Operational Research in Engineering Sciences: Theory and Applications
Vol. 5, Issue 1, 2022, pp. 152-168
ISSN: 2620-1607
eISSN: 2620-1747
DOI: https://doi.org/10.31181/oresta240322136m

NIGHT TRAFFIC FLOW PREDICTION USING K-NEAREST NEIGHBORS ALGORITHM

Dušan Mladenović *, Slađana Janković, Stefan Zdravković, Snežana Mladenović, Ana Uzelac

University of Belgrade, Faculty of Transport and Traffic Engineering, Serbia

Received: 31 December 2021
Accepted: 14 February 2022
First online: 24 March 2022

Original scientific paper

Abstract: *The aim of this research is to predict the total and average monthly night traffic on state roads in Serbia, using the technique of supervised machine learning. A set of data on total and average monthly night traffic has been used for training and testing of predictive models. The data set was obtained by counting the traffic on the roads in Serbia, in the period from 2011 to 2020. Various classification and regression prediction models have been tested using the Weka software tool on the available data set and the models based on the K-Nearest Neighbors algorithm, as well as models based on regression trees, have shown the best results. Furthermore, the best model has been chosen by comparing the performances of models. According to all the mentioned criteria, the model based on the K-Nearest Neighbors algorithm has shown the best results. Using this model, the prediction of the total and average nightly traffic per month for the following year at the selected traffic counting locations has been made.*

Keywords: *machine learning, traffic flow, prediction, K-Nearest Neighbors, Weka.*

1. Introduction

The accelerated urban development is faced with mobility challenges caused by increased transport of passengers and goods. The development of smart cities is based on the analysis of traffic data. They are used in dimensioning of road sections, connections and intersections, as well as dimensioning of road structures, environmental protection measures, economic and financial evaluation of projects, planning of management and maintenance of road infrastructure (Public Enterprise "Roads of Serbia", 2012). Monitoring the road network is one way to collect real-time traffic data. Various sensor technologies prevail in this type of data collection, such as technologies based on inductive loop detectors, laser radar sensors, etc. (Magalhaes et al., 2021).

* Corresponding author.
d.mladenovic@sf.bg.ac.rs (D. Mladenović), s.jankovic@sf.bg.ac.rs (S. Janković), s.zdravkovic@sf.bg.ac.rs (S. Zdravković), snezanam@sf.bg.ac.rs (S. Mladenović), ana.uzelac@sf.bg.ac.rs (A. Uzelac)

The monitoring of traffic flows is important, both because of monitoring of the traffic conditions in real time, and because of predicting the characteristics of traffic flows in the future (Janković et al., 2020). Time determinants, such as: a day of the week, an hour of the day, the dates of state and religious holidays, holiday vacations, and so on, are some of the factors that permanently influence the formation of the usual intensity of traffic flows. Some other factors, such as: weather conditions, road conditions, maintenance of road infrastructure (Sénquiz-Díaz, 2021), use of alternative routes and traffic accidents can influence the characteristics of traffic flows to change for the observed time interval. In the situation where the flow of vehicles exceeds the capacity of the road congestion occurs. Traffic congestion leads to: prolongation of time spent in transport, increase in transport costs, increase in emissions of harmful gases, passenger delays, as well as delays in the delivery of goods. Therefore, the prevention of traffic congestion is one of the most important goals of predicting the characteristics of traffic flows.

Supervised machine learning is a method of predictive analysis that enables prediction of future values of a target variable for independent attributes in the future, based on known values of the same target variable and known values of the same attributes in the past. Collection of traffic data provides opportunities for the development of supervised machine learning models which are going to be used to predict the characteristics of future traffic flows (Zhang et al., 2020; Park et al., 2018; Xu et al., 2013).

The forecasting of traffic flows has been the subject of numerous studies over the last two decades. The second section of this paper contains an overview of the most significant studies related to this subject. The authors of this paper have limited their research to detection of night traffic patterns and the prediction of night traffic (i.e. traffic in the time period from 22.00 hours to 06.00 hours). The purpose of this research is to examine the possibilities of short - term prediction of night traffic volume using the technique of supervised machine learning. The methodology according to which this research has been performed and the basic characteristics of the algorithm that has shown the best results in prediction (K-Nearest Neighbors, K-NN) are presented in the third section of this paper. The fourth section of the paper describes a case study realized within this research. In the case study predictive models have been created and the prediction of the total and average amounts of night traffic per month has been performed on selected road sections in Serbia. The data collected by automatic traffic counters (ATC) have been used in training and testing of machine learning models. The most significant results of the case study and discussion on the results are presented in the fifth section of the paper, while the last sixth section concludes the paper.

2. Literature Review

All models developed for traffic prediction can be broadly classified into three categories: parametric, nonparametric and hybrid types of models. Parametric models are e.g. historical average (Williams et al., 1998) time series models and Kalman filter (Guo & Williams, 2010). Seasonal autoregressive integrated moving average (ARIMA) is a classic parametric time series model used in the study (Williams & Hoel, 2003). In contrast, nonparametric models are mostly data-driven and use empirical prediction methods, including primarily Neural Networks models (Vlahogianni et al., 2005; Yasin

Çodur & Tortum, 2015), nonparametric regression (Marković et al., 2010; Cai et al., 2016), and Support Vector Machine (Zhang & Xie, 2008; Peng & Tang, 2015). In addition, the hybrid approach combines two or more models to generate predictions, e.g. non-linear chaotic prediction model (Wang & Shi, 2013), multiagent prediction model (Ma et al., 2001), modular network model (Vlahogianni et al., 2007), etc. The Karlaftis & Vlahogianni study (2011) compares traffic forecasting models based on parametric (statistical) methods and neural network-based models. Boukerche & Wang (2020) provide a classification and an overview of machine learning models used in traffic flow prediction. According to these authors, the mentioned models are divided into regression models, instance-based models (such as K-NN), kernel-based models (such as Support Vector Machine - SVM and Radial Basis Function - RBF), neural network models (such as Feed Forward Neural Network - FFNN, Recurrent Neural Network - RNN, Convolutional Neural Network - CNN) and hybrid models (combinations of two or more different models).

Shamshad & Sarwr (2020) developed a model for predicting traffic volume at an hourly level, using two machine learning algorithms: Artificial Neural Network (ANN) and SVM. Traffic data obtained with the help of road sensors, as well as data on meteorological conditions have been used to train and later test different machine learning models. This study shows that ANN-based machine learning models show good results in long-term predictions, while SVM-based models show good results in short-term predictions.

Zhang et al. (2013) have developed a nonparametric regression model, based on the K-NN algorithm on the MATLAB platform. The experimental results of this study show that the prediction accuracy of the highway traffic volume, using the K-NN method, is over 90 percent accurate. In the study (Zou et al., 2015) the authors show that, when applying K-NN methods in short-term traffic prediction, a much more accurate prediction is achieved if, in addition to temporal attributes, spatial attributes are included in independent attributes as well. In some studies, the basic K-NN method for short-term traffic prediction has been improved, in some way. For example "Specifically, two screening layers based on shape similarity were introduced in the K-nearest Neighbor non-parametric regression method, and the forecasting results were output using the weighted averaging on the reciprocal values of the shape similarity distances and the most-similar-point distance adjustment method. "(Pang et al., 2016). Zheng & Su (2014) have introduced a time limit when selecting the nearest Neighbors.

In the study (Liu et al., 2018), a short-term prediction of traffic volume has been performed using a hybrid model, based on the ANN and K-NN algorithms. Four types of ANN have been used: back-propagation (BP) neural network, radial basis function (RBF) neural network, generalized regression (GR) neural network, and Elman neural network. The K-NN method has been used to reconstruct a data set on which artificial neural networks have been trained, combining similar traffic flow patterns. By applying these ANNs to real traffic data two important conclusions have been reached: BP and GR neural networks show better prediction performance than the other two types of networks, but are sensitive to changing the scope of the training data set. On the other hand, the RBF and Elman neural networks show prediction results that are fairly stable when increasing the data set for training. The study (Toan & Truong, 2020) shows that applying K-NN methods to a training data set can significantly reduce the size of this data set, thus achieving faster model training using SVM methods, without affecting prediction performance.

In the research (Filipovska & Mahmassani, 2020) different models of machine learning for predicting traffic interruption have been developed and tested and their results have been compared to the results of traditional probabilistic approach.

Stojčić (2018) has given an overview of research in which the ANFIS (Adaptive Neuro-Fuzzy Inference System) model has been used in the prediction of traffic congestion. Zaki et al. (2016), as well as Shankar et al. (2012) take velocity and density as independent attributes and congestion level as a dependent variable in the prediction of congestion using the ANFIS model. Kukadapwar & Parbat (2015), among others, use traffic volume to roadway capacity ratio as an independent variable, while the target variable in their study is congestion index.

Recent research includes the application of deep learning methods in the prediction of traffic flow intensity (Wang et al., 2018). In the study (Lv et al., 2015) the application of a deep learning approach is demonstrated with stacked autoencoders (SAEs) to traffic data sets that have Big Data features. Alshaykha & Shaban (2021) combine the K-NN method and the Broad Learning System (KNN-BLS). "The basic structure of BLS is built on the traditional RVFLNN (Random Vector Functional-Link Neural Network), but unlike RVFLNN that directly uses the original input data to build an enhanced node, BLS first maps the input into a series of mapping nodes, and then uses the mapping node to build an enhanced node, and the mapping node and the enhanced node form joint Nodes, and finally combine the nodes and the output layer to establish a linear connection." (Alshaykha & Shaban, 2021). Mohammed & Kianfar (2018) have investigated the application of four categories of predictive methods in traffic flow prediction. The results obtained using distributed random forest method slightly exceed the results obtained using other methods.

3. Methodology

The machine learning process takes place in the following stages: data preparation, model training, model validation, model testing and prediction. It is an iterative process in which all of the above mentioned phases are repeated as many times as necessary. The repetition of these phases ends when all attribute combinations, all available algorithms and algorithm parameter values are exhausted, or when a satisfactory model performance is reached. Once the model testing shows that the model is successful, the use of the model in the prediction of the selected variable can begin.

The data preparation consists of: cleaning raw data from incomplete records or records with incorrect values, converting data into the appropriate format, etc.

The construction of the prediction models consists of:

1. Selection of the target variable, i.e. an attribute whose value should be projected using a machine learning model;
2. Selection of an algorithm, in accordance with the nature of the target variable and attributes;
3. Selection of relevant attributes of the data set;
4. Preparation of data sets for learning and testing of models, according to the requirements of the selected algorithm;

5. Model adjustment, i.e. values of hyperparameters specific to each type of machine learning algorithm;
6. Model learning – implies obtaining model's hyper-parameters through applying a training data set algorithm on the training data set.

Since the target variables of the data set used in this study are continuous, machine learning models based on the most popular regression algorithms have been built: Linear Regression, K-Nearest Neighbors, Decision Tree, Support Vector Machines for Regression (SMOreg), Neural Network.

In addition to model training and testing, a model validation has been performed in order to select the best type of model among multiple candidates, determine the optimal configuration of model parameters, and avoid problems known as overfitting and underfitting. Excessive matching refers to a situation in which prediction for instances from the training set has been perfectly learned through the model, but there is a very weak ability to predict instances that are slightly different from those learned. Insufficient matching refers to a case when there is failure to approximate training data through the model, so it shows poor performance even on a training data set.

An approach known as cross-validation has been used to validate a model. This approach to model performance evaluation uses only training data and consists of the following phases:

1. The available data set for model training is divided into K equal parts - folds. It is usually divided into 10 subsets (10-fold cross-validation).
2. The model is trained on K-1 subsets of data (e.g. on the first of K-1 subsets).
3. The model is evaluated on the only remaining (K-th) subset of data.
4. Steps 2 and 3 are repeated K times. In each iteration one part of the data is taken for the purpose of model validation, while the rest (K-1 parts) is used for learning. A different subset is always selected to be used for model validation.
5. Model performances are calculated as the arithmetic mean of the performances obtained in K iteration.

Success of the numerical prediction can be evaluated using different metrics (Witten et al., 2017). The projected values of the target variable, obtained for the set of instances for model validation are: p_1, p_2, …, p_n; while the actual values of the target variables are: a_1, a_2, …, a_n.

Mean-squared error - Eq. (1), is the average error.

$$Mean - squared\ error = \frac{(p_1-a_1)^2+\cdots+(p_n-a_n)^2}{n} \quad (1)$$

Mean-absolute error – Eq. (2), is the mean of the absolute value of the errors.

$$Mean - absolute\ error = \frac{|p_1-a_1|+\cdots+|p_n-a_n|}{n} \quad (2)$$

Root mean-squared error – Eq. (3), is calculated in an obvious way.

$$Root\ mean - squared\ error = \sqrt{\frac{(p_1-a_1)^2+\cdots+(p_n-a_n)^2}{n}} \quad (3)$$

Relative-squared error – Eq. (4) is the square root of the mean of the squared errors.

$$Relative - squared\ error = \frac{(p_1-a_1)^2+\cdots+(p_n-a_n)^2}{(a_1-\bar{a})^2+\cdots+(a_n-\bar{a})^2} \quad (4)$$

Root relative-squared error – Eq. (5), is calculated in an expected way.

$$Root\ relative - squared\ error = \sqrt{\frac{(p_1-a_1)^2+\cdots+(p_n-a_n)^2}{(a_1-\bar{a})^2+\cdots+(a_n-\bar{a})^2}} \quad (5)$$

Relative-absolute error – Eq. (6), is the total absolute error, with the same type of normalization.

$$Relative - absolute\ error = \frac{|p_1-a_1|+\cdots+|p_n-a_n|}{|a_1-\bar{a}|+\cdots+|a_n-\bar{a}|} \quad (6)$$

The last measure of prediction accuracy is the correlation coefficient - Eq. (7), which measures the statistical correlation between the values of a and p. The correlation coefficient takes values from 1 for results that are completely correlated, over 0 when there is no correlation, to -1 when the results are in perfect negative correlation.

$$Correlation\ coefficient\ = \frac{S_{PA}}{\sqrt{S_P S_A}} \quad (7)$$

where S_{PA}, S_P and S_A are calculated as shown in (8):

$$S_{PA}\ = \frac{\sum_{i=1}^{n}(p_i-\bar{p})(a_i-\bar{a})}{n-1}, \quad S_P\ = \frac{\sum_{i=1}^{n}(p_i-\bar{p})^2}{n-1}, \quad S_A\ = \frac{\sum_{i=1}^{n}(a_i-\bar{a})^2}{n-1} \quad (8)$$

In the great number of empirical examples, the predictive model which is the best according to one measure is also the best in all other measures of error.

In order to predict the performances of models using unknown data, it is necessary to determine measures of their performance on a data set that did not play any role in model training. This previously uknown data set is entitled as the test data set.

The next phase is comparing the performances of models obtained on the test data set with the performances obtained on the training data set. This type of comparison enables to avoid a problem known as overfitting. If the performance of a model is good on training data but bad on the test data, then there is overfitting.

In order to predict the values of the selected target variables in the future, it is necessary to prepare an appropriate set of data and apply to it the machine learning model chosen as the best. In this research, the best results have been shown by machine learning models based on the K-NN algorithm.

The K-NN algorithm belongs to a class of supervised machine learning algorithms in model learning based on instances (Instance-Based Learning). In this class of algorithms, the classification of a new instance is done by comparing it with the most similar (the closest) instances in the training set (Aha et al., 1991). K is a parameter that indicates the number of most similar instances in the training set, with which the new instance is being compared. The K-NN algorithm belongs to the group of so-called lazy methods, because the decision on classification is postponed until the moment a new instance appears.

The main advantage of lazy methods is that they construct a different approximation of the objective function for each new instance that needs to be classified. Such local assessment of the objective function is suitable for complex objective functions. Because their models are slower to train than some other classes of algorithms, this algorithm is suitable for relatively "small" data sets. This feature of the K-NN algorithm has made it a good candidate for prediction in a case study conducted as part of this research.

In the Weka (Waikato Environment for Knowledge Analysis) software tool used in this study, the K-Nearest Neighbors algorithm has been implemented under the name IBk. Target variable (class), as well as attributes, with this algorithm can be: nominal, numerical, date or binary and missing values of class, as well as missing values of attributes are allowed. Thus, the K-NN algorithm is applicable both in solving classification problems and regression prediction problems. In this research, it has been applied to regression predictive analysis.

4. Case study

A Total of 391 automatic traffic counters have been installed on the network of state roads of the 1st category in the Republic of Serbia. Through automatic traffic counters vehicles are detected and classified in real-time, using inductive loops that are placed in the asphalt layer of the road structure. One such traffic counter is shown in Figure 1.

Figure 1. Automatic traffic counter based on inductive loops

The QLTC-10C counters continuously count and classify vehicles into ten categories, while QLTC-8C counters classify vehicles into eight categories. The QLTC-10C counters, classify vehicles into the following categories: A0 - Motorcycles, A1 - Passenger cars and Passenger cars with trailer, A2 - Combined vehicles and Combined vehicles with trailer, B1 - Light trucks and Light trucks with trailer, B2 – Medium heavy

trucks, B3 - Heavy goods vehicles, B4 - Heavy goods vehicles with trailer, B5 - Semi-trailer trucks, C1 - Buses, C2 - Articulated buses, X - Uncategorized (other) vehicles.

For each vehicle it detects, the counter records: date, time, direction of vehicle movement, ordinal number of the vehicle on that day for the observed direction, traffic lane, vehicle category and vehicle speed. The obtained data is stored on SD (Secure Digital) memory cards.

In this case study data used have been obtained by automatic counting of traffic on state roads in Serbia at 21 counting points (Figure 2), in the period from 1.1.2011 to 31.12.2020. The research was done on 4 sections of the road (IA category (road 1) and IB category (roads 22, 23 and 46)). Selected counting places have the following marks, i.e. names: 1025 (Kraljevo 2), 1026 (Trstenik), 1027 (Pojate), 1046 (Vodice), 1050 (Prijanovci), 1052 (Pridvorica), 1057 (Prijepolje), 1156 (Mojsinje) , 1157 (Mrčajevci), 1183 (Trupale Bg-Ni), 1191 (Ineks), 1193 (Kneževići), 1194 (Zlatibor), 1195 (Kokin Brod 2), 1196 (Nova Varoš), 1198 (Gorjani), 1202 (Međuvršje), 1207 (Prijepolje 2), 1208 (Velika Župa), 1225 (Lučina) and 1270 (Preljina).

Figure 2. Traffic counting locations

The purpose of the case study has been to predict two traffic intensity indicators: total monthly night traffic (TMNT) and average monthly night traffic (AMNT), at selected counting locations, using the method of supervised machine learning. The instances of the available data set are described by the following attributes: counter, year, month, TMNT and AMNT. The TMNT attribute represents the total number of vehicles that are registered by ABS at night (from 22.00 hours to 06.00 hours) during the period of one month. The AMNT attribute represents the average daily number of vehicles that are registered by ABS at night, on a monthly basis. In order to predict the total amount of night traffic per month models of machine learning, whose target variable is the TMNT attribute, have been created, while models whose target variable is the AMNT attribute have been created to predict the average night traffic per month. In both groups of machine learning models, the independent attributes are counter and month. The attribute year is used to classify the instances of the existing data set into two parts: for model training and for model testing. Instances relating to period

from 2011 to 2017 have been selected as a set of data for model training, while instances relating to the period from 2018 to 2020 have been used for model testing.

Training, validation and testing of machine learning models have been performed in the data mining software Weka 3.9.5. This particular software represents a collection of machine learning algorithms used in discovery operations concerning data validity (Witten et al., 2017). It enables the performance of various data mining tasks, such as: data preparation for analysis, classification, regression analysis, clustering, learning through rules of association, selection of relevant attributes and data visualization. Each of these tasks is performed in a separate graphical user interface window of Weka software (Weka Explorer) and is opened by selecting the appropriate tab of Weka Explorer (Figure 3). The Preprocess window, shown in Figure 3, allows you to load and prepare the available data set for later analysis.

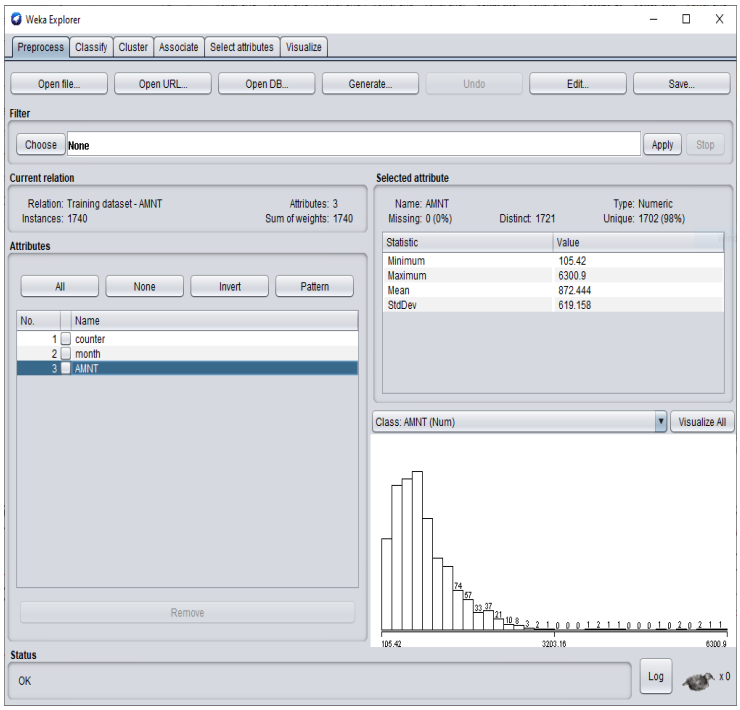

Figure 3. Weka 3.9.5 software tool graphical user interface - data preparation window

5. Results and Discussion

The following eight machine learning algorithms were to predict TMNT on the training data set in the Weka software tool: Linear Regression, Multilayer Perceptron, SMOreg, IBk (K-NN), M5P, Random Forest, Random Tree and REPTree. A 10-fold cross-validation, implemented in Weka software, has been applied to validate the model. The performance of the prediction model, measured on the training data set is shown in Table 1.

Table 1. The performance of eight TMNT prediction models measured on a training data set

Algorithm	Correlation coefficient	Mean-absolute error	Root mean-squared error	Relativ-absolute error (%)	Root relativ-squared error (%)
LinearRegression	0.6417	9718.52	14687.1	73.729	76.637
MultilayerPerceptron	0.6168	10197.2	15161.7	77.360	79.114
SMOreg	0.6373	9430.72	14931.6	71.546	77.914
IBk	0.9803	1985.10	3784.06	15.06	19.745
M5P	0.9434	4124.44	6840.50	31.29	35.694
Random Forest	0.9799	2004.84	3818.77	15.209	19.926
Random Tree	0.9803	1990.11	3784.91	15.098	19.749
REPTree	0.9701	2456.30	4650.40	18.634	24.266

Models based on Multilayer Perceptron, SMOreg algorithms, and Linear Regression have been rejected due to undoubtedly unsatisfactory performance (they had a correlation coefficient of 0.6417, 0.6168 and 0.6373, respectively). Therefore, in the next phase – in testing the machine learning model, the remaining five algorithms have been applied. The performance of these five prediction models, measured on a test data set is shown in Table 2. Comparing the metrics of the selected models, shown in Table 1 and Table 2, it is concluded that none of these models have a problem of overfitting. In addition, in all five models on the test data set, the correlation coefficient has high value.

Table 2. The performances of the top five TMNT prediction models measured on a test data set

Algorithm	Correlation coefficient	Mean-absolute error	Root mean-squared error	Relativ-absolute error (%)	Root relativ-squared error (%)
IBk	0.9391	4473.81	7373.93	32.8912	35.4526
M5P	0.8854	6238.81	10205.8	45.8673	49.0681
Random Forest	0.9382	4495.17	7438.49	33.0482	35.763
Random Tree	0.9391	4473.58	7374.1	32.8895	35.4534
REPTree	0.9303	4893.33	7833.42	35.9755	37.6618

Li & Xu (2021) propose a model for short-term traffic prediction based on the Support Vector Regression (SVR) method. The SVR method is based on the basic principles of the SVM method and is generalized for regression problems. The SVM method is implemented in Weka software called LibSVM. The SVR method in the Weka software tool is obtained by selecting the LibSVM classifier and one of its types: epsilon-SVR or nu-SVR. However, the LibSVM classifier applied to the training data set, in this case study, gave poor results (correlation coefficient: 0.0644 (epsilon-SVR) and 0.0281 (nu-SVR), respectively)). Therefore, the SVR algorithm was rejected in the first phase of this research.

In the research (Filipovska & Mahmassani, 2020) the best performance has been shown by models based on neural networks and SVM, if it is a case of class balancing. Without class balancing, the model based on a Random Forest algorithm has shown

the best results. In this case study, the neural network model (MultilayerPerceptron) was rejected in the first phase because it showed worse results than all other models (Table 1). In contrast, the Random Forest algorithm showed excellent results in this case study, along with the IBk, Random Tree, and REPTree algorithms (Table 1 and Table 2).

The visualization of the prediction results received on the test data set has revealed that the model based on the IBk algorithm (K-NN) gives the results closest to the actual values. Therefore, the model based on the IBk algorithm has been selected as the best prediction model for TMNT. This case study confirmed the results of numerous studies, such as: Zhang et al. (2013), (Zou et al., 2015) and Zheng & Su (2014), which agree that the K-NN (IBk in Weka) algorithm gives excellent results in the short-term prediction of traffic flows.

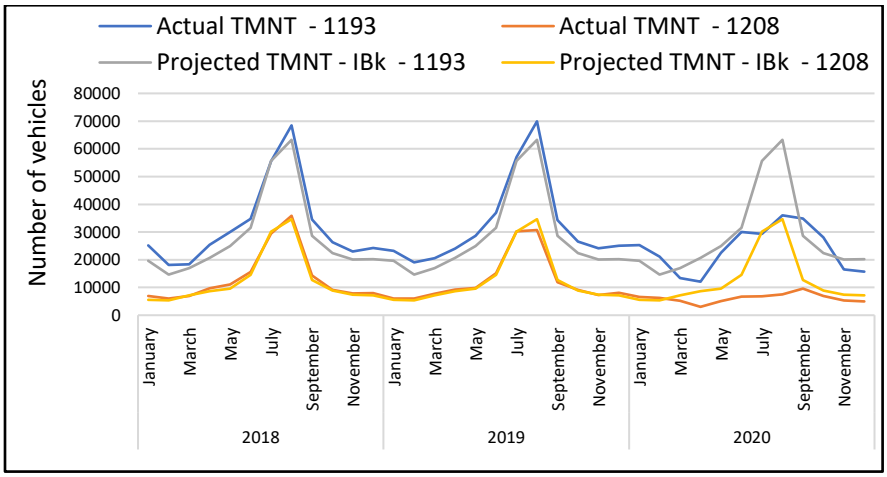

Figure 4. Actual and projected total monthly night traffic (TMNT), at selected counters (ID: 1193 and ID: 1208), for the three selected years (2018, 2019 and 2020)

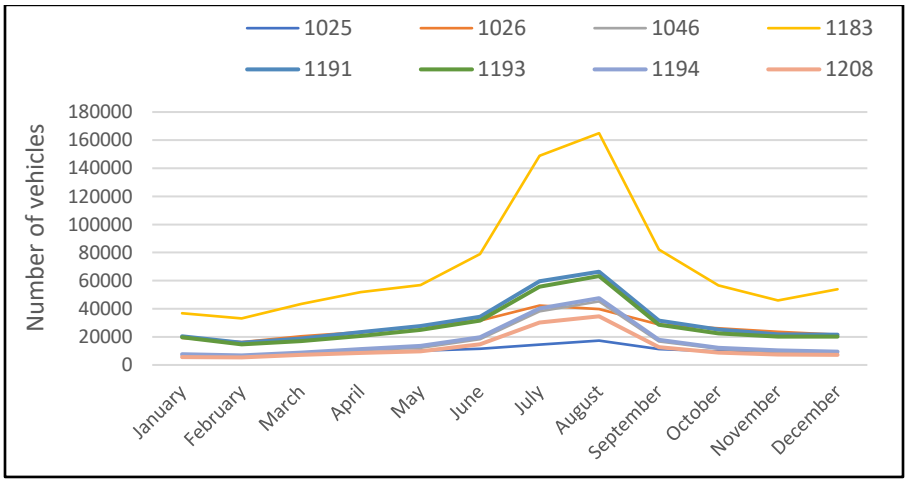

Figure 5. Projected total monthly night traffic (TMNT) at selected counters for 2021

The graph shown in Figure 4 shows the ratio of actual and projected TMNT for two selected traffic counting locations (1193 - Kneževići and 1208 - Velika Župa) and the period from 2018 to 2020. The TMNT projection has been performed using a model based on the IBk algorithm. The graph clearly shows that the TMNT prediction performed on the test data set closely follows the actual TMNT values in the observed period (Figure 4). The results of the TMNT prediction at eight selected traffic counting locations for 2021 are shown in Figure 5.

For AMNT prediction, the same eight machine learning algorithms have been applied to the training data set. The performance of the prediction models, measured on the training data set is shown in Table 3.

Table 3. The performances of eight AMNT prediction models measured on a training data set

Algorithm	Correlation coefficient	Mean-absolute error	Root mean-squared error	Relativ-absolute error (%)	Root relativ-squared error (%)
LinearRegression	0.6346	317.812	478.415	74.3936	77.2268
MultilayerPerceptron	0.608	334.371	494.495	78.2698	79.8224
SMOreg	0.6303	308.949	486.324	72.3191	78.5034
IBk	0.9801	64.9018	122.953	15.1922	19.8474
M5P	0.9445	133.975	220.096	31.3612	35.5284
Random Forest	0.9797	65.5395	124.075	15.3415	20.0285
Random Tree	0.9801	65.069	122.985	15.2314	19.8525
REPTree	0.9694	80.585	152.082	18.8634	24.5494

Models based on the Linear Regression, Multilayer Perceptron and SMOreg algorithms have been rejected due to unsatisfactory performance (correlation coefficients of 0.6346, 0.608 and 0.6303, respectively, have been recorded). Therefore, the remaining five algorithms have been applied in testing the machine learning model. The performance of these five prediction models, measured on the test data set is shown in Table 4. The best AMNT prediction model has been chosen in an identical manner as the best type of TMNT prediction model. The model based on the IBk algorithm has shown the best results this time, as well.

Table 4. The performances of the top five AMNT prediction models measured on a test data set

Algorithm	Correlation coefficient	Mean-absolute error	Root mean-squared error	Relativ-absolute error (%)	Root relativ-squared error (%)
IBk	0.939	146.428	239.814	33.1472	35.5591
M5P	0.8851	204.294	332.391	46.2465	49.2862
Random Forest	0.9381	147.128	241.896	33.3058	35.8678
Random Tree	0.939	146.420	239.819	33.1455	35.5599
REPTree	0.9286	161.299	257.180	36.5137	38.1342

The graph shown in Figure 6 shows the ratio of actual and projected AMNT for two selected traffic counting locations (1026 - Trstenik and 1046 - Vodice) and the period

from 2018 to 2020. The AMNT projection has been performed using a model based on the IBk algorithm. The results of the AMNT prediction at eight selected traffic counting locations for 2021 are shown in Figure 7.

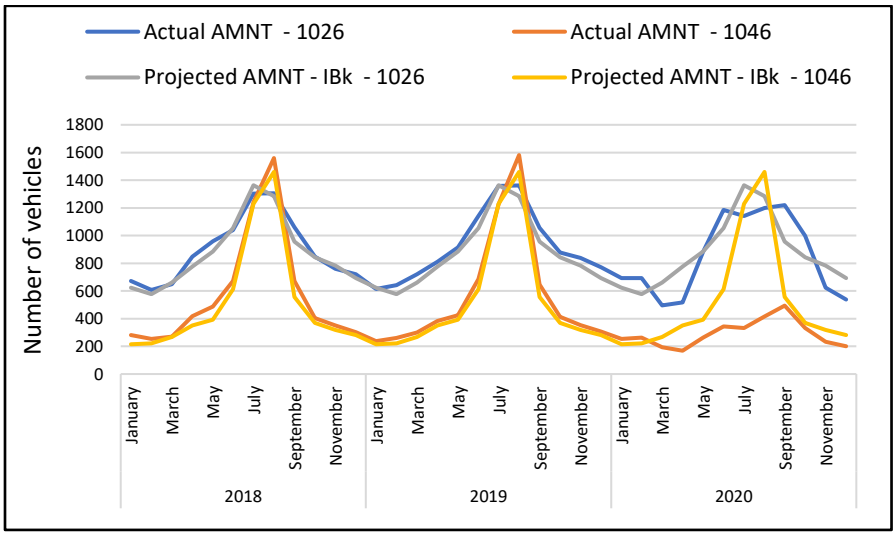

Figure 6. Actual and projected average monthly night traffic (AMNT), at selected counters (ID: 1026 and ID: 1046), for the three selected years (2018, 2019 and 2020)

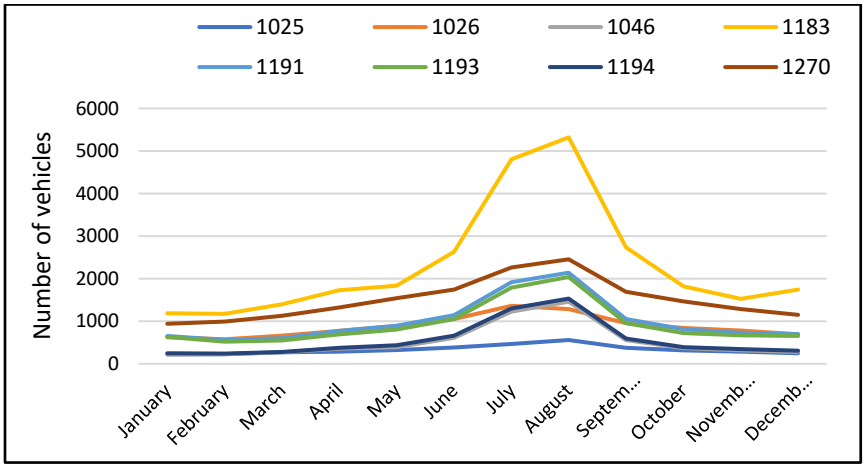

Figure 7. Projected average monthly night traffic (AMNT) at selected counters for 2021

In all the diagrams shown from Figure 4 to Figure 7, it is easy to see that the extreme values of TMNT, as well as of AMNT, occur for the months of July and August. This is because almost all counting places are located on the roads leading to popular tourist destinations, and July and August are the months when most people are on vacation and traveling.

6. Conclusion

The aim of this research has been to train and test predictive models on the existing data set on the volume of night traffic on state roads in Serbia and to predict the total and average amounts of night traffic per month for the following year.

In the conducted case study, using the Weka software tool, machine learning models for prediction of total monthly night traffic (TMNT) and average monthly night traffic (AMNT) have been trained, based on algorithms: Linear Regression, Multilayer Perceptron, SMOreg, IBk, M5P, Random Forest, Random Tree and REPTree. In the training data set, the IBk (K-NN) algorithm-based model and the models based on regression trees have shown a considerably better performance than the models from the functions category (Linear Regression, Multilayer Perceptron and SMOreg). Therefore, only these models have been tested on the test data set. The best performances have been shown by models based on the K-NN algorithm, so the prediction of TMNT and AMNT has been performed using these models. The case study has shown that the K-NN algorithm can be effectively applied in solving the problem of regression analysis of traffic data, even on relatively small data sets.

Future research will include the cluster analysis of traffic flows, especially the analysis of clusters in total and average monthly night traffic. As a result of this analysis, different patterns are expected in the volume of night traffic, on different sections of roads, at different periods of the year.

Acknowledgement: This paper has been partially supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia, within the project number 036012. The data used in the research have been provided by Public Enterprise "Roads of Serbia".

References

Aha, D. W., Kibler, D., & Albert, M. K. (1991). Instance-based learning algorithms. Machine learning, 6(1), 37-66. https://doi.org/10.1007/BF00153759

Alshaykha, A. M., & Shaban, A. I. (2021). Short-Term Traffic Flow Prediction Model Based On K-Nearest Neighbors and Deep Learning Method. Journal of Mechanical Engineering Research and Developments, 44(6), 113-122.

Boukerche, A., & Wang, J. (2020). Machine Learning-based traffic prediction models for Intelligent Transportation Systems. Computer Networks, 181, 107530. https://doi.org/10.1016/j.comnet.2020.107530

Cai, P., Wang, Y., Lu, G., Chen, P., Ding, C., Sun, J. (2016). A spatiotemporal correlative k-nearest neighbor model for shortterm traffic multistep forecasting. Transportation Research Part C: Emerging Technologies, 62, 21-34. https://doi.org/10.1016/j.trc.2015.11.002

Filipovska, M., & Mahmassani, H. S. (2020). Traffic flow breakdown prediction using machine learning approaches. Transportation research record, 2674(10), 560-570. https://doi.org/10.1177%2F0361198120934480

Guo, J., & Williams, B. M. (2010). Real-time short-term traffic speed level forecasting and uncertainty quantification using layered Kalman filters. Transportation Research Record, 2175(1), 28-37. https://doi.org/10.3141%2F2175-04

Janković, S., Zdravković, S., Mladenović, D., Mladenović, S., Uzelac, A. (2020). Traffic Volume Prediction Using Regression Decision Trees. Proceedings of the XLVII International Symposium on Operational Research - SYM-OP-IS '20, Belgrade, Serbia, 287-292.

Karlaftis, M. G., Vlahogianni, E. I. (2011). Statistical methods versus neural networks in transportation research: Differences, similarities and some insights. Transportation Research Part C: Emerging Technologies, 19(3), 387-399. https://doi.org/10.1016/j.trc.2010.10.004

Kukadapwar, S. R., & Parbat, D. K. (2015). Modeling of traffic congestion on urban road network using fuzzy inference system. American Journal of Engineering Research, 4(12), 143-148.

Li, C., & Xu, P. (2021). Application on traffic flow prediction of machine learning in intelligent transportation. Neural Computing and Applications, 33(2), 613-624. https://doi.org/10.1007/s00521-020-05002-6

Liu, Z., Guo, J., Cao, J., Wei, Y., & Huang, W. (2018). A Hybrid Short-term Traffic Flow Forecasting Method Based on Neural Networks Combined with K-Nearest Neighbor. PROMET – Traffic & Transportation, 30(4), 445–456. https://doi.org/10.7307/ptt.v30i4.2651

Lv, Y., Duan, Y., Kang, W., Li, Z., & Wang, F. Y. (2015). Traffic flow prediction with big data: a deep learning approach. IEEE Transactions on Intelligent Transportation Systems, 16(2), 865-873. https://doi.org/10.1109/TITS.2014.2345663

Ma, S. F., He, G. G., Wang, S. T. (2001). A traffic flow forecast supported system based multi-agent. Intelligent Transportation Systems. IEEE Intelligent Transportation Systems Conference Proceedings, 25-29 Aug 2001, Oakland, CA, USA, 620-624.

Magalhaes, R. P., Lettich, F., Macedo, J. A., Nardini, F. M., Perego, R., Renso, C., & Trani, R. (2021). Speed prediction in large and dynamic traffic sensor networks. Information Systems, 98, 101444. https://doi.org/10.1016/j.is.2019.101444

Marković, H., Dalbelo Bašić, B., Gold, H., Dong, F., Hirota, K. (2010). GPS Data-based Non-parametric Regression for Predicting Travel Times in Urban Traffic Networks. Promet – Traffic & Transportation, 22(1), 1-13. https://doi.org/10.7307/ptt.v22i1.159

Mohammed, O., & Kianfar, J. (2018). A Machine Learning Approach to Short-Term Traffic Flow Prediction: A Case Study of Interstate 64 in Missouri. 2018 IEEE International Smart Cities Conference (ISC2).

Pang, X., Wang, C., & Huang, G. (2016). A short-term traffic flow forecasting method based on a three-layer k-nearest neighbor non-parametric regression algorithm. Journal of Transportation Technologies, 6(4), 200-206. http://dx.doi.org/10.4236/jtts.2016.64020

Park, H., Haghani, A., Samuel, S., & Knodler, M. A. (2018). Real-time prediction and avoidance of secondary crashes under unexpected traffic congestion. Accident Analysis & Prevention, 112, 39–49. https://doi.org/10.1016/j.aap.2017.11.025

Peng, T., Tang, Z. (2015). A small scale forecasting algorithm for network traffic based on relevant local least squares support vector machine regression model. Applied Mathematics & Information Sciences, 9(2L), 653-659. http://dx.doi.org/10.12785/amis/092L41

Public Enterprise “Roads of Serbia”. (2012). Manual for road design in the Republic of Serbia. Belgrade: Public Enterprise “Roads of Serbia”.

Shamshad, N., Sarwr, D. (2020). A review of Traffic Flow Prediction Based on Machine Learning approaches. International Journal of Scientific & Engineering Research, 11(3), 126-130.

Sénquiz-Díaz, C. (2021). Transport infrastructure quality and logistics performance in exports. Economics-Innovative and Economic Research, 9(1), 107-124. https://doi.org/10.2478/eoik-2021-0008

Shankar, H., Raju, P. L. N., & Rao, K. R. M. (2012). Multi model criteria for the estimation of road traffic congestion from traffic flow information based on fuzzy logic. Journal of Transportation Technologies, 2(01), 50.

Stojčić, M. (2018). Application of ANFIS model in road traffic and transportation: a literature review from 1993 to 2018. Operational Research in Engineering Sciences: Theory and Applications, 1(1), 40-61. https://doi.org/10.31181/oresta19012010140s

Toan, T. D., & Truong, V.-H. (2020). Support Vector Machine for Short-Term Traffic Flow Prediction and Improvement of Its Model Training using Nearest Neighbor Approach. Transportation Research Record: Journal of the Transportation Research Board, 2675(4), 362–373. https://doi.org/10.1177%2F0361198120980432

Vlahogianni, E. I., Karlaftis, M. G., Golias, J. C. (2005). Optimized and meta-optimized neural networks for short-term traffic flow prediction: a genetic approach. Transportation Research Part C: Emerging Technologies, 13(3), 211-234. https://doi.org/10.1016/j.trc.2005.04.007

Vlahogianni, E. I., Karlaftis, M. G., Golias, J. C. (2007). Spatio-temporal short-term urban traffic volume forecasting using genetically optimized modular networks. Computer-Aided Civil and Infrastructure Engineering, 22(5), 317-325. https://doi.org/10.1111/j.1467-8667.2007.00488.x

Wang, J. & Shi, Q. (2013). Short-term traffic speed forecasting hybrid model based on chaos-wavelet analysis-support vector machine theory. Transportation Research Part C: Emerging Technologies, 27, 219-232. https://doi.org/10.1016/j.trc.2012.08.004

Wang, Y., Zhang, D., Liu, Y., Dai, B., & Lee, L. H. (2019). Enhancing transportation systems via deep learning: A survey. Transportation research part C: emerging technologies, 99, 144-163. https://doi.org/10.1016/j.trc.2018.12.004

Williams, B. M., Durvasula, P. K., & Brown, D. E. (1998). Urban freeway traffic flow prediction: application of seasonal autoregressive integrated moving average and exponential smoothing models. Transportation Research Record, 1644(1), 132-141. https://doi.org/10.3141%2F1644-14

Williams, B. M. & Hoel, L. A. (2003). Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: Theoretical basis and empirical results. Journal of Transportation Engineering, 129(6), 664-672. https://doi.org/10.1061/(ASCE)0733-947X(2003)129:6(664)

Witten, I. H., Frank, E., Hall, M. A. & Pal, C. J. (2017). Data Mining: Practical Machine Learning Tools and Techniques. (4th ed.). Burlington, USA: Morgan Kaufmann.

Xu, Y., Kong, Q. & Liu, Y. (2013). Short-term traffic volume prediction using classification and regression trees. Proceedings of the 2013 IEEE Intelligent Vehicles Symposium (IV), Gold Coast, Australia, 493-498.

Yasin Çodur, M., Tortum, A. (2015). An artificial neural network model for highway accident prediction: A case study of Erzurum, Turkey. Promet – Traffic & Transportation, 27(3), 217-225. https://doi.org/10.7307/ptt.v27i3.1551

Zaki, J. F., Ali-Eldin, A. M. T., Hussein, S. E., Saraya, S. F., & Areed, F. F. (2016). Framework for Traffic Congestion Prediction. International Journal of Scientific & Engineering Research, 7(5), 1205-1210. https://hdl.handle.net/1887/46907

Zhang, L., Liu, Q., Yang, W., Wei, N., & Dong, D. (2013). An Improved K-nearest Neighbor Model for Short-term Traffic Flow Prediction. Procedia - Social and Behavioral Sciences, 96, 653–662. https://doi.org/10.1016/j.sbspro.2013.08.076

Zhang, Y., & Xie, Y. (2007). Forecasting of short-term freeway volume with v-support vector machines. Transportation Research Record, 2024(1), 92-99. https://doi.org/10.3141%2F2024-11

Zhang, Y., Zhou, Y., Lu, H., & Fujita, H. (2020). Traffic Network Flow Prediction Using Parallel Training for Deep Convolutional Neural Networks on Spark Cloud. IEEE Transactions on Industrial Informatics, 16(12), 7369-7380. https://doi.org/10.1109/TII.2020.2976053

Zheng, Z. & Su, D. (2014) Short-term traffic volume forecasting: a k-nearest neighbor approach enhanced by constrained linearly sewing principle component algorithm. Transportation Research Part C: Emerging Technologies, 43, 143-157. https://doi.org/10.1016/j.trc.2014.02.009

Zou, T., He, Y., Zhang, N., Du, R., & Gao, X. (2015). Short-Time Traffic Flow Forecasting Based on the K-Nearest Neighbor Model. Fifth International Conference on Transportation Engineering - ICTE 2015. September 26–27, 2015, Dailan, China.