Original Article

Improved Accuracy of Calculation of Vehicle Crash Severity in Highways using Random Forest over Naive Bayes Algorithm

Vignesh.S[1], Sashi rekha K[2*]

[1]Research Scholar, Department of Computer Science and Engineering , Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamilnadu, India. Pincode: 602105.
[2*]Project Guide, Department of Computer Science and Engineering, Saveetha School of Engineering,Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamilnadu, India. Pincode: 602105

Abstract

Aim: To improve the accuracy rate of vehicle crash severity in highways using Random forest over Naive Bayes. **Materials and Methods:** Random forest and Naive Bayes with sample size of (N=10) is executed with varying training and testing splits for calculating the accuracy for accident crash severity with g power as 75%, threshold 0.000 and confidence interval 95%.. The performance of the classifiers are evaluated based on their accuracy rate using accident severity dataset. **Results:** The accuracy for calculating accident crash severity in Random Forest(91%) and Naive Bayes (71%) is obtained($P<0.005$).**Conclusion:** Prediction of accident crash severity using Random Forest (RF) algorithm appears to be significantly better than Naive Bayes(NVB) with improved accuracy.

Keywords: Crash Severity, Novel Random Forest Algorithm, Naive Bayes Algorithm, Machine Learning, Artificial Intelligence.

DOI:10.47750/pnr.2022.13.S04.177

INTRODUCTION

Accidents are naturally occurring incidents. The increase of vehicles in these years has also increased the number of accidents and deaths.Studying the crashes and the factors which are provided by it will help the traffic engineers and practitioners to predict the severity of accidents and to prevent the loss of life and reduce the amount of injury this method is implemented (Ahmadi et al. 2020). The importance of crash severity prediction in recent years has helped hospitals provide proper medical care as fast as possible when an accident occurs so that proper treatment and the doctors can be ready when the person arrives from the crash place and also to improve road safety to avoid accidents in the future (Zhang et al. 2018).This research helps to predict the crash severity for transportation safety planners so that they can plan the roads and routes much safer,hospitals and agencies can provide emergency care with the help artificial intelligence to provide better and accurate care (Iranitalab and Khattak 2017). It is useful for reducing the consequences of traffic accidents because in this the prediction is made with the datas that can be easily found in the crash/accident site so that the nearest trauma center can dispatch the paramedics to the site quickly (Assi et al. 2020).

A lot of research has been performed on crash severity prediction using machine learning algorithm. 1761 research articles were published in Google Scholar and 37 articles were found in IEEE Xplore. In this work RF, Adaboost, GBDT, Logistic Regression is used to to help the traffic management department to predict the accuracy of accidents with regarding to road infrastructure (Tang et al. 2019).This proposed system can be extend the application for studies on the condition of environments in two way rural highways because of the considering the impacts of crash severity using multiple Decision Trees (Abellán, López, and de Oña 2013).In this various algorithms like ANN with 90%accuracy,SVM with 89% and RF with 92%. During travel we can check constantly for crashes occurring via the probe vehicle based on location and speed it can suggest to take an alternative route with less traffic and good weather to avoid furthermore accidents with the assistance of artificial intelligence (Dogru and Subasi 2018). SVM model is used in this paper to evaluate the sensitive impact of explanatory variables to measure the severity of the crash (Li et al. 2012).The Novel Random Forest Algorithm proved most

efficient based on the accuracy and thus it is chosen as the most preferred algorithm in predicting the severity of accident (Geyik and Kara 2020)

Our institution is passionate about high quality evidence based research and has excelled in various fields (Parakh et al. 2020; Pham et al. 2021; Perumal, Antony, and Muthuramalingam 2021; Sathiyamoorthi et al. 2021; Devarajan et al. 2021; Dhanraj and Rajeshkumar 2021; Uganya, Radhika, and Vijayaraj 2021; Tesfaye Jule et al. 2021; Nandhini, Ezhilarasan, and Rajeshkumar 2020; Kamath et al. 2020). Based on the literature survey it can be concluded that the existing accident severity prediction model was not able to accurately determine severity of crash. In the proposed work the lack of accuracy in predicting the accident crash severity is analyzed and improved using machine learning algorithms such as Novel Random Forest algorithm and Naive Bayes Algorithm.

MATERIALS AND METHODS

The study setting of the proposed work is done in the Image Processing Lab , Department of Computer science and Engineering at Saveetha School of Engineering, Saveetha Institute of Medical And Technical Sciences (SIMATS). The number of groups used for the study is two group classification algorithms. The Group 1 is Random Forest Algorithm and Group 2 is Naive Bayes Algorithm.Using clinical analysis (Geyik and Kara 2020), the analysis of sample size of N = 10 has been carried out with confidence of 95%. The input dataset is collected from kaggle.com (Accident Severity). It consists of two dataset namely test and train. Both the datasets have 17 attributes.

Random Forest Algorithm

Random Forest Algorithm extensively used for both regression and classification problems.This algorithm can be used in various places such as banking,prediction works,health,stock markets,artificial intelligence etc.To get more accurate and stable estimates the random forest will create forest like structures and combines them.Subsets from both datasets and attributes are selected arbitrarily and gets trained. Using this method overfitting of data can be lowered.This algorithm takes lower training time than many other algorithms on large datasets with maintaining precision of the accuracy when a huge part of data is not present. Table 2 contains the pseudocode of Random Forest Algorithm

Naive Bayes Algorithm

The Naive Bayes algorithm belongs to the family of probabilistic classifiers.the target values consist of categorical classes so it gives us easy option to put in and understand it.this algorithm is simple,elegant and is more durable than many other algorithms.it is used in the areas such as filtering of spam messages and for classifying the texts using artificial intelligence.First the Naive Bayes is imported from sklearn and the classifier is fit.This is a scalable algorithm which needs a linear set of parameters.Based on the scaled values precision scores are predicted.The advantage of this algorithm is it only needs small amount of data for training and predicting the parameters which is important for classification. Table 3 contains the pseudocode of Random Forest Algorithm

Statistical Analysis

The independent variables are junction detail,junction control whereas the dependent variables are policing area,week day of collision,hour of collision etc. The software used here for statistics is SPSS version 26. The data set is prepared by using 5 iterations each of both Random Forest and Naive Bayes. The testing variables are accuracy and loss whereas the group ID is given as grouping.

RESULTS

The mean accuracy and loss values using T-test for both the algorithms along with the standard deviation is shown in Table 6. Here RF and NVB classifiers are used.The performances of the classifiers are measured based upon the accuracy value.

Table 1 Contains the description of the dataset, the attribute which are present in the dataset incorporates like collision ref no, policing area, day, month, week, hour, weather conditions etc.Totally it consists of 1550 rows in the datasets which has null and duplicate values. Table 4 contains the accuracy for Random Forest classifier with N = 5 and Table 5 contains the accuracy for Naive Bayes classifier with N = 5. Random Forest Classifier input is taken from the datasets as a class form and gives the output accuracy of 91%. Naive Bayes classifier input is taken as class form from the data set and gives the output accuracy of 71%.Table 8 shows the accuracy values for both Random forest and Naive Bayes algorithm. It can be seen that Random forest has given better results of

accuracy than Naive Bayes. From Table 7 It can be observed that Random Forest has better significance value than Naive Bayes with a value of p = 0.000. Accuracy and the Loss for both the algorithms are presented in a bar graph in Fig. 1.

DISCUSSION

Random forest as convincingly appears better than Naive Bayes with improved accuracy The Random Forest classifier shows some difference in terms of accuracy score speed and performance in comparison with Naive Bayes.Comparison between each algorithm's performance in terms of accuracy and loss to prove that Random Forest is the best overall.

The outcome of this paper is similar to (AlMamlook et al. 2019) where they found that the Novel Random Forest Algorithm has 75% accuracy which is better than Naive Bayes Algorithm. A limitation of this study was that due to lack of appropriate data such as pedestrian details etc, the accident severity was not determined accurately. (Naguib, Abdel-Galil, and AbdelGaber 2020) carried out the experiment by using various text mining techniques for the purpose of predicting who is involved in an accident frequently and what conditions affect an accident. The accuracy of Random Forest turned out to be 87% whereas Naive Bayes gave an accuracy of 79%. (Mafi, AbdelRazig, and Doczy 2018) in this work RF is superior in predicting the severity of the crash more efficiently. (Abeyratne and Halgamuge 2020) used k-fold validation techniques on the collected data. The experiment was carried out in New York City. The findings of the experiment show that the Novel Random Forest algorithm showed the best accuracy when compared to KNN and Naive Bayes algorithm (Iranitalab and Khattak 2017). It also proved that there was an increase in motor vehicle collisions. Hence Random Forest can be used to predict accident crash severity efficiently. (Yassin and Pooja 2020) RF tops other algorithms with the prediction accuracy of (99.86%).

Most of the previous work done is based on the data from the specific dataset. In the proposed work the lack of accuracy in predicting the severity of accident is improved by using machine learning algorithms.In future more attributes in the dataset can be included and images can also be included in the dataset. Artificial intelligence can also be used in future to predict the severity of the crash.

CONCLUSION

The Calculation of accuracy of accident crash severity in highways is carried out using the dataset obtained from Kaggle.It is done by Random Forest and Naive Bayes.The accuracy of Random Forest classifier is 92% and Naive Bayes with 71% accuracy.The accuracy of predicting the accident crash severity is more in Random Forest than Naive Bayes..

DECLARATIONS

Conflict of Interests

No conflict of interest in this manuscript.

Authors Contributions

Author SV was involved in data collection, data analysis, manuscript writing and Author SRK was involved in conceptualization, data validation, and critical review of manuscript.

Acknowledgment

The authors would like to express their gratitude towards Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences (Formerly known as Saveetha University) for providing the necessary infrastructure to carry out this work successfully.

Funding: We thank the following organizations for providing financial support that enabled us to complete the study.

1. Soft Square Solutions, Palavakkam, Chennai.
2. Saveetha University
3. Saveetha Institute of Medical and Technical Sciences
4. Saveetha School of Engineering

REFERENCES

1. Abellán, Joaquín, Griselda López, and Juan de Oña. 2013. "Analysis of Traffic Accident Severity Using Decision Rules via Decision Trees." *Expert Systems with Applications*. https://doi.org/10.1016/j.eswa.2013.05.027.
2. Abeyratne, Dhanushka, and Malka N. Halgamuge. 2020. "Applying Big Data Analytics on Motor Vehicle Collision Predictions in New York City." *Intelligent Data Analysis*. https://doi.org/10.1002/9781119544487.ch11.
3. Ahmadi, Alidad, Arash Jahangiri, Vincent Berardi, and Sahar Ghanipoor Machiani. 2020. "Crash Severity Analysis of Rear-End Crashes in California Using Statistical and Machine Learning Classification Methods." *Journal of Transportation Safety & Security*. https://doi.org/10.1080/19439962.2018.1505793.
4. AlMamlook, Rabia Emhamed, Keneth Morgan Kwayu, Maha Reda Alkasisbeh, and Abdulbaset Ali Frefer. 2019. "Comparison of Machine Learning Algorithms for Predicting Traffic Accident Severity." *2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)*. https://doi.org/10.1109/jeeit.2019.8717393.
5. Assi, Khaled, Syed Masiur Rahman, Umer Mansoor, and Nedal Ratrout. 2020. "Predicting Crash Injury Severity with Machine Learning Algorithm Synergized with Clustering Technique: A Promising Protocol." *International Journal of Environmental Research and Public Health* 17 (15). https://doi.org/10.3390/ijerph17155497.
6. Devarajan, Yuvarajan, Beemkumar Nagappan, Gautam Choubey, Suresh Vellaiyan, and Kulmani Mehar. 2021. "Renewable Pathway and Twin Fueling Approach on Ignition Analysis of a Dual-Fuelled Compression Ignition Engine." *Energy & Fuels: An American Chemical Society Journal* 35 (12): 9930–36.
7. Dhanraj, Ganapathy, and Shanmugam Rajeshkumar. 2021. "Anticariogenic Effect of Selenium Nanoparticles Synthesized Using Brassica Oleracea." *Journal of Nanomaterials* 2021 (July). https://doi.org/10.1155/2021/8115585.
8. Dogru, Nejdet, and Abdulhamit Subasi. 2018. "Traffic Accident Detection Using Random Forest Classifier." *2018 15th Learning and Technology Conference (L&T)*. https://doi.org/10.1109/lt.2018.8368509.
9. Geyik, Buket, and Medine Kara. 2020. "Severity Prediction with Machine Learning Methods." *2020 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*. https://doi.org/10.1109/hora49412.2020.9152601.
10. Iranitalab, Amirfarrokh, and Aemal Khattak. 2017. "Comparison of Four Statistical and Machine Learning Methods for Crash Severity Prediction." *Accident; Analysis and Prevention* 108 (November): 27–36.
11. Kamath, S. Manjunath, K. Sridhar, D. Jaison, V. Gopinath, B. K. Mohamed Ibrahim, Nilkantha Gupta, A. Sundaram, P. Sivaperumal, S. Padmapriya, and S. Shantanu Patil. 2020. "Fabrication of Tri-Layered Electrospun Polycaprolactone Mats with Improved Sustained Drug Release Profile." *Scientific Reports* 10 (1): 18179.
12. Li, Zhibin, Pan Liu, Wei Wang, and Chengcheng Xu. 2012. "Using Support Vector Machine Models for Crash Injury Severity Analysis." *Accident; Analysis and Prevention* 45 (March): 478–86.
13. Mafi, Somayeh, Yassir AbdelRazig, and Ryan Doczy. 2018. "Machine Learning Methods to Analyze Injury Severity of Drivers from Different Age and Gender Groups." *Transportation Research Record: Journal of the Transportation Research Board*. https://doi.org/10.1177/0361198118794292.
14. Naguib, Ahmed Ibrahim, Hala Abdel-Galil, and Sayed AbdelGaber. 2020. "A Model for Traffic Management Based on Text Mining Techniques." *International Journal of Advanced Computer Science and Applications*. https://doi.org/10.14569/ijacsa.2020.0111280.
15. Nandhini, Joseph T., Devaraj Ezhilarasan, and Shanmugam Rajeshkumar. 2020. "An Ecofriendly Synthesized Gold Nanoparticles Induces Cytotoxicity via Apoptosis in HepG2 Cells." *Environmental Toxicology*, August. https://doi.org/10.1002/tox.23007.
16. Parakh, Mayank K., Shriraam Ulaganambi, Nisha Ashifa, Reshma Premkumar, and Amit L. Jain. 2020. "Oral Potentially Malignant Disorders: Clinical Diagnosis and Current Screening Aids: A Narrative Review." *European Journal of Cancer Prevention: The Official Journal of the European Cancer Prevention Organisation* 29 (1): 65–72.
17. Perumal, Karthikeyan, Joseph Antony, and Subagunasekar Muthuramalingam. 2021. "Heavy Metal Pollutants and Their Spatial Distribution in Surface Sediments from Thondi Coast, Palk Bay, South India." *Environmental Sciences Europe* 33 (1). https://doi.org/10.1186/s12302-021-00501-2.
18. Pham, Quoc Hoa, Supat Chupradit, Gunawan Widjaja, Muataz S. Alhassan, Rustem Magizov, Yasser Fakri Mustafa, Aravindhan Surendar, Amirzhan Kassenov, Zeinab Arzehgar, and Wanich Suksatan. 2021. "The Effects of Ni or Nb Additions on the Relaxation Behavior of Zr55Cu35Al10 Metallic Glass." *Materials Today Communications* 29 (December): 102909.
19. Sathiyamoorthi, Ramalingam, Gomathinayakam Sankaranarayanan, Dinesh Babu Munuswamy, and Yuvarajan Devarajan. 2021. "Experimental Study of Spray Analysis for Palmarosa Biodiesel-diesel Blends in a Constant Volume Chamber." *Environmental Progress & Sustainable Energy* 40 (6). https://doi.org/10.1002/ep.13696.
20. Tang, Jinjun, Jian Liang, Chunyang Han, Zhibin Li, and Helai Huang. 2019. "Crash Injury Severity Analysis Using a Two-Layer Stacking Framework." *Accident; Analysis and Prevention* 122 (January): 226–38.
21. Tesfaye Jule, Leta, Krishnaraj Ramaswamy, Nagaraj Nagaprasad, Vigneshwaran Shanmugam, and Venkataraman Vignesh. 2021. "Design and Analysis of Serial Drilled Hole in Composite Material." *Materials Today: Proceedings* 45 (January): 5759–63.
22. Uganya, G., Radhika, and N. Vijayaraj. 2021. "A Survey on Internet of Things: Applications, Recent Issues, Attacks, and Security Mechanisms." *Journal of Circuits Systems and Computers* 30 (05): 2130006.
23. Yassin, Salahadin Seid, and Pooja. 2020. "Road Accident Prediction and Model Interpretation Using a Hybrid K-Means and Random Forest Algorithm Approach." *SN Applied Sciences*. https://doi.org/10.1007/s42452-020-3125-1.
24. Zhang, Jian, Zhibin Li, Ziyuan Pu, and Chengcheng Xu. 2018. "Comparing Prediction Performance for Crash Injury Severity Among Various Machine Learning and Statistical Methods." *IEEE Access*. https://doi.org/10.1109/access.2018.2874979.

TABLES AND FIGURES

Table 1. Dataset Description

Column	**Values(For categorical variables)**	**Type**
Collision_Ref_No	Multiple Values Present	String, Categorical
Weekday_of_Collision	Multiple days present	String, Categorical
Special_conditoin_at_site	1 (Yes), 0 (No)	Numerical, Categorical
Ped_Crossing_HC	1 (Yes), 0 (No)	Numeric, Categorical
Policing_Area	Multiple Values present	String, Categorical
Hour_of_Collision	Multiple Values	Numeric, Categorical

Table 2. Pseudocode for Random Forest Algorithm

Input - Accident-severity (train and Test) Dataset
1. Initialization // input_attributes
2. The Random Forest Classifier is fit on the training set
3. Predict the records based on scaled values.
4. The features are then trained,tested and summarized
5. Predict the precision score
Output: Prediction of Accuracy

Table 3. Pseudocode for Naive Bayes

Input - Accident-severity (train and Test) Dataset
1. Initialization // input_attributes
2. Import Naive Bayes from sklearn
3. The Naive Bayes Classifier is fit
4. Predict the records based on scaled values.
5. Precision scores are predicted
Output: Prediction of Accuracy

Table 4. Random Forest Accuracy and Loss for N = 5

Iterations	Accuracy(%)	Loss(%)
1	91.70	8.30
2	92.10	7.90
3	91.65	8.35
4	91.65	8.35
5	92.24	7.76

Table 5. Naive Bayes Accuracy and Loss for N = 5

Iterations	Accuracy(%)	Loss(%)
1	71.81	28.19
2	71.81	28.19
3	71.80	28.20
4	71.89	28.11
5	70.98	29.02

Table 6. T-Test Group Statistics with Mean, Std.Deviation, Std.Error Mean and Confidence = 95%

	Group	N	Mean	Std. Deviation	Std. Error Mean
Accuracy	Random Forest	5	91.8680	.28084	.12559
	Naive Bayes	5	71.4440	.54007	.24153
Loss	Random Forest	5	8.1320	.28084	.12559
	Naive Bayes	5	28.5560	.54007	.24153

Table 7. Independent Sample T-Test is applied for the data set fixing confidence interval as 95% and level of significance as 0.05

	Levene's test for equality of variances		T-test for equality means with 95% confidence interval						
	f	Sig.	t	df	Sig. (2-tailed)	Mean difference	Std.Error difference	Lower	Upper

Accuracy	Equal variances assumed	13.146	.007	75.024	8	.000	20.42400	.27233	19.796	21.05177
	Equal Variances not assumed			75.024	6.016	.000	20.42400	.27233	19.75830	21.08970
Loss	Equal variances assumed	13.146	.007	-75.024	8	.000	-20.42400	.27233	-21.05177	-19.79623
	Equal variances not assumed			-75.024	6.016	.000	-20.42400	.27233	-21.08970	-19.75830

Table 8. Comparison of the Random Tree Algorithm and Naive Bayes Algorithm with with their accuracy

CLASSIFIER	**ACCURACY(%)**	**LOSS(%)**
Random Forest	91.70	8.30
Naive Bayes	71.81	28.19

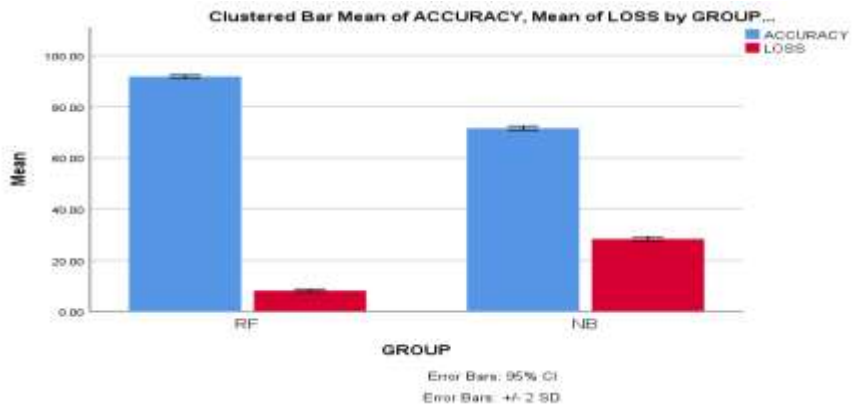

Fig. 1. Comparison of Random Forest algorithm And Naive Bayes in terms of mean accuracy. The mean accuracy of Random Forest is better than Naive Bayes. X axis(Groups): RF VS NVB algorithm, Y axis: Mean accuracy of prediction +/- 2SD