Sinanaj, Luan; Bedalli, Erind; Abazi Bexheti, Lejla

Article

A Classification Model for Predicting Road Accidents Using Web Data

ENTRENOVA - ENTerprise REsearch InNOVAtion

Provided in Cooperation with:
IRENET - Society for Advancing Innovation and Research in Economy, Zagreb

Suggested Citation: Sinanaj, Luan; Bedalli, Erind; Abazi Bexheti, Lejla (2023) : A Classification Model for Predicting Road Accidents Using Web Data, ENTRENOVA - ENTerprise REsearch InNOVAtion, ISSN 2706-4735, IRENET - Society for Advancing Innovation and Research in Economy, Zagreb, Vol. 9, Iss. 1, pp. 60-71,
https://doi.org/10.54820/entrenova-2023-0006

A Classification Model for Predicting Road Accidents Using Web Data

Luan Sinanaj
South East European University, North Macedonia
Erind Bedalli
University of Elbasan, Albania
Lejla Abazi Bexheti
South East European University, North Macedonia

Abstract

The increase in urbanisation and the use of vehicles in recent decades has also led to increased road accidents. The causes of road accidents can be various, including human error, weather conditions or even inadequate road infrastructure. Knowing the causes and areas of road accidents can help prevent them by state institutions taking necessary measures and citizens being informed about the areas of road accidents. The primary purpose of this study is to explore patterns in accident web data in Albania and to construct a classification model using data mining techniques and methods. These techniques have been applied to data obtained from several leading media portals in Albania, including about 30,000 articles from online portals and reports from the state authorities. The constructed classification model is expected to be utilised to predict the accident likelihood according to the locations, weather, and period of the year.

Keywords: data mining; web scraping; classification model; road accident prediction
JEL classification: C3; C6

Paper type: Research article
Received: 20 June 2023
Accepted: 08 September 2023

DOI: 10.54820/entrenova-2023-0006

Introduction

The increase in the use of vehicles in the last decades has also brought an increase in traffic and road accidents. Road accidents occur for a variety of reasons, starting from carelessness or non-compliance with road traffic rules by drivers or even for different reasons related to road infrastructure problems or atmospheric conditions. Part of the purpose of this paper is to study the causes of road accidents or the most problematic places of accidents based on the news data posted online.

Starting from the problems mentioned above about road accidents, it is useful to study the causes of why they happen and to focus on their prevention by the responsible state institutions. Therefore, the analysis of the various causes of accidents can provide important information for state institutions to improve the safety of citizens. This information about road accidents is also necessary for citizens, not only for state institutions. From a review of the literature, there are no publications or applications that study and show the causes or areas of road accidents in the state of Albania. Unlike other states, the state of Albania does not report data on the causes of road accidents but only limits itself to giving the number of accidents during each month of the year (Ministry of Interior, 2023) (INSTAT, 2023). This information is minimal for informing citizens about road accidents or for taking measures to prevent them. For example, unlike the state of Albania, in the publication of a report by the state of Italy (ISTAT, 2023), data on the causes of road accidents and many different statistics about road accidents are shown.

The purpose of this research is to find out the main factors that cause road accidents and to find the possible critical combinations of features that increase the chances of road accidents based on the data obtained from the main online media portals. Knowing the results of this study can help inform citizens to be as careful as possible and to undertake awareness campaigns or prevention of road accidents by relevant state entities such as the state police or civil societies.

The methodology followed for the realisation of this research is divided into four important steps:

(1) building a Python script with Web Scraping techniques for extracting news articles from the online portal and saving them in a CSV file;

(2) building another script in the Python language for processing the items in the previous point by filtering data about traffic accidents and saving them in another CSV file;

(3) checking and manually processing the data in the previous point to having the highest quality dataset for study and the application of Data Mining methods and models;

(4) the application of classification methods such as k-nearest neighbours, naïve Bayes and random forest.

The main research question raised for this case study and based on the purpose of the research are:

- **RQ1**: Which classification model has better performance in predicting the chances of road accidents based on data extracted from the main media portals?

Other supporting research questions intended for an extended version of this work are:

- **RQ2**: In which types of roads are there higher chances of road accidents?
- **RQ3**: Which factor has been mentioned the most for causing road accidents?
- **RQ4**: Is it possible to determine to what gender those who caused the accidents mostly belong?
- **RQ5**: What was the ratio of dead and injured persons in the accidents?

A contribution of this research is the improvement and implementation of two scripts in the Python programming language for extracting articles from online media portals using web scraping techniques, as well as processing them to select articles that talk about road accident data in the state of Albania. Another contribution is the construction of corpora with words of the Albanian language that include all areas of the State of Albania and keywords of accidents. These corpora are used to train the scripts mentioned above to achieve the best possible results. The final output dataset has also been manually checked to increase its quality; this dataset is then applied to Data Mining models to achieve the goal and objectives of this research.

The structure of this paper is organised in this way: section 2 includes a review of the literature related to road accidents; in Section 3, we have the methodology followed for this research describing the details; in Section 4 are the study results and findings. In the end, you will find the conclusions and references for this research.

Literature Review

From a detailed search of scientific publications for the state of Albania, it turned out that there are no studies regarding the prediction of the causes that lead more to accidents or studies on the combination of critical factors that lead more to accidents. The only information that is published about road accidents in the state of Albania is that of the Ministry of the Interior (Ministry of Interior, 2023) and ISTAT (ISTAT, 2023), but they are too few and not detailed.

Scientific research for the prediction of road accidents has been done since the earliest times; the author (Kononov, 2002)for the modelling of the prediction of road accidents and the diagnosis of the causality of accidents has used the pattern recognition algorithm and direct diagnostic methods, alternatively, in the research of the author (Larsen, 2004)who used the multidisciplinary approach and provided a precise knowledge about the factors that lead to road accidents.

In the research of the authors (Lnenicka, Hovad, Komarkova, & Pasler, 2016), an application was created with web scraping techniques for extracting information from the servers of online portals about the crimes that occurred by presenting them on a map and tabular form according to the areas with high crime rate.

In recent years, numerous publications have been made about road accidents in different countries.

To achieve high accuracy of the results, the authors (Yan & Shen, 2022) in their study use a hybrid model that integrates random forest (RF) and Bayesian optimisation (BO), where RF has been used as a basic predictive model and BO to adjust the RF parameters. Experimental results showed that this model achieves higher accuracy than conventional algorithms.

In the scientific paper of the authors (Chen & Chen, 2020), the empirical results for modelling the severity of road accidents showed that the accuracy and specificity of RF are higher than those of LR (logistic regression) and CART (classification and regression tree). RF proved to be the most effective forecasting model among the three models, and this is also consistent with the results of previous studies.

In the next scientific paper of the authors (Elyassami, Hamid, & Habuza, 2020), three machine learning algorithms were used against the Decision Tree, Random Forest and gradient-boosted tree on vehicle crash data provided by the Maryland State Police. The findings of this study showed that ignoring traffic signals and stop signs, road design problems, poor visibility and bad weather conditions are the most important factors in the predictive model of road accidents.

In the authors' study (Siddik, 2021) to predict road accident deaths in Bangladesh, four classification models as Decision Tree, K-Nearest Neighbors (KNN), Naïve Bayes

and Logistic Regression were applied. The model was tested on 1237 road accident data taken from a newspaper. Another study in Bangladesh was done by the authors (Biswas, Mia, & Majumder, 2019), and Random Forest Regression was used to predict the number of road accidents and their victims. The results showed that random forest regression is good enough to predict and apply in this context.
In the state of India, the authors (G & R, 2023), for the prediction of road accidents, are based on four factors: the type of collision, the type of road, the location, and the weather. They used a machine learning model with a random forest regressor and a decision tree regressor, and the analysis results proved that the random forest regressor (RFR) model performs better than the tree regressor model. Decision (DTR).

Methodology

This section explains the methodology followed for this research, the stages that have been passed, the method of data collection and the Data Mining models that have been applied to these data. The methodology followed for the realisation of this project is divided into two main important parts: data preprocessing and classification models based respectively on Nearest Neighbor Classifiers, Naïve Bayes and Random Forests.

Data Preprocessing

One of the most important parts of this study has been the extraction and preparation of data, which were then applied to data mining models for the purposes and objectives of our study. It is worth noting that the difficulties during this study have been many since the scientific papers and studies related to the processing of the text of the Albanian language are very limited.

During the preprocessing phase, three important steps were followed to provide the dataset for the study:

1) **Step 1**: Several online news portals were selected to extract daily posted articles. Web Scraping techniques were used to build Python scripts with the BeautifulSoup library (Richardson, n.d.). The BeautifulSoup library is a Python-based module used to simplify programs to extract data from HTML or XML pages. During this process, about 29288 articles were extracted and saved in a CSV file. In the CSV output file of the articles, the extracted information is divided into columns and contains the following: Number, Date, Title of the article and the body of the article.

2) **Step 2**: With the successful completion of step 1, we moved on to the next important step, which was to select all the articles that talked about the accidents that happened in the state of Albania. For the realisation of this process, a Python script was built again, and libraries such as pandas, numpy and nltk (nltk.org, 2022)were used for processing text data. The construction of the script had its difficulties and challenges since there was a lack of libraries for the processing of Albanian text. It is important to emphasise that at this stage, two corpora of data have been built and used by our script: one that contains words of the Albanian language related to accidents, injury or death (CORPUS1) and another very important corpus that contains all areas and localities of the state of Albania (CORPUS2). By running the script on the 29,288 articles of step 1, it was possible to extract about 1,200 articles that talked about road accidents or other accidents. In the CSV output file of 1200 items, the information is divided into columns containing the following: Number, Date, Area, Address, Type of Road, Injured-Dead, Weight-Accident, Cause and Gender. The type of road where the accident occurred is composed of four categories: rural, interurban, urban and highway. Injured-Dead contains data on whether the accident that occurred had a

dead or injured person. The weight of the accident is divided into degrees according to the risk from 1 (light risk) to 8 (high risk 8). The cause of the accident, when it was possible to find it, is divided into the following categories: humidity, carelessness, non-compliance with the rules, red traffic light, drunken state, wrong overtaking, without license, speed and humidity. For the last data, gender, when possible, we also determined the gender of the person who caused the accident. This was a very important step during the preprocessing phase of the data for which Data Mining models will be applied. In order to have the best quality data for our study, we went to the next step of manual control of the data (Sinanaj & Bexheti, 2023).

3) **Step 3**: after the successful completion of step 2 and in order to have the final data as qualitative as possible, manual control of about 1200 articles was done and from this selection process, we managed to have 829 articles in the final output dataset. During this phase of manual control, an important selection was made that had not managed to be caught and eliminated from the script execution of the second phase. During the manual process, redundant articles talking about the same event and articles talking about different accidents unrelated to road accidents were eliminated. This was one of the most demanding and tiring processes, but it was very important to have the final dataset of the highest quality data.

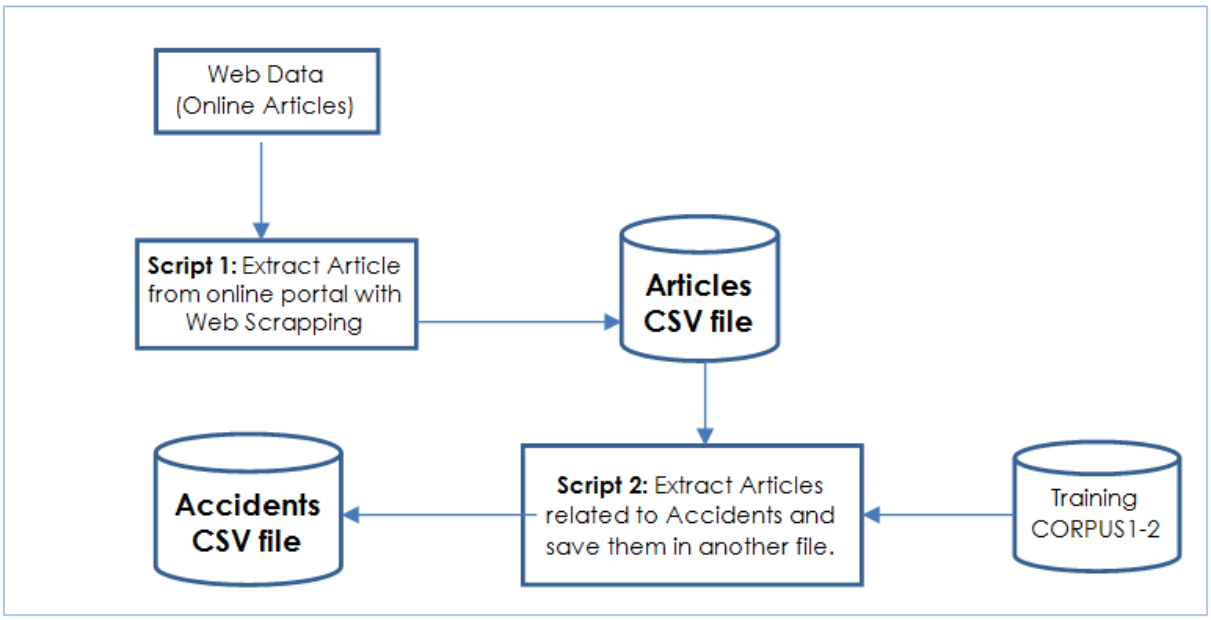
Figure 1 below shows schematically the processes of the steps described above to obtain the final dataset of data related to accidents in the state of Albania.

Script 1 is the execution of the program that reads the articles of the online portal and saves them in the CSV Articles file.

Then Script 2 reads all the articles from the Articles file, filters those that talk about accidents with their data, and saves them in another CSV Accidents file. Script 2 was trained with Albanian language data from the CORPUS1-2 file that we explained above.

Figure 1

Application Model for Data Preprocessing



Source: Authors' work

After preprocessing, the data will contain eight fields: Date, Region, Location, Type of Road, Cause, Gender and Gravity. More detailed descriptions about these fields are presented in Table 1. These data served as the input for the classification models.

Table 1
The variables after the preprocessing stage.

Variables	Description	Type of Variables
Date	*The date of the publication of the accident*	Independent
Region	*The region of the accident*	Independent
Location	*Possible location or address of the accident*	Independent
Type of Road	*Interurban, Urban, Rural, Highway*	Independent
Cause	*None, moisture, Non-compliance, red light violation, drunkenness, velocity, moisture Carelessness, No license, wrong overtaking Velocity, distance*	Independent
Gender	*None, M, F*	Independent
Gravity	*1 (light risk) to 8 (high risk 8)*	Dependent

Classification model

Supervised learning is an important domain of data mining that fundamentally infers a mapping from a set of input variables into an output variable. This mapping is utilised to predict the output variable for new non-categorised data based on their input variables. Two important approaches among the supervised learning algorithms are the classification and regression algorithms. Classification constructs a model that predicts discrete labels for new data based on a known dataset (i.e., training set), while regression aims to predict continuous values of the output variable (Sen, 2020). Classification is used in scenarios where the output variable naturally belongs to a discrete domain, such as spam/non-spam positive/negative (in diagnosis), while regression is used when the output variable is continuous, like temperature, income, and duration. Furthermore, regression is of great relevance even in scenarios where the output variable is discrete and ordinal, such as the quality rank of a product and the severity degree of an accident. Classification and regression are versatile data mining tools used not only on tabular data but also on time series data, ordered data like DNA microarrays, and graph data (Grabocka, 2013).

Several classification algorithms are developed and successfully utilised in practice. In this work, three classification methodologies are used: nearest neighbour classifier, naïve Bayes classifier and random forests. The training set that is obtained after the preprocessing stage is randomly split into training and testing subsets. The results of the classification model are compared to the original training set in order to evaluate the performance. The performance metrics that we are using are precision, recall and accuracy.

Nearest neighbor classifier

The classical K-Nearest Neighbor (KNN) classifier is described as a lazy classification algorithm because it does not utilise the training dataset in order to construct a model. The classification procedure starts only when new data arrives. The classical K-NN algorithm relies on the nearest training samples in the feature space. The training samples are arrays in a feature space of several dimensions. The principle of this classification technique is the majority vote among the K (a constant) nearest neighbours. Typically, the value of K may be tuned and depending on the scenario, its optimal value may vary (Abu Alfeilat, 2019).

The inputs of the K-NN algorithm may be summarised as:

a. $X = \{x_1, x_2, \dots, x_n\}$ the training sample
b. The distance metrics

c. The value y to be classified
d. The value of K *(it may be also be assigned by the algorithm)*

The algorithm may be described by the following pseudo-code (Peterson, 2009):

$FOR(i = 1;\ i \leq n;\ i = i + 1)\{$
Evaluate distance from y to x_i
$IF(i \leq K)\{$
Add the point to the list of K nearest neighbors
$\}ELSE\ IF\ (x_i\ is\ closer\ to\ y\ than\ the\ farthest\ neighbor\ in\ the\ list\ )\{$
Remove the farthest neighbor in the list of K nearest neighbors.
Add x_i *to the list of K nearest neighbors.*
$\}$
$\}$
Find the majority class in the list of K nearest neighbors and classify y.

Naïve Bayes classifier

Naïve Bayes classifier offers a probabilistic framework for solving classification problems. The posterior probabilities for the input values of a new entry are evaluated using the Bayes theorem, and the maximising value is retained as the label forhe new entry (Yang, 2018). According to the Bayes rule, probabilities that an instance belonging to class C_i if its attribute values are $v_1, v_2, \dots, v_n$ are calculated:

$$P(c = C_i \mid a_1 = v_1, a_2 = v_2, \dots, a_n = v_n) = \frac{P(c = C_1)\, P(a_1 = v_1, a_2 = v_2, \dots, a_n = v_n \mid c = C_1)}{P(a_1 = v_1, a_2 = v_2, \dots, a_n = v_n)}$$

The denominator in the above formula serves merely for normalisation, and obviously, it does not depend on the class of the instance. Although the Naïve Bayes classifier represents a fairly simple classification model, it performs reasonably well in practice, making it a significant tool in classification problems.

Ensemble methods – Random forests

The central idea of ensemble methods is to apply several classification models in order to generate several classification results. These results are then intertwined into a single classification result. Several approaches may be used in the intertwining phase, among which the majority vote is one of the fundamental approaches. The ensemble methods are used successfully in both classification and clustering problems, significantly increasing accuracy and stability b at a higher computational cost (Bedalli, 2016).

The ensemble approaches are mainly categorised into three categories: bagging ensembles, stacking ensembles and boosting ensembles. In our approach, the bagging ensemble method of random forests is used. This method merges the results of several decision trees into a single result. This method has been widely adopted due to its ease of use and capability of dealing with both classification and regression problems (Ren, 2016).

For each of the classification models mentioned above, several parameters can be controlled. Table 2 summarises the main parameters and their respective values for each classification model.

Table 2
The main parameters of the deployed classification models

Classification Model	Parameter 1 Name & Value	Parameter 2 Name & Value	Parameter 3 Name & Value
K-NN	K = 5	Distance = L_2	Weight = Uniform
Naïve Bayes	Alpha = 1.0	Fit_prior = True	Class_prior = None
Random Forest	N_estimators = 40	Criterion = gini	Min_samples = 2

Source: Authors' work

Results

Applying the steps described in the Methodology section, the preprocessing stage is carried out first. The outcome of this stage is a table containing processed data of each accident, including the date and time, the region, the location, the type of the road, the primary cause, the gender of the driver and the gravity scale. An overview of this table is provided in Figure 2.

Then, in the second phase, three classification models are applied to the preprocessed data: k-nearest neighbour classification, naïve Bayes and random forests. In each case, the dataset is split into training and testing data and a classification model is constructed based on the training data. The target attribute in each case is the gravity of the accident.

In each case, the performance of the classification model is evaluated, and the significant metrics of accuracy, precision, and recall are tracked. A summary of these metrics is provided in Table 1.

Figure 2
Overview of pre-processing results

	Nr	Date	Region	Location	Type of Road	Cause	Gender	Gravity
0	1	31 12 2021 16:15	durres	rrugor Mamin	interurban	NaN	M	4.0
1	2	30 12 2021 19:34	puke	pukë	urban	moisture	M	5.0
2	3	30 12 2021 18:47	tirane	Rrugën e Arb	interurban	carelessness	M	4.0
3	4	30 12 2021 10:36	thumane	Thumanë	urban	carelessness	M	3.0
4	5	29 12 2021 22:27	maminas	Durrës Mamin	interurban	NaN	M	3.0
...	...	...	...	...	...	...	...	...
824	825	03 01 2021 16:11	milot	autostradën	Highway	velocity	M	5.0
825	826	03 01 2021 10:04	kelcyre	rrugor Këlcy	rural	NaN	M	6.0
826	827	03 01 2021 08:08	Lezhe Milot	Lezhë Milot,	interurban	carelessness	M	6.0
827	828	02 01 2021 14:10	durres	lagja nr 7	urban	NaN	NaN	3.0
828	829	01 01 2021 17:04	librazhd	rrugë kombët	interurban	Non compliance	M	5.0

829 rows × 8 columns

Source: Authors' work

Table 3
Performance evaluation of classification models

Classification Model	Accuracy	Precision	Recall
K-NN	0.51	0.58	0.49
Naïve Bayes	0.54	0.52	0.59
Random Forest	0.66	0.69	0.63

For each classification model, the train-test split procedure was performed randomly with a ratio of 0.75 train – 0.25 test, and it was repeated 100 times. The respective values of accuracy, precision, and recall depicted in the table represent the averages of these 100 executions.

Discussion

The scrapping procedure was handled successfully with the assistance of the BeautifulSoup library, but there were significant challenges in the preprocessing stage as there were no libraries that would properly process the Albanian language. In these conditions, the solution was the usage of common libraries devised for the English language associated with specific scripts for dealing with Albanian keywords.

Constructing a classification model for predicting the likelihood of road accidents according to web scrapping data is naturally a significant challenge. Thus, the performance metrics achieved by the applied classification algorithms are considered decent.

Among the used classification models, the random forest method demonstrated better performance metrics compared to K-nearest neighbours and naïve Bayes. On the other hand, the cons of random forests are related to their higher computational complexity, but this aspect was beyond the scope of our study.

Conclusion

The work presented in this paper was about the construction of a multi-staged model for predicting the likelihood of road accidents according to web data. Although there have been similar works in various countries, this is the first one for the Albanian language. Several classification models were tested in this framework, and the performance evaluation metrics showed that random forest classifiers were the most convenient.

The outcome of this work may assist in the prevention of accidents by knowing in advance the areas of accidents with the highest risk, thus providing important information for both governmental institutions and citizens in order to improve security at the operational and strategic levels.

References

1. Abu Alfeilat, H. A. (2019). Effects of distance measure choice on k-nearest neighbor classifier performance: a review. *Big data, 7*(4), 221-248.
2. Bedalli, M. A. (2016). A heterogeneous cluster ensemble model for improving the stability of fuzzy cluster analysis. *Procedia Computer Science, 102*, 129-136.
3. Biswas, A. A., Mia, M. J., & Majumder, A. (2019). Forecasting the Number of Road Accidents and Casualties using Random Forest Regression in the Context of Bangladesh,. *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)* (pp. 1-5). Kanpur, India: IEEE. doi:10.1109/ICCCNT45670.2019.8944500
4. Chen, M.-M., & Chen, M.-C. (2020). Modeling Road Accident Severity with Comparisons of Logistic Regression, Decision Tree and Random Forest. *Information*, pp. 11(5), 270. doi:https://doi.org/10.3390/info11050270
5. Elyassami, S., Hamid, Y., & Habuza, T. (2020). Road Crashes Analysis and Prediction using Gradient Boosted and Random Forest Trees. *2020 6th IEEE Congress on Information Science and Technology (CiSt)* (pp. 520-525). 2020 6th IEEE Congress on Information Science and Technology (CiSt): IEEE. doi:10.1109/CiSt49399.2021.9357298
6. G, M., & R, R. H. (2023). Prediction of Road Accidents in the Different States of India using Machine Learning Algorithms. *2023 IEEE International Conference on Integrated Circuits and Communication Systems (ICICACS)* (pp. 1-6). Raichur, India: IEEE. doi:10.1109/ICICACS57338.2023.1009951
7. Grabocka, J. B.-T. (2013). Efficient classification of long time-series. *ICT Innovations 2012: Secure and Intelligent Systems* (pp. 47-57). Springer
8. INSTAT. (2023). *Transporti, Aksidentet dhe Karakteristikat e Mjeteve Rrugore*. (Instituti i Statistikave - Tiranë) Retrieved May 2022, from http://www.instat.gov.al/al/temat/industria-tregtia-dhe-sh%C3%ABrbimet/transporti-aksidentet-dhe-karakteristikat-e-mjeteve-rrugor
9. ISTAT. (2023, January). *INCIDENTI STRADALI IN ITALIA*. (Istat – Istituto nazionale di statistica) Retrieved May 2022, from https://www.istat.it/it/archivio/25982
10. Kononov, J. (2002). Road accident prediction modeling and diagnostics of accident causality: A comprehensive methodology. Denver: University of Colorado at Denver
11. Larsen, L. (2004). Methods of multidisciplinary in-depth analyses of road traffic accidents. *Journal of Hazardous Materials, 111*(1-3), 115-122
12. Lnenicka, M., Hovad, J., Komarkova, J., & Pasler, M. (2016). A proposal of web data mining application for mapping crime areas in the Czech Republic. *10th International Joint Conference on Software Technologies (ICSOFT)* (pp. 1-6). Colmar, France: IEEE
13. Ministry of Interior. (2023, January). *Raporti Mujor (Monthly Report)*. (Ministry of Interior in Albania) Retrieved May 2022, from https://mb.gov.al/en/raporti-mujor
14. nltk.org. (2022, June). *Natural Language Toolkit*. (https://www.nltk.org/) Retrieved May 2022, from https://www.nltk.or
15. Peterson, L. (2009). K-nearest neighbor. *Scholarpedia, 4*(2), p. 1883
16. Ren, Y. Z. (2016). Ensemble classification and regression-recent developments, applications and future directions. *IEEE Computational intelligence magazine, 11*(1), pp. 41-53
17. Richardson, L. (n.d.). *Beautiful Soup*. (Crummy) Retrieved March 2023, from https://www.crummy.com/software/BeautifulSoup
18. Sen, P. C. (2020). Supervised classification algorithms in machine learning: A survey and review. *Emerging Technology in Modelling and Graphics: Proceedings of IEM Graph 2018* (pp. 99-111). Singapore: Springer
19. Siddik, M. A. (2021). Predicting the Death of Road Accidents in Bangladesh Using Machine Learning Algorithms. *ICACDS 2021: Advances in Computing and Data Sciences* (pp. 160–171). Springer, Cham. doi:https://doi.org/10.1007/978-3-030-88244-0_1
20. Sinanaj, L., & Bexheti, L. A. (2023). Predicting Road Accidents with Web Scraping and Machine Learning Techniques. *International Scientific Conference on Business and Economics*. Tetovo, North Macedonia

21. Yan, M., & Shen, Y. (2022). Traffic Accident Severity Prediction Based on Random Forest. *Sustainability*, pp. 14(3), 1729. doi:https://doi.org/10.3390/su1403172
22. Yang, F. J. (2018). An implementation of naive Bayes classifier. . *2018 International conference on computational science and computational intelligence (CSCI)* (pp. 301-306). IEEE.

About the authors

Luan Sinanaj is a PhD student at South East European University, North Macedonia, working as a full-time lecturer in the Department of Information Technology at "Aleksandër Moisiu" University of Durrës. He completed his B.Sc. studies in the "Informatics" program at the University of Pisa, Italy. He then pursued a professional master's degree in "Internet Technologies" at the same university. Later, he graduated with a scientific master's degree in the "Economic Informatics" program at the European University of Tirana. PhD(c) Sinanaj is the co-author of the book "Basics of Programming in JAVA" and the author or co-author of several publications at national and international conferences. Also, his work and research experiences and interests are in Programming, Data Mining and Artificial Intelligence. The author can be contacted at **ls30441@seeu.edu.mk**

Erind Bedalli is a lecturer in the Department of Informatics at the University of Elbasan. He has received his B.Sc. degree in Computer Engineering from Hacettepe University, Ankara, and his M.Sc. degree in Informatics from the University of Tirana. He completed his doctoral studies in fuzzy logic and exploratory data analysis at the University of Tirana in 2014. His research experience and interests include fuzzy logic, data mining, artificial intelligence, and large-scale computing. The author can be contacted at **erind.bedalli@uniel.edu.al**

Lejla Abazi Bexheti is an Associate Professor at the Faculty of Contemporary Sciences and Technologies at South East European University in Macedonia. She holds a PhD in Computer Science and has been part of the CST teaching staff since 2002. Her main research activity is in Learning Systems and eLearning, and she has been involved in many international projects and research activities in this area. At SEE University, she was involved in resolving the Learning Management System issue. Currently, she is Pro-rector for academic issues at SEEU. The author can be contacted at **l.abazi@seeu.edu.mk**