2025

# Evaluating the Effectiveness of Machine Learning Techniques in Predicting Accident Severity (Traffic Delay) of Traffic Accident

**AJAKAYE, JESUBUKADE**

**E4216209**

Word Count: 2130

# Evaluating the Effectiveness of Machine Learning Techniques in Predicting Accident Severity (Traffic Delay) of Traffic Accident

Jesubukade Ajakaye

School of Computing, Engineering and Digital Technologies

Teesside University

**Abstract**

*Road traffic accidents continue to pose significant threats to human lives and economic stability. This research focuses on predicting the severity of accidents in terms of their impact on traffic delay using machine learning (ML) models. A subset of 500,000 records from a comprehensive U.S. traffic accident dataset (2016–2023) was preprocessed and transformed for binary classification: low vs. high traffic impact. Key preprocessing steps included handling missing values, recoding the target variable, and feature scaling. Feature selection was guided by correlation analysis and Random Forest feature importance. Four ML algorithms including: Logistic Regression, K-Nearest Neighbour, Naïve Bayes, and Random Forest were implemented and evaluated using accuracy, F1-score, and AUC metrics. Random Forest outperformed others with a testing accuracy of 86.3% and F1-score of 0.87. Weather attributes were found to be the most influential predictors of traffic accident severity, followed by location and selected point-of-interest features. The findings highlight the potential of ML models in enhancing traffic management and emergency response planning by enabling real-time prediction of accident impact. This research contributes to the field by refining predictive techniques and offering insights that could support intelligent transportation systems and urban mobility solutions.*

**Keyword:** Accident Severity, Binary Classification, K-Nearest Neighbour, Logistic Regression, Naïve Bayes, Random Forest, Road Traffic

## 1.0 Introduction

The impact of road accident on life and economy has made it a continued point of interest over the past few decades. The impact and delay caused by traffic accidents causes disruption to activities of other road users. Moosavi *et al.*, (2019b) noted that predicting the impact of accidents assist to optimise public transportation, ensuring recommendations of safer and alternative routes. Related research on accident severity focuses on injuries and fatality when road accident occurs. Çelik and Sevli, (2022) on predicting accident severity compared Logistic Regression, XGBoost, Random Forest, K-Nearest Neigbours, Support Vector Machine and Deep Learning. Using Area Under Curve (AUC). They reported that LR algorithm outperforms the others with an 88.1% accuracy. XGBoost has a better performance with 87.9% accuracy than SVM with 87.4%. An accuracy of 86.0%, 85.8%, and 80.6% is obtained with deep learning, RF, and KNN, respectively. Obasi and Benson, (2023) evaluated effectiveness of machine learning techniques to forecast severity of traffic accident, comparing Logistic Regression, Naïve Bayes, Artificial Neural Network, and Random Forest. The models performed 0.87, 0.80, 0.80 and 0.87 overall accuracy respectively.

This research focused on predicting the impact of accident on traffic delay. This helps road emergency worker to understand the accident complexities and offer alternatives to avoid gridlocks. Moosavi *et al.*, (2019a) developed a framework for data collection to address accident severity in traffic delays. They employed traffic events and points-of-interest data for accident predictions. While the model performed averagely in predicting accident class, it performed well in predicting non accident class with f1-scores showing the same trends.

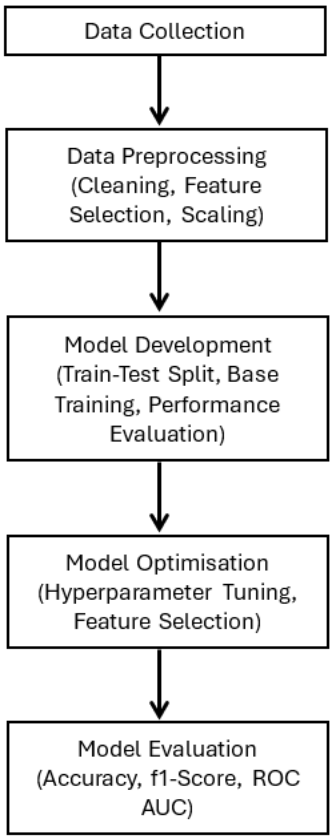
### Research Contributions

- Development of four significant ML algorithms for predicting accident severity (traffic delays)
- Employed Random Forest classifier feature importance to understand sensitivity of predictor variables

- Evaluate model performance and comparing with existing literature in the field.

## 2.0 Research Methodology

This research employed supervised training algorithms to classify the datasets. These ML classifiers were adopted based on several characteristics which include training and testing time, speed on large datasets, interpretability, simplicity of use, works with classification problem, accuracy, among others. The goal of this research is to complete a binary classification. R programming language was used and the process of working with the data are described below.




### 2.1 Data Collection

This research used a dataset meticulously collected on accidents that cover 49 states of the USA between February 2016 and March 2023. The dataset contained approximately 7.7 million accident record, but this data used a sample of 500,000 accidents extracted from the original dataset gotten from the Kaggle data page. Moosavi *etal.*, (2019a) identified some point on interest (POI) attributes based on Open Street Map. All attributes are list as grouped. Information about the attribute's description can be found here.

***Table 1: List of Attributes dataset***

| Group Name | Attributes (46) |
|---|---|
| Traffic Attributes (11) | ID, Sources, Severity, Start_Time, End_Time, Start_Lat, Start_Lng, End_Lat, End_Lng, Distance.mi., and Description |
| Address Attributes (7) | Street, City, County, State, Zipcode, Country, and Timezone |
| Weather Attributes (11) | Airport_Code, Weather_Timestamp, Temperature.F., Wind_Chill.F., Humidity..., Pressure.in., Visibility.mi., Wind_Direction, Wind_Speed.mph., Precipitation.in. and Weather_Condition |
| POI Attributes (13) | Amenity, Bump, Crossing, Give_Way, Junction, No_Exit, Railway, Roundabout, Station, Stop, Traffic_Calming, Traffic_Signal and Turning_Loop |
| Period-of-Day (4) | Sunrise_Sunset, Civil_Twilight, Nautical_Twilight and Astronomical_Twilight |

### 2.2 Ethical, Social, Legal, and Privacy Considerations

This study is grounded in the responsible use of data to improve public safety and traffic management. Predicting traffic delays due to accidents has the potential to improve commuter experiences and enhance road safety. However, it may also introduce social concerns such as the digital divide and unequal access to real-time information. Populations with limited access to smart navigation tools may not benefit equally from the system's recommendations, potentially exacerbating inequalities in mobility and emergency response times. The data used in this research was obtained from publicly available sources, specifically the Kaggle data repository, which is based on the dataset by Moosavi *et al*. (2019a). Legal compliance with data-sharing terms and intellectual property rights was ensured. Finally, Although the dataset does not contain personally identifiable information (PII), it includes location-based data, which may indirectly raise privacy concerns, particularly in high-frequency accident areas. Data anonymisation and aggregation were maintained throughout the research process.

## 2.3 Data Preprocessing

Data preprocessing is an important processing for data preparation to produce a clean dataset for analysis. 829,872 data points had missing values accounting about 3.6%. All these data points were removed reducing the rows to 229,927. The target variable (Severity) shows the severity of the accident, a number between 1 and 4, where 1 indicates the least impact on traffic (i.e., short delay because of the accident) and 4 indicates a significant impact on traffic (i.e., long delay). This was recoded for a binary classification. 1 and 2 were recoded as 0 (low impact) while 3 and 4 were recoded as 1 (high impact). Furthermore, the distribution of the target variable was explored and showed and imbalanced distribution of 94.9% to 5.1%. The Rose package was used to under sample the dominant class therefore row to be considered were reduced to 25,000 making a distribution of 53.4% to 46.6% see figure 1 for distribution.


***Figure 1: Distribution of Target Variable***

New column (Time – duration of the accident) was created to generate new information relevant for analysis. Finally, outliers were checked for (see Figure 4) and a lot of data points were outliers which could be because accidents are unexpected occurrence, and a lot of factors can be contributory. To preserve this information and scale between 0 and 1 for consistency, the min-max scaler was favoured over a robust scaler. Categorical variables were converted to numeric (label encoding) and scaled using the Scale Function in R.

## 2.4 Feature Selection

The research considered information held by each column during data exploration to ascertain if the information carrying will be important to analysis. Columns such as ID, Source (Do not contain information pertain to target variable), Start_Time, End_Time (Transformed to new column "Time" showing total accident time), Start_Lat, Start_Lng, End_Lat, End_Lng (Dropped as "Distance.mi" is in the dataset), Country, Turning_Loop, Roundabout (Only has one value), Description (Unique values for each entries). Furthermore, the research carried out a correlation and multicollinearity analysis.

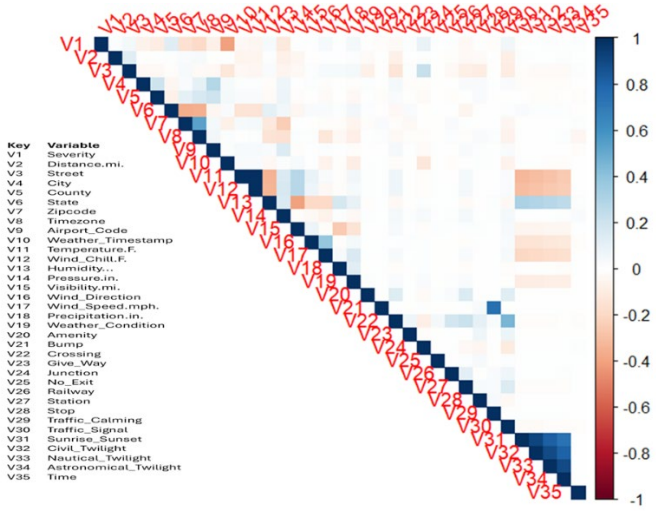



***Figure 2: Correlation Matrix***

All variables in the Period-of-Day Group showed high correlation which is consistent therefore one variable (Sunrise_Sunset) was retained. Temperature.F and Wind_Chill.F were also positively correlated. Virtualisation show that they were carrying the same information therefoe Wind_Chill.F was dropped

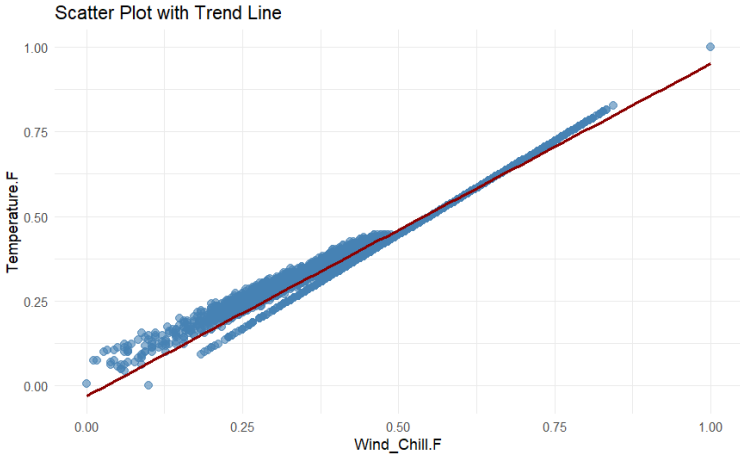



***Figure 3: Scatter plot to check relationship***

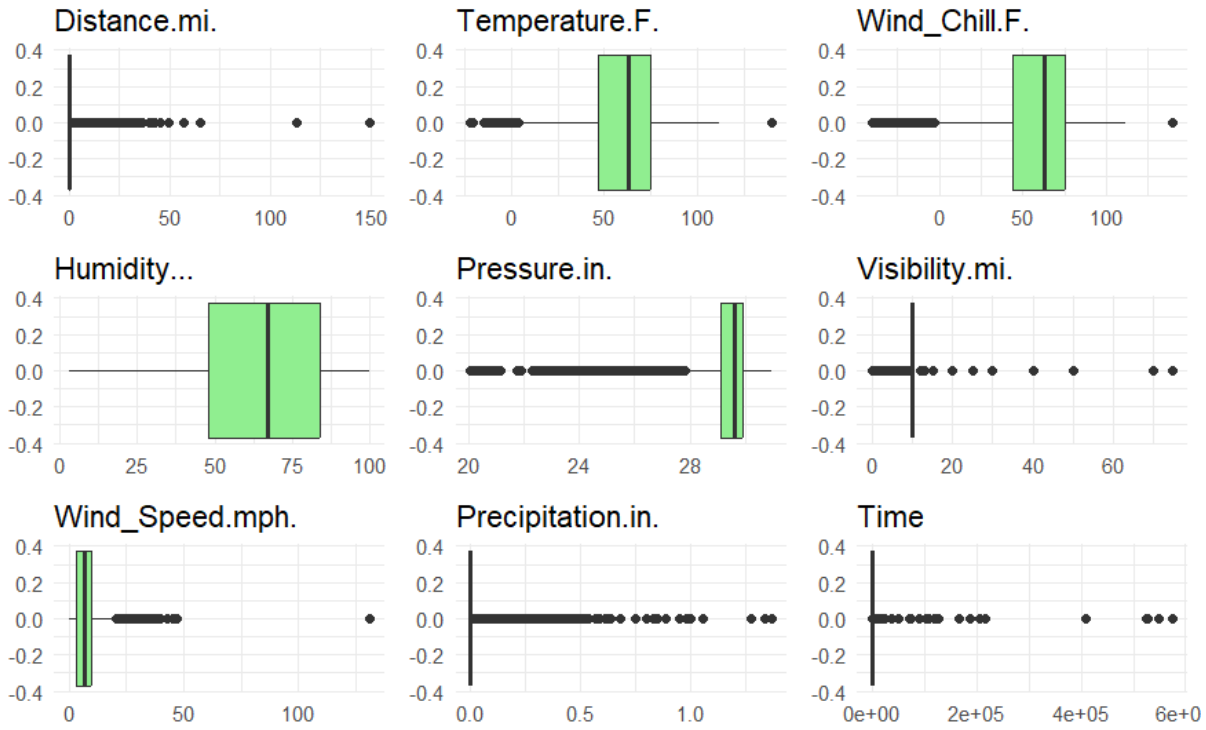



***Figure 4: Outlier detection***

### 2.5 Machine Learning Methods Applied

Four machine learning algorithms was implemented for this research work including: K-Nearest Neighbour (KNN), Logistic Regression (LR), Random Forest Classifier (RF) and Naïve Bayes (NB).

**K-Nearest Neighbour**: KNN posits itself as one of the suitable algorithms for this research work because it is a simple, interpretable algorithm that does not make assumptions about the distribution of the data. Works on every classification data. For optimisation, the "k" (number of neighbours) to consider is tuned the default been 5. The k considered for hyperparameter in this research was between 1 and 150 with a step of 2 (i.e every odd number between 1 and 150).

**Naïve Bayes**: NB is simple, fast and effective. NB treats feature as independent of each other and usually returns optimal accuracy as it sometimes rivals more complex models. Based on Bayes' Theorem, it uses probability to classify data. The key parameters to tune in R include Laplace correction (fL), usekernel and adjust used when usekernel = TRUE. The grid search was used and the values for fL (0, 0.5, 1), usekernel (TRUE, FALSE) and adjust (0.5, 1, 2) were used for tuning.

**Logistic Regression**: LR is well suited for binary classification as it uses logistic function (sigmoid function) to predict the probability of an outcome. LR provides a probability between 0 and 1 representing the possibility of where a data point belongs and makes it ideal for predicting two classes. LR does not have any hyperparameter, but the probability threshold can be moved but its default is 0.5 to predict classes. In this research, the threshold was set between 0.3 and 0.7 to find the optimal threshold.

**Random Forest Classifier**: RF posits itself as a choice model for this research for its ability to mitigate overfitting which is achieved by combining the prediction of multiple decision trees, reducing variance and improving generalisation. RF can handle various data types and provide robust performance with large datasets. RF provides feature importance which was adopted and used for model optimisation in this research. The hyperparameter that was tuned include: mtry (1 – total number of features), ntree (500, 1000, 1500 and 2000) and nodesize (1 – 10).

### 2.6 Train – Test Splitting

After check for missing values again, the data was divided into training set and test set. The training set contains 80% of the processed data used for training the models while 20% was set apart to assess the models' performance and their predictive efficiency.

### 2.7 Performance Metrics and Evaluation

This study compares various metric to determine the overall performance of each model which include overall accuracy, F1-score and ROC/AUC.

- Overall accuracy $(\frac{True\ Negative\ +True\ Positive}{Total})$ shows the proportion of correct classifications made by the model.
- F-Score $(\frac{2\ x\ (Precison\ x\ Recall)}{(Precison+\ Recall)})$ provides a balanced measure of the model's precision and recall.
- ROC AUC is used to evaluate how each model distinguishes between positive and negative classes.

#### 2.7.1 Base Development

The research ensures to be consistent with training of the four models. The Caret package was used for all model development and same seed were set for all models for reproducibility. For the base development, Training Accuracy, Testing Accuracy, and Time Taken were used to compare the four models.

the four models.

***Table 2: Evaluating Base Model***

| Model | Training Accuracy | Testing Accuracy | Training Time Taken |
|---|---|---|---|
| KNN | 0.7030 | 0.7377 | 4 minutes |
| NB | 0.6884 | 0.7155 | 6 minutes |
| LR | 0.7225 | 0.7251 | 12 minutes |
| RF | 0.8549 | 0.8616 | 66 minutes |

Random Forest took the longest time but performed better than all other models.

### 2.8 Model Optimisation

#### 2.8.1 Feature Selection

30 variables were used at the base model development. RF at training returns feature importance see Figure 5. It is important to note that most of the features considered as POI were not important features for prediction of accident severity but variable under the weather attribute ranks highest. The top 20 variables are therefore selected for training in the optimised model.

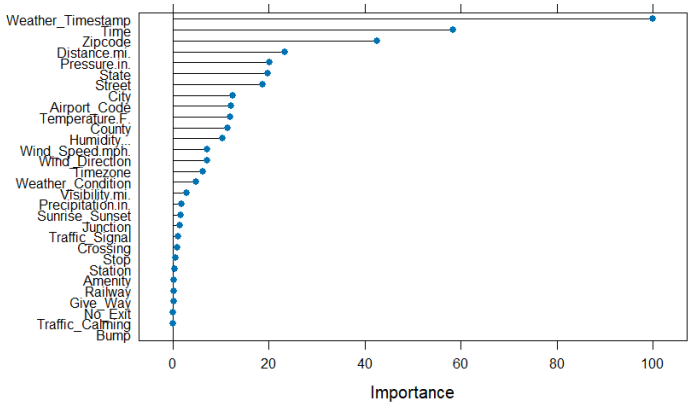



***Figure 2: Feature Importance***

### 3.0 Results

For the optimised models, the same seed was also set as the same for all models and train control function was used to set cross validation which was used across all models. Training time was not considered for evaluation because different hyperparameters was set based on model time and would affect training time. Therefore, training accuracy, testing accuracy, f1 score and area under curve (AUC) were used to compare the model performance.

***Table 3: Evaluating Optimised Models***

| Model | Best Hyper Parameter | Training Accuracy | Testing Accuracy | F1 Score | AUC |
|---|---|---|---|---|---|
| KNN | k = 27 | 0.7347 | 0.7497 | 0.7814 | 0.74 |
| NB | fL = 0<br>usekernel = TRUE<br>adjust = 0.5 | 0.7245 | 0.7285 | 0.7690 | 0.73 |
| LR | threshold = 0.6 | 0.7364 | 0.7437 | 0.7854 | 0.72 |
| RF | ntree = 500<br>mtry = 9 | 0.8600 | 0.8626 | 0.8717 | 0.86 |

**ROC/AUC**

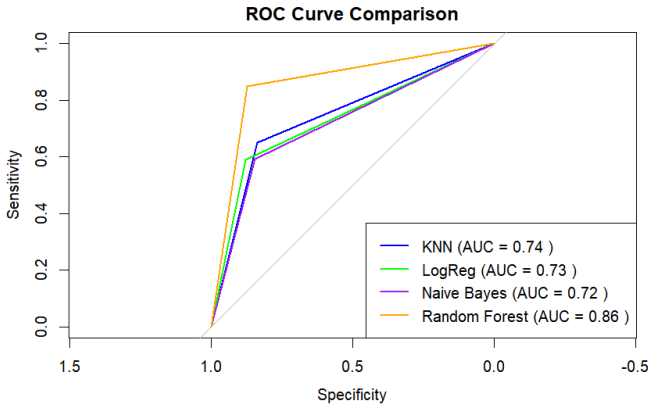



***Figure 3: ROC AUC Comparison of Models***

metrics used in evaluation as revealed in the confusion Matrix in Figure 7.

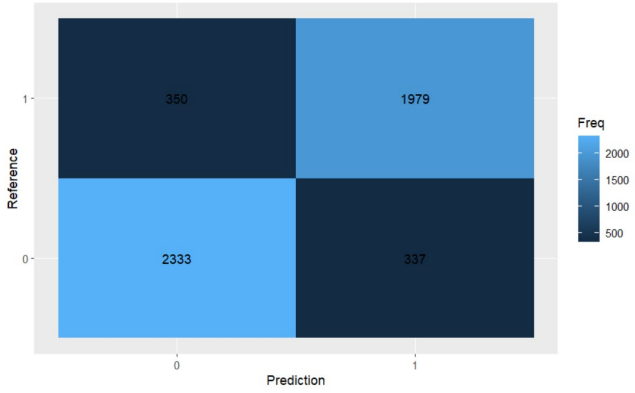



***Figure 4: Confusion Matrix for Random Forest Model***

## 4.0 Discussion and Conclusion

The models four models showed high predictive accuracy making them useful for predicting traffic delays when accident occurs. The Feature Importance as shown in Figure 5 revealed that Weather_Timestamp, Time Taken, Distance, Pressure and Location (State, City) which accident occurred are the most important predictors. Among the point-of-interest features, Junction, Traffic_Signal, Crossing and Stop are more effective than others for accident severity predictions which is in line with the findings of (Moosavi *et al.*, 2019b). For evaluating the models' performance, the research utilised training and testing accuracies, f1-score, the Receiver Operator Characteristics (ROC) and the Area Under the Receiver Operator Characteristics Curve (AUC). Table 3 shows the summary of the metrics used to evaluate the performance of all models while Figure 6 shows the ROC comparison and AUC values. All models indicated good prediction ability for classification, but Random Forest is the best performing model across all

The findings have showed that weather attributes are the top predictors that impact delays when accidents occur, then the address attributes before the point-of-interest attributes. Identification of these important features can help policymakers and first responders plan effectively and mitigate against long delays on the road when accident happens. Consequently, prediction of accident severity (possible delay) in real time can help navigation recommendations to other road users and reduce grid locks on road and can even make accident scene more accessible to those providing helps to accident victims.

## References


Çelik, A. and Sevli, O., 2022. Predicting traffic accident severity using machine learning techniques. *Türk Doğa ve Fen Dergisi*, *11*(3), pp.79-83.

Moosavi, S., Samavatian, M.H., Parthasarathy, S. and Ramnath, R., 2019a. A countrywide traffic

accident dataset. *arXiv preprint arXiv:1906.05409*.

Moosavi, S., Samavatian, M.H., Parthasarathy, S., Teodorescu, R. and Ramnath, R., 2019b, November. Accident risk prediction based on heterogeneous sparse data: New dataset and insights. In *Proceedings of the 27th ACM SIGSPATIAL international conference on advances in geographic information systems* (pp. 33-42).

Obasi, I.C. and Benson, C., 2023. Evaluating the effectiveness of machine learning techniques in forecasting the severity of traffic accidents. *Heliyon*, *9*(8).