DOI 10.7251/ZRPIM2001378M

ORIGINAL SCIENTIFIC PAPER – ORIGINALAN NAUČNI RAD

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

Aleksandar Mamić, Marija Blagojević, Danijela Milošević

University of Kragujevac, Faculty of Technical Sciences Čačak, Svetog Save 65, 32 102 Čačak, acomamic@gmail.com, marija.blagojevic@ftn.kg.ac.rs, danijela.milosevic@ftn.kg.ac.rs

ABSTRAKT

The aim of the research is to examine the possibility of applying Weka software in the process of analysing 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 increasing use of information technology in all spheres of human activity has contributed to the generation of a large amount of digital data, and thus a large number of databases. In order to use the data collected in this way to further improve a particular area, scientists around the world are conducting various experiments on databases. On that occasion, various techniques of machine learning and artificial intelligence are applied over the collected data, all with the aim of extracting the highest quality information, after further processing of information, quality and usable knowledge is obtained. The field of science that deals with the extraction of information and the generation of new knowledge from large databases is called mining (Srivastava J. 2020).

With the help of the information gathered and the acquired new knowledge, the continuous improvement of various scientific fields, but also the spheres of human activities, is being done. The possibilities of predicting the outcome of certain processes are increasing, and additional, until then unknown relations between entities and database attributes are observed. Machines are also enabled to learn and develop new algorithms and execution procedures, and artificial intelligence is also developed (Witten et.al., 2011).

The subject of this research is the use of Weka software in the process of analyzing traffic accidents. Traffic accidents are by their very nature a very complex field because they pose a risk to all road users, whether they are pedestrians, cyclists or motor vehicle drivers. Almost every traffic accident results in material damage, not a small number and injury to one of the participants. Also, a certain number unfortunately ended in death. For that reason, data from previous years are processed and analyzed, so that in the coming period an adequate strategy can be made in order to reduce traffic accidents, which fully corresponds to the goal of this research - examining the possibility of using Weka software in traffic analysis.

After the introductory part, a brief overview of related research in the subject area is given. Chapter 3 presents the theoretical framework of the research, with an overview of the technique used, while Chapter 4 describes the methodology of the research itself. In Chapter 5, the obtained

results are presented and discussed in the context of the results from related research. The conclusion is drawn in the last sixth chapter.

RELATED RESEARCH

In this paper, various papers were used as literature. In them, an experiment using Weka software using the decision tree technique and the j48 algorithm in the process of analyzing traffic accidents was published. Only the most relevant works will be listed in this chapter.

S. Krishnaveni and M. Hemalatha (2011) analyzed traffic accidents in Hong Kong on a sample of 34575. On that occasion, five different algorithms were used, namely Naive Bayesian, J48, AdaBoostM1, PART, Random Forest with three different aspects, ie information about the accident itself, the victims and the vehicle. Bhavna Khatri and Hemendra Patidar (2016) applied Weka software and the j48 decision tree algorithm. A database with 3293 instances was used. Tibebe Beshah and Shawndra Hill (2016) used 3 Weka software techniques - Decision Tree (J48), Naive Bayes and K-Nearest Neighbors. On that occasion, about 18,000 traffic accidents on the territory of Ethiopia were analyzed. Olutayo V.A and Eludire A.A (2014) conducted research on traffic accidents that occurred on one section of the road between the cities of Ibadan and Lagos during 2002 and 2003. The decision tree and neural networks were used in the experiment.

THEORETICAL FRAMEWORK OF THE RESEARCH

Data mining is the process of finding and defining links and patterns in large sets or databases. On that occasion, methods and techniques are used, which are closely related to machine learning, statistics, databases, and even artificial intelligence. All of the above makes data mining an interdisciplinary field of computer and IT science, which has as its fundamental goal the extraction of information, as higher organizational - qualitative structures than the data itself, and their transformation for some further use. In this way, new and high-quality knowledge is practically obtained from a primarily incomprehensible and unstructured database Srivastava J. (2020).

When mining data, their semi-automatic or automatic analysis is performed in order to extract interesting patterns, such as data groups, unusual records, interdependencies, etc. On that occasion, various techniques are used, such as cluster analysis, anomaly detection, decision tree, association rules, etc. The information extracted using the mentioned techniques can later be used in processes such as machine learning and predictive analysis (Han at al. 2011).

In this paper, a database was used, which was taken from the Open Resources Portal, but was located on a computer (local host) during the experiment. The methodology used in this research puts the domain of content mining. The decision tree technique used (J48 algorithm) belongs to the group of classification methods. Classification is the process of sorting information into categories or classes, so that the data can be more clearly analyzed and understood. The classification and decision tree belong to the group of supervised machine learning. It is characteristic of this learning that the algorithm is given data from which it learns and the desired outputs. It is up to the algorithm to provide the appropriate outputs for the given data.

Decision tree and j48 algorithm

The decision tree is a method or tool to support decision making, which uses a chart in the form of a tree. It is most often used in decision analysis in operational research, in the context of identifying the strategy for the most likely goal achievement. It is also actively used in the field of machine learning. In the structure of the decision tree diagram, each internal node represents a "test" for an attribute, each branch represents the test result, and each leaf node represents a class of label (decision made after calculating all attributes). Root to leaf paths represent classification rules. The decision tree consists of 3 types of nodes (Bhargava et al., 2013), (Kamiński, B. et al., 2017):

1. Decision nodes - often represented by squares
2. Nodes of change - represented by circles
3. End nodes - represented by triangles

The decision tree is considered to be one of the most powerful tools in knowledge discovery and data extraction. It incorporates the technology of researching large and complex volumes of data to discover useful patterns. The decision tree offers a number of advantages when mining data, the most important of which are:

- It is easy to understand by the end user
- Supports a large number of data types: nominal, numeric and text
- It is able to process databases that contain errors as well as missing values
- Extremely high performance
- Can be used on different platforms

The J48 association rule algorithm was used in this paper. This algorithm is essentially used to create a truncated decision tree. Each aspect of the information obtained is divided into smaller subsets on the basis of which decisions are made. Smaller subsets are returned by the algorithm. The partition strategy is stopped if the subset has a similar class in all instances. J48 develops a decision node using expected class ratings. The J48 decision tree can deal with specific characteristics, estimating data on lost or missing attributes. Here the accuracy can be extended by pruning (Bhargava et al., 2013).

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.

The data

The retrieved database contains nine attributes, with different types of data entered. The structure specification of the database used is given in the table below.

Table 1. Database structure specification

Attribute name	Data type	Number of entries (instances)
Accident ID number	integer	35956
Police department	string	35956
Municipality	string	35956
Date and time	integer, char	35956
Longitude	real	35956
Latitude	real	35956
Type of traffic accident	string, char	35956
Type of traffic accident	string, char	35956
Detailed description	string, char	35956

It can be seen from Table 1 that the database, in addition to nine attributes, also contains 4 data types, namely integer, string, char and real. Five attributes are defined by one data type, while four use two data types at the same time to describe the entered values. The downloaded database is formatted as a .xlsx type, that is, an MS Excel file. As a given file type cannot be loaded into the web

software, the database needs to be converted to a CSV file type. After the conversion, the file was uploaded to the software and the decision tree techniques and association rules were applied to it.

Toolkit

WEKA (Waikato Environment for Knowledge Analysis) is open source machine learning software, developed at Waikato University in New Zealand. It contains a set of visualization tools, as well as algorithms for data analysis. Weka supports standard data mining techniques, such as clustering, classification, regression, variable selection, decision tree, association rules, etc. It also provides access to SQL databases and can process the result, obtained from the database query (Hussain et al. 2018), (Official website of Waikato University 2020).

Application of decision tree technique

The experiment was performed by selecting the Explorer user interface type in the Weka software. The converted CSV file is then loaded into the program. The J48 decision tree algorithm was then selected from the classify menu, as shown in Figure 1.

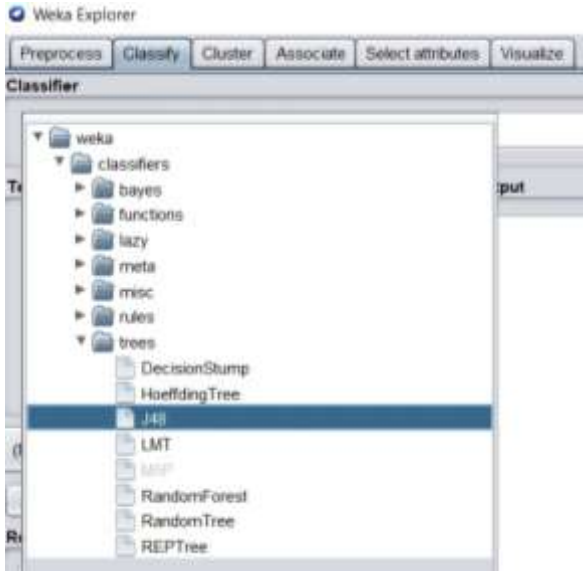

Figure 1. Selected algorithm

When selecting the algorithm, the default parameter values are left. The type of traffic accident was selected as the target attribute. By clicking on the START button, the experiment was realized, and the results are presented in detail in the next chapter.

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.

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.

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**	/

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.

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**

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 2.

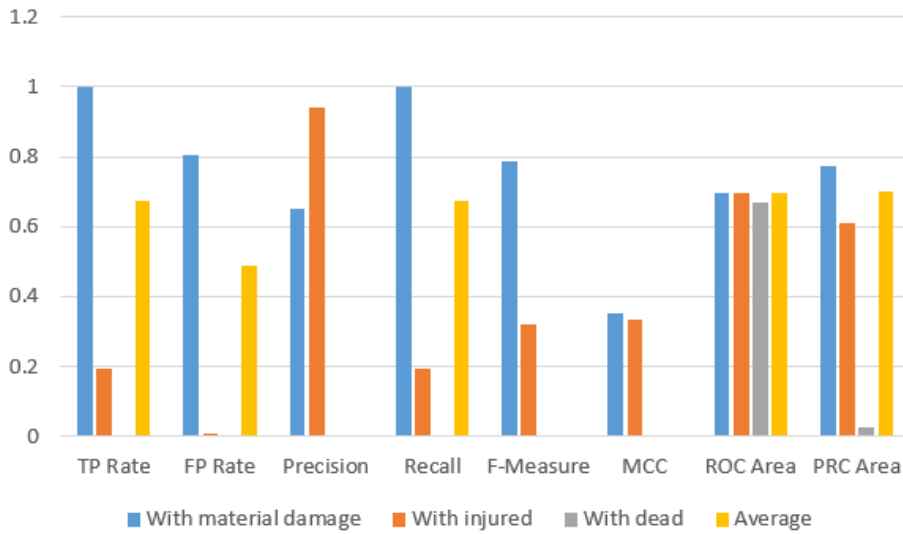

Figure 2. Graph of results from table 3

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.

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**

Comparing results with related research

In the papers listed in Chapter 2, the research was mainly conducted in such a way that the results obtained were compared using different techniques and algorithms. A study that has almost the same structure of results as this paper is Olutayo V.A and Eludire A.A (2014). Therefore, a comparative analysis of this and the mentioned paper will be given in this subchapter.

In the reference work, which was used to compare the results, decision tree and neural network techniques were used. The tables below give a comparative presentation of the obtained results for the decision tree technique, where the data from this paper are in columns marked "obtained", and the data from comparative work in columns marked "reference". Since two stable decision-making algorithms were used in the reference paper, the id3 algorithm was used as a comparison, when the author indicated that it gave the best results.

Table 5. Comparison of results from Summary tables

Category	Obtained	Reference
Correctly Classified Instances	67.4992 %	77.70%
Mean absolute error	0.2697	0.2519
Accuracy rate	0.675	0.777

The results from the Summary tables show that the number of correctly classified instances of the obtained values is about 10% lower than the reference observed ones, which is directly reflected in the Accuracy rate parameter. The mean absolute error is approximately the same.

Table 6. Comparison of results from Tables Detailed Accuracy by Class

Parameter	Obtained	Reference
TP Rate	0.675	0.777
FP Rate	0.487	0.232
Precision	?	0.78
Recall	0.675	0.777
F-Measure	?	0.769
ROC Area	0.696	0.912

Average class values were used to compare the values from the Detailed Accuracy By Class tables. In addition to the TP Rate parameters, which correspond to the correctly classified instances from the previous table, it can be noticed that the FP Rate is more than twice lower in the reference results, which indicates a better database model. Also, the average value for Precision and F-Measure cannot be determined for this work, due to problems in classifying fatal traffic accidents,

which is described at the beginning of the chapter. The ROC Area for the reference values is about 30% higher than the obtained ones, which indicates that an extremely good classifier was used, where the ROC value is close to 1.

Table 7. Comparison of results from Confusion Matrix tables

a	**b**	**c**	**Classified as**
Obtained/ Reference	Obtained/ Reference	Obtained/ Reference	
21601/ 22	46/10	0/0	**a**
11131/ 6	2669/78	0/3	**b**
382/ 2	127/12	0/15	**c**

In the Confusion Matrix table, a better distribution of classified instances is noticeable, where the largest values are located on the main diagonal of the matrix in the reference results. In the obtained results, in addition to having a large number of instances outside the main diagonal of the matrix, there are also classes in which no instance (c) is classified.

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

Krishnaveni, S., Hemalatha, M. (2011), *A Perspective Analysis of Traffic Accident using Data Mining Techniques*, International Journal of Computer Applications (0975 – 8887) Volume 23– No.7

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/

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)