National College of Ireland

Traffic Accidents Prediction Using Ensemble Machine Learning Approach

MSc Research Project
Programme Name

Monisha Lakshme Gowda
Student ID:x18195261

School of Computing
National College of Ireland

Supervisor: Christian Horn

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Monisha Lakshme Gowda		
Student ID:	X18195261		
Programme:	Master's in Data Analytics	**Year:**	2020
Module:	Research Project		
Supervisor:	17/08/2020		
Submission Due Date:	17/08/2020		
Project Title:	Traffic Accident Prediction Using Ensemble Machine Learning Approach		
Word Count:	8873	**Page Count:**	23

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: .....

Date: .....

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Traffic Accidents Prediction using Ensemble Machine Learning Approach

Monisha Lakshme Gowda
X18195261

Abstract

Research paper highlights the significance of various classification strategies in determining the occurrences of traffic accidents that happened throughout the data collection of learning accidents. In this paper, the main focus is to determine the classification using ensemble machine learning models for predicting the accidents when and where it occurs. Historical data for the study is retrieved from the UK transport department website which is open source for the researches. The utilized dataset comes under the huge spatial data for which the clustering is better classification technique. For large number of spatial datasets demands different knowledge regarding the clustering. DBSCAN clustering is better technique as it is capable of clustering the arbitrary shape data which is major requirement for spatial dataset. As the main problem of the research falls in the classification category, but classification requires positive and negative points for classification. The obtained clusters are positive accident spots, hence negative samples are generated by considering the date, time and cluster. The classification model like Random Forest, Ensemble Logistic Regression, AdaBoost Classifier, XGBoost Classifier, Ensemble (DT, SVM, Logistic) for prediction. The Random forest and XGBoost classification performance are pretty good in classifying the accidents based on the place and time when compared to other models. Evaluation metrics explains machine learning models fit better for classifying the accidents. Even though the proposed approach is promising, this study can be improved or extended in future work. Furthermore, in this research can be implemented in the real time traffic accident control concerns for avoiding the future incidents.

Key words: Traffic Accidents, DBSCAN Clustering, Negative Samples, Random Forest, Ensemble Logistic Regression, AdaBoost Classifier, XGBoost Classifier, Ensemble (DT, SVM, LogisticRegession)

1 Introduction

Industrialization and motorization contribute to a rising number of population automobiles and novice road users, leading to an increased road incidents and casualties. Per the study from the World Health Organization, the overall rate of homicides in China has been among the largest in the country around the year 2015. Global official figures show that more and more deadly car accidents take place in highways (Daszykowski and Walczak, 2009). Drivers, lane, automobile and collision characteristics-these are the key factors that influencing the seriousness of the collision. Among some well-famous research methods for analyzing traffic information are machine learning algorithms. Similar approaches can classify the relationship within different parameters that just cannot be specifically calculated utilizing commonplace quantitative mathematical models (Taamneh, Taamneh and Alkheder, 2017) Clustering algorithms make this situation of class findings is interesting. The implementation to wide geographic repositories, furthermore, poses those aforementioned tend to cluster techniques prerequisites:

- Relatively low subject information criteria for deciding the data input, however when working with huge data collection, correct values are sometimes not defined at this stage.

- Exploration of relational data groups, since cluster form in spatial analysis could be circular, sketched-out, dimensional, enlarged, etc.

- Better reliability on massive datasets i.e. larger than some few hundred items on datasets.

These requirements are very difficult to fulfil the above prerequisites. However, DBSCAN is better clustering technique as it is capable of identifying the arbitrary shape clustering (Daszykowski and Walczak, 2009). DBSCAN uses radius to collect the points. It also classifies the noise points if it does not belong to the "Core point" and "Broder point". This process won't stop until the DBSCAN won't check all the points (B *et al.*, 2016). In this paper, DBSCAN technique is used to identify the hot spots of the accidents which occurred in United Kingdom. After clustering process is carried out, machine learning approach is applied as the data consists of numeric values. The classification models like random forest, logistic regression, Adaboost classifier is used to solve the problem statement(Labib *et al.*, 2019). The data set for the research is retrieved Kaggle.com. However, the original source of the data is UK Department of Transport website (data.gov.uk). The license of the dataset is also provided (Open Government License). The data consist of historical traffic accident information from 2012 to 2014. In this dataset district pin-codes is provide, to know the district names another dataset is used which is available in data.gov.uk website. In this paper, implementation of below question is proposed by apply clustering and machine learning approaches.

"Can Supervised models are capable of predicting when and where the accidents in United Kingdom takes places very often?"

So many previous works have utilized traditional case classification system approaches for investigating this challenge. Several of number the researchers have used clustering, machine techniques but the question of classifying important role in the achievement of predicting the traffic accidents remains unresolved. The methods used throughout earlier researches generally have lower measurements or sluggish speeds of handling that weren't really price efficient. The collisions causing traffic were imperceptible especially with the rise in traffic conditions and people and even as a result traffic injury are skyrocketing regularly around the globe. It is therefore very essential to discover the problem of vehicle accident actions through different perspectives and provide a new solution that magnifies several of the limitations of previous novels. The main objective of this research is mentioned below:

- Developing the classification model which is capable of forecasting the traffic accidents with high accuracy.
- Introduce an effective model which is cost effective and also a simple application so that traffic authority can take measures to avoid the accident

The remaining part of paper consist of following sections. Section 2: Previously related work is discussed. Section 3: Explanation of methodology used for the study. Section 4: The architecture of proposed research is explained in this section. Section 5: Implementation of the study is explained. Section 6: Discussion of study is made upon the result. Final conclusion of the research and future work is made in section 7.

2 Related Work

This section of the research investigates the classification and forecasting models for traffic accidents which are not effective. This section is divided into subsections like 1. Prediction using machine learning, 2. Prediction using Deep Learning, 3, Clustering and machine learning techniques for prediction, 4. Conclusion.

2.1 Prediction using Machine Learning

(Shankar, Mannering and Barfield, 1996) This article describes a clustered version of the regression coefficients as a way of assessing the extent of the incident provided that even an incident has taken place. Here various harshness level as been considered. For instance, damages caused during the accidents, injuries happened due to accidents etc. Here multi-nominal regression has been proposed to forecast the accident severities. Seems to provide promising outcomes via the use of parameters like pattern-sobriety, communication and curve-flooring, interplay between surfaces. It appears that ITS can be an efficient way to mitigate adverse nature, individual characteristics and external situations.

(Sameen and Pradhan, 2017) proposed the classification model to analyses the accidents pattern. Traditional methodology has been used to evaluate the model. In classification models, accuracy of the model can be enhanced by implementing the Feature Ranking algorithm, followed by classification algorithms like Random Forest, Decision tree, Naïve Bayes. The relevant analysis of the classification can yield better accuracy is evident in this research.

2.2 Prediction using Deep Leaning

The artificial neural networks performance is better when comes to feature engineering as it is capable of extracting the large amount of data. The neural network models like CNN, RNN and ensemble models are good for predicting using the large dataset. Haitao Zhao et.al proposed the CNN model for predicting the traffic accidents by considering the data which is retrieved from the "Vehicular Ad-hoc Networks". In this paper, neural networks are used to predict the accidents and generate the alerts by sending the alert message of accidents for drivers in real time. The Back Propagation traditional model is compared with the proposed CNN model. As result performance of the CNN model is very high, it also provides the strong analytical base regarding the plan planning optimization and anti-collision (Zhao *et al.*, 2019).

Lu Wenqi et.al (Wenqi, Dongyu and Menghua, 2017) developed a unique TAP CNN model to forecast the traffic accidents by utilizing the various parameters like condition of weather, intensity of traffic etc, which as impact on the accidents. In this paper, different samples have been considered to analyses the accuracy of newly established model. The newly developed TAP CNN model is correlated with traditional model TAP BP by considering the accuracy of both the model. The established model provided better accuracy in forecasting the accidents than traditional model. Even though the proposed model provided good accuracy, there are some drawbacks which need to be enhanced or improved. For instance, accidents may also cause due to the road architecture, size of the road, alignment of the lane this aspect are not considered in this research. Additionally, limited data is test and trained in the model.

The main idea of the Sharaf Alkheder el.at and Al-Radaideh el.at (Al-Radaideh and Daoud, 2018) (Taamneh, Taamneh and Alkheder, 2017) to forecast the severity of the accidents due to traffic jam. The accident severity can be successfully predicted by using classification method like ANN, SVM, Random Forest, Decision tree by analyzing the parameter which has significant impact on the traffic accidents (Lee *et al.*, 2020). Among all the model which used to predict the accidents, ANN model provided better result. In evaluation method traditional evaluation methods like MSE, RMSE to analyses the accuracy. The clustering technique is used to improve the accuracy like K-means and followed by applying the ANN, SVM, Random Forest model(Alkheder, Taamneh and Taamneh, 2017)(Sun and Sun, 2016)(Daszykowski and Walczak, 2009). The accuracy of the ANN model when compared to other model is good which provide accuracy which is completely dependent on the dataset which is used for the model (Alkheder, Taamneh and Taamneh, 2017) but on some dataset ANN model is not able to provide the good accuracy when compared different classification models(Al-Radaideh and Daoud, 2018). By comparing the technique which is proposed by different research it is evident that before applying any prediction model, clustering technique can be applied to enhance the accuracy. Different Clustering algorithm can be used to classify into different clusters like 'Similarity Based Agglomerative Clustering', 'hierarchical clustering', K-means

Clustering. In clustering, the accidents are branched depending on the parameters which influence the accidents the number of clusters can be defined. As the number of clusters increases the accuracy also increases. Hierarchical clustering community information by build- taking a database structure. The whole hierarchy can indeed be developed in two separate forms: clustering - based (bottom-up) where every other piece of evidence constitutes a cluster, which combinations of clusters either combined as it progress up in the hierarchy, or competitive where other trend lines begin through one cluster, and splitting is done as humans progress down the chain of command. The hierarchy clustering is used as it generates the dendrogram which is similar to tree structure (Pettet *et al.*, 2018) (Taamneh, Taamneh and Alkheder, 2017) (Sun and Sun, 2016). In traditional clustering techniques only, numeric values are used to for clustering, but for research it is explained that clustering can be applied on both numeric as well as nominal parameters. It is quite difficult to find the similarity in the both numeric and nominal. This is a kind of analyzing the similarity within the dissimilarity matrix. This process can be carried by using the hierarchy clustering and followed by Bayesian network and ANN model to predict the traffic accidents. The result of both the model are good as the clustering made. Apart from the advantage of the cluster some drawbacks like decreasing of accuracy after hitting the partial large number of clustering. It is uncertain to explain whether the accuracy increases or decreases after reaching the maximum number of cluster(Pettet *et al.*, 2018) (Taamneh, Taamneh and Alkheder, 2017). The traffic accidents also depend on variables like national highway, district roads, accident types etc. When the data is significantly influencing the traffic accidents, then it is easy to carry out the prediction by using the K-means clustering. This technique provides detailed explanation of the accident points when it is visualized by using any simulation tools. This model is also having drawbacks like, difficult to understand the specific pattern of the accidents, explain the similar or dissimilar relationships, classification of unique instance (Kaur and Kaur, 2017).

Most of the research is conducted using CNN model. Apart from CNN, RNN model is also used to predict the severity of the traffic accidents. This model provides better accuracy when it is implemented on the sequential data. To overcome the drawback like vanishing of the RNN, LSTM algorithm is implemented to form dense layer. The compressed layers are then fed to RNN for prediction. To avoid overfitting of the prediction model technique of dropping the layers is used. As result, RNN model yield good accuracy. The model should have been used clustering. As clustering enhance the performance of the model (Sameen and Pradhan, 2017).

2.3 Clustering and Machine learning techniques for prediction

It is uncertain to predict where and when traffic accidents occur. This prediction process requires grouping of regions along with time. This branching can be achieved by implementing DBSCAN algorithm. (Agrawal *et al.*, 2018) studied the DBSCAN algorithm by considering the longitude and latitude of the region, so that clustering can be done on the data. This technique provides the accident region clusters. Once clustering is done predicting model can be applied. (Almjewail *et al.*, 2018)(Daszykowski and Walczak, 2009) conducted research of predicting accidents by using k-means and DBSCAN clustering to identify the hotspots of accidents. The DBSCAN achievement is completely dependent on the intake which is given. DBSCAN is mainly about the grouping the two different dimensions. This clustering is helpful in clustering the geolocation data. K-means is a classification technique, dependent on segmentation. These are centered on clustering artefact's, as well as estimating a co-ordinate meaning. This same coordinate cluster seems to be the square root of both the cluster documents. The ultimate clusters formed by k-means will have a higher cluster formation resemblance as well as less heterogeneity between batches. Both clustering is successfully capable to find the periodic places where actually the accidents used to occur very frequently. Along with the advantage there is a disadvantage like evaluation of this techniques are difficult.

(Daszykowski and Walczak, 2009) proposed a clustering model which consider the DBSCAN clustering as refence to identify the structure of arbitrary. DBSCAN actually needs single scaling factor and assists the consumer in deciding a good estimation for everything. As a result, the DBSACN algorithm performance is better when compared to the newly proposed clustering algorithm. In this research, clustering played important role in checking the density of noise in the cluster. The DBSCAN can be generalized by analyzing the clustering in polygon data. Apart from the advantages of the DBSCAN algorithm as its own drawbacks like algorithm is not capable of cluster the data with various densities(Zhao *et al.*, 2019).(Gutierrez-Osorio and Pedraza, 2020) Such a study gives an overview of both the latest technology in car accident forecasts thru all the data mining approach but instead advanced communication research methods, including such coevolutionary9 deep learning and long-term memory networks, amongst many other learning algorithms. In addition, a new collection and analysis of even the most utilized information sources again for traffic accidents prediction is provided throughout this document. Based on its own sources and attributes, such as open data, measuring methods, board infrastructure and user behavior, a grouping is introduced. The various techniques used to generate forecasts regarding traffic accidents have been described and especially in comparison for both the understanding of the data collected, and also some everyone's enforceability terms of the information gathered, together with the findings acquired along with their allows for easy of observation and understanding.

(Dogru and Subasi, 2015) Research paper simulates a fatal car accident which explores the ability of different classification strategies to identify road accidents. Kinetic energy and location estimates are provided one per automobile, the actions of the automobiles can indeed be examined, and incidents could be identified easily. In some kind of a transportation environment the implementation of the developed architectures is illustrated by projections. Experiments found that information processing software used "DBSCAN and hierarchical clustering" to effectively detect incidents with an overall intelligent surveillance rate of 100 percent and a false negative rate of 0 percent. The K-means, DBSCAN, Hierarchical clustering is compared to analyses the capability of each clustering technique. As a result, DBSCAN performance is way better when compared to the other. Fatal crashes managed to identify utilizing velocity control knowledge, while these strategies are built to recognize clusters of different dimensions like size or shape using DBSCAN. By analyzing the result, it is evident that DBSCAN clustering is capable of analyzing the movement in the traffic.

2.4 Conclusion

The main objective of this study is to be classifying the traffic accidents depending upon the parameters like time and location. Most of the research used clustering technique to group the accidents, specifically DBSCAN clustering is used as the advantage like ability to cluster the arbitrary shape, handling the outliers. The observed part of related work is clustering plays an important role in improving the result of the model. Before training any model, clustering can be applied to increase the effectiveness of the model. Clustering also helps to remove the outliers. As the study is about the classifying the accidents classification model should be applied. According to state of art, the machine learning model are used on numerical data. For classification machine learning models like Random Forest, Binary Logistic Regression, Adaboost classifier, Decision tree, XGboost classifier etc. This research is conducted to propose ensemble approach like Random Forest, Ensemble Logistic Regression, Adaboost classifier, XGBoost Classifier, Ensemble of SVM, Decision Tree, Logistic Regression is used to classify the traffic accident.

3 Research Methodology

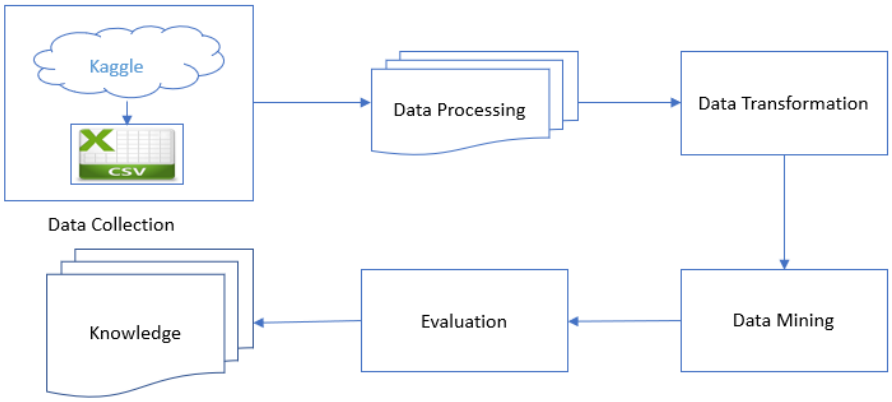

Fig1: KDD Methodology

In this research, Knowledge Discovery in Databases methodology is used to analyse the data set. As KDD guidelines primarily concentrate on the implementation phase instead of just pursuing systems engineering and that it's particularly suitable for identification and Forecast. This data mining method is used to interpret the effect of grouping the accidents using clustering on prediction model accuracy. DBSCAN clustering technique is utilized to make the cluster of the accidents, and machine learning or deep learning model is used develop the accident prediction model. This process consists of following steps: Data Collection, Data Pre-processing, Transformation, Data Mining, Interpretation, Evaluation, Knowledge(Ma *et al.*, 2015).

- Data Collection: Many of the studies has been conducted on the traffic accidents which are based on real time data. The London traffic accident data is retrieved from the official UK transport website. This data is allowed for the public and open source licence is also provided in the website. The data consist of 33 columns and 293779 data entry. To get the district name of London, District Data file is extracted from the same website. Both the data are in CSV format.

- Data Processing: In this step, the two dataset is merged to get the district names using the district code. The data screening should be done in a right manner, so that result is appropriate and not producing the misleading result. The screening should be considered to check missing values, blank values, null values, none etc. By removing the noise data or data which is unreliable, it is possible to obtain the appropriate result.

- Transformation: The column names are renamed for better understanding. The unwanted columns are deleted so make dataset more feasible to analyses. The normalization of data is done to boost the extraction of the data.

- Data Mining: One of the core goals of this study is to classify where the accidents occurs in a particular location of the London. As this research is about classification, ensemble machine learning classification models like Random Forest, Logistic Regression, AdaBoost Classifier, XGBoost classifier, Voting Classification (Ensemble of SVM, Logistic Regression, Decision Tree to classify the traffic accident data.

- Evaluating the Model: The evaluation techniques are applied on the data which are divided training and testing data. Here the data is divided in 7:3 ration. After applying the model, the evaluation of training and testing dataset is made using different evaluation technique. After obtaining the model result, the same is interpreted by following metrics: Accuracy, Recall, Precision, ROC.

4 Design Specification

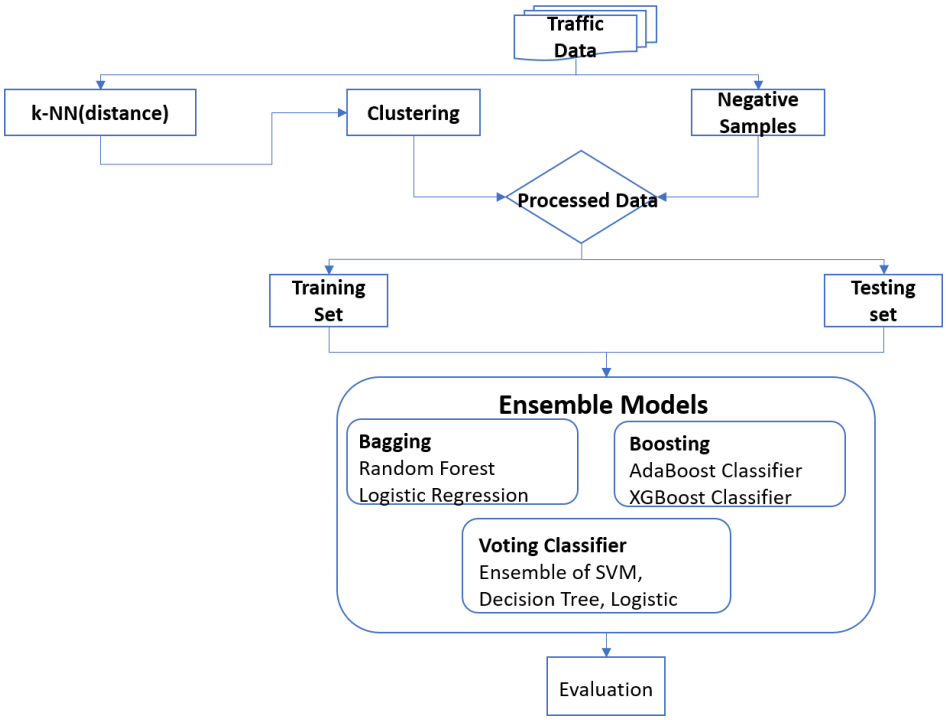

Fig2: Design specification

The design specification diagram explains the adopted design in this research to develop the solution for the traffic accidents. The design is divided into 3 level.

- **Level 1 Architecture**, the retrieved data is processed. Before applying any machine learning model, the required process of data should be made. In this study, processing steps like clustering and negative sample generation is done. The clustering is made to systematically identify and group the traffic accidents happened in different areas of London. K-NN technique is used to identify the distance to define the epsilon in the clustering.

  K-NN Distance: The k-nearest neighbour (k-NN) is traditional discriminant analysis classification algorithm. The k-NN classifier determines the distinctions between some of the position and positions in the training sample to identify an undefined phrase described by certain selected features as just a location in the function space. The feature vector is generally often utilised as a measurement point. The point is then allocated to either the category of its neighbour which is near (T.M. COVER, 2012). Figure 1 depicts the k-NN distance technique.

  DBSCAN clustering: Density-based clustering corresponds to machine learning approaches that classify identifiable gatherings / clusters throughout the data, founded on idea that perhaps a cluster in spatial domain is indeed a clustered large-point intensity area segregated by constituent decreased-point density sections from several other specific clusters(Fong *et al.*, 2014). This technique requires two main parameters min-points and epsilon. Core point is a point where there are m number of points at n distance. Border points is a point where it consists of core point at the distance of n. Noise point which is not core and border point which does not belongs to the group which is present in the figure 4.

Negative sampling: This technique is basically used in classification problem. To avoid the problem of imbalance data. Here random data is generated which is not similar to the original data. This generate data is considered as negative samples.

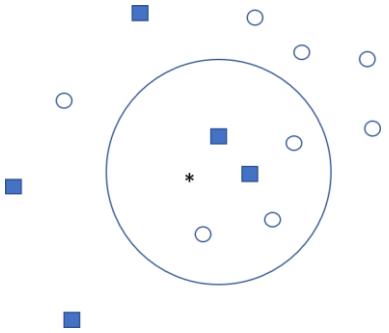

Fig3: k-NN

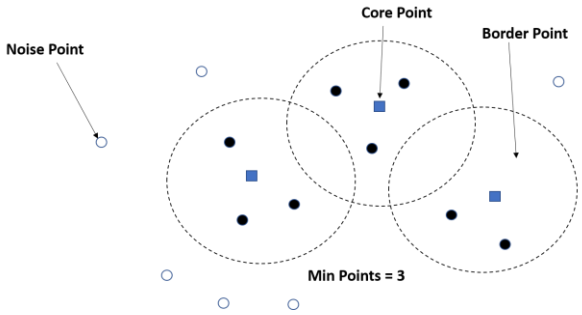

Fig4: DBSCAN Clustering

- **Level 2 Architecture**, the required variables for the predicting is extracted and the same is divided into training and testing data set. After splitting training set consists of 70 percent of data and testing set consists of 30 percent of data.

- **Level 3 Architecture**, once the data is divided the classification of accidents is made by applying the machine learning classification models. The Training set is used to train the model and testing set is used to predicting the accidents. The following classification models of ensemble technique are applied. The main technique of the ensemble models is Bagging and Boosting. Voting classifier is just and approach to decide which model is better among machine learning models.

  Bagging Technique: Another name for the Bagging is Bootstrap aggregation. Here the learning set is split into subset data and encourage to each training sample which are weak.

  - Random Forest Classification: Bagging technology is used in this random forest algorithm. Bagging can be explained as the aggregation of the bootstrap. In this technique, resampling is used in the place of pervious errors. Rather than utilising the bias from either the previous two rounds testing in later phases of measuring as a predictor variables or weight, boosting uses multiple random selection of data throughout the dataset to accommodate several tree(Lee *et al.*, 2020)(Nisbet, Miner and Yale, 2018).

  - Ensemble Logistic Regression: Logistic regression is exceptional situation of linear regression. This model is derived from the huge standard of algorithm called Generalized linear model. Ensemble Logistic the sub samples of the training set are fed to the model more than 2 times to single variables. In this model prediction variable should be categorical and the obtained output will be in categorical. The prediction variable should be in binary value. To stand for binary Categorical performance, forecasts the likelihood of detection of either an activity by applying the logit function to the records. This approach can work Using on issues that weren't ideal for Usage of Linear Regression. In this study, logistic regression is considered to predict the accidents occurrence. Here accident=1 and non-accident = 0(Mathur, Khatri and Sharma, 2018)

  Boosting Technique: Boosting techniques are generally used to encourage the week part of the training set to strong training set.

  - AdaBoost Classifier: This model is also called as “Adaptive Boosting” This classifier model is used along the decision trees which is shorter in length. In this

model each and every samples of training data set are provided with weight. Initially, weight of the instance is set to wi, here wi is weight of the ith train set and n is the value of the nth training set. In the next step, trees are generated and upon each training samples performance is utilised. This process carries out for entire training dataset. The evaluation of the error is done. This error is used to compute the weights. If the prediction is difficult to make, then the weight is increased. If the prediction is easy to make then less weights is provided (Labib *et al.*, 2019)(Nisbet, Miner and Yale, 2018).

$$Weight, w_i = \frac{1}{n}$$

- XGBoost Classifier: The Extra Gradient Boosting is derived from the Gradient Boosting. This approach will generate the new model to predict the errors in previous models and then make a addition of errors to generate the final output of prediction. The technique of this model is "Gradient algorithm" is used to decrease the errors(Nisbet, Miner and Yale, 2018).

Voting Classifier: This technique is used to aggregate the training data which are week by voting the model. There are two voting techniques Hard voting and Soft voting. In hard voting the model which is having majority classes and selecting that model. In soft voting the output is provided in probabilities and the average classes probability of the models are compared and highest probability class are considered as final output. In this research, Soft Voting technique is used(Nisbet, Miner and Yale, 2018).

- Ensemble of SVM, Decision Tree, Logistic Regression
  In this voting classifier, machine learning models like Support Vector Machine, Decision Tree, and Logistic Regression. The probability of classes of all the models is computed and the majority probability class is considered as the output.

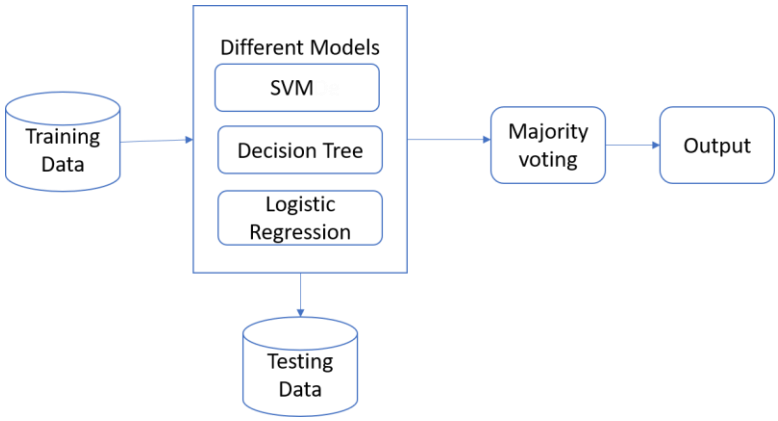

Fig 5: Voting Classifier (Ensemble of SVM, Decision Tree, Logistic Regression)

Evaluation Metrics: After applying the model, the obtained result is evaluated by using the evaluation metrics like accuracy, recall, precision and Roc curve. By analysing metrics, the model performance is explained which model is better to predict the accident. The following equations are used to calculate the Precision, Recall, accuracy and ROC. In equation, TP (True Positive), TN (True Negative), FP (False Positive), FN (False Negative). These values can be obtained by displaying the confusion matrix.

$$\text{Accuracy} = \frac{TP+TN}{(TP+FP+TN+FN)} \qquad \text{Precision} = \frac{TP}{(TP+FP)} \qquad \text{Recall} = \frac{TP}{(TP+FN)}$$

5 Implementation

5.1 Introduction

This paragraph will include explanations of how well the development was carried out in order to achieve competent model for classifying product recommendations dependent on potential efficiency. The step by step methods which are used to classify and predict the traffic accidents is explained. This paragraph will indeed understand the process of abstraction and configuration of the features. In this section details of the dataset are briefly explained. For the development of the model Python Language is used. Three Machine learning models are implemented for prediction and performance is compared to select which model fits better.

5.2 Dataset:

The selection of traffic accident data is a crucial part. The selection of data should be done in appropriate way so that it is possible to obtain the effective result. In this research, traffic accidents recorded by the UK department authority is considered as it is authorised by government. First dataset consists of accident details of entire UK and second dataset consists of pin code of the London districts.

5.3 Data Cleaning:

This step is carried to check whether traffic accident data consists of any missing values, duplicate values, null values, blank values. Presence of unwanted entry may disturb the accuracy or result of the model; hence these values are terminated from the dataset. The dataset consists of 33 columns, but all these variables are not required for the model. These variables are removed from the dataset.

5.4 Feature Engineering:

For the research**,** name of the London Borough's is very important. As the original data doesn't contain name of the Borough's, new column Borough's name is added by extract the data from second dataset. The month, date and year is extracted from 'Date' variable as new columns. The formatting of Time is changed to hour and date format is changed into display the date in dd/mm/yyyy.

5.5 EDA (Exploratory Data Analysis):

This process is used to analyse the dataset completely. In this study, utilization of EDA process is carried to get the insights of traffic data. This step explains the important variables which have direct impact on traffic accidents. It also explains what kind of data is present in the variables and whether the data is normalized or not. To carry out this process different technique is used like visualization, descriptive statistics. The below visualization has been made to analyse the traffic data**.**

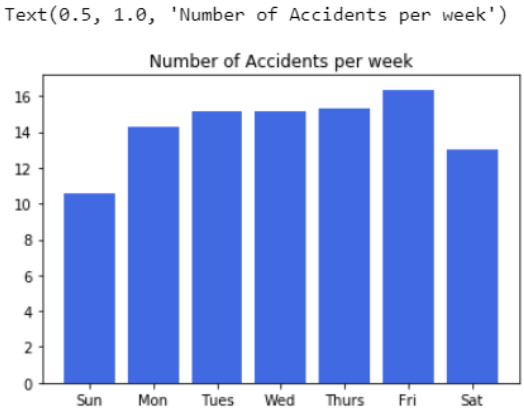

Fig: 6

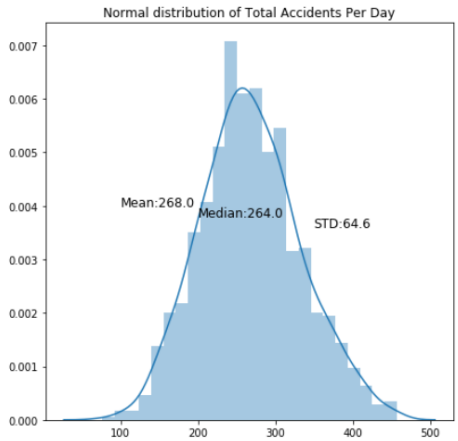

Fig: 7

The analysis of accidents occurred in particular day of the week is made in fig 6. By analysing the bar chart, it can be explained that most of the accidents occurred on Friday. In the other hand, Sunday is having comparatively less accidents. The descriptive analysis is made to understand the normality of the data in fig7. The plotted histogram explains the Mean, Median, and Standard division of the total number of accidents happened in a day.

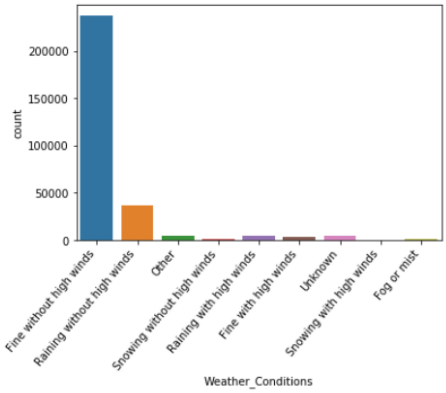

Fig: 8

The data also consists of categorical variables. In the fig 9, the count of accidents is displayed respectively according to the weather conditions. The graph explains most of the accidents occurred when the weather is clear without any high intensity wind. In the other hand, number of accidents in rain, high winds, snowing is less. The correlation between the variables is identified in the fig 10. Plotting heat map is the best way to understand the correlation between the variables. The value of the correlation is displayed in the heat map. If the correlation value is greater than 0.5 then the variable is highly correlated. If the correlation value is less than 0.5 then the variable is not correlated.

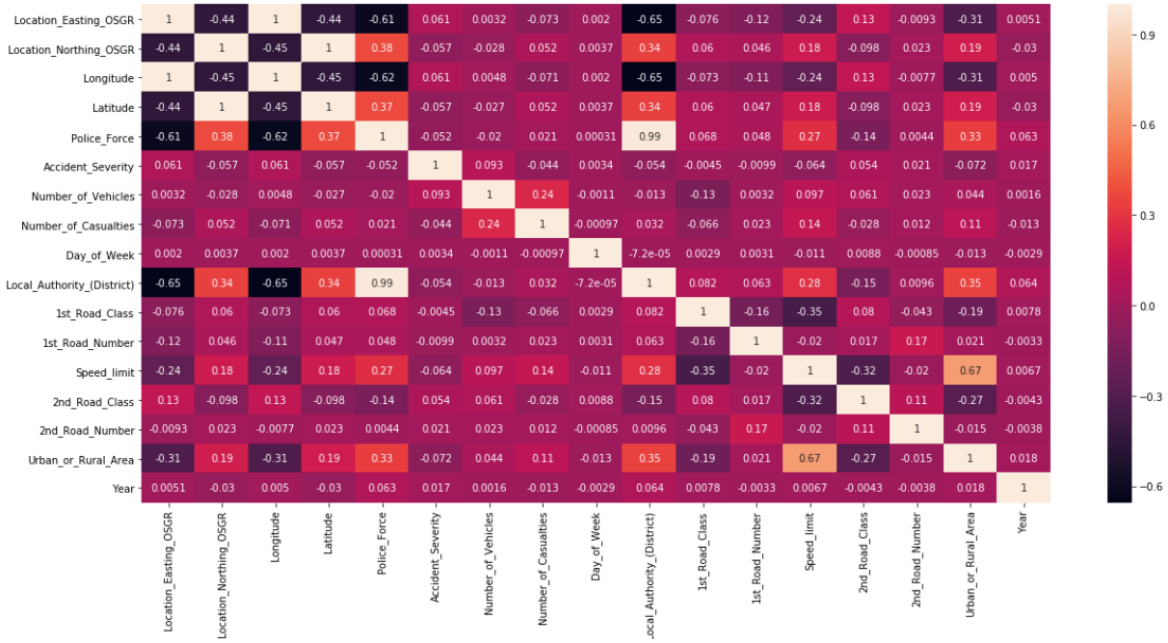

Fig: 10

	Year	Number_of_Casualties
0	2012	156233
1	2013	112313
2	2014	118684

Fig: 11

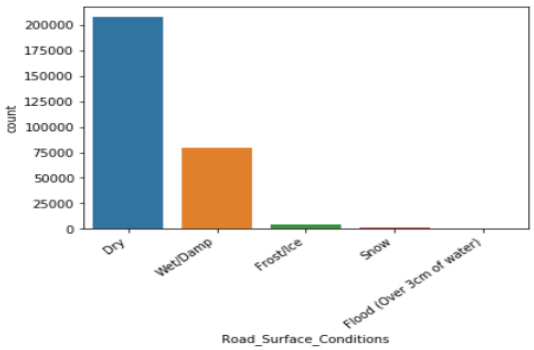

Fig: 12

In the fig 11, total number of accidents per year has been calculated. The total number of accidents was significantly high in 2012. There is not much difference in number of casualties in 2013 and 2014. There is a high possibility of accidents on dry surface rather than wet, frost surface. When it comes to snow surface the accident count is very negligible. Data processing has been done to count the number of accidents which happened on each day of the week along with each month. The calculated accidents values are displayed in the heat map of fig 13. The smaller number of accidents are given light colour background where significant high values are given dark background.

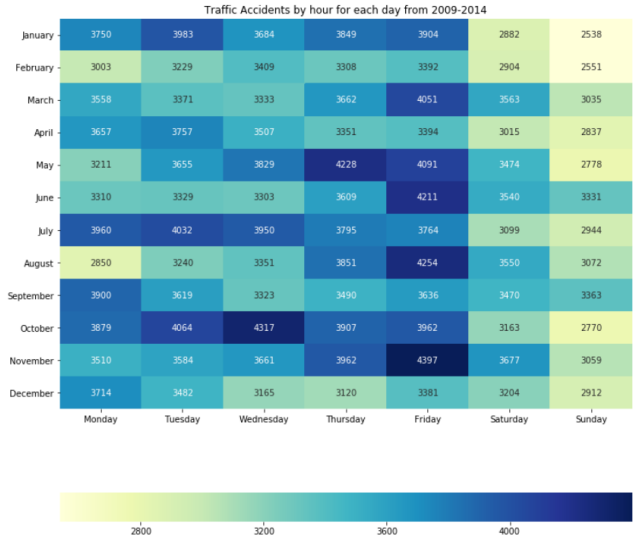

Fig: 13

5.6 Identifying the k-Nearest Neigbor distance :

K-nearestneigboing distance: While appling the DBSCAN clsutering two parameters are very imporatnt. The determination of the Epsilon and min points should be defined. The appropriate value for the epsilon can be identified by using k-NearestNeigbor's distance of the accidents is calacualted by considering number of samples as hyperparametre. Obtained value of epsilon is 0.0033 and number of samples is 4.

5.7 Implementing the DBSCAN (Density-Based Spatial Clustering) :

By analysing the result of EDA process it is clear that the street which are there in London be a site of accident atleast for once. Hence, the statement 'London is indeed a busy city with very regular road accidents (along with obvious characters)' makes sense. While casualties have occurred throughout the region, certain locations seem to be more likely to cause an incidents than most of the others. In this case, the hot spots of the accidents should be identified systematically by using the clustering technique. In this paper, DBSCAN clusterig is used as its processing speed is comparitively high. This clustering is also good in identifying the 'Arbitrarily Shaped' clusters along with the tenacity of outliers. In this paper, DBSCAN to classify city's central incident hotspots. DBSCAN clusters points which are densely packed and classify objects as disturbance outside those clusters. Hence, using this technique, areas where a highest fatality density occurs will indeed be identified as cluster.The places including its incident that fell beyond the clusters are deemed anomalies and removed with our research investigation.

Clustering phase led to 184 clusters with 78497 accidents. While measuring the euclidean distance, DBSCAN needs a measurement to use. To do just that, we must construct a program that takes two equations latitude and longitude and determines the distance within the metres. After finding the distance, the clsutering is applied by using the longitude and latitude. The eps parameter defines the average difference between two measurements for someone to be viewed as in another's neighbourhood. This as well as the min samples factor can indeed be modified to alter the dimension from each cluster and also the amount of locations in it. Points not found inside clusters are marked with the -1 code. The clusters with value -1 are considered as noise. For further process only positive clusters is used. Hence the noise of the clusters is terminated.

5.8 Implementing the Negative Sampling:

By using the clustering technique several point of clustering has been identified which represent the hot spot of the accidents. The hot spots which are identified might not be 100 percent accident spots. Casualty hot spots throughout London's major streets can be

dangerous during the peak periods but slightly safer at slower periods. Through the other extreme, on something like a warm, bright day either a right corner may be relatively innocuous but can become deadly on something like a winter or cloudy day. Thus, it is not clear that on which condition the hot spots are active. Because of this uncertainty the binary classification as became main concentration. This can be solved by checking the combination of different variables like Time, Weather which is actually having impact on activating the hot spot. The classification model requires the positive along with the negative data. The data which is present is only consist of positive samples. The negative samples can be explained as non-accident points which can be achieved as a data. As the study of (Yuan *et al.*, 2017) mentioned the negative samples can be generated randomly. In this step, for every positive samples, the three random samples will be generated by considering the cluster, time, day. negative situations of variations from the three classification groups. The following steps for
The justification to pick certain three characteristics is that their shift will produce multiple every feature are as follows (Yuan *et al.*, 2017):

- Time: If the incident is happened in hour A, then random time for negative sample is taken from [0 – 24] apart from existing hour. This change can also cause the changes in weather
- Day: If the incident is happened on a day A, then random day for negative samples is taken from [1-365] apart from existing day. This may change the feature like time of the accident.
- Cluster: The logic of Time and day is followed for cluster points also.

The non-accidents points generation can be explaining as, if the number of accidents in Ash-Bourne Avenue may be 5, the number of accidents will be incorporated randomly as 15 samples of all spot. During the generation of random samples factors like time, cluster, day of year, Road class, Road number, speed limit, longitude, latitude is considered. The main concentration should be given to make sure that the randomly generated non-accidents point should not have any similarity with the original positive data. The generated non-accidents points are assigned as 0 and accident points are assigned as 1.

5.9 Implementing the Negative Smapling:

By using the clustering technique several point of clustering has been identified which represent the hot spot of the accidents. The hot spots which are identified might not be 100 percent accident spots. Casualty hot spots throughout London's major streets can be dangerous during the peak periods but slightly safer at slower periods. Through the other extreme, on something like a warm, bright day either a right corner may be relatively innocuous but can become deadly on something like a winter or cloudy day. Thus, it is not clear that on which condition the hot spots are active. Because of this uncertainty the binary classification as became main concentration. This can be solved by checking the combination of different variables like Time, Weather which is actually having impact on activating the hot spot. The classification model requires the positive along with the negative data. The data which is present is only consist of positive samples. The negative samples can be explained as non-accident points which can be achieved as a data. As the study of (Yuan *et al.*, 2017) mentioned the negative samples can be generated randomly. In this step, for every positive samples, the three random samples will be generated by considering the cluster, time, day. negative situations of variations from the three classification groups. The following steps for
The justification to pick certain three characteristics is that their shift will produce multiple every feature are as follows (Yuan *et al.*, 2017):

- Time: If the incident is happened in hour A, then random time for negative sample is taken from [0 – 24] apart from existing hour. This change can also cause the changes in weather
- Day: If the incident is happened on a day A, then random day for negative samples is taken from [1-365] apart from existing day. This may change the feature like time of the accident.

- Cluster: The logic of Time and day is followed for cluster points also.

The non-accidents points generation can be explaining as, if the number of accidents in Ash-Bourne Avenue may be 5, the number of accidents will be incorporated randomly as 15 samples of all spot. During the generation of random samples factors like time, cluster, day of year, Road class, Road number, speed limit, longitude, latitude is considered. The main concentration should be given to make sure that the randomly generated non-accidents point should not have any similarity with the original positive data. The generated non-accidents points are assigned as 0 and accident points are assigned as 1.

5.10 Implementing the Model:

After obtaining the processed data is feed to the models. Before feeding to the algorithm the steps like selecting the parameters which are important for the prediction, splitting the dataset into training and testing, and finally applying the model on training and testing dataset. In this study, the important parameters which are selected for the research is day, time, cluster, month, longitude, latitude, day of the week, day of the year. These variables explain the place and time of the accident in London. These variables explain when and where the accidents occur. By predicting the accident using this variable will provide the appropriate solution for the research. Once the variables are selected, the dataset is divided into two parts. One is training set which consists of 70 percent of data and another one is testing set which consists of 30 percent of data. Then training dataset is trained by using the ensemble classification model like Random Forest, Logistic regression, AdaBoost classifier, XGBoost Classifier, Voting Classifier (Ensemble of SVM, Decision Tree, Logistic Regression). In the next step, the trained dataset is used to train the testing dataset for prediction. This process is carried for all the three models. Finally, the accuracy of each model is obtained. The obtained result is evaluated by using the evaluation metrics

6 Evaluation

To explain the performance of each model evaluation should be made on the obtained result. In this section, result of all the models are evaluated to conclude which model is better for predicting the accidents. Table 1 and Figure explains the outcome of the computation conducted to solve the problem of the research.

The Confusion matrix for all the models are displayed in the figure 13, 14, 15. The Confusion matrix of Random forest explains, 4319 observations are non-accidents and classified as non-accidents (True Positi*ve*). 275 entries are non-accidents but classified as accidents (True Negative). 1057 observations are accidents and classified as accidents (False Positive). 498 observations are accidents but classified as non-accidents (False Negative). The confusion matrix of AdaBoost Classifier explains, 4006 observations are non-accidents and classified as non-accidents (True Positive). 588 entries are non-accidents but classified as accidents (True Negative). 1072 observations are accidents and classified as accident (False Positive). 483 observations are accidents but classified as non-accidents (False Negative). The confusion matrix of Ensemble Logistic Regression explains, 4542 observations are non-accidents and classified as non-accidents (TP). 52 entries are non-accidents but predicted as accidents (TN). 1447 observations are accidents and classified as accidents (FP). 108 observations are accidents but classified as non-accidents (FN). The confusion matrix of XGBoost Classifier explains, 4592 observations are non-accidents and classified as non-accidents (TP). 2 entries are non-accidents but predicted as accidents (TN). 1333 observations are accidents and classified as accidents (FP). 222 observations are accidents but classified as non-accidents (FN). The confusion matrix of Ensemble (SVM, Decision Tree, Logistic Regression) explains, 4141 observations are non-accidents and classified as non-accidents (TP). 453 entries are non-accidents but predicted as accidents (TN). 1054 observations are accidents and classified as accidents (FP). 501 observations are accidents but classified as non-

accidents (FN) The obtained confusion matrix values are used to calculate the Accuracy, Precision, Recall.

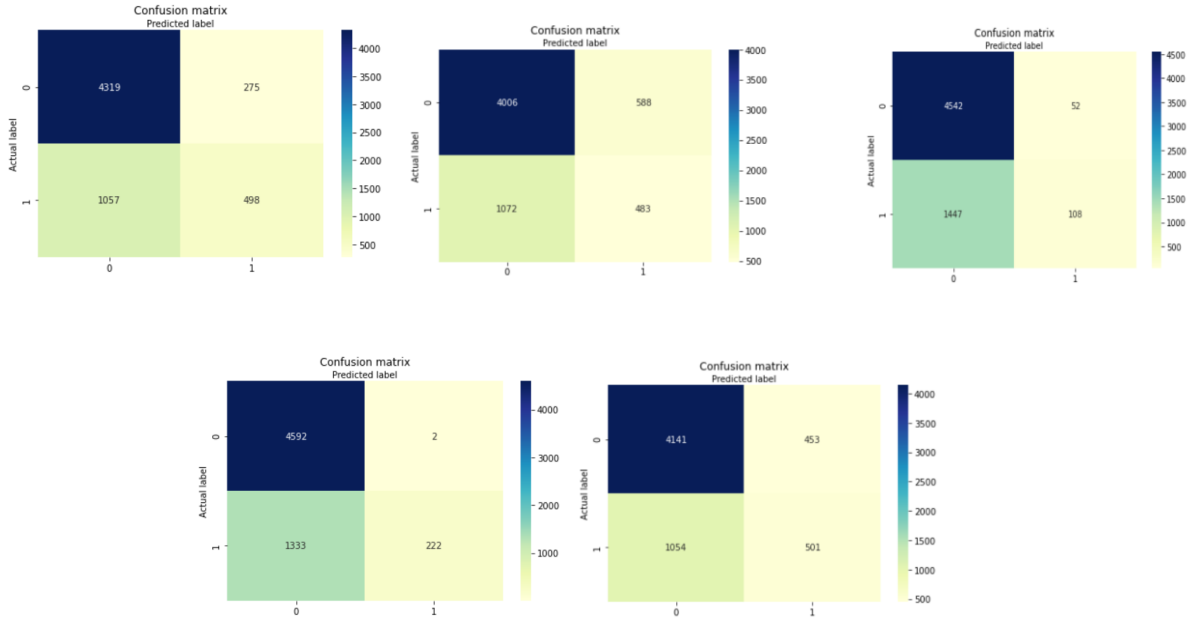

Fig14: Confusion Matrix of Random Forest, AdaBoostClassifier, XGBoostClassifier, Ensemble of Logistic Regression, Ensemble of SVM, Decision Tree, Logistic Regression

Models	Accuracy	Precision	Recall	F1-score
Random Forest	0.782	0.644	0.320	0.427
AdaBoost	0.73	0.450	0.310	0.367
Ensemble Logistic Regression	0.756	0.675	0.069	0.125
XGBoost	0.782	0.991	0.142	0.249
Ensemble(SVM, DT, Logistic)	0.754	0.525	0.322	0.399

Table1: Evaluation metrics for the models

The table evaluation metrics for the models explains the obtained evaluation result of all the three models. The Random Forest and XGBoost accuracy are 78 percent which is better accuracy among all the models. In random forest model the precision rate is about 64 percent which means prediction of positive accident observation is 64% correct when compared total predicted accidents samples. The recall percentage is 32 percent, which means predicting the correct accidents when compared with the actual class. The f1 score is 42% which depicts the mean of precision and recall which is pretty good when compared to XGBoost classifier. Both the result is pretty good. The AdaBoost classifier accuracy is 73%, the precision 45% and recall is 31 percent. The Ensemble Logistic Regression model accuracy is 75%, the precision is 67% and recall is 6% percent. The XGBoost classifier accuracy is 78% which is similar to Random forest, but evaluation metrics of explains that Random Forest is better than XGBoost. For instance, f1-score of random forest is 42% which is better than XGBoost f1-score which is 24%. The Ensemble (SVM, DT, Logistic Regression) model accuracy is 78% and precision is 99% and recall 14%. The f1 score is 39%. By comparing the result of all the models, the performance of the random forest is good when compared to other models.

The ROC curve is also used to explains the performance of the models which is in the fig 16.

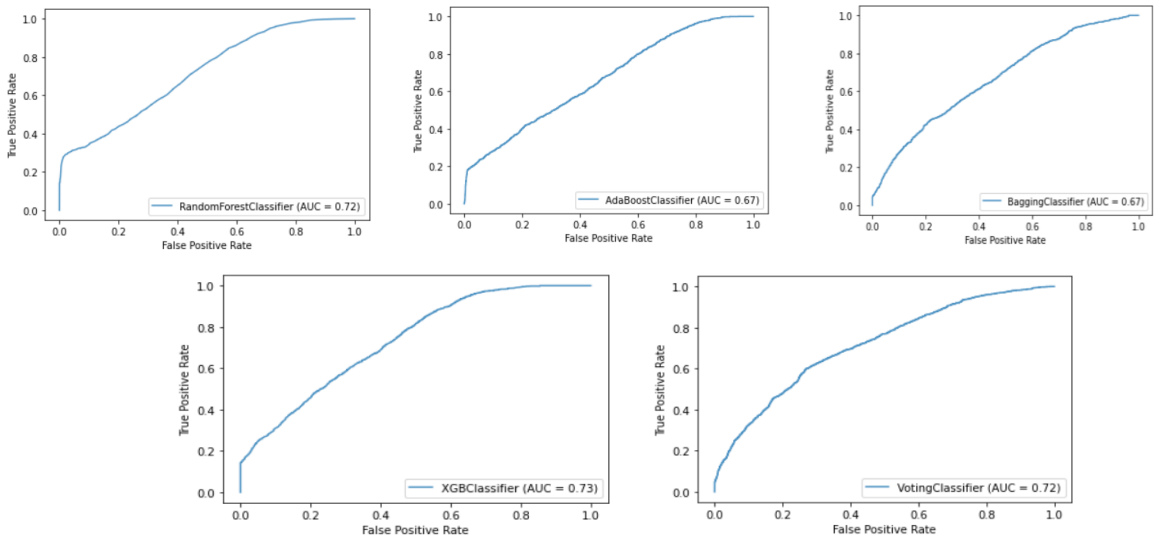

Fig15: ROC of RandomForestClaffisier, AdaBoostClassifier, BaggingClassifier, XGBoostClassifier, Voting Classifier

ROC explains graphically the area covered under the curve of predictive model. Random forest Roc curve value is about 0.72 which explains the model is capable of classification the 70% accidents as accidents (1's as 1's) and non-accidents as non-accidents (0's as 0's). This accuracy of classification is pretty good. AdaBoost classifier Roc curve value is about 0.66 which explains, the model is capable of classifying the 66% of 1's as 1's and 0's as 0's. The Ensemble Logistic Regression model accuracy classification is about 67%. The XGBoost classifier area under the curve is about 73% and Ensemble of SVM, Decision Tree, Logistic Regression model area under the curve is 72%.

7 Discussion

The documentation was obtaining unpredictable just at accuracy of the information happening into even more history. Therefore, the study data were taken into account for restricted three years. The data mining models might be considered relevant mostly with accessibility of even more information collected and indeed the wide variety of knowledgeable prediction might have been continued to increase. Random Forest algorithm may perform better many other types of categorization with both a diverse percentage point throughout this current study. With some of these findings, it is exciting to consider how similar projections of even an uncertain incident feature for less sophisticated classification algorithm might've been produced to practice. The classification models like AdaBoost and Ensemble Logistic regression and XGBoost classifier and Ensemble of SVM, Decision Tree, Logistic Regression also to tried to develop by utilizing the accessible dataset in this study, rather than producing the good result non-acceptable accuracy is obtained. Since the outcomes are calculated during that same measurement model that have become the primary consideration for historically performed work in almost the similar area, the findings in this area of traffic accident forecasting have quite a high degree of trust. Because of the lack of meteorological data, a new parametric analysis just cannot be carried out that would further improve the importance of the whole work in the field of traffic accidents prediction.

The work carried out over the traffic data of certain London districts, because of the restriction of available information. Even the state is having large amount of landscapes. If the possibility of obtaining the weather data along with the accidents of the London city, then by using both accident and weather data would be taken into consideration for the study. This process might take major part in finding the further locations precise of accidents.

8 Conclusion and Future Work

The main focus of the research is to regulate the occurrence of the accidents when and where in the London. By analysing the result of the study, it can be explained that forecasting of traffic accident place and time can be accomplished by using the "Ensemble Machine Learning" algorithms. The proposed approach is suitable for real time solution as classification is made on the basis of the cluster location and time of the historical accidents. Clustering of accidents by analysing the distance of accidents samples plays an important role in optimizing the processing time of model along with the better accuracy by exploring the various parameters which is lacking in state of art. The optimization of the model is very important as it way too informative explanation of the clusters. This research provided the better classification model when compared to previous works. The proposed model lacks in considering density of the traffic whether traffic is high or low. Because, accidents may depend on the volume of the traffic. The used data may not be exact replica of real time traffic conditions.

Current research can be explored and improved by various conditions. As mentioned, there are various parameters which can be considered for the accident classification. Different models can be applied to get better result. Condition of the traffic is completely different when compared to previous years, hence recent data can utilize as it replicates the present traffic situations.

Acknowledgment

I take this opportunity to thank all those who have assisted me throughout my educational career, especially my supervisor Christian Horn for guiding me throughout the research. His guidance as well as recommendation really helped to perform this research.

References

Agrawal, K. *et al.* (2018) 'Ijsrst1848100 |', 4(8), pp. 462–467. Available at: www.ijsrst.com.

Al-Radaideh, Q. A. and Daoud, E. J. (2018) 'Data mining methods for traffic accident severity prediction', *International Journal of Neural Networks and Advanced Applications*, 5(2014), pp. 1–12.

Alkheder, S., Taamneh, M. and Taamneh, S. (2017) 'Severity Prediction of Traffic Accident Using an Artificial Neural Network', *Journal of Forecasting*, 36(1), pp. 100–108. doi: 10.1002/for.2425.

Almjewail, Alaa *et al.* (2018) *5th International Symposium on Data Mining Applications, SDMA 2018*, *Advances in Intelligent Systems and Computing*. Springer International Publishing. doi: 10.1007/978-3-319-78753-4.

Daszykowski, M. and Walczak, B. (2009) 'Density-Based Clustering Methods', *Comprehensive Chemometrics*, 2, pp. 635–654. doi: 10.1016/B978-044452701-1.00067-3.

Dogru, N. and Subasi, A. (2015) 'Comparison of clustering techniques for traffic accident detection', *Turkish Journal of Electrical Engineering and Computer Sciences*, 23, pp. 2124–2137. doi: 10.3906/elk-1304-234.

Fong, S. *et al.* (2014) 'DBSCAN : Past , Present and Future', pp. 232–238.

Gutierrez-Osorio, C. and Pedraza, C. (2020) 'Modern data sources and techniques for analysis and forecast of road accidents: A review', *Journal of Traffic and Transportation Engineering (English Edition)*, (July), pp. 1–15. doi: 10.1016/j.jtte.2020.05.002.

Kaur, G. and Kaur, E. H. (2017) 'Prediction of the cause of accident and accident prone location on roads using data mining techniques', *8th International Conference on Computing, Communications and Networking Technologies, ICCCNT 2017*. doi: 10.1109/ICCCNT.2017.8204001.

Labib, M. F. *et al.* (2019) 'Road Accident Analysis and Prediction of Accident Severity by Using Machine Learning in Bangladesh', *2019 7th International Conference on Smart Computing and Communications, ICSCC 2019*. IEEE, pp. 1–5. doi: 10.1109/ICSCC.2019.8843640.

Lee, Jonghak *et al.* (2020) 'Model evaluation for forecasting traffic accident severity in rainy seasons using machine learning algorithms: Seoul city study', *Applied Sciences (Switzerland)*, 10(1). doi: 10.3390/app10010129.

Ma, Y. *et al.* (2015) 'A data mining model of knowledge discovery based on the deep learning', *Proceedings of the 2015 10th IEEE Conference on Industrial Electronics and Applications, ICIEA 2015*, pp. 1212–1216. doi: 10.1109/ICIEA.2015.7334292.

Mathur, P., Khatri, S. K. and Sharma, M. (2018) 'Prediction of aviation accidents using logistic regression model', *2017 International Conference on Infocom Technologies and Unmanned Systems: Trends and Future Directions, ICTUS 2017*, 2018-Janua, pp. 725–728. doi: 10.1109/ICTUS.2017.8286102.

Nisbet, R., Miner, G. and Yale, K. (2018) 'Model Evaluation and Enhancement', *Handbook of Statistical Analysis and Data Mining Applications*, pp. 215–233. doi: 10.1016/b978-0-12-416632-5.00011-6.

Pettet, G. *et al.* (2018) 'Incident analysis and prediction using clustering and Bayesian network', *2017 IEEE SmartWorld Ubiquitous Intelligence and Computing, Advanced and Trusted Computed, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People and Smart City Innovation, SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI 2017* - , pp. 1–8. doi: 10.1109/UIC-ATC.2017.8397587.

Sameen, M. I. and Pradhan, B. (2017) 'Severity prediction of traffic accidents with recurrent neural networks', *Applied Sciences (Switzerland)*, 7(6). doi: 10.3390/app7060476.

Sun, Jie and Sun, Jian (2016) 'Real-time crash prediction on urban expressways: Identification of key variables and a hybrid support vector machine model', *IET Intelligent Transport Systems*, 10(5), pp. 331–337. doi: 10.1049/iet-its.2014.0288.

T.M. COVER, P. E. H. (2012) 'Nearest Neighbor Pattern Classfication', I, pp. 1–28.
Taamneh, M., Taamneh, S. and Alkheder, S. (2017) 'Clustering-based classification of road traffic accidents using hierarchical clustering and artificial neural networks', *International Journal of Injury Control and Safety Promotion*, 24(3), pp. 388–395. doi: 10.1080/17457300.2016.1224902.

Wenqi, L., Dongyu, L. and Menghua, Y. (2017) 'A model of traffic accident prediction based on convolutional neural network', *2017 2nd IEEE International Conference on Intelligent Transportation Engineering, ICITE 2017*, pp. 198–202. doi: 10.1109/ICITE.2017.8056908.

Yuan, Z. *et al.* (2017) ‘Predicting Traffic Accidents Through Heterogeneous Urban Data : A Case Study’, *Urban Computing*, pp. 1–9. Available at: https://doi.org/10.475/123_4.
Zhao, H. *et al.* (2019) ‘Research on Traffic Accident Prediction Model Based on Convolutional Neural Networks in VANET’, *2019 2nd International Conference on Artificial Intelligence and Big Data, ICAIBD 2019*. IEEE, pp. 79–84. doi: 10.1109/ICAIBD.2019.8837020.