INDUSTRIAL TRAINING

Submitted in partial fulfilment of the

Requirements for the award of the degree

Of

Bachelor of Technology

In

INFORMATION & TECHNOLOGY

By:

HARPREET SINGH MATHARU (03613203118/IT-1/2018)

Department of Information & Technology
Guru Tegh Bahadur Institute of Technology

Guru Gobind Singh Indraprastha University
Dwarka, New Delhi
Year 2018-2022

PREDICTING CAR ACCIDENT SEVERITY USING MACHINE LEARNING

Duration

10th June, 2020 – 28th Aug, 2020

By:

Harpreet Singh Matharu (03613203118/IT-1/2018)

At

IBM, Coursera

DECLARATION

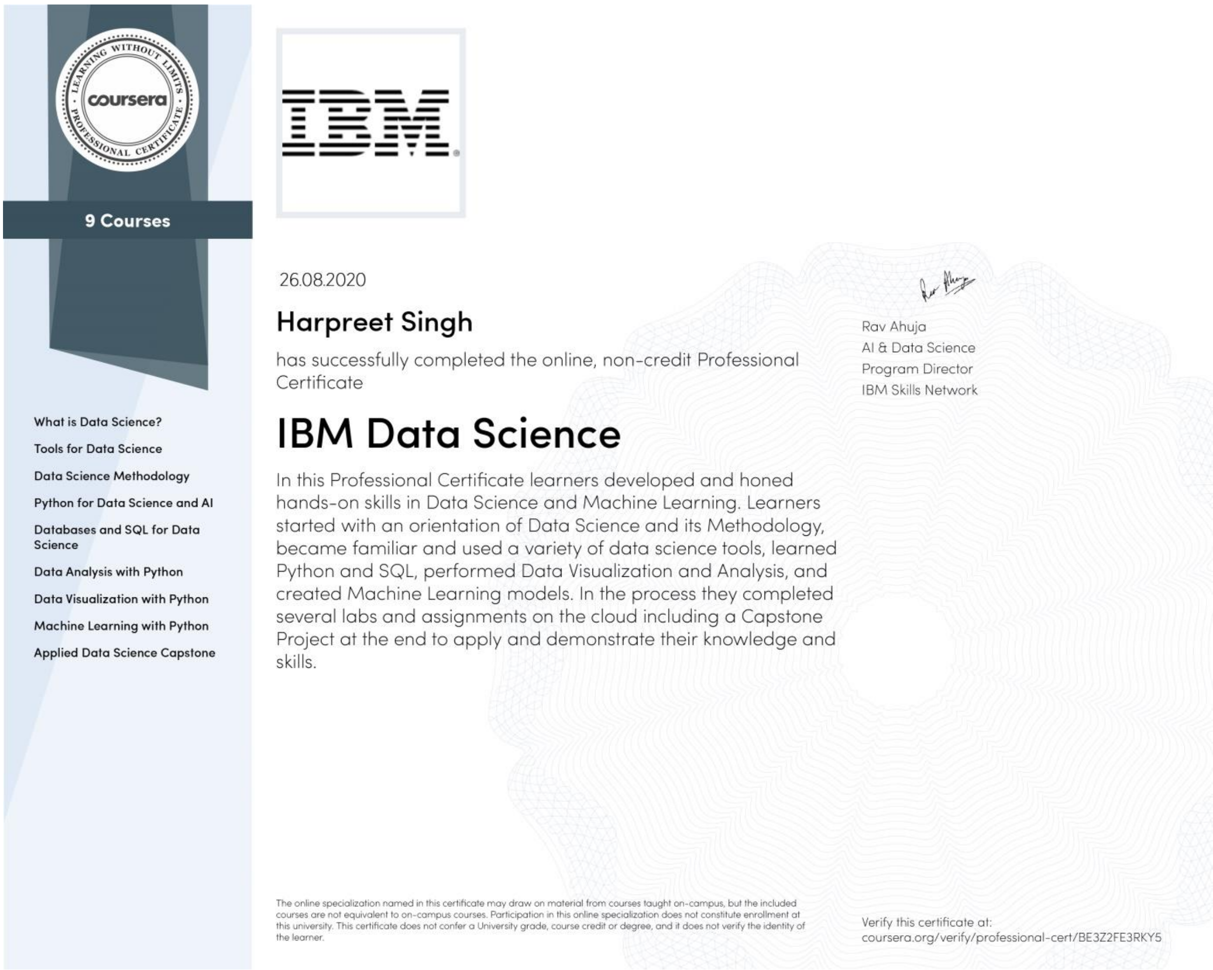
I hereby declare that all the work presented in this Industrial Training Report for the partial fulfilment of the requirements for the award of the degree of Bachelor of Technology in **Information & Technology,** Guru Tegh Bahadur Institute of Technology, affiliated to Guru Gobind Singh Indraprastha University, Delhi comprises only my original work and due acknowledgement has been made in the text to all other material used.

Date: 02-12-2020

Harpreet Singh Matharu (036/IT1/2018)

Email- harpreetmatharu3@gmail.com

CERTIFICATE



9 Courses

- What is Data Science?
- Tools for Data Science
- Data Science Methodology
- Python for Data Science and AI
- Databases and SQL for Data Science
- Data Analysis with Python
- Data Visualization with Python
- Machine Learning with Python
- Applied Data Science Capstone

26.08.2020

Harpreet Singh

has successfully completed the online, non-credit Professional Certificate

IBM Data Science

In this Professional Certificate learners developed and honed hands-on skills in Data Science and Machine Learning. Learners started with an orientation of Data Science and its Methodology, became familiar and used a variety of data science tools, learned Python and SQL, performed Data Visualization and Analysis, and created Machine Learning models. In the process they completed several labs and assignments on the cloud including a Capstone Project at the end to apply and demonstrate their knowledge and skills.

Rav Ahuja
AI & Data Science
Program Director
IBM Skills Network

The online specialization named in this certificate may draw on material from courses taught on-campus, but the included courses are not equivalent to on-campus courses. Participation in this online specialization does not constitute enrollment at this university. This certificate does not confer a University grade, course credit or degree, and it does not verify the identity of the learner.

Verify this certificate at:
coursera.org/verify/professional-cert/BE3Z2FE3RKY5

ACKNOWLEDGEMENT

Every project that is completed is never an individual effort, it consists of the knowledge and guidance of fellow supervisors and instructors that guide and help us reach the path of success and work towards making meaningful use of knowledge.

I owe my deepest gratitude to the instructors of the IBM Data Science professional certification on the Coursera platform for helping me get the foundational knowledge required to bring this project to its completion. The knowledge I gained from them proved to be invaluable for this project.

I would also like to thank the immensely helpful and useful course community that were together in the forums to help me out whenever I encountered any issues and enabled me to take on these challenges head on.

I would like to express our great gratitude towards **Mr. Kanwarjeet Singh** who has given us support and suggestions. Without their help we could not have presented this work up to the present standard. We also take this opportunity to give thanks to all others who gave us support for the project or in other aspects of our study at Guru Tegh Bahadur Institute of Technology.

ABSTRACT

Industrial training is an extremely useful and important in providing an insight on the real-world technical problem-solving aspects and the ability to gain technical training experience. A well planned, properly executed and evaluated industrial training helps a lot in developing a professional attitude. It develops an awareness of industrial approach to problem solving, based on a broad understanding of process. It helps in building professionalism and channelising technical knowledge to useful results.

The aim and motivation of this industrial training is to receive discipline, skills, teamwork and technical knowledge through a proper training environment, which will help me, as a student in the field of Information Technology, to develop a responsiveness of the self-disciplinary nature of problems in information and communication technology. Throughout this industrial training, I have learned new libraries of python programming language that were required for the system, the process of the production lines, data science techniques, the importance and value of data and have been able to implement what I have learnt for the past year as a Bachelor in Information Technology student at GTBIT, GGSIP University, Dwarka, Delhi.

During the development of this project, most of the theoretical knowledge gained during the course of studies was put to test, various effort and processes involved in designing of a logic was studied and understood during the training. During this, I undertook courses – “IBM Data Science Professional Certification” from IBM offered by Coursera.

The training gave me a good experience from the view of implementing my theoretical knowledge in practical aspects. It gave me first-hand experience of working as an engineering professional. It helped me in improving my technical, interpersonal and communication skills, both oral and written. Overall, it a great experience to have industrial training and I truly believe that it will help me in building a successful career.

LIST OF FIGURES

Fig No **Figure Name** **Page**

CONTENTS

Chapter **Page No.**

CHAPTER 1

INTRODUCTION

CHAPTER 1

INTRODUCTION

1.1 Why Machine Learning?

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves. The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. The primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly.

Resurging interest in machine learning is due to the same factors that have made data mining more popular than ever. Things like growing volumes and varieties of available data, computational processing that is cheaper and more powerful, and affordable data storage. All of these things mean it's possible to quickly and automatically produce models that can analyse bigger, more complex data and deliver faster, more accurate results – even on a very large scale. And by building precise models, an organization has a better chance of identifying profitable opportunities – or avoiding unknown risks.

Machine learning specifically is about finding parameters for a model. When the decision source is not human, we can't exactly ask for the correct logic, but we can observe the processes that seem to feed into the final decision. Our observations can be combined into a data set. In all likelihood the information we need to make the decision, or at least an approximation of it, is contained within this dataset. The way we reach these approximations depends on the nature of the problem and the algorithm being used, but in a broad sense they are all statistics. Machine learning approaches are about finding patterns in the data via statistics and transforming those patterns into a decision logic. So, we can now write a pretty complete definition of machine learning. '*Given a decision where the logic is unknown to us, but we have observations about the decision process, it is possible to use statistical analyses of these observations to create an approximation of the decision logic.'*

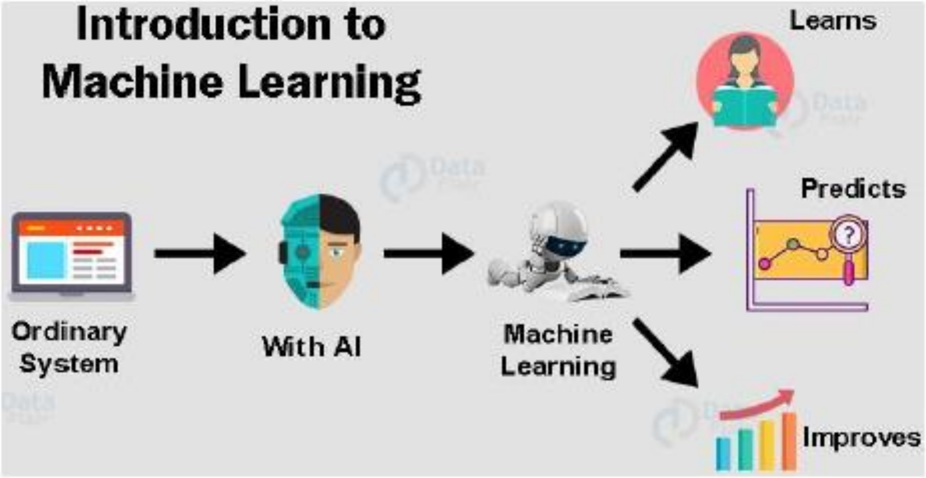


Figure 1.1: Introduction to machine learning

1.2 Types of Machine Learning

As with any method, there are different ways to train machine learning algorithms, each with their own advantages and disadvantages. To understand the pros and cons of each type of machine learning, we must first look at what kind of data they ingest. In ML, there are two kinds of data — labelled data and unlabelled data. Labelled data has both the input and output parameters in a completely machine-readable pattern, but requires a lot of human labour to label the data, to begin with. Unlabelled data only has one or none of the parameters in a machine-readable form. This negates the need for human labour but requires more complex solutions. There are also some types of machine learning algorithms that are used in very specific use-cases, but three main methods are used today:

Supervised learning: Supervised learning is one of the most basic types of machine learning. In this type, the machine learning algorithm is trained on labelled data. Even though the data needs to be labelled accurately for this method to work, supervised learning is extremely powerful when used in the right circumstances. In supervised learning, the ML algorithm is given a small training dataset to work with. This training dataset is a smaller part of the bigger dataset and serves to give the algorithm a basic idea of the problem, solution, and data points to be dealt with. The training dataset is also very similar to the final dataset in its characteristics and provides the algorithm with the labelled parameters required for the problem. The algorithm then finds relationships between the parameters given, essentially establishing a cause-and-effect relationship between the variables in the dataset. At the end of the training, the algorithm has an idea of how the data works and the relationship between the input and the output. This solution is then deployed for use with the final dataset, which it learns from in the same way as the training dataset. This means that supervised machine learning algorithms will continue to improve even after being deployed, discovering new patterns and relationships as it trains itself on new data.

Unsupervised learning: Unsupervised machine learning holds the advantage of being able to work with unlabelled data. This means that human labour is not required to make the dataset machine-readable, allowing much larger datasets to be worked on by the program. In supervised learning, the labels allow the algorithm to find the exact nature of the relationship between any two data points. However, unsupervised learning does not have labels to work off of, resulting in the creation of hidden structures. Relationships between data points are perceived by the algorithm in an abstract manner, with no input required from human beings. The creation of these hidden structures is what makes unsupervised learning algorithms versatile. Instead of a defined and set problem statement, unsupervised learning algorithms can adapt to the data by dynamically changing hidden structures. This offers more post-deployment development than supervised learning algorithms.

Reinforcement learning: Reinforcement learning directly takes inspiration from how human beings learn from data in their lives. It features an algorithm that improves upon itself and learns from new situations using a trial-and-error method. Favourable outputs are encouraged or ‘reinforced’, and non-favourable outputs are discouraged or ‘punished’. Based on the psychological concept of conditioning, reinforcement learning works by putting the algorithm in a work environment with an interpreter and a reward system. In every iteration of the algorithm, the output result is given to the interpreter, which decides whether the outcome is favourable or not. In case of the program finding the correct solution, the interpreter reinforces the solution by providing a reward to the algorithm. If the outcome is not favourable, the algorithm is forced to reiterate until it finds a better result. In most cases, the reward system is directly tied to the effectiveness of the result. In typical reinforcement learning use-cases, such as finding the shortest route between two points on a map, the solution is not an absolute value. Instead, it takes on a score of effectiveness, expressed in a percentage value. The higher this percentage

15value is, the more reward is given to the algorithm. Thus, the program is trained to give the best possible solution for the best possible reward.

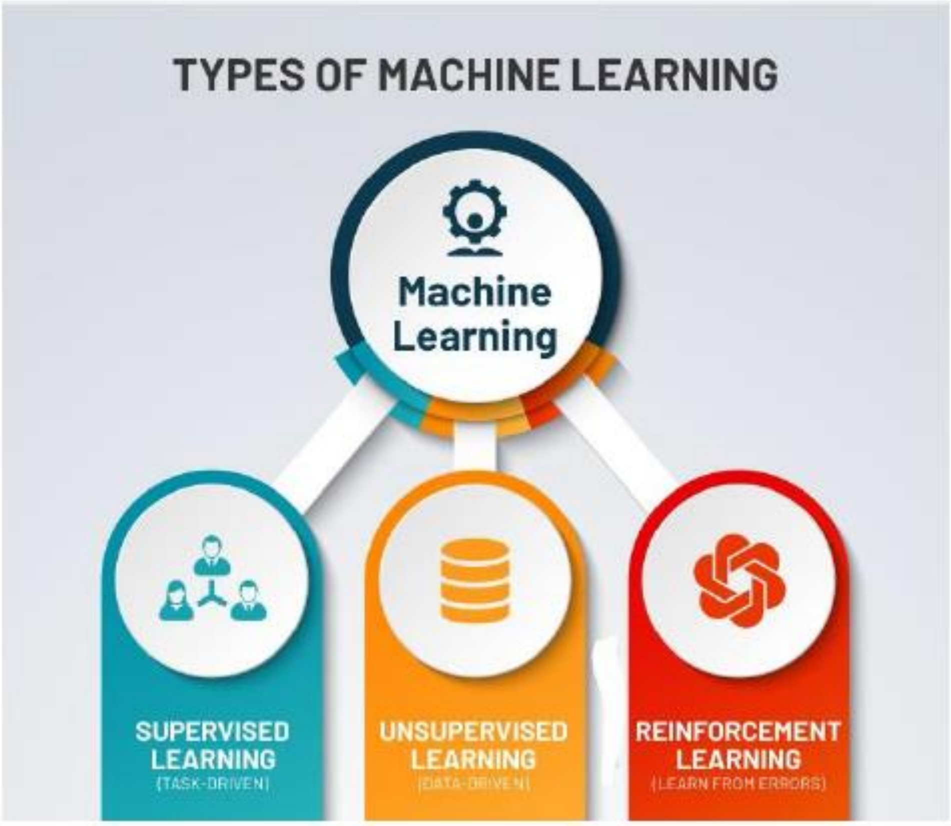


Figure 1.2: Types of machine learning

1.3 Applications of Machine Learning

Machine learning algorithms are used in circumstances where the solution is required to continue improving post-deployment. The dynamic nature of adaptable machine learning solutions is one of the main selling points for its adoption by companies and organizations across verticals. Machine learning algorithms and solutions are versatile and can be used as a substitute for medium-skilled human labour given the right circumstances. Given below are some of the popular fields machine learning is booming in:

Image recognition: Image recognition is one of the most common applications of machine learning. It is used to identify objects, persons, places, digital images, etc. The popular use case of image recognition and face detection is, Automatic friend tagging suggestion: Facebook provides us a feature of auto friend tagging suggestion. Facebook provides us a feature of auto friend tagging suggestion. Whenever we upload a photo with our Facebook friends, then we automatically get a tagging suggestion with name, and the technology behind this is machine learning's face detection and recognition algorithm.

Speech recognition: Speech recognition is a process of converting voice instructions into text, and it is also known as "Speech to text", or "Computer speech recognition." At present, machine learning algorithms are widely used by various applications of speech recognition. Google assistant, Siri, Cortana, and Alexa are using speech recognition technology to follow the voice instructions.

Traffic prediction: If we want to visit a new place, we take help of Google Maps, which shows us the correct path with the shortest route and predicts the traffic conditions. It predicts the traffic conditions such as whether traffic is cleared, slow-moving, or heavily congested with the help of two ways: Real Time location of the vehicle form Google Map app and sensors Average time has taken on past days at the same time. Everyone who is using Google Map is helping this app to make it better. It takes information from the user and sends back to its database to improve the performance.

Product recommendation: Machine learning is widely used by various e-commerce and entertainment companies such as Amazon, Netflix, etc., for product recommendation to the user. Whenever we search for some product on Amazon, then we started getting an advertisement for the same product while internet surfing on the same browser and this is because of machine learning. Google understands the user interest using various machine learning algorithms and suggests the product as per customer interest. As similar, when we use Netflix, we find some recommendations for entertainment series, movies, etc., and this is also done with the help of machine learning.

Self-driving cars: One of the most exciting applications of machine learning is self-driving cars. Machine learning plays a significant role in self-driving cars. Tesla, the most popular car manufacturing company is working on self-driving car. It is using unsupervised learning method to train the car models to detect people and objects while driving

Email Spam and Malware Filtering: Whenever we receive a new email, it is filtered automatically as important, normal, and spam. We always receive an important mail in our inbox with the important symbol and spam emails in our spam box, and the technology behind this is Machine learning. Below are some spam filters used by Gmail: Content Filter Header Filter General blacklists filter Rules-based filters Permission filters Some machine learning algorithms such as Multi-Layer Perceptron, Decision tree, and Naïve Bayes classifier are used for email spam filtering and malware detection.

Virtual Personal Assistant: We have various virtual personal assistants such as Google assistant, Alexa, Cortana, Siri. As the name suggests, they help us in finding the information using our voice instruction. These assistants can help us in various ways just by our voice instructions such as Play music, call someone, open an email, Scheduling an appointment, etc. These virtual assistants use machine learning algorithms as an important part. These assistants record our voice instructions, send it over the server on a cloud, and decode it using ML algorithms and act accordingly. 8. Online Fraud Detection: Machine learning is making our online transaction safe and secure by detecting fraud transaction. Whenever we perform some online transaction, there may be various ways that a fraudulent transaction can take place such as fake accounts, fake ids, and steal money in the middle of a transaction. So, to detect this, Feed Forward Neural network helps us by checking whether it is a genuine transaction or a fraud transaction. For each genuine transaction, the output is converted into some hash values, and these values become the input for the next round. For each genuine transaction, there is a specific pattern which gets change for the fraud transaction hence, it detects it and makes our online transactions more secure.

Stock Market trading: Machine learning is widely used in stock market trading. In the stock market, there is always a risk of up and downs in shares, so for this machine learning's long short-term memory neural network is used for the prediction of stock market trends.

Medical Diagnosis: In medical science, machine learning is used for diseases diagnoses. With this, medical technology is growing very fast and able to build 3D models that can predict the exact position of lesions in the brain. It helps in finding brain tumours and other brain-related diseases easily.

Automatic Language Translation: Nowadays, if we visit a new place and we are not aware of the language then it is not a problem at all, as for this also machine learning helps us by converting the text into our known languages. Google's GNMT (Google Neural Machine Translation) provide this feature, which is a Neural Machine Learning that translates the text into our familiar language, and it called as automatic translation. The technology behind the automatic translation is a sequence to sequence learning algorithm, which is used with image recognition and translates the text from one language to another language.

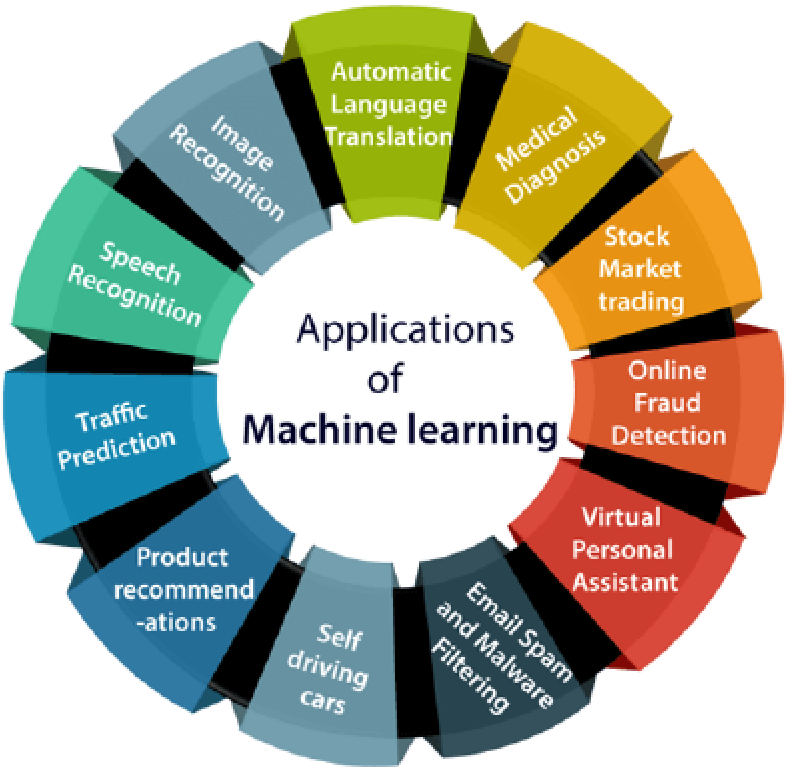

Applications of Machine learning
Automatic Language Translation
Medical Diagnosis
Stock Market trading
Online Fraud Detection
Virtual Personal Assistant
Email Spam and Malware Filtering
Self driving cars
Product recommend -ations
Traffic Prediction
Speech Recognition
Image Recognition

Figure 1.3: Applications of Machine Learning

1.4 Data Science

Data science is the field of study in which a data scientist extracts meaningful and previously hidden information from data that is used for making software more efficient, products better, increase revenue and prediction of certain things like user churn. Information is extracted from data in various ways: by exploratory data analysis, visualising the given data for visual insights and running statistical models for prediction etc. In reality, data science is evolving so fast and has already shown such enormous range of possibility that a wider definition is essential to understanding it. And while it's hard to pin down a specific definition, it's quite easy to see and feel its impact. Data science, when applied to different fields can lead to incredible new insights.

Digital data is information that is not easily interpreted by an individual but instead relies on machines to interpret, process, and alter it. The words you are reading on your computer screen are an example of this. These digital letters are actually a systematic collection of ones and zeros that encodes to pixels in various hues and at a specific density. In recent years, digital information has gotten so pervasive and essential that we've almost become unwilling to handle anything that isn't in a digital form. The digital elements of information have become essential. We cannot do modern work without them. Billions of datapoints are being acquired everyday and their use becomes easier and more productive everyday due to the following factors:

Easy availability of high-performance computing: With the advent of high-performance computing platforms like GPUs FPGAs TPAs We have been able to process such a voluminous amount of data. We are able to analyse and draw insights from this data owing to these advanced computational systems. However, despite all these advancements, data remains a vast ocean that is growing every second.

Exponential increase in speed and storage capabilities: Speed, the second prong of the big data revolution involves how, and how fast we can move around and compute with data. Advancements in speed follow a similar timeline to storage, and like storage, are the result of continuous innovation around the size and power of computers.
It's gotten so easy to write data, and so cheap to store it, that sometimes companies don't even know what value they can get from that data. They just think that at some point they may be able to do something and so it's better to save it than not. And so, the data is everywhere. About everything. Billions of devices. All over the world. Every second. Of every day. This is how you get to zetabytes. This is how you get to big data.

1.5 Data Science Methodology

The Data Science Methodology is an iterative system of methods that guides data scientists on the ideal approach to solving problems with data science, through a prescribed sequence of steps. The people who work in Data Science and are busy finding the answers for different questions every day comes across the Data Science Methodology. Data Science Methodology indicates the routine for finding solutions to a specific problem. This is a cyclic process that undergoes a critic behaviour guiding business analysts and data scientists to act accordingly.

In a nutshell, the Data Science Methodology aims to answer 10 basic questions in a prescribed sequence, that cover the four main aspects of data science projects. These aspects are:

1. From Problem to Approach:

Every customer's request starts with a problem, and Data Scientists' job is first to understand it and approach this problem with statistical and machine learning techniques. The **Business Understanding** stage is crucial because it helps to clarify the goal of the customer. The next step is the **Analytic Approach**, where, once the business problem has been clearly stated, the data scientist can define the analytic approach to solve the problem. This step entails expressing the problem in the context of statistical and machine-learning techniques, and it is essential because it helps identify what type of patterns will be needed to address the question most effectively. If the issue is to determine the probabilities of something, then a predictive model might be used; if the question is to show relationships, a descriptive approach may be required, and if our problem requires counts, then statistical analysis is the best way to solve it. For each type of approach, we can use different algorithms.

2. From Requirements to Collection:

Once we have found a way to solve our problem, we will need to discover the correct data for our model. **Data Requirements** is the stage where we identify the necessary data content, formats, and sources for initial data collection, and we use this data inside the algorithm of the approach we chose. In the **Data Collection** 20 **Stage,** data scientists identify the available data resources relevant to the problem domain. Usually, premade datasets are CSV files or Excel.

3. From Understanding to Preparation:

Now that the data collection stage is complete, data scientists use descriptive statistics and visualization techniques to understand data better. Data scientists, explore the dataset to understand its content, determine if additional data is necessary to fill any gaps but also to verify the quality of the data. In the **Data Understanding** stage, data scientists try to understand more about the data collected before. We have to check the type of each data and to learn more about the attributes and their names. In the **Data Preparation** stage, data scientists prepare data for modelling, which is one of the most crucial steps because the model has to be clean and without errors. In this stage, we have to be sure that the data are in the correct format for the machine learning algorithm we chose in the analytic approach stage. The dataframe has to have appropriate columns name, unified Boolean value (yes, no or 1, 0). We have to pay attention to the name of each data because sometimes they might be written in different characters, but they are the same thing; for example (WaTeR, water), we can fix this making all the value of a column lowercase. Another improvement can be made by deleting data exceptions from the dataframe because of their irrelevance.

4. From Modelling to Evaluation

Once data are prepared for the chosen machine learning algorithm, we are ready for modelling. In the **Modelling** stage, the data scientist has the chance to understand if his work is ready to go or if it needs review. Modelling focuses on developing models that are either descriptive or predictive, and these models are based on the analytic approach that was taken statistically or through machine learning. Descriptive modelling is a mathematical process that describes real-world events and the relationships between factors responsible for them, for example, a descriptive model might examine things like: if a person did this, then they’re likely to prefer that. Predictive modelling is a process that uses data mining and probability to forecast outcomes; for example, a predictive model might be used to determine whether an email is a spam or not. For predictive modelling, data scientists use a training set that is a set of historical data in which the outcomes are already known. This step can be repeated more times until the model understands the question and answer to it. In the **Model Evaluation** stage, data scientists can evaluate the model in two ways: Hold-Out and Cross-Validation. In the Hold-Out method, the dataset is divided into three subsets: a training set as we said in the modelling stage; a validation set that is a subset used to assess the performance of the model built in the training phase; a test set is a subset to evaluate the likely future performance of a model.

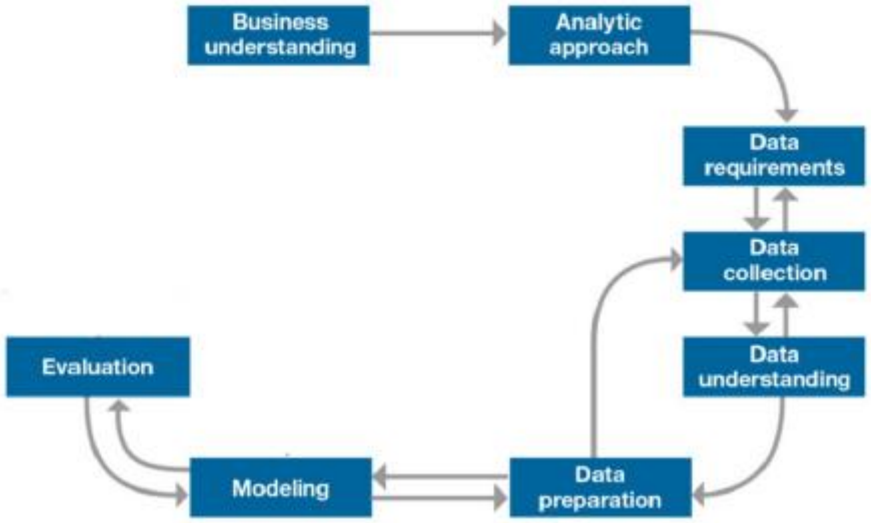


Figure 1.4: Data Science Methodology

1.6 Pros and Cons of Data Science

The field of Data Science is massive and has its own fair share of advantages and limitations. So, here we will measure the pros and cons of Data Science:

Pros:

1. **It's in Demand:**

Data Science is greatly in demand. Prospective job seekers have numerous opportunities. It is the fastest growing job on LinkedIn and is predicted to create 11.5 million jobs by 2026. This makes Data Science a highly employable job sector.

2. **Abundance of Positions:**

There are very few people who have the required skill-set to become a complete Data Scientist. This makes Data Science less saturated as compared with other IT sectors. Therefore, Data Science is a vastly abundant field and has a lot of opportunities. The field of Data Science is high in demand but low in supply of Data Scientists.

3. **A Highly Paid Career:**

Data Science is one of the most highly paid jobs. According to Glassdoor, Data Scientists make an average of $116,100 per year. This makes Data Science a highly lucrative career option.

4. Data Science is Versatile:

There are numerous applications of Data Science. It is widely used in health-care, banking, consultancy services, and e-commerce industries. Data Science is a very versatile field. Therefore, you will have the opportunity to work in various fields.

5. Data Science Makes Data Better:

Companies require skilled Data Scientists to process and analyse their data. They not only analyse the data but also improve its quality. Therefore, Data Science deals with enriching data and making it better for their company.

6. Data Scientists are Highly Prestigious:

Data Scientists allow companies to make smarter business decisions. Companies rely on Data Scientists and use their expertise to provide better results to their clients. This gives Data Scientists an important position in the company.

7. No More Boring Tasks:

Data Science has helped various industries to automate redundant tasks. Companies are using historical data to train machines in order to perform repetitive tasks. This has simplified the arduous jobs undertaken by humans before.

8. Data Science Makes Products Smarter:

Data Science involves the usage of Machine Learning which has enabled industries to create better products tailored specifically for customer experiences. For example, Recommendation Systems used by e-commerce websites provide personalized insights to users based on their historical purchases. This has enabled computers to understand human-behavior and take data-driven decisions.

9. Data Science can Save Lives:

Healthcare sector has been greatly improved because of Data Science. With the advent of machine learning, it has been made easier to detect early-stage tumours. Also, many other health-care industries are using Data Science to help their clients.

Cons:

1. Mastering Data Science is near to impossible:

Being a mixture of many fields, Data Science stems from Statistics, Computer Science and Mathematics. It is far from possible to master each field and be equivalently expert in all of them. While many online courses have been trying to fill the skill-gap that the data science industry is facing, it is still not possible to be proficient at it considering the immensity of the field. A person with a background in Statistics may not be able to master Computer Science on short notice in order to become a proficient Data Scientist. Therefore, it is an ever-changing, dynamic field that requires the person to keep learning the various avenues of Data Science.

2. Large Amount of Domain Knowledge Required

Another disadvantage of Data Science is its dependency on Domain Knowledge. A person with a considerable background in Statistics and Computer Science will find it difficult to solve Data Science problem without its background knowledge. The same holds true for its vice-versa. For example, A health-care industry working on an analysis of genomic sequences will require a suitable employee with some knowledge of genetics and molecular biology. This allows the Data Scientists to make calculated decisions in order to assist the company. However, it becomes difficult for a Data Scientist from a different background to acquire specific domain knowledge. This also makes it difficult to migrate from one industry to another.

3. Arbitrary Data May Yield Unexpected Results

A Data Scientist analyzes the data and makes careful predictions in order to facilitate the decision-making process. Many times, the data provided is arbitrary and does not yield expected results. This can also fail due to weak management and poor utilization of resources. You Must Read – Difference Between Data Science and Data Analytics.

4. Problem of Data Privacy

For many industries, data is their fuel. Data Scientists help companies make data-driven decisions. However, the data utilized in the process may breach the privacy of customers. The personal data of clients are visible to the parent company and may at times cause data leaks due to lapse in security. The ethical issues regarding preservation of data-privacy and its usage have been a concern for many industries

CHAPTER 2

SCOPE OF WORK

CHAPTER 2

SCOPE OF THE WORK

2.1 Motivation

2.1.1 Adversity of road accidents

Road accidents are one of the most serious adversities and harmful aspects of the modern world. Road vehicle accidents cause a capricious amount of damage to public health and infrastructure. According to WHO:

- Approximately 1.35 million people die each year as a result of road traffic crashes. Road traffic crashes cost most countries 3% of their gross domestic product.
- More than half of all road traffic deaths are among vulnerable road users: pedestrians, cyclists, and motorcyclists.
- 93% of the world's fatalities on the roads occur in low- and middle-income countries, even though these countries have approximately 60% of the world's vehicles.
- Road traffic injuries are the leading cause of death for children and young adults aged 5-29 years.

The following are the causes of road vehicle accidents:

- **Speeding:** An increase in average speed is directly related both to the likelihood of a crash occurring and to the severity of the consequences of the crash. For example, every 1% increase in mean speed produces a 4% increase in the fatal crash risk and a 3% increase in the serious crash risk. The death risk for pedestrians hit by car fronts rises rapidly (4.5 times from 50 km/h to 65 km/h). In car-to-car side impacts the fatality risk for car occupants is 85% at 65 km/h.
- **Driving under the influence of alcohol and other psychoactive substances:** Driving under the influence of alcohol and any psychoactive substance or drug increases the risk of a crash that results in death or serious injuries. In the case of drink-driving, the risk of a road traffic crash starts at low levels of blood alcohol concentration (BAC) and increases significantly when the driver's BAC is ≥ 0.04 g/dl. In the case of drug-driving, the risk of incurring a road traffic crash is

increased to differing degrees depending on the psychoactive drug used. For example, the risk of a fatal crash occurring among those who have used amphetamines is about 5 times the risk of someone who hasn't.

- **Non-use of motorcycle helmets, seat-belts, and child restraints:** Correct helmet use can lead to a 42% reduction in the risk of fatal injuries and a 69% reduction in the risk of head injuries. Wearing a seat-belt reduces the risk of death among drivers and front seat occupants by 45 - 50%, and the risk of death and serious injuries among rear seat occupants by 25%. The use of child restraints can lead to a 60% reduction in deaths.
- **Distracted driving:** There are many types of distractions that can lead to impaired driving. The distraction caused by mobile phones is a growing concern for road safety. Drivers using mobile phones are approximately 4 times more likely to be involved in a crash than drivers not using a mobile phone. Using a phone while driving slows reaction times (notably braking reaction time, but also reaction to traffic signals), and makes it difficult to keep in the correct lane, and to keep the correct following distances. Hands-free phones are not much safer than hand-held phone sets, and texting considerably increases the risk of a crash.
- **Unsafe road infrastructure:** The design of roads can have a considerable impact on their safety. Ideally, roads should be designed keeping in mind the safety of all road users. This would mean making sure that there are adequate facilities for pedestrians, cyclists, and motorcyclists. Measures such as footpaths, cycling lanes, safe crossing points, and other traffic calming measures can be critical to reducing the risk of injury among these road users.
- **Unsafe vehicles:** Safe vehicles play a critical role in averting crashes and reducing the likelihood of serious injury. There are a number of UN regulations on vehicle safety that, if applied to countries’ manufacturing and production standards, would potentially save many lives. These include requiring vehicle manufacturers to meet front and side impact regulations, to include electronic stability control (to prevent over-steering) and to ensure airbags and seat-belts are fitted in all vehicles. Without these basic standards the risk of traffic injuries – both to those in the vehicle and those out of it – is considerably increased.
- **Inadequate post-crash care:** Delays in detecting and providing care for those involved in a road traffic crash increase the severity of injuries. Care of injuries after a crash has occurred is extremely time-sensitive: delays of minutes can

make the difference between life and death. Improving post-crash care requires ensuring access to timely prehospital care, and improving the quality of both prehospital and hospital care, such as through specialist training programmes.

- **Inadequate law enforcement of traffic laws**: If traffic laws on drink-driving, seat-belt wearing, speed limits, helmets, and child restraints are not enforced, they cannot bring about the expected reduction in road traffic fatalities and injuries related to specific behaviours. Thus, if traffic laws are not enforced or are perceived as not being enforced it is likely they will not be complied with and therefore will have very little chance of influencing behaviour. Effective enforcement includes establishing, regularly updating, and enforcing laws at the national, municipal, and local levels that address the above-mentioned risk factors. It includes also the definition of appropriate penalties.

Motivated by all these factors of the impact of road accidents on daily lives and how adversely road vehicle accidents affect our daily lives, this project will mainly focus on discovering and applying algorithm that allow accurate, fast, and efficient way of determining the intensity of a traffic congestion caused by a road vehicle accident. It will cover data acquisition, data processing, data visualisation, feature extraction, model training, results analysis, and future works.

2.2 Problem Statement

Classification problems have forever been one of the pillars in machine learning use case scenarios. Considering the immense problem of the adverse effect of traffic congestions and road accidents in our daily lives, the aim is to devise a classification problem approach that classifies the severity of an accident using the various types of datapoints or information that describe the scenario during the accident and the intensity along with it.

Machine learning approach is particularly useful as the amount of data is gigantic and processing of large-scale data, or, big data is the sweet spot to use data science algorithms in. Machine learning approach towards classification is gaining more and more importance due to its ease of handling data and using data in an efficient and fast way to determine new information and alongside keep on learning to constantly update and improve results.

2.3 Objectives

In this project, the aim will be to take unformatted raw data containing information about the various conditions present during the time of the accident in various scenarios and do the following:

Data pre-processing: We will take the data and filter it out for important data attributes, outliers, unknown data points.

Data visualisation: We will visualise the given data in various ways to uncover visual insights.

Exploratory Data Analysis: We will run various functions on processed data to uncover new information about the dataset.

Model Building: Create a classification machine learning model that defines the intensity of a traffic congestion caused by a road vehicle accident.

Evaluation: The model will then be evaluated using various metrics to evaluate its accuracy and define its ability to be used in real world applications.

CHAPTER 3

PROPOSED WORK

CHAPTER 3

PROPOSED WORK

3.1 Dataset

A data set consists of roughly two components. The two components are rows and columns. Additionally, a key feature of a data set is that it is organized so that each row contains one observation.

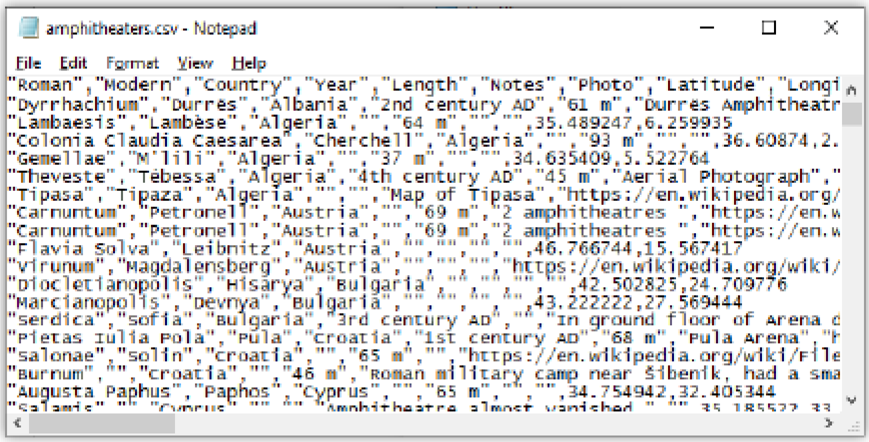
Several characteristics define a data set's structure and properties. These include the number and types of the attributes or variables, and various statistical measures applicable to them, such as standard deviation and kurtosis. The values may be numbers, such as real numbers or integers, for example representing a person's height in centimetres, but may also be nominal data (i.e., not consisting of numerical values), for example representing a person's ethnicity. More generally, values may be of any of the kinds described as a level of measurement. For each variable, the values are normally all of the same kind. However, there may also be missing values, which must be indicated in some way. In statistics, data sets usually come from actual observations obtained by sampling a statistical population, and each row corresponds to the observations on one element of that population. Data sets may further be generated by algorithms for the purpose of testing certain kinds of software. Some modern statistical analysis software such as SPSS still present their data in the classical data set fashion. If data is missing or suspicious an imputation method may be used to complete a data set.

3.1.1 Dataset Formats:

A dataset can be in various formats. Every type of dataset has a different approach to deal with. Some datasets need to be completely converted before using to actually access the values. Some datasets are very small in size and some datasets are very large to process hence they need to be compressed before using. The following are some common formats for datasets:

CSV: A Comma Separated Values (CSV) file is a plain text file that contains a list of data. These files are often used for exchanging data between different applications. For example, databases and contact managers often support CSV files. These files may sometimes be called Character Separated Values or Comma Delimited files. They mostly use the comma character to separate (or delimit) data, but sometimes use other characters, like semicolons. The idea is that you can export complex data from one application to a CSV file, and then import the data in that CSV file into another application.

A CSV representation of a shopping list with a header row, for example, looks like this:



```
"Roman","Modern","Country","Year","Length","Notes","Photo","Latitude","Longi
"Dyrrhachium","Durrës","Albania","2nd century AD","61 m","Durrës Amphitheatr
"Lambaesis","Lambèse","Algeria","","64 m","","",35.489247,6.259935
"Colonia Claudia Caesarea","Cherchell","Algeria","","93 m","","",36.60874,2.
"Gemellae","M'lili","Algeria","","37 m","","",34.635409,5.522764
"Theveste","Tébessa","Algeria","4th century AD","45 m","Aerial Photograph","
"Tipasa","Tipaza","Algeria","","","Map of Tipasa","https://en.wikipedia.org/
"Carnuntum","Petronell","Austria","","69 m","2 amphitheatres ","https://en.w
"Carnuntum","Petronell","Austria","","69 m","2 amphitheatres ","https://en.w
"Flavia Solva","Leibnitz","Austria","","","","",46.766744,15.567417
"Virunum","Magdalensberg","Austria","","","","https://en.wikipedia.org/wiki/
"Diocletianopolis","Hisarya","Bulgaria","","","","",42.502825,24.709776
"Marcianopolis","Devnya","Bulgaria","","","","",43.222222,27.569444
"Serdica","Sofia","Bulgaria","3rd century AD","","In ground floor of Arena d
"Pietas Iulia Pola","Pula","Croatia","1st century AD","68 m","Pula Arena","h
"Salonae","Solin","Croatia","","65 m","","https://en.wikipedia.org/wiki/File
"Burnum","","Croatia","","46 m","Roman military camp near Šibenik, had a sma
"Augusta Paphus","Paphos","Cyprus","","65 m","","",34.754942,32.405344
"Salamis","","Cyprus","","","Amphitheatre almost vanished ","",35.185522,33
```

Figure 3.1: Example of a CSV file

CSVs are the most common of the file formats available on Kaggle and are the best choice for tabular data. CSV files will also have associated column descriptions and column metadata. The column descriptions allow you to assign descriptions to individual columns of the dataset, making it easier for users to understand what each column means. Column metrics, meanwhile, present high-level metrics about individual columns in a graphic format.

3.1.2 Data Understanding:

Seattle also known as the Emerald city is Washington largest city and has many international people. Due to increasing population, number of vehicles in Seattle has increased and it causes often traffic jams. In addition, Seattle’s rainy weather is adding fire to the bad traffic conditions.

Road accidents have been major cause of concern across Seattle City, claiming about thousand lives per year.

Next to this intolerably high number of lives lost, about millions are injured in road traffic crashes. As a result, our societies bear a huge cost. Road traffic injuries are the leading cause of death among young people in the region and are predicted to increase in countries with low or medium income as they become more highly motorized.

The fact that effective preventive strategies do exist makes this situation unacceptable. The success of some other states in reducing the toll of deaths and injuries on their roads clearly demonstrates strong commitment. Much can be learned from these experiences, innovative approaches and can be reapplied to various situations. With this project, our objective is to reduce road accidents by warning traffic ahead of time.

Road collisions data consist of information related to severity of the road collisions along with various factors that could cause road collisions. Injury Collision and Property Damage Collision are two the severe collisions recorded by Traffic management. Other major causes and details of collisions recorded includes Weather, Road condition, Light condition, junction information, speeding, vehicle count, person count and so on.

Data set contains **194673 rows × 38 columns**. Data captured by traffic management has lot of empty fields, NAN (not a number) and other invalid values. These values won’t be considered for data modeling as it may result in incorrect information. Classification algorithm will be used for modeling as target variable SEVERITY is categorical variable with discrete value ‘Property damage’ and ‘Injury collision’. This dataset consists data from 1990 to 2013 of accidents happened in Seattle city.

3.2 Geographical Map to show accident numbers:

Below map gives the good overview of total number of accidents in Seattle city. More you zoom into the map; detailed area wise information will be revealed. Traffic management can use this map to identify accidents hot spots and provide more frequent warning to the traffic in accident prone areas.

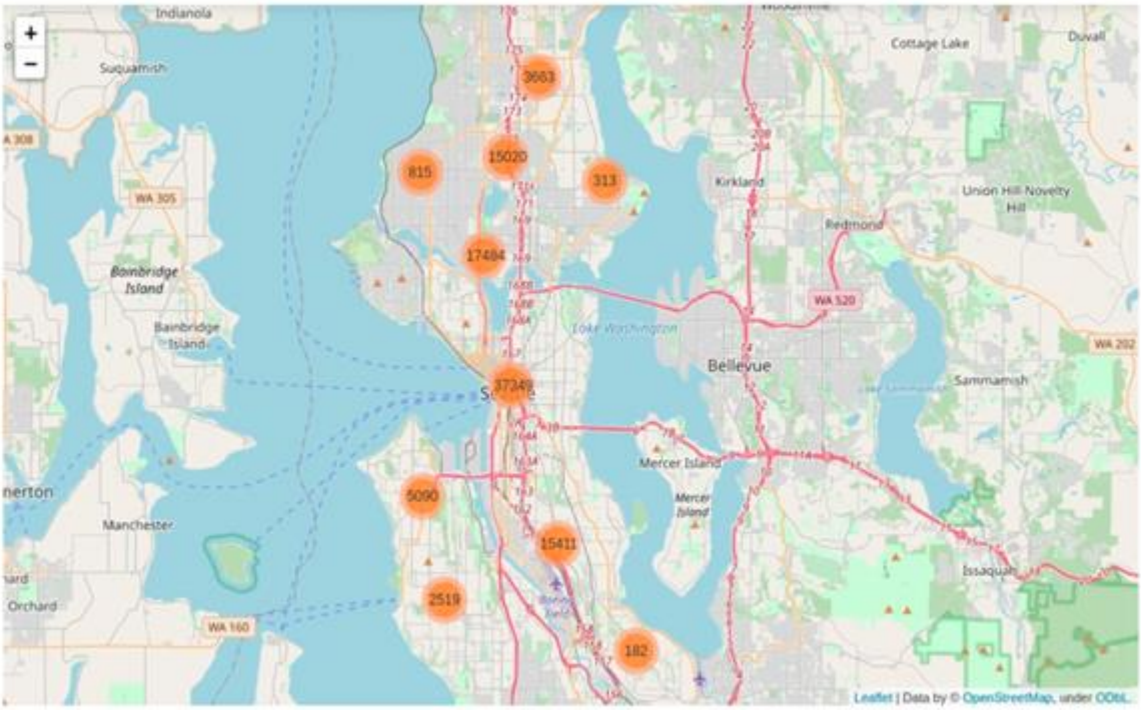
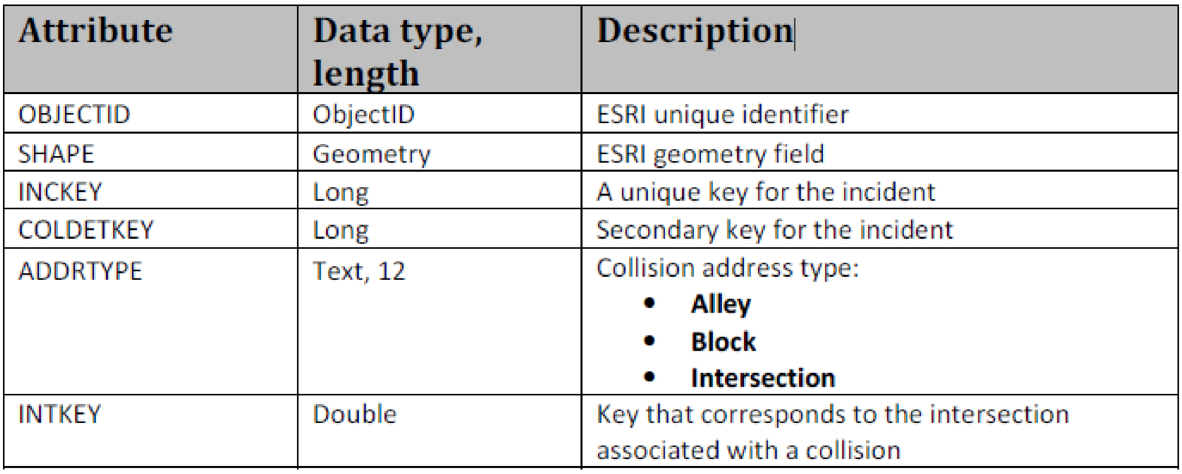
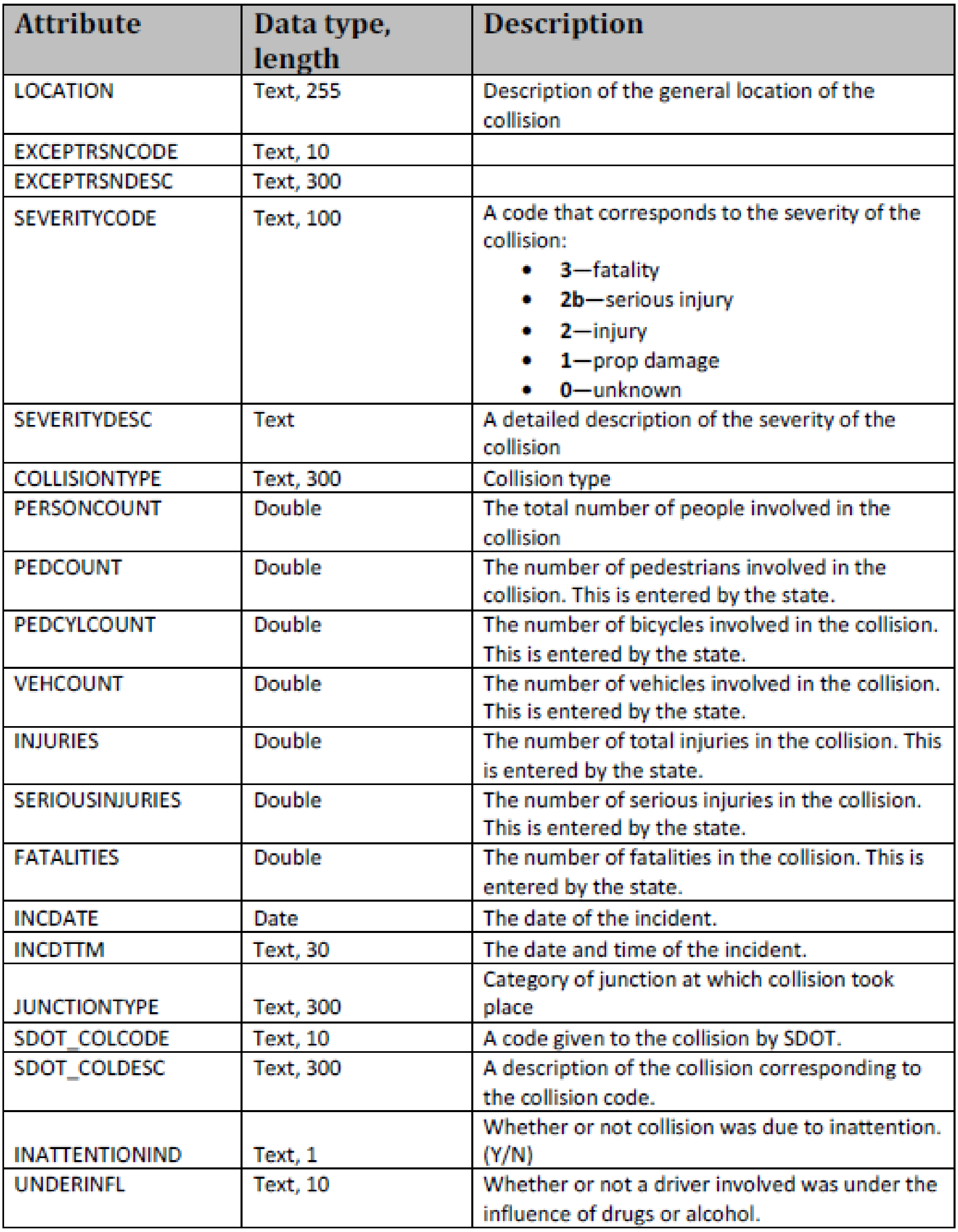


Fig 3.2 Map of Seattle city showing number of accidents in an area

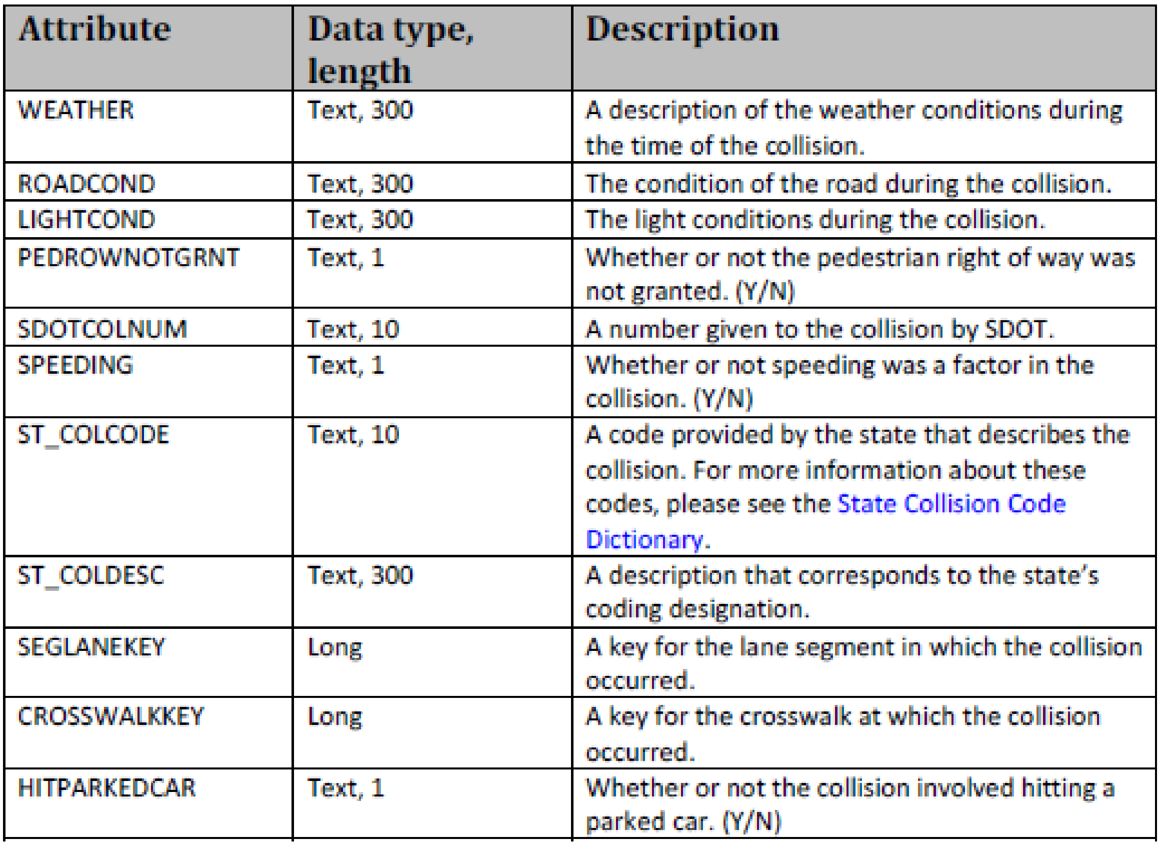
The following table explains all of the features in the dataset:



Attribute	**Data type, length**	**Description**
OBJECTID	ObjectID	ESRI unique identifier
SHAPE	Geometry	ESRI geometry field
INCKEY	Long	A unique key for the incident
COLDETKEY	Long	Secondary key for the incident
ADDRTYPE	Text, 12	Collision address type: • **Alley** • **Block** • **Intersection**
INTKEY	Double	Key that corresponds to the intersection associated with a collision



Attribute	Data type, length	Description
LOCATION	Text, 255	Description of the general location of the collision
EXCEPTRSNCODE	Text, 10	
EXCEPTRSNDESC	Text, 300	
SEVERITYCODE	Text, 100	A code that corresponds to the severity of the collision: • **3**—fatality • **2b**—serious injury • **2**—injury • **1**—prop damage • **0**—unknown
SEVERITYDESC	Text	A detailed description of the severity of the collision
COLLISIONTYPE	Text, 300	Collision type
PERSONCOUNT	Double	The total number of people involved in the collision
PEDCOUNT	Double	The number of pedestrians involved in the collision. This is entered by the state.
PEDCYLCOUNT	Double	The number of bicycles involved in the collision. This is entered by the state.
VEHCOUNT	Double	The number of vehicles involved in the collision. This is entered by the state.
INJURIES	Double	The number of total injuries in the collision. This is entered by the state.
SERIOUSINJURIES	Double	The number of serious injuries in the collision. This is entered by the state.
FATALITIES	Double	The number of fatalities in the collision. This is entered by the state.
INCDATE	Date	The date of the incident.
INCDTTM	Text, 30	The date and time of the incident.
JUNCTIONTYPE	Text, 300	Category of junction at which collision took place
SDOT_COLCODE	Text, 10	A code given to the collision by SDOT.
SDOT_COLDESC	Text, 300	A description of the collision corresponding to the collision code.
INATTENTIONIND	Text, 1	Whether or not collision was due to inattention. (Y/N)
UNDERINFL	Text, 10	Whether or not a driver involved was under the influence of drugs or alcohol.



Attribute	Data type, length	Description
WEATHER	Text, 300	A description of the weather conditions during the time of the collision.
ROADCOND	Text, 300	The condition of the road during the collision.
LIGHTCOND	Text, 300	The light conditions during the collision.
PEDROWNOTGRNT	Text, 1	Whether or not the pedestrian right of way was not granted. (Y/N)
SDOTCOLNUM	Text, 10	A number given to the collision by SDOT.
SPEEDING	Text, 1	Whether or not speeding was a factor in the collision. (Y/N)
ST_COLCODE	Text, 10	A code provided by the state that describes the collision. For more information about these codes, please see the State Collision Code Dictionary.
ST_COLDESC	Text, 300	A description that corresponds to the state's coding designation.
SEGLANEKEY	Long	A key for the lane segment in which the collision occurred.
CROSSWALKKEY	Long	A key for the crosswalk at which the collision occurred.
HITPARKEDCAR	Text, 1	Whether or not the collision involved hitting a parked car. (Y/N)

3.3 Installations

3.3.1 Pandas

Fig 3.3 Pandas

Pandas is a high-level data manipulation tool developed by Wes McKinney. It is built on the Numpy package and its key data structure is called the DataFrame. DataFrames allow you to store and manipulate tabular data in rows of observations and columns of variables. There are several ways to create a DataFrame. One way is to use a dictionary.

Pandas is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language.

3.3.2 Seaborn

Fig 3.4 Seaborn

Seaborn is a library for making statistical graphics in Python. It builds on top of matplotlib and integrates closely with pandas data structures. Seaborn helps you explore and understand your data. Its plotting functions operate on data frames and arrays containing whole datasets and internally perform the necessary semantic mapping and statistical aggregation to produce informative plots. Its dataset-oriented, declarative API lets you focus on what the different elements of your plots mean, rather than on the details of how to draw them.

3.3.3 Matplotlib

Fig 3.4 Matplotlib

Matplotlib is an amazing visualization library in Python for 2D plots of arrays. Matplotlib is a multi-platform data visualization library built on NumPy arrays and designed to work with the broader SciPy stack. It was introduced by John Hunter in the year 2002. One of the greatest benefits of visualization is that it allows us visual access to huge amounts of data in easily digestible visuals. Matplotlib consists of several plots like line, bar, scatter, histogram etc.

3.3.4 SKLearn

Figure 3.8: SKLearn

Scikit-learn (formerly scikits.learn and also known as sklearn) is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.

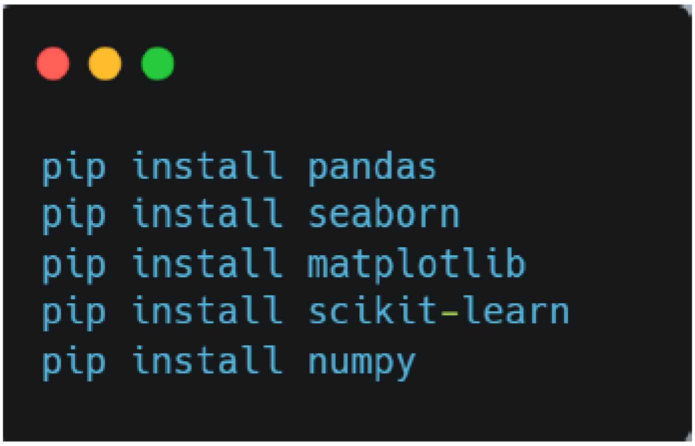
Scikit-learn is largely written in Python, and uses NumPy extensively for high-performance linear algebra and array operations. Furthermore, some core algorithms are written in Python to improve performance. Support vector machines are implemented by a Python wrapper around LIBSVM; logistic regression and linear support vector machines by a similar wrapper around LIBLINEAR. In such cases, extending these methods with Python may not be possible. Scikit-learn integrates well with many other Python libraries, such as matplotlib and plotly for plotting, NumPy for array vectorization, panda’s data frames, SciPy, and many more.

3.3.5 NumPy

Figure 3.9 NumPy

NumPy is the fundamental package for scientific computing in Python. It is a Python library that provides a multidimensional array object, various derived objects (such as masked arrays and matrices), and an assortment of routines for fast operations on arrays, including mathematical, logical, shape manipulation, sorting, selecting, I/O, discrete Fourier transforms, basic linear algebra, basic statistical operations, random simulation and much more.

All the above libraries mentioned were installed using the Anaconda Command Prompt using the following code:



```
pip install pandas
pip install seaborn
pip install matplotlib
pip install scikit-learn
pip install numpy
```

Figure 3.10: Installing the libraries

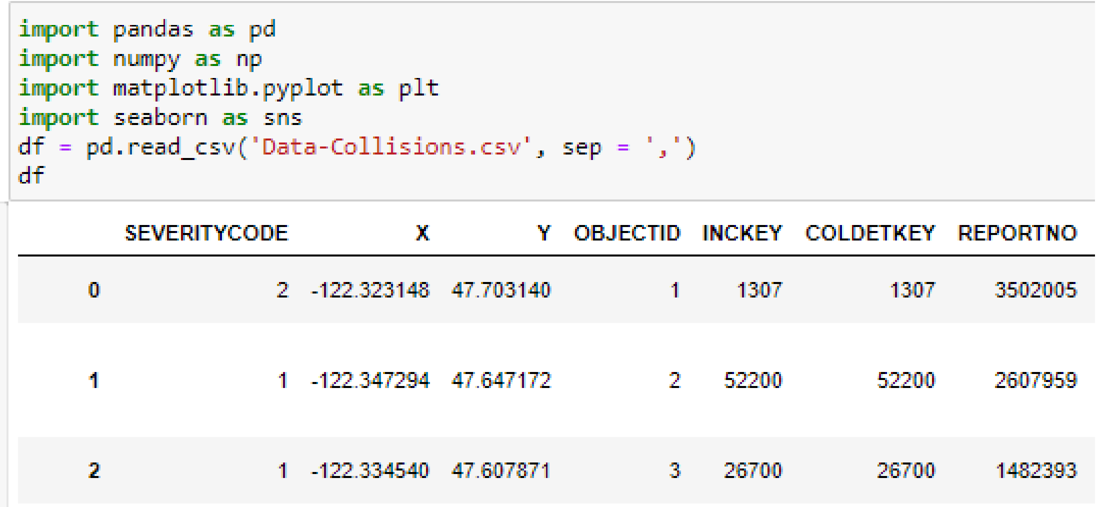
3.4 Data Pre-Processing

On a predictive modelling project, such as classification or regression, raw data typically cannot be used directly. This is because of reasons such as:

- Machine learning algorithms require data to be numbers. Some machine learning algorithms impose requirements on the data.
- Statistical noise and errors in the data may need to be corrected. Complex nonlinear relationships may be teased out of the data.
- As such, the raw data must be pre-processed prior to being used to fit and evaluate a machine learning model. This step in a predictive modelling project is referred to as “data preparation”.

We can define data preparation as the transformation of raw data into a form that is more suitable for modelling.

Hence, first all the necessary libraries are imported in the notebook and the data is loaded.

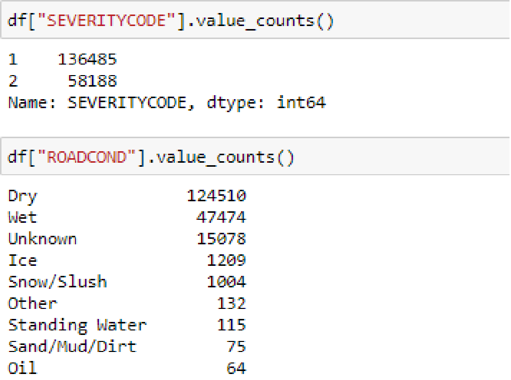


```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
df = pd.read_csv('Data-Collisions.csv', sep = ',')
df
```

	SEVERITYCODE	X	Y	OBJECTID	INCKEY	COLDETKEY	REPORTNO
0	2	-122.323148	47.703140	1	1307	1307	3502005
1	1	-122.347294	47.647172	2	52200	52200	2607959
2	1	-122.334540	47.607871	3	26700	26700	1482393

Fig 3.11 Importing libraries and loading data

Further checking the value counts of important features:



```
df["SEVERITYCODE"].value_counts()
```

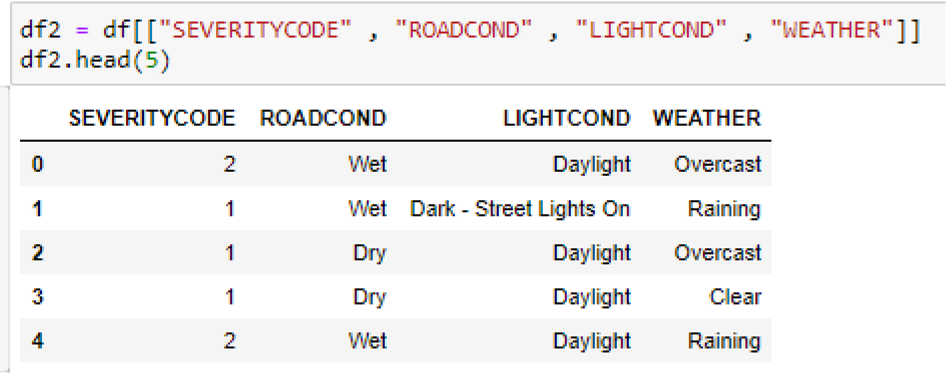
```
1    136485
2     58188
Name: SEVERITYCODE, dtype: int64
```

```
df["ROADCOND"].value_counts()
```

```
Dry               124510
Wet                47474
Unknown            15078
Ice                 1209
Snow/Slush          1004
Other                132
Standing Water       115
Sand/Mud/Dirt         75
Oil                   64
```

Fig 3.12 Checking value counts

3.4.1 Feature Selection



```
df2 = df[["SEVERITYCODE" , "ROADCOND" , "LIGHTCOND" , "WEATHER"]]
df2.head(5)
```

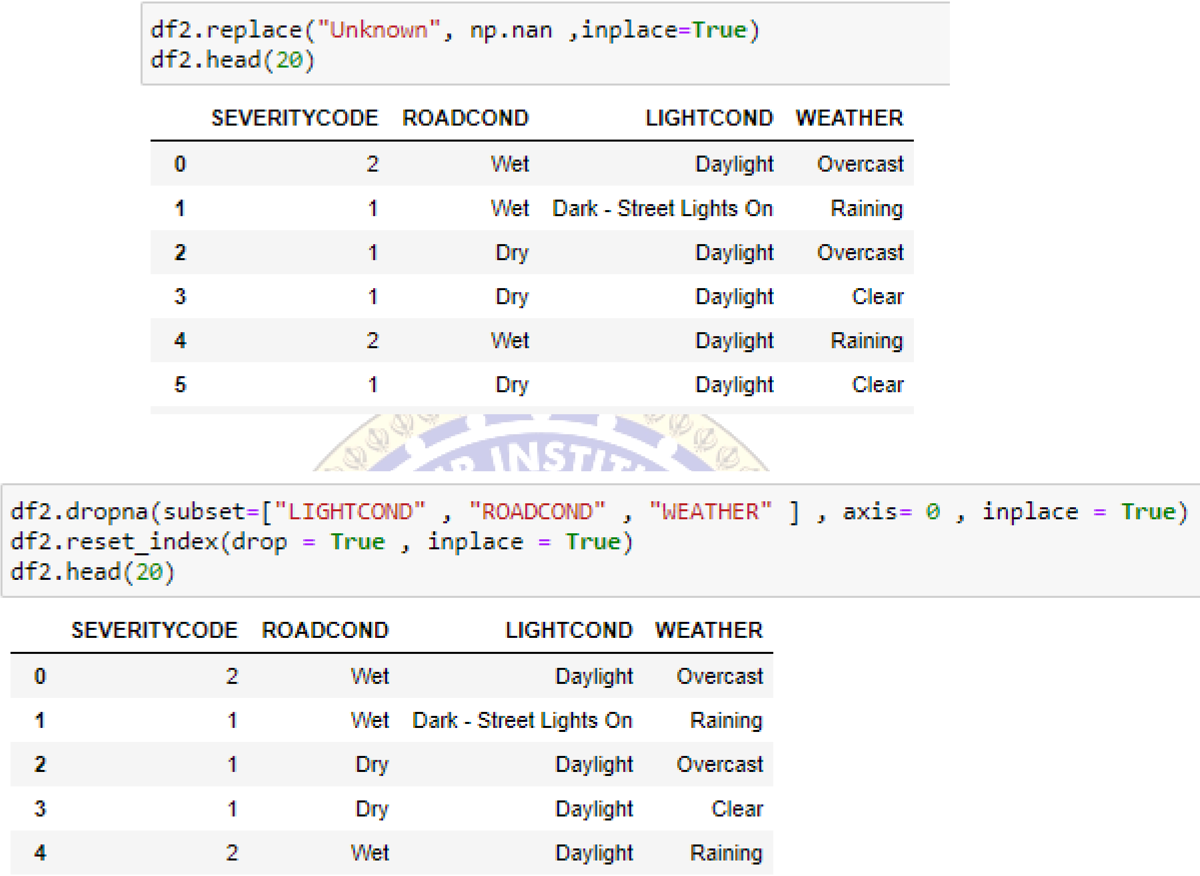
	SEVERITYCODE	ROADCOND	LIGHTCOND	WEATHER
0	2	Wet	Daylight	Overcast
1	1	Wet	Dark - Street Lights On	Raining
2	1	Dry	Daylight	Overcast
3	1	Dry	Daylight	Clear
4	2	Wet	Daylight	Raining

3.12 Feature Selection

Feature selection plays important role in predictive modeling. Correct feature selection will result in greater accuracy. Feature selection is performed using mutual information method of scikit learn library.After analyzing the feature, it is observed that COLLISION TYPE feature affects the most in determining the collision severity (target variable). However, COLLITION TYPE, ADDRTYPE and JUNCTION TYPE features can’t predict the severity beforehand. These features can only be known after the accident. Hence, COLLISION TYPE, ADDRTYPE and JUNCTION TYPE features will not be considered for modeling.

Four features WEATHER, ROADCONDITION, LIGHTCONDITION, and SEVERITYCODE are used to for modeling. Rest of the features doesn't affect much on severity and won't be considered for modeling.

3.4.2 Dropping Unknown Values



```
df2.replace("Unknown", np.nan ,inplace=True)
df2.head(20)
```

	SEVERITYCODE	ROADCOND	LIGHTCOND	WEATHER
0	2	Wet	Daylight	Overcast
1	1	Wet	Dark - Street Lights On	Raining
2	1	Dry	Daylight	Overcast
3	1	Dry	Daylight	Clear
4	2	Wet	Daylight	Raining
5	1	Dry	Daylight	Clear

```
df2.dropna(subset=["LIGHTCOND" , "ROADCOND" , "WEATHER" ] , axis= 0 , inplace = True)
df2.reset_index(drop = True , inplace = True)
df2.head(20)
```

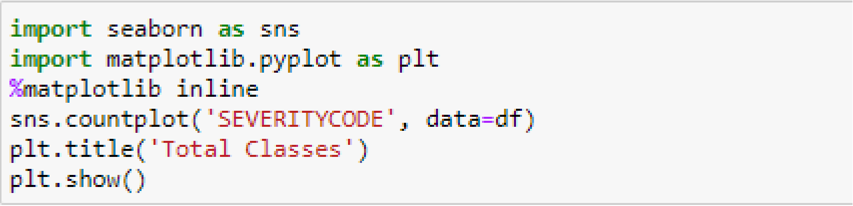
	SEVERITYCODE	ROADCOND	LIGHTCOND	WEATHER
0	2	Wet	Daylight	Overcast
1	1	Wet	Dark - Street Lights On	Raining
2	1	Dry	Daylight	Overcast
3	1	Dry	Daylight	Clear
4	2	Wet	Daylight	Raining

Fig 3.13 Dropping Nan Values

The unknown values are dropped in the dataset for higher accuracy of the model by first converting the unknowns to NaN and then dropping all the rows which consists NaN in the feature selected column.

3.4.3 Resampling the Dataset

Severity of the property damage (value=0) accidents is almost as double as the one involving injuries (value=1). RandomUnderSampler resampling technique is used to balance the data set in order to improve the accuracy. Below you can see how data set is balanced:



```
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
sns.countplot('SEVERITYCODE', data=df)
plt.title('Total Classes')
plt.show()
```

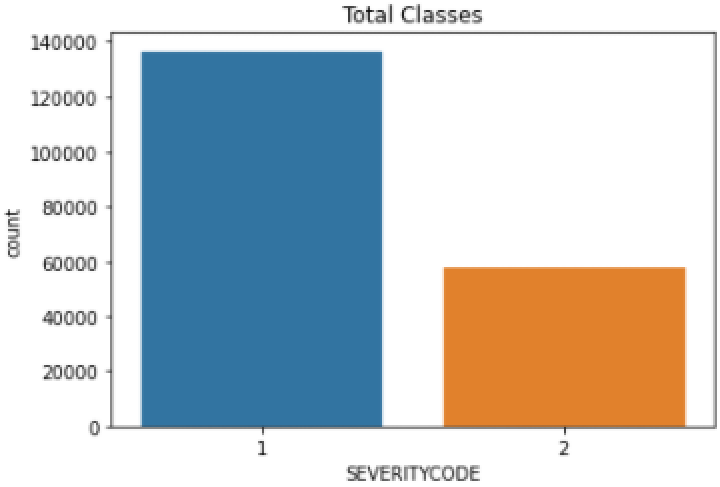
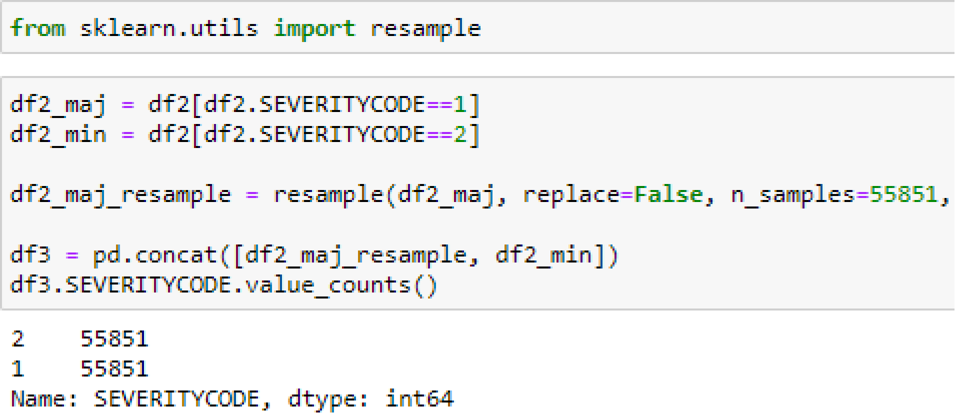


Fig 3.14 Before Resampling



```
from sklearn.utils import resample
```

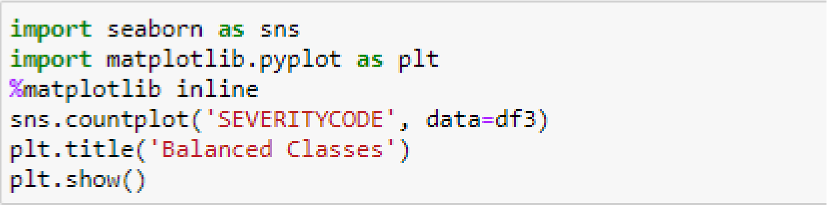
```
df2_maj = df2[df2.SEVERITYCODE==1]
df2_min = df2[df2.SEVERITYCODE==2]

df2_maj_resample = resample(df2_maj, replace=False, n_samples=55851,

df3 = pd.concat([df2_maj_resample, df2_min])
df3.SEVERITYCODE.value_counts()
```

```
2    55851
1    55851
Name: SEVERITYCODE, dtype: int64
```

Fig 3.15 Resampling



```
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
sns.countplot('SEVERITYCODE', data=df3)
plt.title('Balanced Classes')
plt.show()
```

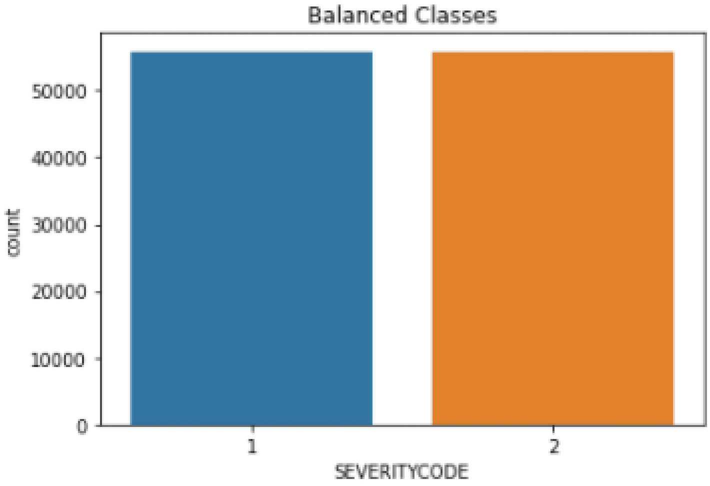


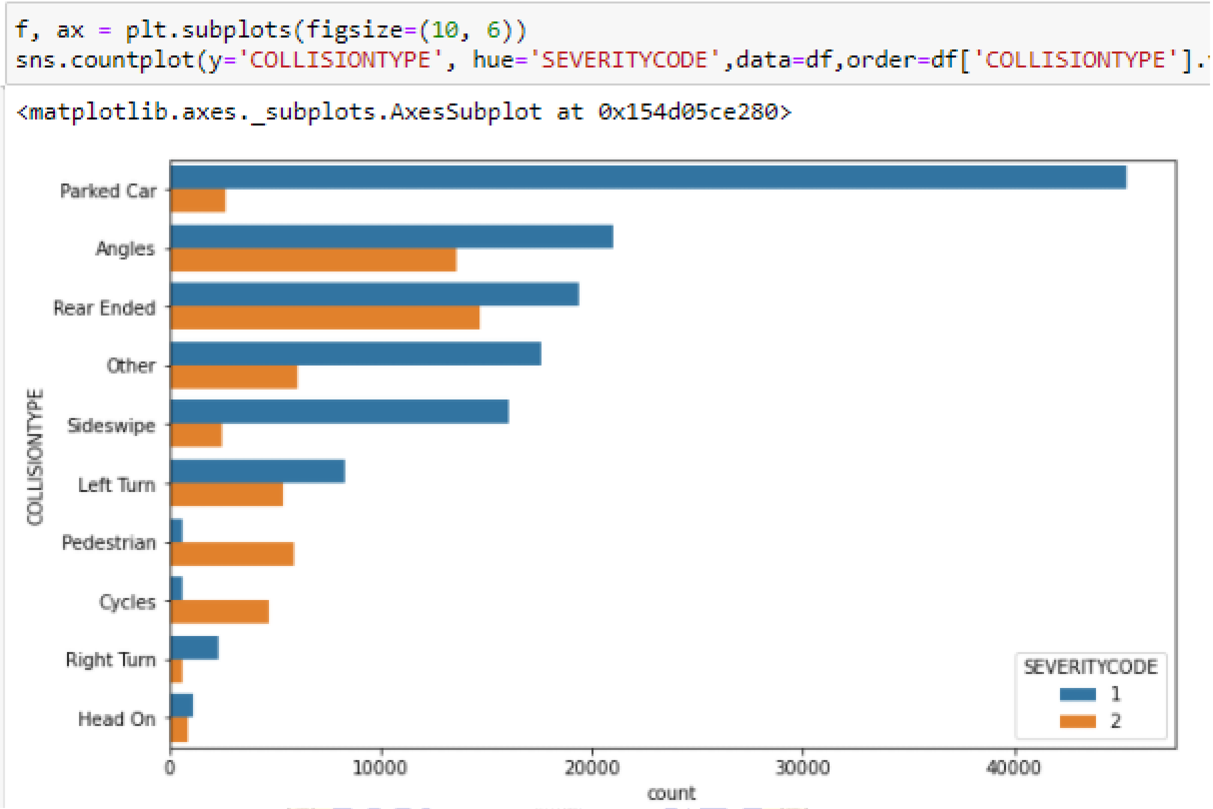
Fig 3.16 After Resampling

3.5 Data Visualization and Analysis

Data visualization is the representation of data or information in a graph, chart, or other visual format. It communicates relationships of the data with images. This is important because it allows trends and patterns to be more easily seen. With the rise of big data upon us, we need to be able to interpret increasingly larger batches of data. Machine learning makes it easier to conduct analyses such as predictive analysis, which can then serve as helpful visualizations to present.

3.5.1 Relation between Severity and Collision

Below histogram shows most injury collision happens in Angles, Rear end and Left Turn. There is less risk of property damage by Pedestrian, Cycles and Head on collision. However, Pedestrian and Cycles also causes injury collision. It is also being notice that Left turn is risker than Right turn. Parked cars cause the most property damage.

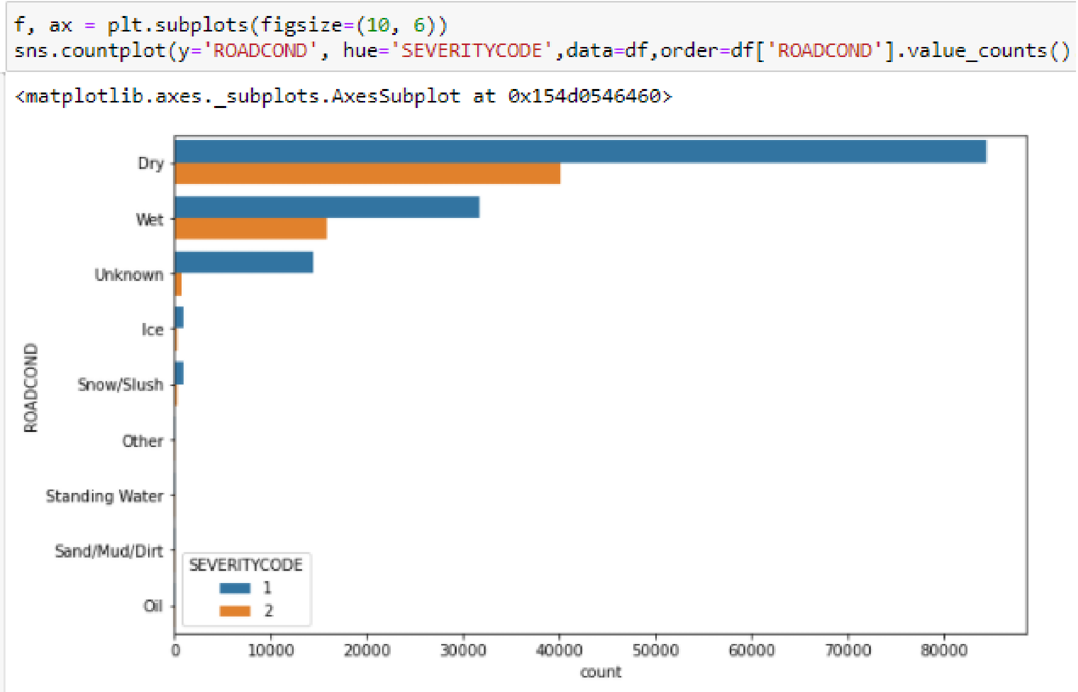


```
f, ax = plt.subplots(figsize=(10, 6))
sns.countplot(y='COLLISIONTYPE', hue='SEVERITYCODE',data=df,order=df['COLLISIONTYPE'].
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x154d05ce280>
```

Fig 3.17 Relation between Collison and Severity

3.5.2 Road Conditions



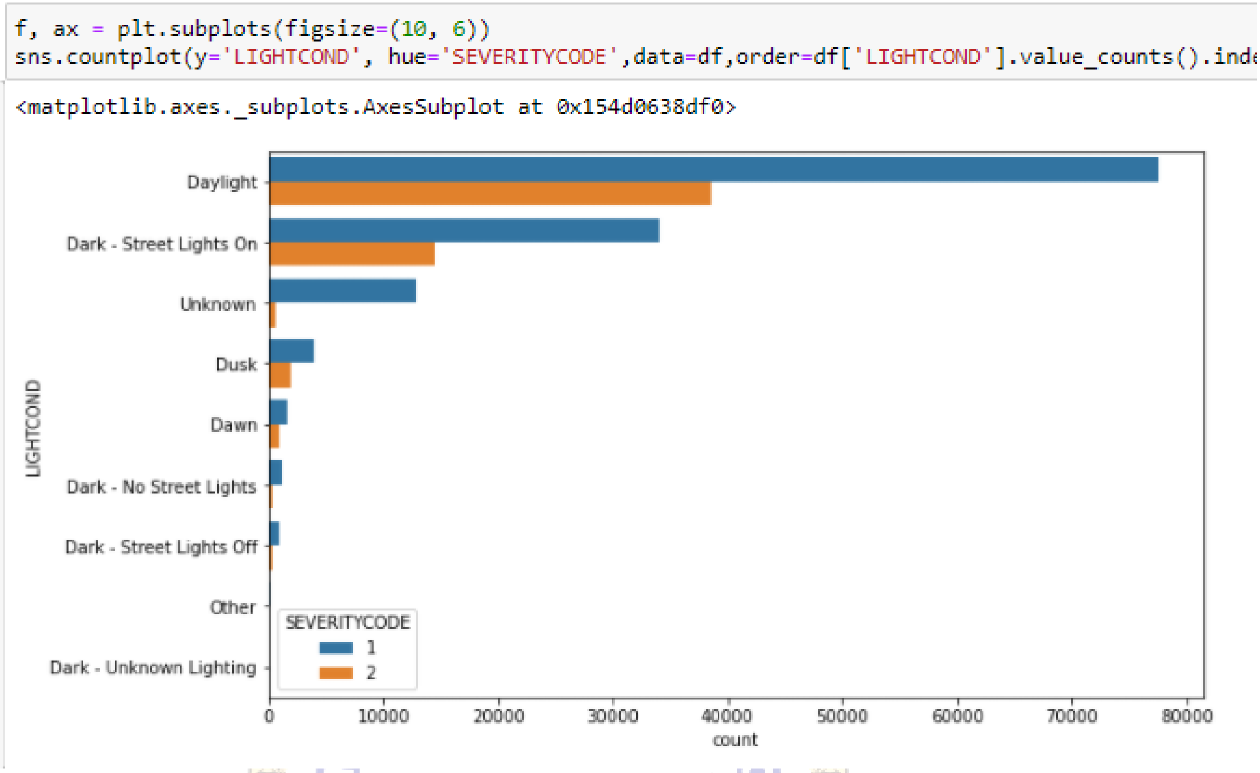
```
f, ax = plt.subplots(figsize=(10, 6))
sns.countplot(y='ROADCOND', hue='SEVERITYCODE',data=df,order=df['ROADCOND'].value_counts()
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x154d0546460>
```

Fig 3.18 Road Conditions

Property damage and Injury collision both are higher in Dry Roads than the Wet Roads.

3.5.3 Light Conditions



```
f, ax = plt.subplots(figsize=(10, 6))
sns.countplot(y='LIGHTCOND', hue='SEVERITYCODE',data=df,order=df['LIGHTCOND'].value_counts().inde
```

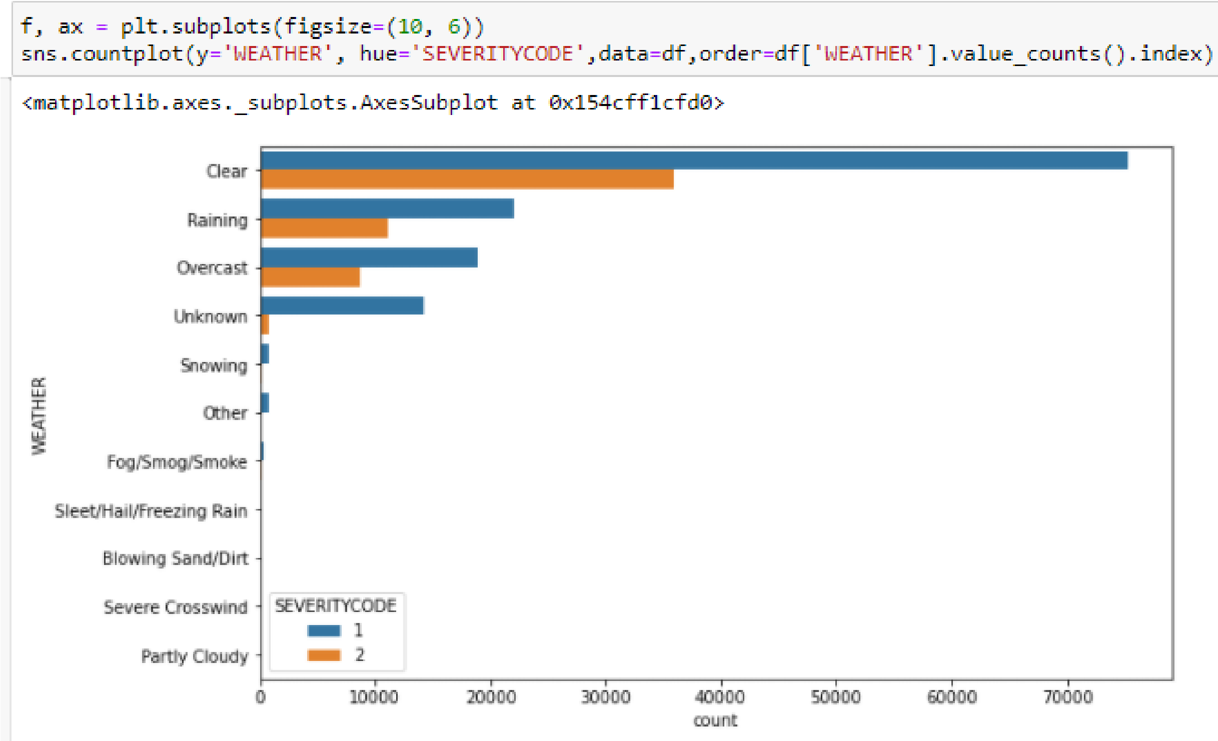
```
<matplotlib.axes._subplots.AxesSubplot at 0x154d0638df0>
```

Fig 3.19 Light Conditions

After seeing this graph it shows that more accidents occur during day Light than during night.

3.5.4 Weather

Similar to Road condition, accidents are more frequent during clear weather than during raining and overcast.



```
f, ax = plt.subplots(figsize=(10, 6))
sns.countplot(y='WEATHER', hue='SEVERITYCODE',data=df,order=df['WEATHER'].value_counts().index)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x154cff1cfd0>
```

Fig 3.20 Weather

3.6 Modeling and Evaluation

Classification algorithm – KNN, Decision Tree, Logistic Regression, and SVM models are used for predictive modeling. Target variable severity is a categorical variable with a discrete value ‘0’ property damage and ‘1’ injury collision. As explained in the feature selection section, WEATHER, ROAD CONDITION and LIGHT CONDITION features are chosen features for modeling.

Data Standardization is performed before predictive modeling using standard scalar preprocessing technique.

Next step is to perform Out of Sample Accuracy using Train/Test split approach. Out of Sample Accuracy is the percentage of correct predictions that the model makes on the data set that the model has NOT been trained on. Train/Test Split involves splitting the dataset into training and testing sets respectively, which are mutually exclusive. After which, data is trained with the training set and tested with the testing set.

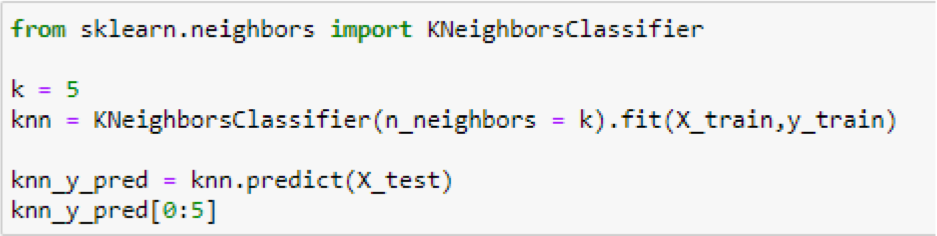
Model Accuracy will provide Jaccard score, Logloss and f1-score results.

- **F1-Score:** f1-score is a measure of accuracy of the model, which is the harmonic mean of the model’s precision and recall. Perfect precision and recall is shown by the f1-score as 1, which is the highest value for the f1-score, whereas the lowest possible value is 0 which means that either precision or recall is 0.
- **Jaccard Score:** The Jaccard index, or Jaccard similarity coefficient, defined as the size of the intersection divided by the size of the union of two label sets, is used to compare set of predicted labels for a sample to the corresponding set of labels in y_true.
- **Logloss:** Logloss, aka logistic loss or cross-entropy loss. This is the **loss** function used in (multinomial) logistic regression and extensions of it such as neural networks, defined as the negative log-likelihood of a logistic model that returns y_pred probabilities for its training data y_true.

3.6.1 KNN

K-Nearest Neighbors is an algorithm for supervised learning. Where the data is 'trained' with data points corresponding to their classification. Once a point is to be predicted, it takes into account the 'K' nearest points to it to determine it's classification.

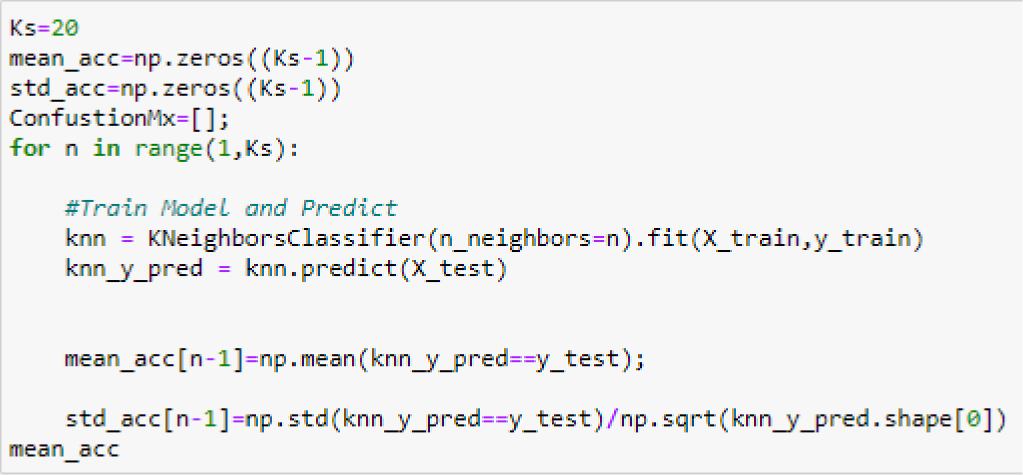
Below graph shows the best accuracy is when K is equal to 18. So we will take K = 18.



```
from sklearn.neighbors import KNeighborsClassifier

k = 5
knn = KNeighborsClassifier(n_neighbors = k).fit(X_train,y_train)

knn_y_pred = knn.predict(X_test)
knn_y_pred[0:5]
```



```
Ks=20
mean_acc=np.zeros((Ks-1))
std_acc=np.zeros((Ks-1))
ConfustionMx=[];
for n in range(1,Ks):

    #Train Model and Predict
    knn = KNeighborsClassifier(n_neighbors=n).fit(X_train,y_train)
    knn_y_pred = knn.predict(X_test)


    mean_acc[n-1]=np.mean(knn_y_pred==y_test);

    std_acc[n-1]=np.std(knn_y_pred==y_test)/np.sqrt(knn_y_pred.shape[0])
mean_acc
```

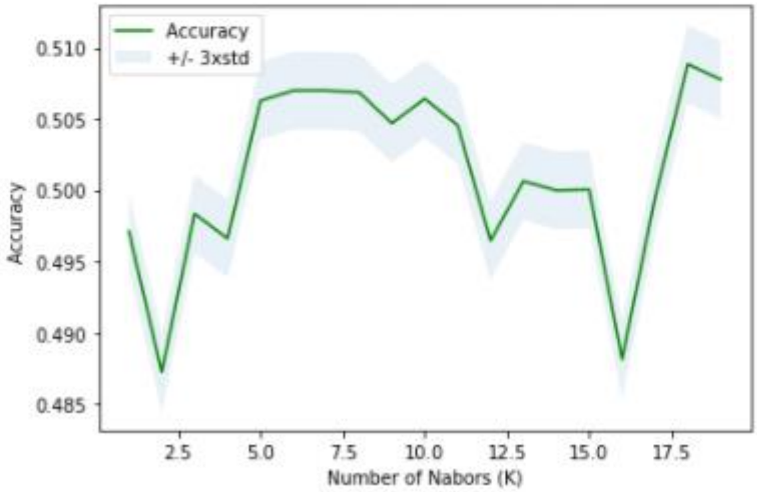
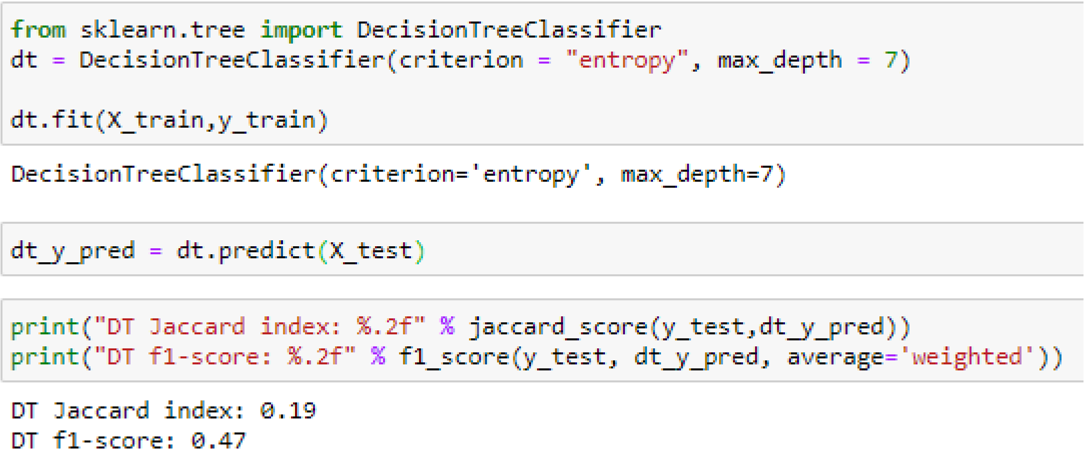


Fig 3.21 KNN

3.6.2 Decision Tree

Decision trees are built by splitting the training set into distinct nodes, where one node contains all or most of one category of the data. Instance of the decision tree classifier is created. Entropy Criterion is selected inside the classifier to see the information gain.



```
from sklearn.tree import DecisionTreeClassifier
dt = DecisionTreeClassifier(criterion = "entropy", max_depth = 7)

dt.fit(X_train,y_train)
```

```
DecisionTreeClassifier(criterion='entropy', max_depth=7)
```

```
dt_y_pred = dt.predict(X_test)
```

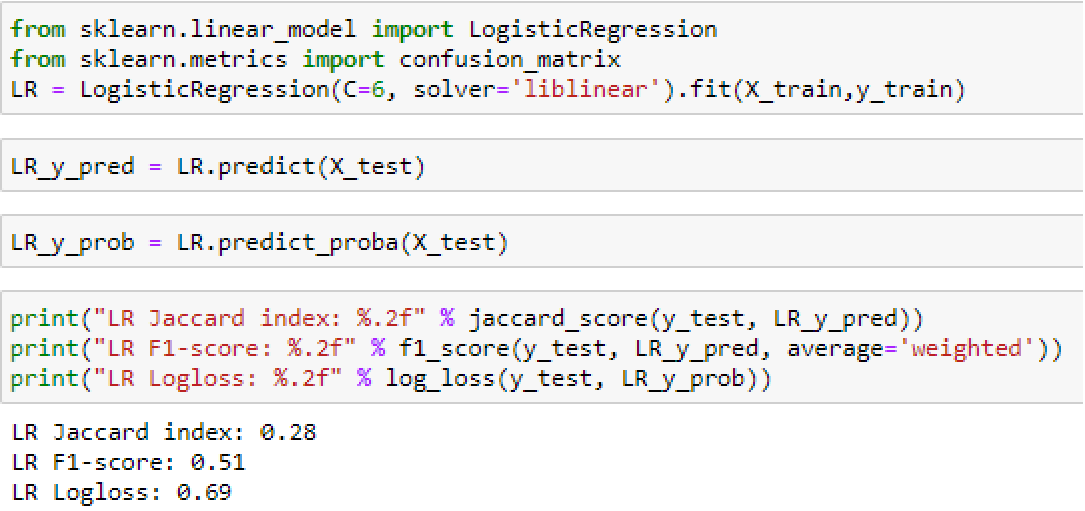
```
print("DT Jaccard index: %.2f" % jaccard_score(y_test,dt_y_pred))
print("DT f1-score: %.2f" % f1_score(y_test, dt_y_pred, average='weighted'))
```

```
DT Jaccard index: 0.19
DT f1-score: 0.47
```

Fig 3.22 Decision Tree

3.6.3 Logistic Regression

Logistic Regression is a variation of Linear Regression, useful when the observed dependent variable, **y**, is categorical. It produces a formula that predicts the probability of the class label as a function of the independent variables. We have used the inverse of regularization strength in C = 0.1 for model prediction.



```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import confusion_matrix
LR = LogisticRegression(C=6, solver='liblinear').fit(X_train,y_train)
```

```
LR_y_pred = LR.predict(X_test)
```

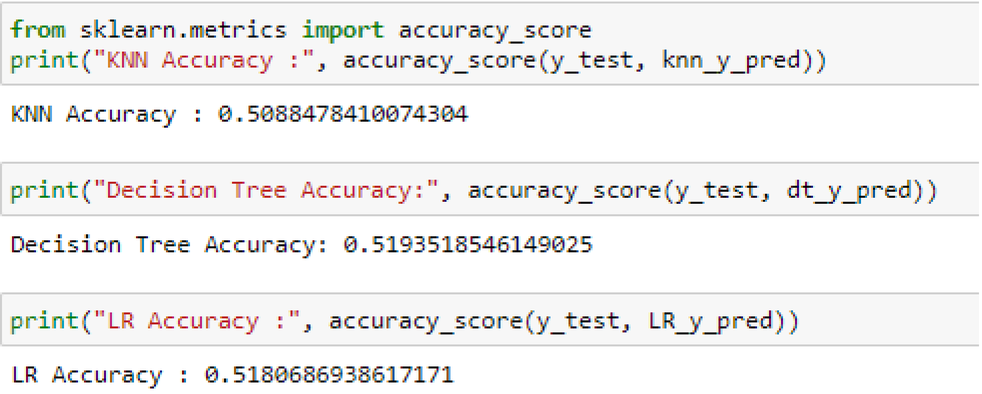
```
LR_y_prob = LR.predict_proba(X_test)
```

```
print("LR Jaccard index: %.2f" % jaccard_score(y_test, LR_y_pred))
print("LR F1-score: %.2f" % f1_score(y_test, LR_y_pred, average='weighted'))
print("LR Logloss: %.2f" % log_loss(y_test, LR_y_prob))
```

```
LR Jaccard index: 0.28
LR F1-score: 0.51
LR Logloss: 0.69
```

Fig 3.23 Logistic Regression

3.7 Accuracy



```
from sklearn.metrics import accuracy_score
print("KNN Accuracy :", accuracy_score(y_test, knn_y_pred))
```

```
KNN Accuracy : 0.5088478410074304
```

```
print("Decision Tree Accuracy:", accuracy_score(y_test, dt_y_pred))
```

```
Decision Tree Accuracy: 0.5193518546149025
```

```
print("LR Accuracy :", accuracy_score(y_test, LR_y_pred))
```

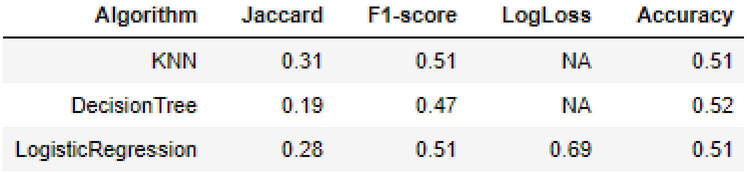
```
LR Accuracy : 0.5180686938617171
```

Fig 3.24 Accuracy

The final accuracy is calculated by taking in consideration these three scores:

- Jaccard score
- F1 score
- Log loss

RESULTS



Algorithm	Jaccard	F1-score	LogLoss	Accuracy
KNN	0.31	0.51	NA	0.51
DecisionTree	0.19	0.47	NA	0.52
LogisticRegression	0.28	0.51	0.69	0.51

Fig 4.1 Results

The model has a ~52% in sample accuracy and a ~51% out of sample accuracy.
In this study I have analysed the relationship between the various aspects like weather conditions, road conditions, light conditions etc. when a road accident occurs and how severe the accident will be. I built a classification model that takes in all these factors affecting the severity of traffic due to an accident and predicts how severe the traffic will be.

A machine learning classification is successfully trained using ~ 20 thousand datapoints that define the various attributes of a scenario when a road vehicle accident occurs which has been used beneficially to produce significant results.

CONCLUSION

In order to build road accidents possibilities model, we have first cleaned the data provided by Seattle traffic management system. Data is also balanced on severity level using resampling technique. Feature selection using mutual information classifier is performed which resulted in selecting four features Weather, Road condition, Light condition and severity code. We have used KNN, Decision Tree, and Logistic Regression algorithm for predictive modeling. Accuracy of all the models are around 51 percent which won't be enough for prediction. More data is needed with severity equal to injury collision in order to find contributing features involved. More data addition to Road collision dataset will result in greater accuracy to warn traffic about road accidents.

We also found about relation between accident severity and other features. Below points should be considered by traffic management before warning traffic.

- Angles, Rear end and Left turn causes more injury collisions.
- Parked cars are more prone towards property damage.
- Left turn is riskier than right turn.
- Injury occurred more for pedestrian and cyclist than the property damage.
- Most of the road accidents happens in clear weather. However, caution should be taken during bad weather condition.
- Most of the road accidents happens during day light.

REFRENCES

[1]Eric Matthes,”Python Crash Course”,No Starch Press, 2016

[2]Paul Barry,”Head-First Python”,2nd Edition,O’Reilly, 2016

[3]Al Sweigart,”Invent Your Own Computer Games with Python”,4th Edition,No Starch, 2017

[4]Allen B. Downey,”Think Python: How to Think Like a Computer Scientist”,2nd Edition,O’Reilly, 2015

[5]Zed A. Shaw,”Learn Python 3 the Hard Way”,Addison-Wesley, 2016 [6]Dan Bader,”Python Tricks: A Buffet of Awesome Python Features”,dbader.org, 2017

[7]Brett Slatkin,”Effective Python: 59 Ways to Write Better Python”,Addison-Wesley, 2015

[8]Mark Lutz,”Python Pocket Reference 5ed: Python in Your Pocket”,5th Edition,O′Reilly,2014

[9]John Zelle,”Python Programming: An Introduction to Computer Science”,3rd Edition,Ingram short title,2016

[10]R. Nageswara Rao ,”Core Python Programming”,2nd Edition,Dreamtech Press,2018

[11]Andrew Stetsenko,”What do companies expect from Python devs in 2019?”,Article

[12]Serdar Yegulalp,”Python 2 EOL: How to survive the end of Python 2”,Article 33

[13]David Bolton,”Asynchronous Programming in Python: A Walkthrough”,Article

[14]

Martin Chikilian.”Buggy Python Code: The 10 Most Common Mistakes That Python Developers Make”,Article 27

[15]Nick Heath,”Python programming language gets speed boost from latest PyPy interpreter”,Article

[16]Emily Chang and Nils Bunge,”How to collect, customize, and centralize Python logs”,Article

[17]Moshe Zadka,”Format Python However You Like With Black”[18]J Brownlee,”Basic Concepts in Machine Learning”[19]J Brownlee,”What is Deep Learning? Retrieved” [20]S Chen,”A Basic Machine Learning Workflow In Production”