**INTRODUCTION**

**CHAPTER 1**

**INTRODUCTION**

In recent years, the road accident has become a global problem and marked as the ninth prominent cause of death in the world. Due to the enormous number of road accidents every year, it has become a major problem in Bangladesh. It is entirely inadmissible and saddening to allow its citizen to kill by road accidents. Consequently, to handle this overwhelmed situation, a precise analysis is required. This research paper has been done to analyze traffic accidents more deeply to determine the intensity of accidents by using machine learning approaches

In this paper we are using an approach known as Data Mining which can be used in various areas the cause of accident like Time of Accident,Position,Rural/Urban, Location, Nature of Accident Position, Nature of Accident, Vehicle Involved, Type of Accident, Causes, Road Feature,Road condition, Intersection Type And Control, Weather Condition , Vehicle Responsible, Classification of Accident, No of Affected Persons. We predict by using different Machine learning Algorithms like logistic, Navy Bayes, Random Forest, and KNN Algorithm

**1.1 Objectives**

The main objective of the road accident prediction system:

* Analyze the previously occurred accidents in the locality which will help us to determine the most accident-prone area and help us to set up the immediate required help for them.
* To make predictions based on constraints like weather, pollution, road structure, etc.

**1.1.1 Problem Statement**

There are several problems with current practices for prevention of the accidents occurred in the localities. The database we will use is available officially by many institutes and government websites. The data collected will be analysed, integrated and grouped together based on different constraints using the best suited algorithm. This estimation will be helpful to analyse and identify the flaw and the reasons of the accidents. It will also be helpful while making roads and bridges as a reference to avoid the same problems faced before. The predictions made will be very much useful to plan the management of such problems.

**1.2 Project Undertaken**

**1.2.1 Problem Definition**

In recent years, the road accident has become a global problem and marked as the ninth prominent cause of death in the world. Due to the enormous number of road accidents every year, it has become a major problem in Bangladesh. It is entirely inadmissible and saddening to allow its citizen to kill by road accidents. Consequently, to handle this overwhelmed situation, a precise analysis is required. This research paper has been done to analyze traffic accidents more deeply to determine the intensity of accidents by using machine learning approaches .With increase in urbanization and socio-economical growth, the number of vehicles in major metropolitan cities is increasing day by day. Therefore, traffic congestion is becoming a major concern of metropolitan cities all over the world. This results in tremendous air pollution, loss of valuable time and money of citizens. Hence, traffic congestion monitoring of different road segments is very essential for analyzing the problem associated with smooth mobility. Identifying the problematic road segments within the city is one of the important job for the transport authority to assess the road condition. That will assist the government agencies or policy makers to optimize traffic rules and regulations. This work identifies traffic congestion pattern which can classify the different road segments based on traffic density and average speed of vehicles.

**1.2.2 Scope Statement**

Comparative study of different Machine Learning algorithms will be done. The most efficient algorithm of them will be found and used. Create a Machine Learning program trained with a different data set

**1.3 Organization of Project Report**

The overall report revolves around the objective of prediction of heart disease based on user’s data.

First chapter deals with introduction of Prediction offake user using Machine Learning and its need. In that we have included background and basics, literature survey and Project Undertaken.

Second chapter deals with project planning and managements. In that we include details of System Requirements Specification (SRS), Project Process Modeling, Cost and Effort Estimates and finally Project scheduling.

Third chapter deals with Analysis and Design. In this we include Idea Matrix, Mathematical Model, Feasibility Analysis and UML Diagrams.

Fourth chapter deals with testing section.

**LITERATURE SURVEY**

**CHAPTER 2**

**LITERATURE SURVEY**

**Many kinds of research studies have been conducted to predict results for CKD related problems using various data mining techniques.**

Analyzing road accident data using machine learning paradigms

Priyanka A. Nandurge, Nagaraj V. DharwadkarPublished in International Conference on I… 2017

DOI:10.1109/i-smac.2017.8058251

To determine the main factors associated with road traffic accidents is one of main objectives of accident data analysis. Due heterogeneity nature of road accident data makes analysis tricky. To overcome heterogeneity of data partitioning is used. The proposed method uses k-means clustering method as the main task of segmentation of road accident data. Further, association rule mining is applied to discover the situations related with the occurrence of the whole data set and the occurrence of clusters recognized by the k-means clustering

Road Accident Analysis and Prediction of Accident Severity by Using Machine Learning in Bangladesh

Md. Farhan Labib, Ahmed Sady Rifat, +2 authors Faria Nawrine Published in 7th International Conference… 2019

DOI:10.1109/ICSCC.2019.8843640

In recent years, the road accident has become a global problem and marked as the ninth prominent cause of death in the world. Due to the enormous number of road accidents every year, it has become a major problem in Bangladesh. It is entirely inadmissible and saddening to allow its citizen to kill by road accidents. Consequently, to handle this overwhelmed situation, a precise analysis is required. This research paper has been done to analyze traffic accidents more deeply to determine the intensity of accidents by using machine learning approaches in Bangladesh. We also figure out those significant factors that have a clear effect on road accidents and provide some beneficent suggestions regarding this issue. Analysis has been done, by using Decision Tree, K-Nearest Neighbors (KNN), Naïve Bayes and AdaBoost these four supervised learning techniques, to classify the severity of accidents into Fatal, Grievous, Simple Injury and Motor Collision these four categories. Finally, the best performance is achieved by AdaBoost.

Analysis of road accidents in India using data mining classification algorithms

E. Suganya, S. Vijayarani Published in International Conference on… 2017

DOI:10.1109/icici.2017.8365315

Abstract-Classification is a model finding process which is used for segmenting the data into different classes based on some constraints. This work analyzes the road accidents in India data set using classification algorithms namely linear regression, logistic regression, decision tree, SVM, Naïve Bayes, KNN, Random Forest and gradient boosting algorithm. Performance measures used are accuracy, error rate and execution time. This analysis is done in R data mining tool

**2.1 Project Planning and Management**

**2.1 .1 Detail System Requirement Specification (SRS)**

This chapter covers the project planning and management details. It also covers System Requirement specifications. SRS is considered as the base for the effort estimations and project scheduling.

**2.1.2 System Overview**

It is a web-based machine learning application which is trained by a UCI dataset. The user inputs it’s specific medical details to get the prediction of heart disease for that user. The algorithm will calculate the probability of presence of fake user. The result will be displayed on the webpage itself. Thus minimizing the cost and time required to predict the fake user.

**Hardware Requirements**

The web applications has no extra hardware requirements**.**

**Software Requirements**

* Operating System : Windows 7/10, Ubuntu 17.04
* Anaconda Python
* Matplotlib
* Seaborn
* Pandas

**2.2 Project Process Modeling**

**2.2.1 Incremental Model**

The incremental model is a method of software development where the product is designed, implemented and tested incrementally until the product is finished. It involves both development and maintenance. This process goes on until it satisfies all of its requirements of the user/client. This model is the combination of the elements of the waterfall model with the iterative philosophy of prototyping. The product is decomposed into a number of components, each of which is designed and built separately. Each component is delivered to the client/user when it is complete. This allows partial utilization of the product and avoids a long development time. It also avoids a large initial capital outlay and subsequent long waiting period.

**Characteristics of Incremental Model**

* System is broken down into many small development projects.
* Partial systems are built to produce the final system.
* Highest priority requirements are tackled first.
* The requirement of a portion is frozen once the incremented portion is developed.

**Advantages**

* After each iteration, regression testing should be conducted. During this testing, faulty elements of the software can be quickly identified because few changes are made within any single iteration.
* It is generally easier to test and debug than other methods of software development because relatively smaller changes are made during each iteration. This allows for more targeted and rigorous testing of each element within the overall product.
* Customer can respond to features and review the product for any needed or usefulchanges.
* Initial product delivery is faster and also lowers the initial delivery cost.

**2.3 Cost Effective Estimates**

Cost of the project will be the cost of hardware (mainly CPU) plus the cost of work that isput in.

**2.3.1 Basic COCOMO**

* **Project class**

However it determines that the project fits in the characteristics of organic Mode as project is college level and requirements are rigid and less than rigid.

* **Number of code Lines**

The project will have approximately 1000 Delivered Source instructions. So, the Basic COCOMO model equations are as follows:

Effort Applied (E) = ab (KLOC) bb [man months]

Development Time (D) =cb (E) db [months]

People Required (P) = E/D [count] Where: KLOC (Kilo lines of code) is the estimated

number of delivered lines (in thousands) of code for project. E is the effort applied per

person per month. D is the development time in consecutive months. The coefficients

ab; bb; cb; db

are predetermined according to Semi-detached mode project class.

For Semi-detached mode:

ab = 1.0

bb = 1:12

cb = 2:5

db = 0:35

* **Calculations**

So, this project comes under the organic mode.

Effort Applied (E) = 3.0\* (1) (1.12) [man-months]

E = 3.36

Development Time (D) = 2.5\* (3.36) (0.35) [months]

D = 2.94

People Required (P) = 22.31/2.94 [count]

P = 7.5

**THEORETICAL BACKGROUND**

**CHAPTER 3**

**THEORETICAL BACKGROUND**

**3.1 Overview of Machine Learning**

Machine learning is a subfield of computer science and a rapidly up surging topic in today’s context and is expected to boom more in coming days. Our world is flooded with data and data is being created rapidly every day all around the world. According to Big Data and Analytics Solutions company CSC, it is expected by 2020, that the data amount will be 44 times bigger than in 2009 [2]. Therefore, it is necessary to understand data and gain insights for better understanding of a human world. The data amount is so huge today that traditional methods cannot be used. Analysing data or building predictive models manually is almost impossible in some scenarios and also time consuming and less productive. Machine learning, on other hand, produces reliable, repeatable results and learns from earlier computation.

Data used for machine learning are basically of two types labelled data and unlabelled data. Labelled data is the data where attributes are provided. It has some sort of tag or meaning attached to the data therefore used in supervised learning. Labelled attribute can be numerical or categorical. Numerical data are used in regression to predict the value while categorical data are used in classification. Unlabelled data is the data where there are only data points and no labelling to assist. Unlabelled data are used in unsupervised learning so that machine can identify the patterns or any structure present in the data set. [3, 3]

The labelled data and unlabelled data are used with supervised learning and unsupervised learning respectively. Supervised learning entails a learning map between a set of input variables X and an output variable Y and applying this mapping to predict the output for unseen data [3]. After learning the dataset, algorithms generalise the data and formulates the hypothetical value H for the given dataset.

Supervised learning is further categorised into two types: Regression and Classification. According to the business dictionary, a regression is a technique for determining the statistical relationship between two or more variables where a change in dependent variable is associated with, and depends on a change in one or more independent variables [4]. Classification is a task that occurs very frequently in everyday life. Essentially it involves dividing up objects so that each is assigned to one of a number of mutually exhaustive and exclusive categories known as classes. The term ‘mutually exhaustive and exclusive’ simply means that each object must be assigned to precisely one class, i.e. never to more than one and never to no class at all [5,5].

Unsupervised learning studies how systems can learn to represent particular input patterns in a way that reflects the statistical structure of the overall collection of input patterns. By contrast, with supervised learning there are no explicit target outputs or environmental evaluations associated with each input; rather the unsupervised learner brings prior biases as to what aspects of the structure of the input should be captured in the output. [6, 858]

**3.2 Dimension Reduction Techniques**

Dimensionality reduction is simply the process of decreasing the number of input random variables without the loss of information. A greater number of input variables or dimensions and large data samples increase the complexity of the dataset. To reduce the memory and computational time dimensionality of data is reduced. Dimensionality reduction also helps to eliminate unnecessary input variables like duplicate variables or variables with a very low significance level. [7, 109-110]

There are two types of dimensionality reduction techniques: Feature Selection and Feature Extraction are described below.

**3.2.1 Feature Selection**

In feature selection, k dimensions are selected out of d dimensions that gives most information and discard the (d-k) dimensions. In other words, feature selection is also called subset selection. The best subset contains the least number of dimensions that contribute most to the accuracy. The best subset is found with suitable error function.

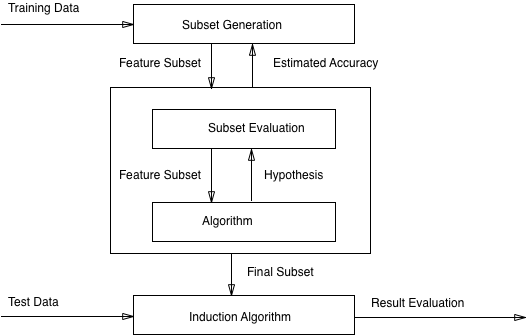


Figure 3.1: Feature Selection. Adapted from Md Rahat Hossian

In Figure 1, Training data is put through a certain process of subset generation, for example sequential backward selection. The resulting subset is now put through the algorithm to test its performance if the performance meets the expected criteria, then it will be selected as the final subset. Otherwise, the resulting subset will again be put through the process of subset generation for more fine-tuning.

There are two different approaches to feature selection: Sequential Forward Selection and Sequential Backward Selection which are explained below.

**Sequential Forward Selection:**

Sequential Forward Selection begins with a model containing no predictors, and then predictors are added to the model, one at a time, until all of the predictors are in the model. In particular, at each step the variable that gives the greatest additional improvement to the fit is added to the model [9,207].

Let us denote a set by P, with variables Xi, i = 1,……,d. E(P) is the error incurred in the test sample. Sequential Forward Selection starts with empty set with no variables P = { φ }. At each step, a single variable is added to the empty set and a model is trained, and error E(P ∪ Xi) is calculated on test set. Error criteria is set as per requirement, for example, the least square error and misclassification error. From all the errors, the input variable causing the least error Xj, is selected and added to the empty set P. The model is trained again with remaining number of variables and the process continues to add variables to P, if E(P ∪ Xi) is less than E(P).

**Sequential Backward Selection:**

Sequential Backward Selection is an efficient alternative for best subset solution but it begins with full set of features unlike Sequential Forward Selection. It removes the least significant features one at a time iteratively. [9,207.]

Sequential Backward Selection starts with a full set of variables P = {1,2,3,…..,d}. At each step, the model is trained with a full set of variables and the error is calculated in test set, and the variable with highest error Xj, is removed from the set P. The model is trained again with a new set of variables P, and the process continues to remove variables from P, if E(P-Xj) is less than E(P).

**3.2.2 Feature Extraction**

In the feature extraction technique, features or independent variables from the data set are transformed into new independent variables known as new feature space. Newly constructed feature space explains the data most and only significant data are selected.

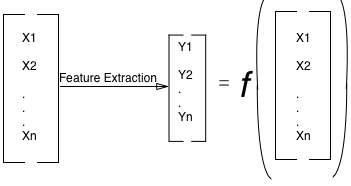
Let, there are n features of X1…….Xn. After feature extraction there are m attributes where (n > m) and this feature extraction is done with some mapping function, F.

Figure 3.2:Feature Extractipon

Referring to Figure 2, Xn set of independent features or dimensions are reduced to Yn set of independent features. In the feature extraction process, a technique such as principal component analysis is used. It will take only non-redundant and significant features from Xn and change into new feature space Yn [11]. With the feature extraction, capability of interpretation is lost since, Yn features obtained after feature extraction is not same as Xn meaning it is not a direct subset of Xn.

**3.3 Machine Learning**

Machine learning is a growing technology which enables computers to learn automatically from past data. Machine learning uses various algorithms for **building mathematical models and making predictions using historical data or information**. Currently, it is being used for various tasks such as **image recognition**, **speech recognition**, **email filtering**, **Facebook auto-tagging**, **recommender system**, and many more.

This machine learning gives you an introduction to machine learning along with the wide range of machine learning techniques such as **Supervised**, **Unsupervised**, and **Reinforcement** learning. You will learn about regression and classification models, clustering methods, hidden Markov models, and various sequential models.

**3.3.1 What is Machine Learning:**

In the real world, we are surrounded by humans who can learn everything from their experiences with their learning capability, and we have computers or machines which work on our instructions. But can a machine also learn from experiences or past data like a human does? So here comes the role of **Machine Learning**.

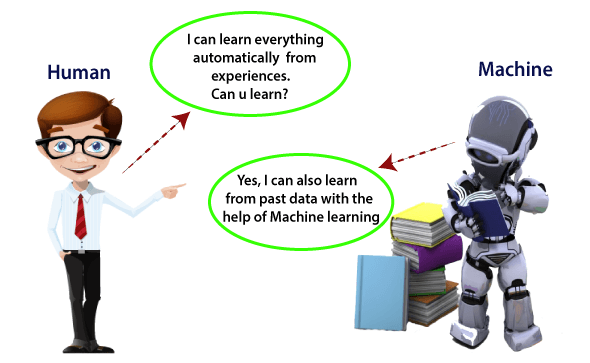
****

Figure 3.3: **Machine Learning**

Machine Learning is said as a subset of **artificial intelligence** that is mainly concerned with the development of algorithms which allow a computer to learn from the data and past experiences on their own. The term machine learning was first introduced by **Arthur Samuel** in **1959**. We can define it in a summarized way as:

“Machine learning enables a machine to automatically learn from data, improve performance from experiences, and predict things without being explicitly programmed.”

With the help of sample historical data, which is known as **training data**, machine learning algorithms build a **mathematical model** that helps in making predictions or decisions without being explicitly programmed. Machine learning brings computer science and statistics together for creating predictive models. Machine learning constructs or uses the algorithms that learn from historical data. The more we will provide the information, the higher will be the performance.

**A machine has the ability to learn if it can improve its performance by gaining more data.**

## 3.3.2 How does Machine Learning work

A Machine Learning system **learns from historical data, builds the prediction models, and whenever it receives new data, predicts the output for it**. The accuracy of predicted output depends upon the amount of data, as the huge amount of data helps to build a better model which predicts the output more accurately.



## Figure 3.3.2: Machine Learning working

## Features of Machine Learning:

* Machine learning uses data to detect various patterns in a given dataset.
* It can learn from past data and improve automatically.
* It is a data-driven technology.
* Machine learning is much similar to data mining as it also deals with the huge amount of the data.

## Need for Machine Learning

## The need for machine learning is increasing day by day. The reason behind the need for machine learning is that it is capable of doing tasks that are too complex for a person to implement directly. As a human, so for this, we need some computer systems and here comes the machine learning to make things easy for us.

## The importance of machine learning can be easily understood by its uses cases, Currently, machine learning is used in **self-driving cars**, **cyber fraud detection**, **face recognition**, and **friend suggestion by Facebook**, etc.

## ****Following are some key points which show the importance of Machine Learning:****

* Rapid increment in the production of data
* Solving complex problems, which are difficult for a human
* Decision making in various sector including finance
* Finding hidden patterns and extracting useful information from data.

**3.4 Classification of Machine Learning**

At a broad level, machine learning can be classified into three types:

1. **Supervised learning**
2. **Unsupervised learning**
3. **Reinforcement learning**

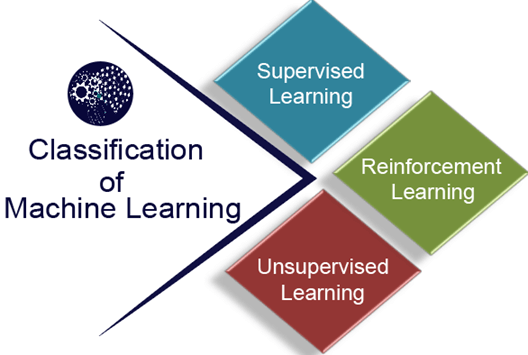


Figure 3.4: Classification of Machine Learning

### 3.4.1 Supervised Learning

Supervised learning is a type of machine learning method in which we provide sample labeled data to the machine learning system in order to train it, and on that basis, it predicts the output.

The system creates a model using labeled data to understand the datasets and learn about each data, once the training and processing are done then we test the model by providing a sample data to check whether it is predicting the exact output or not.

The goal of supervised learning is to map input data with the output data. The supervised learning is based on supervision, and it is the same as when a student learns things in the supervision of the teacher. The example of supervised learning is **spam filtering**.

Supervised learning can be grouped further in two categories of algorithms:

* **Classification**
* **Regression**



### Figure 3.4.1: Supervised Learning

**1. Regression**

Regression algorithms are used if there is a relationship between the input variable and the output variable. It is used for the prediction of continuous variables, such as Weather forecasting, Market Trends, etc. Below are some popular Regression algorithms which come under supervised learning:

* Linear Regression
* Regression Trees
* Non-Linear Regression
* Bayesian Linear Regression
* Polynomial Regression

**2. Classification**

Classification algorithms are used when the output variable is categorical, which means there are two classes such as Yes-No, Male-Female, True-false, etc.

Spam Filtering,

* Random Forest
* Decision Trees
* Logistic Regression
* Support vector Machines

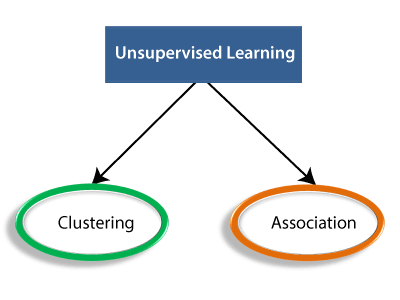
### 3.4.2 Unsupervised Learning

Unsupervised learning is a learning method in which a machine learns without any supervision.

The training is provided to the machine with the set of data that has not been labeled, classified, or categorized, and the algorithm needs to act on that data without any supervision. The goal of unsupervised learning is to restructure the input data into new features or a group of objects with similar patterns.

In unsupervised learning, we don't have a predetermined result. The machine tries to find useful insights from the huge amount of data. It can be further classifieds into two categories of algorithms:

* **Clustering**
* **Association**



### Figure 3.4.2: Unsupervised Learning

* **Clustering**: Clustering is a method of grouping the objects into clusters such that objects with most similarities remains into a group and has less or no similarities with the objects of another group. Cluster analysis finds the commonalities between the data objects and categorizes them as per the presence and absence of those commonalities.
* **Association**: An association rule is an unsupervised learning method which is used for finding the relationships between variables in the large database. It determines the set of items that occurs together in the dataset. Association rule makes marketing strategy more effective. Such as people who buy X item (suppose a bread) are also tend to purchase Y (Butter/Jam) item. A typical example of Association rule is Market Basket Analysis.

Unsupervised Learning algorithms:

Below is the list of some popular unsupervised learning algorithms:

* **K-means clustering**
* **KNN (k-nearest neighbors)**
* **Hierarchal clustering**
* **Anomaly detection**
* **Neural Networks**
* **Principle Component Analysis**
* **Independent Component Analysis**
* **Apriori algorithm**
* **Singular value decomposition**

### 3.4.3 Reinforcement Learning

Reinforcement learning is a feedback-based learning method, in which a learning agent gets a reward for each right action and gets a penalty for each wrong action. The agent learns automatically with these feedbacks and improves its performance. In reinforcement learning, the agent interacts with the environment and explores it. The goal of an agent is to get the most reward points, and hence, it improves its performance.

The robotic dog, which automatically learns the movement of his arms, is an example of Reinforcement learning.

**3.5 Machine Learning Algorithms That You Can Use In Your Own Projects As Well, Including:**

1. k-Nearest Neighbors (k-NN)
2. Naïve Bayes.
3. Logistic Regression.
4. Decision Trees.
5. Random Forests.

**3.5.1 k-Nearest Neighbors (k-NN)**

The K-nearest neighbors (KNN) algorithm is a type of supervised machine learning algorithms. [KNN](https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm) is extremely easy to implement in its most basic form, and yet performs quite complex classification tasks. It is a lazy learning algorithm since it doesn't have a specialized training phase. Rather, it uses all of the data for training while classifying a new data point or instance. KNN is a non-parametric learning algorithm, which means that it doesn't assume anything about the underlying data. This is an extremely useful feature since most of the real world data doesn't really follow any theoretical assumption e.g. linear-separability, uniform distribution, etc.

In this article, we will see how KNN can be implemented with Python's [Scikit-Learn](https://stackabuse.com/tag/scikit-learn/) library. But before that let's first explore the theory behind KNN and see what are some of the pros and cons of the algorithm.

## Theory:

The intuition behind the KNN algorithm is one of the simplest of all the supervised machine learning algorithms. It simply calculates the distance of a new data point to all other training data points. The distance can be of any type e.g Euclidean or Manhattan etc. It then selects the K-nearest data points, where K can be any integer. Finally it assigns the data point to the class to which the majority of the K data points belong.

Let's see this algorithm in action with the help of a simple example. Suppose you have a dataset with two variables, which when plotted, looks like the one in the following figure.

Your task is to classify a new data point with 'X' into "Blue" class or "Red" class. The coordinate values of the data point are x=45 and y=50. Suppose the value of K is 3. The KNN algorithm starts by calculating the distance of point X from all the points. It then finds the 3 nearest points with least distance to point X. This is shown in the figure below. The three nearest points have been encircled.

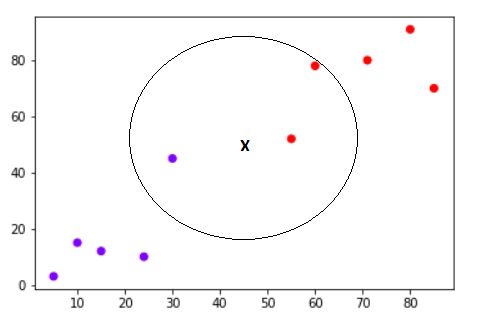


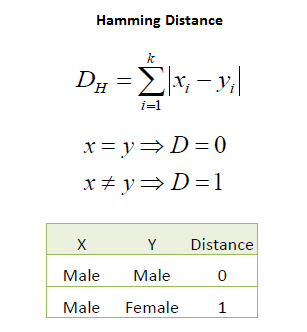
Figure 3.5: K Nearest Neighbour

The final step of the KNN algorithm is to assign new point to the class to which majority of the three nearest points belong. From the figure above we can see that the two of the three nearest points belong to the class "Red" while one belongs to the class "Blue". Therefore the new data point will be classified as "Red".

The Euclidean distance between two points can be calcu lated by following formula:



For continuous data , for categorical data use hamming distance



|  |
| --- |
| We can now use the training set to classify an unknown case (Age=48 and Loan=$142,000)  using Euclidean distance. If K=1 then the nearest neighbor is the last case in the training set with Default=Y. |
|  |
| D = Sqrt[(48-33)^2 + (142000-150000)^2] = 8000.01  >> Default=Y |

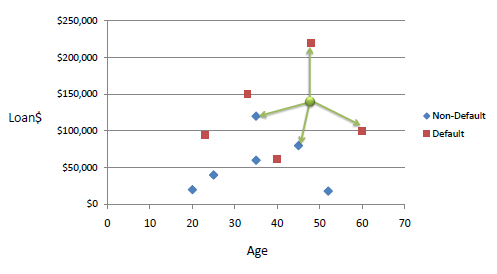


Figure 3.5.1: K-Nearest Neighbour and Dynamic Loan

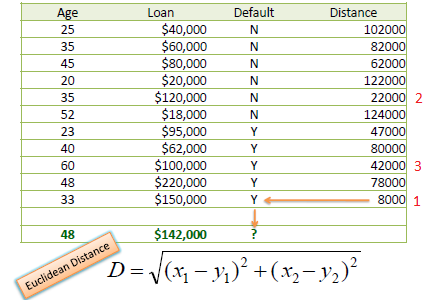


Figure 3.5.2: Data set on loan

# 3.5.2 Naïve Bayes Classifier Algorithm

* Naïve Bayes algorithm is a supervised learning algorithm, which is based on **Bayes theorem** and used for solving classification problems.
* It is mainly used in text classification that includes a high-dimensional training dataset.
* Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.
* **It is a probabilistic classifier, which means it predicts on the basis of the probability of an object**.
* Some popular examples of Naïve Bayes Algorithm are **spam filtration, Sentimental analysis, and classifying articles**.

## Why is it called Naïve Bayes?

The Naïve Bayes algorithm is comprised of two words Naïve and Bayes, Which can be described as:

* **Naïve**: It is called Naïve because it assumes that the occurrence of a certain feature is independent of the occurrence of other features. Such as if the fruit is identified on the bases of color, shape, and taste, then red, spherical, and sweet fruit is recognized as an apple. Hence each feature individually contributes to identify that it is an apple without depending on each other.
* **Bayes**: It is called Bayes because it depends on the principle of [Bayes' Theorem](https://www.javatpoint.com/bayes-theorem-in-artifical-intelligence).

## Bayes' Theorem:

* Bayes' theorem is also known as **Bayes' Rule** or **Bayes' law**, which is used to determine the probability of a hypothesis with prior knowledge. It depends on the conditional probability.
* The formula for Bayes' theorem is given as:

Naïve Bayes Classifier Algorithm

**Where,**

**P(A|B) is Posterior probability**: Probability of hypothesis A on the observed event B.

**P(B|A) is Likelihood probability**: Probability of the evidence given that the probability of a hypothesis is true.

**P(A) is Prior Probability**: Probability of hypothesis before observing the evidence.

**P(B) is Marginal Probability**: Probability of Evidence.

### Applications of Naïve Bayes Classifier:

* It is used for **Credit Scoring**.
* It is used in **medical data classification**.
* It can be used in **real-time predictions** because Naïve Bayes Classifier is an eager learner.
* It is used in Text classification such as **Spam filtering** and **Sentiment analysis**.

# 3.5.3 Logistic Regression in Machine Learning

* Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.
* Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, **it gives the probabilistic values which lie between 0 and 1**.
* Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas **Logistic regression is used for solving the classification problems**.
* In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).
* The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc.
* Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets.
* Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification. The below image is showing the logistic function:



Figure 3.5.3: logistic function

**Assumptions for Logistic Regression:**

* The dependent variable must be categorical in nature.
* The independent variable should not have multi-collinearity.

**Logistic Regression Equation:**

The Logistic regression equation can be obtained from the Linear Regression equation. The mathematical steps to get Logistic Regression equations are given below:

* We know the equation of the straight line can be written as:

Logistic Regression in Machine Learning

* In Logistic Regression y can be between 0 and 1 only, so for this let's divide the above equation by (1-y):

Logistic Regression in Machine Learning

* But we need range between -[infinity] to +[infinity], then take logarithm of the equation it will become:

Logistic Regression in Machine Learning

The above equation is the final equation for Logistic Regression.

**Type of Logistic Regression:**

On the basis of the categories, Logistic Regression can be classified into three types:

* **Binomial:** In binomial Logistic regression, there can be only two possible types of the dependent variables, such as 0 or 1, Pass or Fail, etc.
* **Multinomial:** In multinomial Logistic regression, there can be 3 or more possible unordered types of the dependent variable, such as "cat", "dogs", or "sheep"
* **Ordinal:** In ordinal Logistic regression, there can be 3 or more possible ordered types of dependent variables, such as "low", "Medium", or "High".

# 3.5.4 Decision Tree Classification Algorithm

* Decision Tree is a **Supervised learning technique**that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where**internal nodes represent the features of a dataset, branches represent the decision rules** and **each leaf node represents the outcome.**
* In a Decision tree, there are two nodes, which are the **Decision Node** and**Leaf Node.** Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.
* The decisions or the test are performed on the basis of features of the given dataset.
* **It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.**
* It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.
* In order to build a tree, we use the **CART algorithm,** which stands for **Classification and Regression Tree algorithm.**
* A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into subtrees.
* Below diagram explains the general structure of a decision tree:



Figure 3.5.4: General Structure of A Decision Tree

**Why use Decision Trees?**

There are various algorithms in Machine learning, so choosing the best algorithm for the given dataset and problem is the main point to remember while creating a machine learning model. Below are the two reasons for using the Decision tree:

* Decision Trees usually mimic human thinking ability while making a decision, so it is easy to understand.
* The logic behind the decision tree can be easily understood because it shows a tree-like structure.

**Example:**

Suppose there is a candidate who has a job offer and wants to decide whether he should accept the offer or Not. So, to solve this problem, the decision tree starts with the root node (Salary attribute by ASM). The root node splits further into the next decision node (distance from the office) and one leaf node based on the corresponding labels. The next decision node further gets split into one decision node (Cab facility) and one leaf node. Finally, the decision node splits into two leaf nodes (Accepted offers and Declined offer). Consider the below diagram:



Figure 3.5.5: Decision Node Splits into Two Leaf Nodes

# 3.5.5 Random Forest Algorithm

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of **ensemble learning,** which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

As the name suggests, ***"Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset."*** Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

**The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.**

The below diagram explains the working of the Random Forest algorithm:

****

Figure 3.5: Working of Random Forest Algorithm

**Why use Random Forest?**

Below are some points that explain why we should use the Random Forest algorithm:

<="" li="">

* It takes less training time as compared to other algorithms.
* It predicts output with high accuracy, even for the large dataset it runs efficiently.
* It can also maintain accuracy when a large proportion of data is missing.

**Applications of Random Forest:**

There are mainly four sectors where Random forest mostly used:

1. **Banking:** Banking sector mostly uses this algorithm for the identification of loan risk.
2. **Medicine:** With the help of this algorithm, disease trends and risks of the disease can be identified.
3. **Land Use:** We can identify the areas of similar land use by this algorithm.
4. **Marketing:** Marketing trends can be identified using this algorithm.

**System Analysis**

## CHAPTER 4

## SYSTEM ANALYSIS

## 4.1 Existing System:

* In the current existing system so many researchers working with data mining algorithms in different survey techniques.
* In these survey techniques they are using different sources but they are unable to analyze data with visualization techniques to identify the correlation between different attributes.
* Using limited techniques and they are unable to optimize by increasing efficiency.

## 4.1.1 Disadvantages of Existing System:

## Increasing traffic congestion, air pollution, and fuel consumption

## Increase in use of less-adequate roads to avoid traffic signs

## Excessive delay due to time allocated by the traffic signals

## Some drivers disobey certain rules

**4.2 Proposed System:**

In our proposed system we are collecting data from different data sources and using advanced data visualization modules in python and analyze data with different visualization techniques and we are using different Machine learning Algorithms and evaluate increase performance and accuracy.

**4.2.1 Advanatges of Proposed System:**

* After each iteration, regression testing should be conducted. During this testing, faulty elements of the software can be quickly identified because few changes are made within any single iteration.
* It is generally easier to test and debug than other methods of software development because relatively smaller changes are made during each iteration. This allows for more targeted and rigorous testing of each element within the overall product.
* Customer can respond to features and review the product for any needed or useful changes.
* Initial product delivery is faster and also lowers the initial delivery cost.

**SYSTEM DESIGN**

## CHAPTER 5

## SYSTEM DESIGN

**5.1 Modules:**

* User
* Admin
* System
* Database

**5.1.1 User:**

The client enter data or created using accident data from records . In many projects the end user is the one who needs the solution and gets the solution and that’s the total of their involvement.   **user** is the term used for those people who actually use a product .For this project the users are public , apps and security .

* + 1. **Admin:**

Admin to login and also collect data from UCI dataset. The **UCI** Machine Learning Repository is a collection of databases, domain theories, and **data** generators that are used by the machine learning community for the empirical analysis of machine learning algorithms.

* + 1. **System:**

It can illustrated how statistical method based on directed graphs and also print the result

* + 1. **Database:**

A database is a data structure that stores organized information.

* 1. **SYSTEM ARCHITECTURE:**

A **system architecture** or **systems architecture** is the conceptual model that defines the structure, behavior, and more views of a **system**. An **architecture** description is a formal description and representation of a **system**, organized in a way that supports reasoning about the structures and behaviors of the **system**.

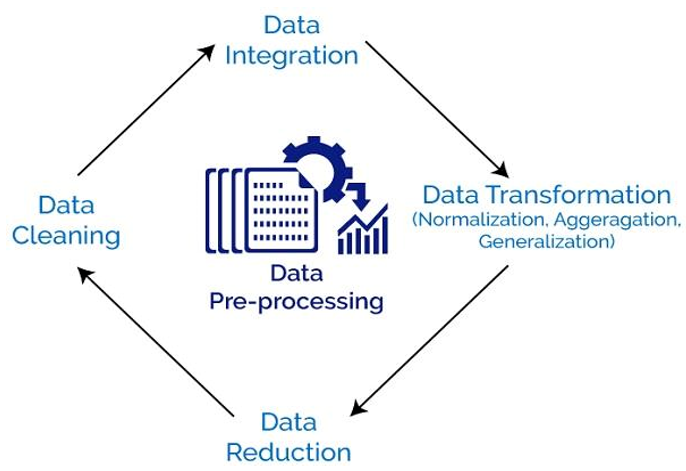
****

Figure 5.2: System Architecture

**5.3 UML Diagrams:**

UML stands for Unified Modeling Language. UML is a standardized general-purpose modeling language in the field of object-oriented software engineering. The standard is managed and was created by the Object Management Group. The goal is for UML to become a common language for creating models of object oriented computer software. In its current form UML is comprised of two major components: a Meta-model and a notation. In the future, some form of method or process may also be added to; or associated with, UML.

The Unified Modeling Language is a standard language for specifying, Visualization, Constructing and documenting the artifacts of software system, as well as for business modeling and other non-software systems. The UML represents a collection of best engineering practices that have proven successful in the modeling of large and complex systems.

**GOALS:**

The Primary goals in the design of the UML are as follows:

1. Provide users a ready-to-use, expressive visual modeling Language so that they can develop and exchange meaningful models.
2. Provide extendibility and specialization mechanisms to extend the core concepts.
3. Be independent of particular programming languages and development process.
4. Provide a formal basis for understanding the modeling language.
5. Encourage the growth of OO tools market.
6. Support higher level development concepts such as collaborations, frameworks, patterns and components.
7. Integrate best practices.

**5.3.1 Architecture Design:**

* Design of a system consists of classes, interfaces, and collaboration.
* UML provides class diagram, object diagram to support this.
* Implementation defines the components assembled together to make a complete physical system.
* UML component diagram is used to support the implementation perspective.

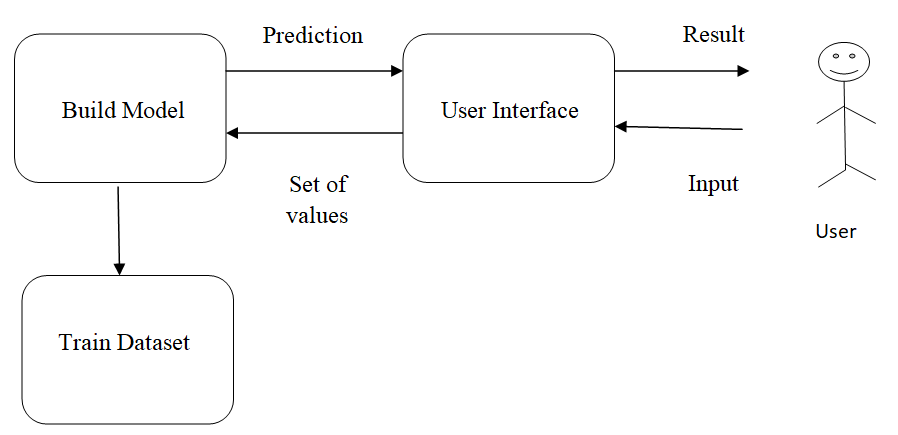


Figure 5.3.1: Architecture Design

**5.3.2 Activity Diagram:**

* Activity diagram is another important diagram in UML to describe the dynamic aspects of the system.
* Activity diagram is basically a flowchart to represent the flow from one activity to another activity.
* The activity can be described as an operation of the system.

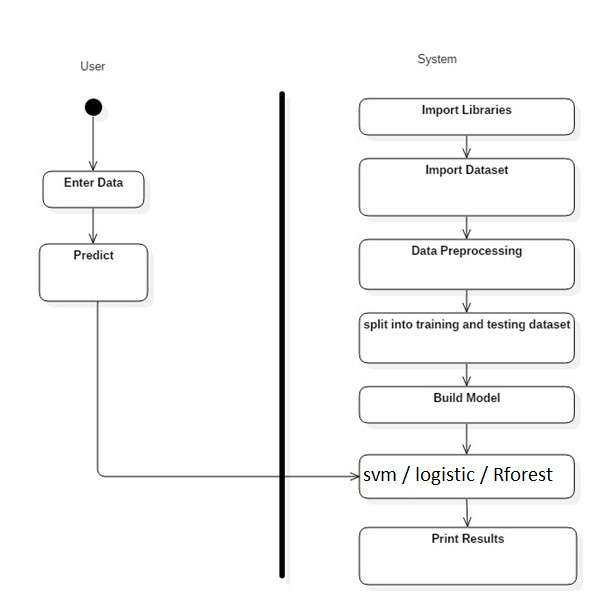
****

Figure 5.3.2: Activity Diagram

* + 1. **Sequence Diagram:**
* A sequence diagram is a type of interaction diagram because it describes how and in what order a group of objects works together.
* These diagrams are used by software developers and business professionals to understand requirements for a new system or to document an existing process.
* It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams.

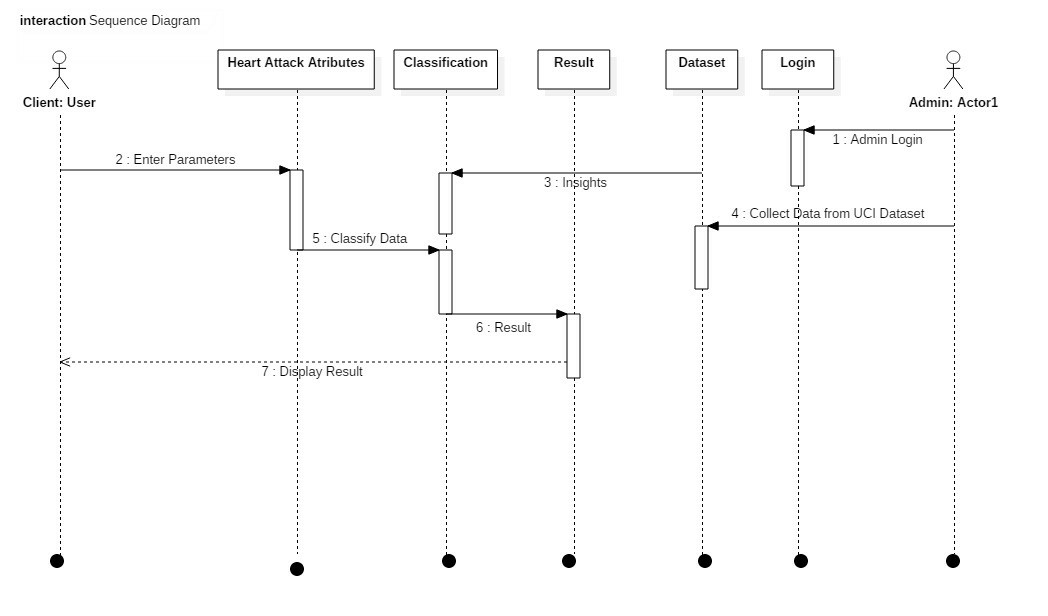
****

Figure 5.3.3: Sequence Diagram

* + 1. **Interface Diagram:**
* In UML modeling, interfaces are model elements that define sets of operations that other model elements, such as classes, or components must implement.
* You can use interfaces in class diagrams and component diagrams to specify a contract between the interface and the classifier that realizes the interface.

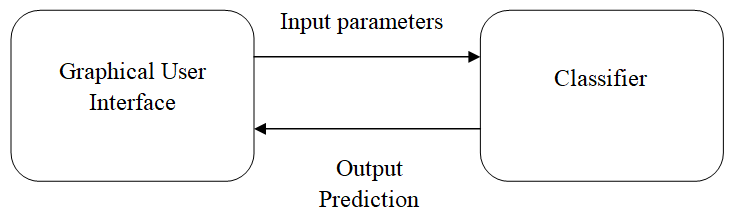
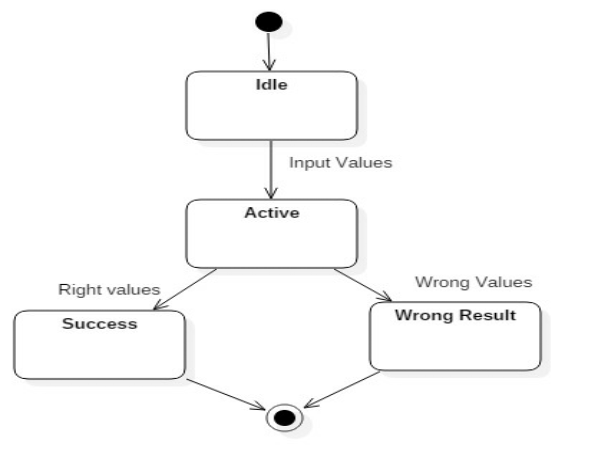


Figure 5.3.4: Interface Diagram

* + 1. **State Machine Diagram:**
* State machine diagram is a behavior diagram which shows discrete behavior of a part of designed system through finite state transitions.
* State machine diagrams can also be used to express the usage protocol of part of a system.

Figure 5.3.5: State Machine Diagram

**SYSTEM REQUIREMENTS**

## CHAPTER 6

**SYSTEM REQUIREMENTS**

**6.1 HARDWARE REQUIREMENTS:**

For developing the application, the following are the Hardware Requirements:

## System : i5 Processor

## Hard Disk : 500 GB.

## Input Devices : Keyboard, Mouse

## Ram : 4GB

## 6.2 SOFTWARE REQUIREMENTS:

## For developing the application, the following are the Software Requirements:

## IDE : Anaconda Python, pandas, Jupyter Notebook

## Operating systems : Windows 7/10

## Coding language : Python

**SYSTEM IMPLIMENTATION**

## CHAPTER 7

**SYSTEM IMPLIMENTATION**

**7.1 MODULES DESCRIPTION:**

* Data Pre-Processing
* Classification with logistic regression

**7.1.1 Data Pre-Processing:**

Before actually applying the artificial intelligence algorithms to the data, it should be preprocessed. First of all it was decided to use only the statements themselves for classification purposes. This means that none o

f the metadata provided is used for classification. The classification algorithm might actually be improved in the future by taking into account this metadata

**7.1.2 Classification with logistic regression:**

Logistic regression is a statistical method for analyzing a data set in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes). For the cases when there are more than two labels, the strategy, which is called “One versus all”, is used. In this strategy every category is binary classified against its inverse (a fictional category that states that the example does not belong to the current category). The category with the highest score is picked as a result of a classification.

* 1. **Software and Languages used:**
     1. **Jupyter**

Jupyter exists to develop open-source software. It is used for open-standards, and services for interactive computing across dozens of programming languages. It is an opensource web application that allows you to create and share documents and code live. Which is a very big advantage of Jupyter. It can be used for data cleaning and transformation, numerical simulation, statistical modeling, machine learning and much more. We used Jupyter to run the algorithm.

* + 1. **Python**

Python is an interpreted, high-level and a general-purpose programming language. Created by Guido van Rossum and first released in 1991, Python has a design philosophy that emphasizes code readability, notably using significant whitespace. It is the most used programming language presently. It provides constructs that enable clear programming on both small and large scales. The logistic regression in the system is implemented in jupyter and the algorithm is written in python language.

**7.2.3 HTML,CSS,JSCRIPT**

These are the most commonly used webpage development languages. The user interface designed for the prediction system is been developed using these languages. The website acts as an interface which takes the input those are the various constraints from the users and passes to the program to work upon.

**7.3 WORKING**

The working of the project is divided in to parts.

**7.3.1 Data Set Selection**

Data is the most import part when you work on prediction systems. It plays a very vital role your whole project i.e., you system depends on that data. So selection of data is the first and the critical step which should be performed properly, For our project we got the data from the government website. These datasets were available for all. There are other tons of websites who provide such data. The dataset we choose wad selected based on the various factors and constraints we were going to take under the consideration for our prediction system.

**7.3.2 Data Cleaning and Data Transformation**

After we have selected the dataset. The next step is to clean the data and transform it into the desired format as it is possible the dataset we use may be of different format. It is also possible that we may use multiple datasets from different sources which may be in different file formats. So to use them we need to convert them into the format we want to or the type that type prediction system supports. The reason behind this step is that it is possible that the data set contains the constraints which are not needed by the prediction system and including them makes the system complicated and may extend the processing time. Another reason behind data cleaning is the dataset may contain null value and garbage values too. So the solution to this issue is when the data is transformed the garbage values are replaced. There are many methods to perform that.

**7. 3.3 Data Processing and Algorithm Implementation**

After the data is been cleaned and transformed it’s ready to process further. After the data has been cleaned and we have taken the required constraints. We divide the whole dataset int o the two parts that can be either 70-30 or 80-20. The larger portion of the data is for the processing. The algorithm is applied on that part of data. Which helps the algorithm to learn on its own and make prediction for the future data or the unknown data. The algorithm is executed in which we take only the required constraints from the cleaned data. The output of the algorithm is in ‘yes’ and ‘no’. It gives the error rate and the success rate.

**7.3.4 Output and User Side Experience**

After the prediction system is ready to use. The Website is developed for the user. The user just has to fill a form which consists of different options they need to select. They are like the type of climate, the type of vehicle and so on. Once the user submits the form the algorithm is triggered and the input given by the user is passed to the prediction system. The user is given how accident prone the road can be in percentage.

**7.4 Input And Output Design**

**7.4.1 Input Design**

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. Input Design considered the following things:

* What data should be given as input?
* How the data should be arranged or coded?
* The dialog to guide the operating personnel in providing input.
* Methods for preparing input validations and steps to follow when error occur

**7.4.2 Objectives**

1.Input Design is the process of converting a user-oriented description of the input into a computer-based system. This design is important to avoid errors in the data input process and show the correct direction to the management for getting correct information from the computerized system.

2. It is achieved by creating user-friendly screens for the data entry to handle large volume of data. The goal of designing input is to make data entry easier and to be free from errors. The data entry screen is designed in such a way that all the data manipulates can be performed. It also provides record viewing facilities.

3.When the data is entered it will check for its validity. Data can be entered with the help of screens. Appropriate messages are provided as when needed so that the user will not be in maize of instant. Thus the objective of input design is to create an input layout that is easy to follow

**7.4.3 OUTPUT DESIGN**

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs.

In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user.

Efficient and intelligent output design improves the system’s relationship to help user decisionmaking. Designing computer output should proceed in an organized well thought out manner; the right output must be developed while ensuring that each output element is designed so that people will find the system can use easily and effectively. When analysis design computer output, they should Identify the specific output that is needed to meet the requirements.

1. Select methods for presenting information.

2. Create document, report, or other formats that contain information produced by the system.

3. The output form of an information system should accomplish one or more of the following objectives.

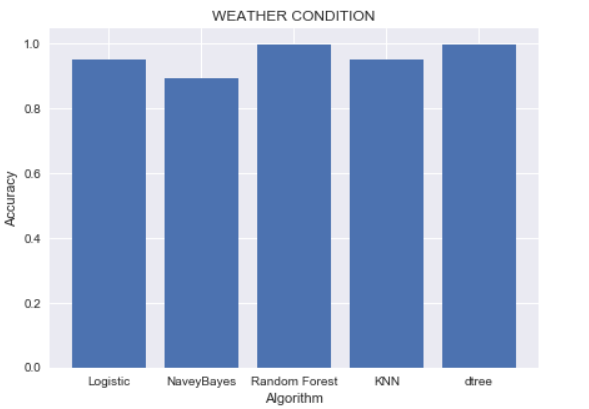
* Convey information about past activities, current status or projections of the Future.
* Signal important events, opportunities, problems, or warnings.
* Trigger an action.
* Confirm an action.

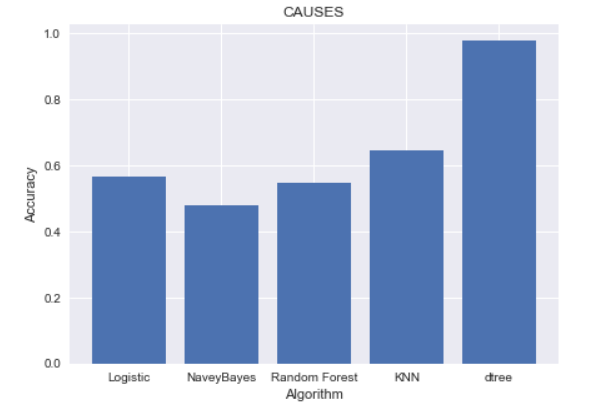
EVALUATION AND RESULTS:

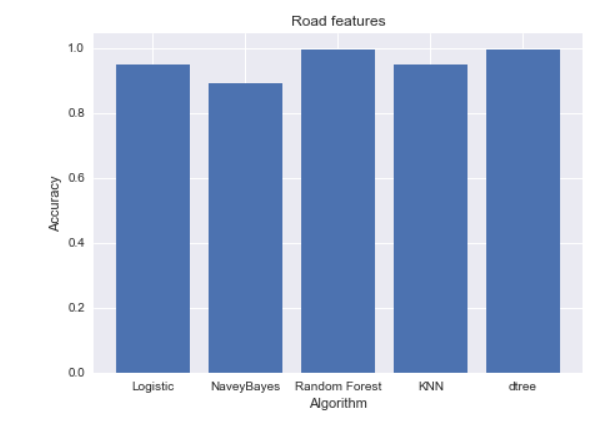
In that classification algorithm we will use Logistic Regression Algorithm The logistic algorithm will make the prediction in terms of percentage, to find accuracy level in percentage and Error percentages. This Algorithm is only for the yes and no type of result or successful and unsuccessful. The equation for combinations of all 15 input variables.

The classification algorithm of the entire dataset. In the Road Accident prediction final result is to find the percentage of accident in particular area. Having lower number of features helps the algorithm to converge faster and increases accuracy.

In the Road Accident prediction final result is to find the percentage of accident in particular area. Then we apply logistic regression on these features and obtain the least error.



****

****

For training and testing of the dataset, 80% of records are selected to train the model and the other 20% of records are selected to test the trained model.

Once the model has been built, that appear to have a high quality based on whichever loss functions have been selected, these need to be tested to ensure they generalize against unseen data and that all key business issues have been sufficiently considered.

Here,

• Class 1 : Positive

• Class 2 : Negative

**Definition of the Terms:**

• Positive (P) : Observation is positive (for example: is an apple).

• Negative (N) : Observation is not positive (for example: is not an apple).

• True Positive (TP) : Observation is positive, and is predicted to be positive.

• False Negative (FN) : Observation is positive, but is predicted negative.

• True Negative (TN) : Observation is negative, and is predicted to be negative.

• False Positive (FP) : Observation is negative, but is predicted positive.

**Accuracy =**

AUC - ROC curve is a performance measurement for classification problem at various thresholds settings. ROC is a probability curve and AUC represents degree or measure of separability. It tells how much model is capable of distinguishing between classes. Higher the AUC, better the model is at predicting 0s as 0s and 1s as 1s. By analogy, Higher the AUC, better the model is at distinguishing between patients with disease and no disease

**SYSTEM TESTING**

## CHAPTER 8

**SYSTEM TESTING**

* Testing is that the debugging program is one amongst the leading crucial aspects of the pc programming triggers, while not programming that works, the system would ne'er turn out relate in Nursing output of that it had been designed .
* Testing is best performed once user development is asked to help in characteristic all errors and bugs.
* The sample knowledge are used for testing. It is not amount however quality of the information used the matters of testing .
* Testing is aimed toward guaranteeing that the system was accurately relate in Nursing with efficiency before live operation commands

**Testing objectives:-**

* The most objective of testing is to uncover a bunch of errors, consistently and with minimum effort and time.
* Stating formally ,we can say, testing may be a method of corporal punishment a program with intent of finding miscalculation.
* A decent legal action is one that has likelihood of finding miscalculation, if it exists.The check is insufficient to find probably gift errors.

**8.1 TYPES OF TESTING**

* Unit testing we have a tendency to test every module separately and integrate with the general system. Unit testing focuses verification efforts on the littlest unit of code style within the module. this is often conjointly called module testing.
* The module of the system is tested individually. as an example the validation check is completed for variable the user input given by the user that validity of the information entered. it's terribly straightforward to search out error rectify the system. Every Module will be tested victimization the subsequent 2 Strategies: recording machine Testing and White Box Testing.

**8.1.1 BLACK BOX TESTING**

* Recording machine cheking may be a code testing techniques during which practicality of the code below test (SUT) is tested while not staring at the interior code structure
* implementation details and data of internal ways of the code .This type of testing is predicated entirely on the code needs and specifications.
* In recording machine Testing we have a tendency to simply concentrate on inputs and output of the package while not bothering concerning internal data of the code program.
* The on top of recording machine will be any package you wish to check. For example, Associate in Nursing software like Windows, a web site like Google ,a information like Oracle or maybe your own custom application .
* Under recording machine testing, you can check these applications by simply that specialize in the inputs and outputs while not knowing their internal code implementation.

**Types of Black Box Testing**

There are many varieties of recording machine Testing however following ar the outstanding ones.

**Functional testing:** This recording machine testing kind is said to purposeful needs of a system; it's done by code testers.

**Non-Functional testing:** This sort of recording machine testing isn't associated with testing of a selected practicality, however non-functional needs like performance, measurability, usability.

**Egression Testing:** Regression testing is completed once code fixes, upgrades or the other system maintenance to visualize the new code has not affected the prevailing code.

**8.1.2. WHITE BOX TESTING**

* White Box Testing is that the testing of a code solution's internal committal to writing and infrastructure.
* It focuses totally on Traffic Redundancy Elimination ngthening security, the flow of inputs and outputs through the applying, and rising style and value.
* White box testing is additionally called clear, open, structural, and glass box testing. It is one amongst 2 elements of the "box testing" approach of code testing.

**System Testing:**

Once the individual module testing is completed, modules are assembled and integrated to perform as a system. The top down testing, that began from higher level to lower level module, was allotted to visualize whether or not the whole system is playacting satisfactorily. There are 3 main types of System testing:

1. Alpha Testing
2. Beta Testing and
3. Acceptance Testing.

**Alpha Testing:**

This refers to the system checking that's allotted by the test team with the Organization.

**Beta Testing:**

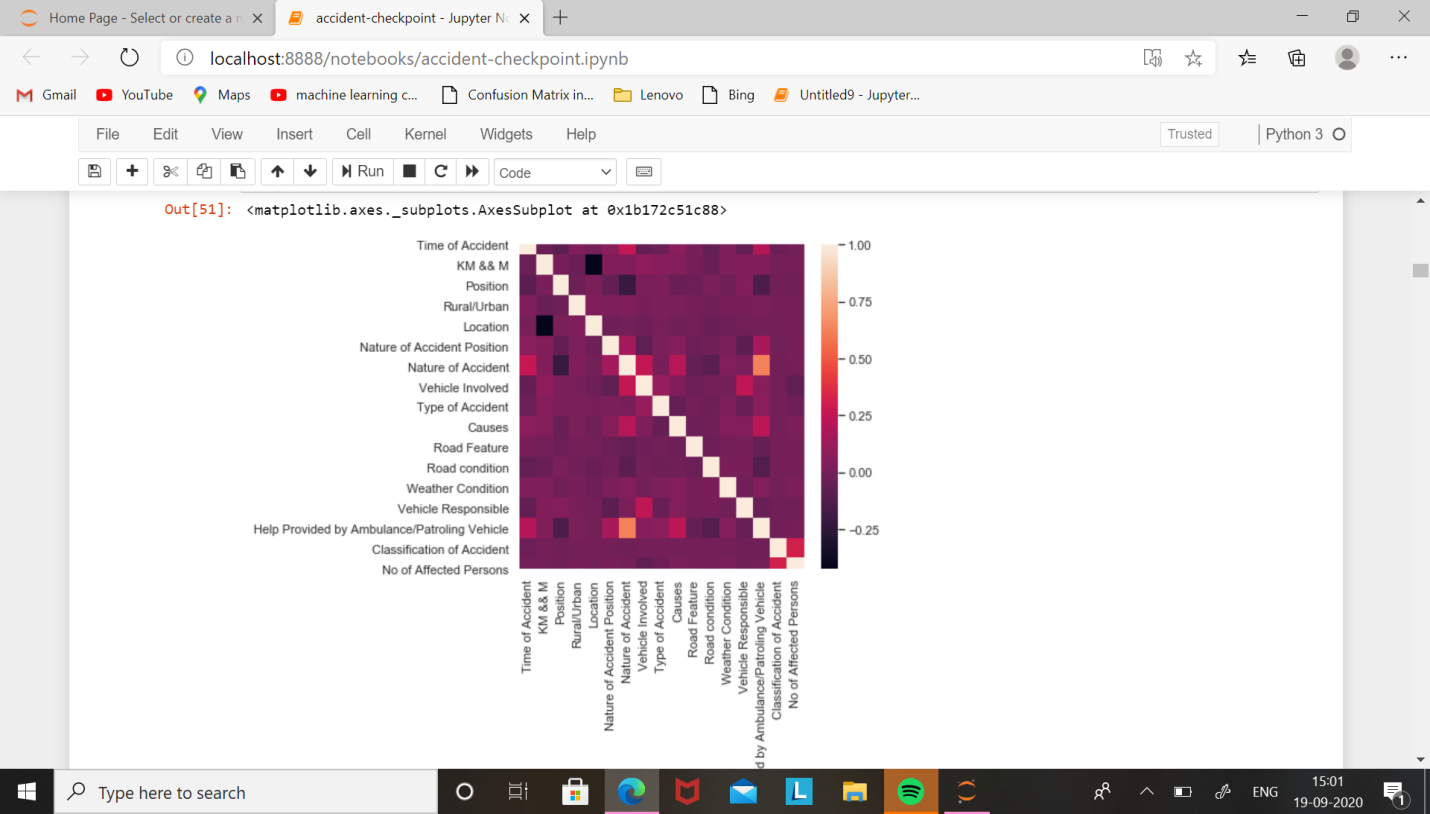
This refers to the system testing that's performed by a particular cluster of friendly customers.

**Acceptance Testing:**

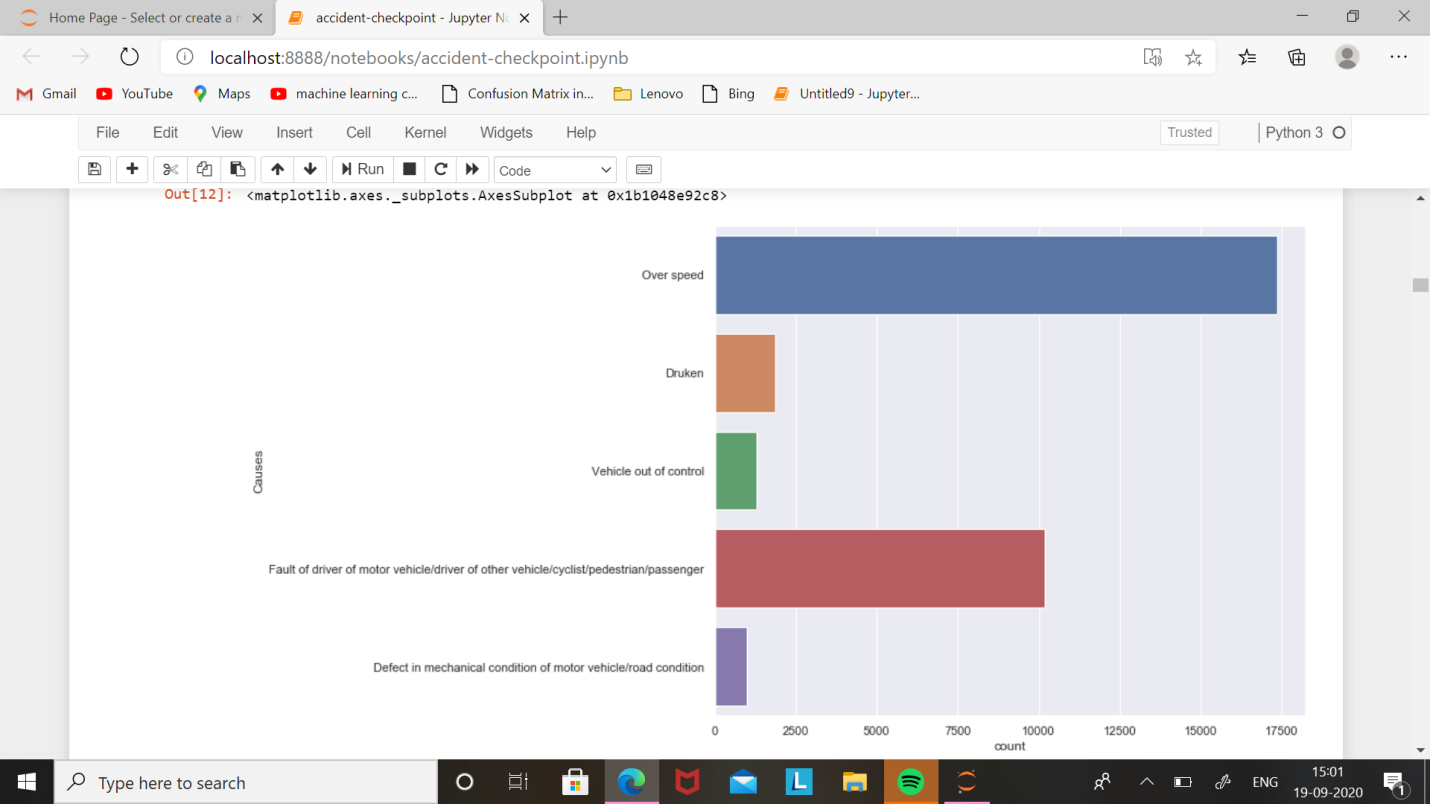
This refers to the system testing that's performed by the client to see whether or not or to not settle for the delivery of the system.

**OUTPUT SCREENS**

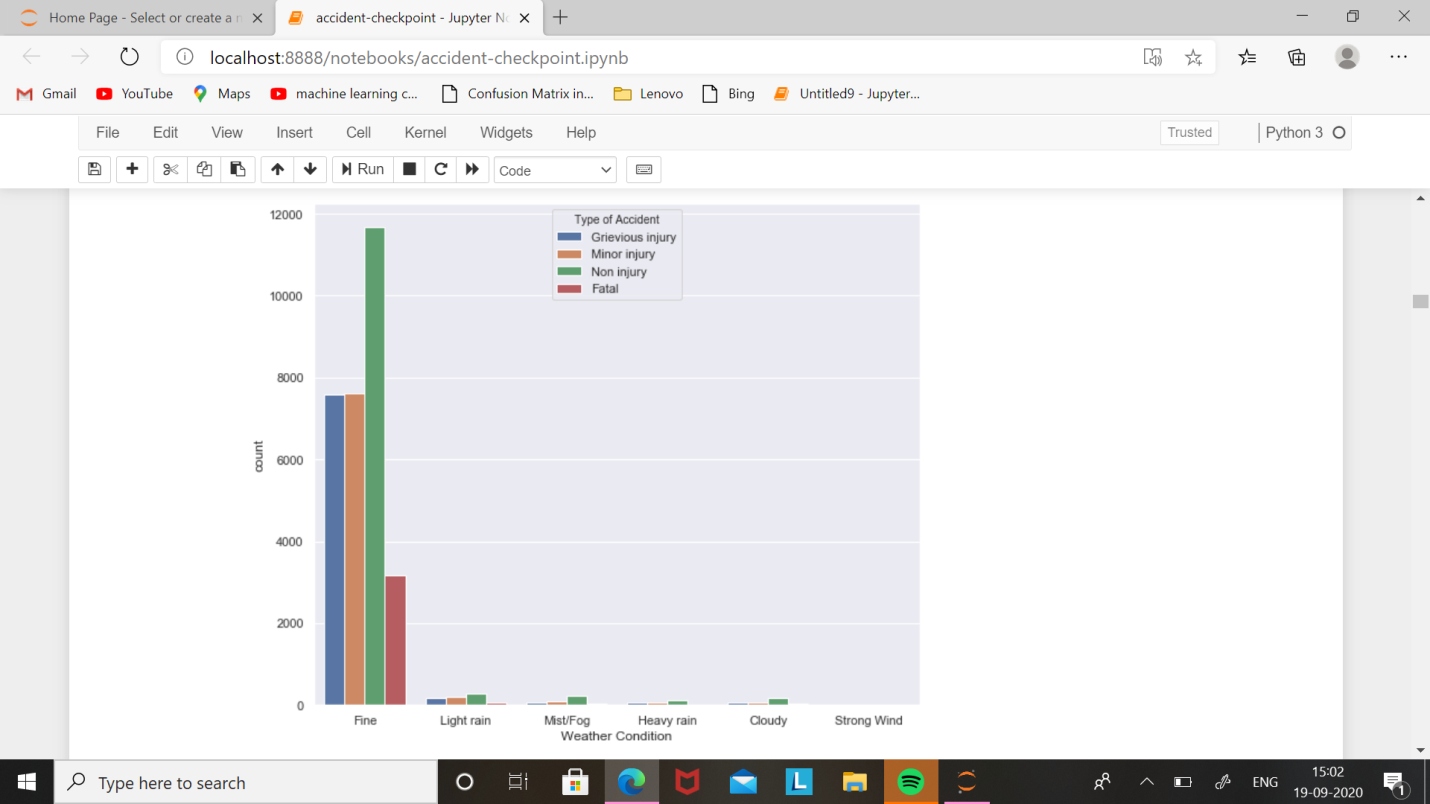
## CHAPTER 9

**OUTPUT SCREENS**

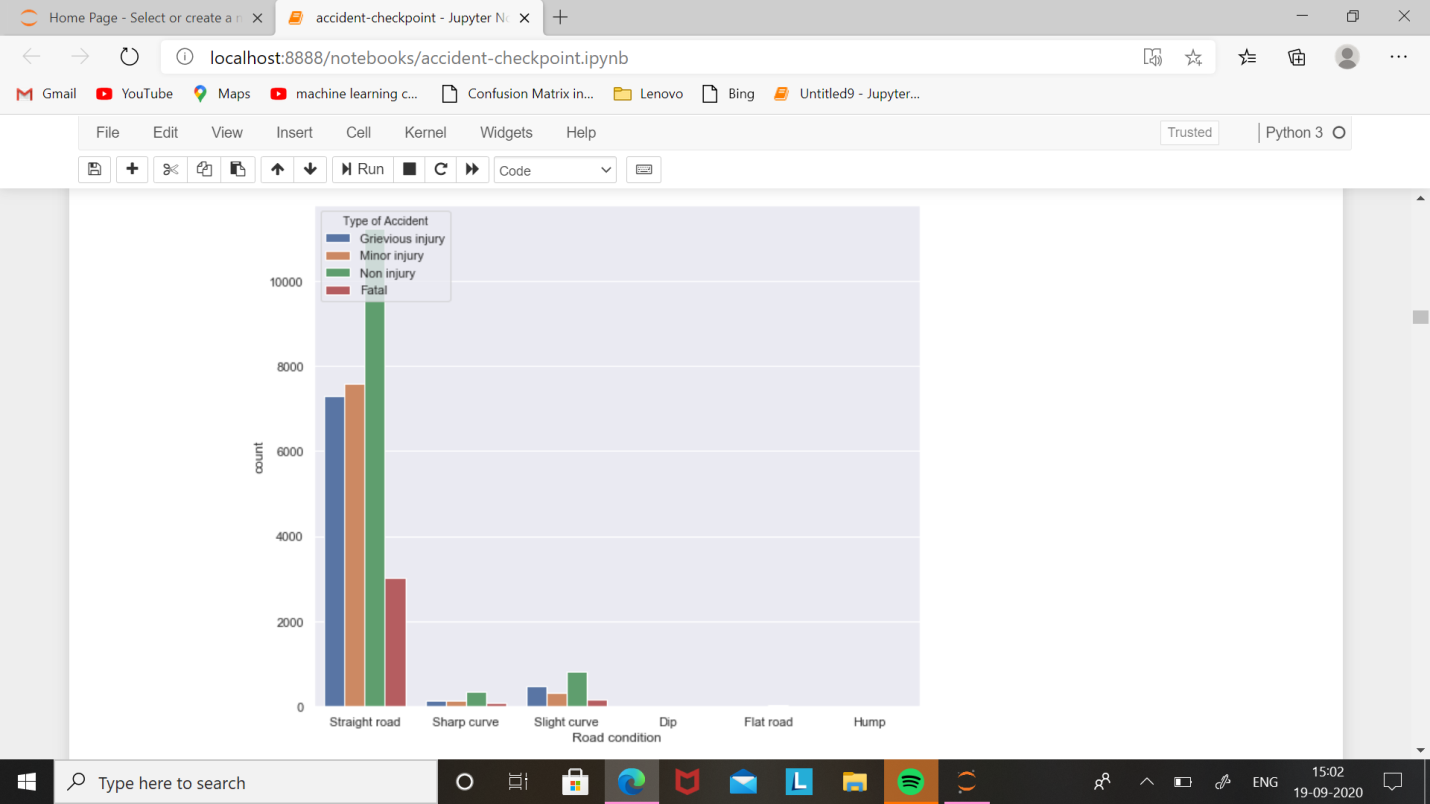
**Fig 9.1: Correlation matrix with Heatmap**



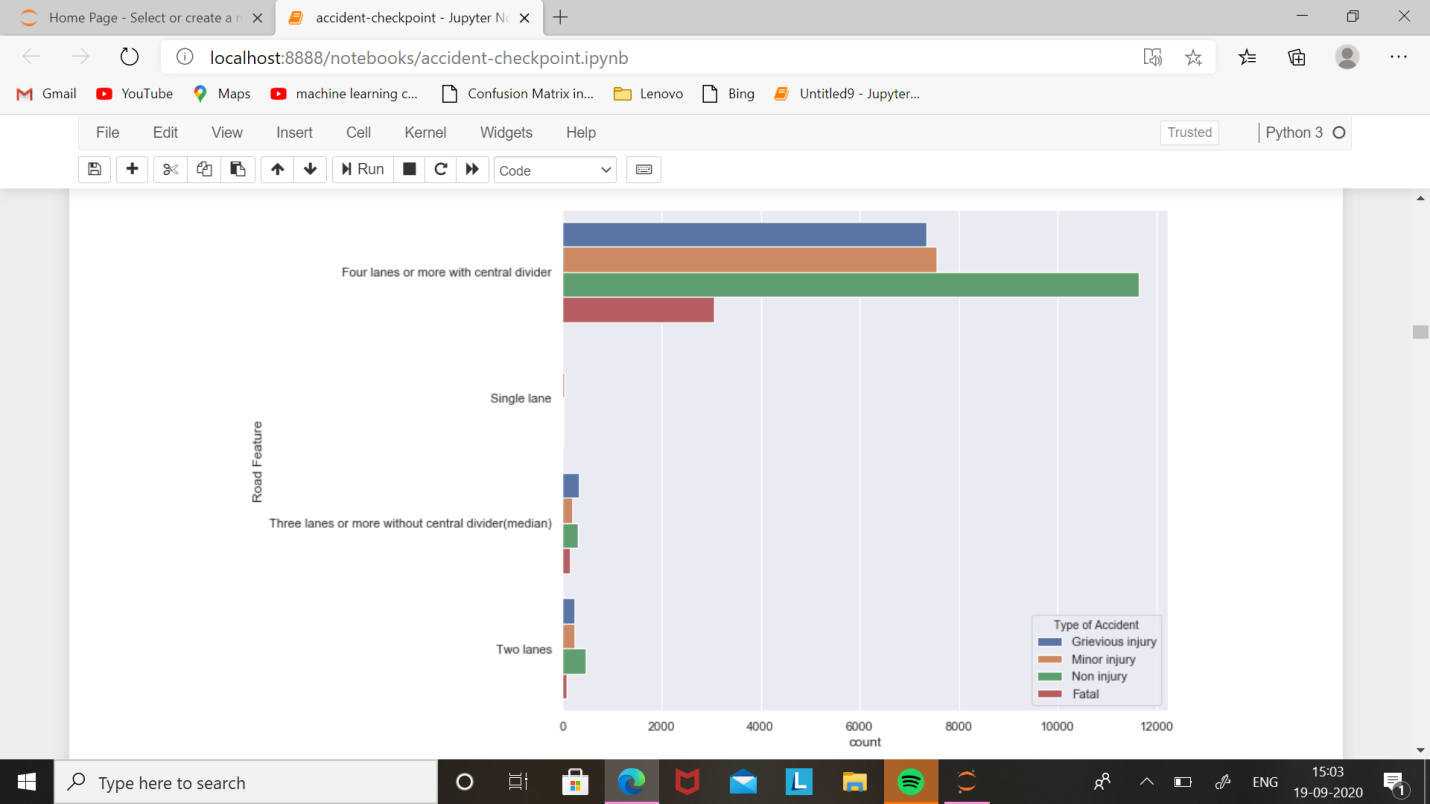
**Fig 9.2: Causes with Countplot on Seaborn**



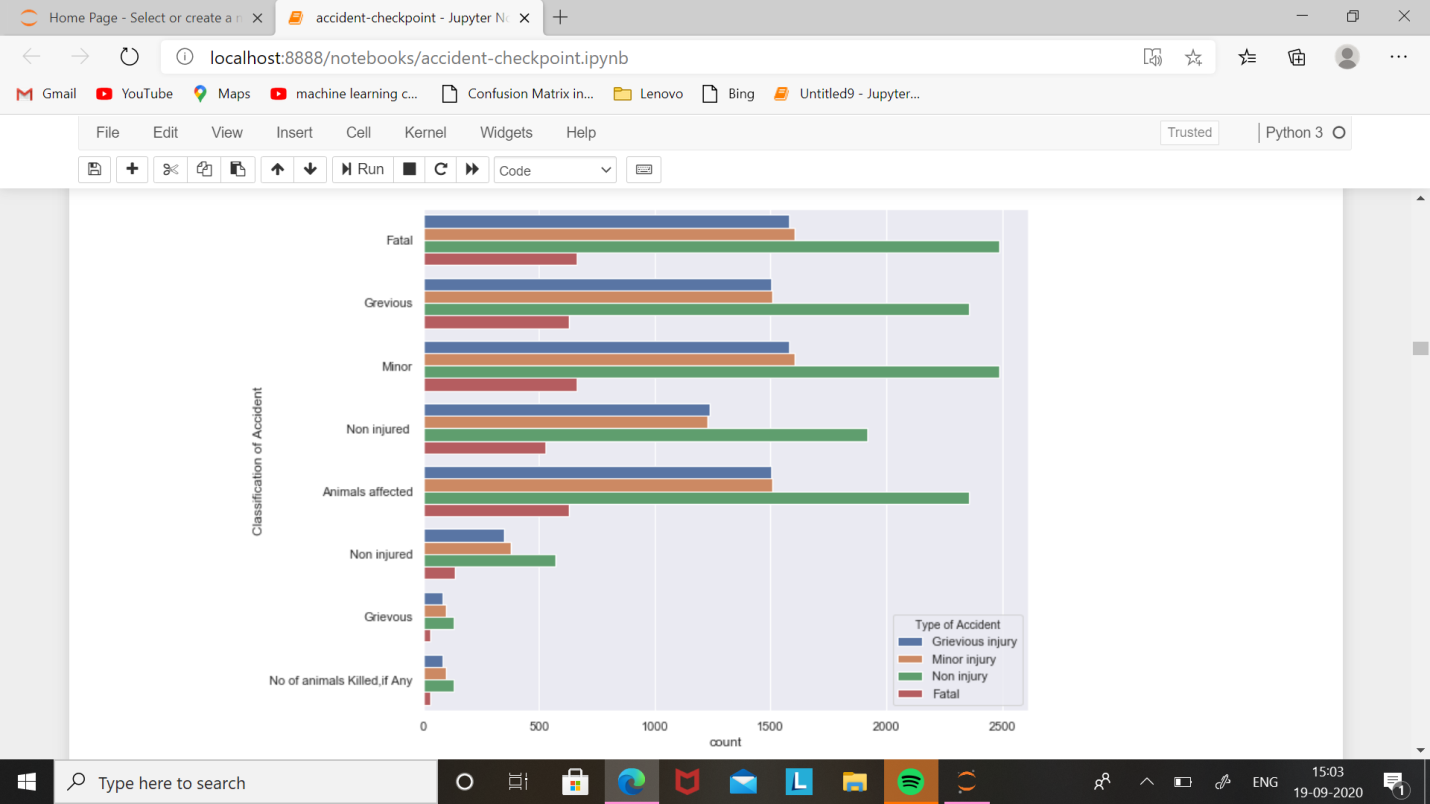
**Fig 9.3: Weather Condition with Countplot on Seaborn**



**Fig 9.4: Road Condition with Countplot on Seaborn**



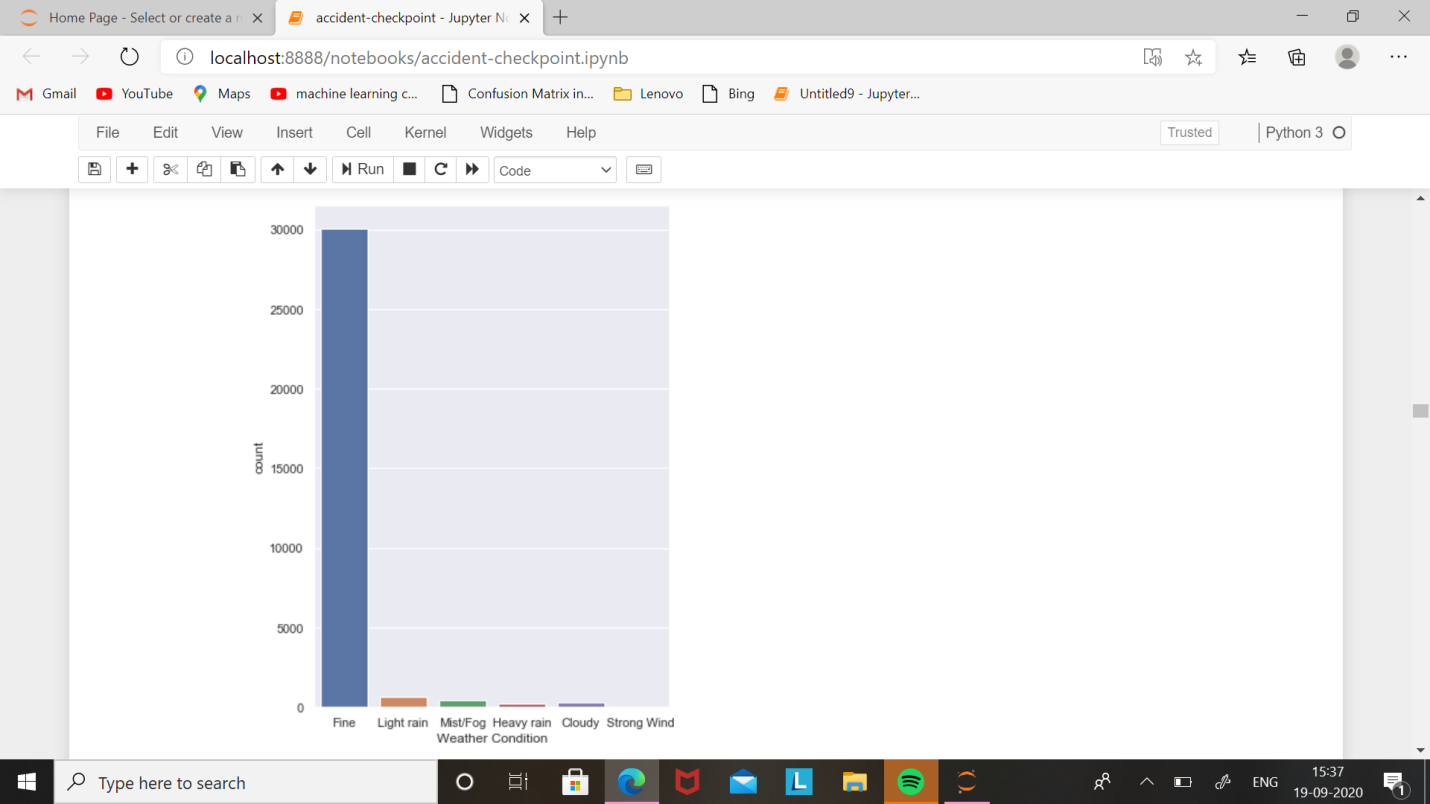
**Fig 9.5: Road Feature with Countplot on Seaborn**



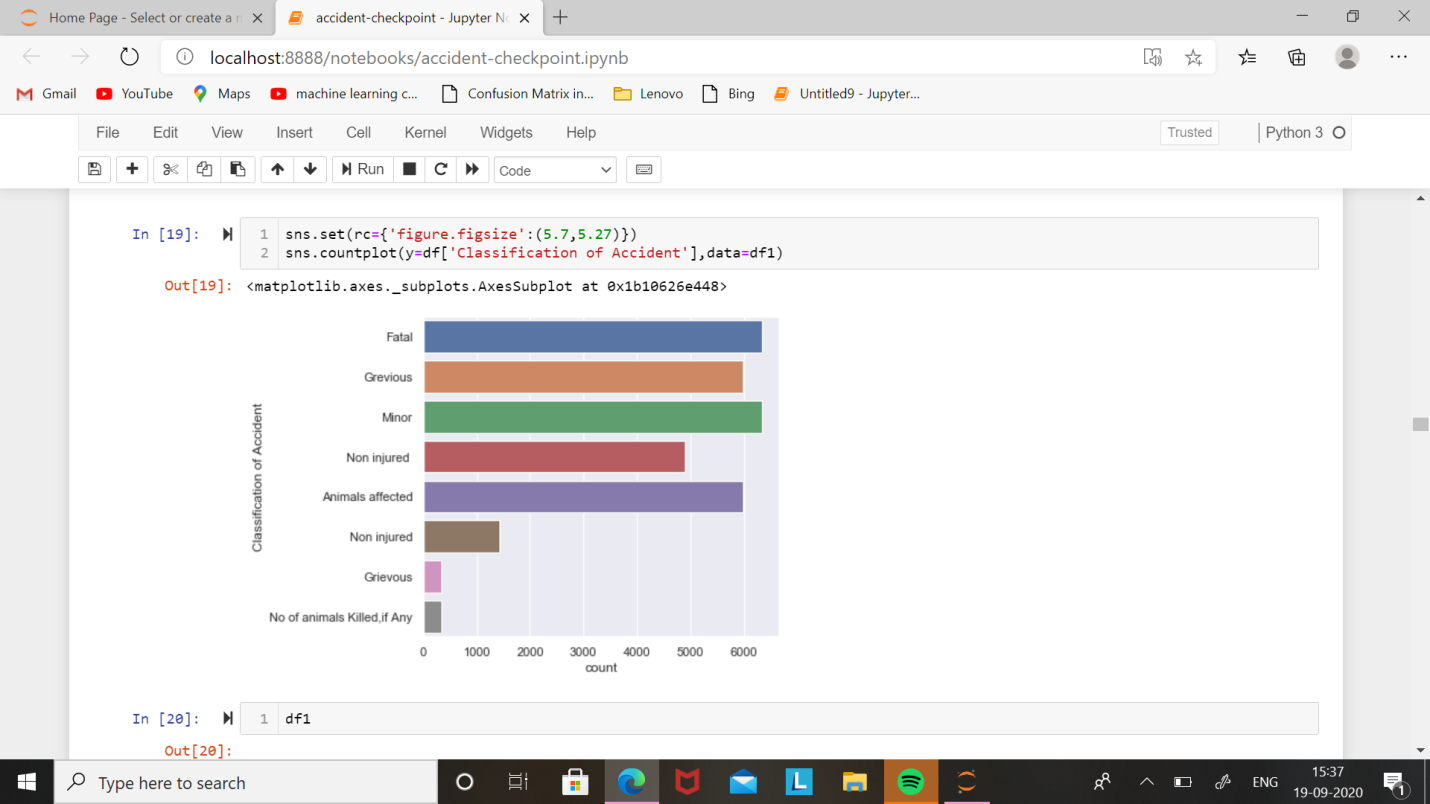
**Fig 9.6: Classification of Accident with Countplot on Seaborn**



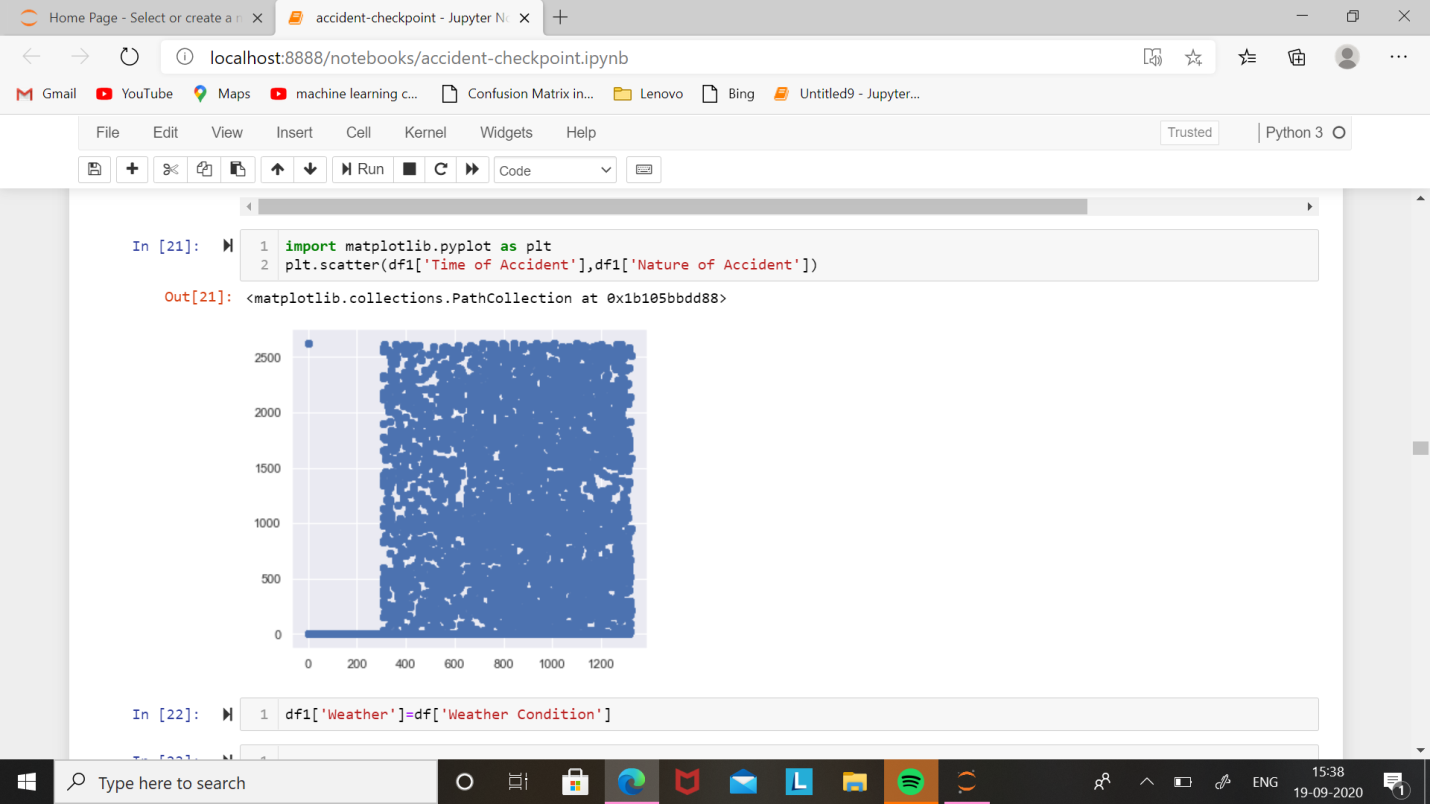
**Fig 9.7: Vehicle Involved with Countplot on Seaborn**



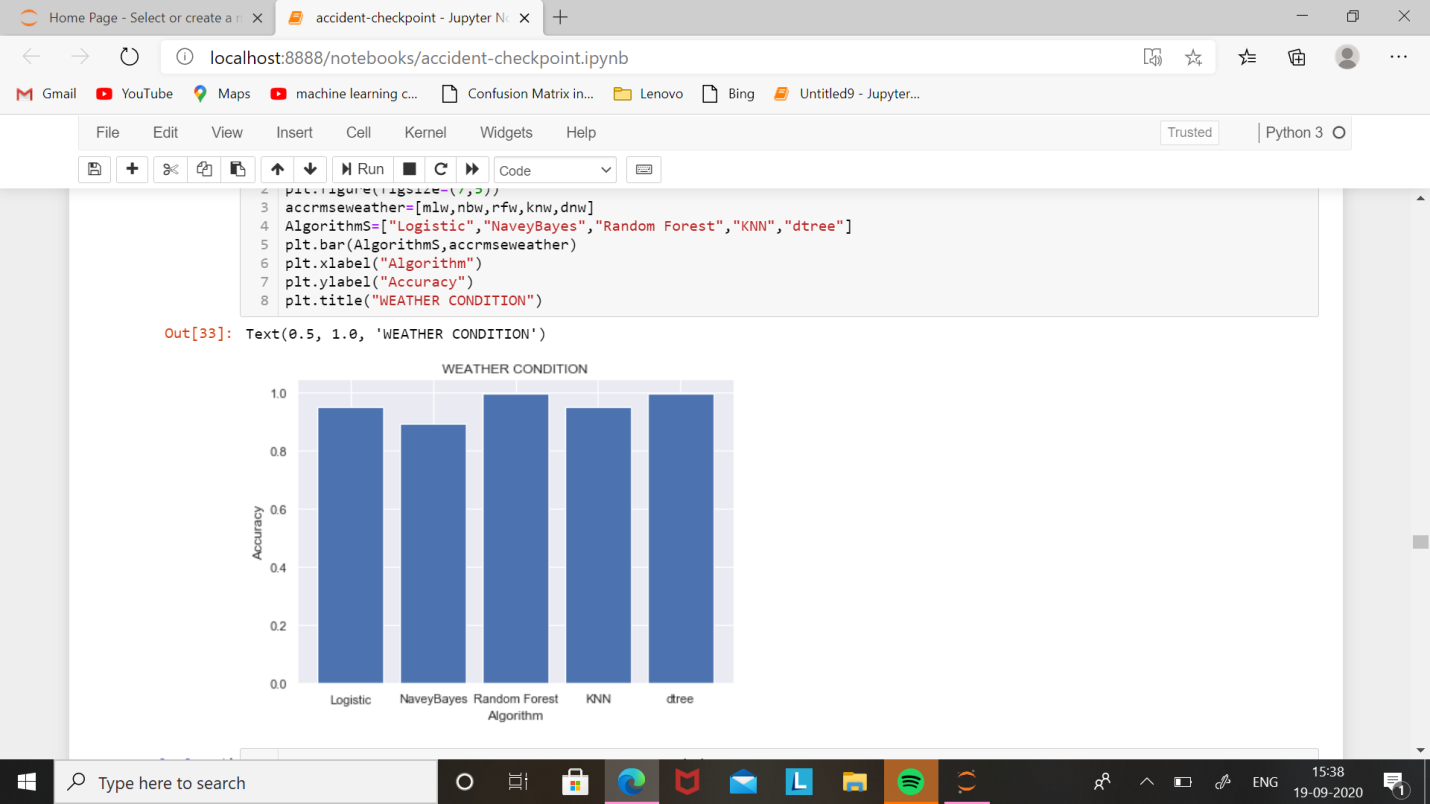
**Fig 9.8: Weather Condition with Countplot on Seaborn**

****

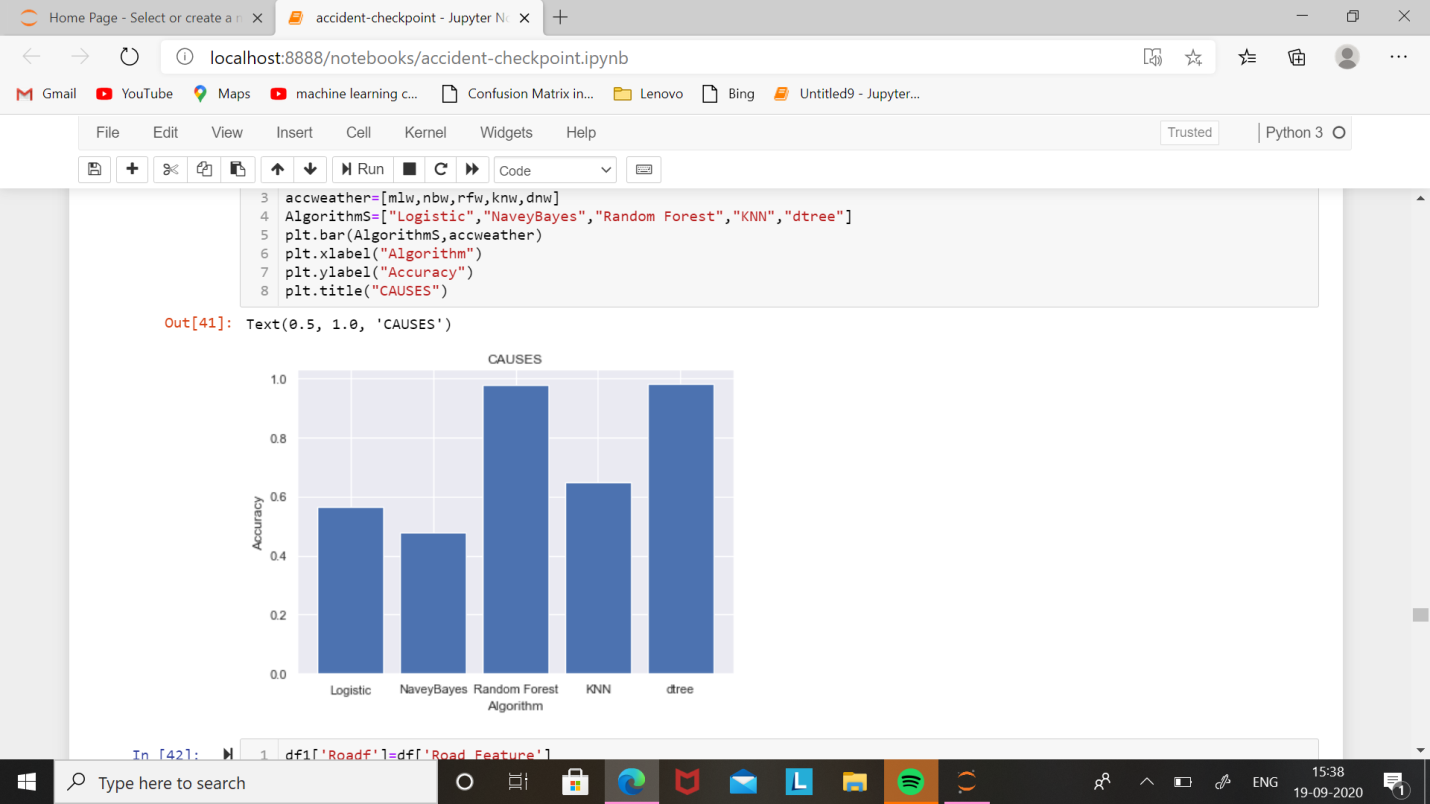
**Fig 9.9: Classification of Accident with Countplot on Seaborn**

****

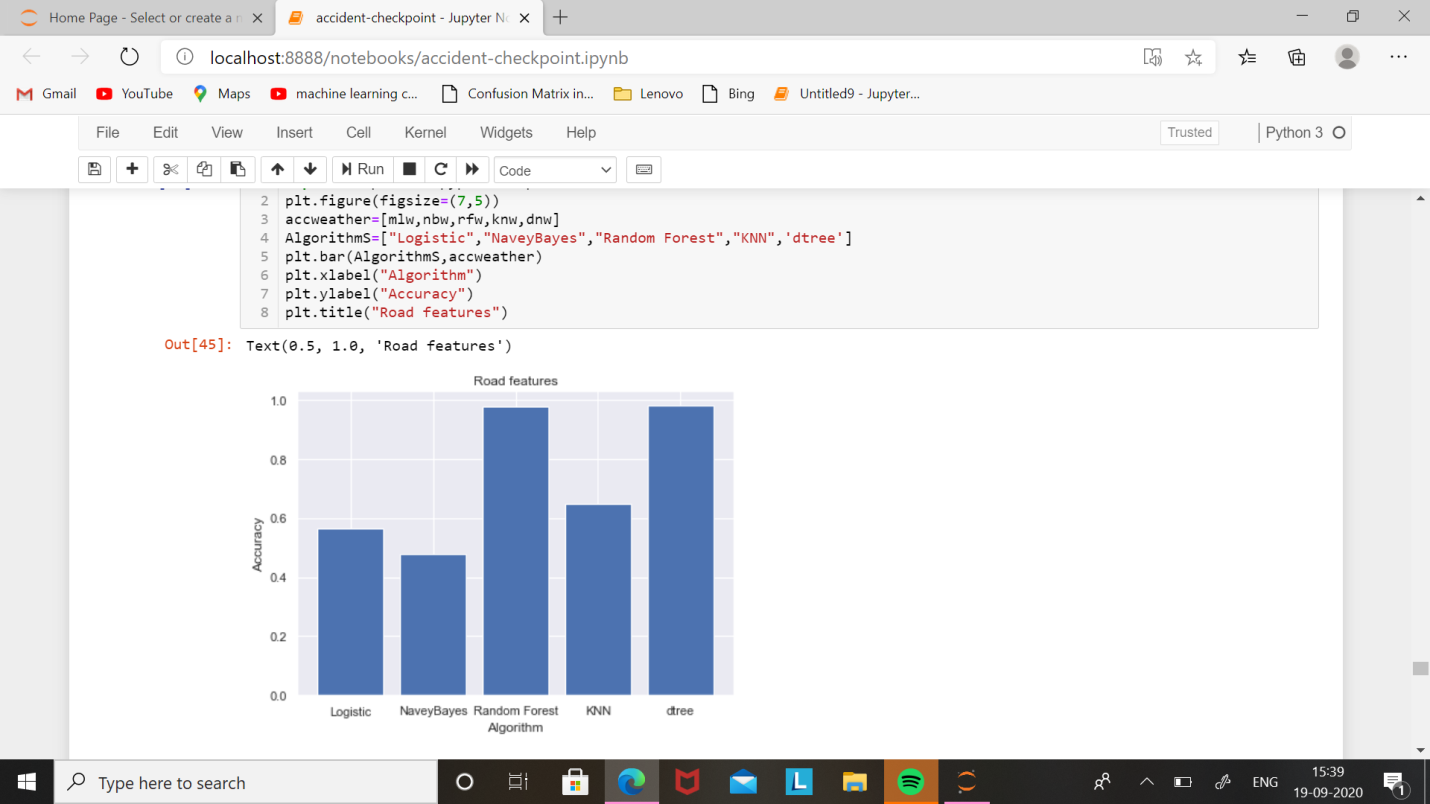
**Fig 9.10: Time of Accident & Nature of Accident with Scatterplot on Matplotlib**

****

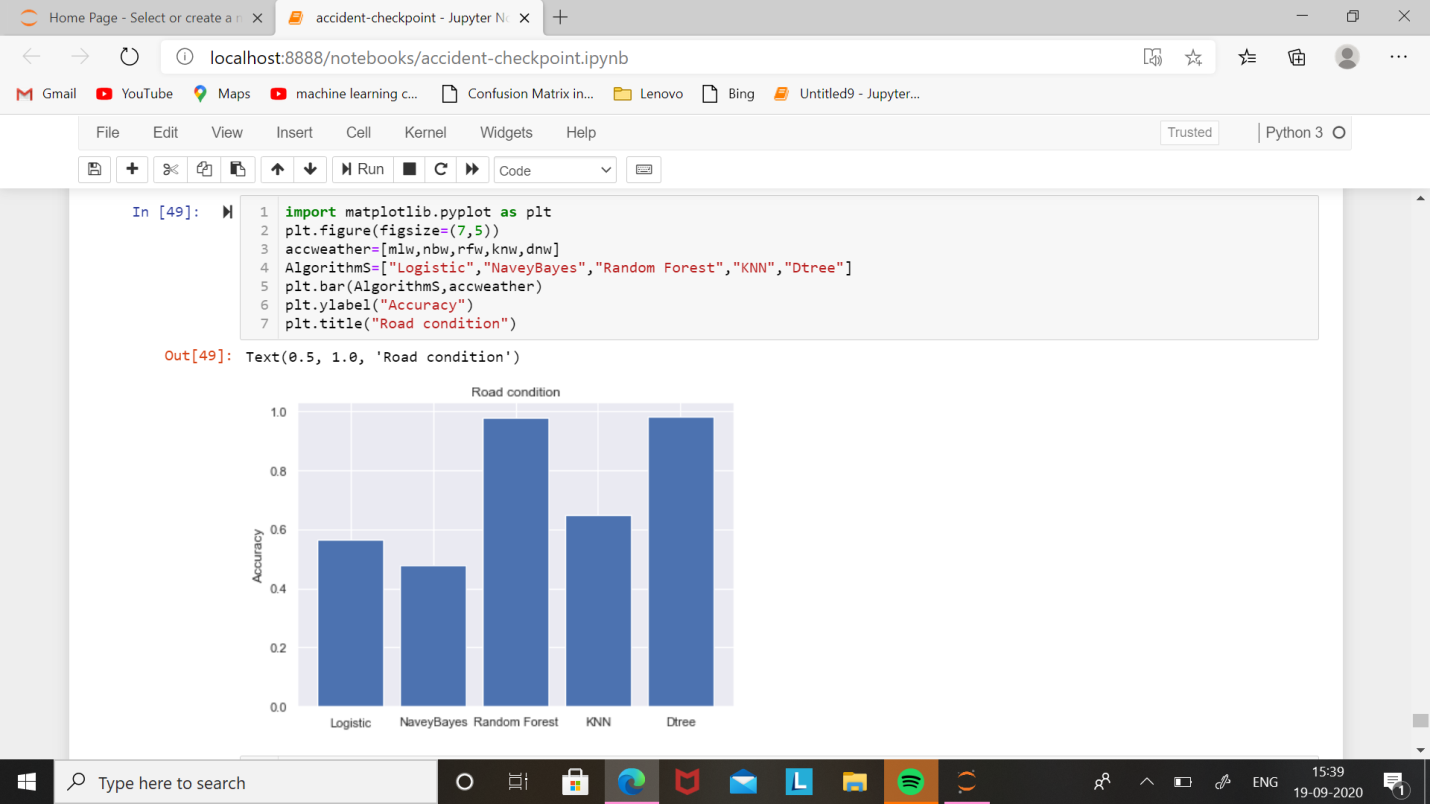
**Fig 9.11: Weather Condition Accuracy with barplot on Matplotlib**

****

**Fig 9.12: Accuracy with Algorithams on barplot in Matplotlib**

****

**Fig 9.13: Road features Accuracy with Algorithams on barplot in Matplotlib**

****

**Fig 9.14: Road condition Accuracy with barplot on Matplotlib**

**CONCLUSION**

**CONCLUSION**

Road Accidents are caused by various factors. By going through all the research papers it can be concluded that Road Accident cases are hugely affected by the factors such as types of vehicles, age of the driver, age of the vehicle, weather condition, road structure and so on. Thus we have build an application which gives efficient prediction of road accidents based on the above mentioned factors.

**REFRENCES**

**REFRENCES**

1. Peden, M. (2004) "World report on road traffic injury prevention". Geneva: World Health Organization
2. M. Chang, L. Y., &amp; Chen, W. C. (2005). "Data mining of tree- based models to analyse freeway accident frequency". Journal of Safety Research, 36(4), 365-375
3. Tesema, T. B., Abraham, A., & Grosan, C. (2005). "Rule mining and classification of road traffic accidents using adaptive regression trees". International Journal of Simulation, 6(10), 80-94.
4. Maze, T. H., Agarwai, M., & Burchett, G. (2006). "Whether weather matters to traffic demand, traffic safety, and traffic operations and flow".
5. Cao X, David MF, Theodore H. Detecting Clusters of Fake Accounts in Online Social Networks. Proceedings of the 8th ACM Workshop on Artificial Intelligence and Security. 2015: 91-101.
6. Buket E, Ozlem A, Deniz K, Cyhun A. Twitter Fake Account Detection. IEEE 2nd International Conference on Computer Science and Engineering 2017: 388-392.
7. Naman S, Tushar S, Abha T, Tanupriya C. Detection of Fake Profile in Online Social Networks Using Machine Learning. IEEE International Conference on Advances in Computing and Communicaton Engineering. 2018: 231-234
8. Sarah K, Neamat E, Hoda M. O. M, . Detecting Fake Accounts on Social Media. IEEE Intenational Conference on Big Data. 2018: 3672-3681.
9. Yeh-Cheng C, Shyhtsun F W. FakeBuster: A Robust Fake Account Detection by Activiy Analysis. IEEE 9th International Symposium on Parallel Architectures, Algorithms and Programming. 2018: 108-110.
10. Myo MS, Nyein NM, Fake Accounts Detection on Twitter using Blacklist. IEEE 17th International Conference on Computer and Information and Information Science. 2018: 562-566
11. Qiang C, Michael S, Xiaowei Y, Tiago P. Aiding the Detection of Fake Accounts in Large Scale Social Online Services. 9th USENIX Conference on Networked Systems Design and Implementation. 2012: 1-14
12. Mauro C, Radha P, Macro S. Fakebook: Detecting Fake Profiles in Online Social Networks. IEEE International Conference on Advances in Social Networks Analysis and Mining. 2012: 1071-1078.
13. R. Kaur and S. Singh. A survey of data mining and social networkanalysis based anomaly detection techniques. Egyptian informatics journal.2016; 17(2):199–216.
14. Y. Boshmaf, D. Logothetis, G. Siganos, J. Ler´ıa, J. Lorenzo, M. Ripeanu, K. Beznosov, and H. Halawa, “´Integro: Leveraging victimprediction for robust fake account detection in large scale osns,”Computers & Security.2016; 61: 142–168.
15. Tsikerdekis M, Zeadally S. Multiple Account Identity Deception Detection in Social Media using Non Verbal Behaviour. IEEE Transactions on Information Forensics and Security. 2014; 9(8): 1311-1321.
16. How fake news and hoaxes have tried to derail jakarta’s election. Internet draft(online).
17. Political advertising spending on facebook between 2014 and 2018 Internet draft.
18. [Online].Available:[https://www.statista.com/statistics/891327/politicaladvertisingspending- facebook by-sponsor-category/](https://www.statista.com/statistics/891327/politicaladvertisingspending-%20facebook%20by-sponsor-category/).2018
19. Statista.twitter: number of monthly active users 2010-2018. Internet draft. [Online].Available:[https://www.statista.com/statistics/282087/number-of-monthlyactive- twitter-users/](https://www.statista.com/statistics/282087/number-of-monthlyactive-%20twitter-users/).2018
20. Cbc.facebook shares drop on news of fake accounts. Internet draft. [Online]. Available:http://www.cbc.ca/news/technology/facebook-shares-drop-onnews-of-fake-accounts-1.1177067.2012