Nigerian Journal of Technology (NIJOTECH)
Vol. 40, No. 5 September, 2021, pp.888-900
www.nijotech.com

Print ISSN: 0331-8443
Electronic ISSN: 2467-8821
http://dx.doi.org/10.4314/njt.v40i5.15

Predicting Road Traffic Crash Severity in Kaduna Metropolis using some Selected Machine Learning Techniques

A. E. Evwiekpaefe[1, *], S.M. Umar[2]

[1,2] *Department of Computer Science, Nigerian Defence Academy, Kaduna, NIGERIA*

Abstract

Road Traffic Crash (RTC) is among the leading causes of death in the world and has a significant impact on the socio–economic development in a society. Generally, RTC can be caused by one or a combination of the following factors: Human, environment and vehicle. This study utilized five data mining algorithm classifiers (Decision Tree (DT), K-Nearest Neighbor (KNN), J-Repeated Incremental Pruning to Produce Error Reduction (JRIP), Naïve Bayes (NB), and Multi-layer Perceptron (MLP)) to classify the severity of RTC and identify the significant causes of RTC in Kaduna State, Nigeria. The RTC data used in this study included 26 RTC attributes with 1580 instances from 2016 to 2018 that covered fatal, serious and minor cases obtained from the Federal Road Safety Corps, Kaduna sector command. Two sets of experiments were performed on the classifiers (without and with feature selection). The study results showed that among the five data mining algorithms used, K-NN had the best accuracies of 94.8% and 96.1% respectively for the without and with feature selection experiments.

Keywords: Classifier, Data mining algorithm, Federal Road Safety Corps, Feature Selection, Road Traffic Crash

1.0 INTRODUCTION

Movement is one of the general characteristics of living things, as well as extremely vital part of the socio-economic development of the society in general. Accident is anything which occurs by chance, anything taking place unexpectedly [1]. Traffic crash may result in fatal, serious and vehicle damage as well as financial burden [2]. RTC happens when the road users (vehicle, motorcycles or pedestrian) collide with one another which may be due to mechanical deficiency, human or environmental conditions. According to the World Health Organization (WHO) about 1.35 million people die worldwide yearly, up to 50 million people are injured and many remain disabled for life due to RTC. Therefore, it is imperative to discover and eliminate as many RTC as possible and its causes. The types of RTC crashes are fatal, serious and minor [1].

RTC causes are paramount to the Road Safety Agency as it will assist them come up with rules and regulations on how to prevent the crash.

RTC is one of the major socio-economic concerns in many countries all over the world since it causes loss of lives and properties [4]. The classification of RTC mainly deals with categorization of severity of injury rate using recorded data of RTC by utilizing the identified contributory factors. Generally, from previous research, the factors that contribute to RTC plunge into three categories: human, environment and vehicle [1]. Predicting the major contributory factor for crash severity of injuries has also become a centre focus for researchers since it will assists the government and policy makers find a general solution to the reduction and elimination of crash severity bringing it to a zero fatal crash society.

The risk of road traffic severe injuries varies significantly from different countries and regions of the world. Africa is the region with the highest rate of road traffic death 26.6% for every 100,000 populations [3]. Consequently, there is high morbidity and mortality rate in the region as well as socio–economic loss in its societies and nations in general. As in many other developing countries, the RTC is one of the major challenges in Nigeria. The chance of the RTC to kill someone in Nigeria is 47 times advanced than in Britain [5].

In Nigeria RTC has become a daily event. Hardly a day goes by without information on RTC. The road users in Nigeria also continue to use outdated vehicles which have contributed in the increase of RTC in Nigeria. The

*****Corresponding author **(Tel:** +234 (0) 8035600524**)**

Email addresses: *aeevwiekpaefe@nda.edu.com* (A.E Evwiekpaefe), *umarsalisumohammed@gmail.com* (S.M Umar)

three major factors of RTC which are Human, Environment and Vehicle need to be thoroughly highlighted to classify what combinations of crash severity can be predicted as fatal, serious or minor so as to easily classify and predict the type of road traffic crash. Several studies related to RTC have been conducted over a decades but they address different aspect of RTC. Data mining techniques have been widely chosen to address the global problem of RTC [6]. However, little or no work has been done to address severity of RTC using Machine Learning Techniques in Kaduna State. This research is to predict the most significant influential types of RTC severe injuries (fatal, serious and minor) among human, environment and vehicle factors using some selected Machine Learning Techniques.

2.0 RELATED WORKS

[7] applied data mining technologies to link recorded road characteristics to accident severity in Ethiopia and developed a set of rules that could be used by the Ethiopian Traffic Agency to improve safety. To predict accident severity, various classification models were built using decision tree, naive Bayes, and K-nearest neighbor classifiers. In the first experiment, the 18,288-accident dataset with 10 attributes, including 9 independent variables and one dependent variable (the class-label attribute 'Accident Severity'), were fed to WEKA's explorer. The J48 classifier was used and an accuracy of 80.221 was achieved. In the second and third experiments, the same input, instances, and attributes were fed to WEKA. Using the naive Bayes classifier, an accuracy of 79.9967 was achieved. Using the K-nearest neighbors classifier (IBK), an accuracy of 80.8281 was achieved. Thus IBK surpassed the other classifiers.

[8] used data mining techniques to classify locations of RTC in one of the busiest highways in Nigeria. Neural Network and Decision Tree classification algorithms were used with 21 attributes of road traffic crash data collected from FRSC, Nigeria. The suggested model classified and predicted the normal location of road crash and identified that the most important causes of crash in the location are tyre burst, loss of control and over speeding. The neural network algorithm achieved accuracy of 54.73% and Decision Tree 77.70% for classification of road crash location. The study concentrated on one location and only predicted crashes that are fatal.

[9] proposed a new fuzzy granular decision tree to generate road collision rules to apply to the discrete and continuous data stored in collision databases. To improve the efficiency of the algorithm, the fuzzy rough set feature selection was applied .The major highways in California were considered as a case study to examine the proposed approach. The experimental results demonstrated that the proposed method is more accurate and efficient than the traditional decision tree methods, with the less redundancy in constructing the decision tree.

[10] stated that weather condition such as rain, mist, fog and sunrays are likely to increase the rate of RTC in Nigeria. The study further observed that the entire reluctance of Nigerian command to regularly improve the condition of the roads which has resulted to the present frequency of potholes on the highway is another significant factor affecting RTC in Nigeria. The study focused only on environmental factor as a key to RTC in Nigeria.

[11] discussed RTC data collected from the UK from 2014 to 2015 that consisted of nine (9) attributes of 81,690 crash cases. The authors proposed a model to predict and classify road crashes into fatal, serious and minor using three classification algorithms: Bayesian Network, Multilayer Perceptron and Decision Tree. The proposed model achieved accuracy results of predicting road crash classes BayesNet 72.3%, MLP 72.62% and Decision Tree 71.70%. The researchers used vehicle and environment factors to predict the common factors of crash severity and identify that road type, light condition and vehicle maneuver are the most causes of crash severity. The study used few attributes and only two factors of RTC Human and environment.

[12] applied Artificial Neural Network (ANN) to predict RTC based on 5973 traffic crash cases that happened in Abu Dhabi from 2008 to 2013. For each crash record, 48 different features were collected at the time of the crash. After data preprocessing, the data features were reduced to 16 features and four injury classes. The data mining tool Waikato Environment for Knowledge Analysis (WEKA) was used to build the ANN classifier. The complete crash data was used for the training and validating of the data set. While 90% of the data were used for training, the remaining 10% were used for testing. The results show that the developed ANN classifier can predict crash injury with reasonable accuracy. The overall model prediction performance for the training and testing data is 81.6% and 74.6%, respectively. The study only predicted the serious crash and used only one classification algorithm.

[13] analyzed road traffic crashes in Anambra State, Nigeria with the idea of achieving accurate predictive classification for forecasting RTC in the State. The result showed that ARIMAX model outperformed the ARIMA model generated when their performances were compared using the decrease Bayesian information standard, mean ultimate percentage error, root mean square error; and higher coefficient of determination (R-

Squared). The results of this study revealed that human, vehicle and environmental factors in time series analysis of the crash data set produced stronger predictive classifiers than solely used collective crash count. The study provided awareness on road traffic safety and forecasted road crash using human, vehicle and environmental factors. The study believed that if the result achieved in the research is applied, it will assist in minimizing road traffic crash. The researchers based their work on comparison between ARIMA and ARIMAX forecasting models.

[14] used support vector machine classifier to predict the road crash severity injuries using human, environment and vehicle contributory factors. The main significance of this research was the development of a model for classification of the road traffic crash severity injuries in Nigeria. The general prediction performance accuracy was 73% for training data set and 74% for testing data set. The work combined the three main contributory factors identified in the data set which resulted more attributes or features (23 features). Feature selection was suggested in the future since it will identify the most significant causes of the accident.

[15] applied comparison methodology approaches which consisted of text classification algorithm to identify traffic related tweets. These traffic communications were then categorized and further classified into positive, negative, and neutral classes using sentiment analysis. In addition, stress and repose strength detection was performed with the means to further analyzed users sensation within the tweet. Future work was suggested to be carried out to apply the proposed framework. No implementation of machine learning tools was carried out. [16] applied established models (classifiers) to identify accident factors and to predict traffic crash severity using earlier recorded traffic data. Data mining decision tree (J48, ID3, CART) and Naïve Bayes classifiers were built to model the injury crash using the WEKA tool. The classification result shows that the accuracy of J48 classifier was higher than others. In the study the researchers concentrated only on injury crash with few causes of RTC classification.

[17] used five years period of two automobile crash data collected at the metropolitan area of Miami and Florida. The results indicated that cost-sensitive learning classifiers were higher than regular classifiers in accurately predicting serious and fatality crashes. Among the cost sensitive models used, Random Forest outclassed Decision tree and International Baccalaureate models in predicting driver injury severity for four categories of drivers. The models displayed substantial differences in injury severity determinants across the age and gender partners. The vehicle factor was not included in the study.

2.1 Research Gap

Based on the extensive review of related works on RTC, it was discovered that little or no research has been done on RTC classification that include all the three contributory factors which are: Human, Environment and Vehicle with special focus on crash severity of fatal, serious and minor in Nigeria using Machine Learning Techniques. This study will therefore fill this gap.

3.0 METHODOLOGY

3.1 Research Framework

The proposed research framework has three phases as shown in figure 1. The first phase identifies and formulates the research problem. The second phase involves the data preparation where the data obtained from Federal Road safety Corps (FRSC), Kaduna Sector Command in form of Excel document was cleaned in order to remove the noise (unnecessary information). The final phase is the development of the classifiers which included the training, testing and validation of the classifiers. Feature selection approach was used to extract the most relevant features and the model evaluated using accuracy, true positive rate, false positive rate, recall and precision on JRip, Decision tree, KNN, Multilayer Perceptron and Naïve Bayes classifiers.

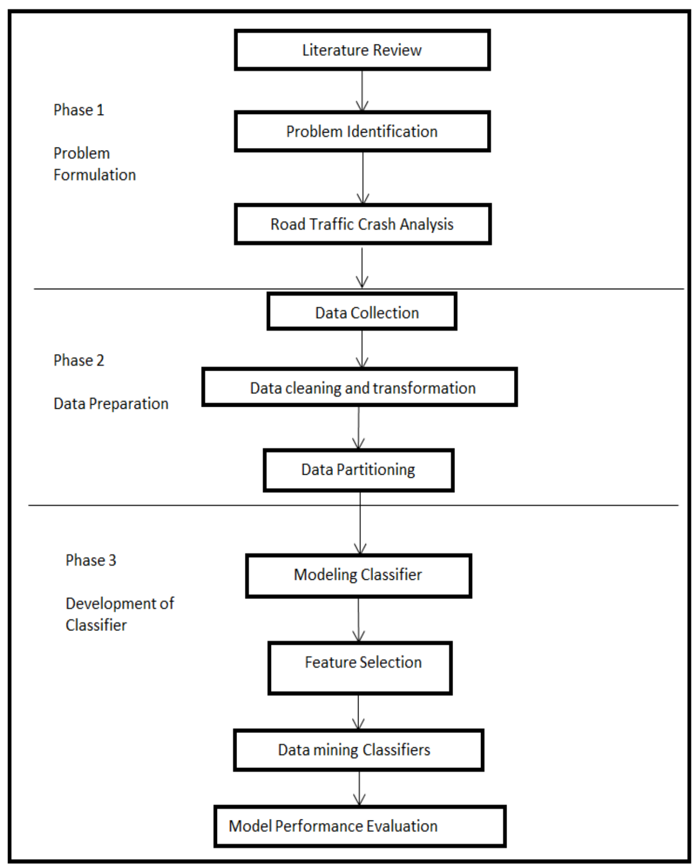

Figure 1: The Proposed Framework

3.2 Data Collection

The dataset used in this research was obtained from Federal Road safety Corps of Nigeria Kaduna State command (the agency responsible for eradicating road traffic crashes and creating safe motoring environment in Nigeria). The Kaduna sector command covers the 12 sectors across State. The data contains the three RTC contributory factors of human, environment and vehicle with severity class as fatal, serious and minor for the period of three years from 2016 to 2018 with 1580 crash cases and 26 attributes.

3.3 Description of Data Attributes

The dataset of this study has 26 attributes in total within three factors (Human, Environment and Vehicle). Human factors are considered the most vital factor contributing to RTC. Human related factors include over speeding, dangerous driving, drink driving, the use of drugs, inexperience and unqualified drivers. Environmental factors include the poor condition of the roads, severe weather,and inadequate road support facilities such as road signs and drains. Another Environment factor are weather condition such as rain, mist, fog and sunrays which are likely to increase the rate of RTC and poor weather such as heavy rain could increase the hazard condition of the road with potholes. Vehicle factors is the another contributory factor of road traffic crash. The condition of the vehicle is a key factor to the causes of crashes. Vehicle related factors include the brake system, tyres condition, vehicle body, lighting system and the engine. The RTC data was classified into three classes which include fatal, serious and minor. Fatal is defined as a road traffic crash that involves one or more persons killed in a crash.

In other words any road traffic crash involving loss of life is said to be a fatal crash. Whilst, a class serious is a type of crash that the victim sustain multiple injuries or may become completely disabled. In a minor class RTC, the victim does not sustain any serious injury and no loss of life in the crash; it may only be loss of property or vehicle damage. The detailed description of the attributes used in the study is provided in Table 1.

Table 1: Crash Attributes

S/N	Attributes	RTC Factors	Description	Value
1	Command	Environment	The area administrative office in charge of safety activities	RS1.1, RS1.12, RS1.13, RS1.14, ZEBRA35, ZEBRA40…………ETC
2	Crash Time	Environment	An approximate time of the crash	Valid time of crash
3	Year	Environment	The period the crash was reported	2016, 2017, 2018
4	Route	Environment	The path taken to move from a starting point to a destination	KZ, KA, KB, ZT, KS, MK, BT, MS, DK, KKD………. ETC
5	Vehicle Category	Vehicle	The categories of the vehicles involved in the crash e.g. Private, Commercial and Government.	C, P, G
6	Vehicle Name	Vehicle	The product or make of the vehicle involved in the crash e.g. Honda, Toyota, etc.	TOYOTA (TYT), PEUGOUT (PGT), VOLKSWAGEN (VW), BAYERISCHE MOTOREN WERKE(BMW), TC ,MC, GOLF, MAN, BX, HD, INNOSON……………ETC
7	Vehicle Type	Vehicle	The types of vehicle involved in the crash e.g. car, bus, Trailer, etc.	B, C, T1, T2, TR, MC
8	Over Speeding	Human	The crash occurs due to the over speeding of the driver	0,1
9	Dangerous Overtaking	Human	The crash occurs during dangerous overtaking by the driver	0,1
10	Dangerous Driving	Human	The road crash occurs as a result of dangerous driving of the vehicle by the driver.	0,1

S/N	Attributes	RTC Factors	Description	Value
11	Loss of Control	Human	The road crash occurs as a result of lost control of the vehicle on the road by the driver.	0,1
12	Overloading	Human	The crash occurs as a result of excess overloading of the vehicle with either passenger or load by the driver	0,1
13	Route Violation	Environment	The road crash occurs as a result of the driver violating the route on the highway.	0,1
14	Sleeping on steering	Human	The driver was sleeping when the crash occurred	0,1
15	Use of phone while driving	Human	The road crash occurs as a result of the driver using his phone while driving. That is, his concentration was not on the driving.	0,1
16	Wrongful overtaking	Human	The crash happens when the driver wrongly overtakes in a corner, sharp bend, etc. without seeing ahead.	0,1
17	Bad Road	Environment	The crash occurs as a result of bad road or a black spot, pothole, sharp bend etc.	0,1
18	Road Obstruction	Environment	The road crash happens as a result of obstructions on the highway	0,1
19	Poor weather	Environment	The crash occurs as a result of poor weather in the area like during rain, haze, etc.	0,1
20	Brake failure	Vehicle	Crash occurs as a result of failure of the brake of the vehicle.	0,1
21	Mechanical deficiency	Vehicle	The crash happens as a result of the vehicle mechanical deficiency.	0.1
22	Sign light violation	Vehicle	The road crash occurs as a result of a violation of sign light such as traffic lights, zebra crossing, etc.,	0,1
23	Tyre burst	Vehicle	Crash happens as a result of flat tyre or tyre outburst in the vehicle.	0,1
24	Driving under Alcohol/Drug	Human	The driver was under the influence of alcohol/drug while driving which led to the road crash.	0,1
25	Fatigue	Human	This road crash occurs as a result of exhaustion or tiredness on the part of the driver.	0,1
26	Class	Fatal, Serious or Minor	The result of crash severity when it happens	

3.4 Data Preprocessing

The preparation process includes data selection, data cleaning and transformation, attribute selection and partitioning the data into training and testing for model development.

3.5 Data normalization

The machine learning algorithms accept real numbers for the classification and building of models. After presenting the data in nominal and numerical format, the data was saved in CSV file and imported into

WEKA for normalization ranging from 0 to 1. Also oversampling and under sampling of the dataset was performed to resample the dataset to cater for imbalanced data. This was done to help achieve better performance matrices so as to gain better or higher accuracy.

3.6 Data Cleaning and Transformation

Data cleanness is one of the qualities for accurate classification in data mining. If the data set used is not clean and transformed for quality mining the result of prediction may not be accurate. Therefore, all the noisy data are identified and removed, the missing field in the data also addressed and the possible attribute predictors of crash severe injury were selected from the three contributory factors.

3.7 Data Partitioning

The data was partitioned into two sets, one for training and another set for testing data using 90/10 rule. The training dataset contained 90% of the data while the testing data contained the remaining 10% of the dataset. The training dataset was used to train the proposed classifier while the testing dataset used to measure and validate the performance accuracy and efficiency of the classifier. The 10 – fold cross validation techniques was used to validate the classifiers in the JRip, KNN, Multilayer Perceptron, Decision Tree and Naïve Bayes algorithms.

3.8 Development of classifier

The development of a classifier in this study includes the following stage, initial model was built with the standard machine learning classification algorithms using the WEKA tool. Feature selection methods were applied to improve the model development by extracting the most significant possible predictors of crash severe injuries in the classifier.

3.9 Model Performance Evaluation

The model performance was evaluated by generating the confusion matrix which utilizes the class True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) to find the Percentage Accuracy, Precision, Recall and F- measure of each predicted and classified class in the proposed algorithm.

Table 2: Confusion matrix

Actual Class	Predictive Class	
	Positive	Negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

Here the labels that constitute the confusion matrix are True Positive (TP) which represents the number of data that are correctly classified in the positive class. False Positive (FP) is number of instances wrongly classified as positive class. False Negative (FN) represents the number of instances supposed to be classified in positive class but wrongly classified as negative and True Negative (TN) is number correctly classified in the negative class. The description of the performance measures are as follows:

Accuracy is used to measure the percentage of correctly classified result in the overall road traffic crash dataset [18].

$$A = \frac{TP + TN}{TP + FP + FN + TN} \tag{1}$$

Precision is used to measure the percentage of the correctly predicted positive values to the overall predicted positive values [18].

$$P = \frac{TP}{TP + FP} \tag{2}$$

Recall is used to measure what percentage of tuples in the crash data the classifier label as positive in the class. It is also known as the true positive rate, ie., the number of correct predictions [18].

$$R = \frac{TP}{TP + FN} \tag{3}$$

F – Measure is the accuracy of harmonic mean of precision and recall, that is, the weighted average of the class [18].

$$F = \frac{2 \times PRECISION \times RECALL}{PRECISION + RECALL} \tag{4}$$

4.0 RESULTS AND DISCUSSION

4.1 Preprocessing Results

The preprocessing results included the 26 attributes that were used in the study either as nominal or numeric so as to get good results in the analysis.

Command sector represents the area administrative offices of the FRSC in Kaduna State. There are 15 area command offices in Kaduna State with RS1.1 as head office. RS1.16 has the highest reported crash cases of 288 while Zebra40 has the lowest reported crash cases of 10. The results are shown in Figure 2. Year of Crash is the period the crash cases is being reported and used in the study from 2016 to 2018. The analysis discovered that in 2016, 2017 and 2018 there are reported crash cases of 446, 719 and 415 respectively. The result is shown in Figure 2. Route is the path taken in moving from the starting point to the destination. Based on dataset there are 21 routes across Kaduna state. Kaduna-Doka is the highest route with crash cases of 356 while Saminaka- Jos is the lowest with 7 crash cases. The results are shown in Figure 2. Vehicle type is the kind of vehicles that involved in the crash as reported by FRSC. There are 6 type of vehicles based on the dataset. Cars have the highest number of crash cases with 781 cases while tricycle with lower cases of 23 crashes.

The results are shown in Figure 2. Vehicle category is the category of vehicles in the crash. There are private, commercial and government vehicles. The crash cases show that commercial vehicles have the highest number of cases with 950 then private with 564 and government vehicles with 66 crash cases. The result is shown in Figure 2. Vehicle Name is the product of the vehicle involved in the crash. Based on the data collected from FRSC there are 44 different vehicle products that were involved in the road crash. Toyota product has the highest crash of 286 cases. The result is show in Figure 2. Over speeding is the type of crash that occurs as a result of excessive speed while driving. The result is show in figure 2. Dangerous driving is the crash that occurs as a result of driving recklessly. The result is shown in Figure 2.

Dangerous overtaking is the crash that occurs as a result of illegal overtaking. The result is shown in Figure 2. Class categories is the crash severiy cases as fatal, serious or minor, the crash cases is 560, 912 and 108 respectively. The result is shown in Figure 2. The visualization of attributes showing the complete output of the 26 attributes used in this study is shown in Figure 2.

4.2 Performance Results of the Classifiers

The results of the experiment which was guided by the performance metrics discussed in the methodology are demonstrated here.

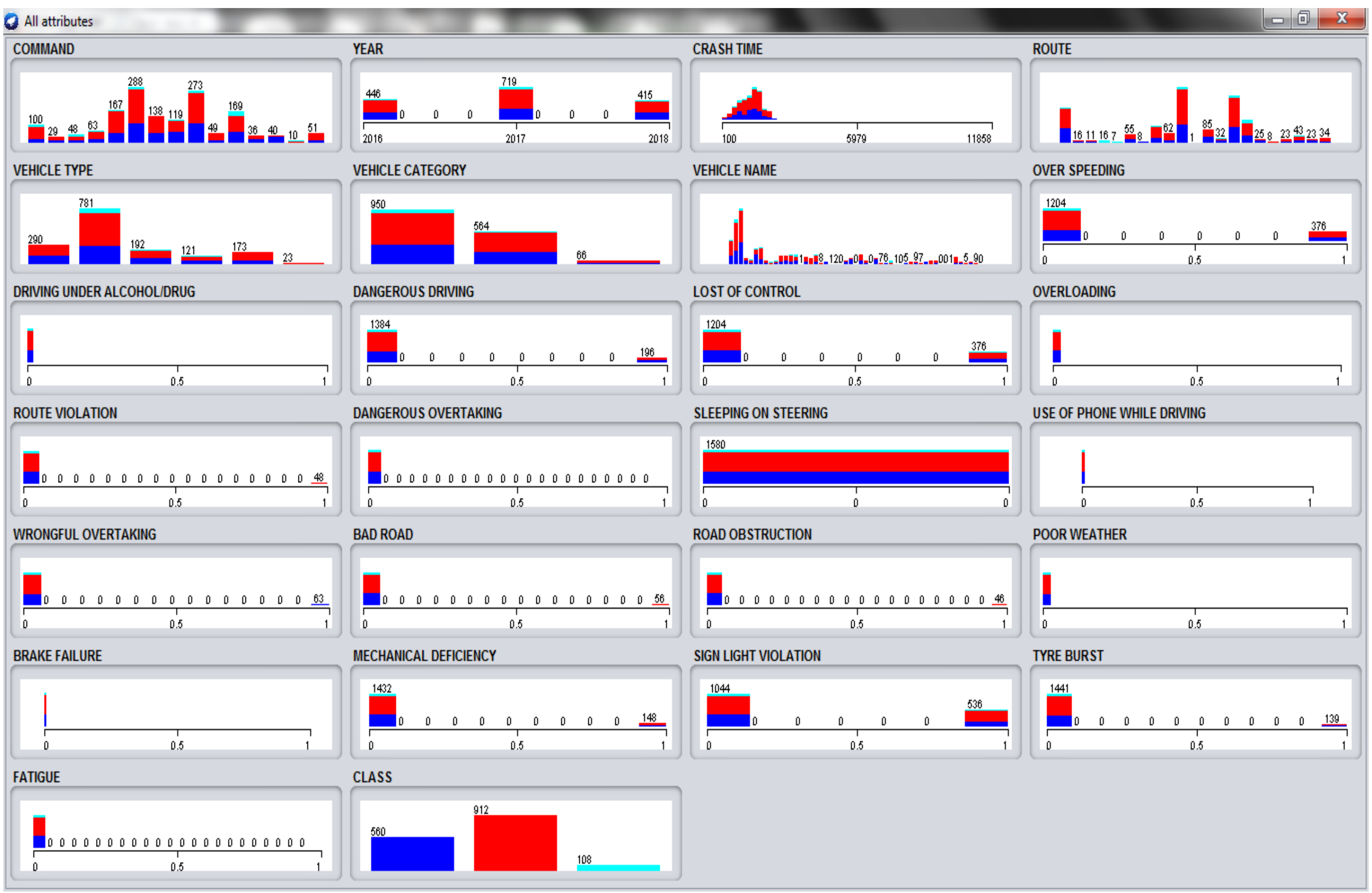

Figure 2: Attributes visualization

4.3 *Experimental Results of Classifiers without Feature Selection*

The experimental results of the data mining algorithms executed on the RTC dataset without feature selection technique are shown in this section. The five selected algorithms used in this study are: Decision tree, K-NN, MLP, Jrip and Naïve Bayes algorithm.

J48 Decision Tree algorithm had 90.3% accuracy without feature selection. The result of Decision Tree algorithms is show in Figure 3. The comparison of Percentage Accuracy, Precision, Recall and F – measure of Decision tree using confusion matrix are shown Table 3. K- Nearest Neighbor had 94.8% accuracy without feature selection. JRip classifier had 86.3% accuracy without feature selection method. Naïve Bayes classifier had 63.7% accuracy without feature selection method. Multilayer Perceptron had 86.7% accuracy without feature selection method. The results of Percentage Accuracy, Precision, Recall and F – measure of these classifiers are shown in Table 3.

The Table 3 summarizes clearly the result of the five selected algorithms that were used to predict RTC severity as fatal, serious or minor without applying feature selection technique using the 26 attributes and the number instances of 1580.

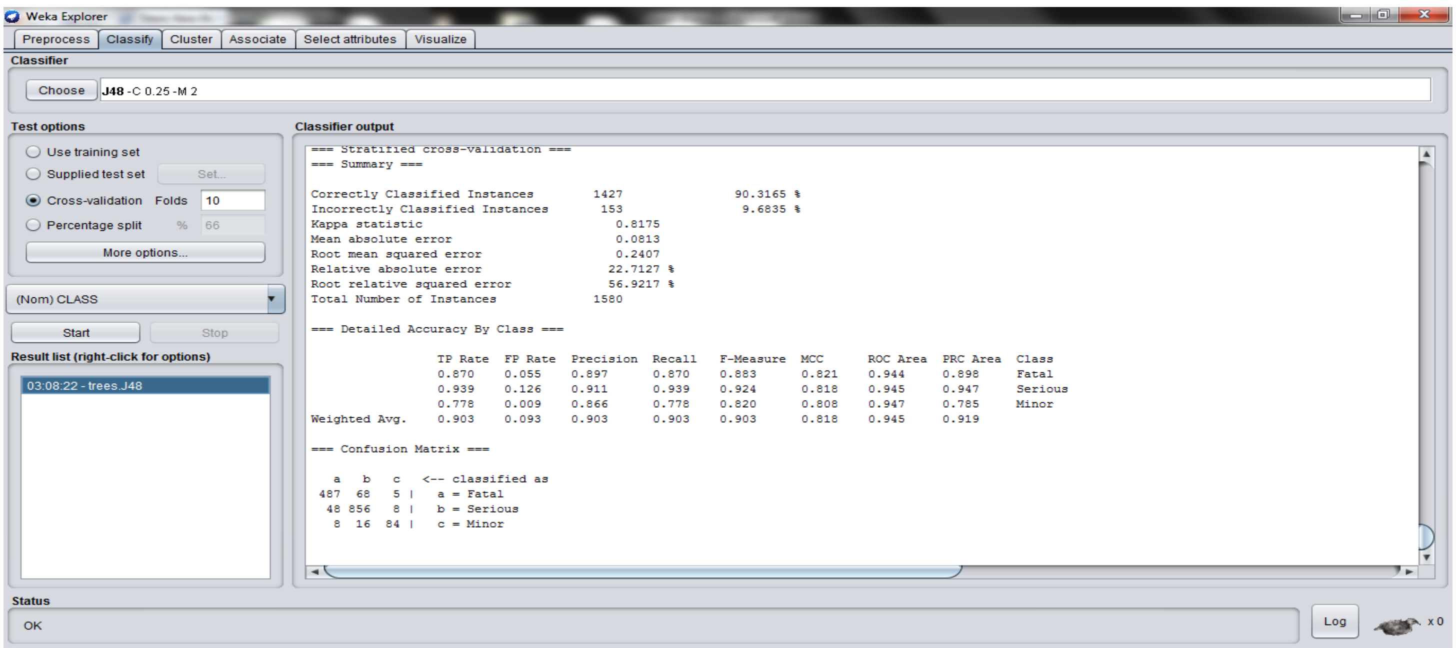

Figure 3: Decision Tree result without feature selection

Table 3: summary result of algorithms without feature selection

S/N	Algorithms	Class	TP	FP	Precision	Recall	F-Measure	Result (%)
		Fatal	0.870	0.055	0.897	0.870	0.883	
1	Decision Tree	Serious	0.939	0.126	0.911	0.939	0.924	90.3165
		Minor	0.778	0.009	0.866	0.778	0.820	
		Fatal	0.941	0.032	0.941	0.941	0.941	
2	K-NN	Serious	0.957	0.055	0.959	0.957	0.901	94.8101
		Minor	0.907	0.088	0.891	0.907	0.892	
		Fatal	0.823	0.076	0.855	0.823	0.834	
3	Jrip	Serious	0.913	0.178	0.875	0.913	0.894	86.3924
		Minor	0.657	0.012	0.798	0.657	0.721	
		Fatal	0.479	0.184	0.588	0.479	0.528	
4	Naïve Bayes	Serious	0.764	0.460	0.694	0.764	0.728	63.7975
		Minor	0.398	0.052	0.358	0.398	0.377	
		Fatal	0.850	0.091	0.837	0.850	0.843	
5	Multilayer Perceptron	Serious	0.904	0.148	0.893	0.904	0.898	86.7722
		Minor	0.657	0.012	0.807	0.657	0.724	

From The experimental results in Table 3, K-NN algorithm had the highest marginal accuracy score of 94.8% for the prediction of RTC classification without feature selection technique. Naïve Bayes (NB) algorithm had the lowest score of 63.8% accuracy.

4.4 Experimental Results of Classifiers with Feature Selection

Here the experimental results of the five data mining algorithms (Decision tree, K-NN, MLP, JRip and NB) on the RTC dataset using feature selection technique are demonstrated. After feature selection using Principal Component Analysis (PCA) technique, the number of attributes were reduced to 15 attributes which involves Command, Year, Crash time, Route, Vehicle type, Vehicle category, Vehicle name, Over speeding, Driving under alcohol, Dangerous driving, Loss of control, Over loading, Route violation, Dangerous overtaking and Sleeping on steering. Decision tree algorithm had 91.3% accuracy with feature selection method. The result of decision tree algorithm is shown in Figure 4. Table 4 shows its confusion matrix with feature selection. The comparison of Percentage Accuracy, Precision, Recall and F – measure of Decision tree with feature selection are shown Table 6.

The comparison of Percentage Accuracy, Precision, Recall and F – measure of KNN with feature selection are shown Table 7. Table 5 shows the K-NN confusion matrix with feature selection. The comparison of Percentage Accuracy, Precision, Recall and F – measure of K-NN with feature selection are shown Table 6

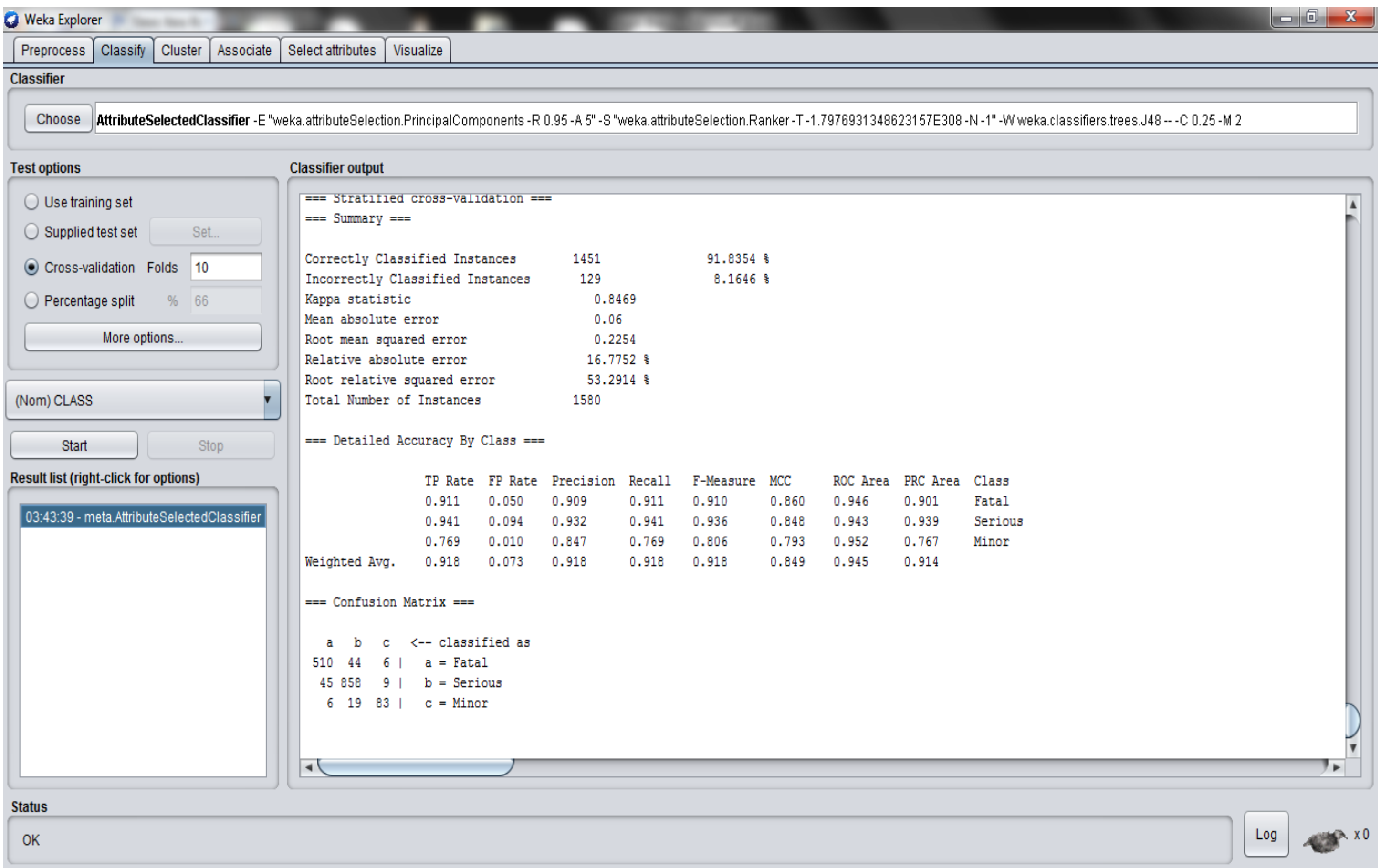

Figure 4: Decision Tree result with feature selection

Table 4: Decision tree confusion matrix with feature selection

A	B	C	Class
510	44	6	A= Fatal
45	858	9	B= Serious
6	19	83	C= Minor

K- Nearest Neighbor had 96.1% accuracy with feature selection method. The result of K-NN is shown in Figure 5.

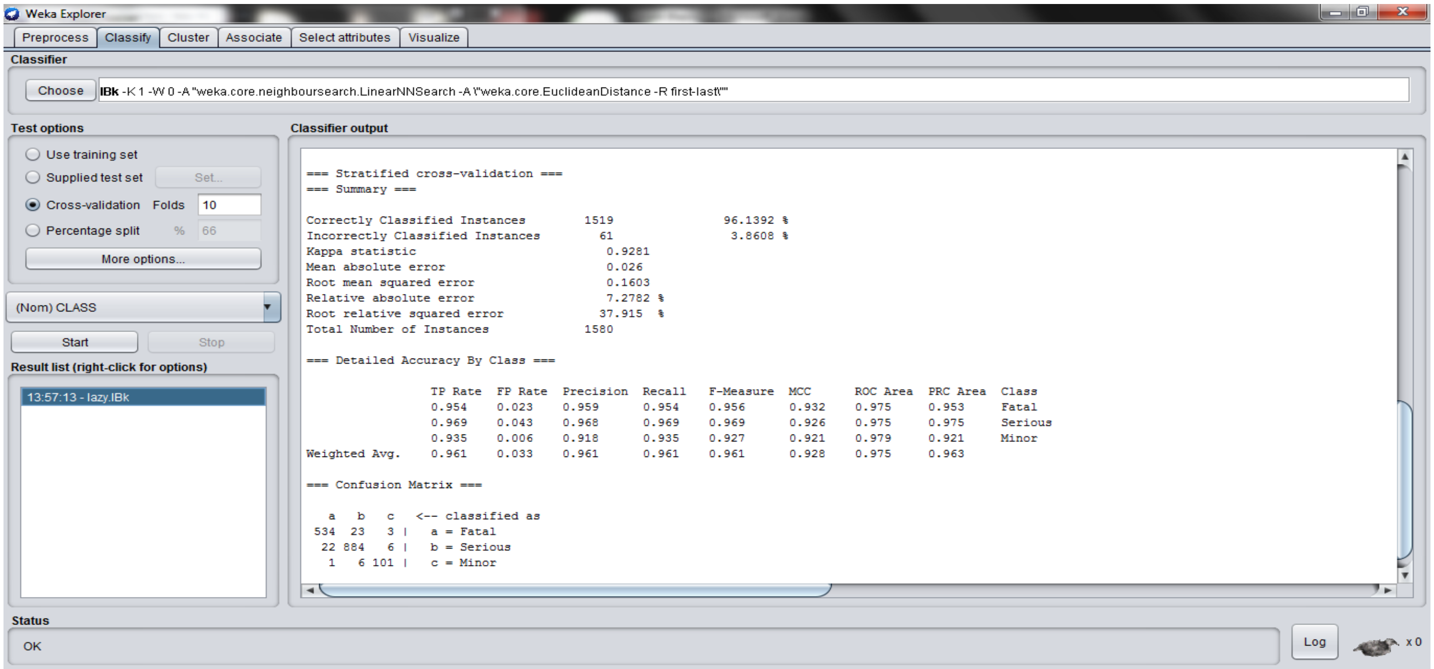

Figure 5: K-NN result with feature selection

Table 5: K-NN confusion matrix with feature selection

A	B	C	Class
534	23	3	A= Fatal
22	884	6	B= Serious
1	6	101	C= Minor

JRip algorithm had 88.6% accuracy with feature selection method. The result obtained from JRip algorithms is shows in Figure 6.

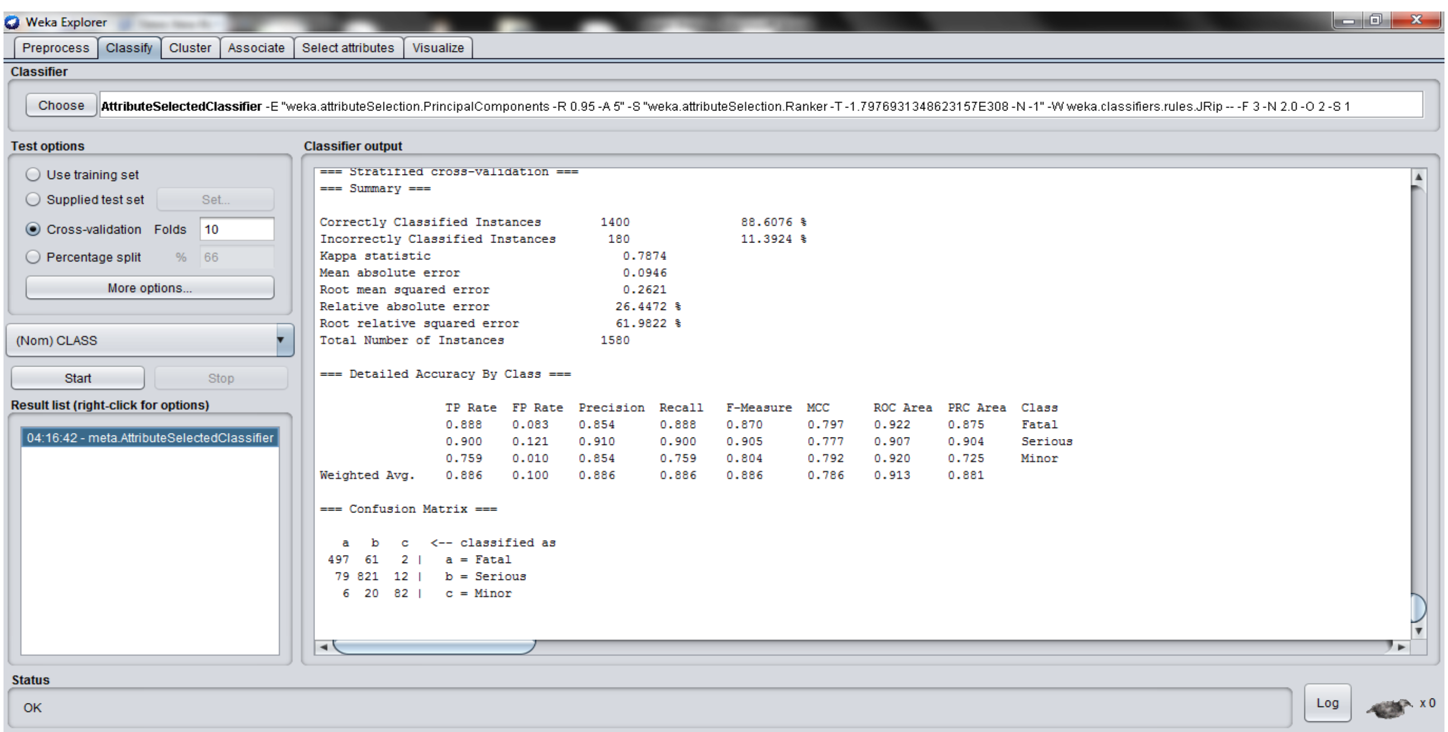

Figure 6: JRip algorithm result with feature selection

Naïve Bayes algorithm had 91.3% accuracy with feature selection method. The result of naïve Bayes algorithms is shown in Figure 7

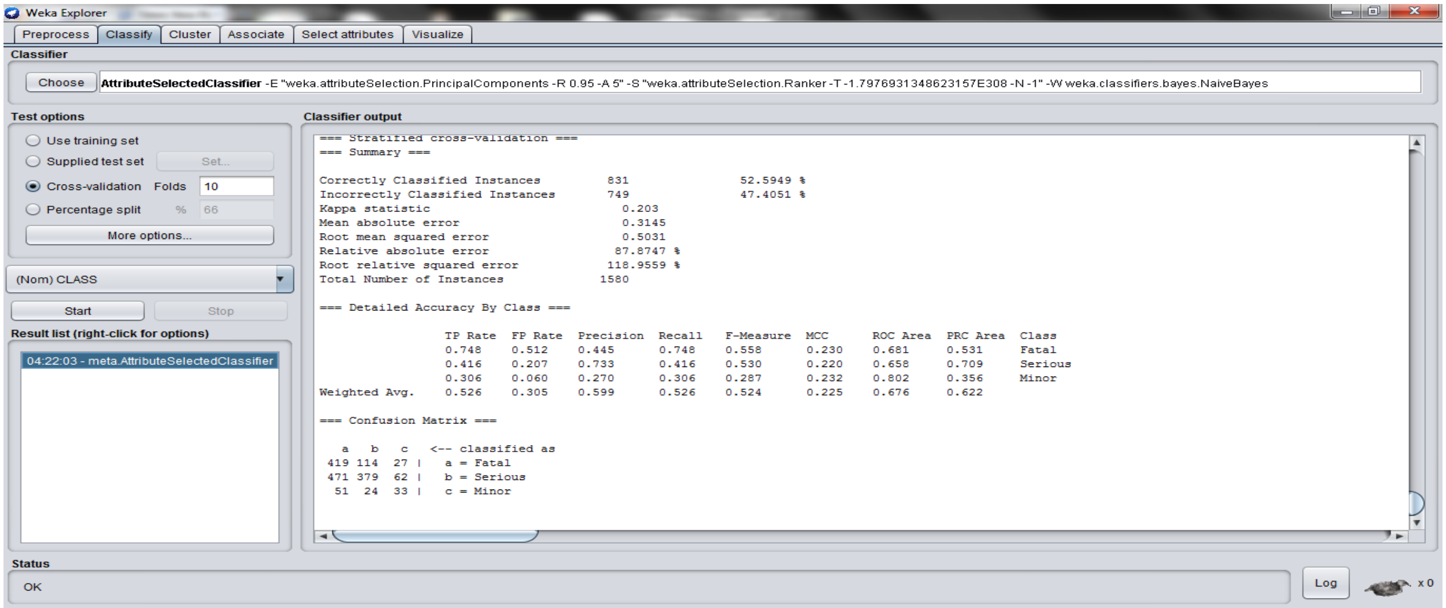

Figure 7: Naïve Bayes algorithm result with feature selection

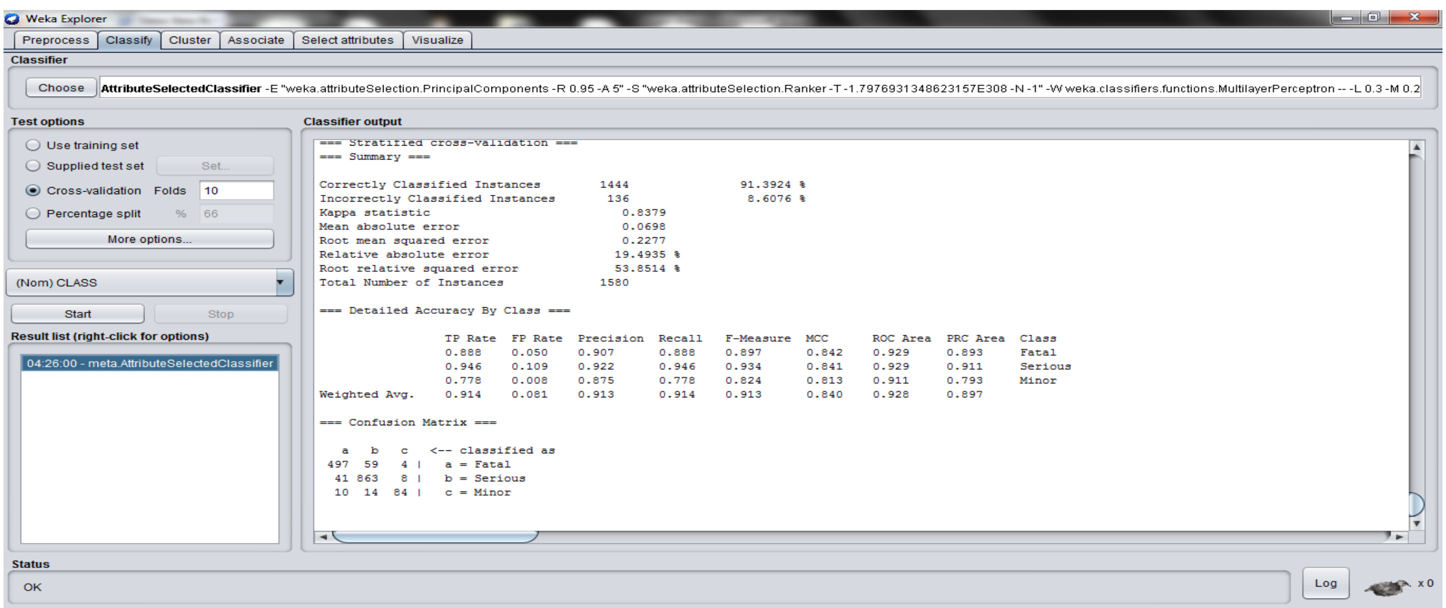

Figure 8: Multilayer Perceptron algorithm result with feature selection

Multilayer Perceptron algorithm had 86.7% accuracy with feature selection method. The result of multilayer perceptron algorithms is shown in Figure 8.

Table 4 and Table 5 indicate the confusion matrices of Decision Tree and K-NN classifiers with feature selection of the RTC dataset for the Fatal, Serious and Minor classes. The decision tree classifier was able to predict accurately 510 out of 560 fatal crash, 858 out of 912 serious crash and 83 out of 108 as minor crash. This gave an accuracy of 91% for fatal crash, 94% for serious crash and 76.9% for minor crash prediction. KNN classifier predicted accurately 534 out of 560 fatal crash, 884 out of 912 serious crash and 101 out of 108 minor crash giving accuracies of 95.3% for fatal crash, 96.9% for serious and 93.5% for minor crash prediction.

Therefore, KNN performed better than other classifiers in this study. The table 6 summarizes the results of the five selected algorithms that were used to predict RTC severity as fatal, serious or minor with applying feature selection technique using the 15 attributes and the number instances of 1580.

Table 6: Summary of performance result algorithms with feature selection

S/N	Algorithms	Class	TP	FP	Precision	Recall	F-Measure	Result (%)
1	Decision Tree	Fatal	0.911	0.050	0.909	0.911	0.910	91.8354
		Serious	0.941	0.094	0.932	0.941	0.936	
		Minor	0.769	0.010	0.847	0.769	0.806	
2	K-NN	Fatal	0.939	0.031	0.943	0.939	0.941	96.1368
		Serious	0.957	0.054	0.960	0.957	0.959	
		Minor	0.907	0.010	0.867	0.907	0.806	
3	JRip	Fatal	0.888	0.083	0.854	0.888	0.870	88.6076
		Serious	0.900	0.121	0.910	0.900	0.905	
		Minor	0.759	0.010	0.854	0.759	0.804	
4	Naïve Bayes	Fatal	0.748	0.512	0.445	0.748	0.558	52.5949
		Serious	0.416	0.207	0.733	0.416	0.530	
		Minor	0.306	0.060	0.270	0.306	0.287	
5	Multilayer Perceptron	Fatal	0.850	0.091	0.837	0.850	0.843	91.3924
		Serious	0.904	0.148	0.893	0.904	0.898	
		Minor	0.657	0.012	0.807	0.657	0.724	

As demonstrated in table 6 K-NN algorithm had the highest accuracy score of 96.1% above the other algorithms for the prediction of RTC classification with Feature Selection Technique, while the Naïve Bayes (NB) had the lowest accuracy score of 52.6%.

4.5 Performance Evaluation

This research provided a comprehensive data mining research that focused on predicting RTC with focus on crash severity of fatal, serious and minor. The study was addressed to fully explore the data mining domain and was targeted to achieve interesting results related to the RTC classification. The results were in line and agree with most of the related literature.

The study also provided a systematic review of literature related to the subject. In the study the best performing classifier on RTC as fatal, serious or minor was K-NN with 96.1% accuracy. The Table 7 shows the performance of other related works compared to this study.

Table 7: Performance Evaluation with Related Studies

S/N	Author	Classifier	Dataset	Method	Software	Result %
1	[7]	decision tree, naive Bayes, and K-nearest neighbor	Ethiopian RTC	Feature selection	WEKA	80.8
2	[9]	Fuzzy granular DT.	California RTC	Fuzzy expression	Python	95.2
3	[11]	Bayesnet, MLP, DT and ANN	United kingdom RTC			72.6
4	[12]	ANN	Abu Dhabi RTC	Feature selection	WEKA	81.6
5	[14]	SVM	FRSC, Birnin Yero Unit Command	No Feature selection	WEKA	74
6	This study (2021)	DT, NB, JRip, MLP, K-NN.	Local data from FRSC in Kaduna State.	Feature selection	WEKA	96.1

5.0 CONCLUSION

This study identified the major significant factors of RTC as well as the significant attributes affecting road crashes in the targeted case study. This study used the three severity classes of fatal, serious and minor to classify the RTC dataset using some selected machine learning techniques. The results showed that K-NN outperformed the other classifiers in the prediction of RTC. The study is believed will assist the policy maker in making rules and regulations that will help the road users eradicate RTC fatality.

REFERENCES

[1] FRSC. "Federal Road Safety Corps 2018 Annual Report". Available online @ https://frsc.gov.ng/wp-content/uploads/2021/01/FRSC-2018-annual-report.pdf. Accessed on March 20, 2021.

[2] Mohammed, A., Ambak K., Mosa M. and Syamsunur, D. "Expert System in Engineering Transportation", *Journal of Transport Technology* Review, 14(2), (2019), 23-31.

[3] WHO. "Global status report on road safety 2018", Geneva:World Health Organization, file:///C:/Users/User-PC/AppData/Local/Temp/9789241565684-eng.pdf, Accessed on December 15, 2020.

[4] Ratanavaraha, V. and Suangka, S. "Impacts of accident severity factors and loss values of crashes on expressways in Thailand", *IATSS research*, 37(2), (2014), 130-136.

[5] Atubi, A. O. and Onokala, P. C. "Contemporary analysis of variability in road traffic accidents in Lagos State, Nigeria". *African Geographical Review*, 28(1), (2019), 11-41.

[6] Martín, L., Baena, L., Garach, L., López, G. and de Oña, J. "Using data mining techniques to road safety improvement in Spanish roads", Procedia - *Social and Behavioral Sciences*, 160(2014), 607 – 614.

[7] Beshah, T. and Hill, S. "Mining Road Traffic Accident Data to Improve Safety: Role of Road-related Factors on Accident Severity in Ethiopia", *In AAAI Spring Symposium: Artificial Intelligence for Development,* 24, (2010), 1173–1181.

[8] Olutayo, V. and Eludire, A. "Traffic Accident Analysis Using Decision Trees and Neural Networks", *International Journal of Information Technology and Computer Science (IJITCS),* 6(2), (2014), 22-28.

[9] Moghaddam, H. K. and Wang. X. "Vehicle Accident Severity Rules Mining Using Fuzzy Granular Decision Tree", *In International Conference on Rough Sets and Current Trends in Computing,* (2014), 280 – 287.

[10] Williams, K., Idowu, A. P. and Olonade, E. "Online Road Traffic Accident Monitoring System for Nigeria", *Transactions on Networks and Communications,* 3(1), (2015), 10-15.

[11] Castro, Y. and Kim, Y. J. "Data mining on road safety: factor assessment on vehicle accidents using classification models", *International Journal of Crashworthiness,* 21(2), (2016), 104-111.

[12] Sharaf, A., Madhar, T., and Salah, T. "Severity Prediction of Traffic Accident Using an Artificial Neural Network". *Journal of Forecasting Department of Computer Science,* 4, (2016), 23-35.

[13] Chimba, D., Ruhazwe, E., Allen S. and Waters, J. "Digesting the safety effectiveness of cable barrier systems by numbers", *Journal of Transportation Research Part A: Policy and Practice,* 95, (2017), 227-237.

[14] Isah, S.B., Radzi, N. H. M. and Mustaffa, N. H. "Road Traffic Crash Severity Classification Using Support Vector Machine". *International Journal of Innovative Computing and Applications,* 7(1), (2017), 15 – 18.

[15] Salas, A. G., and Nwagboso, A. A. Traffic Event Detection Framework Using Social Media. *IEEE International Conference on Smart Grid and Smart Cities*. Wolverhampton: IEEE. (2017), 45-79

[16] Kebede, T., Singh D. K. & Tessfaw E. A. "Comparative Study on Data Mining Classification Algorithms for Predicting Road Traffic Accident Severity", *Proceedings of the 2nd International Conference on Inventive Communication and Computational Technologies,* 8, (2018), 24-27.

[17] Somayeh, M., Yassir, A., and Ryan, D. "Machine Learning Methods to Analyze Injury Severity of Drivers from Different Age and Gender Groups", *Transportation Research Record* (2018), 1–13.

[18] Evwiekpaefe, A.E., Bitrus, E. and Ajakaiye, F. "Selecting Forward Players in a Football Team using Artificial Neural Networks", *International Journal of Computer Applications*, 176(28), (2020), 8-13.