International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 10, October-2021

A Comparative Study on Machine Learning-based Approaches for Improving Traffic Accident Severity Prediction

Jovial Niyogisubizo[1], Evariste Murwanashyaka[2], Eric Nziyumva[1]
[1]Fujian Key Lab for Automotive Electronics and Electric Drive, Fujian University of Technology, Fuzhou, China.
[2]Institute of Rock and Soil Mechanics, University of Chinese Academy of Sciences, Wuhan, China

***Abstract*: This Traffic accidents are the leading cause of many deaths, property damages, injuries, and fatalities as well as financial losses every year. Accurate traffic accident severity prediction would be very crucial to evaluate the major determinants associated with road accidents, offer precautions before occurrence based on the predicted outcomes and thus minimize all negative impacts caused by accidents. In the past decades, traditional techniques and machine learning have been used to predict traffic accidents. However, machine learning models are criticized because they perform like “black box” and lack interpretations for humans. The main purpose of this research is to employ machine learning-based approaches to predict crash injury severity and analyze the most influential factors contributing to road crashes as well as giving recommendations to concerned stakeholders. In this study, four classification approaches were employed: Random Forest (RF), Multinomial Naïve Bayes (MNB), K-Means Clustering (KC), and K-Nearest Neighbors (KNN) to predict accident severity and analyze feature importance. On the road accident dataset from 2015 to 2020 provided by the State of Victoria in Australia, the RF outperformed the remaining methods in terms of accuracy, precision, recall, and F1 Score. Month, time of day, female drivers, male drivers, total persons, speed zone, day of the week, passengers, etc., were found as the major determinants of accident severity. The accuracy enhanced model can help in giving recommendations such as safe route planning, preparing emergency vehicle allocation, reducing property damage, placing additional signage where necessary, and roadway design to concerned stakeholders to eradicate the number of fatalities and injuries resulting from traffic accidents.**

Keywords:* *Traffic accident severity; Random forest, Multinomial Naïve Bayes; K-Nearest Neighbors; Traffic safety and Feature importance.

I. INTRODUCTION

Globally, traffic accidents constitute the major cause of injuries, death, property damage, and disability with a disproportionate number occurring in developing countries. According to World Health Organization, around 1.2 million people die each year due to road accidents and nearly half of them are pedestrians, motorcyclists, and cyclists who are less protected. According to several reports, it is expected that traffic accidents will become the leading cause of fatalities by 2030 due to the lack of sustainable transportation[1].

Traffic safety studies are very crucial to several stakeholders to avoid delays for road users, property damage, reduce health costs and ensure better transportation safety. In the past decades, several studies of traffic safety focused on traffic injury severity prediction and analysis of significant factors influencing traffic crash severity. Traditionally, different classical statistical techniques have been used to predict traffic accidents severity. Among statistical methods, the Ordered Probit (OP) model was developed to analyze crash injury severity on datasets with different sample sizes from the 2003 National Automotive Sampling System General Estimates System. The overall results showed that the Bayesian OP outperformed the OP using a small sample size[2].

Besides, the three commonly used approaches such as OP, Multinomial logit (MNL), and mixed logit (ML) models were compared for crash severity modeling on sample size requirements. The results showed that huge sample size is required for the ML model, small sample size is required for the OP model while the MNL model requires the sample size located between the OP and MNL models [3]. Furthermore, ordered multiple-choice was developed to predict moto vehicle crash injury severity using the dataset provided by New South Wales, Australia. The results showed that the rises in the age of the victim and vehicle speed lead to slight increases in the probabilities of serious crashes and fatalities while other factors such as vehicle type, blood alcohol level, seating position, vehicle make and type of collision also have a significant impact on crash severity[4].

Classical parametric techniques such as probit and logit models have been widely utilized to predict traffic accident severity because the severity of vehicular crashes is random. However, these parametric techniques suffer from several limitations. For example, when the dataset contains missing values and different outliers, the output of the prediction model will be negatively affected. Besides, these techniques need a predefined mathematical form to function accurately. To handle the limitations arising from the use of traditional statistical techniques, machine learning (ML) approaches have been employed to deal with nulls and missing values in the dataset. ML models have the ability to dig useful information from huge traffic accident datasets for several road networks.

In this section, the literature about the ML approaches relating to accident severity prediction is presented. A comparative study on different ML algorithms such as logistic regression (LR), classification and regression tree (CART), and random forest (RF) was conducted to model road accident severity and identify the significant variables that influence accident severity. The results showed that RF produces improved prediction performance in terms of accuracy, sensitivity, and specificity[5]. Similarly, the performance of two classical statistical techniques namely OP and MNL models were with four ML methods such as the k-nearest neighbor (KNN), decision tree (DT), random forest (RF), and support vector machine (SVM) to predict crash injury severity.

On the traffic crash dataset of Florida, the results show that ML approaches produce better prediction performance than the classical statistical techniques in terms of prediction accuracy. However, some ML approaches suffered from the issues of overfitting. Among all methods, the RF achieved the overall enhanced results while the OP is the poor performer amongst the group[6].

Besides, the KNN method was employed to predict real-time highway traffic crashes. Before categorizing road patterns, the traffic crash precursors and their calculation time slice duration were determined. The conclusions of this study demonstrated that the KNN produced better results when compared to the conventional C-means clustering approach[7]. Moreover, the KNN was compared with hazard-based models to predict the incident duration. Using an incident dataset from the BBC for the Greater London area, both KNN and hazard-based models have demonstrated the ability to produce accurate incident duration prediction. However, these methods failed to show comprehensiveness in illustrating the performance of these two methods[8].

To account for heterogeneity in accident data K-Means Clustering algorithm was used to analyze patterns of vehicle collisions before and after analysis[9]. Moreover, the KC and Kernel density estimation (KDE) was used to identify road accident hotspots. On the road accidents data in UK, London provided by the Metropolitan Police from 1999 to 2003, Geographical Information Systems and KDE were employed to explore the spatial patterns of accident-related factors. The KC has demonstrated the ability to analyze the major determinants of accident severity in different hotspot cells[10].

Although machine learning approaches have demonstrated the ability to outperform classical statistical techniques, they are criticized because they employ a "black box" tactic to predict traffic accident severity and lack proper interpretation of the model for humans[11]. Comparing ML models and traditional techniques, ML approaches are more accommodating with no or little presumptions for explanatory variables[12].

To address the issues presented in the literature, the goal of this research is to assess the application of Random Forest (RF), Multinomial Naïve Bayes (MNB), K-Means Clustering (KC), and K-Nearest Neighbors (KNN) models to predict accident severity. Besides, feature analysis is conducted through feature importance to identify and examine potential factors contributing to crash injury severity using traffic accident data. Apart from the analysis of feature importance, several recommendations such as safe route planning, preparing emergency vehicle allocation, reducing property damage, placing additional signage where necessary, and roadway design are provided to concerned stakeholders to eradicate the number of fatalities and injuries resulting from traffic accidents.

The main contributions of this paper are summarized as follows:

- This study is aimed at filling the gap in the absence of implementing machine learning approaches in accident severity prediction. The innovation behind this contribution is that the model like Multinomial Naïve Bayes known in the field of text mining is firstly developed to predict traffic accident severity,
- A feature importance study is conducted in this research to analyze the significant factors contributing to accident severity and provide recommendations to concerned stakeholders to ensure better safety. This contribution mitigates the issue of lack of interpretability presented in the literature as the major intrinsic limitation of tree-based classifiers,
- An ensemble-based method considered among the robust machine learning approach is employed to enhance the prediction performance and easily bring about nonlinear classification methods with better generality.

II. METHODOLOGY

The purpose of the methodology proposed in this paper is to enhance the prediction ability of traffic accidents using the best-performing model (KNN, MNB, RF, and KC). The overall design of the procedure followed to build our model is illustrated in Fig. 1. Firstly, the traffic accident dataset is collected. After collecting the accident dataset, the most important stage is data preprocessing where the dataset is imputed by replacing NaN and missing values with the most frequent values of the corresponding column. Additionally, all the categorical values have been labeled by integers from 0 to n for each column in the given dataset. Accident date has been converted to a categorical feature with 2 values i.e., month, time of day. Moreover, the dataset is visualized for correlation. The negatively correlated variables are selected to be removed. The next stage after data preprocessing is feature selection. At this stage, feature importance is plotted graphically to visualize the effect of contributing factors, and only attributes with high importance are taken into consideration for accident severity prediction and model building.

Due to the limited number of fatal and serious injury accidents, we decided to merge the minor class for accurate prediction. Therefore the derived new severity levels are injury and Serious/Fatal accidents. Fig. 2 represents the severity levels used in the analysis and prediction of traffic accidents. From this figure, 44909 (58.01%) are injury accidents while 32507 (41.99%) are Serious/Fatal accidents. Before moving to the next stage of model development, 80% of the accident dataset was selected to be used as a training set and the remaining 20% was used as a testing set. Later, the best performing models are developed based on parameter optimization and 10-fold cross validation was implemented. The overall prediction results are compared based on four performance indicators namely, accuracy, precision, recall, and F1 score.

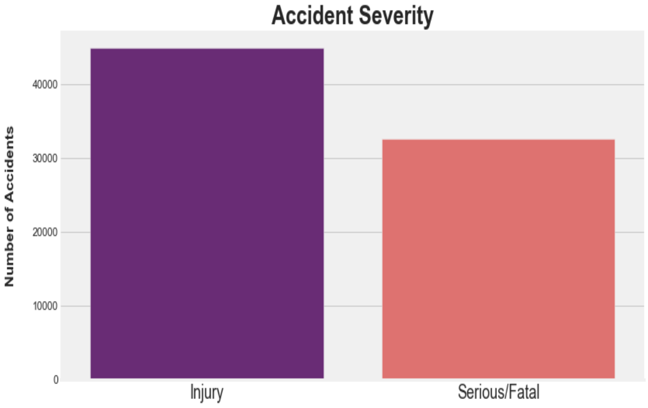

Fig. 2. Accident severity level.

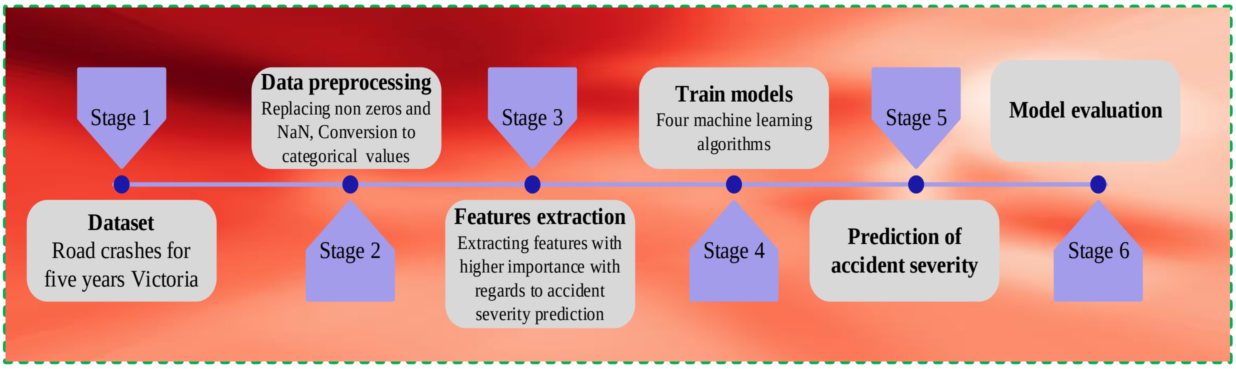
Stage 1
Dataset
Road crashes for five years Victoria
Data preprocessing
Replacing non zeros and NaN, Conversion to categorical values
Stage 2
Stage 3
Features extraction
Extracting features with higher importance with regards to accident severity prediction
Train models
Four machine learning algorithms
Stage 4
Stage 5
Prediction of accident severity
Model evaluation
Stage 6

Fig. 1. Model building processing

III. DATASET SOURCE AND DESCRIPTION

The road accident data used in this study contains accidents statistics for the last five years from 2015 to 2020 for the State of Victoria in Australia. The dataset was provided by the department of transport through its open data website. The raw data consisted of 77416 traffic accidents with 71 features. The accident severity label was classified into four levels: other injury, serious injury, fatal and non-injury accidents. Among all the traffic accidents, 44906 (58.01%) were other injury accidents, 30916 (39.93%) were serious injury accidents, 1591 (2.06%) were fatal accidents and, only 3(0.001%) were non-injury accidents. Table 1 describes the statistical description of all screened attributes including one target variable (Severity) and potential contributing features such as accident status, accident type, light conditions, day of the week, road geometry, etc.

TABLE I. ATTRIBUTES DESCRIPTION.

Variables	**Description**	**Accident severity level**		
		Injury (%)	**Serious/Fatal (%)**	**Total**
Severity		44909 (58.01)	32507 (41.99)	77416
Accident status	Discarded	14 (0.02)	9 (0.01)	23
	Finished	43532 (56.23)	31723 (40.98)	75255
	Private Property	617 (0.80)	679 (0.88)	1296
	Reopened	12 (0.02)	4 (0.01)	16
	Unfinished	734 (0.95)	92 (0.12)	826
Alcohol time	No	29718 (38.39)	19842 (25.63)	49560
	Yes	15191 (19.62)	12665 (16.36)	27856
Accident type	Collision with a fixed object	5680 (7.34)	6797 (8.78)	12477
	Collision with vehicle	29757 (38.44)	17308 (22.36)	47065
	Fall from or in moving vehicle	377 (0.49)	364 (0.47)	741
	No collision and no object struck	2170 (2.80)	1676 (2.16)	3846
	Other accident	50 (0.06)	43 (0.06)	93
	Struck Pedestrian	3725 (4.81)	3733 (4.82)	7458
	Struck animal	470 (0.61)	297 (0.38)	767
	Vehicle overturned (no collision)	2126 (2.75)	1878 (2.43)	4004
	Collision with some other object	554 (0.72)	411 (0.53)	965
Day of week	Friday	7035 (9.09)	5049 (6.52)	12084
	Monday	6449 (8.33)	4360 (5.63)	10809
	Saturday	4501 (5.81)	4394 (5.68)	8895
	Sunday	6164 (7.96)	4842 (6.25)	11006
	Thursday	7066 (9.13)	4782 (6.18)	11848
	Tuesday	6825 (8.82)	4374 (5.65)	11199
	Wednesday	6869 (8.87)	4706 (6.08)	11575
Hit run flag	No	41721 (53.89)	31024 (40.07)	72745
	Not known	202 (0.26)	159 (0.21)	361
	Yes	2986 (3.86)	1324 (1.71)	4310
Light condition	Dark No street lights	1955 (2.53)	2333 (3.01)	4288
	Dark Street lights off	85 (0.11)	78 (0.10)	163
	Dark Street lights on	6468 (8.35)	5244 (6.77)	11712
	Dark Street lights unknown	556 (0.72)	249 (0.32)	805
	Day	30342 (39.19)	21438 (27.69)	51780
	Dusk/Dawn	3927 (5.07)	2712 (3.50)	6639
	Unknown	1576 (2.04)	453 (0.59)	2029
Police attend	No	15069 (19.46)	3995 (5.16)	19064
	Not known	195 (0.25)	89 (0.11)	284
	Yes	29645 (38.29)	28423 (36.71)	58068
Road geometry	Cross intersection	10011 (12.93)	5932 (7.66)	15943
	Dead end	57 (0.07)	48 (0.06)	105
	Multiple intersection	929 (1.20)	599 (0.77)	1528
	Not at intersection	22022 (28.45)	18067 (23.34)	40089
	Private property	786 (1.02)	600 (0.78)	1386

	Road closure	2(0.00)	2 (0.00)	4
	T intersection	9779 (12.63)	6580 (8.50)	16359
	Unknown	1219 (1.57)	622 (0.80)	1841
	Y intersection	104(0.13)	57 (0.07)	161
Run off road	No	38461(49.68)	24953 (32.23)	63414
	Yes	6448 (8.33)	7554 (9.76)	14002
Pillion	0.0	44747 (57.80)	32261 (41.67)	77008
	1.0	161(0.21)	242 (0.31)	403
	2.0	1 (0.00)	4 (0.01)	5
Alcohol related	No	44085 (56.95)	30907 (39.92)	74992
	Yes	824 (1.06)	1600 (2.07)	2424
Unlicensed	0.0	43555 (56.26)	31254 (40.37)	74809
	1.0	1346 (1.74)	1243 (1.61)	2589
	2.0	8 (0.01)	10 (0.01)	18
Divided	Divided	16519 (21.34)	10982 (14.19)	27501
	Undivided	28390 (36.67)	21525 (27.80)	49915
Time of day	Morning Rush (6-10)	8540 (11.03)	5856 (7.56)	14396
	Day (10-12)	5049 (6.52)	3720 (4.81)	8769
	Lunch Rush (12-14)	5706 (7.37)	3914 (5.06)	9620
	Afternoon (14-16)	6751 (8.72)	4610 (5.95)	11361
	After Work Rush (16-18)	7456 (9.63)	4952 (6.40)	12408
	Evening (18-22)	7415 (9.58)	5501 (7.11)	12916
	Night (22-6)	3992 (5.16)	3954 (5.11)	7946

IV. CLASSIFICATION MACHINE LEARNING MODELS

Since we aim at predicting accident severity class, classification machine learning models are the best alternative for attaining our goal. Classifiers are supervised machine learning techniques used for assigning an accident class to new hidden observations. To predict a given class of any accident, supervised machine learning algorithms are trained based on crash data of already recognized observations. This means that the input features and the target attribute are provided during the prediction period. It is worth mentioning that the implementations and experiments are executed using the python programming language and its popular package known as such as Sklearn.

Below is a brief introduction of classification machine learning techniques employed in the research.

A. K-Nearest Neighbor (KNN)

KNN was first introduced by Cover and Hart[13]. This model aims at categorizing an instance by considering the closest K observations found in the feature space. To achieve an accurate prediction using KNN, four procedures are followed. Firstly, the distance between the target variable and other popular observations is computed. Besides, the K neighbors that are close to the target considering the calculated distance are extracted. The next process is searching the value of K neighbors and finally, the predicted value of the target attribute is calculated. Generally, the classification output of the predicted variable is the mode value of the K neighbors. During the prediction phase, it is advised to select the suitable K value to ensure better prediction performance. Furthermore, the Euclidean distance must be employed as the physical distance function to calculate distance between two observations[14], as shown in the following equation:

$$d_{ij} = \sqrt{\sum_{k=1}^{m} \left(x_{ik} - x_{jk}\right)^2} \qquad (1)$$

Where m is the number of independent features, x_{ik} and x_{jk} are the values of the k^{th} independent features for observations i and j respectively.

B. Random Forest (RF)

RF is an ensemble learning approach for classification and regression which generates many classifiers and aggregates their results at training time[15]. This method was developed as an efficient prediction tool composed of a set of tree-structured classifiers with independent identically distributed random vectors, while each tree casts a unit vote for the most popular class at input[16]. RF consists of multiple tree predictors and uncorrelated decision trees functioning as an ensemble classifier to enhance the prediction outcome. In RF a bootstrap aggregation principle is followed to get uncorrelated trees. This means that a subset of training samples is created through replacement. In RF model, the Cross validation is employed to reduce estimation error, the out-of-bag error, and build the most reliable trees. Furthermore, all features are exploited utilizing the randomness approach. One of the advantages of RF is to grow a large number of trees to produce trees that have large variance and reduce the issues of bias. After growing trees, the class of new observation is created by averaging the class assignment to all decision trees[17].

C. K-means Clustering (KC)

KC is one the most popular unsupervised machine learning algorithm used in signal processing, image processing, statistical data analysis, and information retrieval, which combines the observations of a dataset in K clusters, based on the physical distance of the observations from clusters' means[18]. When the dataset consists of m variables, the KC assigns the observations to the clusters with the nearest average in m-dimensional space. Hewson [14] revealed that when a number of m random starting points is provided, the dataset variables are nucleated, the averages recomputed and the same procedure continues until the stability point is attained. As was stated above in the KNN, the same procedure of selecting the suitable K value is followed to ensure better prediction performance. Furthermore, the Euclidean distance function (Eq. (1)) was used as the physical distance function to calculate the distance between two observations.

D. Multinomial Naïve Bayes (MNB)

MNB classifier is a supervised machine learning approach that uses probability and focused on the domain of text mining. In the MNB the principle of multinomial distribution in conditional probability is followed[19]. Besides, the method has played a significant role in classifying semi-automated document tasks such as spam mails detection. The MNB has the potential to transform text cases to a nominal form that can be computed with an integer value. Though the development of the MNB considers the naïve assumptions, there is more proof that this approach is very accurate in practice. One of the major differences between the Naïve Bayes (NB) and the MNB is that the MNB operates on the frequency distribution of all packet sizes at once whereas the NB estimates the probability of class membership using Gaussian kernels, thus choosing the class, whose occurrence frequencies of the various packet sizes match best with the observed values in the test instance[20]. Eventually, the calculation of conditional probabilities in the MNB is denoted by Eq. 2 as shown below:

$$P(\mathbf{f}|c_i) \sim \prod_{j=1}^{m} P\left(X = x_j|c_i\right)^{f_{xj}} \qquad (2)$$

Such that m represent the unique sizes present within the sum of all training instances for every class. The overall probability is proportional to the product of $P\left(X = x_j|c_i\right)$, which denotes the probability that any packet size x_j is taken from the aggregated multiset of the totality packet size value counts of the training examples of class c_i. The individual conditional probabilities will significantly affect f_{xj} times to the outcome, for which f_{xj} represents the number of circumstances of packet size x_j in the unlabeled test example.

E. Performance measurement

In this research, the contingency table (confusion matrix) and its related evaluation metrics such as accuracy, precision, recall, and F1 Score are the parameters used to assess the classifier accuracies proposed in this paper. A column in the confusion matrix represents the predicted class instances, a row represents the actual class instances, while the diagonal denotes the prediction accuracy. Table 2 demonstrates the confusion matrix used to compute the metrics used to assess model performance

TABLE II. CONFUSION MATRIX.

		Predicted class	
		Yes	**No**
Actual class	Yes	True Positive (TP)	False Negative (FN)
	No	False Positive (FP)	True Negative (TN)

The formulas used to calculate the metrics used in this study are demonstrated in the following equations:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FN + FP} \qquad (3)$$

$$\text{Precision} = \frac{TP}{TP + FP} \qquad (4)$$

$$\text{Recall} = \frac{TP}{TP + FN} \qquad (5)$$

$$\text{F1 Score} = \frac{2 * Recall * Precision}{Recall + Precision} \qquad (6)$$

Where TP and TN are correctly classified instances. A FP is when the output is wrongly classified as “Yes” while a FN is when the output is wrongly classified as “No”[21]

V. EXPERIMENTAL RESULTS AND DISCUSSION

This section presents the analysis and experimental results for the four classifiers machine learning models namely, Random Forest, Multinomial Naïve Bayes, K-Means Clustering, and K-Nearest Neighbors. A comparative analysis is conducted using four evaluation metrics to see the method which provides better performance on the prediction of crash injury severity. In the experiments carried out in this study, a brief introduction on how the algorithms proposed have been parameterized is described below:

To train the RF, some parameters need to be optimized. For example, the number of trees to grow and the number of variables randomly sampled as candidates at each split are important parameters that need to be calibrated. In this study, when approaching 500 trees during the training process, the best accuracy result was produced and we decided to use the RF model constructed with 500 trees. For MNB, hyper-parameter tuning has not been conducted since the model has no parameter to tune. Concerning the KNN and KC models, different K values are tested to evaluate the prediction performance and find the enhanced prediction results. In this research, we calculated the results by increasing the K value from 1 to 25 and the best accuracy was achieved when K=15.

The overall classification accuracies for each model used in this study are shown in Table 3. The overall training accuracy results of KNN, MNB, KC, and RF are 74.25, 73.62, 60.93, and 81.65 respectively, in which the RF model achieved better training accuracy of 81.65 while the lowest training accuracy is 60.93 for the KC. Besides, the testing accuracy for all models is 60.03, 73.68, 61.07, and 76.72 respectively, in which the RF model produced better testing accuracy of 76.72 while the lowest testing accuracy was 60.03 for the KNN. The KNN suffers from overfitting.

TABLE III. CLASSIFICATION OF DIFFERENT MODELS.

Model	**Training Accuracy (%)**	**Testing Accuracy (%)**
K-Nearest Neighbors	74.25	60.03
Multinomial Naïve Bayes	73.62	73.68
K-means clustering	60.93	61.07
Random Forest	**81.65**	**76.72**

Fig. 3 shows graphically the training and testing accuracy described above for all models.

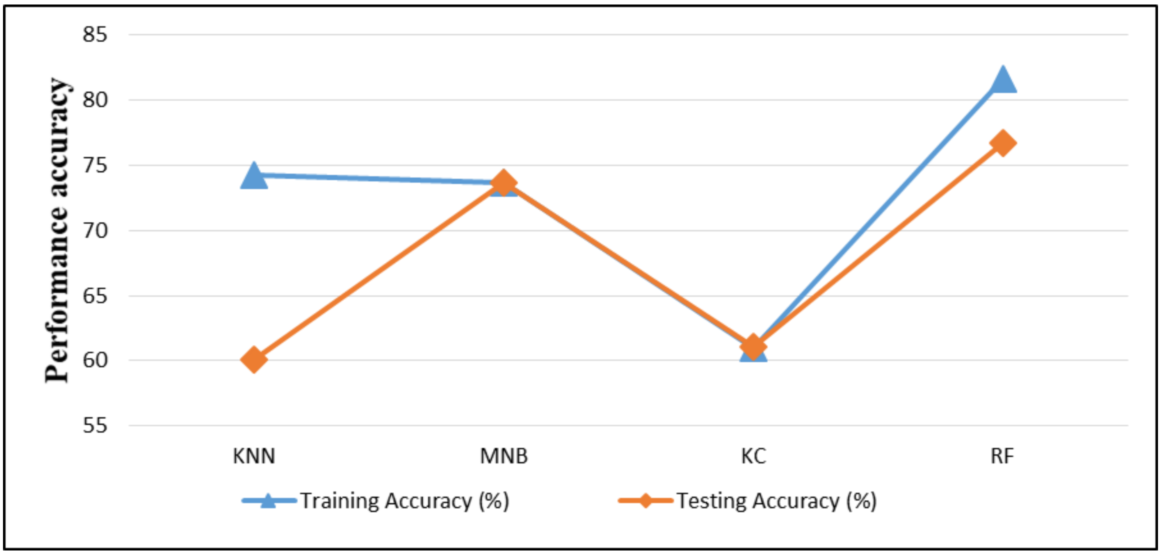

Fig. 3. Training and testing accuracy for all models.

Though accuracy is a performance indicator that shows the performance of a single model, considering only this metric to evaluate model performance can be misleading. For example, the model performance might be biased toward the major class and ignore the minor class. To address this issue, other performance metrics such as recall, precision, and F1 score were determined. These evaluation indicators determine the prediction performance of each severity level, offering better insights into proposed algorithms. The performance results of these metrics for injury accidents class are represented in Table 4. The overall precision results of injury accidents range from 0.65 to 0.73. The MNB and KC show the same prediction results. The RF outperforms all models in the group while the KNN produced lower precision results.

Additionally, the recall values range from 0.64 to 0.72, in which the RF also produces better results. Generally, the precision indicator measures the quality or exactness of the model, whereas the recall measures the quantity or the completeness of the model. This means that high recall indicates that the model produced the most relevant performance results while high precision shows that the model produced more relevant prediction results than irrelevant results and vice versa. Furthermore, the F1 Score which employs both recall and precision is considered an efficient evaluation indicator while interpreting the model's performance. In this research, the F1 Score for KNN, MNB, and KC are almost similar. The RF produces better results among all models. The performance results of the RF show acceptable results and without hesitation, the RF is an algorithm of choice for this kind of data.

TABLE IV. PERFORMANCE MEASURES OF MODELS FOR INJURY ACCIDENTS.

Model	Precision	Recall	F1 Score
K-Nearest Neighbors	0.65	0.68	0.66
Multinomial Naïve Bayes	0.68	0.70	0.67
K-means clustering	0.68	0.64	0.66
Random Forest	0.73	0.72	0.69

Fig. 4 shows graphically performance measures for the test dataset of injury accidents described above for all models.

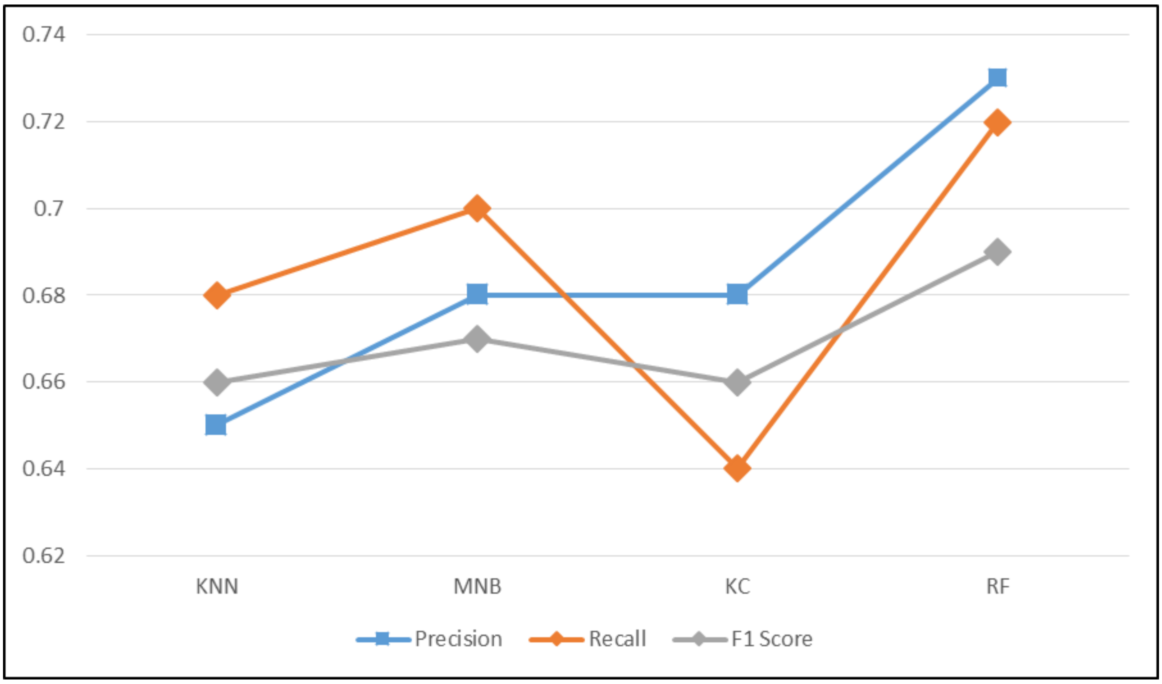

Fig. 4. Performance measures for test dataset of injury accidents.

The performance results of all models for Serious/Fatal accidents are illustrated in Table 5. The precision results of all models for Serious/Fatal accidents range from 0.52 to 0.79, in which the RF produced better precision (0.79) while the KNN achieved lower precision (0.52). Similarly, the recall value ranges from 0.49 to 0.76, in which the RF produced a better recall of 0.76. Practically, the recall value of 0.76 produced by RF on Serious/Fatal accidents means that we are able to predict nearly 76% of Serious/Fatal accidents while the precision value of 0.79 on the same severity level means that we are correct about those predictions about 79%. Furthermore, for the F1 Score which is the harmonic mean of recall and precision or the tradeoff between recall and precision, the RF also produced better results (0.77) while the KNN produced lower results among the group (0.52) on the test dataset. The improved prediction performance of the RF model indicates that it is the model of choice for predicting accident severity.

TABLE V. PERFORMANCE MEASURES OF MODELS FOR SERIOUS/FATAL ACCIDENTS.

Model	Precision	Recall	F1 Score
K-Nearest Neighbors	0.52	0.49	0.51
Multinomial Naïve Bayes	0.68	0.70	0.69
K-means clustering	0.53	0.58	0.55
Random Forest	0.79	0.76	0.77

Fig. 5 shows graphically performance measures for the test dataset of Serious/Fatal accidents described above for all models.

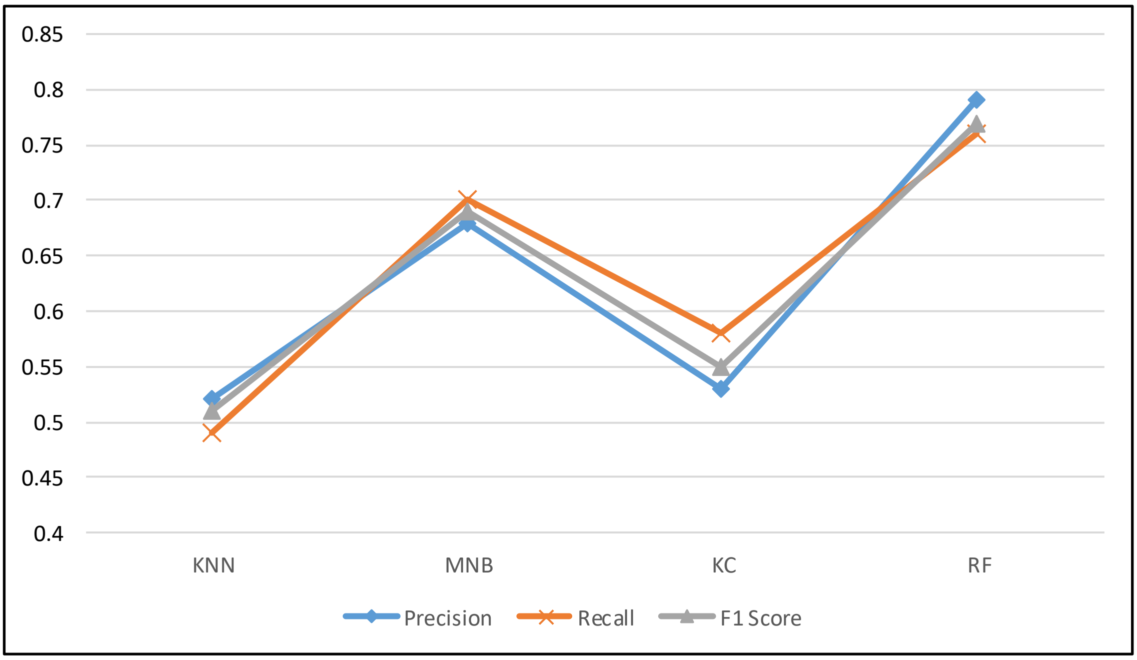

Fig. 5. Performance measures for the test dataset of Serious/Fatal accidents

VI. FEATURE IMPORTANCE

One of the main objectives of this study is to assess the relative feature importance with random forest to interpret the contribution of every feature on model performance while predicting accident severity. Generally, the results of tree-based models such as random forests are not interpreted easily by humans. Fig. 6 represents the relative feature importance conducted to tackle the issue of lack of interpretability for humans while using RF.

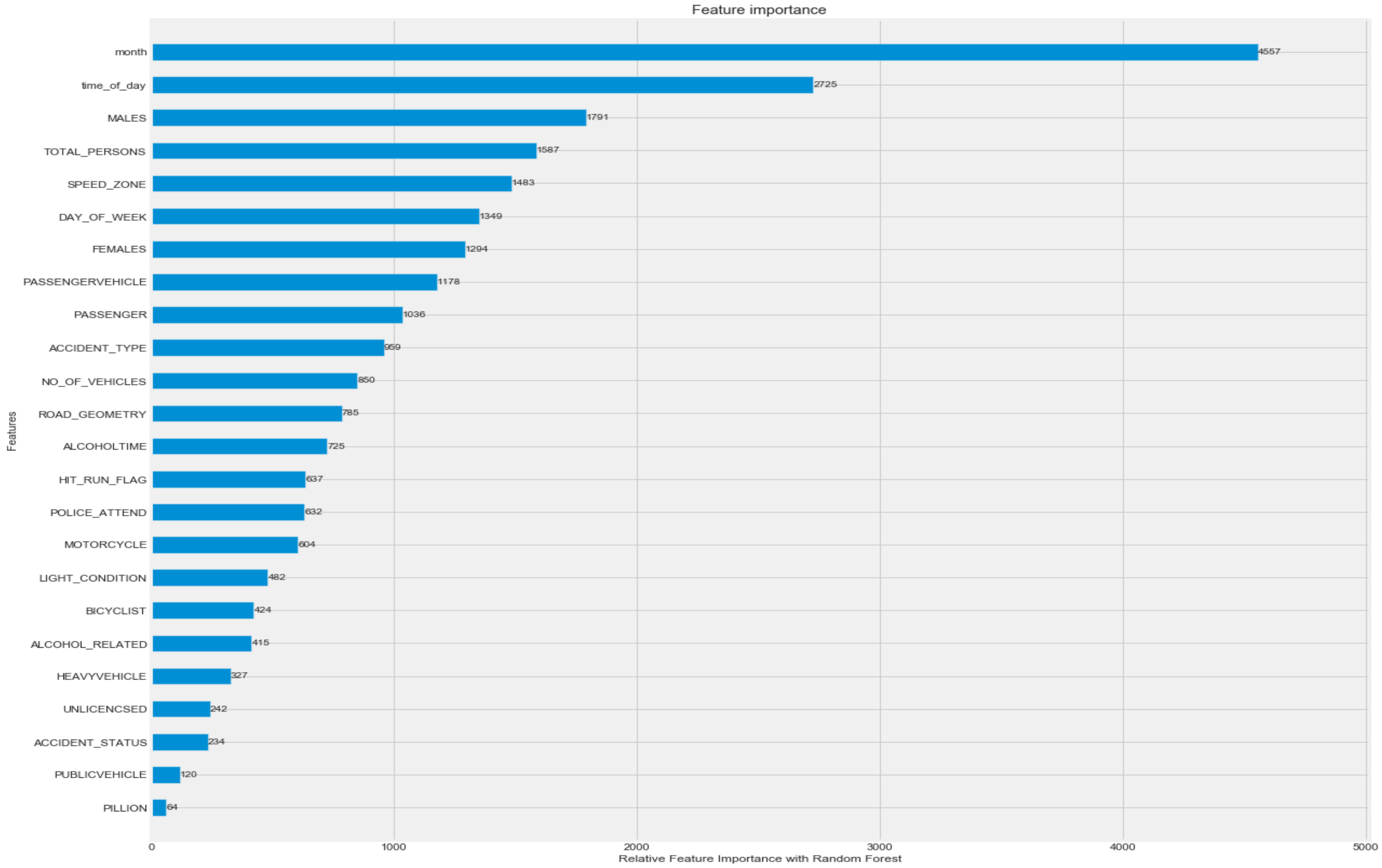

Fig. 6. Relative feature importance with Random Forest

The month of the year prevails as the most important feature when predicting accident severity. This makes sense because certain months see poorer weather conditions, less focus by drivers. Therefore different stakeholders should consider safety programs, increased traffic accident avoidance efforts in problematic times of the year. Besides, the time of day is the second most influential variable. Naturally, the time of day influences accident severity due to traffic congestions occurring at different times of the day. For example during the morning when different people are moving to their jobs, schools, and during the evening time when they are coming back.

The number of male drivers, total persons, speed zone, day of the week, female and passengers is obviously influential on severity. As a recommendation on these factors, the responsible authorities should consider options to limit passengers in a vehicle during certain times of the month and day, to minimize the magnitude of risk. In addition, factors such as accident type, number of vehicles, road geometry, and alcohol time have a clear influence in predicting severity. This is not surprising but not actionable by itself. The remaining factors such as light conditions, police attend, pillion, accident status, etc., are of minor importance

VII. CONCLUSIONS

Traffic accident severity prediction is a crucial task to ensure better transportation safety and management. Recently machine learning models are emphasized in the literature as the non-parametric techniques employed in the transportation domain to provide recommendations of saving human lives. However, there is still a gap in the use of machine learning methods in crash injury severity prediction. For example, some methods like Multinomial Naïve Bayes have not been widely used to analyze crash injury severity. Moreover, the accuracy produced by some models is low, other models suffer from overfitting issues while others lack interpretability for humans.

On the road accident dataset from 2015 to 2020 provided by the State of Victoria in Australia, this study applied Random Forest (RF), Multinomial Naïve Bayes (MNB), K-Means Clustering (KC), and K-Nearest Neighbors (KNN) models to predict and classify traffic accident severity. Parameters were optimized to improve the overall prediction performance. Besides, a feature importance study is conducted in this research to analyze the significant factors contributing to accident severity and provide recommendations to concerned stakeholders to ensure better safety. The confusion matrix and its related evaluation metrics such as accuracy, precision, recall, and F1 score were used to assess the classification accuracies. The overall results of this study revealed that the RF method outperformed other approaches in predicting accident severity. After cross-validated, the RF model produced better testing accuracy, followed by the MNB, KC, and KNN respectively. The KNN suffered from overfitting issues.

Furthermore, in feature importance study, month, time of day, female drivers, male drivers, total persons, speed zone, day of the week, passengers, etc., were found as the major determinants of accident severity. According to the findings of models used in this study and feature analysis, recommendations such as safe route planning, preparing emergency vehicle allocation, reducing property damage, placing additional signage where necessary, and roadway design are provided to concerned stakeholders to eradicate the number of fatalities, property damages and injuries resulting from traffic accidents. This study offers few recommendations for future research. Firstly, hybrid models, deep learning, and stacking framework should be introduced to compare the overall prediction performance for accident severity prediction. Furthermore, different datasets from different locations of the world should be used to evaluate the effectiveness of the model.

REFERENCES

[1] World Health Organization, *Global status report on road safety 2018: summary*, World Health Organization, 2018.

[2] Y. Xie, Y. Zhang, and F. Liang, "Crash injury severity analysis using Bayesian ordered probit models," *Journal of Transportation Engineering*, vol. 135, no. 1, pp. 18-25, 2009.

[3] F. Ye, and D. Lord, "Comparing three commonly used crash severity models on sample size requirements: Multinomial logit, ordered probit and mixed logit models," *Analytic methods in accident research,* vol. 1, pp. 72-85, 2014.

[4] C. O'donnell, and D. Connor, "Predicting the severity of motor vehicle accident injuries using models of ordered multiple choice," *Accident Analysis Prevention* vol. 28, no. 6, pp. 739-753, 1996.

[5] M.-M. Chen, and M.-C. Chen, "Modeling road accident severity with comparisons of logistic regression, decision tree and random forest," *Information,* vol. 11, no. 5, pp. 270, 2020.

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

[7] Y. Lv, S. Tang, and H. Zhao, "Real-time highway traffic accident prediction based on the k-nearest neighbor method." pp. 547-550.

[8] B. N. Araghi, S. Hu, R. Krishnan, M. Bell, and W. Ochieng, "A comparative study of k-NN and hazard-based models for incident duration prediction." pp. 1608-1613.

[9] R. Mauro, M. De Luca, and G. Dell'Acqua, "Using a k-means clustering algorithm to examine patterns of vehicle crashes in before-after analysis," *Modern Applied Science* vol. 7, no. 10, pp. 11, 2013.

[10] T. K. Anderson, "Kernel density estimation and K-means clustering to profile road accident hotspots," *Accident Analysis Prevention* vol. 41, no. 3, pp. 359-364, 2009.

[11] S. H.-A. Hashmienejad, and S. M. H. Hasheminejad, "Traffic accident severity prediction using a novel multi-objective genetic algorithm,"*Internationaljournalof crashworthiness,* vol. 22, no. 4, pp. 425-440, 2017.

[12] J. Tang, F. Liu, Y. Zou, W. Zhang, and Y. Wang, "An improved fuzzy neural network for traffic speed prediction considering periodic characteristic," *IEEE Transactions on Intelligent Transportation Systems,* vol. 18, no. 9, pp. 2340-2350, 2017.

[13] T. Cover, and P. Hart, "Nearest neighbor pattern classification," *IEEE transactions on information theory,* vol. 13, no. 1, pp. 21-27, 1967.

[14] P. J. Hewson, "Multivariate statistics with R," *URl: http://local.disia.unifi.it/rampichini/analisi_multivariata/Hewson_ notes. pdf,* 2009.

[15] A. Liaw, and M. Wiener, "Classification and regression by randomForest," *R news,* vol. 2, no. 3, pp. 18-22, 2002.

[16] L. Breiman, "Random forests," *Machine learning* vol. 45, no. 1, pp. 5-32, 2001.

[17] A. Iranitalab, and A. Khattak, "Comparison of four statistical and machine learning methods for crash severity prediction," *Accident Analysis Prevention*, vol. 108, pp. 27-36, 2017.

[18] S. S. Yassin, "Road accident prediction and model interpretation using a hybrid K-means and random forest algorithm approach," *SN Applied Sciences,* vol. 2, no. 9, pp. 1-13, 2020.

[19] I. Mogotsi, "Christopher d. manning, prabhakar raghavan, and hinrich schütze: Introduction to information retrieval," Springer, 2010.

[20] D. Herrmann, R. Wendolsky, and H. Federrath, "Website fingerprinting: attacking popular privacy enhancing technologies with the multinomial naïve-bayes classifier." pp. 31-42.

[21] L. Wahab, and H. Jiang, "A comparative study on machine learning based algorithms for prediction of motorcycle crash severity," *PLoS one,* vol. 14, no. 4, pp. e0214966, 2019.