©2021 Published in 5th International Symposium on Natural Hazards and Disaster Management
Sakarya Uygulamalı Bilimler Üniversitesi Sakarya TURKEY 5-7 NOV 2021
DOI: doi.org/10.33793/acperpro.04.02.54

Analysis of Pedestrian-Vehicle Crashes Using Artificial Learning Methods: City of Sakarya Case Study

[1]Zeliha Cagla Kuyumcu, [2]Suhrab Ahadi and *[1]Hakan Aslan
[1]Faculty of Engineering, Department of Civil Engineering Sakarya University, Turkey
[2]Faculty of Engineering, Department of Geodesy Engineering Jawzjan University, Afghanistan

Abstract

The lives of approximately 1.3 million people are cut short every year as a result of road traffic crashes. Up to 50 million people suffer non-fatal injuries, with many incurred disabilities as a result of their injury. The risk of dying in a road traffic crash is more than 3 times higher in low-income countries than in high-income countries [1]. In Turkey, 18% of traffic accidents was related to pedestrian-vehicle collisions in urban roads according to the figures of 2020. In addition, 20% of death toll caused by accidents is pedestrians in 2020 [2]. This study deals with the some of the classifiers to forecast the pedestrian lapses of traffic accidents. The classifier's performance ratios were also examined.

Key words: traffic accidents, pedestrian crashes, data mining algorithms

1. Introduction

Pedestrians, cyclists, and riders of motorized 2- and 3- wheelers and their passengers are collectively known as "vulnerable road users" and account for half of all road traffic deaths around the world. A higher proportion of vulnerable road users die in low-income countries than in high-income countries [1].

For reducing the fatalities of pedestrians,

- Drivers should be more careful to see pedestrians, especially in bad weather conditions and at night when visibility is low.
- Drivers should slow down when entering pedestrian crossings.
- Drivers should stop at the stop line and give them the opportunity to see pedestrians and stop at the crosswalk for the driver in the other lane.
- Drivers should pay attention to pedestrians, especially to small children.
- Pedestrians must be in predictable places, obey traffic signs and rules, and cross the road through pedestrian crosswalks.
- Pedestrians should walk facing and as far away from traffic as possible if there is no sidewalk.
- Pedestrians should be mindful of the traffic moving around them, this is not the time to text or talk on a cell phone.

* Corresponding author: Address: Faculty of Engineering, Department of Civil Engineering Sakarya University, 54187, Sakarya TURKEY. E-mail address: haslan@sakarya.edu.tr, Phone: +902642955752

- Pedestrians should make eye contact with drivers as the vehicle approaches, never assuming that the driver sees them.
- Pedestrians should wear bright clothing during the day and reflective materials (or use a flashlight) at night.
- Pedestrians should look left-right-left (right-left-right in some countries) before crossing the street [3].

This paper mainly involves the specific investigation of the classifiers` performances for pedestrian-vehicle crashes.

2. Data Description and Processing

2.1. Raw data and study area

The raw data for this study is related with the pedestrian-involved accidents in city of Sakarya and was obtained from the Turkish Republic General Directorate of Security.

2.2. Data processing

The data analysed includes 473 pedestrian-vehicle accidents occurred in 6 years' time period from 2006-2010 in Sakarya, Turkey. Age categories were classified into four groups: child, young, middle aged, and elderly.

Some variables were excluded from the data set as they were very small as a result of the sample distribution. Furthermore, meaningless data were also removed from the dataset.

2.3. Structured Dataset Construction

For the analysis purposes, twelve variables and 443 reported crashes were filtered from 473 reported crashes as far as data processing is concerned. The description of the twelve variables are illustrated in Table 1.

3. Methodology

3.1. Classifiers

In this study, five of the machine learning algorithms were used. Three of them are tree algorithms (ID3, J48 and Simple CART).

J48 is a simple C4.5 decision tree algorithm creating a binary tree. C4.5 builds decision trees from a set of training data using the concept of information entropy [4]. Decision tree models have several advantages over traditional statistical approaches. These algorithms are non-parametric models not necessitating the normality premises. The models can handle different types of variables

for instance nominal, ordinal, continuous variables [5]. WEKA machine learning programme, known as one of the most efficient and well known software, was chosen for this study to analyse the data.

Table 1. Description of variables in the original dataset

NO.	Variable	Converted Variable Type	Label
1	Hour	nominal	1,2
2	Road type	nominal	1,2
3	Day light	nominal	1,2,3
4	Traffic sign	nominal	1,2
5	Traffic signal	nominal	1,2
6	Illumination	nominal	1,2
7	Driver education	nominal	1,2,3,4
8	Pedestrian age	nominal	1,2,3,4
9	Pedestrian lapses	nominal	1,2
10	Pedestrian gender	nominal	1,2
11	Road direction	nominal	1,2
12	Road Surface	nominal	1,2

Naive Bayes is a machine learning algorithm based on Bayes' probability theorem. In the Naive Bayes method, when the class to which each data instance belongs is clear, the aim is to establish a rule that will determine the class label of the next data instance to come [6]. This situation is also called conditional probability. It is based on the principle of the value received by the corresponding attribute when the data class is labelled.

Bayesian networks are also known by other names such as Bayes nets or belief nets and considered from the family of graphical models. Bayesian networks often use directed acyclic graph (DAG), considered as the qualitative parameters of the model, to estimate the conditional probabilities among random variables as each random variable is represented in the Bayesian network by a node [7].

4. Results

Classifications were carried out according to the pedestrian lapses of the accidents. While the original data set in the first case were analysed, the data set created by eliminating the variables with high correlation with each other was studied in the second case. Following, the success levels of various classification methods were compared.

Pedestrian lapses and violation of right of way of crossing over are included in the dataset. Although results were obtained for all these stated classes in this study, only the classification results of pedestrians suddenly appearing on the roadways were illustrated.

All the attributes involved in 443 records of all accident dataset were classified by Naïve Bayes, Bayes Net, ID3, J48 and Simple CART classifiers. While the obtained performance values are listed in Table 2, for the nine attributes performance values are listed in Table 3.

Table 2. Comparison of used classifiers for original dataset

Method	Accuracy	AUC	Precision	Recall	F-measure
Naive Bayes	0.63	0.65	0.68	0.73	0.70
Bayes Net	0.63	0.65	0.68	0.72	0.70
ID3	0.56	0.56	0.64	0.64	0.64
J48	0.65	0.53	0.66	0.83	0.74
Simple CART	0.65	0.64	0.67	0.80	0.73

Some of these attributes are irrelevant or inadequate to express the severity of the model. For this reason, Principal Components Analysis (PCA)is employed to investigate the potential attributes, which lead to better classification. In this method, the main purpose is to transform the related features into non-related features, also called fundamental components, using mathematical procedures. Here, the number of principal components is equal to or less than the initial number of features [8].

Table 3. Comparison of used classifiers for 6-attributes dataset

Method	Accuracy	AUC	Precision	Recall	F-measure
Naive Bayes	0.65	0.66	0.69	0.75	0.72
Bayes Net	0.65	0.66	0.69	0.75	0.72
ID3	0.53	0.58	0.59	0.70	0.64
J48	0.61	0.58	0.64	0.79	0.71
Simple CART	0.60	0.61	0.63	0.79	0.70

After the PCA, 6 attributes are remained for classification. The attributes, significant for classification, are; traffic signs, traffic signal, illumination, pedestrian lapses, pedestrian age, and driver education. The related description of these attributes are listed in Table 4.

Table 4. Description and information field of variables

NO.	Variable	Variable Type	Information Fields	Percentage
1	traffic sign	nominal	1=not available	0.71
			2=available	0.29
2	traffic signal	nominal	1=not available	0.90
			2=available	0.10
3	illumination	nominal	1=not available	0.51
			2=available	0.49
4	Driver education	nominal	1=primary school	0.46
			2=secondary school	0.08
			3=high school	0.30
			4=university	0.16
5	pedestrian lapses	nominal	1=abrupt appearance on the road	0.59
			2=violation of crossing over	0.41
6	pedestrian age	nominal	1=child	0.29
			2=young	0.22
			3=middle aged	0.33
			4=old	0.16

$$\text{Accuracy} = \frac{TP+TN}{(TP+TN+FP+FN)} \qquad \text{AUC} = \frac{TPR-TNR}{2} \qquad \text{Precision} = \frac{TP}{(TP+FP)}$$

$$\text{F-measure} = \frac{2*P*R}{P+R} \qquad \text{Recall} = \frac{TP}{(TP+FN)}$$

TP=True positive TN=True negative FP=False positive FN=False negative

P=Precision R=Recall TPR=True positive rate TNR=True negative rate

Some of classification models, Naive Bayes, Bayes Net, ID3, J48 and Simple CART classifiers, to predict the pedestrian lapses on traffic accidents were compared for classifying the pedestrian lapses for various traffic accidents. The final results illustrate that the ID3 outperforms the other four algorithms for original dataset and Bayes Net outperforms the other four algorithms for reduced dataset.

The ROC curve is used to evaluate the balance between accuracy and sensitivity. The area under the ROC curve is defined as the ROC score. The ROC curve is plotted based on varying classification thresholds of true positives as a function of false positives. A ROC score of "1"

indicates excellent separation of positives from negatives. If the ROC score is "0", it means that no positives were found [9].ROC analysis is commonly used for binary-class problems, but is also suitable for multi-class problems. For multi-class problems, a two-class approach is used. While one of these approaches is the "one versus the one" approach, the other is "one versus all" approach. In the "one against one" approach, each class is compared in pairs with each of the other classes. In the "one versus all" approach, however, for class "t", all other classes are marked as "not t" and compared to "t" [10]. In this study, a multiclass ROC analysis was applied using a "one against one" approach. Precision is the rate at which the predicted positive class is actually positive. Recall is the proportion of true positives predicted correctly. Since these two metrics are also important performance metrics, the F1-score is also calculated using these two metrics [11].

ROC curves and complexity matrices of the most successful method obtained with all datasets Table 5-6 and Figure 1-2 shown throughout.

Table 5. Original dataset J48

True Class		Abrupt appearance on the road	Violation of crossing over
	Abrupt appearance on the road	217	45
	Violation of crossing over	111	70

Predicted Class

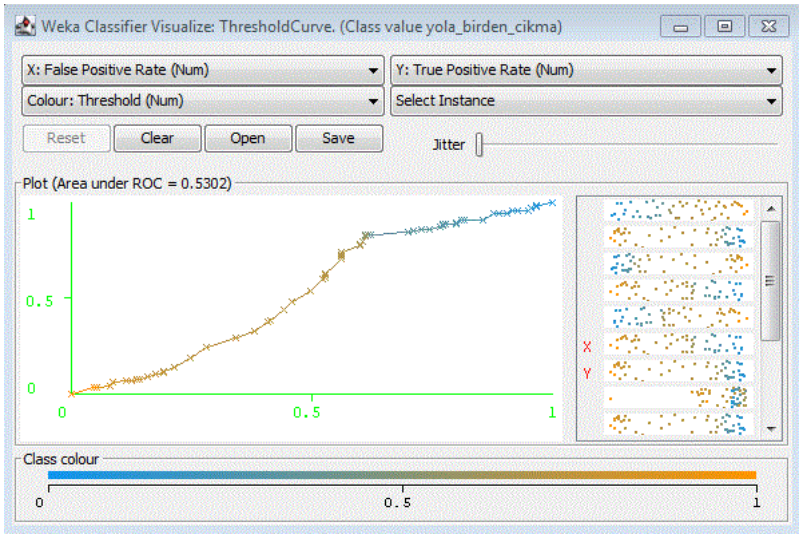

Figure 1. Area under curve (AUC) for J48 classifier

Table 6. Reduced dataset Naïve Bayes

True Class		Abrupt appearance on the road	Violation of crossing over
	Abrupt appearance on the road	196	66
	Violation of crossing over	88	93

Predicted Class

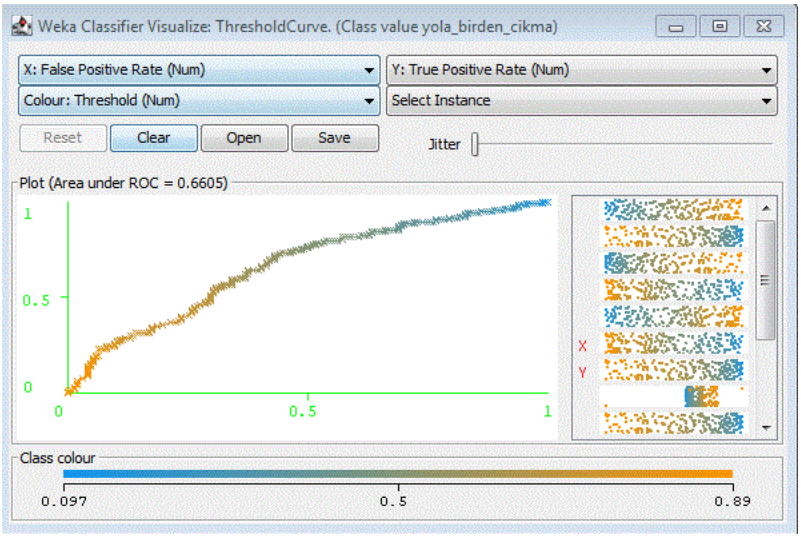

Figure 2. Area under curve (AUC) for Naïve Bayes classifier for reduced dataset

4. Conclusions

In this study, performance of some classification algorithms were investigated for traffic accidents. Pedestrian lapses were analysed with these algorithms and the related performances were compared. J48 and Simple CART algorithms are the most predictive for original dataset. Naive Bayes and Bayes Net algorithms are the most predictive for reduced dataset. It is obvious that more data may have contributed to better analysis to measure the performance. Therefore, classification performances will be examined with data sets with high quality and large number of variables in the future studies.

References

[1] https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries

[2] Traffic Accidents Summary, General Directorate of Highways-Republic of Turkey.

[3] NHTSA_pedestrian_aug2013_9718.pdf. (t.y.). United States Department of Transportation. Geliş tarihi 05 Kasım 2021, gönderen https://www.nhtsa.gov/sites/nhtsa.gov/files/documents/s1n_pedestrian_aug2013_9718.pdf

[4] Krishnaveni, S., & Hemalatha, M. (2011). A Perspective Analysis of Traffic Accident using Data Mining Techniques. International Journal of Computer Applications, 23(7), 40-48. https://doi.org/10.5120/2896-3788

[5] AlKheder, S., AlRukaibi, F., & Aiash, A. (2020). Risk analysis of traffic accidents' severities: An application of three data mining models. ISA Transactions, 106, 213-220. https://doi.org/10.1016/j.isatra.2020.06.018

[6] Wu, X., Kumar, V., Quinlan, JR, Ghosh, J., Yang, Q., Motoda, H., ... & Zhou, Z.H., Top 10 algorithms in data mining, Knowledge and information systems, 14 (1), 1-37, 2008.

[7] Ben-Gal I. Bayesian networks. In: Encyclopedia of statistics in quality and reliability. 2008, W475W9360.

[8] https://docs.rapidminer.com/, accessed on 17 Jan 2021.

[9] Hand D. J., Till R. J., A simple generalization of the area under the ROC curve for multiple class classification problems, Machine Learning, Vol.45 pp:171-186, 2001.

[10] Wandishin M. S., Mullen S.J., Multiclass ROC analysis, Weather and Forecasting Vol.24, DOI: 10.1175 / 2008WAF2222119.1, 2008.

[11] Ramzan F., Khan M. U. G., Rehmat A., Iqbal S., Saba T., Rehman A., Mehmood Z., A deep learning approach for automated diagnosis and multi-class classification of Alzheimer's disease stages using resting-state fMRI and residual neural networks, Journal of Medical Systems vol.44(37), 2020.