

ACCIDENT SEVERITY ESTIMATION
FROM ENVIRONMENTAL FACTORS
WITH THE PROPOSED DATASET SEPARATION METHOD

Kerim Karadağ[#], Mehmet Ayhan Çetin

Received on May 16, 2022

Presented by Ch. Roumenin, Member of BAS, on July 19, 2022

Abstract

Traffic accidents can be thought of as being the whole combination of events that usually occur due to reasons caused by humans, vehicles, roadway, and environment, which are the basic elements of transportation, and these events most likely result in injuries, death, and huge financial damage. In this study, by making use of a dataset, received from the New Zealand Transportation Agency, it is estimated whether or not traffic accidents occurring on roads end up with minor or serious damage. Predictions are made by exploiting machine learning algorithms. A new method has been proposed for separating the dataset and the classification success results made in this way gave the most successful scores by 89%. Based on the high performance rates, it is shown that such methods can be used to assist in preventing possible traffic accidents.

Key words: road accidents, environmental factors, machine learning algorithms, classification, performance evaluation

Introduction. Unfortunately, it is an inevitable fact that technological development across the globe has negative effects besides its positive aspects. While the rise of the number and diversity of vehicles in use today is a crucial sign of technological improvement, it is also one of the main reasons that gives rise to increasing number of traffic accidents simultaneously. Growing number of traffic

[#]Corresponding author.

DOI:10.7546/CRABS.2022.11.12

accidents is one of the major hazards for people which threaten their life. Therefore, it is important to predict accidents in advance and react to them quickly to reduce such dangers and damage [1]. Traffic accidents can be considered to be the whole combination of events that usually occur due to reasons caused by humans, vehicles, roadway, and environmental conditions, which are the basic elements of transportation, and, these events most likely result in injuries, death, and huge financial damage. Among these factors, driver faults alone are responsible for 57% of accidents while road and vehicle defects together cause 93% of accidents [2]. Traffic accidents are an issue that causes serious economic losses. Hence, a detailed analysis of the factors affecting accidents is needed to reduce the accident rate [3]. Since traffic accidents may occur because of many different reasons, ensuring traffic safety is a relatively complicated problem. The reasons for the occurrence of traffic accidents must be determined correctly in order to use the resources of the country, which may be limited, to accurately produce effective results. Hence, people, vehicles, road and weather conditions come to the fore as the major reasons for the occurrence of traffic accidents [4]. Additionally, traffic jam leads to negative impacts on human health, and also it might cause fuel wastage and unproductive use of time spent on roads [5]. Luckily, there is an innovative approach which exploits the machine learning based prediction methods to forecast accidents that may occur in the future [3]. However, there may be some issues to create such a prediction system. Since data needed to estimate traffic accidents beforehand is very large in volume, collection and analysis of all the traffic data require a huge amount of hardware resources [1]. There are many accident prediction models developed in the literature with various methods. In these studies, it is seen that either machine learning based methods or statistical methods are used. In his study in 2018, VASAVI [6] examined the traffic accident data of major national highways crossing the Krishna region using machine learning techniques. In their study, SINGH et al. [7] examined the traffic accidents by using naive Bayes method, k-nearest neighbour algorithm, decision trees, and support vector machines. SINGH et al. [8] examined and studied geometry of roads, traffic, and environment related data records obtained from Police Department. These records include 15 input parameters as well as 222 data points, regarding the traffic accidents which happened on the highways in rural regions of India, in sum. IRANITALAB and KHATTAK [9], in their study, made predictions of collision severity using machine learning methods by exploiting accident data received from Nebraska in USA between 2012–2015. In addition, data of the first two years is used for training whereas the rest of the data is used for testing. In their study, CHONG et al. [10] applied machine learning techniques by making use of parameters such as vehicle brand and model, driver's age, driver's gender, alcohol consumption of the driver, speed, illumination on roads, road surface quality, accident collision type, restraint systems, speed limit, and decommission that have a huge impact on the intensity of injuries in traffic accidents. In their study, ZOU

and YUE [11] have shown that the Bayesian network model can effectively reveal the complex logical relationship in traffic accidents and express the uncertain relationship between the relevant variables. In their work, BABIČ and ZUSKÁČOVÁ [12] tried to identify the factors causing traffic accidents with machine learning methods such as decision trees using data that were effective in traffic accidents in the United Kingdom between 2005 and 2015. YAN et al. [13] in their study, unearthed the effect of driver behaviour (phone, cigarette, food, etc.) in traffic accidents by using convolutional neural networks. In their study, BÜLBÜL et al. [14] aimed to determine the most appropriate machine learning classification techniques for traffic accident prediction by means of data mining. NARASIMHAN et al. [3] made a traffic accident prediction by utilizing the python and tableau methods in their studies that exploit the data of traffic accidents between 2011 and 2015. TAAMNEH et al. [15] by utilizing the traffic data in Abu Dhabi between 2008 and 2013, endeavored to forecast the severity of injuries in traffic accidents using artificial neural networks and hierarchical clustering methods. So as to analyze the factors that affect the intensity of accidents for two-wheel vehicles, KUMAR and TOSHNIVAL [16] in their study, used naive Bayes, support vector machines, and decision trees. In their study, PARK et al. [1] tried to create a prediction model using traffic data between 2011 and 2013 in Korea. CONDURAT et al. [5] in their work, tried to estimate traffic density using machine learning with data obtained from sensors placed on vehicles. SOHN and SHIN [17], using artificial neural network, logistic regression, decision tree methods, investigated the reasons affecting the severity of traffic accidents in Korea.

In recent years, it has been seen that many studies focused on traffic accident analysis and forecasts have been done. However, in these studies, statistical models act according to predefined fixed assumptions between dependent and independent variables, and the results may change if some assumptions are ignored. These problems were prevented by using machine learning methods, and results with high accuracy rates were obtained in traffic accident estimations. In this study, in addition to the studies mentioned in the literature, the dataset created with different normalization methods were compared with different performance criteria, and a method, which is shown to increase the success rate of the performance criterion, was proposed.

Material and method. Dataset used in our study. In this study, we benefit from the data, which was recorded at the time of accidents occurred, obtained from New Zealand Transportation Agency [18]. By exploiting the different 88 pieces of information related to road, vehicle, and driver at the time of accident, it was tried to predict the severity of accidents. In order to create our dataset, 17 different features, such as slope of road, number of lanes, illumination, speed limit, wetness or dryness of the road, existence of pedestrians, riders, or trees just to name a few, which are relatively effective for the occurrence of accidents are used. As a consequence, we produce either of two outputs, namely, Fatal-Serious (FS)

and Minor-Non-injury (MN), to forecast the severity of accident. Records that consist of 561492 samples and 88 features are gained between January 2000 and October 2018. Upon gathering data, records are divided into four different categories, namely, Fatal (F), Serious (S), Minor (M), and Non-injury (N). However, in our study, we combine Fatal and Serious categories, and call the new category Fatal-Serious (FS). In the same manner, we combine Minor and Non-injury categories, and call the new one Minor-Non-injury (MN). Therefore, in this way, we now have two categories instead of four categories. Moreover, rather than using all 88 pieces of information regardless of their direct or indirect effect on the accidents, we solely take into account 17 prominent features that have higher impact compared to others for the classification. Additionally, textual data is converted into numerical data, and then, normalized. When the dataset is divided into two categories, since the number of samples of the FS accidents is 4858, other samples are blended among themselves so that approximately equal number of samples that represent MN category are created and processed. Finally, our dataset has 9780 samples and 17 features in total. Of all samples, 4858 are FS accidents whereas 4922 are MN accidents. Parameters and features used in our dataset in this study are shown in Table 1.

T a b l e 1
The properties of datasets

No	Parameters	Features
1	Pedestrian	0-1
2	Bicycle	0-1-2-3
3	Car Stations	0-1-2-3-4-5-6-7-8
4	Weather	Fine, Heavy Rain, Mist, Light Rain, Snow
5	Number of Lanes	0-1-2-3-4-5-6-7-8
6	Direction of the road	East West North South
7	The road is flat or hill	Flat, Hill
8	Light	Overcast, Twilight, Bright Sun, Dark
9	Dark light	Dark, Light
10	Multi-vehicle	Bike only, Multi-vehicle, Other, Single vehicle-Vehicle + bicycle, Many other vehicles, Vehicle + pedestrian
11	Road Curvature	Low slope, Medium slope, Multi-slope, Straight road
12	Road Lane	1-2
13	Road Marking	Raised Island, Centre Line, No Marks, No Passing Lines, Painted Islands, Pedestrian Crossing,
14	Road Wet	Dry, Icy, Snowy, Wet
15	Speed Limit	20-30-40-50-60-70-80-90-100
16	Tree	0-1-2
17	Urban	Open road, Urban
18	Crash Severity	Fatal-Serious (FS), Minor-Non-injury (MN)

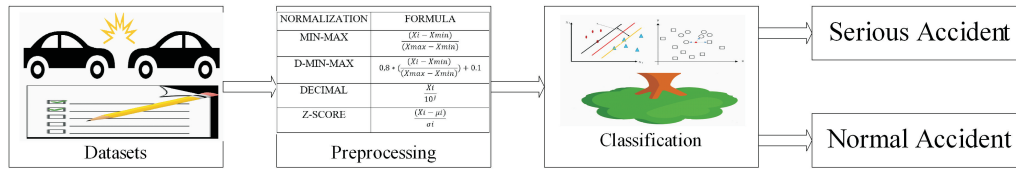


Fig. 1. Flow chart of the study

Proposed method for the prediction of accident severity. In the scope of our study, the flow chart of the proposed system for the prediction of the severity of accidents is given in Fig. 1. At first, obtained data is conveyed to computers, and then, feature vector of the records is created. Subsequently, data is normalized to [0,1] interval, and is classified by means of Support Vector Machines (SVM), K-Nearest Neighbour (KNN), Ensemble Tree (ET), and Decision Tree (DT). Computation of success parameters are done on a computer with 2.8 GHz Intel Core i7 processor and 8 GB RAM hardware properties.

Application and achievements. In this study, from the data received from the New Zealand Transportation Agency, it is aimed to predict the FS and MN accidents by machine learning algorithms. The dataset created by normalization is blended and classified into three different cases. In the first case 10-fold cross validation (10-fold CV) mean is used. In the CV, the dataset is divided into 10 (9 training and 1 testing). This process is repeated by shifting 10 times ($I_1, I_2, \dots, I_{10}$). The formula for CV is given in eq. 1.

$$(1) \quad I_{CV} = \frac{1}{10} \sum_{i=1}^{10} I_i.$$

In the second case, for the 10-fold Highest Average Cross Validation (HACV), average of the highest True Positive (HTP) and the highest False Positive (HFP), values are taken and classified. The formula for HACV is given in eq. 2.

$$(2) \quad A_{HACV} = \frac{A_{HTP} + A_{HFP}}{2}.$$

Finally, in equivalent to the second case, 80% of the data is used for training and 20% of the data is used for testing for the completion of the classification.

Performance gained using SVM, KNN, ET and DT classification methods is obtained by using accuracy, sensitivity, specificity, precision, F-score, MCC and ROC performance criterions. The interception points of the ROC curve are chosen according to Youden Index (YI) and mean values.

The training and test values were created in three different ways and the number of true and false values obtained with the classification methods used is given in Table 2.

T a b l e 2
True and false values

(10-fold CV)											
DT	4080	777	SVM	4053	804	KNN	4156	701	ET	4104	753
	622	4301		628	4295		807	4116		620	4303
(10-fold HACV)											
DT	835	135	SVM	821	150	KNN	838	113	ET	826	111
	108	878		118	876		141	844		141	878
(% 80 Train- %20 Testing)											
DT	802	170	SVM	815	157	KNN	823	149	ET	807	165
	144	840		167	817		189	795		137	847

In the study, which was processed by separating the training and test data with 10-fold CV, HACV and %80 training-%20 testing; different performance criteria such as accuracy, sensitivity, specificity, precision, F-score, MCC, YI and mean are examined and the success rates are given in Table 3.

When the obtained success rates are compared for accuracy, the highest score obtained using the ET classification method and the success rate is 87.6%. For the sensitivity, the highest score obtained by the KNN and ET classification methods and its success rate is 88.1%. For specificity, the highest score obtained by the DT classification method, and this score is 89.0%. For precision, the highest score obtained using the DT classification method and this score is 88.5%. For F-score, the highest score obtained by DT classification method and this score is 87.5%. For MCC, the highest score obtained by DT classification method and this score is 75.17%. Considering the methods used to determine the interception point of the ROC curve, the highest score obtained using DT classification method and this score is 75.1%. For mean, the highest score obtained by DT classification method and this score is 87.6%. When classification methods are compared, the success rates obtained with the DT method are higher than the other classification methods.

In three applications for classification estimation, the highest scores are achieved using the recommended performance criteria. The success rate for accuracy is 87.6 using the DT method, and the highest success rate for sensitivity is 88.8 with the ET method.

In a study using the same dataset, regression analysis is made between the number of accidents and the number of injuries and deaths, and evaluations are made on fatal, serious, minor and simple non-injury accidents [19]. In addition, the same dataset was used by ÇETİN [20] in his graduate study. However, in this study, different performance evaluation was used and a new method was proposed to separate the dataset. And the success rates obtained with the proposed method have been higher. In this study, the estimation of severe and minor accidents

T a b l e 3

Achievement performance

Application	Classification Methods	Accuracy	Sensitivity	Specificity	Precision	F-Score	MCC	ROC	
								YI	Mean
CV	DT	0.857	0.840	0.873	0.867	0.856	0.714	0.713	0.857
	SVM	0.854	0.834	0.872	0.866	0.852	0.707	0.707	0.853
	KNN	0.846	0.846	0.836	0.837	0.841	0.691	0.692	0.846
	ET	0.860	0.845	0.874	0.869	0.859	0.719	0.719	0.859
HACV (proposed method)	DT	0.876	0.861	0.890	0.885	0.875	0.7517	0.751	0.876
	SVM	0.864	0.845	0.881	0.874	0.863	0.727	0.727	0.863
	KNN	0.869	0.881	0.857	0.856	0.869	0.738	0.738	0.869
	ET	0.871	0.881	0.862	0.854	0.871	0.742	0.743	0.871
80% training 20% test	DT	0.839	0.825	0.854	0.848	0.839	0.680	0.679	0.839
	SVM	0.834	0.838	0.830	0.830	0.834	0.669	0.669	0.834
	KNN	0.827	0.847	0.808	0.813	0.827	0.655	0.655	0.827
	ET	0.846	0.830	0.861	0.855	0.845	0.691	0.691	0.845

from the current accident values is made using classification methods. Thus, the parameters affecting the accidents can be analysed and the severity of the accidents can be reduced. In addition, the proposed system can be employed in performance criteria used in classification methods.

Conclusion. In this study, dataset was created by using normalization method by selecting the features that are thought to have an impact on accident occurrence and severity. Accidents were divided into two categories, namely, FS and MN, and were compared with different performance criteria by using ML algorithms such as SVM, KNN, ET and DT. When evaluated, in general, the estimation methods used were found to be successful and it was observed that the ET and DT methods were more successful in this study. Based on the results obtained, it can be said that this study will be useful in terms of giving ideas to road and transportation experts in order to ensure reducing or taking preventive precautions for traffic accidents. In addition, some situations that have an impact on the severity of the accident can be added to the pre-prepared 'Accident report form' to give supportive ideas to the insurance companies and the crasher in the expert reports.

REFERENCES

- [1] PARK S. H., S. M. KIM, Y. G. HA (2016) Highway traffic accident prediction using VDS big data analysis, *The Journal of Supercomputing*, **72**(7), 2815–2831.
- [2] RUMAR K. (1985) The Role of Perceptual and Cognitive Filters in Observed Behavior. In: Evans L., Schwing R.C. (Eds) *Human Behavior and Traffic Safety*, Springer, Boston, MA, https://doi.org/10.1007/978-1-4613-2173-6{_}8.
- [3] NARASIMHAN G., B. G. EPHREM, S. CHERIYAN, N. BALASUPRAMANIAN (2017) Predictive analytics of road accidents in Oman using machine learning approach. In: 2017 International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT), 1058–1065, IEEE.
- [4] BAŞ F. İ. (2015) The Effect of Fatigue and Insomnia from Temporary Physical Characteristics on Driver Behavior, Master Thesis, Atatürk University, Erzurum, Turkey.
- [5] CONDURAT M., A. M. NICUTA, R. ANDREI (2017) Environmental Impact of Road Transport Traffic. A Case Study for County of Iaşi Road Network, *Procedia Engineering*, **181**, 123–130.
- [6] VASAVI S. (2018) Extracting Hidden Patterns within Road Accident Data Using Machine Learning Techniques. In: Mishra D., Azar A., Joshi A. (Eds), *Information and Communication Technology, Advances in Intelligent Systems and Computing*, vol. 625, Springer, Singapore.
- [7] SINGH J., G. SINGH, P. SINGH, M. KAUR (2019) Evaluation and Classification of Road Accidents Using Machine Learning Techniques. In: Shetty N., Patnaik L., Nagaraj H., Hamsavath P., Nalini N. (Eds), *Emerging Research in Computing, Information, Communication and Applications, Advances in Intelligent Systems and Computing*, vol. 882, Springer, Singapore.

- [8] SINGH G., S. N. SACHDEVA, M. PAL (2016) M5 model tree based predictive modeling of road accidents on non-urban sections of highways in India, *Accident Analysis & Prevention*, **96**, 108–117.
- [9] IRANITALAB A., A. KHATTAK (2017) Comparison of four statistical and machine learning methods for crash severity prediction, *Accident Analysis & Prevention*, **108**, 27–36.
- [10] CHONG M., A. ABRAHAM, M. PAPRZYCKI (2005) Traffic accident analysis using machine learning paradigms, *Informatica (Slovenia)*, **29**(1), 89–98.
- [11] ZOU X., W. L. YUE (2017) A Bayesian network approach to causation analysis of road accidents using Netica, *Journal of Advanced Transportation*, Article ID 2525481.
- [12] BABIČ F., K. ZUSKÁČOVÁ (2016) Descriptive and predictive mining on road accidents data. In: 2016 IEEE 14th International Symposium on Applied Machine Intelligence and Informatics (SAMII), 87–92.
- [13] YAN S., Y. TENG, J. S. SMITH, B. ZHANG (2016) Driver behavior recognition based on deep convolutional neural networks. In: 2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), 636–641.
- [14] BÜLBÜL H. İ., T. KAYA, Y. TULGAR (2016) Analysis for Status of the Road Accident Occurrence and Determination of the Risk of Accident by Machine Learning in Istanbul. In: 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), 426–430.
- [15] TAAMNEH M., S. TAAMNEH, S. ALKHEDER (2017) Clustering-based classification of road traffic accidents using hierarchical clustering and artificial neural networks, *International Journal of Injury Control and Safety Promotion*, **24**(3), 388–395.
- [16] KUMAR, S., D. TOSHNIWAL (2017) Severity analysis of powered two wheeler traffic accidents in Uttarakhand, India, *European Transport Research Review*, **9**(1), 24.
- [17] SOHN S. Y., H. SHIN (2010) Pattern recognition for road traffic accident severity in Korea, *Ergonomics*, **44**(1), 107–117.
- [18] <https://opendata-nzta.opendata.arcgis.com/datasets/207f07e874de4220a6da63fa0e29f229>
- [19] <https://colab.research.google.com/gist/shakasom/382a5cc0c59e24fdae29679de6d7e88a/crash-analysis.ipynb{\#}scrollTo=m218tlnzKeFH>
- [20] ÇETİN M. A. (2020) Accident Severity Estimation by Using the Factors of Traffic Accident Formation Using Machine Learning Methods, Master Thesis, Harran University Institute of Science and Technology, Şanlıurfa.

Electrical & Electronics Engineering Department
Faculty of Engineering
Harran University
Osmanbey Campus
63000 Haliliye, Şanlıurfa, Turkey
e-mail: k.karadag@harran.edu.tr
cetinayhan2@hotmail.com