Comparison of Machine Learning Algorithms for Predicting Traffic Accident Severity

Rabia Emhamed Al Mamlook, Ph.D. Candidate
Dept. of Industrial Engineering
Western Michigan University
Kalamazoo, MI.
Email: rabiaemhamedm.almamlook@wmich.edu

Keneth Morgan Kwayu, Ph.D. Candidate
Dept. of Civil and Construction Engineering
Western Michigan University
Kalamazoo, MI.
Email: kenethmorgan.kwayu@wmich.edu

Maha Reda Alkasisbeh, Assistant Professor
Dept. of Civil Engineering
Hashemite University
Al-Zarqa, Jordan.
Email: malkasasbeh@hu.edu.jo

Abdulbaset Ali Frefer, Associate Professor
Dept. of Mechanical and Industrial Engineering
University of Tripoli
Tripoli, Libya
Email: A.Frefer@uot.edu.ly

***ABSTRACT*: Traffic accidents are among the most critical issues facing the world as they cause many deaths, injuries, and fatalities as well as economic losses every year. Accurate models to predict the traffic accident severity is a critical task for transportation systems. This investigation effort establishes models to select a set of influential factors and to build up a model for classifying the severity of injuries. These models are formulated by various machine learning techniques. Supervised machine learning algorithms, such as AdaBoost, Logistic Regression (LR), Naive Bayes (NB), and Random Forests (RF) are implemented on traffic accident data. SMOTE algorithm is used to handle data imbalance. The findings of this study indicate that the RF model can be a promising tool for predicting the injury severity of traffic accidents. RF algorithm has shown better performance with 75.5% accuracy than LR with 74.5%, NB with 73.1%, and AdaBoost with 74.5% accuracy.**

Keywords- Random Forest, Logistic* Regression, *Naïve Bayes, AdaBoost, Traffic Accident Severity.

I. INTRODUCTION

Traffic accidents are a daily source of death, injury, and property damage on roadways resulting in huge losses at economic and social levels. According to the World Health Organization (WHO), in 2017 around 1.5 million different road users die every year from traffic accidents and half of them die due to traffic crashes. It is also expected that in the absence of sustainable traffic, traffic crashes will become the leading cause of death by 2030 [1]

As the demand for vehicles rises, the number of vehicles on the road and traffic jams increase, particularly during rush hours. Therefore, road traffic accidents are among the leading causes of death and injury worldwide. According to the report by the Michigan Traffic Crash Decade-At-A-Glance, [2], there were over 314,921 traffic accidents in the US in 2017, which cost Americans over $230 billion each year. Over 1,028 people lost their lives while over 78,394 people were injured. Classification methods are among the most commonly used techniques in mining traffic accidents, where the goal is building classifiers that can predict the accidents. These classifiers are built using training sets of data in which accidents factors are known. Thus, the investigative and predictive methods, such as machine learning algorithms are vital to making smart decisions that will eradicate avoidable accidents on freeways. Can machine learning algorithms assist in saving lives? This inspires the researchers of this study to use machine learning algorithms to predict and analyze freeway crashes based on the roadway, human, and environmental factors. The primary objective of this study is to achieve the accuracy and identify the factors behind Traffic Accident Severity that could be helpful to reduce accident frequency and severity in near future, thus saving many lives and wealth, as well as many other things. Additionally, the study aimed to establish models to select a set of influential factors and to build up a model for classifying the severity of injuries that can be used by the Michigan Traffic Agencies (MTA). This model will help MTA and other responsible agencies in Michigan to be more proactive in combating high-risk areas on freeways.

II. RELATED WORK

The costs of fatalities and injuries due to traffic accidents have a great impact on society. In recent years, road traffic accidents, especially severe vehicle crashes have increased because of the rapid growth of road traffic. Indeed, in recent years much attention has been paid to determine factors that significantly affect the severity of traffic accidents and several approaches have been used to study this problem [3]. The factors that are related to traffic accidents, include; environmental (i.e. weather conditions and road signs), vehicle type & its safety, and the characteristics of

978-1-5386-7942-5/19/$31.00 ©2019 IEEE

traffic users. Moreover, some of these factors are more important in determining the accident severity than others. Therefore, it is apparent that the analysis of the determinant factors of accident severity will help reveal more patterns and knowledge that can be used in the prevention of accidents [4].

The predictive traffic accidents models are considered to be vital to making smart decisions that can lead to avoid accidents on freeways. With the advancements of information technology, machine learning becomes increasingly mature, and useful information without preconditions can be found in databases. Several studies, such as Krishnaveni and Hemalatha [5] Beshah and Hill [6] Chen et al. [7], have investigated machine learning algorithms in transportation-related applications of the causes of accidents. Krishnaveni and Hemalatha [5] conducted a prospective analysis of 34,575 traffic accident events in Hong Kong. In their study, Naive Bayes, AdaBoostM1, J48, PART, and Random Forest classifiers were employed to predict and detect the severity of injury and causes of accidents using WEKA tool. According to the comparison results, the Random Forest classifier outperformed all other algorithms. There are no percentage results. Beshah and Hill [6], employed Naive Bayes, Decision Tree (J48), and K-Nearest Neighbors classifiers to build prediction models to assess the injury severity that were used to analyze and predict the role of road-related factors for traffic accident severity. Moreover, they utilized the PART algorithm to present the knowledge in the form of rules, with the accuracy of 79.94% using the WEKA tool [6]. Chen et al. [7] used the SVM models to investigate driver injury severity patterns in rollover crashes using two-year crash data collected in New Mexico. The results showed that the support vector machine (SVM) models produce reasonable predictions and the polynomial kernel outperforms the Gaussian RBF kernel.

Dong et al. [15] used two modules, an unsupervised feature and a supervised fine-tuning module to perform traffic crash prediction. The results showed that the feature learning section classifies interactive information between the explanatory variables and the feature representations, which decreases the dimensionality of the input and preserves the original information.

Therefore, there is a need to perform a comprehensive analysis that aims to understand the relationship between the influence factors and traffic crash outcomes. In this study, the researchers utilized the full traffic dataset and used big data technology and tools to gain better insight and achieve more accurate results. So, this study focuses on using machine learning techniques to predict accident severity in Michigan, USA.

Despite all the strategies applied to reduce traffic accidents, fatalities and Injuries, the severity of the accidents on Michigan freeways is still a real problem that requires paying more attention, analyzing the nature and extent of the problem, and developing & implementing specific and suitable measures and responses.

III. METHODOLOGY

The purpose of the methodology that was used in this research study is to build the prediction classification rules of the best performing model (AdaBoost, LR, NB, and RF). The overall procedure for studying the classification is shown in Fig. 1.

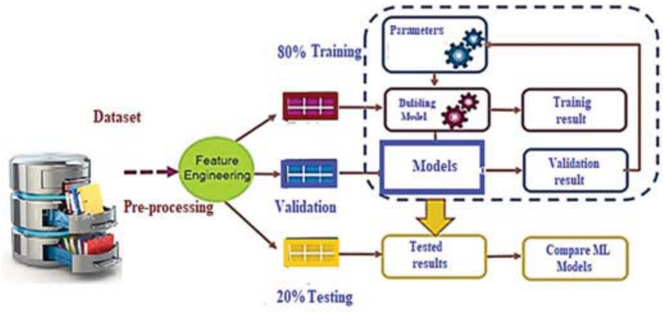


Figure 1. Model building processing

This section discusses the methods used in this research study, including data collection, attribute selection, building the classification models, and extracting the required knowledge.

➢ *Data Source and Description.*

The data used in this study was provided by the Office of Highway Safety Planning (OHSP). This office shares raw crashes on annual basis with Western Michigan University (WMU), Transportation Research Center for Livable Communities (TRCLC). In the present study, the dataset contains information about road crashes occurred on Michigan during the period ranging from 2010 to 2016. Additionally, the dataset for the study contains traffic accident records for the same years, having the total number of 271,563 traffic crashes.

Table 1. Freeway crashes in Michigan (2010-2016)

Years	Fatal & serious injuries (KA)	Other injuries (BCO)	Total Crashes (KABCO)	% of KA in Total Crashes
2010	403	35886	36289	1.11%
2011	378	41242	41620	0.91%
2012	408	39870	40278	1.01%
2014	469	50108	50577	0.93%
2015	431	49552	49983	0.86%
2016	539	52277	52816	1.02%
2010-2016	2628	268935	271563	0.97%

Accident severity, including the number of fatalities, number of injuries, and property damage, as well as accident duration were forecasted with the aforementioned models. Furthermore, the number of fatalities in 2016 was the highest in a decade as can be seen in Table 1.

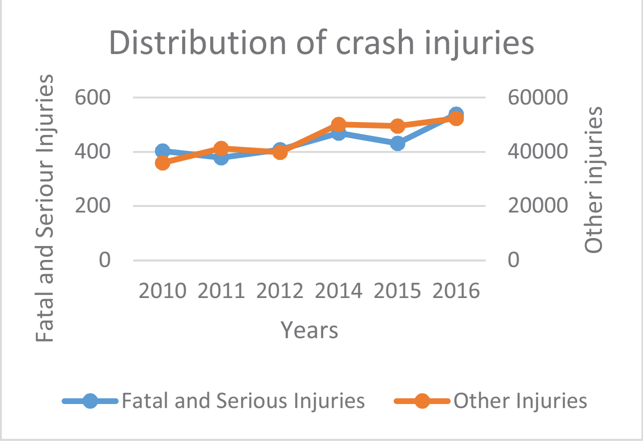


Figure 2. Distribution of fatalities

Over 539 people lost their lives while driving on freeways. Fig. 2 shows distribution of fatalities and serious injuries in Michigan. Crashes on freeways. They have increased considerably over the years.

➢ *Data Pre-Processing.*

Data preparation was performed before each model construction. The process involves various steps that include, cleaning, normalization, feature selection, and transformation. In this study, the researchers applied Class-Imbalanced data solution and feature selection task on the dataset. The dataset used contains integer values for the entire attributes. Similar transformations have been done to have the categorical variables. Also, the researchers chose factors which are related to accidents that include, roadway, environmental, vehicle, and human factors. As for the feature selection, ten attributes were selected as possible predictors of crash instances. The response variable was binary with one (1) indicating a fatal or severe crash (KA) while zero (0) indicating minor,

Possible or property damage crashes (BCO) [8-11]. The database was rearranged, and 3 continuous variables and 6 categorical variables were selected: Continuous variables were; (1) car manufacturing year, (2) age, (3) traffic volume; while categorical variables were; (4) gender, (5) alcohol or drug-related crash (6) lighting condition-dark, (7) weather condition-clear, (8) Driver hazard action, and (9) seatbelt usage.

➢ *Machine Learning Classification Model.*

In order to determine the effectiveness of the machine learning techniques, the algorithms are trained on a portion of the data and then evaluated on a testing set. The purpose of splitting the data in this manner is to determine how the developed models work on new data that has not been seen before. For the purposes of this research, the data was split into 70% training and 30% testing. In order to compare the results of the prediction models in a more robust way, 10-fold cross validation was implemented. At the first step, the researchers used SMOTE resampling strategy to handle data imbalance as there was a small percentage of fatal and serious injuries compared to other injuries. The SMOTE algorithm generates synthetic positive instances to increase the proportion of minority class [12]. The next step is to build a model for predicting the Fatal and serious Injuries by comparing the accuracies between them. For the developing the model for the prediction of injury severity, the following four popular machine learning classification techniques were studied:

- Logistic Regression (LR):

Logistic Regression is a classification model. The idea of the algorithm is to map the results of linear functions to sigmoid functions. The linear regression model is a simple mathematical model and easy to implement.

- Random Forest Model (RF):

The Random Forest model is an ensemble learning method which constructs a series of decision trees at training time and outputs the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. The minimum number of samples required to split a node was set to two, and the minimum samples per leaf are set to one.

- Naïve Bayesian Classifier (NBC):

Naive Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong (naive) independence assumptions between the features. Since the feature set contains continuous variables, the Gaussian NB was chosen.

- AdaBoost Classification Tree:

AdaBoost is a classification algorithm which calls a given weak learner algorithm repeatedly in a series of rounds. It is a binary boosting algorithm that maybe the

Most significant one representing the simple milestone of many other classification algorithms; including, Boost by Majority, LP Boost, and Logit Boost [13].

➢ *Performance Measurement*

Five measures were used to compare the performances of the four machine learning techniques: precision, recall/sensitivity, f-measure, the Receiver Operator Characteristic (ROC), and the Area under the Receiver Operating Characteristic Curve (AUC). For the first three performance metrics, a confusion matrix must be used to perform the calculations and compare the performances. In this step, the evaluation of the performance of the classification AdaBoost, LR, NB, and RF algorithms is performed and compared.

Table 2: Results of Performance Measurements with different ML techniques

Model	Accuracy	Precision	Recall	F1Score
LR	0.745	0.858	0.586	0.696
NB	0.731	0.949	0.489	0.645
Boosted	0.745	0.873	0.573	0.692
RF	0.755	0.881	0.591	0.707

To evaluate the results, the researchers used the measures accuracy, precision, recall and F-score that can be derived from the confusion matrix as can be seen in Table 2.

IV. RESULTS AND DISCUSSION

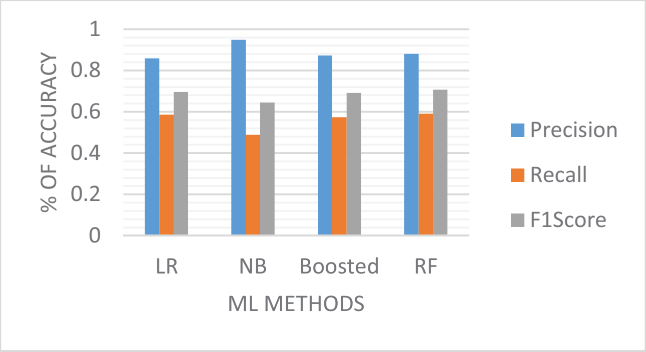
This section presents and discusses the experiments and the results for the four different classifiers Decision tree; namely, Random Forest, Logistic Regression, Naive Bayes, and AdaBoost algorithms. Various comparisons and analysis were discussed to see which of the five approaches provide better performance on prediction traffic accident severity.

➢ *Classification Accuracy*

Precision is a measure of exactness or quality, whereas recall is a measure of completeness or quantity. High recall means that an algorithm returned most of the relevant results. High precision means that an algorithm returned more relevant results than irrelevant. After an accurate comparison between the used algorithms, the results indicated that RF achieved a higher efficiency of 0.82. As shown in Fig 3, the Random Forest classifier achieves the highest accuracy among all the classifiers. Therefore, there is no doubt that the Random Forest classifier has an advantage over the other three representative classifiers in this sample.

The result demonstrates that the best machine learning technique is Random forest. Compared with other classifiers, Random Forest has a good ability to resist noise due to the application of randomly selecting variables and data to generate plenty of classification trees. It can process not only discrete data, but also continuous data.

Figure 3. Comparison of Accuracy, Precision, Recall and F1score with different ML



➢ *The Area under ROC Curve (AUC).*

The value is the area under the ROC curve and is a ratio between 0 and 1, where a value of 1 indicates a perfect classifier, while a value close to 0.5 indicates a bad model, since that is equivalent to a random classification [14].

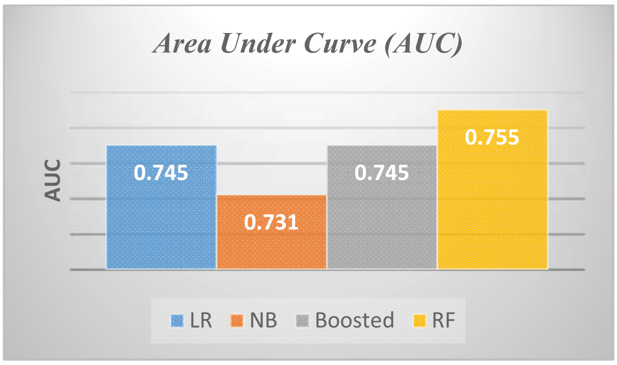


Figure 4. Compare of Area under Curve (AUC)

Generally, AUC ranges from 0.5 to 1 and the greater the AUC, the better the classification performance. When the AUC is greater than 0.7, it indicates that the model has an excellent prediction ability in pattern recognition. RF algorithm has shown better performance with 75.5% accuracy than LR with 74.5%, NB with 73.1%, and AdaBoost with 74.5% accuracy in Fig.4. These encouraging AUCs give a statistical proof of the excellent classification capacity of the Random Forest in this study.

V. SUMMARY AND CONCLUSIONS

This study investigated the efficiency of the four machine learning algorithms to build classifiers that are precise and reliable. This includes the Random Forest (RF), Logistic Regression (LR), Naïve Bayesian Classifier (NB), and AdaBoost algorithms. Based on the confusion matrix F1-Score, the test results show that the Random Forest seemed to perform better than the other models. This research study shows that the algorithms can predict accidents with a 75.50% accuracy. This study can help provide useful information for highway engineers and transportation designers to design safer roads. Further studies should be done to collect related information and investigate the impacts of these factors. It is recommended that Random forest (the best model for predicting freeway crashes) to be applied in monitoring

Fatal and serious injuries. The recommended predictive model can be used to rapidly and efficiently identify the Key factor causing traffic crashes. One limitation of the current study is that some of the factors (i.e. characteristics of the driver, passenger, and pedestrian, along with traffic conditions) may have possible effects on accident severity and duration, which are not considered because of the lack of suitable data.

ACKOWLEDGMENT

This work was supported by Western Michigan University College of Civil Engineering.

REFERENCES

[1] WHO | Road traffic injuries, 2017. WHO.

[2] Michigan State Police, Michigan Traffic Crash Decade-At-A-Glance, 2018.

[3] M. Chong, A. Abraham, M. Paprzycki, “Traffic accident data mining using machine learning paradigms”, Fourth International Conference on Intelligent Systems Design and Applications (ISDA'04), Hungary, 2004, pp. 415- 420.

[4] T. Dejene, A. Ajith, V. Snášel, and P. Krömer, “Knowledge discovery from road traffic accident data in Ethiopia: data quality, ensembling and trend analysis for improving road safety”, Neural Network World, vol. 22, no. 3, 2012, pp. 215–244.

[5] S. Krishnaveni and M. Hemalatha, “A perspective analysis of traffic accident using data mining `techniques,” International Journal of Computer Applications, vol. 23, no. 7, 2011, pp.40-48.

[6] T. Beshah and S. Hill, “Mining road traffic accident data to improve safety: role of road-related factors on accident severity in Ethiopia,” AAAI Spring Symposium, 2010.

[7] G. Chen, Z. Zhang, R. Qian, R. A. Tarefder, and Z. Tian, “Investigating Driver Injury Severity Patterns in Rollover Crashes Using Support Vector Machine Models,” Accident Analysis and Prevention, vol. 90, 2016, pp. 128–139.

[8] S. Nazneen, M. Rezapour, and K. Ksaibati, “Determining causal factors of severe crashes on the fort peck Indian reservation,” Journal of Advanced Transportation’ Montana, 2018, pp. 1-8.

[9] S. Dissanayake and U. Roy, “Crash Severity Analysis of Single Vehicle Run-off-Road Crashes,” Journal of Transportation Technologies, vol. 4, 2014, pp. 1-10.

[10] D. Shinstine, S. Wulff, and K. Ksaibati, “Factors associated with crash severity on rural roadways in Wyoming,” Journal of Traffic and Transportation Engineering, vol. 3, no. 4, August 2016, pp. 308-323.

[11] M. I. Ratnayake, “Effectiveness of seat belts in reducing Injuries: a different approach based on KABCO injury severity scale,” Midwest Transportation Consortium, Iowa, 2006.

[12] N. Chawla, K. Bowyer, L. Hall, and W. KegelMeyer, “SMOTE: synthetic minority oversampling technique,” Journal of Artificial Intelligence Research, vol. 16, 2002, pp. 321-357.

[13] L. Lei and X.-D. Wang, “Improved ad boost ensemble approach based on loss function,” Journal of Computer Applications, vol. 32, no. 10, 2013, pp. 2916–2919.

[14] G. James, D. Witten, T. Hastie, and R. Tibshirani, 2013. “An introduction to statistical learning with applications in R,” Springer Texts in Statistics Series, New York, 2013.

[15] Dong. C, Shao. C, Li. J, and Xiong. Z, “An improved deep learning model for traffic crash prediction,” Journal of Advanced Transportation, 2018, pp.1-13.