INTERNATIONAL CONFERENCE ON ADVANCES IN TECHNOLOGY AND COMPUTING (ICATC-2021)
Faculty of Computing and Technology (FCT), University of Kelaniya, Sri Lanka
18th December 2021

Index No: TL-27-41

Road Accident Severity Prediction in Mauritius using Supervised Machine Learning Algorithms

Jameel Ahmad Sowdagur (1st Author)
University of Technology Mauritius (UTM)
Pointe-aux-Sables
Republic of Mauritius.
jsowdagur@umail.utm.ac.mu

B. Tawheeda B. Rozbully-Sowdagur (2nd Author)
University of Technology Mauritius (UTM)
Pointe-aux-Sables
Republic of Mauritius.
brozbullysowdagur@umail.utm.ac.mu

Geerish Suddul (3rd Author)
University of Technology Mauritius (UTM)
Pointe-aux-Sables
Republic of Mauritius.
g.suddul@umail.utm.ac.mu

***Abstract* — Road accidents with high severities are a major concern worldwide, imposing serious problems to the socio-economic development. Several techniques exist to analyse road traffic accidents to improve road safety performance. Machine learning and data mining which are novel approaches are proposed in this study to predict accident severity. Support Vector Machine (SVM), Gradient Boosting (GB), Logistic Regression (LR), Random Forest (RF) and Naïve Bayes (NB) were applied to perform effective data analysis for informed decisions using Python programming language. The gradient boosting outperformed all the other models in predicting the severity outcomes, yielding an overall accuracy of 83.2% and an AUC of 83.9%.**

Keywords — Data mining, machine learning, accident severity

I. INTRODUCTION

Road accidents comprise significant public health and development threats. Each year, 1.35 million deaths occur due to road traffic accidents worldwide and in many countries, it accounts for 3% of the Gross Domestic Product (GDP) [1]. Among the victims, more than half are young adults aged between 15 and 44. An increase in fatal road accidents is anticipated by the World Health Organization (WHO) if road safety is not addressed [1]. A National Observatory for road safety in Mauritius for the period 2016-2025 has been set up by the concerned Ministry with one of its objectives of curbing the number of road fatalities to 50% by the year 2025 in line with the target set by the WHO [2]. However, figures published by Statistics Mauritius indicates 21.4% increase in road accidents from the year 2013 to 2020 resulting in a rise of the fatality index from 3.8 to 4.5 during the same period. This increasing trend has made it a necessity to handle the problem in a more scientific approach.

This paper presents road accident severity prediction using a machine learning approach and evaluation of the classification models to identify the most performant one among support vector machine, gradient boosting, logistic regression, random forest and naïve Bayes.

The remaining part of the paper is structured as follows: Section II will cover the literature review on the features related to injury accidents and machine learning.

Approaches to predict accident severity. Section III will address the data collection procedures. Section IV will deal with the data preprocessing. The different machine learning algorithms and their performance metrics will be highlighted in the methodology in section V. The results obtained will be discussed in section VI and ultimately, the paper will be concluded in section VII together with future work

II. LITERATURE REVIEW

Data mining is a diagnostic operation that integrates artificial intelligence, statistics, and machine learning which focuses on pattern detection, prediction, and forecasting [3]. Further, [4] added that machine learning is effective in road traffic injury predictions and can help to mitigate road accidents. The literature review highlights some contributing features identified and machine learning techniques used in this particular field.

A. Features contributing to injury accidents

According to [5], the multiple factors affecting road crashes are human causes, weather circumstances, road designs, traffic characteristics and vehicle conditions. In the study of [6], it was concluded that the use of data mining can determine and forecast leading factors amidst human, vehicle and environment. The study of [7] revealed that junction type, road type, location, signposting, the hour of the day, license type, driver's age, day of the week and vehicle type are all significantly related to injury severity.

B. Machine learning techniques to model road accidents

The utilization of machine learning classifiers provides alternatives to traditional data mining techniques for generating higher results and accuracy, as highlighted in [8]. The study of [9] compared machine learning algorithms to predict the severity of motorcycle crashes. J48 decision tree classifier, random forest and instance-based learning with parameter "k" (IBk) were used in modelling the severity outcome. These models were validated employing the

technique of 10-fold cross-validation. Comparisons were made with each other and with a statistical model namely the multinomial logit model
(MNLM). Their experimental results revealed that the prediction performances of the machine learning algorithms were better than MNLM. The random forest showed its superiority with the experimental data for its optimization and extrapolation capability. In [10], the authors employed J48, rule induction (PART), naïve Bayes and multilayer perceptron (MLP) which showed similar conclusions from the algorithms apart from the naïve Bayes classifier which exhibited less accuracy.

A comparative study in [11] was made with four models for crash severity prediction using multinomial logit (MNL), nearest neighbor classification (NNC), SVM and RF. The results demonstrated that NNC performed satisfactorily overall and in high severity crashes. NNC's best achievement was since this method does not require a distributional assumption of the data. The limitation in this study was the forced removal of part of the probably paramount predictors, like speed and impact type due to the presence of unavailable information in them. The study of [12] employed gradient boosted, decision tree (DT) and RF to identify hazard factors and injury severity prediction of drivers. The gradient boosted had the highest accuracy of 73.3%. Similarly, in [13], six algorithms namely LR, DT, SVM, neural network (NN), RF and extreme gradient boosting (XGBoost) were compared in predicting injury severity levels. The XGBoost outperformed the other algorithms with an accuracy of 74.4%, followed by the RF (73.8%).

In [14], to predict injury severity crashes, a comparison of the SVM with the ordered probit (OP) model was made. It was deduced that the SVM model had improved predicting power (48.8%) for severity predictions over the OP model (44.0%). As limitations, the authors suggested that apart from the basic radial basis function (RBF), different kernel functions can be experimented with to enhance the results of the models. Another study comparing SVM models with polynomial and Gaussian RBF kernels was carried out by [15] to enquire about driver's injury severity. It was found that the SVM model gave feasible prediction achievement and the polynomial kernel surpassed the Gaussian RBF kernel. The authors stated that SVM algorithms are a common non-parametric classification technique that has been extensively used in the transportation field, but are yet somewhat new in the road accident analysis field. The polynomial SVM classifier worked best on the majority of instances. It was also confirmed that transforming multi-categorical variables into numerical ones is an effective way to enhance the ability of the classification model. The limitation was however the small sample sizes for each type of injury analyzed.

III. DATA COLLECTION

Road accidents data including information on accidents circumstances, vehicles and casualties collected by the local authority in Mauritius was used.

The dataset initially comprised 12,523 instances and 49 variables, including the target variable "Accident_Severity". The quality of the data was assessed by investigating if the type of the variables was correct and the presence of any missing data. The data preparation phase consisted of removing unwanted and conflicting variables, handling missing data, treatment of class variables, renaming of factors, binning, dimension reduction/feature selection, association test and handling of imbalance class.

A. Excluded variables

Unwanted variables which don't add intrinsic value to the dataset and were not influential to the target variable were removed. Some variables were: "Accident key", "Vehicle reference number" among others. The response from the variables "Casualty injury" or "Driver injury" or both depending on whichever is higher, determine the overall accident severity type. These two conflicting variables were removed after the strength of their associations were tested using the chi-square test.

B. Handling missing data

Associated entries containing missing values were not dropped, as this would have resulted in considerable loss of information. The missing values were replaced by the next highest code to be further labelled by "Unknown" for the categorical variables and only drivers above 15 years were considered.

C. Treatment of class variables and renaming of factors

The variables were converted to their appropriate classes and renamed accordingly. The numerical variables were converted to an integer and the categorical variables to factors.

D. Binning

The "Time" variable was converted into a factor category containing three levels: 'Night', 'Morning' and 'Afternoon'. The variables "Casualty age" and "Driver age" were binned in the range of five years.

E. Feature selection

Feature selection is primarily based on the exclusion of non-informative or irrelevant predictors from the model and is also a dimensionality reduction technique. This technique helps in improving the performance of some predictive modelling and computational time in the case of large datasets [16]. After one hot encoding of the categorical variables to be used for the SVM, the most important features were selected according to their *k* highest score.

F. Handling of imbalance class

As highlighted in [17], machine learning algorithms are influenced by a high imbalance class between dominant and classes of the minority, which may result in favour of the majority (negative) class during prediction. The authors further stated that ratios of 75:25 and 65:35 are classified as slight imbalance, whereas the ratio of 90:10 is regarded as moderately imbalance. Fig. 1 shows that the dataset under study had two minority classes (fatal and serious) suggesting an imbalanced dataset.

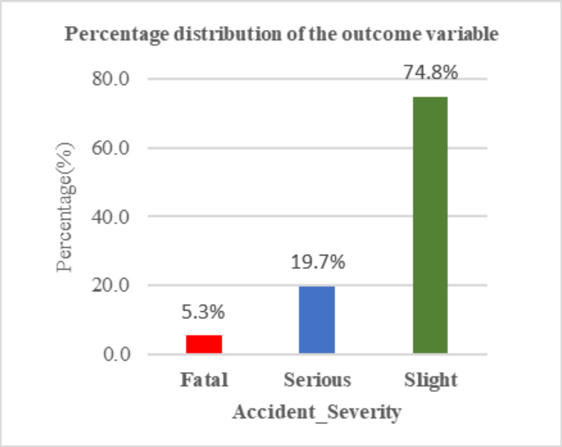

Fig. 1. Distribution of the target variable.

As such, these classes were grouped to form the positive class KSI (Killed and serious injury) which later for classification purposes in model prediction will be allocated a value of “1”. The negative class or majority slight class will be assigned a value of “0”. The result obtained after merging the two minority classes is shown in Fig. 2.

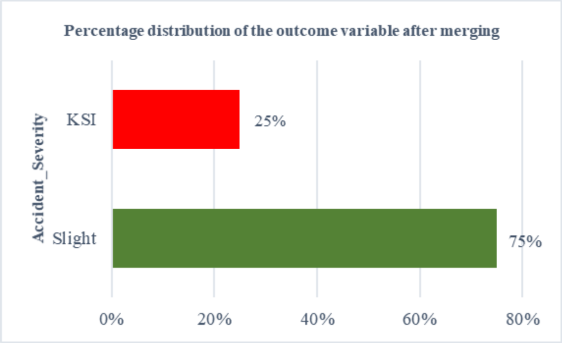

Fig. 2. Distribution of the merged classes.

A slightly imbalanced dataset was obtained with a ratio of 25:75, from which the classification models were built.

G. *Final dataset*

We now have a clean dataset with 12,111 observations and 30 variables, including the outcome.

IV. METHODOLOGY

In this work, the experimentation was based on five machine learning algorithms namely support vector machine, gradient boosting, logistic regression, random forest and naïve Bayes. Different split ratios (75:25, 80:20 and 90:10) were employed for the training and test set respectively. SVM requires numerical inputs. Hence, the categorical variables were one-hot encoded. The numerical variables were scaled using the min-max normalization. To assess the performance of each model, the confusion matrix was considered.

A. *Logistic regression*

Logistic regression utilizes the power of regression to classify using maximum likelihood to fit a sigmoid curve on the target variable distribution. In this study, the logit is the natural logarithm of the odds or the likelihood ratio that the response is ‘1’ (KSI) opposing to ‘0’ (Slight). The probability ‘p’ of a KSI accident is given by:

$$Y = logit = ln\, ln \left(\frac{p}{1-p}\right) = \beta X. \qquad (1)$$

Where $'Y'$ is the dependent variable (Accident_Severity; Y=1, if every severity is KSI and Y=0 if the severity is Slight), ‘β’ is a vector of parameters to be calculated and ‘X’ is a vector of the predictors.

B. *Naïve Bayes*

The naive Bayes classifier assumes that the existence (or lack) of one class feature (i.e., attribute) has no bearing on the presence (or absence) of any other feature. The model uses trained data to compute the probability of each class and the conditional probability of each class given each ‘x’ value. It is effective for a large range of complicated problems [18]. The study of [19] concluded naïve Bayes to be more accurate than decision tree algorithm in predicting the severity of an accident.

C. *Decision tree*

The decision tree can be employed to solve problems of regression and classification. ID3, classification and regression trees algorithm, J48, alternating decision tree (ADTree) form part of the decision tree algorithms [20]. It can be used to visually and explicitly represent decisions and decision-making based on the lowest Gini index and highest information gain.

D. *Random Forest*

It is an ensemble of many decision trees and can be used in both classification and regression. RF is viewed as an improved method of the decision tree [21]. Being non-parametric, RF does not require any formal distributional assumption. It can handle many predictor variables and missing data [22].

E. *Support vector machine*

This machine learning model was initially used in training data where it could be separated without errors and later was extended to non-separable training data [23]. It is another supervised algorithm used in the analysis of both classification and regression. In this study, we experimented with the linear SVM.

F. *Gradient boosting*

The Gradient Boosting method works by constructing a series of successive decision trees, each of which seeks to improve on the one before it by optimizing errors. It is a stage-wise additive model, in which each weak learner is added one at a time, previous weak trees are left unmodified and the model is trained by iteratively improving each tree [12].

G. *Confusion matrix*

The classifying ability of the learned models was assessed as per the confusion matrix in Fig. 3.

		Actual values	
		KSI	Slight
Predicted values	KSI	TP	FP
	Slight	FN	TN

Fig. 3. (2x2) confusion matrix.

True Positive (TP) refers to the predicted positive outcome by the model that corresponds to the positive real value. True Negative (TN) is the predicted negative result that fits the actual value that was negative. False Positive (FP) is also termed as Type 1 error. It is the predicted positive value by the model that was incorrectly predicted as the real value was negative. False Negative (FN) also referred to as Type 2 error is the forecasted negative value by the model that was wrongly predicted as the real value was positive.

In the context of this study, to gauge the performance of the algorithms, the accuracy and area under the receiver operating characteristic (ROC) curve were considered. Accuracy is a measure of the actual positives plus the actual negatives that are correctly classified in the test data. The higher is the area under the curve, the better is the classifier.

V. RESULTS

Table 1 provides a holistic view of the metrics used to gauge the performance of the different classification models with their numerical results.

Table 1. Accuracy of classifiers by different split ratios.

Classifier algorithm	Ratio of training set: test set		
	75:25	*80:20*	*90:10*
SVM (linear)	76.8%	77.0%	77.8%
Gradient boosting	82.9%	83.0%	83.2%
Logistic regression	76.8%	77.0%	77.7%
Random forest	78.7%	78.5%	77.4%
Naïve Bayes	67.7%	67.1%	66.7%

It is observed from Table 1 that the gradient boosting has the best performance irrespective of the ratio used when compared with the other classifiers. However, its performance was the best for the split ratio of 90:10 yielding an accuracy of 83.2%, which concur with the results of [12] and [13]. The naïve Bayes was the only algorithm with an accuracy of less than 75% and was therefore not considered for other metrics calculation. This result is in contrast with the findings of [19]. The ability of the other remaining classifiers was appraised in terms of the AUC as shown in Fig. 5 which further confirms the superiority of the gradient boosting with an AUC value of 0.839. Although the SVM and the LR had better accuracy than the RF, the latter had a better AUC value of 0.687 as compared to 0.652 for the other two classifiers.

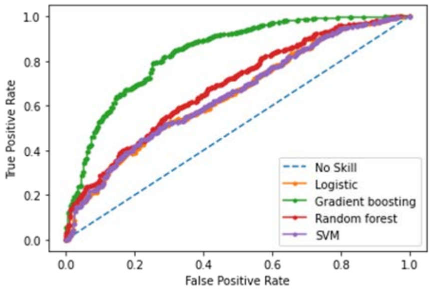

Fig. 5. Area under curve of classifiers.

VI. CONCLUSION AND FUTURE WORK

The primary objective of this study was to apply several accident severity prediction models using machine learning. The dataset was preprocessed to deal with all possible issues that affect the performance of machine learning algorithms. The different algorithms were made to learn from the training set and their performance was assessed on the test set. The two minority classes were merged to deal with the imbalanced problem. The classifications were made according to a 2x2 table through a confusion matrix and a model's performance was established based on accuracy and the area under curve. The gradient boosting with an accuracy of 83.2% and an AUC of 83.9% showed better performance in terms of classifying the minority and majority classes. As an improvement to this work, it is proposed that the k-fold cross validation be carried out as a means to detect any over/underfitting. Other models like the artificial neural network (ANN), SVM with different kernels and other deep learning models may be implemented and compared.

REFERENCES

[1] "Road traffic injuries," Who.int. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries, Accessed on: Jul 21, 2021.

[2] *Govmu.org*. [Online]. Available: http://www.govmu.org/French/News/Pages/Création-d'un-Observatoire-national-pour-la-sécurité-routière-à-Maurice.aspx, Accessed on: Jul 22, 2021.

[3] F. M. Nafie Ali and A. A. Mohamed Hamed, "Usage Apriori and clustering algorithms in WEKA tools to mining dataset of traffic accidents," *J. Inf. Telecommun.*, vol. 2, no. 3, pp. 231–245, 2018.

[4] D. Donchenko, N. Sadovnikova, and D. Parygin, "Prediction of road accidents' severity on Russian roads using machine learning techniques," in *Lecture Notes in Mechanical Engineering*, Cham: Springer International Publishing, 2020, pp. 1493–1501.

[5] M. Vilaça, E. Macedo, and M. C. Coelho, "A rare event modelling approach to assess injury severity risk of vulnerable road users," *Safety (Basel)*, vol. 5, no. 2, p. 29, 2019.

[6] S. Ramya, S. K. Reshma, V. D. Manogna, Y. S. Saroja and G. S. Gandhi, "Accident Severity Prediction Using Data Mining Methods," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 5, no. 2, 2019.

[7] L. Eboli, C. Forciniti, and G. Mazzulla, "Factors influencing accident severity: an analysis by road accident type," *Transp. res. procedia*, vol. 47, pp. 449–456, 2020.

[8] A. Venkat, M. Gokulnath, V. K. P. Guru, S. T. Irish and D. Ranjani, "Machine learning based analysis for road accident prediction," *International Journal of Emerging Technology and Innovative Engineering*, vol. 6, no. 2, pp. 31-37, 2020.

[9] 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, p. e0214966, 2019.

[10] M. Taamneh, S. Alkheder, and S. Taamneh, "Data-mining techniques for traffic accident modeling and prediction in the United Arab Emirates," *J. Transport. Safety Security*, vol. 9, no. 2, pp. 146–166, 2017.

[11] A. Iranitalab and A. Khattak, "Comparison of four statistical and machine learning methods for crash severity prediction," *Accid. Anal. Prev.*, vol. 108, pp. 27–36, 2017.

[12] S. Elyassami, Y. Hamid, and T. Habuza, "Road crashes analysis and prediction using gradient boosted and random forest trees," in *2020 6th IEEE Congress on Information Science and Technology (CiSt)*, 2020, pp. 520–525.

[13] M. K. Nour, A. Naseer, B. Alkazemi, and M. Abid, "Road traffic accidents injury data analytics," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 12, 2020.

[14] Z. Li, P. Liu, W. Wang, and C. Xu, "Using support vector machine models for crash injury severity analysis," *Accid. Anal. Prev.*, vol. 45, pp. 478–486, 2012.

[15] C. Chen, G. Zhang, Z. Qian, R. A. Tarefder, and Z. Tian, "Investigating driver injury severity patterns in rollover crashes using support vector machine models," *Accid. Anal. Prev.*, vol. 90, pp. 128–139, 2016.

[16] D. Bhalla, "Feature selection: Select important variables with Boruta package," *Listendata.com*. [Online]. Available: https://www.listendata.com/2017/05/feature-selection-boruta-package.html, Accessed on: Aug 13, 2021.

[17] T. Hasanin, T. M. Khoshgoftaar, J. L. Leevy, and N. Seliya, "Examining characteristics of predictive models with imbalanced big data," *J. Big Data*, vol. 6, no. 1, 2019.

[18] G. Shobha and S. Rangaswamy, "Machine Learning," in *Handbook of Statistics*, vol. 38, V. N. Gudivada and C. R. Rao, Eds. Elsevier, 2018, pp. 197–228.

[19] V. M. Ramachandiran, P. N. Kailash Babu, and R. Manikandan, "Prediction of Road Accidents Severity using various algorithms," Acadpubl.eu. [Online]. Available: https://acadpubl.eu/hub/2018-119-12/articles/6/1546.pdf, Accessed on: Aug 29, 2021.

[20] A. K. Sharma and S. Sahni, "A Comparative Study of Classification Algorithms for Spam Email Data Analysis," *International Journal on Computer Science and Engineering*, vol. 3, no. 5, pp. 1890-1895, May 2011.

[21] J. Wu, Z. Gao, and C. Hu, "An empirical study on several classification algorithms and their improvements," in *Advances in Computation and Intelligence*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 276–286.

[22] S. Richmond, "Algorithms exposed: Random forest," Org.au. https://bccvl.org.au/algorithms-exposed-random-forest/, Accessed on: Jul 30, 2021.

[23] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn*., vol. 20, no. 3, pp. 273–297, 1995.