Temperature prediction and analysis of the impact of weather conditions on the number of accidents in Belgrade

Ivana Tomić
Computing and automation
Faculty of Technical Sciences, University of Novi Sad
ivana1996tomic@gmail.com

Mihajlo Levarski
Computing and automation
Faculty of Technical Sciences, University of Novi Sad
mlevarski@gmail.com

Abstract— This paper deals with the prediction of temperature, the influence of weather conditions on traffic accidents in the territory of Belgrade, as well as the prediction of the type of traffic accident. The goal of this research is the automation of temperature prediction as well as the reduction of the number of traffic accidents. It was created for the purposes of the courses Systems for Research and Data Analysis and Neural Networks. Prediction performance is represented by several different machine learning models namely linear regression, DNN, MLP, RFR, LSTM, GRU, SVM, RFC, naive Bayes, CNN. The results for temperature prediction were obtained using historical weather data from the territory of Belgrade and cities from the region. After arranging the collected data, the models were trained, and then the evaluation of the obtained results was carried out. Based on the evaluation, it can be seen that the MLP model and the random forest model (RFR), as well as LSTM and CNN, are the most suitable for this problem. Data related to weather conditions in the territory of Belgrade have been combined based on date and time with data on traffic accidents for 2018. After that, the attributes of interest were graphically displayed, analyzed, and then the type of traffic accident was predicted. Through the analysis, we came to the conclusion that weather conditions and other factors have the greatest influence on the number of traffic accidents. Based on the results for predicting the type of traffic accident, we concluded that RFC and SVM are almost equally good. The basic solution and analysis are presented, as well as some expansion and improvement possibilities.

Keywords— weather forecast, prediction, analysis, traffic accidents

I. INTRODUCTION

The progress of technology is visible in almost all fields. One of the consequences of this is the facilitation of jobs. Most of the traditional processes have been replaced by processes performed by computers. This paper describes a solution that provides the application of technology in temperature prediction based on historical data on the territory of Belgrade and cities from the region (Budapest, Sofia, Skopje, Bucharest, Podgorica, Sarajevo, Zagreb). It is well known how necessary it is to be informed about the weather conditions in the future both for some minor needs and for more complex matters. However, we must be aware that due to the large number of factors that influence the weather, this process is very complex. Therefore, to be as accurate as possible, it requires domain knowledge in this area.

The idea of this solution was to use various methods and historical data for this work, as well as historical data of cities from the region, to get some more or less good results that will be described in the rest of the work.

In addition to the solution related to the prediction of the temperature in Belgrade, this paper also provides an analysis of the impact of weather conditions on traffic accidents, as well as a prediction of the type of traffic accident. The analysis and prediction were made on the basis of traffic accident data for 2018. Accordingly, data on weather conditions for this year in the territory of Belgrade were also used. This is another possibility of applying new technologies. By analyzing this data, we come to important conclusions that can help us to ensure that things like traffic accidents happen less in the future. That is, from this data we can conclude when greater precautions, better visibility and adjustment of speed are necessary. In addition, by combining these features, we predicted the outcome of traffic accidents.

As the temperature is not the only valid parameter for road conditions and traffic conditions, descriptions of weather conditions, for example "cloudy", "clear", etc., were also used in the analysis.

After the introduction, in the second section, an overview of related research is given, followed by a description of the dataset. Then, in the fourth section, an explanation of the methodology used in this research is provided. In the fifth section, the data analysis is given, and in the sixth, the prediction results are described. The seventh section provides concluding remarks and directions for future research.

II. SGENDER RESEARCH

In the following, the works dealing with the prediction of temperature will be described, followed by the works in which the analysis of the influence of weather conditions on traffic accidents was carried out.

Paper [1] solves the problem of weather forecasting based on historical data of meteorological stations from the target area as well as neighboring areas. Data was collected for the city of Nashville and surrounding cities. Unique weather records at specific times from specific areas. Each record contains data on temperature, humidity, wind direction, atmospheric pressure and atmospheric conditions. 68 in total of records for each day from July 1, 2018 to September 7, 2018. The target variable of the records is temperature, so the prediction for the next day is made based on the observation of the current day's weather within the same time period. Methods used in the work: Linear regression, Support Vector Regressor, Multi-Layer Perceptron Regressor (MLPR), Random Forest Regressor (RFR), Extra-Tree Regressor (ETR). It has been shown that these machine learning models can predict the temporal characteristics well enough. It was concluded that the use of historical data of the surrounding areas in relation to the target area shows a more efficient and accurate prediction of the weather forecast. It was noticed that the points on the graph are much more scattered when using data related to one city compared to data from 10 cities, which means that the possibility of error is much smaller in the second case. Based on these results, we can apply the same type of methodology to our research data. The training set contains data for a period of two months (starting from July 1, 2018), while the test set contains data for a period of seven days (from September 1 to 7, 2018). The root mean square error was used to estimate the model.

In the paper [2], Holmstrom et al. is applying machine learning to weather forecasting to potentially create more accurate weather forecasts for longer time periods as these techniques are more robust to disturbances compared to physical models of the atmosphere. Collected daily data (maximum and minimum temperature, mean humidity, mean atmospheric pressure) in the period from 2011 to 2015 for Stanford, California. Which means that it contains 1826 records, each referring to a single day. The data from the first four years were used to train the algorithms, and the data from the last year were used for the test. Data classified as clear, moderately cloudy, very cloudy and precipitation. They used linear regression methods as well as a variation of functional regression. They calculated the mean square error for the results obtained by linear regression, variation of functional regression, but also for professional weather forecasting. The results show that linear regression as a model does not depend so much on the parameters themselves, but on the amount of information compared to functional regression, which shows better performance with appropriately selected parameters.

Abrahamsen et al. in paper [3] they develop a model for forecasting the temperature in a period of one, three, six or twelve hours. Two experiments were performed in which AR-NN (auto-regressive neural network), ARX-NN (auto-regressive with exogenous input) and Multi Layer Perceptron (MLP) were used. The data was collected in the period from 2016 to 2017 from a meteorological station located in Porsgrund, Norway. In the first example, only temperature is used as data, and in the second, temperature and precipitation. The experiments concluded that the error increases in proportion to the number of hours for which the temperature prediction is made. If additional data, such as precipitation information, is included in the data set, improvements in performance are observed. The data is split so that 60% goes to training, 20% to validation and 20% to testing. Calculation of error using mean square error. The obtained results are presented graphically with realistically measured temperatures at a certain time. By observing the displayed test results, we come to the conclusion that the error increases in proportion to the increase in the number of hours for which we make a temperature prediction. Paper [4] provides a detailed report investigating the links between various factors that may be associated with road accidents in the United Kingdom.

In addition to weather conditions, the author also takes into account the type of road, lighting conditions, etc. Data for 2016 were used, which are placed in three tables. The first table is related to accident data and contains 32 columns and a total of 136621 records. Some of these columns are: location, time, date, lighting, weather and road surface conditions, number of participants, road type and other variables. A more detailed description of the victims is given in the second table. This table has 16 columns and 181384 rows. Each row in this table contains information about one person injured in a traffic accident. The third table gives a more detailed description of the vehicles involved in traffic accidents. In addition, government estimates of distances traveled on different types of UK road, relative road traffic density every hour of every day of the week and UK population data at local authority level are also used. Correlation between features was calculated using Pearson's and Spearman's coefficients. They used multiple linear regression methods as well as classification (K-nearest neighbors). The author had expectations that driving in the dark and in bad weather conditions would be shown as dangerous. That is, that the results will show that e.g. a higher crash rate during winter and in areas where it is generally colder. However, the results do not prove this assumption. As for weather conditions, the results show the opposite. The author concluded that bad weather discourages people from driving, which leads to fewer drivers, less congested roads and lower speeds, and therefore fewer accidents. The number of accidents cannot be linked to a certain month or period of the year, because the number is similar on a monthly basis.

The paper [5] gives us the details of the research that provides different analyzes of data on traffic accidents that occurred in the territory of the State of Qatar. Data were used on a monthly basis for a period of 7 consecutive years, ranging from 2010 to 2016. Thanks to this, it is possible to compare frequencies in a certain period of time. The method of analysis suitable in this case is the analysis of time series, such as the decomposition method. Along with time series decomposition analysis, multivariate analysis of variance is used to investigate the relationship between seasonal weather variations and the number of traffic accidents. The results of the analysis revealed that drivers and pedestrians involved in collisions with serious injuries or fatalities are significantly affected by seasonal weather changes, mostly in the winter and autumn seasons. The summer season is said

to be calmer because few pedestrians go out and walk the streets due to the unbearable temperatures, in contrast to the autumn and winter periods when the number of pedestrians increases due to more pleasant and milder temperatures. It is said that the autumn and winter period is characterized by fog that occurs at dawn and early morning due to northeast or southeast winds that cause humidity, which also affects the number of traffic accidents. In this analysis, traffic accidents are classified into four different groups: property damage, minor injuries, serious injuries and fatality.

The paper [14] describes the application of machine learning methods in order to predict the risk assessment in traffic accidents. The dataset consists of traffic accident records that occurred in the city of Porto Alegre, Brazil during the year 2013 and were described by 44 features. There are 20798 records in the dataset. Due to duplicate and invalid data it was necessary arrangement of the meeting. The data set is divided into a training set and a test set. Training data make up 60% and test data 40% of the total number. The following algorithms were used in this solution: logical regression, method of support vectors (SVM - Support Vector Machines), naive Bayes classifier, K-nearest neighbors, random forest (Random Forests). The used measures of model quality are: the area under the ROC (Receiver Operating Characteristic) curve or AUC (Area Under the Curve), then precision, response and F1 measure (harmonic mean of precision and response). The results are presented tabularly and graphically. Their analysis leads to the conclusion that predictive models for injury risk assessment can be created with high precision. The best and almost the same results were provided by logistic regression and the method of support vectors, followed by random forest.

The paper [18] describes a solution that uses recurrent neural networks for temperature prediction. They use a two-layer architecture, one of which is LSTM. The second layer is the dense layer which is connected with LSTM. The prediction is made on the basis of data downloaded from IoT stations that have sensors for temperature, humidity and some of the gases. The obtained data is compared with the data that was actually measured. LSTM input layer consists of eight neurons, contains three hidden layers of 100 neurons each and one output neuron.

III. DATA SET

In this chapter, the data set used in the prediction and analysis will be described.

The weather datasets were taken from [7]. The data was created in the meteorological station Belgrade/Batajnica. In addition to data recorded at the target location, the set also contains data recorded in the following cities: Budapest, Sofia, Skopje, Bucharest, Podgorica, Sarajevo, Zagreb. Historical data was collected using HTML code retrieval and parsing techniques using the BeautifulSoup API. For each of the cities, information from January 1, 2018 to January 1, 2020 was pulled at 30-minute intervals and converted to CSV files for further processing. Sets contain the following attributes: date, time, temperature, feel of temperature, wind, gust, relative humidity, dew point, pressure, icon, and description.

By analyzing the sets, it was decided that due to the inconsistency of the records of certain attributes, only the following attributes should be used: date and time, temperature, air humidity and pressure. Further analysis of the data revealed that not all cities have weather records at a specific time interval or moment. In order to harmonize the data as well as the time points in which the records are valid, the excess data of the neighboring cities was thrown out and harmonized in relation to the data for Belgrade. By merging and sorting all data in relation to date and time, we get a total of 25025 records.

The traffic accident data set was taken from [8] within the ODS file. The data refer to the territory of Belgrade in 2018. They contain the following attributes: identifier, date and time, type, vehicle number and description of the traffic accident. The type of traffic accident can be with material damage, injury or death. The vehicle number refers to a moving vehicle, a parked vehicle or a pedestrian. Data analysis removed the records related to the attributes description and identifier, considering the inconsistency and negligible contribution in the analysis of the impact of weather conditions on traffic accidents.

Categorical variables were converted to numerical ones. After sorting the data, the total number of records is 18064. In order to analyze the impact of weather conditions on traffic accidents, as well as to predict the type of traffic accident, we combined the data from both sets. We used data about weather conditions in the territory of Belgrade, which include the following attributes: date, time, temperature, pressure, humidity and description of weather conditions. The data is joined by date and time using a Pandas API function that joins rows based on the closest values for a common feature.

After processing, the data set of weather conditions in Belgrade and surrounding cities is divided into subsets. The training set contains 80% of the data, while the test set contains 20%. By applying scaling, the distribution of variable values is made between -1 and 1, where the mean value is close to 0. In this way, we reduce the cost of training the model, and affect the performance of the algorithm.

IV. METHODOLOGY

*A. Algorithms used:*The two basic types of supervised learning problems we solve are regression and classification. Regression is a problem of predicting a continuous target variable. Classification is a problem of predicting a categorical target variable that takes a finite number of values [15].

Linear regression is a method that allows us to study the relationship between independent and dependent variables. The input represents the feature vector Xi (independent variables), that is, the data that influence the prediction of the target feature. Y (dependent variable) is the outcome variable, that is, the target variable [6]. In this solution, the target variable is the temperature in Belgrade, while the other data represent the values contained in the vector Xi. Linear regression, depending on the number of independent variables, can be simple or complex. Simple linear regression has one independent variable, and multiple regression has two or more. Multiple linear regression was used in this solution. Figure 1 shows simple and multiple linear regression formulas with marked dependent and

independent variables. Parameters from b0 to bn, i.e. b1 in the case of simple linear regression, indicate the unknown parameters to be estimated. It is a measurement error, i.e. a residualε.

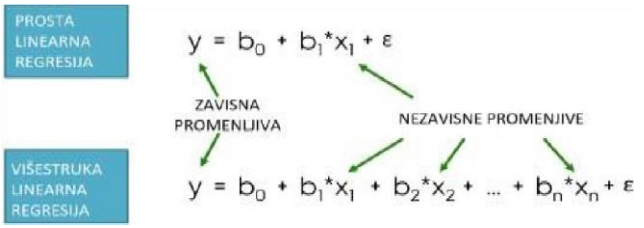

Figure 1. Simple and complex linear regression formulas

Deep neural networks (DNN-Deep Neural Network) can be viewed as a form of artificial neural network with multiple hidden layers (Figure 2). A neural network has an input and an output layer of neurons. The layers between the input and the output are called hidden layers and deep neural networks must have more of these layers. The number of neurons in the layers can be different. In this solution, different combinations of the number of neurons and layers were used for DNN. The combination of four layers with the following number of neurons: 256, 128, 64 and 32 proved to be the best of the ones tested. Each layer performs certain types of data sorting and editing, so they are suitable for dealing with unlabeled or unstructured data. Compared to shallow architectures, deep ones are more efficient when it comes to the number of parameters and computing elements for representing functions [9]. Deep neural networks can use loops. Multilayer perceptron (MLP - Multilayer perceptron) is an artificial deep neural network that generates output values based on a set of input variables. MLP connects several layers in a directed graph, which means that the signal path through the nodes goes only in one direction, that is, there are no feedback links [11]. It uses a Backpropagation algorithm that can adjust the weights in the hidden layers. Each output from one layer is used as inputs for the next layer (Figure 2), this does not apply to the output layer where the outputs represent final values. MLP gave the best results in combination with three layers of 256, 128 and 64 neurons. The ReLU (Rectified Linear Unit) activation function was used in the DNN and MLP models.

ReLU always returns a value greater than or equal to zero. It is currently one of the most popular activation functions for neurons in the hidden layer because it is easy to calculate, and thus we get faster training. The output from the neural network is one and represents our target feature, i.e. temperature, and we have 23 inputs representing other features from the described data set.

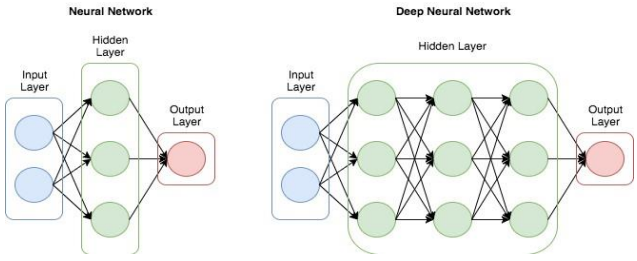

Figure 2. An example of a neural network and a deep neural network (Taken from [10])

Random Forest (RF - Random Forest) is a supervised learning algorithm that uses an ensemble of decision trees as a learning method for solving classification and regression problems. It uses the packing technique (Bootstrap aggregation) as one of the ensemble methods. In this way, it includes multiple decision trees that are run in parallel and do not interact with each other. Each of the trees makes a prediction that is finally combined to achieve the final result, shown in Figure 3. If it is a question of classification, the modality of all classes or the mean value of the prediction in the case of regression is performed. The number of features that can be shared on each node is limited to some percentage of the total number (hyperparameter). In this way, the model does not rely too heavily on any single feature and makes fair use of all potentially predictive features. Each tree takes a random sample from the original data set when splitting the nodes, introducing additional randomness that prevents overfitting [12].

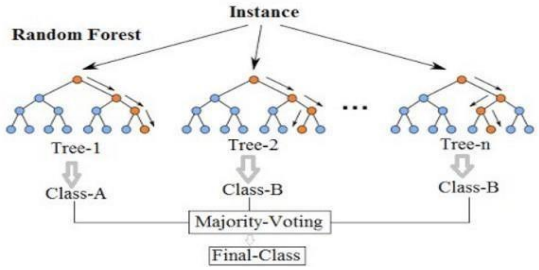

Figure 3. RF classification structure (Taken from [13])

Naive Bayes algorithm is a probabilistic model of machine learning that is used to solve classification problems, using the Bayesian formula (Figure 4).

$$p(y|X) = \frac{p(X|y)p(y)}{p(X)}$$

Figure 4. Conditional probability formula

It represents the probability that event y will occur if event X has occurred. X represents evidence, while y is a hypothesis. It is assumed that the attributes are independent, that their values do not influence each other and that they are all equally important. Hence the name "Naive Bayes". The variable y refers to a specific discrete value (class variable), while the variable X represents features (X=x1....xn). Based on that, we get the formula in Figure 5.

$$p(y|x_1, x_2 \dots x_n) = \frac{p(x_1|y)p(x_2|y) \dots p(x_n|y)p(y}{p(x_1)p(x_2) \dots p(x_n)}$$

Figure 5. Formula for conditional probability with vector x

We note that for each data set, p(x) is always the same and does not depend on y. In this way, p(x) is not calculated, so only the value in the numerator remains, and thus we get proportionality (Figure 6).

$$p(y|x_1, x_2 \dots x_n) = \frac{p(x_1|y)p(x_2|y) \dots p(x_n|y)p(y)}{p(x_1)p(x_2) \dots p(x_n)}$$

Figure 6. Proportionality

We conclude that, if xi is a categorical value, the relation is reduced to the calculation of the frequencies of its different values (Figure 7) [15] [16].

$$p(y|x) \sim p(y) \prod_{i=1}^{n} P(x_i|y)$$

Figure 7. Discriminative model

The Support Vector Machine (SVM) is a supervised learning model that can be used to solve both classification and regression problems. In this solution, it was used to solve the classification problem. The goal of the SVM classifier is to construct a hyperplane that will separate data classes with the largest possible margin. This hyperplane is called the optimal hyperplane. If we were to draw parallels through the nearest point on both sides of the hyperplane hyperplane with the optimal one, they would actually get two support vectors, after which this method got its name. The distance between the optimal hyperplane and the support vectors is equal on both sides. In practice, we can very rarely expect such a linear separation of classes, so we must be prepared to accept mistakes. With the support vector method, we can choose between the number of errors, that is, the number of points that will end up on the wrong side of the hyperplane, and the size of the margin. We introduce a parameter C that gives a certain weight to each point that is on the wrong side. A small value of the parameter C means that the margin will be large, but even a large number of cases will be allowed to be misclassified, while a large value of the parameter C means that we ignore the size of the margin and will have a small number of points that are misclassified [15] . In this solution, we tried different values for the parameter C and decided on a value of one thousand.

Recurrent neural networks have a kind of internal state (memory) because the neuron's output is fed back to its input. The method of long short term memory (Long short term memory network-LSTM) is an artificial recurrent neural network. It differs from an ordinary recurrent neural network in that it has additional components (input gate, forget gate, output gate). The memory, that is, the internal state is visible only through the activation function. GRU (Gated Recurrent Unit-GRU) is a modification of LSTM. Instead of input and forget gates, it has update and reset gates. No output gate. Another difference is that the internal state is visible from the outside because it does not have an activation function.

Convolutional neural networks (Convolutional neural network-CNN) consist of input, convolutional layer, activation function, aggregation (pooling) layer and MLP. In this solution, a 1D CNN was used whose kernel moves in one direction. The most common application of convolutional networks is in image processing, where they have proven to be extremely successful [15].

Based on research papers on this topic, we concluded that it would be convenient to use Linear Regression, MLP and RFR. By observing the results obtained in the works, which we chose as related to this problem, we concluded that these models are satisfactory for predicting the temperature height based on historical data. Also, based on the results achieved in the paper [14], we decided to use SVM, RFC and naive Bayes for the classification problem in the case of predicting the type of traffic accident.

*B. Evaluation methods:*Mean Squared Error (MSE) is calculated according to the formula shown in Figure 8 (formula for MSE). The input dimension is represented as n, Yi represents the i-th value of the obtained output, and Ŷi is the i-th value of the expected, i.e., correct output. The same applies to the variables in the formula for the Mean Absolute Error (MAE) found in Figure 8 below the formula for MSE. The advantage of the squared error is that it is differential at all points, but it is sensitive to extreme values. The opposite is true for the absolute error, that is, it is not sensitive to extreme values because it moves away from zero more slowly than the squared error, but it is not differential at all points. MSA and MSE are used in regression problems.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \hat{Y}_i|$$

Figure 8. Formula for calculating MSE and MAE

For the measures of success in the classification problem, in this solution we used accuracy, precision, recall and the F1 method. Accuracy represents the proportion of instances correctly classified in relation to the total number of instances. This method does not give accurate results if the number of data between classes is unbalanced. Precision represents the ratio between true positives and the sum of true positives and false positives, while response gives us the ratio between true positives and the sum of true positives and false negatives. The F1 gauge combines precision and responsiveness. It represents their harmonious environment.

V. DATA ANALYSIS

Analyzing the graphs in Figure 9, we can conclude that the largest number of traffic accidents with material damage occurred in February, although the weather conditions were favorable during that period, i.e. it was mostly cloudy without precipitation. The largest number of people injured in traffic accidents occurred in March and December, when driving conditions were unfavorable, i.e. when there was snow and ice. The greatest number of tragedies occurred in the summer period when it was cloudy with periodic precipitation.

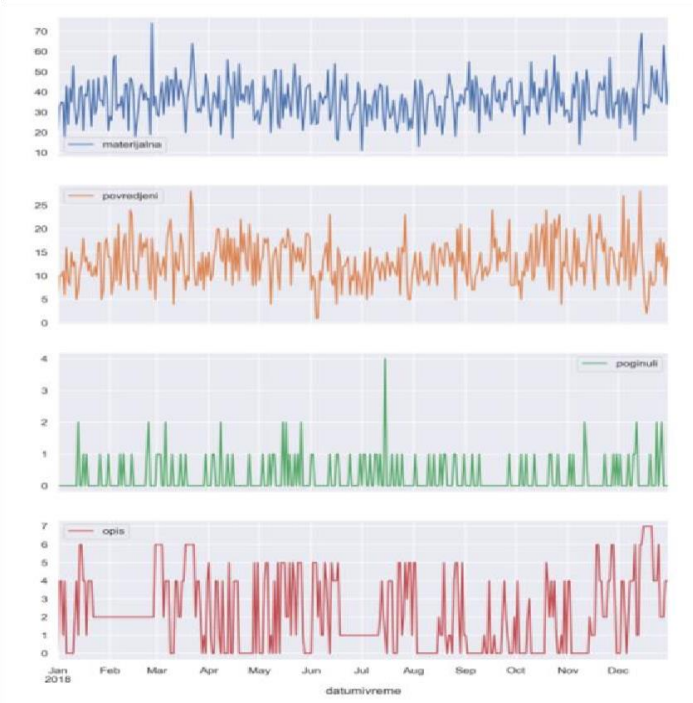

Figure 9. Number of traffic accidents by month

The data in Figure 10 represent the percentage of traffic accidents by description of weather conditions. The data is shown by summarizing the number by group (material, injured and dead), for a certain weather condition, divided by the total number of recorded records for that weather occasion. In this way, we show the influence of the description of weather conditions on traffic accidents, regardless of the frequency of the weather phenomenon. Here we see that the largest number of traffic accidents were without victims and that was in cloudy weather and icy fog.

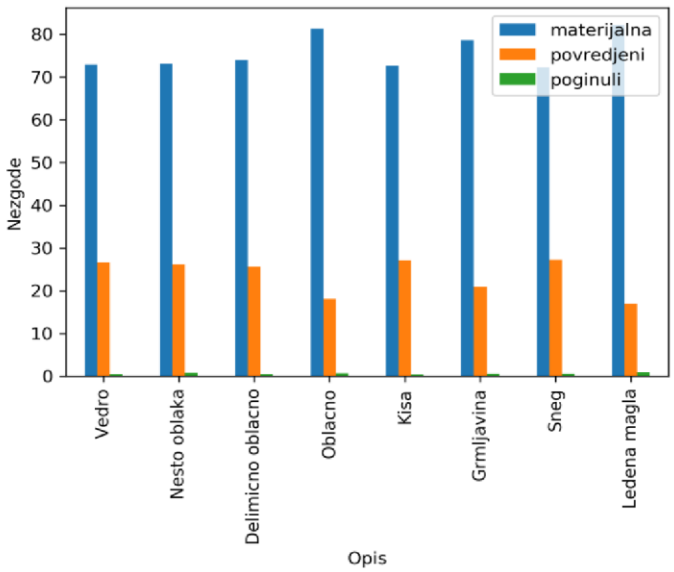

Figure 10. The number of traffic accidents divided into three classes based on the description of weather conditions

In picture number 11, we have a more precise insight into more serious traffic accidents. We can confirm the conclusion we reached by analyzing Figure 9, but also that there is a danger when there is no precipitation. We assume that this is due to the increased number of pedestrians when there is no precipitation. The data in figure number 11 is presented in the same way as in figure number 10.

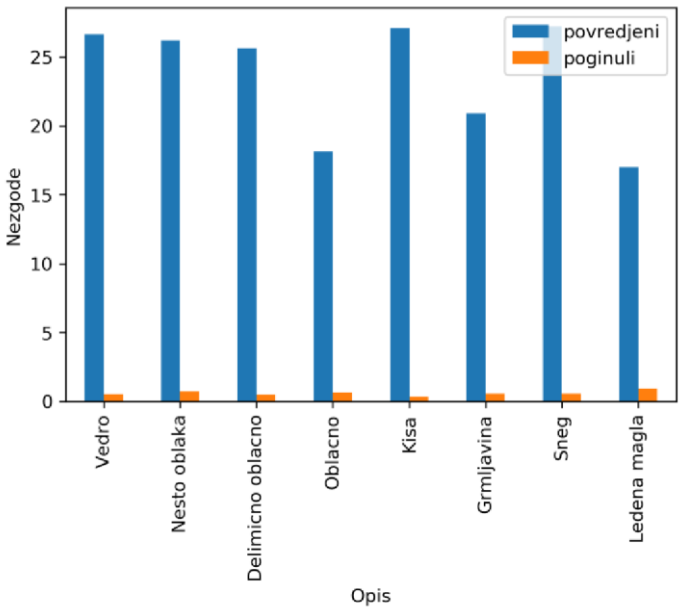

Figure 11. Number of traffic accidents with excluded accidents that resulted in material damage based on the description of weather conditions

From the graphics shown in Figure 12, we can conclude that accidents (observing all three groups) occurred in more extreme cold weather as well as more pleasant temperatures, which additionally confirms the previous conclusion.

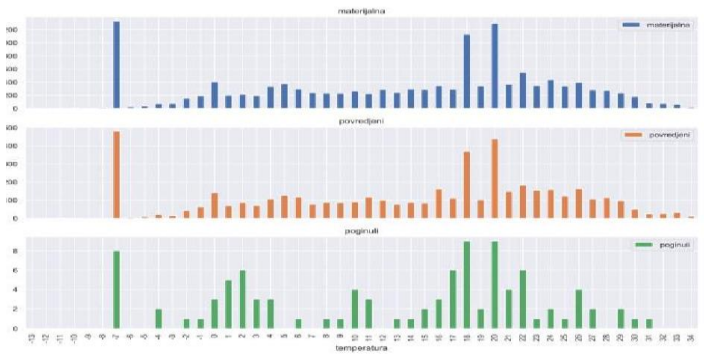

Figure 12. Total number of traffic accidents in relation to temperature

Figure 13 shows the total number of traffic accidents by location. We can conclude that traffic accidents occur more often on roads with a higher frequency of traffic, that is, in the city center and the surrounding area.

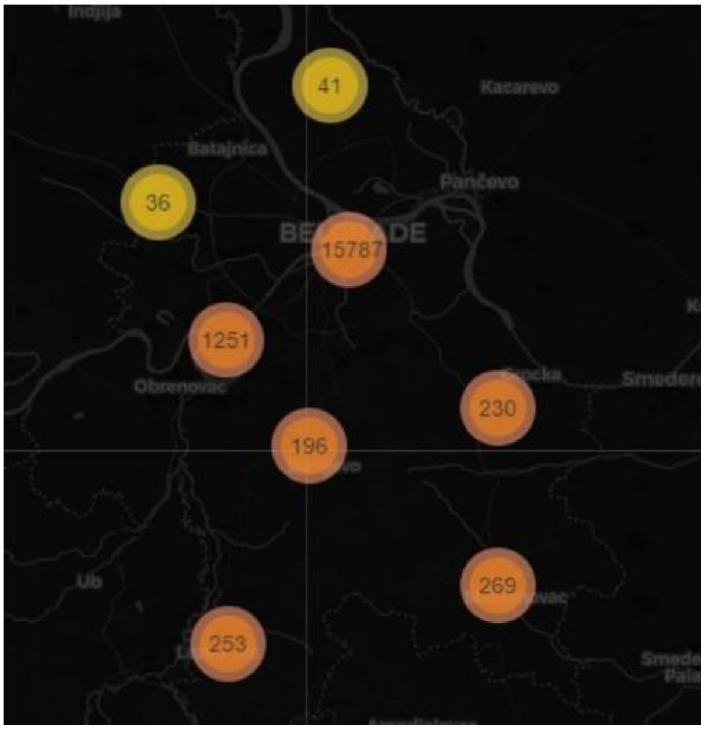

Figure 13. Display of the number of traffic accidents by region

Figure 14 shows the dependence between the temperature in Belgrade and other characteristics. The correlation was performed on the data set that was used to predict the temperature in Belgrade. We can see that the highest correlation has the temperature in the surrounding cities, followed by pressure and humidity.

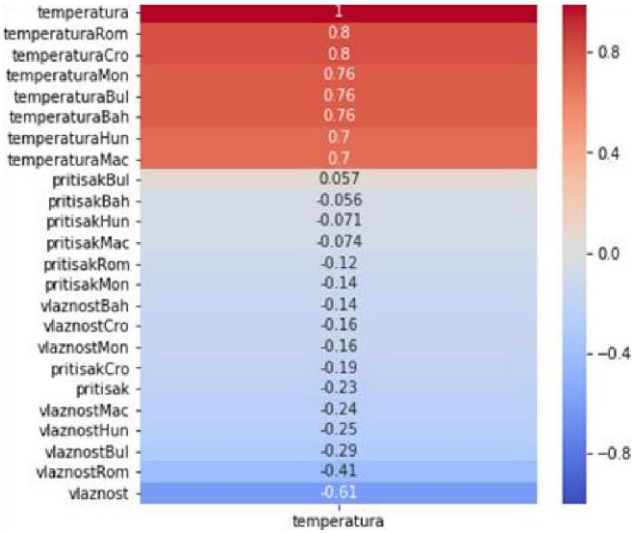

Figure 14. Pearson coefficient of mutual correlation of temperature with other features

In the data set that contains weather conditions, we can notice that the largest number are clear and cloudy weather conditions, which is also characteristic of our area. As for the set with traffic accidents, we can conclude that it is unbalanced. That is, we have a drastically higher number of material traffic accidents compared to the other two classes.

VI. RESULTS AND ANALYSIS

In this part, the results of the prediction of temperature and type of traffic accident achieved with different models will be presented. The model is expected to generalize well, that is, to rarely make large errors when predicting the values of the target variable based on the attribute values.

The check will be performed by comparing the predicted temperature with the actual settled temperature, since it is a supervised learning. The difference between the obtained and the correct temperature is considered an error. The mean square error method and the mean absolute error method were used to calculate the error.

In Table 1, we see that MAE and MSE are not large for linear regression, that is, that the prediction is done satisfactorily. However, this model is not entirely suitable for prediction because it requires a linear relationship between the dependent and independent variables.

	MAE	MSE
Linear regression	1.75	5.11
Linear regression T	3.79	22.6
Linear regression V	5.38	44.46
Linear regression P	6.02	53.61
DNN	2.86	13.98
DNN T	2.73	13.41
DNN V	4.53	32.91
DNN P	6.4	59.67
MLP	0.66	0.73
MLP T	2.26	9.21
MLP V	3.43	21.13
MLP P	2.7	12.41
RFR	0.65	0.94
RFR T	1.62	5.73
RFR V	1.90	7.53
RFR P	0.97	2.16

Table 1. Presentation of the mean absolute and mean square error by model

In Table 1, the T next to the model name indicates that only temperature features were used for its training. The same applies to other cases where P stands for pressure and V for humidity. Data from Belgrade and cities from the region were used for model training, the results of which are shown in Table 1.

Based on the results from Table 1, we can conclude that the MLP and RFR methods proved to be the best for training the model when we use features related to temperature, pressure and humidity. We can see how the accuracy of the model decreases by excluding features from it. The trained models confirm the results obtained by correlation. That is, we see that the error is smaller for the models trained by temperature, than the models for which we used only humidity or only pressure for training. In addition to these modifications in model training, changes were also made with the number of layers and neurons in them. For example, the MLP model trained with three layers of one hundred neurons had MAE=1.98 and MSE=6.13, while three layers with 256, 128 and 64 neurons gave us MAE=0.66 and MSE=0.73 as a result. The DNN model was trained with different number of layers and neurons, we concluded that we get more accurate. results if we use more layers. The results obtained with three layers of one hundred neurons are MAE=3.89 and MSE=20.69, and for two layers of 50 neurons, MAE=4.06 and MSE=25.94 were obtained. Guided by these results, we increased the number of layers to four and the number of neurons per layer to 256, 128, 64 and 32 and obtained the result shown in the table. Experiments on the number of layers and neurons were performed on the complete set of features.

If only historical weather data on the territory of Belgrade is used as a set of data, we notice that the precision of the prediction decreases. By linear regression, we get MAE=5.8, and MSE=48.78. If we compare this result with the result from the table that was trained with the complete set of features, we can conclude that the data from the cities in the region are of great importance

on the accuracy of the model. The same applies to other methods. MLP gave the following accuracy over the set of features that only includes data from Belgrade: MAE=5.23 and MSE=40.43. By using the LSTM method, we obtained satisfactory results. MSE and MEA were approximate for different model types. The model was trained on historical data related only to the temperature in Begorad, then on data containing, in addition to temperature, humidity and pressure, as well as on the entire data set containing this information for the surrounding cities. Errors of the type of model that gave the least errors over the data set containing temperature, humidity and pressure in the territories of Belgrade is MAE=0.65, and MSE=0.87. The model, which was trained on a data set containing only the temperature in the territory of Belgrade, used historical data from the previous three days in order to predict the temperature in thirty minutes. The best type of this model gave MAE=0.61 and MSE=0.71 results. Over the complete data set, which also includes data from cities in the region, the best type of model has MSE=1.11 and MAE=0.79. The code for these examples is available at [17], as well as for the examples whose results are described below.

Almost all types of LSTM performed better than the GRU model. The GRU model is trained on a dataset that includes data on surrounding cities. The obtained result is MAE=0.79, and MSE=1.26.

The convolutional neural network model was trained on a set of data related only to the territory of Belgrade.

In addition to temperature, the data also includes values for air humidity and pressure. For this model, the MAE is 0.13.

	Precision	Response	F1-method
Material RFC	0.77	0.99	0.87
Violated RFC	0.89	0.22	0.35
RFC casualties	0.00	0.00	0.00
Accuracy RFC			0.78
Average RFC	0.80	0.78	0.72
Material NB	0.74	0.94	0.83
Injured NB	0.47	0.13	0.20
Killed NB	0.00	0.00	0.00
Accuracy NB			0.71
Average NB	0.66	0.71	0.65
Material SVM	0.76	0.98	0.86
Injured SVM	0.79	0.19	0.31
SVM died	0.00	0.00	0.00
Accuracy of SVM			0.76
Average SVM	0.77	0.76	0.70

Table 2. Presentation of the results obtained for the prediction of the type of traffic accident

Table 2 shows the results obtained for predicting the type of traffic accident. Three presented classification models were used. By comparing the results, we see that the method of support vectors (SVM) and random forests (RFC) have approximate results. The results obtained are as expected. If you could enrich the data set with more features and extend it to a larger time interval, the quality of these models would increase. By analyzing the data, we saw that this set lacks characteristics that have a greater correlation with the type or outcome of the traffic accident. Using the naive Bayes method, we got the weakest results, which we expected based on the features of the algorithm described in chapter 4.

Landmark	**Important**
Vehicle number	0.69
Latitude	0.10
Longitude	0.09
Pressure	0.04
Humidity	0.03
Temperature	0.03
Description	0.01

Table 3. Importance of attributes within the RFC model

Based on table 3, we can see the importance of other attributes in relation to the target, specifically the type of

traffic accident. Using the random forest method, we get this attribute importance table. We conclude that the number of vehicles has the greatest influence.

VII. Conclusion

In this paper, it has been shown that it is possible to achieve satisfactory results of temperature prediction and traffic accident outcome prediction using machine learning methods. In order to improve the results, it is necessary to include reliable data and people with domain knowledge in the field of meteorology in this research. In addition to improving the results, this work could be upgraded by predicting additional parameters related to weather conditions, such as the amount of precipitation. The difficulty we encountered when creating this project was the lack of data for certain time periods.

Nowadays, while traffic accidents are happening all the time, traffic safety has become the main focus of contemporary social issues. Many factors influence the number of traffic accidents and their outcome. Some of them are listed in this paper such as the place of the accident, the period of the year and weather conditions. This paper does not solve the issue of traffic accidents, but it indicates some important factors. It is necessary to raise the awareness of all road users in order to be more careful in situations that are described as the most characteristic for traffic accidents. In addition, the introduction of increased traffic flow control in more urban areas would contribute to increased safety.

This research could be extended by additional analysis of human factors involved in traffic accidents, as well as vehicle and road category data. Data that would be useful in this research related to human factors are the number of years, gender, how long it has been owned

driver's license, but also additional parameters related to the driver's condition (alcohol influence, drowsiness, etc.). Vehicle characteristics of importance would be age, make, model, cubic capacity and the like. An additional attribute related to the traffic accidents themselves could be the cause. Thanks to this, they would have insight into whether the traffic accident occurred due to the carelessness of the driver, aggressive driving or a violation of traffic regulations.

Literature

[1] AHM Jakaria, Md M. Hossain, MA Rahman, "SmartWeather Forecasting Using Machine Learning: A Case Study in Tennessee", 2019, Available on the: https://www.researchgate.net/publication/330369173_Smart_Weather_Forecasting_Using_Machine_Learning_A_Case_Study_in_Tennessee[Accessed 2 May 2020].

[2] M. Holmstrom, D. Liu, C. Vo, "Machine Learning Applied to Weather Forecasting", Stanford University 2016, Available at: http://cs229.stanford.edu/proj2016/report/HolmstromLiuVoMachineLearningAppliedToWeatherForecasting-report.pdf [Accessed 2 May 2020].

[3] EB Abrahamsen, OM Brastein, B. Lie, "Machine Learning in Python for Weather Forecast based on Freely Available Weather Data", Proceedings of The 59th Conference on Simulation and Modeling (SIMS 59), 26-28 September 2018, Oslo Metropolitan University, Norway. Available on the:https://www.ep.liu.se/ecp/153/024/ecp18153024.pdf [Accessed May 2, 2020].

[4] Toby Staines, "Environmental Conditions and Road Traffic Collisions in the UK", 2018, Available on the:https://github.com/tobystaines/RoadAccidentsPODS/blob/master/Environmental%20Conditions%20and%20Road%20Traffic%20Collisions%20in%20the%20UK%20v1.2.pdf [Accessed 2 May 2020].

[5] Ch.Timmermans, W.Alhajyaseen, A. Al Mamun, T. Wakjira, M. Qasem, M. Almallah, H. Younis, "Analysis of road traffic crashes in the State of Qatar", 2019, Available at: https://www.tandfonline.com/doi/full/10.1080/17457300.2019.1620289# [Accessed 3 May 2020].

[6] E. Marsden, "Regression analysis using Python", Available at:https://risk-engineering.org/static/PDF/slides-linear-regression.pdf [Accessed 4 May 2020].

[7] FreeMeteo - Weather forecasts., Available at:https://freemeteo.rs/ [Accessed 4 May 2020].

[8] Republic Serbia Portal open data., Available on the:https://data.gov.rs/sr/datasets/podatsi-about-traffic-accidents-Forterritory-of the city-Belgrade/ [Accessed 4 May 2020].

[9] A. N George, "Deep Neural Network Toolkit & Event Spotting in Videousing DNN features", Department of computer science and engineering Indian Institute of Technology Madras,.May 2015. Available at: https://www.academia.edu/20434118/Deep_Neural_Network_Toolkit_And_Event_Spotting_in_Video_using_DNN_features [Accessed 4. May 2020].

[10] "Simple tutorial to write deep neural network by TensorFlow". Available at:http://marubon-ds.blogspot.com/2017/09/simple-tutorialthat-write-deep-neural.html [Accessed 4 May 2020].

[11] NK Kain, "Understanding of Multilayer Perceptron (MLP)". Available at:https://medium.com/@AI_with_Cain/understanding-of-multilayerperceptron-mlp-8f179c4a135f [Accessed 4 May 2020]. [12] Afroz Chakure, "Random Forest Regression". Available at:https://towardsdatascience.com/random-forest-and-and TS-implementation-71824ced454f [Accessed May 5, 2020]

[13] Afroz Chakure, "Random Forest Classification". Available at: https://towardsdatascience.com/random-forest-classification-and-and TSimplementation-d5d840dbead0 [Accessed May 5, 2020]

[14] Christian S. Perone, "Injury risk prediction for traffic accidents in Porto Alegre/RS, Brazil (draft)", Pontif˙ıcia Universidade Catolica do Rio Grande do Sul, February 2015, Available at: https://arxiv.org/pdf/1502.00245.pdf [accessed 25 May 2020]

[15] M. Nikolić, A. Zečević., "Machine Learning", Belgrade, 2019, Available at: http://ml.matf.bg.ac.rs/readings/ml.pdf [accessed May 30, 2020]

[16] R. Gandhi., "Naive Bayes Classifier", 2018, Available at: https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c [accessed 30 May 2020]

[17] Github Repository, Available at:https://github.com/TomicI/NM

[18] Ben AOI, Boudhir AA, Astito A., Bassam Z., Bouhorma M., "Deep Learning architecture for temperature forecasting in an IoT LoRa based system", NISS19, March 27–29, 2019, Rabat, Morocco 2019 Association for Computing Machinery , Available at: https://www.researchgate.net/publication/333258384_Deep_Learning_architecture_for_temperature_forecasting_in_an_IoT_LoRa_based_system [accessed 14 June 2020]