Road Accidents Analysis Using Comparative Study & Application of Machine Learning Algorithms

ARNAV SAINI, NIPUN GAUBA, HARDIK CHAWLA, JABIR ALI,
SRMIST, Delhi-NCR Campus, Ghaziabad, INDIA

Abstract- Traffic Collisions are one of the major sources of deaths, injuries & property damage every year. Road accidents are one of the most difficult real world problems to tackle with, due to its high order of unpredictability. The persistence as well as existence of this problem may be prevalent to a different degree for each & every place. The consequences of this may result in loss of human life & capital. To avoid this, every place needs to tackle the problem with a customized approach depending on the causes that are responsible for the accidents. Even in today's world, where the mass operation of autonomous vehicles is still grim or out of sight, the possibility of predicting a road accident before it takes place, is practically impossible. The only idea or approach that can help to decrease the number of road accidents, is to analyze the reasons that lead to these accidents. The concepts of Data Analysis, Data Visualization & Machine Learning help to tackle real world problems, by exploring & deriving valuable insights, which in turn help in taking measures to solve the targeted problem & drive business growth. In this research study, the dataset pertaining to road mishaps that occurred in UK over time period 2005 - 2015 will be analyzed using these concepts. The defined approach can help the concerned authorities & respective government, to take every possible step & amendment, & hence mitigate the identified causes & scenarios that lead to road accidents.

Keywords: Data Analysis, Data Exploration, Data Cleaning, Data Exploration, Machine Learning, SVM, Random Forest.

Received: March 4, 2021. Revised: July 9, 2021. Accepted: July 12, 2021. Published: July 21, 2021.

1. Introduction

The rapid development & progress of the automobile industry has proved to be extremely comfortable to mankind & is also responsible for several economic benefits & greater communal prosperity. However, the recurrence of traffic accidents has also proved to be quite detrimental to the human society. According to the statistics of **WHO [1]**, road traffic crashes lead to the deaths of 1.35 million people annually worldwide. Another 20 to 50 Million people were victims of non-fatal injuries that often result in some form of permanent or temporary disability. Furthermore, most countries suffer a loss of 1% to 3% of their total GDP due to road accidents. People, their family & nations as a whole suffer massive economic losses due to road accidents every year.

The research work carried out by **Chen. and C. [2]** was primarily focussed on consolidating the data for traffic accidents that took place in Shanghai. The geographical data was converted to longitude & latitude numerical values, & the road accidents data was superimposed on the map of Shanghai. For this purpose they used "Remap", which is a powerful tool that helps in visualizing data obtained from a dynamic map & is based on the JavaScript plugin called "Echarts". An accident heat map was generated which helped them identify the distribution of cluster of accidents & potential hotspots over the city. **Chen. and C. [2]** also made use of various machine learning algorithms, namely, Fisher Linear Discriminant, Random Forest & Bagging Decision Tree, to predict different types of accident occurrence probability. They used K fold validation & the results attained by the machine learning models were compared on the basis of Accuracy & Kappa Values.

Many researchers have tried to investigate the rate of occurrence of road mishaps & the role of various factors that contribute to it. According to one such study conducted by **Saranya Krishna et al. [3]**, there exists a lot of potential in the analysis of massive data that was related to transportation & accidents. Also,

they approached to find if there existed a concealed relationship between the various parameters of data that could be used to draw useful conclusions. They were of the opinion that accident data sets could be used to find & address the root problems of traffic mishaps which mainly caused accident fatalities & road blockage. The patterns & conclusions drawn from their study could assist law makers & professionals to enhance the current transportation system in a sophisticated manner & enforce new stricter rules. Their study was also able to reveal some of the common misconceptions that currently exist about road accidents. It can be inferred from their study that flow of traffic & decisions related to safety are strongly affected by human conduct. They concluded that age & sex which are driver related characteristics that were previously thought to be inconsequential could be predicted accurately up to 70% if other characteristics related to a casualty or accident were provided under the required labels.

Another research undertaken by **Ismael et al. [4]** focussed on how the weather conditions could affect the probability for occurrence of an accident. The research was confined to Greater Manchester in UK & data pertaining to road mishaps over the period 2011 - 2015, was consolidated from the records maintained by police & administrative authorities. Focal point of the analysis in the research was to find how the weather conditions could have an effect on driving behaviour on wet roads & dry roads, & how road safety can be handled at bend roads. The association between road surface & road accidents was examined using chi square statistical method.

On the other hand, the studies conducted by **Jian Zhang et al. [5]**, indicated that crash injury or accident severity prediction was also an encouraging area of research in traffic or road safety. According to them, many traffic safety researchers had shown substantial interest as well as concern regarding an increased effort that needs to be put in for creation of crash injury or accident severity prediction models. If such a model could be successfully developed & tested with sufficient prediction accuracy, hospitals would be able to provide/ ensure spontaneous & appropriate medical care. Additionally, by studying the intensity with which accidents could occur, an attempt can be made to identify & understand the causes that contribute to the severity of an accident injury. This could also help us reduce the intensity with which accidents occur in the future & also improve general traffic safety. They were able to measure the intensity of accidents by various distinct categories that included: damage caused to property only, potential injury, non-incapacitating injury, incapacitating injury & lethal injury.

Conventionally, various statistical methods have been used for modelling accident severity. However in recent years, due to their good rate of prediction, methods based on machine learning concepts have become quite popular. In their paper, **Jian Zhang et al. [5]** aimed at comparing the prediction results & accuracy metric of various statistical & machine learning methods for analysis of accident severity. Their study also incorporated & took into account the role of various variables/ factors that lead to road mishaps, & ensured that each method had a different modelling logic. They estimated that Multinomial Logit Model & Ordered Probit Model were the most frequently used statistical methods & Random Forest, KNN, Decision Trees & SVM were the most frequently used machine learning methods. The overall prediction rate they had calculated for each individual accident severity level was quite accurate. Their results indicated that the predictions made by machine learning based methods were more accurate than the predictions made by statistical methods. It should be noted however, that overfitting was a recurring issue in machine learning based methods. It was concluded that the best overall accuracy was given by Random Forest method & the weakest was made by Ordered Probit method.

In this comparative study, Road Accidents, which is one of the major concerns in today's world, will be analyzed using Data Analysis & Visualization. The main focus of our work is to explore the various causes & scenarios that lead to road accidents, identify valuable insights & important trends, & generate representations from the data. The dataset that is used comprises of various features & attributes about road accidents, which took place in United Kingdom, over time period 2005 – 2015. Different features from the data will be enlisted, for possible application of Machine Learning algorithms & models, which could help in predictive analysis & generation of correlation matrix & confusion matrix, which in-turn would help in understanding the underlying distribution of the data & calculation of performance metrics like Accuracy. The analysis of the dataset is organized into five phases, namely, Business Understanding, Data Acquisition & Data Understanding, Data Exploration

& Data Cleaning, Feature Engineering & Feature Selection, & Machine Learning. Also, we have used Microsoft Power BI, which is an advanced data analytics & visualization tool, to generate high level visual representations. The results obtained in our study can be used by organizations to analyze possible situations that lead to loss in human life, & come up with solutions that can mitigate such scenarios.

2. Data Source

The data is soured from the UK Government's Data Repository, & is published by the Department of Transport. The dataset comprises following files:

- AccInfoData.csv: Every record present in the file pertains to a unique traffic accident (identified by the "Accident_Index" column). The columns refer to the different features related to the road accidents. [Year Range: 2005 – 2015]
- VehInfoData.csv: Every record present in the file refers to the involvement of a unique vehicle in a particular traffic accident. Different features related to vehicles are represented by the columns. [Year Range: 2005 – 2015]
- Casualties0515.csv: This file consists of data related to casualties that occurred due to the road accidents. In this, each casualty is uniquely identified by the column "Accident_Index", identifying the particular road accident which resulted in the particular casualty. The columns represent elaborated details about each casualty. [Year Range: 2005 – 2015]

The column "Accident_Index" is the column which can be used to link the above mentioned csv files.

3. Methodology

In this study, four machine learning algorithms were applied, namely KNN, Random Forest, Support Vector Machine & Multiple Linear Regression (MLR) [6-9]. Amongst these, the first three algorithms are used in classification problem for the attributes "Accident Severity": accidents are classified into different categories based on the severity, "Rural or Urban": accidents are classified into two categories based on whether the accident took place in a rural area or urban area, & "Sex of Casualty": casualties are classified based on gender. MLR is applied for prediction of the attribute "Age of Casualty", which is the dependent variable. All the machine learning models applied in our study are supervised in nature, i.e. they are trained on a labelled dataset.

3.1 KNN

K Nearest Neighbors, an algorithm which is supervised in nature, uses the labelled sample data to train & develop a function, & generates the most suitable output/ response when it's provided with new data which is unlabelled. KNN is usually used for classification problems, but it can be used for regression as well. This algorithm presumes nothing regarding how the statistical information is actually distributed underneath; hence it's non-parametric in nature. Also, it doesn't make any generalizations, which means that it keeps all the training data & uses all of it during the testing phase, therefore it's termed as a lazy learning algorithm.

KNN stocks all the labelled input data for training. The classification of the new unlabelled data depends on the "k" nearest observations, which is decided using a similarity measure like a distance function. Therefore, the classification of the new data point is based on nearest set (k) of previously classified data points. The main parameters that determine the performance of the algorithm are appropriate selection of distance function which is taken into consideration & "k" value. The value used for parameter "k" is resolved using hit & try method, & observing which value gives the highest accuracy. The minkowski distance with power parameter 2, which is equivalent to the Euclidean distance, is used as the distance function.

The formula used for calculation of minkowski distance between two data points O & R is as follows:

$$Dist\ (O,R) = \left(\sum_{j=1}^{n} \quad |x_j - y_j|^z \right)^{1/z}$$

Where z = power parameter.

3.2.Random Forest

Random Forest (RF) is an ensemble learning method & a supervised algorithm, in which a large number of Decision Trees work together as a group to generate the most suitable prediction results. These decision trees act as the building blocks of a forest. At every node, the decision tree searches through the features to find that particular value that splits the tree in a clean & the most efficient way. The criteria for splitting the nodes can be either "Gini Impurity" or "Information

Gain". The process of splitting is recursively repeated till the tree reaches a maximum depth, or the leaf node contains an output or samples from the same class. The RF algorithm will use all the results generated by every tree which is part of the forest, to generate the final prediction. Each individual decision tree will generate a class prediction for the input. The algorithm will use the output of all the decision trees, & the class with the most votes will be the predicted class. Some of the advantages of using RF is that, it prevents over fitting, requires less training time, can efficiently handle large & multidimensional dataset, & maintains high accuracy even when the dataset consists missing values. The two most important parameters that may determine the performance of RF, are the criteria which is used for judging/ computing the quality of the split & number of trees. As part of our study, the criterion used for splitting is "entropy", which is used in calculation of information gain. The feature that is found to produce the most valuable insight is selected as the decision node.

The formula for Gini Impurity is as follows

$$G = \sum_{l=1}^{F} p(l) * (1 - p(l))$$

Where F is the total no. of classes & p (l) is the probability of picking a data point with class "l".

The formula for Entropy is as follows:

$$E\ (S) = \sum_{l=1}^{f} - p_l p_l$$

Where S denotes sample of data, f is the total no. of classes & p_l is the probability of picking a data point with class "l".

The formula for Information Gain is as follows:

$$Gain\ (M, N) = Entropy\ (M) - Entropy\ (M, N)$$

Where M is the target & N is the particular feature for which information gain is being calculated.

3.3 Support Vector Machine

SVM works on transforming the data to n dimensional space for finding an optimal hyperplane between the different classes of data points. As part our study, we have used it for classification, as it produces significant accuracy & requires less computation power. In SVM, the algorithm plots each data item in an h – dimensional space, where "h" specifies the no. of features. The objective of the algorithm is to find the best possible hyperplane which can classify the given data points with the utmost accuracy. There may exist multiple possible hyperplanes for this purpose, but the algorithm needs to be tuned to find the most optimal one. The aim of SVM is to identify that particular hyperplane which results in the max. margin or the max. possible distance b/w the inputs of the two groups. This is the best hyperplane, which is known as the decision boundary or Maximum Marginal Hyperplane (MMH). The data points from both the clusters which lie nearest to the hyperplane on either side, helps in determining the maximum margin. These data points are termed as Support Vectors.

For linearly separable data, a single straight line can be used as the hyperplane. This is known as Linear SVM. But if the data is nonlinear, a straight line hyperplane cannot be used. For this, a tuning parameter "kernel" is used to map the data to another dimension so that a nonlinear optimal hyperplane could be found. Different types of kernels are used in SVM, namely polynomial kernel, linear kernel, rbf or "radial basis function" kernel, sigmoid kernel etc. As part of our study, we have used "rbf" kernel.

Different tuning parameters for SVM include Kernel, Regularization & Gamma. Regularization (C) specifies the degree of importance that is given to misclassifications or the degree to which the misclassification of training examples should be avoided. If the value of C is kept high, a small margin hyperplane will be chosen if it's able to do a better job at classifying each data point correctly. On the contrary, if the value of C is very small, the algorithm will search for a hyperplane which generates a larger margin, even if it leads to higher misclassification of the data points.

The Gamma parameter is used to define the distance upto which the data points should be considered for finding the separation line. High value for Gamma denotes that only nearby points are considered. Low value for Gamma means far away points are also considered. A tradeoff is made between all the tuning parameters to determine which set of values for the parameters give the best accuracy or performance.

3.4 Multiple Linear Regression

A regression model usually works on the basis of fitting a line to the observed data using the identified &

described relationships among the given variables. It helps us to keep a track of how the dependent variable changes as independent features/ variables change. In Simple Linear Regression (SLR), only a single independent variable is worked upon to foretell the value of the selected dependant variable. The relation between the single independent variable & single dependent variable can be represented using a straight line in two dimensional space. MLR uses multiple number of independent variables (2 or more) to forecast the value of the dependent variable. MLR aims to represent the linear relationship b/w multiple explanatory independent variables & response or dependent variable. In our study, we have used MLR for prediction of the variable "Age of Casualty".

The assumptions made while using MLR are that there exists a linear relationship between independent & dependent variables, the independent variables are not very highly correlated to each other, no hidden relationships exist among variables, residuals must be normally distributed & there should be no significant outliers present in the data.

The mathematical eq. for Multiple Linear Regression is given below:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \in$$

Where, y_i = dependent variable

x_i = explanatory variables, β_0 = y-intercept (constant term)

β_p = slope coefficients for each explanatory variable

$\in$ = the model's error term (also known as the residuals

4. Results

4.1. Data Visualization

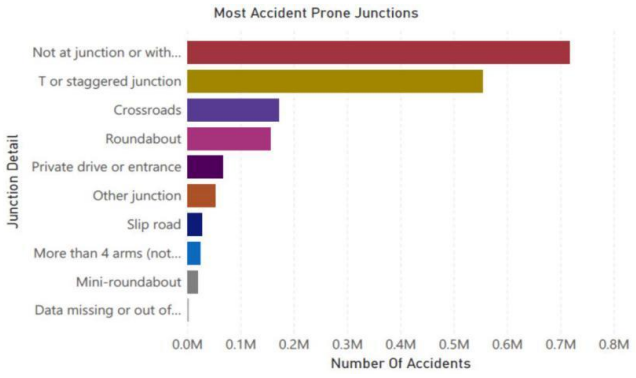


Figure 1: Accidents by Junction Detail

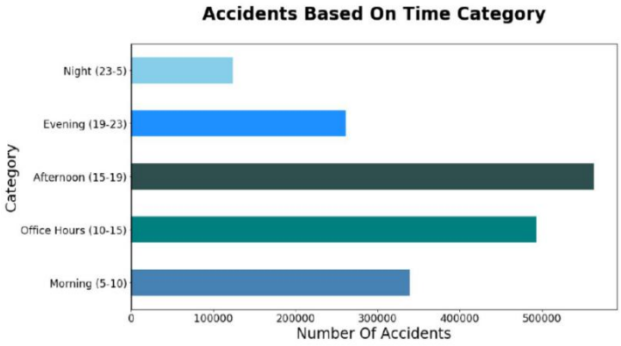


Figure 2: No. of Accidents Based on Time Category

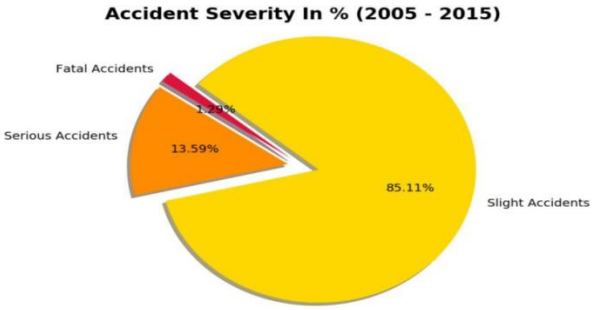


Figure 3: Percentage of Accidents by Accident Severity

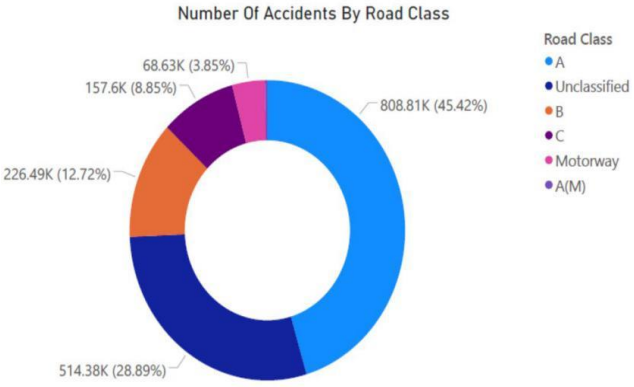


Figure 4 (a): Accidents by Road Class

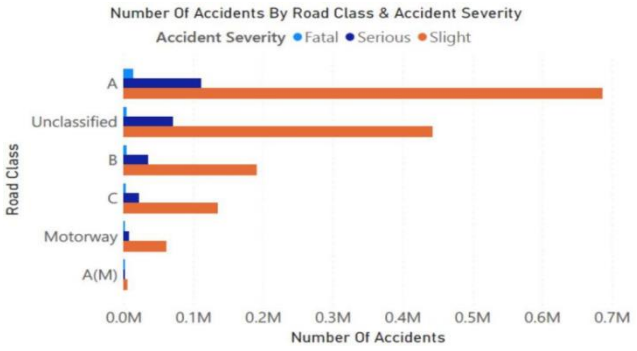


Figure 4 (b): No. of Accidents by Road Class & Accident Severity

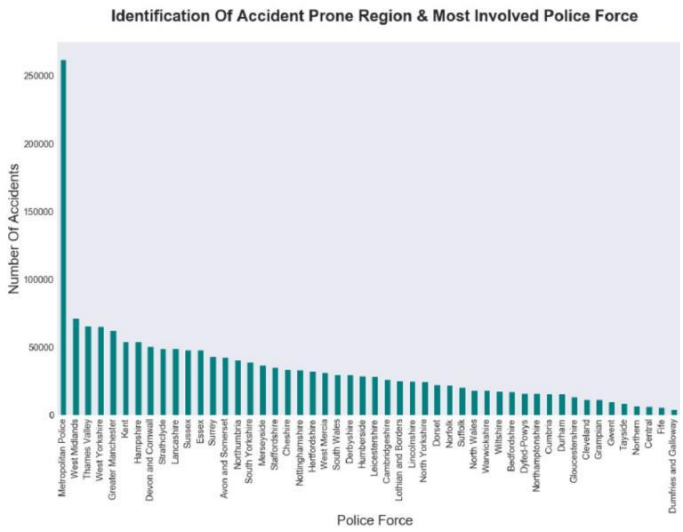


Figure 5 (a): Accident Cases Dealt by Police Force

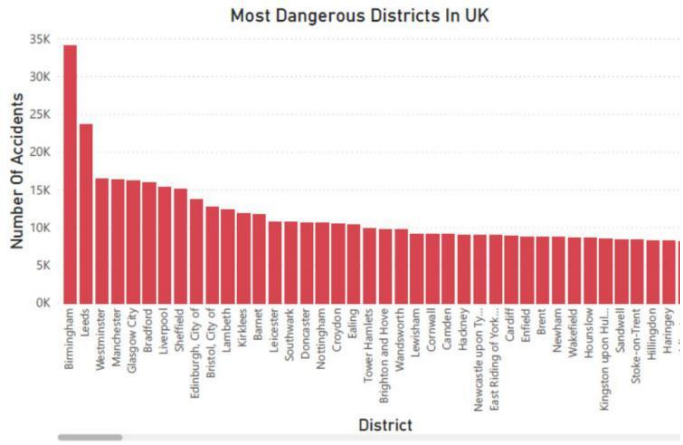


Figure 5 (b): Most Dangerous Districts in UK

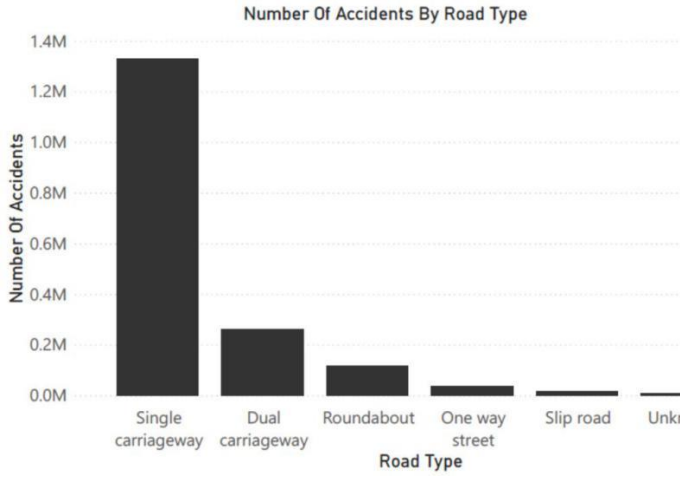


Figure 6: Correlation Between Accidents & Road Type

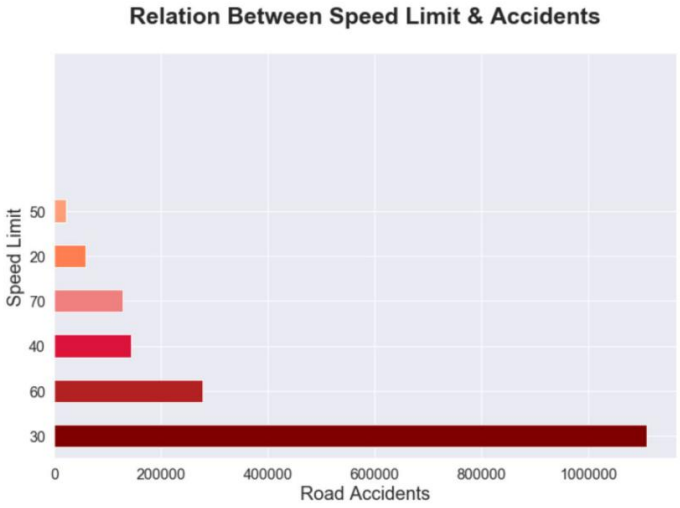


Figure 7: Relation between Speed Limit & Number of Accidents

4.2 Model Estimation

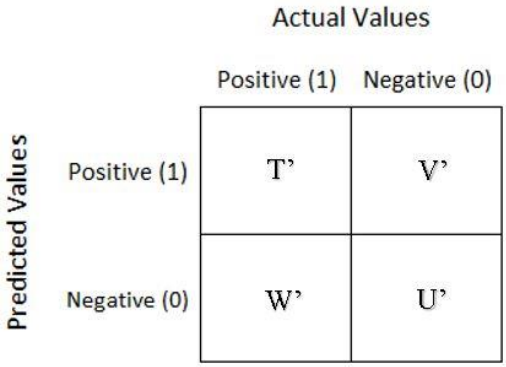
The UK road accidents dataset is first cleaned, leading to removal of any null values, missing values, incorrect domain data like Nan values, etc. In the preprocessing phase, Label Encoder & OneHot Encoder are used to convert categorical text data present in the columns/ features, into model understandable numerical data. Standard Scaler is used for standardization of the dataset & ensures that the individual attributes pertain to standard & normal distribution of data, which is required, so that the estimators of respective machine learning algorithms don't behave badly. Using this, the data is transformed in a way, such that it satisfies the two conditions of mean equaling to 0 & value of standard deviation being 1. The process of eliminating the mean & escalating the same to unit variance for standardization of each attribute, is also taken care of using the Standard.

Standard Scaler changes each feature column $f_{:,i}$ to

$$f'_{:,i} = \frac{f_{:,i} - mean\ (f_{:,i})}{std\ (f_{:,i})}$$

Simple Imputer is applied to the dataset, which acts as an Imputation transformer for completion of missing values. The imputation strategy used is 'mean', i.e. replacing missing values using the mean along each column.

The dataset, at the end of this phase, is split randomly into train set & a test set (test_size = 0.1), with machine learning algorithms & model being trained on the training dataset. After this, model is applied to the testing dataset for prediction & calculation of performance metrics of various machine learning algorithms, for respective shortlisted variables, namely Accident Severity, Rural or Urban, Sex of Casualty & prediction for Age of Casualty. The accuracy for prediction marks one of the best metric for evaluating performance of a machine learning algorithm. The accuracy is calculated using the confusion matrix, which is a matrix that provides information about T' (True P), U' (True N), V' (False P) & W' (False N).



T': predicted positive & it is true

U': the prediction made by the model is negative and it is correct

V': the prediction made by the model is positive but it is false or incorrect

W': the prediction made by the model is negative and it is false or incorrect.

The formula for calculation of accuracy for a machine learning algorithm, using the generated confusion matrix, is as follows:

$$Accuracy = \frac{T' + U'}{T' + U' + V' + W'}$$

Different values were tried using hit & trial method for various parameters of respective machine learning algorithms, in order to achieve the best accuracy & predictive performance. The value of "k" (no. of neighbors) was set to 5 within the KNN model. The distance metric used is "minkowski", with power parameter 2, which is equivalent to the standard Euclidean distance. In Random Forest algorithm, sub samples of dataset are identified, & various decision tree classifiers are applied on them. All the results are averaged over to improve the predictive accuracy. In this way, RF also helps control overfitting. In the RF model applied in our study, the number of trees (value of n_estimators) is set to 10. The value for "criterion" plays a vital role in judging the standard of a split, & is set to entropy for which supported criteria is information gain. For the SVM model, the kernel chosen is rbf (Radial Basis Function), & the gamma value is kept at a default 1.

4.3.Comparison of Prediction Accuracy

	Attribute		
	Accident Severity	Rural or Urban	Sex of Casualty
KNN	86.64%	92.78%	58.45%
Random Forest	86.86%	94.58%	62.39%
SVM	87.73%	94.18%	64.84%

Table 1: Comparison Chart of Prediction Accuracy

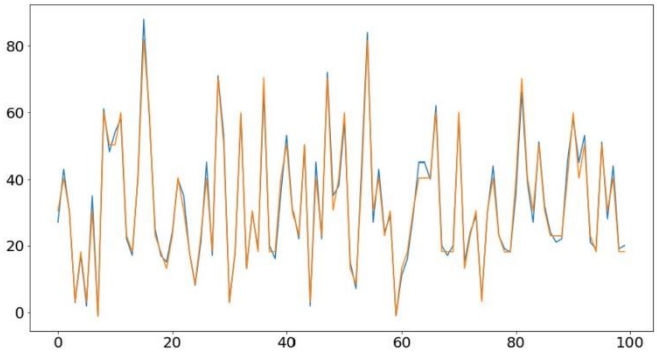


Figure 8: Age of Casualty Prediction

(Blue line: actual data, Orange line: predicted data)

The overall accuracy for Accident Severity ranges between 86.64% & 87.73%. Of all the machine learning algorithms applied, SVM gave the highest accuracy of 87.73%, closely followed by Random Forest & KNN, with 86.86% & 87.73% respectively. The overall accuracy for the attribute Rural or Urban, as shown in table1, ranges between 92.78% & 94.58%. The most robust performance for accuracy is given by Random Forest, i.e. 94.58%. The algorithms KNN & SVM also

showed promising results, with an accuracy of 92.78% & 94.18% respectively.

The accuracy for the attribute Sex of Casualty, as shown in table1, ranges between 58.45% & 64.84%. SVM gave the better performance of all the algorithms applied, with an accuracy 64.84%. KNN & Random Forest gave an accuracy of 58.45% & 62.39% respectively.

5. Conclusion

The horizontal bar plot (Figure 1) infers that the T Union is the most accident prone junction in UK. Other than this, Crossroads & Roundabouts are the next most dangerous junctions. Hence, different measures can be taken by the government, like installation of convex mirrors at blind cuts or dangerous junctions, necessary stop & check signs etc. to prevent road accidents.

From Figure 2, it can clearly be inferred that a large no. of accidents took place during the Afternoon Hours (15-19), the time during which the traffic is at its peak & roads are fairly congested. The next time category with the highest number of accidents is Office Hours (10 - 15). Least number of accidents have taken place during the night i.e. between the time period (23 - 5).

The pie plot in Figure 3 shows the proportion of accidents by severity, namely Slight, Serious & Fatal that took place in UK, over the time period 2005 - 2015. It clearly infers that a fairly large amount of accidents i.e. 85.11% of the total accidents, were Slight Accidents. Whereas, the proportion of Serious Accidents & Fatal Accidents comes out to be 13.59% & 1.29% respectively.

As in India, it's well known that the roads are classified like NH (National Highways), SH (State Highways) etc. Similarly, the roads in UK are classified as A, B, C etc. From the plots, Figure 4 (a) & Figure 4 (b), it can be inferred that A Class Roads are the most dangerous roads, with 45.42% of all the accidents taking place on these roads. The A Class roads are followed by Class B & C, with 12.72% & 8.85% of the total accidents, having taken place on these roads. The pictorial representation Figure 3 (b), clearly infers that most of the accidents, irrespective of their severity, happened on Class 'A' Roads. Hence, it can be concluded that the roads classified by the UK Government as 'A' Class, are the most dangerous & accident prone roads. The Motorway roads, are the safest roads in comparison to all the other road classes. This is supported by the inference that can be made from the graph, as it can be clearly seen that, in the time period 2005 - 2015, that is a fairly long time duration, no fatal accident has taken place on a Motorway. Also, the proportion of accidents that took place on Motorways, to all the accidents that occurred in the defined time period, i.e. 3.85%, is very less as compared to the other road classes.

The plot Figure 5 (a), shows that the most number of road accident cases were handled by Metropolitan Police. This also goes to show that the region that comes under or is handled by Metropolitan Police, is the most accident prone area. The regions in UK where the Metropolitan Police operates, are one of the biggest & densely populated areas of UK, i.e. London & Birmingham. This observation is well in accordance with the Figure 5 (b). The next most involved police force is of West Midlands, which is closely followed by the police force o Greater Manchester.

From Figure 6, we can conclude that single carriageway roads are the most accident prone. The obvious reason for this is that people don't drive in the same way on single way roads as they do on roads with a divider, i.e. dual carriageway roads. Overtaking is also riskier on single roads due to blind spots from the opposite ends, hence leading to high number of road accidents. Dual carriageway roads are much safer than single carriageway roads, as it's well evident in Figure 6, with dual carriageway roads experiencing only 1/6th of all the accidents that occurred on single way roads. Hence, the government & concerned authorities should make sure for implementing stricter measures for single way roads, like installation of speed cameras, declaration of no overtaking zones, virtual fluorescent stripes that can dual up as divider for single way roads etc.

From Figure 7, it's clearly evident that a majority no. of accidents occurred on the roads, where the specified speed limit was only 30 km/hr. Chances of an accident should be fairly less on such roads, as the speed limit specification is very less. So, it can be inferred that the drivers disobey the rules by not driving under the specified speed limit, which results in dangerous circumstances. This also helps to give an important insight, that rash driving on roads, & disobedience by the drivers to follow the rules to drive under the specified speed limit on every respective road, is one of the major cause leading to road mishaps. Hence, the government & concerned authorities must ensure stricter implementation of rules & regulations related to speed limit on the roads, & make sure all the people follow them.

In Figure 8, Multiple Linear Regression is applied for prediction of Age of Casualty. In this, MLR uses several independent variables/ features to predict the outcome or

response for the dependent variable, i.e. Age of Casualty. The actual data is depicted using the "blue" line & the predicted data is depicted using the "orange" line. As it's evident from the graph, the algorithm is able to predict the expected data to a large extent, as at most of the points in the graph, the lines for the actual data & predicted data overlap.

As shown in Table 1, the performance metric accuracy is calculated for all the machine learning models, applied on the respective attributes. The overall accuracy for Accident Severity ranges between 86.64% & 87.73%. Amongst the ML models executed, SVM gave the highest accuracy of 87.73%, closely followed by RF & KNN, with 86.86% & 87.73% respectively.

The overall accuracy for the attribute Rural or Urban, as shown in table1, ranges between 92.78% & 94.58%. The most robust performance for accuracy is given by Random Forest, i.e. 94.58%. The algorithms KNN & SVM also showed promising results, with an accuracy of 92.78% & 94.18% respectively.

The accuracy for the attribute Sex of Casualty, as shown in table1, ranges between 58.45% & 64.84%. SVM gave the better performance of all the algorithms applied, with an accuracy 64.84%. KNN & Random Forest gave an accuracy of 58.45% & 62.39% respectively.

All the machine learning models gave very low or weak performance for the attribute Sex of Casualty. This may be due to the reason that this attribute has a high covariance & varies due to high degree of dependency on other attributes of the dataset. Due to this, another problem faced while performing the analysis for this particular attribute was that, there was duplication of rows at the time of loading the dataset.

The findings of our study can be used for analysis & prediction of various attributes & scenarios that lead to road accidents & other hazardous situations in the transport industry. The government & concerned authorities can use these results & analysis to better understand the various causes that result in road accidents & enforce strict rules & regulations to prevent such situations from taking place in the future. Also, the hospitals & medical emergency services could be boosted in areas that are most prone to road accidents, so that immediate treatment or help is available.

In our current analysis, we have mainly focused on identifying the reasons that lead to road accidents & identification of important insights/ trends that can be inferred from the data. As part of our future plans for research in this field, we will make an attempt to apply multivariate modelling techniques & other ML methods, such as Artificial NN, Deep Learning etc. which may help us to resolve the stated problem & identify the accident prone area to a greater extent.

References

[1] WHO (2018). "Global status report on road safety 2018 (violence and injury prevention)." Geneva, Switzerland.

[2] Chen. And C. (2017). "Analysis and forecast of traffic accident big data." ITM Web of Conferences EDP Sciences, 12, 04029.

[3] Krishna, S., S, S. K., S, S. K., and Mungara, D. J. (2017). "Traffic management using big data analytic tool." International Journal of Scientific Research In Computer Science Engineering And Information Technology (IJSRCSEIT), 2, 777-781.

[4] Ismael, K.S., and Razzaq (2017). "Traffic accidents analysis on dry and wet road bends surfaces in greater Manchester – UK." Kurdistan Journal of Applied Research, 2(3), 284-291.

[5] Zhang, J., Li, Z., Xu, and C. (2018). "Comparing prediction performance for crash injury severity among various machine learning and statistical methods." IEEE Access, 6, 60079-60087

[6] Zheng, W., & Tropsha, A. (2000). Novel variable selection quantitative structure− property relationship approach based on the k-nearest-neighbor principle. *Journal of chemical information and computer sciences*, *40*(1), 185-194.

[7] Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., & Feuston, B. P. (2003). Random forest: a classification and regression tool for compound classification and QSAR modeling. *Journal of chemical information and computer sciences*, *43*(6), 1947-1958.

[8] Joachims, T. (1999). Svmlight: Support vector machine. *SVM-Light Support Vector Machine http://svmlight. joachims. org/, University of Dortmund*, *19*(4).

[9] Tranmer, M., & Elliot, M. (2008). Multiple linear regression. *The Cathie Marsh Centre for Census and Survey Research (CCSR)*, *5*, 30-35.