Data science applied to extract insights from data - weather data influence on traffic accidents

Milana Novkovic, Marko Arsenovic, Srdjan Sladojevic, Andras Anderla, Darko Stefanovic
University of Novi Sad, Faculty of Technical Sciences
Trg Dositeja Obradovica 6, 21000 Novi Sad
Serbia
mnovkovic@uns.ac.rs, arsenovic@uns.ac.rs, sladojevic@uns.ac.rs, andras@uns.ac.rs, darkoste@uns.ac.rs

Abstract— **An increasingly growing amount of publicly available data opens countless possibilities for research and analysis in the field of data science. Data mining enables us to summarize large amounts of data in a manner that allows discovery of hidden patterns in collected data. Based on those patterns it is possible to generate predictive models and improve decision-making processes. This study illustrates how data mining techniques are applied to a dataset consisting of meteorological and traffic accident data for the purpose of determining the correlation between different weather factors and traffic accident occurrences. We rely on publicly available meteorological data and web scrapping as a form of gathering large amount of traffic accident data in a period of 15 years. Rainfall, daily changes in temperature, cloudiness, and humidity have been found to be in correlation with the occurrences of traffic accidents.**

Keywords-data mining; traffic accidents; weather data influence; machine learning algorithms; predictive models

I. INTRODUCTION

Prediction of future weather conditions has significant impact on social and economic areas of human life. By gathering the weather data, meteorology opens the possibility of analyzing significant patterns in large amounts of data. Over 150 thousand lives annually are claimed due to the climate changes in temperature and precipitation trends [1].

Climate changes also affect traffic flow. By changing external conditions in which transport takes place and which affect the health or concentration of driver's unfavorable meteorological conditions can lead to traffic accidents, injuries, and death. World Health Organization indicates that the number of road traffic deaths is troublesome and has plateaued at 1.25 million per year [2].

This study tries to determine the correlation between weather conditions and traffic accident occurrences by analyzing collected data. Data analysis is one of the activities of data science focused on obtaining important information from collected data.

Data science based projects rely on a set of activities. The first activity is acquisition of data, which focuses on gathering the necessary data by using different techniques as interviews, questionnaires, observations, focus groups, documents, records and case studies [3]. Unceasing development of computer science and information has led to new ways of collection and presentation of digital data. This study relies on the growing concept of open data. Open data is publicly available data on the Internet that can be downloaded, copied, and analyzed without any barriers and with maximum freedom to use and reuse [4]. Weather and traffic accident data for this study was primarily located on the web. Due to the lack of publicly available data present in this region, acquisition of data concerning traffic accidents was done by using web scraping tool.

Web scraping represents an automated way of extracting information from web pages. It transforms website content into structured data that can be stored in different forms [5]. With this technique, we collect only specified information from given web pages.

Acquisition of data is then being followed by preparation of data, which refers to the cleaning and transforming the data through activities such as removing invalid records or duplicate entries, removing the inconsistencies in data or combining two different datasets into a single table [3].

Analysis is the third activity focused on finding connections and patterns in collected data. This study uses data mining techniques and tools to discover hidden patterns in the dataset and determine the correlation between weather conditions and traffic accident occurrences. Data mining is an automatic or a semiautomatic process of discovering significant patterns in data, which enable predictions on the basis of new data [6]. There are different approaches in discovering patterns in the dataset, and one of them is machine learning. Machine learning represents a sub-field of computer science which relates to the design and development of algorithms that can learn from data without being programmed and make predictions based on new data [7].

Final stage after data analysis is a presentation of the findings concerning most significant patterns and answers to previously defined questions [3]. Most significant findings of this study are presented in the results section of this paper. The rest of this paper is organized as follows: Section II deals with the related published work. Section III describes the applied approach, dataset description and gathering process. Section IV provides achieved results and related discussion. Section V presents our conclusions.

II. RELATED WORK

Relationship between different weather conditions and many variables of traffic safety has been explored by several papers. Van den Bossche, Wets and Brijs discuss the high impact of weather conditions like rain and thunderstorm on the decrease in traffic safety [8]. By using multiple regression analysis on a dataset consisting of monthly exposure, weather, law and economic conditions data they determined high relationship between weather conditions and number of people seriously injured or killed in traffic. Brodsky and Hakkert have also associated substantial risk of a road accident injuries with occasional rains [9]. Based on the data collected from traffic injury accident file in Israel and U.S. fatal accident file they try to estimate the risk of a traffic accident during the rainy season by using regression analysis. For their estimations, they use two datasets consisting of daily weather, accident, and travel data. The study puts focus on the fact that rain makes the surface more slippery at curves and during certain maneuvers, which is being caused by the reduction in friction between the road surface and vehicle's tires. The assessments made in our study were carried over dataset consisting of daily weather and accident data during several years. By matching daily weather data with traffic accident data, we try to associate different weather factors including rainfall with traffic accidents.

Poisson regression and a linear probability are used by Leard and Roth in [10] to determine the positive relationship between temperature and fatalities. They associated warmer temperatures with significant increases in fatal accidents. The study found that high temperatures are more likely to lead to motor vehicle deaths and that temperatures above 80 degrees Fahrenheit are associated with a 9.5% increase in fatality rates. The study also shows the elevated risk of an accident with property damage or fatality when temperatures are below freezing and rainfall or snowfall are present. Through examining the effects of temperature on traffic incidents Yannis and Karlaftis established that higher temperatures lead to an increase in the number of accidents by associating rise in temperature with larger pedestrian traffic [11]. Statistical data show that one-quarter of all crashes on U.S. public roads are weather-related [12]. As shown in the experimental part of this paper our study also found a correlation between daily temperature, rainfall and traffic accident occurrences.

Andreescu and Frost in [13] through calculation of correlation coefficients and regression equations found that rain, snow, and temperature are in significant correlation with traffic accidents in Montreal. Meteorological data as daily values of temperature, humidity, precipitation, cloud cover, wind direction and speed were gathered from monthly and annual publicly available meteorological publications. We also use meteorological publications consisting of weather data to form dataset consisting of daily data for 6 main cities in Serbia. Mentioned daily values were tested to determine correlation with traffic accident occurrences.

Chang and Chen used accident data in the period from 2001. to 2002. to conduct data mining research to analyze the frequency of freeway accidents on National Freeway 1 in Taiwan [14]. Results of the analysis showed that the risk of potential accidents rises with the number of vehicles and present rainy conditions that are affecting the visibility and vehicle maneuvering due to wet pavement conditions. In [15] Krishnaveni and Hemalatha apply different data mining techniques based on classification models for predicting the severity of possible injuries that occur during traffic accidents. They applied Naive Bayes, AdaBoostM1, Partial Decision Three, J48 and Random Forest Classifier. In our study, we have also used Random Forest ensemble learning method for classification and regression which gave the best performance with high accuracy as in previously mentioned research.

III. MATERIALS AND METHODS

The complete procedure of developing the predictive model is described further in detail. The entire process is divided into several stages, including data collection, attribute and algorithm selection, and finally training and validating the models.

A. Dataset collection

Daily weather information in the time span from 2000. to 2015. is gathered from publicly available Climatological yearbooks published by the Republic Hydrometeorological Service of Serbia [16]. Maximum and minimum air temperature are read in 21 hours and recorded for the day. Daily precipitation relates to a period of 24 hours: 7 hours the previous day to the 7 hours of the current day when they register. The height of snow cover is measured at 7 pm. The yearbook contains measurements of 6 stations: Belgrade, Novi Sad, Vranje, Zlatibor, Loznica and Niš and is available in PDF document format. Daily data referring to air pressure, air temperature, relative humidity, water vapor saturation, wind direction, wind speed, insolation, cloudiness, precipitation and snow cover depth are presented in the form of tables. Table data was extracted from PDF documents by using a PDF document parser written in C# language. Its purpose was to extract all data from PDF tables and store it in Microsoft SQL Server Database table.

Accident data are obtained by using a Web Scraper, Chrome browser extension which provides data extraction from specified web pages [17]. Data extraction is based on sitemaps, a way of navigating the site and scrapping site elements previously marked by the user for scraping. A selected site for scrapping represents a site that collects news from different sources and enables advanced news search based on entering keywords [18]. Advanced news search available on this site is used to navigate to relevant news, which refer to traffic accidents on location of Belgrade, Novi Sad, Vranje, Zlatibor, Loznica and Niš. Keywords used for advanced news search are presented in Table 1. Defined keywords are combined with city names in order to target traffic accidents related to specified cities.

TABLE I. KEYWORDS USED FOR NEWS SEARCH

No.	Keywords	Keywords
1.	*lancani sudar*	car pileup
2.	*poginulo u sudaru*	killed in crash

No.	Keywords	Keywords
3.	*pregazio pesaka*	run over pedestrian
4.	*saobracajka*	car crash
5.	*saobracajna nesreca*	traffic accident
6.	*saobracajne nesrece*	traffic accidents
7.	*sudar automobila*	car collision
8.	*udes*	accident

Data concerning news title, text, source, date and location were scraped and exported in the form of comma separated values (CSV) documents. Obtained CSV documents are stored in database table after being parsed by C# parser created for this purpose. The dataset was filtered from duplicate news and manually exported in the form of CSV file. The document was then imported to Weka workbench, a collection of machine learning algorithms that can be applied on datasets [19]. Fig. 1 shows the complete process of gathering the dataset including model development.

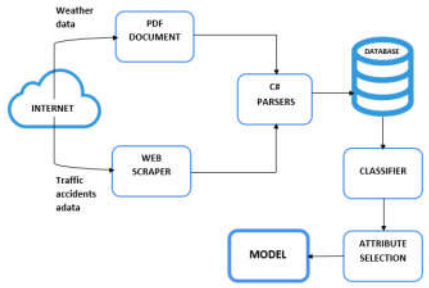

Figure 1. Process of model developemnt

B. *Attribute Selection*

Gathered dataset consists of values from 40 different attributes which are presented in Table 2.

TABLE II. ATTRIBUTES OF GATHERED DATASET

	Attribute	Description
1.	City	City
2.	AirPresure07, AirPresure14, AirPresure21	Air pressure in three terms: 7, 14 and 21 hours
3.	AirPresureAver	Average air pressure
4.	MaxTemp and Min Temp	Extreme maximal and minimal daily air temperature
5.	Amp	Value amplitude – air temperature
6.	Min	Air temperature in degrees Celsius 5cm
7.	TempPeriod07, TempPeriod14, TempPeriod21	Temperature in three terms: 7, 14 and 21 hours
8.	TempAver	Average temperature
9.	HumidPeriod07, HumidPeriod14, HumidPeriod21	Relative humidity in three terms: 7, 14 and 21 hours
10.	AirHumidityAver	Average relative humidity
11.	VoltageStream07, VoltageStream14, VoltageStream21	Water vapor saturation in three terms: 7, 14 and 21 hours
12.	VoltageStreamAver	Average voltage stream
13.	WindDirVelocity07A, WindDirVelocity14A, WindDirVelocity21A	Wind direction in three terms: 7, 14 and 21 hours
14.	WindDirVelocity07B, WindDirVelocity14B, WindDirVelocity21B	Wind speed in three terms: 7, 14 and 21 hours
15.	WindDirVelAver	Average wind speed
16.	Insolation	Insolation
17.	Cloudiness07, Cloudiness14, Cloudiness21	Cloudiness in three terms: 7, 14 and 21 hours
18.	CloudinessAver	Average cloudiness
19.	Rainfall	Amount of rainfall in millimeters
20.	SnowM	Overall snow cover depth in centimeters
21.	SnowN	New snow cover depth in centimeters in the evening
24.	Occur	Indicates did traffic accident occur on a specified date
25.	DateID	Date identification
26.	NewsID	Traffic accident news identification
27.	Day	Date

Weka, data mining software, is used for the process of the attribute selection, training and evaluating the models [20]. All the further stages described are also replicated in Rapidminer, Data Science tool [21]. Due to the very small variation in the results, same attributes and algorithms proved the best performance and accuracy, process of the replication will not be discussed separately in this paper.

Attribute selection represents a process of searching the most influential subset of the attributes in the dataset. The attribute evaluator represents the method for assessing the subset of the attributes. Weka provides the tools for the attribute selection. The process in Weka is divided in two parts: attribute evaluator and search method. Search method is used for finding the probable subsets of the attributes.

This stage of development, attribute selection, is initiated by manually removing the unnecessary attributes that are known that are not significant for further data analysis, such as the DateID, Day, NewsID.

After that, attribute selection techniques described in this section are exploited on the dataset described in order to remove the redundant attributes with the goal of reducing the overfitting and improving the accuracy. This is achieved by the cross-validating the rankings and removing the attributes which influenced the model to overfit. The result of this procedures is the reduced number of the attributes used for developing the final model. This stage was repeated every time the new training algorithm was tested, attributes that proved the highest accuracy for the final chosen algorithms, Random Forest and J48, are presented in Table 3.

TABLE III. ATTRIBUTES USED FOR TRAINING THE FINAL MODEL

No.	Attributes	Unit	Min	Max	Aver.
1.	City	-	-	-	-
2.	TempPeriod07	degrees Celsius	-25.4	34.6	8.9
3.	TempPeriod14	degrees Celsius	-17.8	43	16.1
4.	TempPeriod21	degrees Celsius	-19.6	34.3	11.1
5.	Cloudiness07	tenths of sky coverage with clouds	0	10	5.6
6.	Cloudiness14	tenths of sky coverage with clouds	0	10	5.9
7.	Cloudiness21	tenths of sky coverage with clouds	0	10	4.7
8.	Humidity07	percents	14	100	82.5
9.	Humidity14	percents	10	100	57.3
10.	Humidity21	percents	18	100	75.4
11.	MinTemp	degrees Celsius	-28.7	27.1	7.0
12.	MaxTemp	degrees Celsius	-16.6	44.2	17.4
13.	Rainfall	millimeters	0	1219	2.1
14.	Occur	number	-	-	-

C. *Algorithm Selection*

Algorithms selected for the training on the dataset from this paper are one of the most popular in the supervised learning including: C4.5 (J48), SVM, KNN and AdaBoost [22]. Dataset used for developing the predictive models in this paper is highly imbalanced. According to the research by Khoshgoftaar et al. in the paper [23], where they highly recommend using the Random Forest classifiers when having the imbalanced dataset, Random Forest was also included in the selection process.

C4.5 is the extension of the ID3 algorithm which generates a decision tree where each node splits the classes based on information gain [24]. J48 is a Java open source implementation of the C4.5 algorithm. Support Vector Machine (SVM) calculates the hyperplane based on the labeled data for classifying the new data [25]. K nearest neighbors (KNN) stores all the data and classify new data by calculating the distance functions such as Euclidian, Manhattan, Minkowski and etc [26]. AdaBoost is a type of ensemble learning methods, which functions by selecting an initial algorithm first and then improving it iteratively by accounting for the incorrectly classified data from the dataset [27]. Random Forest is also an ensemble learning method for classification and regression problems. To make a classification for the new input vector, the input vector is put down on every tree in the forest where each tree gives a classification from which forest chooses the classification that outputs the most [28].

Training the algorithms for predictive purposes in this paper was followed by the attribute selection for every each of them as described in the subsection III B. After the initial evaluation of the performance of the algorithms on this dataset, only J48 and Random Forest classifier showed higher accuracy than the baseline. The results will be described in detail in the further section 4.

IV. RESULTS AND DISCUSSION

Algorithms from the subsection III C were trained on the dataset with the selected attributes and their performance was evaluated using 10-fold cross-validation, along with observation of the area under the ROC Curve and confusion matrix. Due to the fact that the data is correlated with the city attribute, two analyses were made, where all cities are involved and only the ones with a significant number of occurrences. Fig. 2 represents number of occurrences per city. The main reason for that is to evaluate the real influences of the selected attributes on the occurrences, so only instances with Belgrade and Novi Sad for the city value were used for the second analysis.

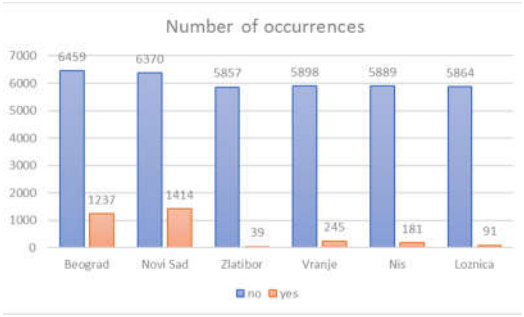

Figure 2. Number of occurrences per city

J48 and Random Forest will be involved in further discussion due to initial higher performance values than others mentioned comparing to the baseline. The models were built on training the algorithms on the attributes displayed in Table 3 selected by the techniques described in subsection III B. First analysis includes all the cities. Baseline was calculated using ZeroR algorithm, which only predicts the major class. The baseline accuracy when all cities are included is 91.1743%. Confusion matrix for ZeroR is displayed in Table 4.

TABLE IV. BASELINE CONFUSION MATRIX FOR THE FIRST ANALYSIS

Classified as A	*Classified as B*	
33130	0	A = no
3207	0	B = yes

The overall accuracy of the model trained with the pruned J48 on the same attributes, tested with 10-fold cross-validation, is 92.16%, which is slightly higher than the baseline. The confusion matrix for the developed model is displayed in Table 5 and ROC Curve in Fig. 3.

TABLE V. J48 CONFUSION MATRIX FOR THE FIRST ANALYSIS

Classified as A	*Classified as B*	
32317	813	A = no
2052	1155	B = yes

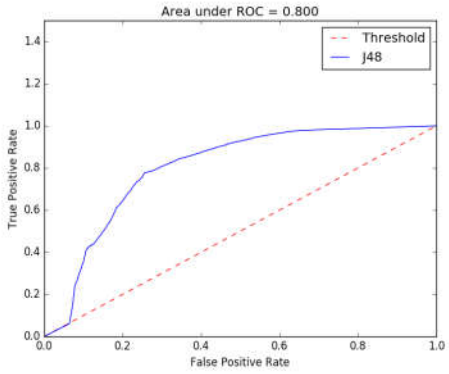

Figure 3. J48 ROC curve for the first analysis

From the confusion matrix, it could be seen that the model has good accuracy in predicting negative occurrences as expected, but there are still more false predictions when it is positive occurrences.

When applying the Random Forest on the same dataset, overall prediction accuracy reached 96.1%. The confusion matrix for the developed model is displayed in Table 6 and ROC Curve in Fig. 4.

TABLE VI. RANDOM FOREST CONFUSION MATRIX FOR THE FIRST ANALYSIS

Classified as A	***Classified as B***	
33062	68	A = no
1349	1858	B = yes

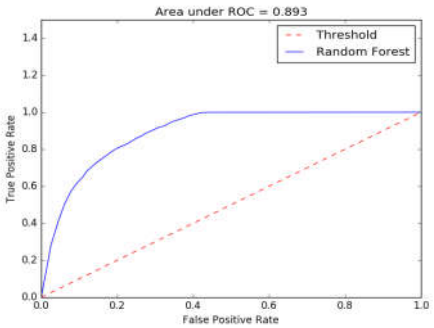

Figure 4. Random Forest ROC curve

From the confusion matrix, it could be seen that Random Forest has more successful rate in predicting the positive occurrences and has better ROC value, as the proven in the paper [26]. Due to that, best performance was achieved with default values for parameters in Weka *numTrees* and [log_2M + 1] for numFeatures, where M is the total number of attributes, as also suggested in the paper [26].

The second analysis dismissed the instances where are cities with a small number of negative occurrences, as mentioned above in this section. As it was in the first analysis, baseline was calculated first using ZeroR. Baseline accuracy is 79.34%. Confusion matrix for ZeroR is displayed in Table 7.

TABLE VII. ZEROR CONFUSION MATRIX FOR THE SECOND ANALYSIS

Classified as A	***Classified as B***	
10178	0	A = no
2651	0	B = yes

Model trained using J48 achieved an overall accuracy of 82.31%. The confusion matrix for the developed model is displayed in Table 8 and ROC Curve in Fig. 5.

TABLE VIII. J48 CONFUSION MATRIX FOR THE SECOND ANALYSIS

Classified as A	***Classified as B***	
9238	941	A = no
1328	1323	B = yes

As could be seen from the confusion matrix, better accuracy even with J48 is gained when observing the cities with more negative occurrences.

Model trained with Random Forest gave the best performances with overall accuracy of 91.82% where baseline is 79.34%. Confusion matrix for the developed model is displayed in Table 9 and ROC Curve in Fig. 6.

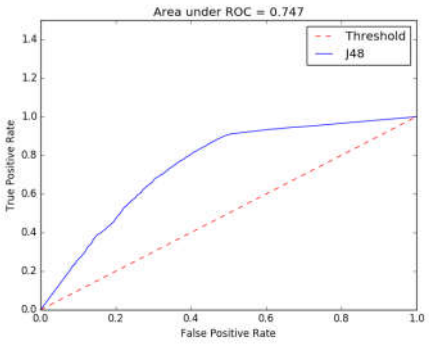

Figure 5. J48 ROC curve for second analysis

TABLE IX. RANDOM FOREST CONFUSION MATRIX FOR THE SECOND ANALYSIS

Classified as A	***Classified as B***	
10126	52	A = no
997	1654	B = yes

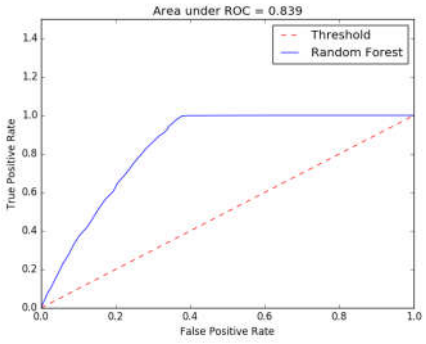

Figure 6. Random Forest ROC curve for second analysis

By inspecting the confusion matrix, it could be seen that there is not only a higher accuracy in predicting the negative occurrences, but a descent percentage in predicting the positive occurrences despite of the very imbalanced dataset.

Our research found that Random Forest is most suitable for dealing with traffic accident dataset and significant correlation between the amount of rainfall, daily changes in temperature, cloudiness, and humidity with the occurrence of traffic accidents, similar as in [15].

V. CONCLUSION

In this study, we estimate the correlation between different weather factors and traffic accident occurrences. Dataset composed of publicly available weather data and accident data is used for training machine learning algorithms with the purpose of creating predictive models.

The findings of previously mentioned studies show that rain, snow, extremely low and high temperatures are in correlation with higher risk of traffic accidents. According to the results obtained in the experimental part of this paper, it is proven that some attributes are in high correlation with the accident occurrences. Amount of rainfall, daily changes in temperature, cloudiness, and humidity have been found to be in correlation with the occurrence of traffic accidents. Besides that, it is confirmed that ensemble method such as Random Forest has very high performance in dealing with imbalanced datasets such was the one used in this paper.

In terms of future work, expansion of existing dataset would open possibilities for further research. Existing data can be supplemented by meteorological and traffic accident data that includes other major cities or regions. More detailed data concerning traffic accidents could be gathered from police reports created on scene of a traffic accident including weather conditions on traffic accident location. New parameters as road quality, driving speed, traffic volume, time of day or night when the accident happened, the number of people injured or killed open numerous directions on which further research could be based on.

REFERENCES

[1] Patz, Jonathan A. "Impact of regional climate change on human health." Nature 438.7066 (2005): 310-317.

[2] "Global Status Report On Road Safety 2015." Geneva, Switzerland: WHOPress, 2015.

[3] Fred Blackburn, Sullivan, Guerra, Zutavern, Escaravage. "The field guide to data science, Second Edition.". Booze Allen Hamilton, (2013): 21-37.

[4] Buchholtz, Sonia, Maciej Bukowski, and Aleksander Śniegocki. "Big and open data in Europe: A growth engine or a missed opportunity." Warsaw Institute for Economic Studies Report Commissioned by demosEUROPA 10, 2014.

[5] Hanretty, Chris. "Scraping the web for arts and humanities." University of East Anglia, 2013.

[6] Witten, Ian H., and Eibe Frank. "Data Mining: Practical machine learning tools and techniques. " Morgan Kaufmann, 2005.

[7] Shalev-Shwartz, Shai, and Shai Ben-David. "Understanding machine learning: From theory to algorithms." Cambridge University Press, 2014.

[8] Van den Bossche, Filip, Geert Wets, and Tom Brijs. "Role of exposure in analysis of road accidents: a Belgian case study." Transportation Research Record: Journal of the Transportation Research Board 1908 (2005): 96-103.

[9] Brodsky, Harold, and A. Shalom Hakkert. "Risk of a road accident in rainy weather." Accident Analysis & Prevention 20.3 (1988): 161-176.

[10] Leard, Benjamin, and Kevin Roth. "Weather, Traffic Accidents, and Climate Change." Resources for the Future Discussion Paper (2015): 15-19.

[11] Yannis, George, and Matthew G. Karlaftis. "Weather effects on daily traffic accidents and fatalities: a time series count data approach." Proceedings of the 89th Annual Meeting of the Transportation Research Board. 2010.

[12] Pisano, Paul A., Lynette C. Goodwin, and Michael A. Rossetti. "US highway crashes in adverse road weather conditions." 24th Conference on International Interactive Information and Processing Systems for Meteorology, Oceanography and Hydrology, New Orleans, LA. 2008.

[13] Andreescu, Mircea-Paul, and David B. Frost. "Weather and traffic accidents in Montreal, Canada." Climate Research 9.3 (1998): 225-230.

[14] Chang, Li-Yen, and Wen-Chieh Chen. "Data mining of tree-based models to analyze freeway accident frequency." Journal of Safety Research 36.4 (2005): 365-375.

[15] Krishnaveni, S., and M. Hemalatha. "A perspective analysis of traffic accident using data mining techniques." International Journal of Computer Applications 23.7 (2011): 40-48.

[16] "Index Of /Podaci/Meteo_Godisnjaci". Hidmet.gov.rs. N.p., 2016. Web. 10 Oct. 2016.

[17] "Chrome Web Store". Chrome.google.com. N.p., 2016. Web. 11 Oct. 2016.

[18] "Naslovi.Net | Najnovije Vesti". Naslovi.net. N.p., 2016. Web. 11 Oct. 2016.

[19] Ian H. Witten and Eibe Frank. "Data mining. Practical Machine Learning Tools and Techniques, Second Edition". Elsevier, 2005.

[20] Hall, Mark, et al. "The WEKA data mining software: an update." ACM SIGKDD explorations newsletter 11.1 (2009): 10-18.

[21] Hofmann, Markus, and Ralf Klinkenberg, eds. RapidMiner: Data mining use cases and business analytics applications. CRC Press, 2013.

[22] Wu, Xindong, et al. "Top 10 algorithms in data mining." Knowledge and information systems 14.1 (2008): 1-37.

[23] Khoshgoftaar, Taghi M., Moiz Golawala, and Jason Van Hulse. "An empirical study of learning from imbalanced data using random forest." 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007). Vol. 2. IEEE, 2007.

[24] Quinlan, J. Ross. C4. 5: programs for machine learning. Elsevier, 2014.

[25] Suykens, Johan AK, and Joos Vandewalle. "Least squares support vector machine classifiers." Neural processing letters 9.3 (1999): 293-300.

[26] Zhang, Min-Ling, and Zhi-Hua Zhou. "ML-KNN: A lazy learning approach to multi-label learning." Pattern recognition 40.7 (2007): 2038-2048.

[27] Rätsch, Gunnar, Takashi Onoda, and K-R. Müller. "Soft margins for AdaBoost." Machine learning 42.3 (2001): 287-320.

[28] Breiman, Leo. "Random forests." Machine learning 45.1 (2001): 5-3