Assignment – 2
Group -J

Group Members	Student ID
Megha Navin	C0751827
Akhila Saladi	C0752116
Gaurav Manchanda	C0752979
Manpreet Kaur	C0752299
Sumanth Mohan	C0752365

Table of Contents

Objectives

- **Gain experience of handling data from various sources like Social media, databases (example, UCI database) or real-time data**
- **Exploring data collecting and pre-processing methods**
- **Gain experience of data storage**
- **Gain experience of using various data mining algorithms**
- **Gain experience of data visualization and presentation.**

Introduction

Data mining is the process by which we can discover design patterns in large data sets involving methods at the intersection of machine learning, statistics and database systems. It is an analysis step of knowledge discovery in databases.

Datasets:

A phishing website (also called a spoofed site) that seems like or makes us believe it to be a legitimate website that tries to steal our account details and confidential information. If we receive an email asking about our personal information such as password and social security number. This problem is considered a vital issue in industry especially e-banking and e-commerce taking the number of online transactions involving payments. We have identified different features related to legitimate and phishy websites and collected 1353 different websites from different sources. Phishing websites were collected from Phishtank data archive (www.phishtank.com), which is a free community site where users can submit, verify, track and share phishing data. The legitimate websites were collected from Yahoo and starting point directories using a web script developed in PHP. The PHP script was plugged with a browser, and we collected 548 legitimate websites out of 1353 websites. There are 702 phishing URLs and 103 suspicious URLs. Some of the features related to our data are (Fruhlinger, 2020):
Address based features:

- Using the IP address instead of the URL.
- Using a long URL to hide the suspicious part.
- URL's have a "@" method.
- Redirecting using "//"
- Adding prefix and suffix separated by (-) to the domain
- Sub Domains and Multi Sub Domains
- HTTPS
- Domain registration length

- Using Non-Standard Port

HTML and JavaScript based features:

- Website forwarding
- Status bar customization
- Disabling right click
- Using pop-up window

Domain based features:

- Age of Domain
- DNS record
- Website traffic

Attributes:

collected features hold the categorical values, these values have been replaced with numerical values 1,0 and -1 respectively. Details of each feature are mentioned in the research paper mentioned below:

- URL Anchor
- Request URL
- SFH
- URL Length
- Prefix/Suffix
- IP
- Sub Domain
- Web traffic

```
train_X,test_X,train_Y,test_Y=train_test_split(X,Y,test_size=0.2,random_state=2)
```

```
print(train_X.shape)
print(test_X.shape)
print(train_Y.shape)
print(test_Y.shape)
```

```
(8844, 31)
(2211, 31)
(8844, 1)
(2211, 1)
```

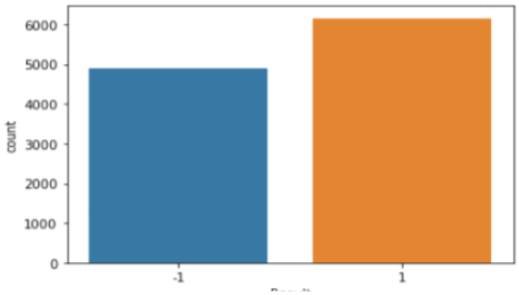
```
#display the results
diff_models
```

	Model	Top 10 features
0	Support Vector Machine	94.00
1	Logistic Regression	92.34
2	Random Forest	95.06
3	Perceptron	91.62
4	CART	94.81

```
In [8]: print(a,"times 0 repeated in Result")
        print(b,"times -1 repeated in Result")
        print(c,"times 1 repeated in Result")
        sns.countplot(data['Result'])

        0 times 0 repeated in Result
        4898 times -1 repeated in Result
        6157 times 1 repeated in Result

Out[8]: <matplotlib.axes._subplots.AxesSubplot at 0x1d154b98>
```



Algorithms

K Nearest Neighbour

KNN algorithm is one of the simplest classification algorithms and it is one of the most used algorithms. It can be used both for classification as well as regression, but it's mainly used for classification predictive problems in industry (SRIVASTAVA, 2018). It is:

- Lazy Learning Algorithm: because it does not have a specialized training phase and uses all the data for training while classification.
- Non-parametric Learning Algorithm: because it does not assume anything about the underlying data

```
print(classification_report(test_Y,knn_predict))
```

	precision	recall	f1-score	support
-1	0.60	0.60	0.60	987
1	0.68	0.68	0.68	1224
accuracy			0.64	2211
macro avg	0.64	0.64	0.64	2211
weighted avg	0.64	0.64	0.64	2211

Decision Tree

A Decision tree is a flow-chart like structure in which each internal node represents a test on an attribute (0 or 1), each branch represents the outcome of the test and each leaf node represents a class label (decision taken after computing all attributes). The path from root to leaf represents classification rules.
Tree based algorithms are considered as one of the best and mostly used supervised learning models. Tree based models empower predictive models with high accuracy, stability and ease of interpretation (Chauhan, 2019).

```
print(classification_report(test_Y,tree_predict))
```

	precision	recall	f1-score	support
-1	0.95	0.96	0.95	987
1	0.96	0.96	0.96	1224
accuracy			0.96	2211
macro avg	0.96	0.96	0.96	2211
weighted avg	0.96	0.96	0.96	2211

Logistic Regression

Logistic Regression is an appropriate regression analysis to conduct when the dependent variable is binary. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables (Brownlee, 2016).

```
print(classification_report(logreg_predict,test_Y))
```

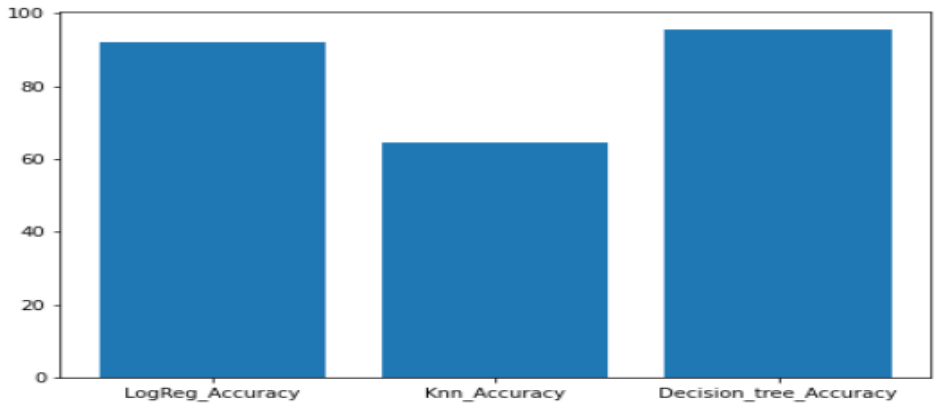
```
              precision    recall  f1-score   support

          -1       0.91      0.91      0.91       984
           1       0.93      0.93      0.93      1227

    accuracy                           0.92      2211
   macro avg       0.92      0.92      0.92      2211
weighted avg       0.92      0.92      0.92      2211
```

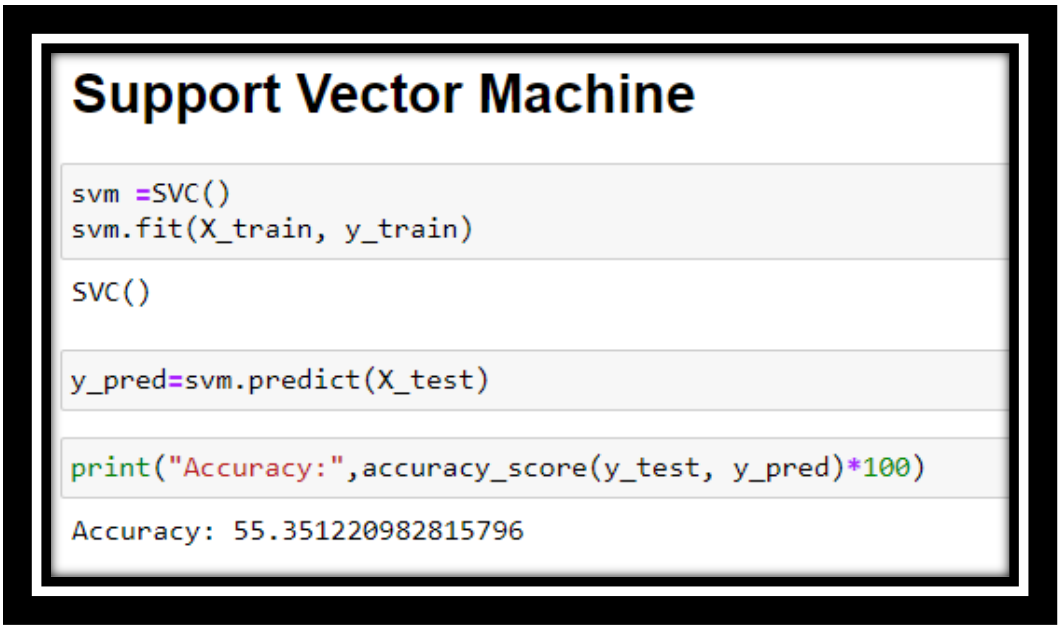
Output of three algorithm

```
fig = plt.figure()
ax = fig.add_axes([0,0,1,1])
algorithms = ['LogReg_Accuracy', 'Knn_Accuracy', 'Decision_tree_Accuracy']
result = [LogReg_Accuracy*100,Knn_Accuracy*100,Decision_tree_Accuracy*100]
ax.bar(algorithms,result)
plt.show()
```



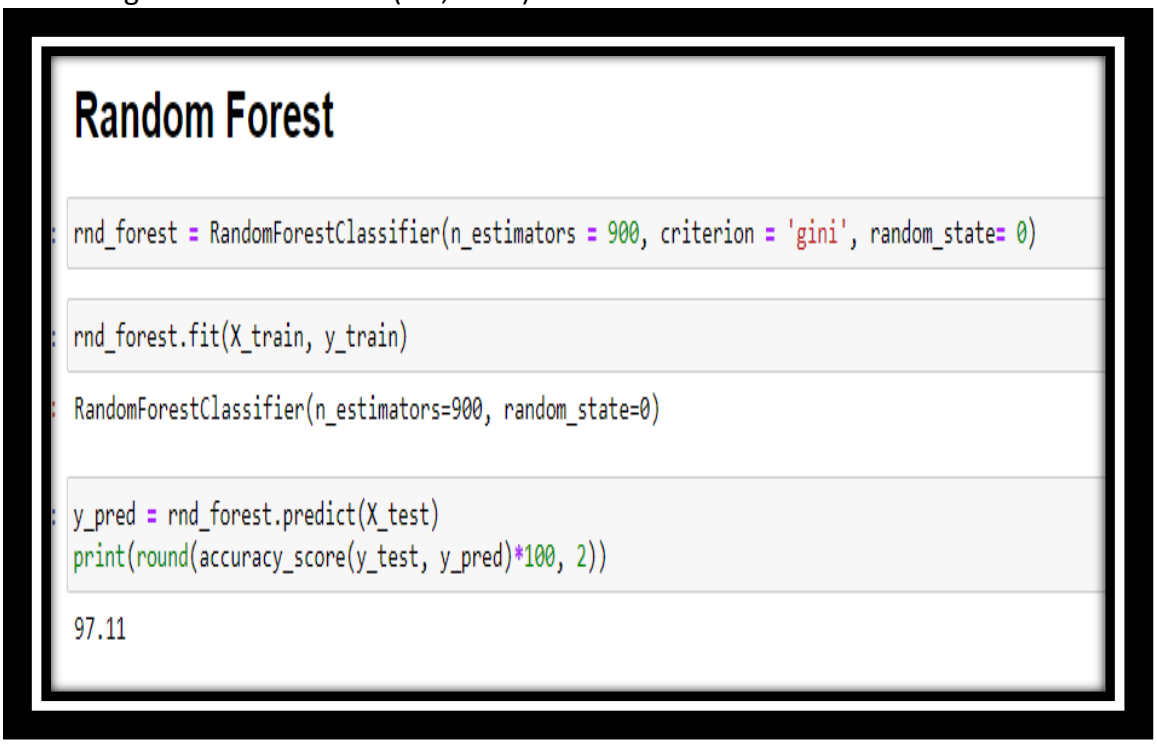
Support Vector Machine

A support vector machine (SVM) is a supervised machine learning model that uses classification algorithms for two-group classification problems. After giving an SVM model sets of labeled training data for each category, they're able to categorize new text. The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space(N- the number of features) that distinctly classifies the data points (Ray, 2017).



Random Forest

A supervised learning algorithm which can be used for both regression and classification. It works by creating decision trees on random data samples. It predicts the outcome by selecting the best solution through voting. Random forest classifier will handle the missing values and maintain the accuracy of a large proportion of data. If there are more trees, it won't allow overfitting trees in the model (Yiu, 2019).



Natural Language Processing

Natural Language Processing is the technology used to aid computers to understand the human's natural language. It is a new branch of Artificial Intelligence (AI) that allows machines to break down and understand human language. NLP techniques can be used to interpret text data that we are working with for analysis. Firstly, we need text pre-processing techniques, machine learning techniques and python libraries for NLP (Shetty, 2018).

Text pre-processing techniques include tokenization, text normalization and data cleaning. Once in a standard format, various machine learning techniques can be applied for a better understanding of the data. This includes using the most popular modelling techniques to classify spam detection. Newer, more sophisticated techniques can also be used, such as topic modelling, word embeddings or text generation with deep learning.

NLP libraries in python include NLTK, TextBlob, spaCy and genism along with standard machine learning libraries, including pandas and scikit-learn.

How are we using NLP in our Project?

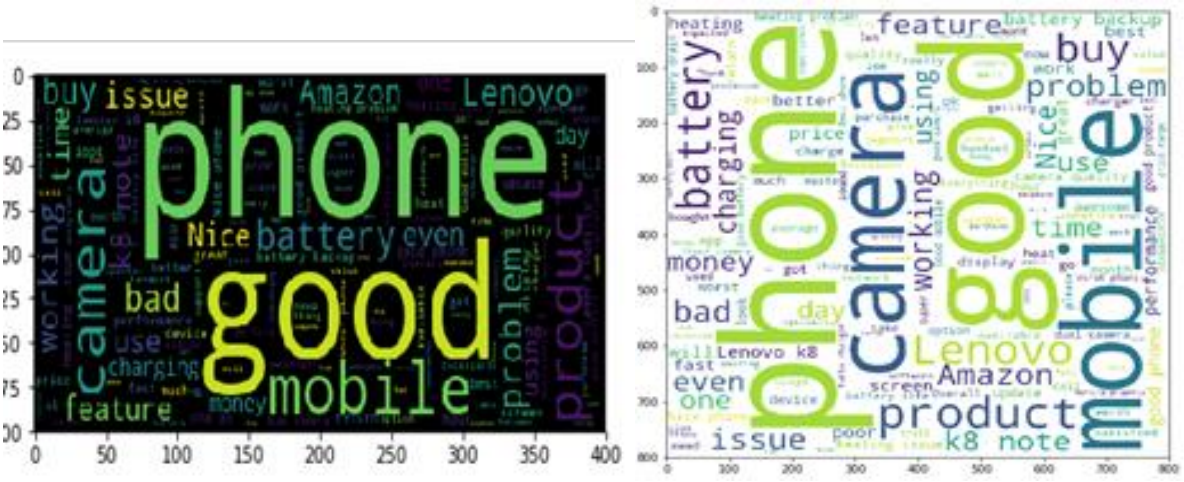
Dataset:

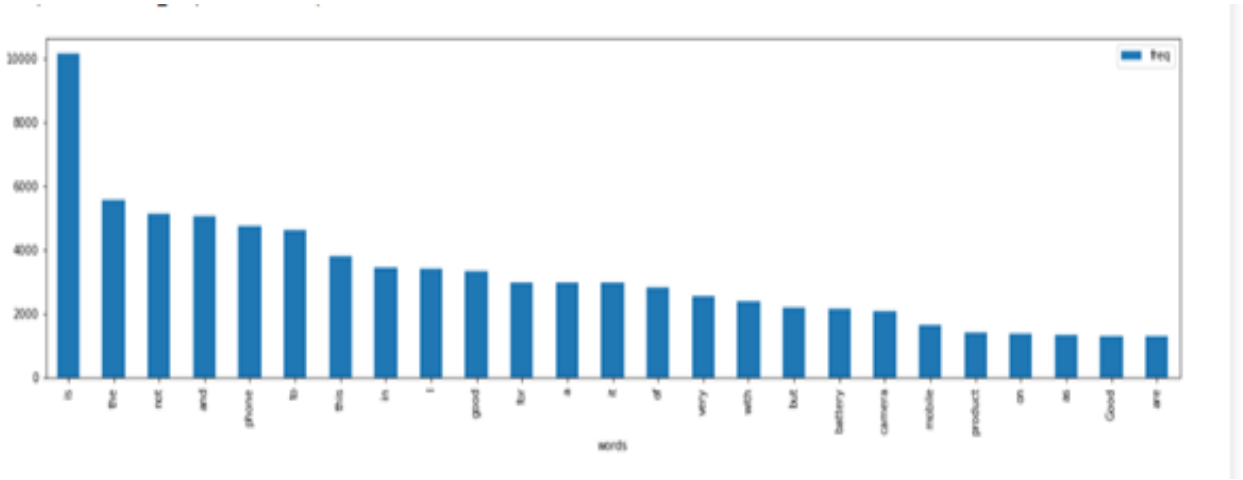
This dataset is scrapped from amazon about Lenovo K8 mobile phones. While the users give the reviews, they also give the ratings 1,2,3 -> negative sentiment has a value 0 in the sentiment column Ratings 4,5 -> positive sentiment has a value 1 in the sentiment column.

Operations Performed for NLP in our Project:

- Collection of the Data, naming and conventions
- Data Cleaning
- Case normalization and tokenizing
- visualizing the frequency distribution
- Usage of Lemmatizer and #Stemmer Functions
- AxesSubplot

Output Graphs for NLP:





Conclusion:

Based on all the Algorithms we have used and thus displayed all the results based on the Validation and verification of data, we conclude that the Website Phishing Dataset reacts differently based on the training set Data. As represented in the above algorithms, we received different accuracy for a different set of data.

Also, the report suggests the usage of Natural Language Processing as a machine learning tool. Graphs, along with the steps followed for the NLP, have beeN Included.

We have professionally cited all the references and thank our professor for giving us this opportunity to learn a plethora of data mining concepts.

References

Brownlee, J. (2016, April 1). *Logistic Regression for Machine Learning*. From machinelearningmastery: https://machinelearningmastery.com/logistic-regression-for-machine-learning/

Chauhan, N. S. (2019, December 24). *Decision Tree Algorithm — Explained*. From towardsdatascience: https://towardsdatascience.com/decision-tree-algorithm-explained-83beb6e78ef4

Fruhlinger, J. (2020, April 07). *What is phishing? How this cyber attack works and how to prevent it*. From CSO: https://www.csoonline.com/article/2117843/what-is-phishing-how-this-cyber-attack-works-and-how-to-prevent-it.html

Ray, S. (2017, September 13). *Understanding Support Vector Machine(SVM) algorithm from examples (along with code)*. From analyticsvidhya: https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/

Shetty, B. (2018, November 24). *Natural Language Processing(NLP) for Machine Learning*. From towardsdatascience: https://towardsdatascience.com/natural-language-processing-nlp-for-machine-learning-d44498845d5b#:~:text=NLP%20is%20a%20field%20in,finds%20relevant%20and%20similar%20results).

SRIVASTAVA, T. (2018, March 26). *Introduction to k-Nearest Neighbors: A powerful Machine Learning Algorithm (with implementation in Python & R)*. From analyticsvidhya: https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/

Yiu, T. (2019, June 12). *Understanding Random Forest*. From towardsdatascience: https://towardsdatascience.com/understanding-random-forest-58381e0602d2