=\columnwidth]{3.jpg}
88: \begin{figure}[h]
89: \includegraphics[width=\columnwidth]{4.jpg}
91: \begin{figure}[h]
92: \includegraphics[width=\columnwidth]{5.jpg}
94: \begin{figure}[h]
95: \includegraphics[width=\columnwidth]{6.jpg}
```

```
figures 6
```

```
tables 0
```

```
includegraphics 6
```

```
labels 0
```

```
refs 0
```

```
floats 6
```

```
True : ref check passed: (refs >= figures + tables)
```

```
True : label check passed: (refs >= figures + tables)
```

```
True : include graphics passed: (figures >= includegraphics)
```

```
True : check if all figures are refered to: (refs >= labels)
```

```
Label/ref check
```

```
passed: True
```

```
When using figures use columnwidth
```

```
[width=1.0\columnwidth]
```

```
do not change the number to a smaller fraction
```

```
find textwidth
```

---

passed: True

below\_check

---

WARNING: table and above may be used improperly

103: The above sections bring out the idea about how to detect the spammer, and the method is pretty sample because of it just to use for proof my idea is feasible. To rise the filter wrong information to the big data level, we can using database to storage our data instead of storage the data into a txt files. We can make a connection between API and mysql, and the programming will automatically storage the data inside each table for different data category. Also the authorized code can be automatically get from the authorized page, so user no need to enter it by hand. All the analysis will be integration into one application, so user only need to copy and paste the feed link that they want to see the true for that feed, and the application will decode the feed id and collecting all kind of data into the mysql database. For collecting huge size of data like over ten thousand comments, we can using different virtual machine to get data from the API, so we don't need to worry about the daily API usage any more.

WARNING: code and above may be used improperly

103: The above sections bring out the idea about how to detect the spammer, and the method is pretty sample because of it just to use for proof my idea is feasible. To rise the filter wrong information to the big data level, we can using database to storage our data instead of storage the data into a txt files. We can make a connection between API and mysql, and the programming will automatically storage the data inside each table for different data category. Also the authorized code can be automatically get from the authorized page, so user no need to enter it by hand. All the analysis will be integration into one application, so user only need to copy and paste the feed link that they want to see the true for that feed, and the application will decode the feed id and collecting all kind of data into the mysql database. For collecting huge size of data like over ten thousand comments, we can using different virtual machine to get data from the API, so we don't need to worry about the daily API usage any more.

bibtex

---

label errors

bibtex errors

```
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
Warning--I didn't find a database entry for "intro01"
Warning--I didn't find a database entry for "intro02"
Warning--I didn't find a database entry for "target01"
(There were 3 warnings)
```

bibtex\_empty\_fields

---

entries in general should not be empty in bibtex

find ""

---

```
382: % editor =      "",  
383: % number =     "",  
385: % type =       "",  
386: % chapter =    "",  
387: % pages =      "",  
388: % address =    "",  
389: % month =      "",  
  
passed: False
```

ascii

---

```
non ascii found 65292
non ascii found 65306
non ascii found 8220
non ascii found 8221
non ascii found 65306
non ascii found 65306
non ascii found 22238
non ascii found 22797
non ascii found 65292
non ascii found 22238
non ascii found 22797
```

```
=====
The following tests are optional
=====
```

Tip: newlines can often be replaced just by an empty line

```
find newline
-----
```

```
passed: True
cites should have a space before \cite{} but not before the {
```

```
find cite {
-----
```

```
passed: True
```

# CMD5 Plugin to Create a Docker Swarm Cluster on 3 Raspberry PIs

Andres Castro Benavides  
Indiana University  
107 S. Indiana Avenue  
Bloomington, Indiana 43017-6221  
acastrob@iu.edu

Uma M Kugan  
Indiana University  
107 S. Indiana Avenue  
Bloomington, Indiana 43017-6221  
umakugan@iu.edu

## ABSTRACT

Information technologies are evolving from mainly one-host environments to more distributed environment. Docker Swarm makes it possible to avoid having a single point of failure and instead, have multiple nodes that can be properly balanced and contain replicas of the information. Currently, Dockers must be individually downloaded, installed and configured on each physical computer in order for the desired computers to work in swarm mode. This paper details the development of a plug-in that would allow CloudMesh to deploy a Docker Swarm cluster. The creation of this plug-in would be the first step towards the development of a tool which would allow larger debian based networks to work as container oriented virtual environments with optimized usage of resources.

## KEYWORDS

Raspberry Pi, Cloudmesh, CMD5, Big Data, Big Data, i523, HID305, HID323

## 1 INTRODUCTION

### 1.1 Docker: Swarm mode, Current Use, Installation and Configuration

Docker is the technology used for containerization for software development. It is an open source tool which makes it easy to deploy applications. Applications are packaged in containers and then it is shipped to all the platforms that is supposed to work with. Applications are divided into manageable sizes and all the dependent functions are added and individually packaged. Both Linux and Windows are supported by Docker.

Docker Swarm is a clustering and scheduling tool for Docker containers. A swarm is nothing but multiple Docker hosts which run in swarm mode and act as managers to manage delegation and workers will run swarm services). A given Docker host can be a manager or a worker or it can perform both roles. If any of the worker node becomes unavailable, manager schedules that node's tasks on other nodes. A node is an instance of the docker engine participating in the swarm [4].

A swarm is made up of multiple nodes. We need to execute docker swarm init to enable swarm mode and to make current machine a swarm manager, run docker swarm join on other machines to add them to the swarm as workers and run docker node ls on the manager to view the nodes in this swarm.

Docker Swarms are used to orchestrate processes, optimizing the use of resources across clusters. In other words, the use of Docker Swarms allow individual computers to work as a cluster, sharing their RAM, processors, physical memory, among other features

or abilities. The docker, when used in swarm mode, evaluates the assets across the network and manages tasks in real time. Each computer can contribute its assets to complete tasks in the most efficient way. It is dynamic and adapts based on the available resources and current demands.

In order to set up a Docker Swarm, there needs to be direct access to each machine that will be used as a node (an instance of Docker that will be part of the swarm). In order to set up the nodes, the docker must be independently installed and configured on each machine. Then, each machine must be added to the swarm, allowing it to communicate or interact with the other nodes.

This process not only requires human resources (technicians working on installation and configuration) but also demands these actions be repeated manually on each individual node or manager. While this can be done virtually, it still requires individual attention in the setup of each machine. In order to optimize the setup of Docker Swarms, CloudMesh could be utilized to centralize installation and configuration of every node and manager.

*Inside Docker.* The four main internal components of docker are Docker Client and Server, Docker Images, Docker Registries, and Docker Containers.

*Docker Client and Server.* The docker server gets the request from the docker client and then processes it accordingly. Docker server and docker client can either run on the same machine or a local docker client can be connected with a remote server running on another machine [8]. Fig. 1 Docker architecture [8].

*Docker Images.* Base images are the operating system images such as Ubuntu 14.04 LTS, or Fedora 20 which creates a container to run an operating system. The docker file contains a list of instructions to build an image. When using docker, we start with a base image, boot up, create changes and those changes are saved in layers forming another image [7].

*Docker Registries.* Docker images are placed in docker registries. It is similar to source code repositories where images can be pushed or pulled from a single source.

*Docker Containers.* Docker image creates a docker container. Containers have everything for the application to run on its own.

#### 1.1.1 Benefits of using Docker.

*Open Source Technology.* The Docker containers are based on open standards which means that anyone can contribute to the Docker tool and at the same time customize it for their needs, if the features they are looking for is not already available.

*Portability.* Docker makes distributed applications to be dynamic and portable which can be run anywhere which makes it extremely popular among developers.

*Sharing.* Docker is integrated with a software sharing and distribution mechanism that allows for sharing and using container content which helps the tasks of both the developer and the operations team.

*Elimination of Environmental Inconsistencies.* Any changes made in one environment will be shared across other environments or all the applications can exist in the single environment.

*Resource Isolation.* Resource isolation adds to the security of running containers on a given host. Docker uses Namespaces technology to isolate work spaces called containers. Namespace is created when container is run and access is limited to that namespace only. Every container in Docker will have its own work space which makes it easier debug if there are issues with any particular container.

*Easy Integration.* Docker can be easily integrated into a variety of infrastructure tools like Amazon Web Services, Ansible, IBM Bluemix, Jenkins, Google Cloud Platform, Oracle Container Cloud Service, Microsoft Azure to name a few.

*Better Security.* Docker provides a interface for developers and IT teams to define and manage their security configurations for applications as it navigates from one stage to another.

*Docker - Use Cases.* The Docker platform is the only container platform to build, secure and manage the variety of applications from development to production both on premises and in the cloud. It also creates room for innovation, increases time to market, highly agile. Docker supports diverse set of applications and infrastructure for both developers and IT. It transforms IT without having to re-tool, re-code or re-vamp existing applications, policies or staff [5].

*DevOps.* The main goal of DevOps is to eliminate the gap between the developers and IT operations team. Docker with DevOps get the developers and operations team to work together so that they both understand the challenges faced by each other, apply DevOps practices [5].

*CI/CD.* Continuous Integration and Continuous Deployment (CI/CD) are the most common use cases of Docker. Continuous Integration testing and Continuous Deployment allows developers to build codes, test them in any environment. Docker integration with Jenkins and GitHub making it easier for developers to build codes, test them in GitHub and trigger a build in Jenkins and adding the image in Docker registries [5].

*Docker Containers As A Service.* Docker help any organization to modernize their application architecture. It can deploy scalable services securely on a wide variety of platforms, improving flexibility and maximizing capacity. Best use case for Docker installation is the US Government where they enhanced their applications and made their components and services of their system and easily transportable/shareable with other agencies within the government [5].

#### 1.1.2 Docker - Services.

*Docker Engine.* Docker Engine is the foundation for the application platform which is used for creating and running Docker containers. It is supported on Linux, Windows, Cloud and Mac OS. It is lightweight, open source and integrated with a work flow to build and containerize applications. User interface is very simple and it makes the environment easily portable from single container on single host to multiple applications on a many number of hosts [5].

*Docker Enterprise.* Docker Enterprise provides an integrated platform for both developers and IT operations team where container management and deployment services are together for end-to-end agile application portability. It is easy to manage, monitor and secure images both within the registry and those deployed across various clusters [5].

*Docker Hub.* Docker Hub functions as a hosted registry service that helps you store, manage, share and integrate images across various developer work flows. Integration testing is done each time when the image is shared [5].

*Docker Compose.* Docker Compose is a tool that developers deploy to define and run all multi-container Docker applications. Single host can be used to isolate multiple environments, even if they are of the same name. Data volume is copied automatically from old container whenever a new container is created. Compose uses the previous configuration to create the new container which reduces the time for replicating the same changes to the environment [5].

## 1.2 CloudMesh

CloudMesh is an innovative tool that allows communication and interaction between cloud based solutions. Not all clouds are docker based and there are different types of virtual and cloud environments. Through CloudMesh, data can be shared and utilized by cloud solutions that are not otherwise programed to communicate with each other. Cloud mesh does not just manage a series of clouds, but centralizes and deploys them as one main system that manages the data resources.

\*\*Quote teacher= Cloudmesh is a project to easily manage virtual machines and bare metal provisioned operating systems in a multicloud environment. We are also providing the ability to deploy platforms\*\*

## 1.3 Creating CloudMesh plug-ins

*what it currently does and has the potential to do.* By creating CloudMesh plug-ins, it is possible to extend its potential from different kinds of cloud based environments interconnection to deployment of a container management system, in this case, Docker .

Utilizing CloudMesh to Centralize Docker Swarm Installation Cloud Mesh does not have a plug in that allows you to deploy container solutions on physical networks. Create a plug in that would allow Cloud Mesh to deploy container solutions (in this case, the Swarm mode of Docker) to a physical Debian based network (in this case, a series of raspberry pies). Could be used as a model to deploy other types of container oriented solutions. It is taking a simple network (Debian based network and allowing it to centralize

resources and assigning tasks and optimizing different functions by installing a container management system, called Docker Swarm.

In order to simulate the deployment of a Docker Swarm cluster, this Cloudmesh project develops a Cloudmesh plug in, that deploys a Docker Swarm cluster on three Raspberry Pi, allowing them to be part of this multi cloud environment.

The cloud mesh allows Methods you to deploy the Docker Swarms (are container management tools) to the raspberry pies.

## 2 METHODS: PROPOSED SOLUTION

About the current solution:

This solution was created for a specific type of hardware and software, but is modular enough to be extended to different environments with similar features, such as basic architecture -which include but is not limited to ARM single boarded computers- and an operating system based on Debian, such as Debian, Raspbian, Ubuntu, etc.

### 2.1 Hardware

For the current proposed solution, the different pieces of hardware were chosen based on criteria such as Compatibility and Price.

The following is a list of the hardware that was used and below that list there is a description of each piece of hardware that was used.

- 3 Raspberry Pi
- 3 Micro SD Cards (64 GB)
- 3 USB to Micro USB Cables for power supply to the Raspberry Pi
- 1 External monitor (for the configuration only).

*2.1.1 Raspberry Pi.* For this experiment, the 3 machines that were used were Raspberry Pi 3 Model B. Raspberry Pi are single boarded computers, that come in a small presentation. They have been developed with education and extension in mind, making them very popular in the academic and entrepreneur communities. The specifications of the model that has been used for this experiment are the following:

- CPU: 1.2 GHZ quad-core ARM Cortex A53 (ARMv8 Instruction Set)
- GPU: Broadcom VideoCore IV @ 400 MHz
- Memory: 1 GB LPDDR2-900 SDRAM
- USB ports: 4
- Network: 10/100 MBPS Ethernet, 802.11n Wireless LAN, Bluetooth 4.0

[2]

The Raspberry Pi are interacting with each other using a private wireless network, and they have been assigned static Internet Protocol Addresses. In this case 192.168.1.85, 192.168.1.86 and 192.168.1.87.

*2.1.2 Micro SD Cards.* Because of its architecture, Raspberry devices require the use of Micro SD Cards to contain the Operative system and other files. They emulate the Hard drive resource used on other kinds of computers. The reason that it is required to have at least 16 GB of memory, is because there will be several pieces of software installed in the devices, each one of them with different requirements:

Docker Memory Requirements [4]:

- 8GB of RAM for manager nodes or nodes running DTR.
- 4GB of RAM for worker nodes.
- 3GB of free disk space.

So at least 12 of the GB would be required for Docker and 4 GB used for the proper functioning of Raspbian. [6]

Taking these requirements in consideration, there should be a minimum of 16GB of free space in the MicroSD in order to perform this experiment.

The Micro SD cards used were San Disc Memory Cards with a 64GB capacity.

*2.1.3 Micro USB Cables.* 3 USB to Micro USB Cables for power supply to the Raspberry Pi Since these small computers don't use the regular power supply chords, they are equipped with MicroUSB ports to power the device. All of these devices are plugged to a main power outlet that allows to charge multiple devices at the same time. There are other options to power the devices include, such as attaching them to external batteries.

*2.1.4 External monitor.* Since the Raspberry Pi are headless machines, they require to be accessed directly for the initial set up and after that it is possible to continue the configuration and installation process using any kind of remote access, like SSH or RealVNC. For this initial connection, any kind of screen that is HDMI compatible is useful. In this case the initial setup of the Raspberry Pi was performed on a Toshiba 55 inch HDTV with HDMI port. After that they were accessed from a Laptop computer with Linux Ubuntu 17.10, using Remmina via ssh (XORG).

*2.1.5 Initial input devices.* In order to set up the devices. The Raspberry Pi will require a set of initial input devices attached to each computer. For this exercise, a USB enabled standard keyboard and a USB enabled standard mouse were used.

### 2.2 Software

*2.2.1 Raspbian.* Currently, the default way to deploy the operating system to the Raspberry Pi is by using an Operating System installation Manager called Noobs -which stands for fiNew Out Of Box Softwarefi-. This manager can be downloaded directly from the Raspberry Pi website and it includes several Operating system options, among them:

- Raspbian
- Pidora
- LibreELEC
- OSMC
- RISC OS
- Arch Linux

Since Raspbian is the default Operating system and most commonly used, this experiment decided to use it. This is also helpful because there is material available in different websites with instructions on how to install Docker in Debian based Machines. Raspbian is Debian based. Another important reason is that Docker has as a requirement that the Linux kernel version on which it will be installed is 3.10 or higher. The Kernel version of the version of Raspbian that was used is 4.9.

The version of Raspbian that was used has the following specifications:

- **Kernel version:** 4.9
- **Release date:** 2017-11-16

2.2.2 *Docker.* There are several versions of Docker available. Each version with their own advantages and disadvantages. Because of the architecture used by Raspberry Pi -ARM instead of AMD-, the Docker version used is **Docker for Debian ARM**. With the following Specifications:

- Version 17.09.0-ce
- Release 2017-09-26

This version of Docker is Community Edition (CE), which means that it is available for free and can be installed on bare metal or cloud infrastructure. This flexibility is good for the experiment, because it will be installed on Raspberry Pi, which are considered physical devices or bare metal Machines. [4]

### 3 PREREQUISITES

There are several reasons to have the pre requisites that the user will find in this document. They will be explained in a separate section. Before using the proposed solution, the userfis environment needs to meet the following requirements:

3.0.1 *Raspbian Installed.* Raspbian must be installed and configured on all Micro SD Cards. For this, the user may download Noobs from <https://www.raspberrypi.org/> and copy it to a formatted Micro SD Card. Once the Raspberry Pi has the MicroSD loaded with noobs in place and has the input devices and display attached to it, the user may follow the OS installation guide found on: <http://raspbian.org/>

It is advisable to be hooked up to the network where the user is planning on implement this solution before running Noobs for the first time. This will allow the user to download newer packages or Raspbian and avoid interruptions in the process.

This requirement exists because there is a function that is being explored to capture Raspberry Pi images to be deployed later on and avoid the present pre requisite, but it is not ready yet.

3.0.2 *Update OS repositories.* In order to ensure that the user is accessing the latest version available of the software, it is important to update the Raspbian repositories. In this case, the user can access the Terminal and enter the following commands:

"**sudo apt-get update**" to update the list of available repositories and then "**sudo apt-get upgrade**" to upgrade the available packages.

The first time that the user runs one of these commands, the root password will have to be entered. This process might take a few minutes. [3]

3.0.3 *Remote access setup.* Enable SSH on the Raspberry Pi. After Raspbian installation, enable SSH on all your Raspberry Pi machines.

To do this, the user has to add a line in the file "**sshd\_config**" found in the directory "**/etc/ssh/**" The line has to go at the end of in the "**Authentication section**". It has to contain the following string "**PermitRootLogin yes**". [1]

3.0.4 *Changing hostnames.* In order to keep the three Raspberry Pi organized it is highly advisable to assign an exclusive and distinctive hostname to each Raspberry Pi. The three Raspberry Pi have the following IP addresses:

- (1) pi85 - 192.168.1.85
- (2) pi86 - 192.168.1.86
- (3) pi87 - 192.168.1.87

By default, all Raspberry Pi devices will have the same Host Name.

To change this feature on each machine, the user will have to modify the line that contains "**127.0.1.1**" and as hostname it includes the string "**raspberrypi**" in "**/etc/hosts**" file, in most of the cases it is the last line in the file. Then, the user may type the desired hostname instead of the word raspberrypi and save the file and close it. This part can be done by using the text editor that comes by default with Raspbian, an editor called "**nano**". It is not advisable for the users to modify the rest of the entries, at least as part of this project.

Once the file is modified, the user will have to initialize the hostname with the **hostname.shfi** script this can be done using the following line in the Terminal: "**sudo /etc/init.d/hostname.sh**"

To check if the modification has worked as expected, the user may check the hostname of the machine from the Terminal by running the command: "**hostname -I**"

## 4 STEPS FOLLOWED

### 4.1 Testing shell commands prior to integrations with Cloudmesh

Since Raspberry pi is not currently listed under the supported operative systems for Docker or Cloudmesh, The process of deploying Docker and configuring the swarm Mode was successfully tested on the Raspberry Pi first using the commands that are intended for Debian. Once the Swarm was configured, the three Raspberry Pi devices were left on for over 24 hours and it was not observed any kind of abnormal behavior, like looping services in the OS or overheating.

### 4.2 Purchasing the hardware

The different hardware components were purchased via Amazon.com and took anywhere between 2 to 5 days to arrive. The different components can also be purchased through multiple on line sources or local electronics stores.

### 4.3 Installing the components via ssh into every node.

The following steps were followed on each device: Usig the TV as an external monitor, the USB input devices (keyboard and mouse), and the Raspberry Pi with Raspbian installed. An ssh key was generated and the device was accessed using Remmina via a XORG connection from a computer equipped with Linux Ubuntu 17.10 (Artful Aardvark). The components were installed in the following order:

Updated the **Raspbian** packages Installed **Python 3.6.2** and **Python 2.7.13** via PIP and also Installed **Cloudmesh**: following the instructions found in: <https://github.com/cloudmesh>/ Installed **Docker CE ARM** via Terminal using the command `curl -sSL https://get.docker.com -sh` as suggested in <https://www.raspberrypi.org/>

#### 4.4 Installing and configuring Docker Swarm

**4.4.1 Manager.** Since Docker requires at least one computer to be a Manager and Cloudmesh also requires at least one main configured piece of equipment, a Raspberry Pi was chosen to be the main device (in this case, the Raspberry Pi with the IP address 192.168.1.85). The following command was run on the Terminal or that device to set it as the manager: `sudo docker swarm init --advertise-addr 192.168.1.85`

#### 4.5 Workers

The other two Raspberry Pi devices (in this case, the Raspberry Pi with the IP address 192.168.1.86 and the one with 192.168.1.86) were defined as simple worker nodes. To define the workers, the following command was used: `sudo usermod -a -G docker USER` and to work as part of the swarm the command used was: `docker swarm join --token *** 198.168.1.85:2377` As a last step, it was confirmed that all the nodes were added by using the following command: `sudo docker node ls`

#### 4.6 Additional Research

**4.6.1 Other functions considered.** Initially, for this case, it was considered an option to developed a function called CaptureImage and a second function called DeployRaspbian. As their names suggest, the first one intended to capture an image or backup of a Raspberry Pi. This first function would receive the IP address or hostname of the desired machine and the desired location to store the captured image, alongside the corresponding credentials and wrap a `dd` shell command similar to the following:

`dd if=/dev/mmcblk0 bs=1M -gzip - - dd of=imageDir`

Among the challenges faced, this line was returning an invalid syntax, most likely because of the use of the variables. Since there was not a lot of time, the team decided to postpone this function.

The second function was called DeployRaspbian and would receive the route and name where the image would be deployed (i.e. `/dev/bkp`) and image name and route (i.e. `/Desktop/raspbian.gz`). The shell command that would be wrapped would be:

`gzip -dc diskNm - sudo dd of=imageName bs=1m conv=noerror,sync`

More information on this topic can be found in the section called **Backup** [www.raspberrypi.org](https://www.raspberrypi.org/).

Among the challenges, there is no clarity on whether the image can be deployed over a lan connection and this point there is not enough time to run tests. Also, since this is a copy of a previously used Raspbian, there is a chance that there might be conflicts related to the IP addresses that might be stored in different files of the OS.

**4.6.2 Final code.**

## 5 CONCLUSIONS

1. It is possible to create the plug in. 2. Since this was 3. The fact that the passwords would have to be either hard coded or transferred in plain text has to be seen as a vulnerability, that has to be addressed either by adding an encryption/decryption module or finding another way to safely access the root of the target device.

### 5.1 Evaluation

### 6 OTHER OPTIONS CONSIDERED

Other options of coding were considered during the development of this solution. Since all of the deployment can successfully be done via terminal in Raspbian, two main options were considered:

Option 1. A bash script for every part of the deployment and wrap it in python. This option would have been less dynamic and wouldn't make the best use of the available resources, but at the same time it could have been easier to adapt to linux Operating systems other than Raspbian. Option 2. Use the SH subprocess included in python 2.5-3.5. This is the option that was chosen by the team.

### ACKNOWLEDGMENTS

The authors would like to thank Dr. Gregor von Laszewski for his support and suggestions on this project.

### REFERENCES

- [1] Laura Bailey, Laura Novich, Tim Hildred, and David Jorm. 2012. Red Hat Customer Portal. (2012). [https://access.redhat.com/documentation/en-us/red\\_hat\\_enterprise\\_linux/6/html/v2v\\_guide/preparation\\_before\\_the\\_p2v\\_migration-enable\\_root\\_login\\_over\\_ssh](https://access.redhat.com/documentation/en-us/red_hat_enterprise_linux/6/html/v2v_guide/preparation_before_the_p2v_migration-enable_root_login_over_ssh)
- [2] Brian Benchoff. 2016. Introducing the Raspberry Pi 3. (Feb 2016). <https://hackaday.com/2016/02/28/introducing-the-raspberry-pi-3/>
- [3] Debian. 2017. DebianPackageManagement. (2017). <https://wiki.debian.org/DebianPackageManagement>
- [4] Docker. 2017. Docker CE release notes. (Dec 2017). <https://docs.docker.com/release-notes/docker-ce/>
- [5] Hackernoon. 2017. Docker-the Popular Containerization Technology for an Effective Software Development. (2017). <https://hackernoon.com/docker-the-popular-containerization-technology-for-an-effective-software-development-4e2cddc5>
- [6] Raspberry Pi and Raspberry Pi-Teach. [n. d.]. SD cards. ([n. d.]). [https://www.raspberrypi.org/documentation/installation/installation\\_sd-cards.md](https://www.raspberrypi.org/documentation/installation/installation_sd-cards.md)
- [7] Babak Bashari Rad, Harrison John Bhatti, and Mohammad Ahmadi. 2017. An Introduction to Docker and Analysis of its Performance. *International Journal of Computer Science and Network Security (IJCSNS)* 17, 3 (March 2017), 228.
- [8] James Turnbull. 2014. *The Docker Book: Containerization is the new virtualization*. James Turnbull, New York, USA.

## A WORK BREAKDOWN

Uma Kugan.

Andres Castro Benavides.

**Editing:** Andres Castro Benavides and Uma Kugan.

```
bibtext report
```

---

```
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
I was expecting a ',' or a '}'---line 20 of file report.bib
:
: url={http://paper.ijcsns.org/07_book/201703/20170327.pdf},
(Error may have been on previous line)
I'm skipping whatever remains of this entry
Warning--empty year in rpicards2017
(There was 1 error message)
make[2]: *** [bibtex] Error 2
```

```
latex report
```

---

```
[2017-12-11 13.29.50] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflatex)
Missing character: ""
Typesetting of "report.tex" completed in 1.1s.
./README.yml
20:81      error      line too long (85 > 80 characters)  (line-length)
24:1      error      trailing spaces  (trailing-spaces)
```

---

```
Compliance Report
```

---

```
name: Uma M Kugan
hid: 323
paper1: Review Date 11.10.2017
paper2: Review Date 11.06.2017
project: Dec 04 17 0%
```

```
yamlcheck
```

---

```
wordcount
```

---

```
6
wc 323 project 6 4190 report.tex
wc 323 project 6 4120 report.pdf
wc 323 project 6 204 report.bib
```

```
find "
```

---

```
323: \textbf{\textit{dd if=/dev/mmcblk0 bs=1M | gzip -" | dd
      of=imageDir}}\\
```

```
passed: False
```

```
find footnote
```

---

```
passed: True
```

```
find input{format/i523}
```

---

```
4: \input{format/i523}
```

```
passed: True
```

```
find input{format/final}
```

---

```
passed: False
```

```
floats
```

---

```
figures 0
tables 0
includegraphics 0
labels 0
refs 0
floats 0
```

```
True : ref check passed: (refs >= figures + tables)
True : label check passed: (refs >= figures + tables)
True : include graphics passed: (figures >= includegraphics)
True : check if all figures are refered to: (refs >= labels)
```

```
Label/ref check
passed: True
```

```
When using figures use columnwidth
[width=1.0\columnwidth]
do not change the number to a smaller fraction
```

```
find textwidth
```

---

```
passed: True
```

```
below_check
```

---

```
bibtex
```

---

```
label errors
```

```
bibtex errors
```

```
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
I was expecting a ',' or a '}'
:
line 20 of file report.bib
```

```
: url={http://paper.ijcsns.org/07_book/201703/20170327.pdf},  
(Error may have been on previous line)  
I'm skipping whatever remains of this entry  
Warning--empty year in rpicards2017  
(There was 1 error message)
```

```
bibtex_empty_fields
```

---

```
entries in general should not be empty in bibtex
```

```
find ""
```

---

```
passed: True
```

```
ascii
```

---

```
non ascii found 8217  
non ascii found 8217  
non ascii found 8220  
non ascii found 8221  
non ascii found 8217  
non ascii found 8220  
non ascii found 8221
```

```
=====
```

```
The following tests are optional
```

---

```
Tip: newlines can often be replaced just by an empty line
```

```
find newline
```

---

```
passed: True  
cites should have a space before \cite{} but not before the {
```

```
find cite {
```

---

```
passed: True
```