**SCHOOL OF INFORMATION SCIENCES**

**Manipal University, Manipal.**

****

**Data Analysis on Road Traffic Injuries**

**Submitted by**

**Amarendranath Sunkara**

**171046025**

|  |  |
| --- | --- |
|  |  |

**ME “Big Data and Data Analytics”**

**Under the guidance of**

**Sandeep Patil,**

**Applied Cognition Systems Pt. Ltd.**

**Big Data and Data Analytics,**

**School of Information Sciences,**

**Manipal University.**

**INDEX**

1. Abstract

2. Objective

3. Introduction

4. Operational Flow

5. Specifications

6. Experimental Results

7. Classification

8. Plan of action

**1.Abstract:**

Every year we are witnessing around 1.25 million road traffic injuries globally. In this survey we have collected the datasets which are available for a given period. The main Objective of this project is to predict the severity of the accidents based on the specified conditions and to analyse the road traffic accidents which were occurred over past years using data visualization techniques.

**2. Objective**

The objective of the project is to visualize and analyse the road traffic injury data to reach people with best visualization techniques. We are trying to reach the people around the globe and by letting them to known facts, figures and causes for the road traffic injury that has happened.

**Machine Learning:**

The objective of this part of the project is to classify the severity of accidents by classification model, here we are using Decision Tree classifier for multi-class classification on a dataset.

* Multi-class AdaBoost: this uses algorithm for multi-class boosting which is called SAMME — Stage wise Additive Modelling using a Multi-class Exponential loss function.

**Data visualization:**

* The objective of this part of project mainly mains in visual representation of outputs or results that makes ease of understanding to audience what we are trying to convey. We used tools like matplotlib and Highmaps.

**Individual contribution to the project:**

As an individual, my contribution to this project starts from the scratch to delivering the model. I worked for the requirement from the in-house project, right from searching through the data sets which are available. Collected the data set from data.gov.uk. Next step was how to analyse and visual from the dataset. Visualization of the data, I have plotted different kind of charts/graphs which includes Stacked Bar, Multiline, normal pie charts, Bar charts, Boxplot chart, HighMaps. The challenging part was applying Machine learning concepts which suits this dataset to classify the supervised model data. Decision tree classifier is used to train the model and classify multi class classification using SciKit learn packages.

**3. Introduction**

There are a lot of vehicles driving on the roadway every day, and traffic accidents could happen at any time anywhere. Some accident involves fatality, means people die in that accident. As human being, we all want to avoid accident and stay safe. To find out the causes, data visualization techniques could be applied on the traffic accident dataset to find out some valuable information, thus give driving suggestion.

Road traffic safety refers to the methods and measures used to prevent [road users](https://en.wikipedia.org/w/index.php?title=Road_user&action=edit&redlink=1) from being killed or seriously injured. Typical road users include [pedestrians](https://en.wikipedia.org/wiki/Pedestrian), [cyclists](https://en.wikipedia.org/wiki/Cyclist), [motorists](https://en.wikipedia.org/wiki/Driving), vehicle passengers, and passengers of on-road [public transport](https://en.wikipedia.org/wiki/Public_transport) (mainly [buses](https://en.wikipedia.org/wiki/Bus) and [trams](https://en.wikipedia.org/wiki/Tram)).

Classification in data analysis methodology aims at constructing a model (classifier) from a training data set that can be used to classify records of known class labels. The Decision Tree Multi class classifier technique is one of the basic accuracy-based methods for classification.

**4. Operational Flow**

**Data Analysis**

**Data Collection**

**Data Pre-processing**

**Data Visualization**

The above figure shows the operational flow of our project. First, we collected the data sets from the Data.gov.uk. In the second step, we have modified the datasets as per our requirements. In third step, we stored it in database, and wrote queries. From queries we visualised the data.

**Dataset collection:**

This phase deals with collecting the required datasets based on the factors that need to be analyzed in theproject. The dataset collected with respect to road traffic injury. Data is obtained from the data.gov.uk website.

**Data pre-processing:**

Data cleaning refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data. Administratively, incorrect or inconsistent data can lead to false conclusions and misdirected investments on both public and private scales. The data set we collected from data.gov.uk was cleaned data, we did some pre-processing by removing un wanted columns to the limited number as we were concerned with the objective of this project as stated earlier.

**Data Visualization:**

Data visualization helps people understand the significance of data by placing it in a visual context. Patterns, trends and correlations that might go undetected in text-based data can be exposed and recognized easier with data visualization. Primary goal of data visualization is communicating information clearly and efficiently via statistical graphics, plots and information graphics. Visualization is central to advanced analytics.

* The total number of causalities categorized by male and female.
* The total number of causalities categorized by severity of accidents.
* The total number of accidents happened categorized by urban and rural areas.
* Year wise report of accidents happened in a day of week.
* Number of accidents happening based on weather conditions
* Number of accidents happening based on light conditions.
* The total number of accidents happened categorized by age group and gender.
* The total number of accidents happened categorized by severity of accident and gender.
* probability distribution of Speed limits of the vehicles when accidents happened.
* The average number of vehicles which are mostly involved in each accident.
* Plotted the total number of accidents happened by city wise using high maps.

**5. Specification**

The software requirements of the project are as follows:

**Software Requirement:**

• JavaScript (Front-end)

• Python (Business logic)

• Python Packages: NumPy, Pandas, Plotly, MYSQL dB and Matplotlib

• MySQL (Database)

**6. Experimental Results**

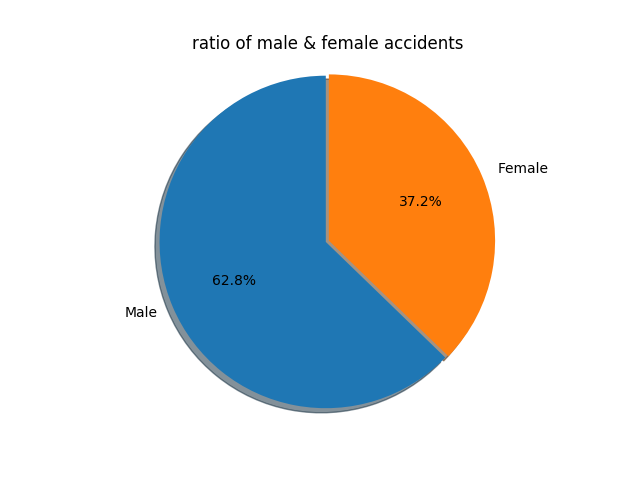


Fig 1. Ratio of male and female

**Gender Ratio comparison**: The percentage of accident happened by gender as shown in Fig 1. The maximum number of accidental death happened to the male.

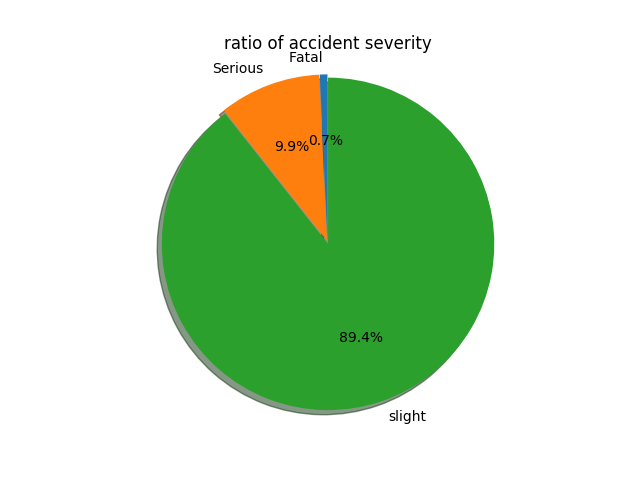


Fig 2. Ratio of accident severity

**Severity Ratio comparison**: The percentage of accident happened by severity of accidents as shown in Fig 2. The maximum number of accidents are happened with the severity slight. This is understandable from the figure that serious accidents are less in number.

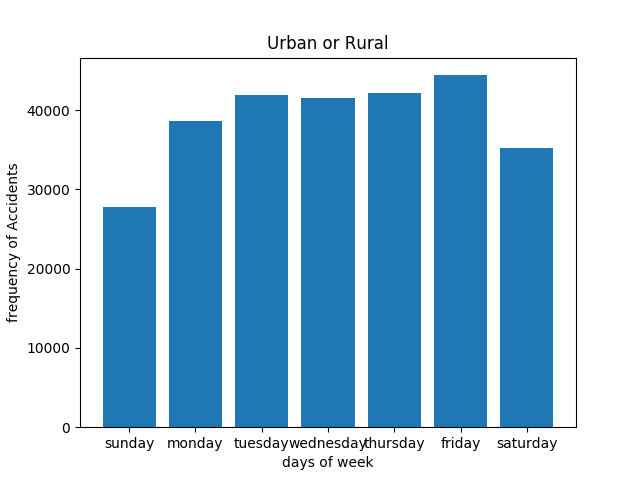


Fig 3. Urban and Rural areas by day of week

**Accidents based on urban and rural areas:** The frequency of accident happened on each day of the week is shown in Fig 3, Most accidents happened on Friday.

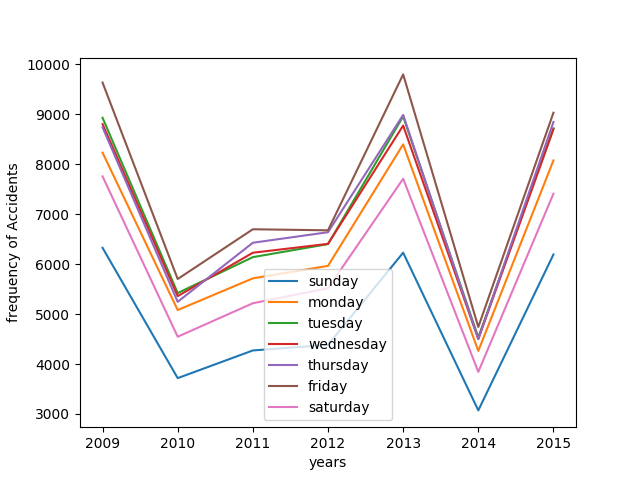


Fig 4. Frequency of accidents over years

**Accidents of each day over years**: The frequency of accidents happened on different day is shown in Fig 4, in comparison with years. Most number of accidents happened on each Friday of the week in the year 2013.

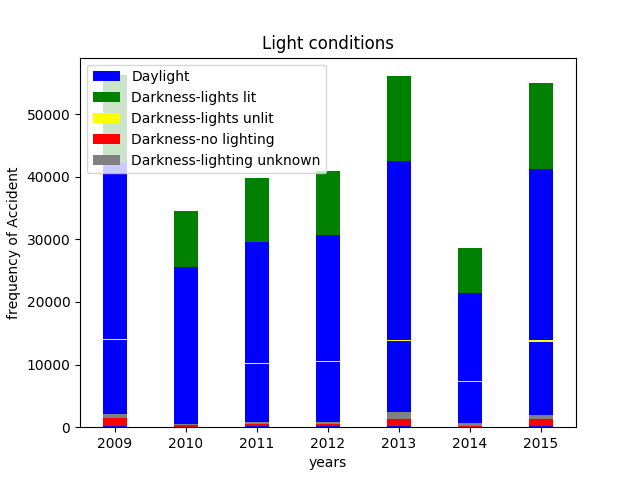


Fig 5. Light conditions over years

**Light Condition**: The percentage of severity of accidents happened on different light condition in comparison over years are shown in Fig 5. Unsurprisingly, most of the accidents happen in day light condition because much more roadway traffic happens in day time other than at night.

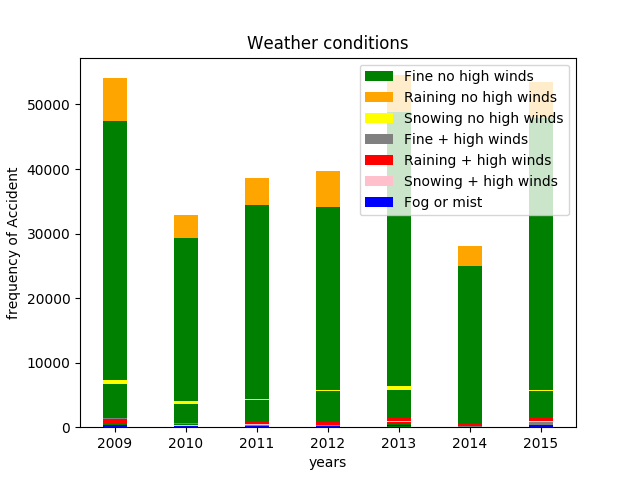


Fig 6. Weather conditions over years

**Weather Condition**: The frequency of accidents happened on different weather is shown in Fig 6, in comparison over years. Most of accidents happened at Fine no wind weather conditions. This is understandable because fine and no winds is the most usual case of weather condition.

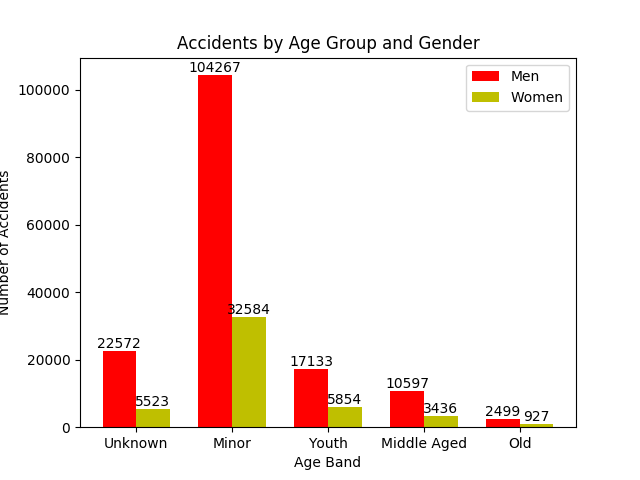


Fig 7. Accidents by age group and gender

**Comparison by age and gender:** The plot represents the frequency of the accidents happened based on the different age groups. The ‘Gender of the driver’ are also classified on different age groups. Here age of the driver is classified into four categories. That is explained as follows. If

Age is less than or equal to 18 years then they are classified as ‘Minor’

Age is greater than 18 years and less than or equal to 40 then they are classified as ‘Youth’

Age is greater than 40 years and less than or equal to 60 then they are classified as ‘Middle Aged’

Age is greater than 60 years then they are classified as ‘Old’

Age is equal to -1 then they are classified as ‘Unknown’.

The Fig 7 clearly illustrates that most of the accidents happened when the driver belongs to a Minor age group and also explains male drivers are involved in more number of accidents compared to female drivers.

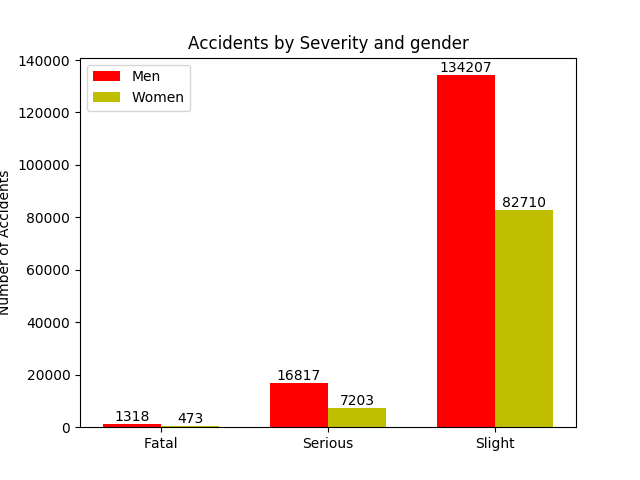


Fig 8. Accidents by gender and severity

**Comparison by severity and gender:** The plot represents the frequency of the accidents happened based on the severity of the accident. The ‘Gender of the driver’ are also classified on among severity of the accident.

In the above figure Fig 8. It is representing that most of the accidents are slight accidents. The number of accidents are high if the driver is male.

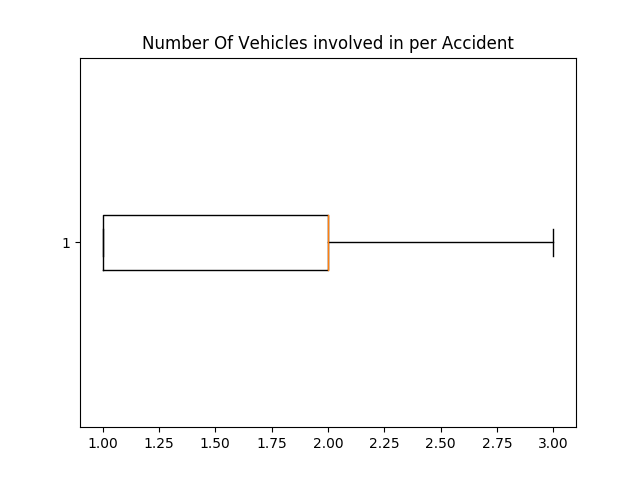


Fig 9. Box plot for average number vehicle involved per accident

**Box Plot:** The box plot representing that the number of vehicles are involving per accident. The following table illustrates the above boxplot. In Fig 9, we can infer that two vehicles are involving in most of the accidents

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Minimum | Quartile1 | Quartile2 | Quartile3 | Maximum |
| 1 | 1 | 2 | 2 | 3 |

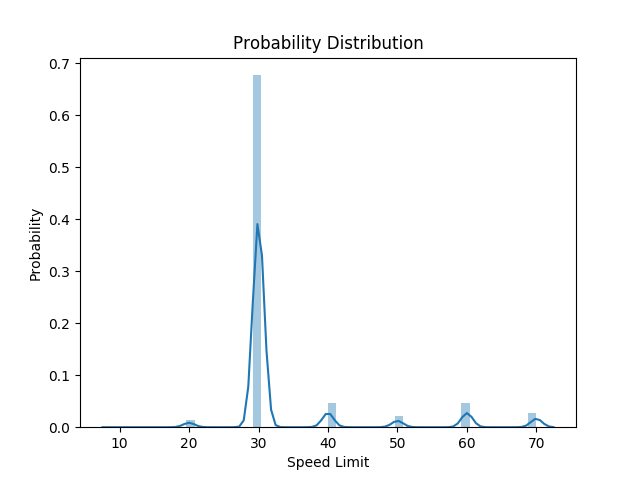


Fig 10. Probability distribution of speed limit

**Speed Limit:** The probability distribution of speed limit illustrating that the probability of accidents are more when travelling on a road of 30 speed limit. The safest roads may be a road with speed limit 10 based on the dataset.

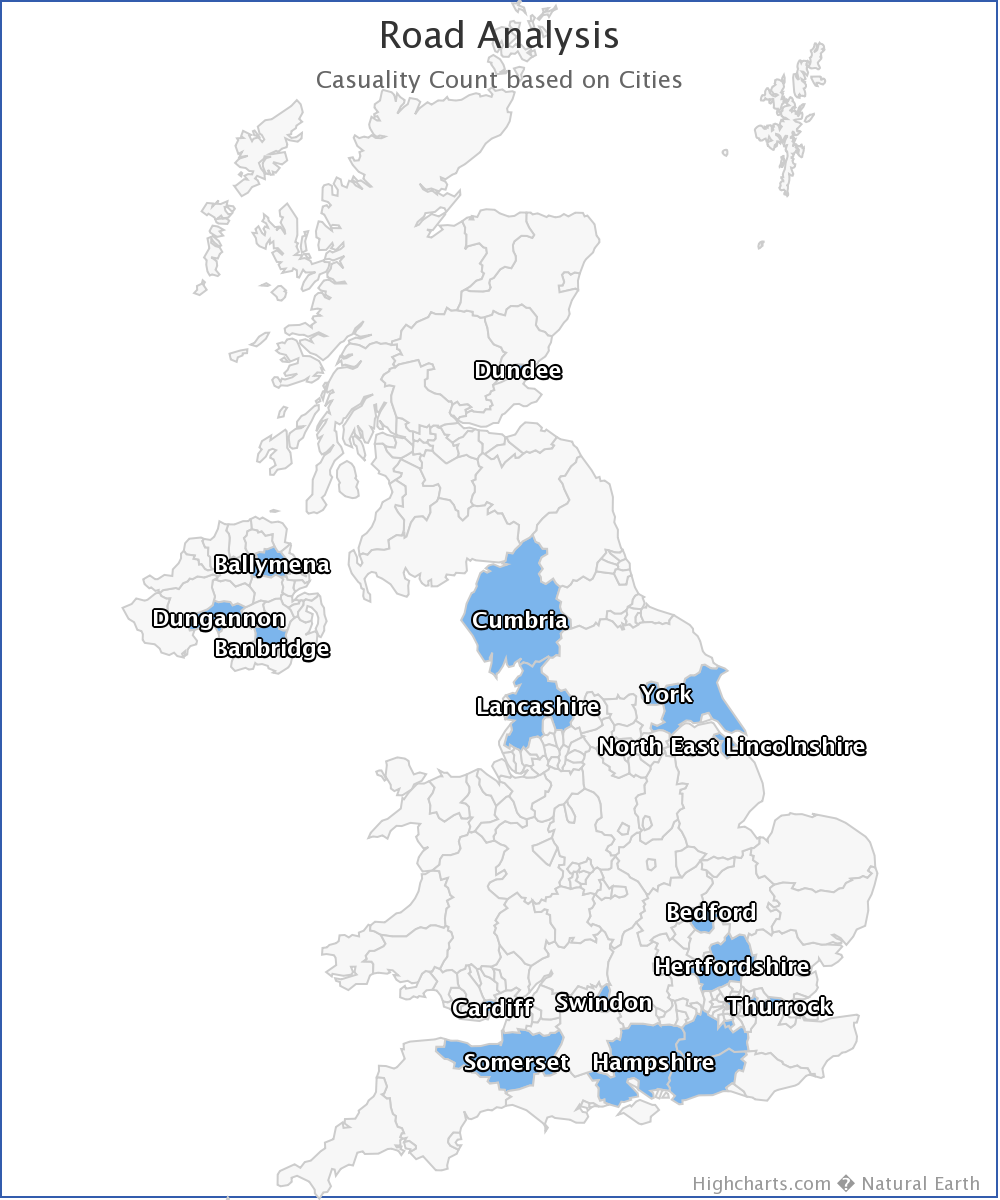


Fig 11. High maps representing city wise accidents

**High Maps representing the frequency of accidents***:* The number of accidents happened on different weather and light conditions and many reasons is shown in Fig 11, is aggregated by city wise so it makes easy comparison with neighbour cities. Most number of accidents happened in a city is highlighted more with colour intensity.

We have the latitude and longitude values in the data set, which are internally converted to the name of the city by google maps latitude and longitude to name converter. Then the output is stored in json format.

**7.Classification**

Boosting is one of the coolest Machine Learning techniques I’ve ever used. Essentially, boosting allows you to build a strong learning by combining the outputs of a set of weak learners. How cool is that?

# Stage wise Additive Modelling using Multiclass Exponential loss (bagging and boosting algorithms)

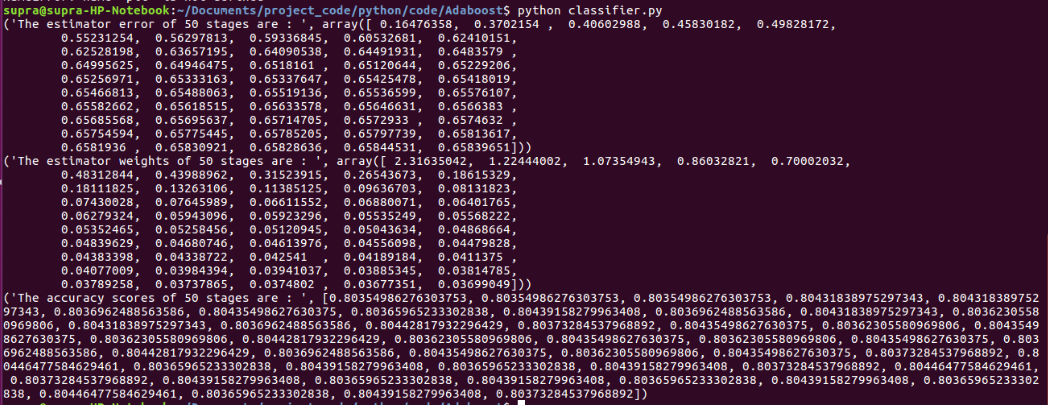


Fig 12. Output of Multi class classification showing accuracy, estimated errors

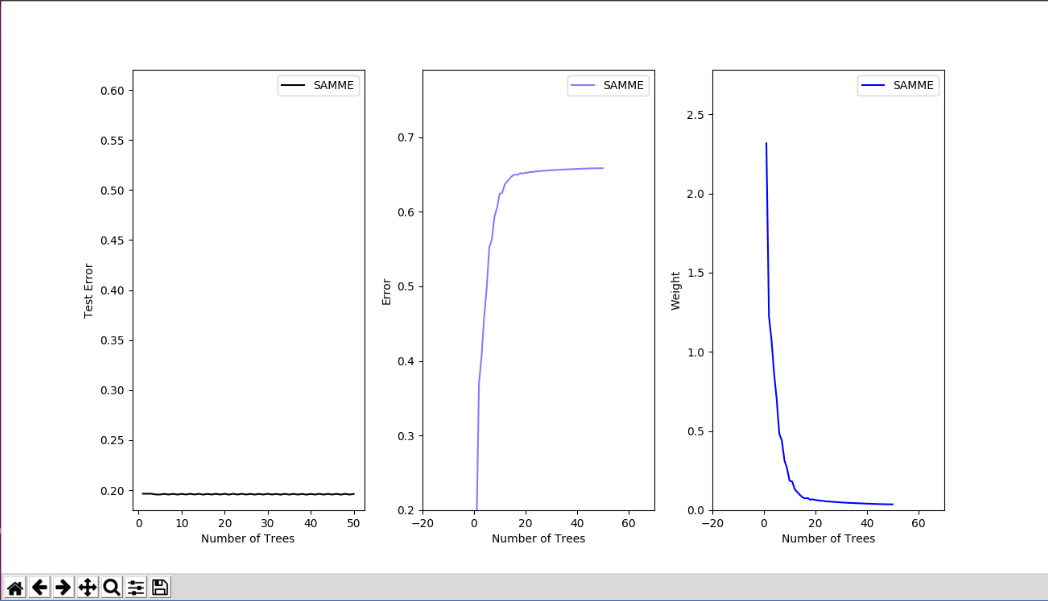
****

Fig 13. Plots showing test error, error and weight

Multi class ADABoost classifier using SAMME algorithm was built on the cleaned data. Of the total 1,37,287 records, out of which 1,09,994(80%) was taken to train the model and rest 20% were considered as test dataset. This model has given a 80.37% accuracy rate.

**8. Plan of Action**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **SL.No.** | **From Date** | **To Date** | **Work** | **Status** |
| 1 | 28th August 2017 | 12th September 2017 | Requirement analysis related Dataset collection. | Done |
| 2 | 18th September 2017 | 24th September 2017 | Data pre-processing and required software installation. | Done |
| 3 | 25th September 2017 | 3rd October 2017 | Setting-up of project, Database creation. | Done |
| 4 | 4th October 2017 | 9th October 2017 | Code work and visualization of data by plotting graphs in python. | Done |
| 5 | 9th October 2017 | 10th October 2017 | Submission of 1st part of project. | Done |
| 6 | 19th October 2017 | 21st October 2017 | Plotting of maps using high charts. | Done |
| 7 | 22nd October 2017 | 28th October 2017 | Classification model using decision tree | Done |
| 8 | 29th October 2017 | 5th November 2017 | Web Scrapping to infer the data. | Half Done, future work. |
| 9 | 6th November 2017 | 7th November 2017 | Final Submission of project. | Done. |