APPLICATION OF THE DECISION TREE TECHNIQUE IN THE ANALYSIS OF TRAFFIC ACCIDENTS

Aleksandar Mamić[1], Marija Blagojević[1], Danijela Milošević[1]
acomamic@gmail.com, marija.blagojevic@ftn.kg.ac.rs, danijela.milosevic@ftn.kg.ac.rs
University of Kragujevac, Faculty of Technical Sciences Čačak

Abstract—The aim of the research is to examine the possibility of applying Weka software in the process of analyzing traffic accidents that occurred in the territory of the Republic of Serbia. During the analysis, the J48 algorithm of the decision tree technique was applied. The database of traffic accidents from 2019, in which a total of 35,956 accidents were recorded, was taken over from the portal of open resources. The analysis of the same in Weka software, and the application of the mentioned algorithm, it came to the results, which showed that a large number of instances of the downloaded database were incorrectly classified. The reason for this is the inadequately standardized database model used in the research. In conclusion, the dominant thesis is that, in order to obtain useful information from databases, they need to have a clear and logical structure, as well as standardized elements for entering the value of attributes.
Keywords-Mining; extraction; traffic accidents; decision tree; J48

INTRODUCTION

The aim of the research is to examine the possibility of applying Weka software in the process of analyzing traffic accidents that occurred in the territory of the Republic of Serbia. During the analysis, the J48 algorithm of the decision tree technique was applied. The database of traffic accidents from 2019, in which a total of 35,956 accidents were recorded, was taken over from the portal of open resources. The analysis of the same in Weka software, and the application of the mentioned algorithm, it came to the results, which showed that a large number of instances of the downloaded database were incorrectly classified. The reason for this is the inadequately standardized database model used in the research. In conclusion, the dominant thesis is that, in order to obtain useful information from databases, they need to have a clear and logical structure, as well as standardized elements for entering the value of attributes.

RESEARCH METHODOLOGY

In this paper, the research was conducted in such a way that the database on traffic accidents, which occurred on the territory of the Republic of Serbia in 2019, was first downloaded from the Internet. The download was made from the internet address https://data.gov.rs/sr/datasets/ (Open data sets, 2019). An experiment was performed on the mentioned database, using Weka software and the j48 decision tree algorithm. The methodology is described in more detail in the following subsections.

RESULTS AND DISCUSSION

The decision tree algorithm generated data for the target attribute of the type of traffic accident, which can be divided into three categories: Summary, Detailed Accuracy By Class and Confusion Matrix. A detailed overview of these data is given in Tables 2, 3 and 4.

Table 2. Summary

Category	Absolute value	Percentage
Correctly Classified Instances	24270	**67.4992 %**
Incorrectly Classified Instances	11686	**32.5008 %**
Kappa statistic	0.2172	/
Mean absolute error	0.2697	/
Root mean squared error	0.3673	/
Relative absolute error	/	**82.5626 %**
Root relative squared error	/	**90.8702 %**
Total Number of Instances	**35956**	/

Table 3. Detailed Accuracy by Class

Class/ Parameter	With material damage	With injured	With dead	Average
TP Rate	0.998	0.193	0.000	**0.675**
FP Rate	0.805	0.008	0.000	**0.487**
Precision	0.652	0.939	?	**?**
Recall	0.998	0.193	0.000	**0.675**
F-Measure	0.789	0.321	?	**?**
MCC	0.351	0.335	?	**?**
ROC Area	0.697	0.695	0.668	**0.696**
PRC Area	0.772	0.612	0.028	**0.700**

It can be seen from Table 2 that the number of correctly classified instances is 24270, which makes about 67.5%, while the number of incorrectly classified instances is 11686 or about 32.5%. The statistics cap is essentially a value that shows the chance of randomly guessing which class something belongs to. Since the value is greater than zero, it means that the classifier is more accurate than random guessing. Mean absolute error is a value used to measure how close predictions or predictions are to possible outcomes. Root mean squared error is a measure of the difference between the values predicted by the model and the values actually observed. It represents a standard experimental deviation of the differences between the predicted and observed values. It is a good measure of accuracy, but only for comparing the prediction errors of different models for a particular variable, and not between variables, because it depends on the scale. It is also called the root square deviation. Relative absolute error is calculated as the quotient of the mean absolute error and the error of the classifier used. It is expressed in percentages the same as Root relative squared error which is the quotient of Root mean squared error and the error of the classifier used. At the very end, the total number of instances of 35956 is shown.

The results shown in Table 3 give us an overview of the classification of instances by 8 parameters, as follows:

1. TP Rate (True positive rate) – shows the rate of true positive values, that is, values that are accurately classified as a particular class.
2. FP Rate (False positive rate) - shows the rate of false positive values, that is, values that are incorrectly classified as a certain class.
3. Precision - the percentage of specimens that are actually classes divided by the total specimens classified as that class.
4. Recall - the share of examples classified as a certain class divided by the actual total amount in that class, equivalent to TP Rate.
5. F-Measure – Combined measure for Precision and Recall, calculated as 2 * Precision * Recall / (Precision + Recall).
6. MCC - used in machine learning as a measure of the quality of binary (double) classifications. It takes into account true and false positive and negative evaluations and is generally considered a balanced measure that can be used even if the classes are substantially different sizes.
7. ROC (Receiver Operating Characteristics) Area - this is one of the most important values that Weka produces. The "optimal" classifier will have ROC range values approaching 1, and 0.5 will be comparable to "random guessing" (similar to the Kappa statistic of 0).
8. PRC Precision-Recall Plot Area – is a significantly more relevant parameter than ROC, when it comes to testing binary classifiers on unbalanced data sets.

From Table 3, based on the obtained values of the parameters TP Rate and FP Rate, it is noticeable that the database used is not largely consistent and correct. For the type of traffic accident with material damage, the J48 algorithm used correctly classified almost all instances, while it incorrectly classified as many as 80% of instances. For the type of traffic accident with injured outcomes, about 19% of instances were correctly classified, while the number of incorrectly classified instances was less than 1%, which indicates that a large number of instances remained unclassified. For cases with dead persons, the algorithm did not classify any instance, and a visual inspection of the database clearly shows that such cases exist. The parameters Precision, Recall, F-Measure and MCC are indirectly obtained from the initial ones, so the previous constance also applies to them. The values of ROC and PRC Area are within optimal limits, or around the middle of the range (0.5-1). The values obtained from Table 3 can be visually viewed in the graph below.

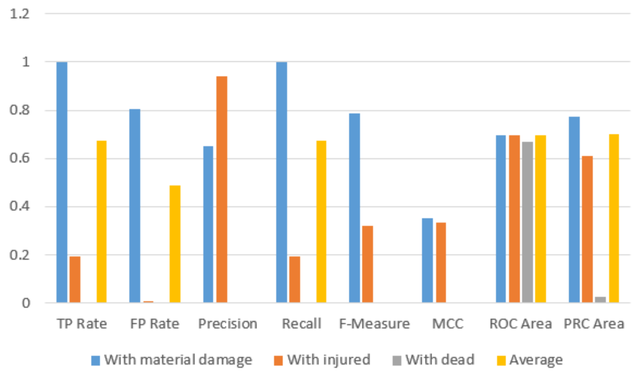

Table 4. Confusion Matrix

a	b	c	Classified as
21601	46	0	**a = With material damage**
11131	2669	0	**b = With injured**
382	127	0	**c = With the deceased**

The Confusion Matrix, gives us an overview of how many instances are correctly and how many are misclassified and in which misclassification it "ended". Table 4 shows that out of the total number of accidents with material damage (21601), a large number (21601) were classified correctly, while 46 were classified as accidents with an injured person. Out of the total number of accidents with injured persons (13800), a higher percentage of as much as 80% is classified as an accident with material damage, while only 20% of the relevant instances are correctly classified. In the species with dead persons, no correct instance out of a total of 507 was classified, where 382 (about 75%) were classified as with material damage, and 127 or about 25% as with an injured person. All of the above further supports the claim that this is a bad model, that is, the database on which the research was conducted.

CONCLUSION

The results obtained by applying the mentioned technique showed that there are certain inconsistencies in the database model itself, which primarily reflected on the results of the TP and FP Rate parameters, and later on the secondary parameters. It previously conditioned that a large number of instances could not be correctly classified, which is best seen in Table 4 - Confusion Matrix.
In addition to the obtained results, this paper also pointed out the importance of data quality and database organization. Namely, in order to obtain useful information from them, they need to have a clear and logical structure, as well as standardized elements for entering attribute values.
The subject of our future work in this area will relate to the use of other Weka software techniques in accident research, with an emphasis on neural networks, which are widely represented in scientific papers and in some way represent the future of data mining. Also, a certain part of the research will be dedicated to the analysis of different types of databases with identical attribute structure.

LITERATURE

Bhargava, Sharma and Mathuria, (2013), *Decision Tree Analysis on J48 Algorithm for Data Mining*, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 6, June 2013, ISSN: 2277 128X
Bhavna Khatri & Hemendra Patidar (2016), *Road Traffic Accidents with Data Mining Techniques*, International Journal of Information Engineering and Technology Vol. 2, Issue 1, 1-6
Han, Kamber, Pei, Jaiwei, Micheline, Jian (2011), *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufmann. ISBN 978-0-12-381479-1.
Hussain, Najoua and Dahan, (2018) *Educational Data Mining and Analysis of Students' Academic Performance Using WEKA*, researchgate, Article February 2018 DOI: 10.11591/ijeecs.v9.i2.pp447-459
Kamiński, B.; Jakubczyk, M.; Szufel, P., (2017), *A framework for sensitivity analysis of decision trees*, Central European Journal of Operations Research. 26 (1): 135–159. doi:10.1007/s10100-017-0479-6. PMC 5767274. PMID 29375266
Official website of Waikato University, https://www.cs.waikato.ac.nz/ml/weka/, accessed in May 2020.
Olutayo V.A and Eludire A.A (2014), *Traffic Accident Analysis Using Decision Trees and Neural Networks, I.J. Information Technology and Computer Science*, 02, 22-28 Published Online January 2014 in MECS (http://www.mecs-press.org/) DOI: 10.5815/ijitcs.2014.02.03.
Open data sets, (2019), retrived in May 2020 from https://data.gov.rs/sr/datasets/
S. Krishnaveni, M. Hemalatha (2011), *A Perspective Analysis of Traffic Accident using Data Mining Techniques*, International Journal of Computer Applications (0975 – 8887) Volume 23– No.7
Srivastava J. (2020), *Web Mining: Accomplishments & Future Directions*, University of Minnesota USA
Tibebe Beshah, Shawndra Hill (2016), *Mining Road Traffic Accident Data to Improve Safety: Role of Road- elated Factors on Accident Severity in Ethiopia*, Addis Ababa University, Ethiopia
Witten, Ian H.; Frank, Eibe; Hall, Mark A.; Pal, Christopher J. (2011), *Data Mining: Practical machine learning tools and techniques*, 3rd Edition, Morgan Kaufmann, San Francisco (CA)