Materials and its Characterization
Vol: 2(3), 2023
REST Publisher; ISBN: 978-81-948459-0-4
Website: https://restpublisher.com/book-series/mc/

Car Price Prediction using Machine Learning Techniques

Nisha Sharmi, Mallika
PSN Institute of technology and science, Tirunelveli, TamilNadu, India

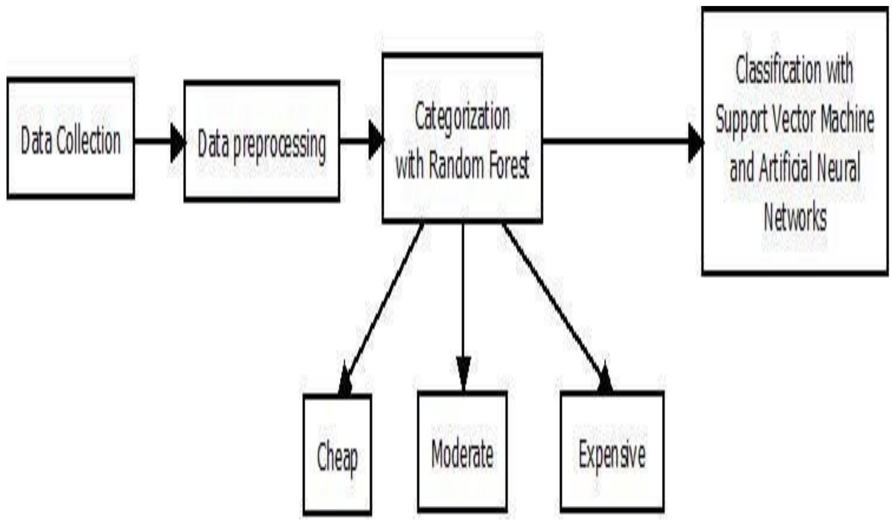
Abstract:*A car price prediction has been a high- interest research area, as it requires noticeable effort and knowledge of the field expert. Considerable number of distinct attributes is examined for the reliable and accurate prediction. To build a model for predicting the price of used cars in Bosnia and Herzegovina, we applied three machine learning techniques (Artificial Neural Network, Support Vector Machine and Random Forest). However, the mentioned techniques were applied to work as an ensemble. The data used for the prediction was collected from the web portal autopijaca.ba using web scraper that was written in PHP programming language. Respective performances of different algorithms were then compared to find one that best suits the available data set. The final prediction model was integrated into Java application. Furthermore, the model was evaluated using test data and the accuracy of 87.38% was obtained.*
Keywords:*car price prediction, support vector machines, classification, machine learning.*

INTRODUCTION

Car price prediction is somehow interesting and popular problem. As per information that was gotten from the Agency for Statistics of BiH, 921.456 vehicles were registered in 2014 from which 84% of them are cars for personal usage [1]. This number is increased by 2.7% since 2013 and it is likely that this trend will continue, and the number of cars will ncrease in future. This adds additional significance to the problem of the car price prediction.Accurate car price prediction involves expert knowledge, because price usually depends on many distinctive features and factors. Typically, most significant ones are brand and model, age, horsepower and mileage. The fuel type used in the car as well as fuel consumption per mile highly affect price of a car due to a frequent changes in the price of a fuel. Different features like exterior color, door number, type of transmission, dimensions, safety, air condition, interior, whether it has navigation or not will also influence the car price. In this paper, we applied different methods and techniques in order to achieve higher precision of the used car price prediction.This paper is organized in the following manner: Section II contains related work in the field of price prediction of used cars. In section III, the research methodology of our study is explain. Section IV elaborates various machine learning algorithms and examine their respective performances to predict the price of the used cars. Finally, in section V, a conclusion of our work are given, together with the future works plan.

RELATED WORK

Predicting price of a used cars has been studied extensively in various researches. Listian discussed, in her paper written for Master thesis [2], that regression model that was built using Support Vector Machines (SVM) can predict the price of a car that has been leased with better precision than multivariate regression or some simple multiple regression. This is on the grounds that Support Vector Machine (SVM) is better in dealing with datasets with more dimensions and it is less prone to overfitting and underfitting. The weakness of this research is that a change of simple regression with more advanced SVM regression was not shown in basic indicators like mean, variance or standard deviation.Another approach was given by Richardson in his thesis work [3]. His theory was that car producers produce more durable cars. Richardson applied multiple regression analysis and demonstrated that hybrid cars retain their value for longer time thantraditional cars. This has roots in environmental concerns about the climate and it gives higher fuel efficiency.



Wu et al. [4] conducted car price prediction study, by using neuro-fuzzy knowledge-based system. They took into consideration the following attributes: brand, year of production and type of engine. Their prediction model produced similar results as the simple regression model. Moreover, they made an expert system named ODAV (Optimal Distribution of Auction Vehicles) as there is a high demand for selling the cars at the end of the leasing year by car dealers. This system gives insights into the best prices for vehicles, as well as the location where the best price can be gained. Regression model based on k-nearest neighbor machine learning algorithm was used to predict the price of a car. This system has a tendency to be exceptionally successful since more than two million vehicles were exchanged through it [5].

Gonggie [6] proposed a model that is built using ANN (Artificial Neural Networks) for the price prediction of a used car. He considered several attributes: miles passed, estimated car life and brand. The proposed model was built so it could deal with nonlinear relations in data which was not the case with previous models that were utilizing the simple linear regression techniques. The non-linear model was able to predict prices of cars with better precision than other linear models.

Furthermore, Pudaruth [7] applied various machine learning algorithms, namely: k-nearest neighbors, multiple linear regression analysis, decision trees and naïve bayes for car price prediction in Mauritius. The dataset used to create a prediction model was collected manually from local newspapers in period less than one month, as time can have a noticeable impact on price of the car. He studied the following attributes: brand, model, cubic capacity, mileage in kilometers, production year, exterior color, transmission type and price. However, the author found out that Naive Bayes and Decision Tree were unable to predict and classify numeric values. Additionally, limited number of dataset instances could not give high classification performances, i.e. accuracies less than 70%.

Noor and Jan [8] build a model for car price prediction by using multiple linear regression. The dataset was created during the two-months period and included the following features: price, cubic capacity, exterior color, date when the ad was posted, number of ad views, power steering, mileage in kilometer, rims type, type of transmission, engine type, city, registered city, model, version, make and model year. After applying feature selection, the authors considered only engine type, price, model year and model as input features. With the given setup authors were able to achieve prediction accuracy of 98%.In the related work shown above, authors proposed prediction model based on the single machine learning algorithm. However, it is noticeable that single machine learning algorithm approach did not give remarkable prediction results and could be enhanced by assembling various machine learning methods in an ensemble.

MATERIALS AND METHODS

Approach for car price prediction proposed in this paper is composed of several steps, shown in Fig. 1.

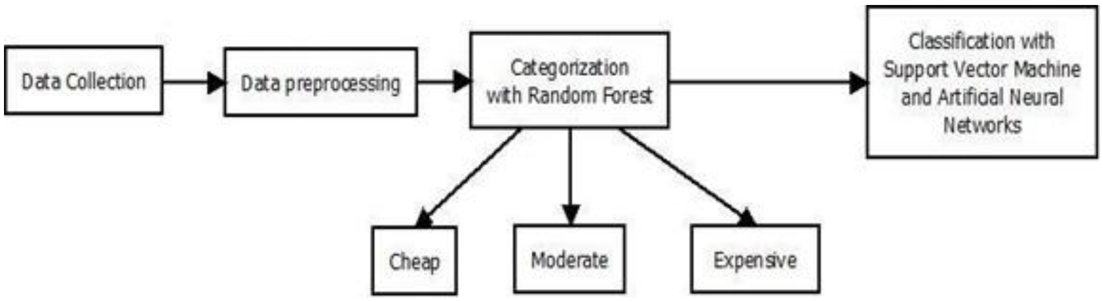


FIGURE 1. Block diagram of the overall classification process

Data is collected from a local web portal for selling and buying cars autopijaca.ba [9], during winter season, as time interval itself has high impact on the price of the cars in Bosnia and Herzegovina. The following attributes were captured for each car: brand, model, car condition, fuel, year of manufacturing, power in kilowatts, transmission type, millage, color, city, state, number of doors, four wheel drive (yes/no), damaged (yes/no), navigation (yes/no), leather seats (yes/no), alarm (yes/no), aluminum rims (yes/no), digital air condition (yes/no), parking sensors (yes/no), xenon lights (yes/no), remote unlock (yes/no), electric rear mirrors (yes/no), seat heat (yes/no), panorama roof (yes/no), cruise control (yes/no), abs (yes/no), esp (yes/no), asr (yes/no) and price expressed in BAM (Bosnian Mark).Since manual data collection is time consuming task, especially when there are numerous records to process, a "web scraper" as a part of this research is created to get this job done automatically and reduce the time for data gathering. Web scraping is well known technique to extract information from websites and save data into local file or database. Manual data extraction is time consuming and therefore web scrapers are used to do this job in a fraction of time. Web scrapers are programed for specific websites and can mimic regular users from website's point of view.After raw data has been collected and stored to local database, data preprocessing step was applied. Many of the attributes were sparse and they do notcontain useful information for prediction. Hence, it is decided to remove them from the dataset. The attributes "state", "city", and "damaged" were completely removed.

TABLE 1. Processed data set sample in CSV format

brand	model	fuel	powerink ilowatt's	year ofman	miles	leath er	cruisec ontrol	price
volkswagen	golf2	Diesel	45-55	l7	l4	no	no	0-1500
volkswagen	golf2	Gasoline	0-45	l7	l4	no	no	0-1500
ford	escort	Gasoline	45-55	l7	l1	no	no	0-1500
ford	fiesta	Gasoline	55-65	l4	l2	no	no	0-1500
mercedes-benz	190	Gasoline	45-55	l7	l4	no	no	0-1500
volkswagen	jetta	Diesel	0-45	l7	l5	no	no	0-1500
ford	focus	Gasoline	55-65	l6	l4	no	no	0-1500
fiat	punto	Diesel	65-75	l5	l4	no	no	0-1500
volkswagen	golf2	Gasoline	65-75	l7	l4	no	no	0-1500

The collected raw data set contains 1105 samples. Since data is collected using web scraper, there are many samples that have only few attributes. In order to clean these samples, PHP script that is reading scraped data from database, perform cleaning and saves the cleaned samples into CSV file. The CSV file is later used to load data into WEKA, software for building machine learning models [10].After cleanup process, the data set has been reduced to 797 samples. In particular, all brands that have less than 10 samples and where the price is higher than 60 000 BAM were removed due to the skew class problem.The color of the cars was normalized into fixed set of 15 different colors. Continuous attributes such as "millage", "year of manufacturing", "power in kilowatts" and "price" are converted into categorical values using predefined cluster intervals. The millage is converted into five distinct categories, the year ofmanufacturing has been converted into seven categories and the power in kilowatts is converted into eleven categories. The price attribute has been categorized into 15 distinct categories based on price range. These categories are shown in Table 2 and similar principle was applied to other attributes. This data transformation process converted regression prediction machine learning problem into classification problem.The whole dataset creation process is shown in the Fig. 2.

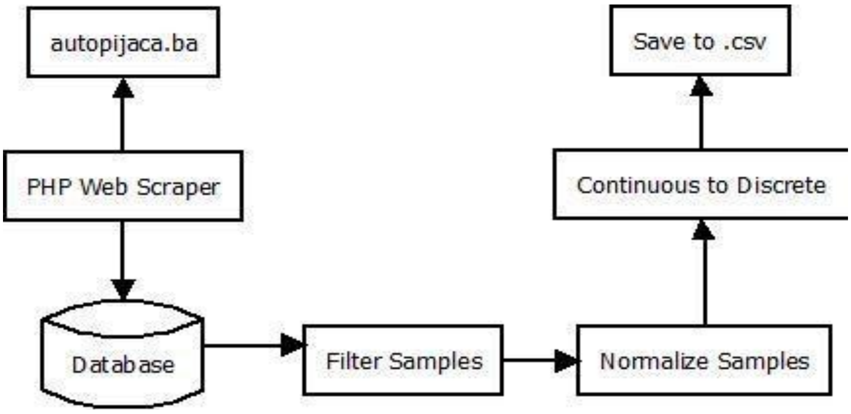


FIGURE 2. Data gathering and transformation workflow diagram

TABLE2. Priceclassificationbasedonpriceranges

From	To	Class
500	2000	500-2000
2000	3500	2000-3500
3500	5000	3500-5000
5000	6500	5000-6500
6500	8000	6500-8000
8000	9500	8000-9500
9500	11000	9500-11000
11000	14000	11000-14000
14000	17000	14000-17000
17000	20000	17000-20000
20000	25000	20000-25000
25000	30000	25000-30000
30000	60000	30000-60000

MODELIMPLEMENTATIONANDEVALUATION

Single machine learning classifier approach thathas been used in all previous researches was alsotestedinthisresearch.Thewholedatasetcollectedin this research has been split into training (90%) andtesting (10%) subsets and Artificial Neural Network,SupportVectorMachineandRandomForestclassifiersmodels were built. Randomforest(RF)alsoknownasrandomdecision forest belongs to the category of ensemblemethods.RFcanbeusedforclassificationandregression problems. The algorithm was developedbyHoasanimprovementforoverfittingofthedecisiontreealgorithms[11].ArtificialNeuralNetworks is the machine learning model that tries tosolve problems in the same way as the human braindoes. Instead of neurons, the ANN is using artificialneuronsalsoknownasperceptron.Inthehumanbrain,neuronsareconnectedwithaxonswhileinANN the weighted matrices are used for connectionsbetweenartificialneurons.Informationtravelsthroughneuronsusingconnectionsbetweenthem,fromonen euroninformationtravelstoalltheneuronsconnectedtoit.Adjustingtheweightsbetween neurons system can be trained from inputexamples [12]. Support Vector Machine can be usedforsolving classification and regression problems.Forinputdataset,theSVMcanmakeabinarydecisionanddecideinwhichamongthetwocategoriestheinputsa mplebelongs.TheSVMalgorithmistrainedtolabelinputdataintotwocategoriesthataredividedbythewidestareapossibl ebetweencategories[12].Incaseswheninput data is not labeled, SVM algorithm can not beapplied. For unlabeled data, it is necessary to applyunsupervisedlearningmethodandSVMhasitsimplementationcalledSupportVectorClustering(SVC) [13][14].

TABLE3. Singleclassifierapproachaccuracyresults

Classifier	Accuracy	Error
RF	41.18%	8.04%
ANN	42.35%	7.05%
SVM	48.23%	10.53%

ResultsshowninTable3.confirmthatsinglemachine learning classifier approach is not reliablefor prediction of car prices. Therefore, in this paperensemblemethodforcarpricespredictionwasdata set. This attribute divides cars into three pricecategories: cheap (price < 12 000 BAM), moderate(12000BAM<=price<24000BAM)andexpensive(24000 BAM<=price).Ensemble method combines three machinelearningalgorithmsthatwereappliedinthefirstexperiment as single classifiers: RF, SVM, and ANN.RandomForestalgorithmwasappliedonthewholedataset,totesthowaccuratelytheclassifiercancategorizesamplesintocheap,moderateandexpensivecarclasses.RFisametaestimatorthatfitsanumberofdecisiontreeclassifiersonvarioussub-samplesofthedatasetanduseaveragingtoimprovethe predictive accuracy and control over-fitting [15].Thefollowingfeatureswereusedtobuildmodel:brand,model,carcondition,fuel,age,kilowatts,transmission, miles, color, doors, drive, leather seats,navigation,alarm,aluminumrims,digitalAC,manual AC, parking sensors, xenon, remote unlock,seat heat, panorama roof, cruise control, abs, asr, espandprice.Before model training step, numeric attribute pricewasconvertedinto nominalclasses showninTable4.

TABLE4. Nominalcategoriesofcarpriceattribute

From	To	Class
0	12000	cheap
12000	24000	moderate
24000	….	expensive

Then,RFclassifierisapplied,andresultsareobtained(Table 5.).Table5.ClassificationresultswithRFclassifier

Typeof evaluation	of correctlyclassified
Cross validationwith10folds	85.82
90% percentagesplit	88.75

Bothclassifiers,SVMandANNarefurtherappliedtoeachpricecategorydataset:cheap,moderateand expensive carsdatasets.

ApplyingclassificationoncheapdatasetusingSVMand ANN algorithms

Cheap dataset was divided into 2 nominal classes,shown inTable 6.

TABLE6. NominalclassesinCheapdataset

From	To	Class
0	6000	0-6000
6000	12000	6000-12000

proposed. To apply ensemble of machine learningclassifiers a new attribute “price rank” with values:cheap,moderateand expensivehasbeenadded totheIn total, 230 samples of Cheap dataset were inputtoSVM and ANNalgorithms.After running SVM and ANN on given dataset,followingresults were obtained:

TABLE7. AccuracyresultsforSVMandANNonCheapdataset

Typeof evaluation	SVM	ANN
validation with10 folds	86.96	83.91
90%percentagesplit	86.96	73.91

ApplyingClassificationonModeratedatasetusingSVMandANN algorithms

ThemodelisfurthertrainedontheModeratedataset.For thispurpose, attribute price isrankedinto2 classes,showninTable 8.

TABLE8. NominalclassesinModeratedataset

From	To	Class
12000	15000	12000-18000
18000	21000	18000-24000

After applying Multilayer Perceptron algorithm ondataset,we got the followingresults

TABLE 9. Accuracy results for SVM and ANN on Moderatedataset

Typeof evaluation	SVM	ANN
Crossvalidation with10folds	78.65	76.41
90%percentage split	83.33	86.11

ApplyingClassificationonExpensivedatasetusingSVMalgorithm

As for the previous datasets, the model is trainedontheExpensivedataset.Forthispurpose,theattributepriceis groupedinto2 classes.

TABLE10. Nominalclassesfor Expensivedataset

From	To	Class
24000	28000	24000-32000
32000	36000	32000-...

TABLE 11. Accuracy results for SVM and ANN onExpensivedataset

Typeof evaluation	SVM	ANN
validation with10 folds	**79.72**	**75**
90%percentagesplit	**90.48**	**85.71**

After models are built, they have been assembledinto the final prediction system, shown in Fig. 3. Forthecaseof90%datasetsplit,SVMachievedthehighestaccuracyinCheapandExpensivesubsets,whileANNperform edbetter inModeratesubsetThe final prediction system has been incorporatedinto the Java swing GUI application for the car priceprediction.ThesimpleapplicationGUI,showninFig. 4. enables potential car buyers to estimate thepriceofthe desired car.Theproposedpredictionmodelhasbeenevaluatedonthetestsubsetandmodelachievedoverallaccuracyof87.38%.T hisprovesthatcombination of multiple machine learning classifiersstrengthenstheclassificationperformanceoverall.

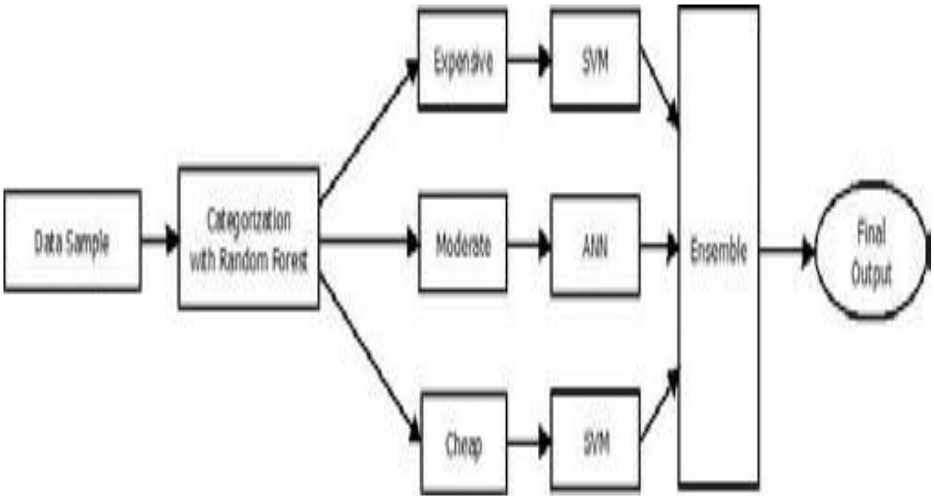


FIGURE 3. Prediction model for 90% split case

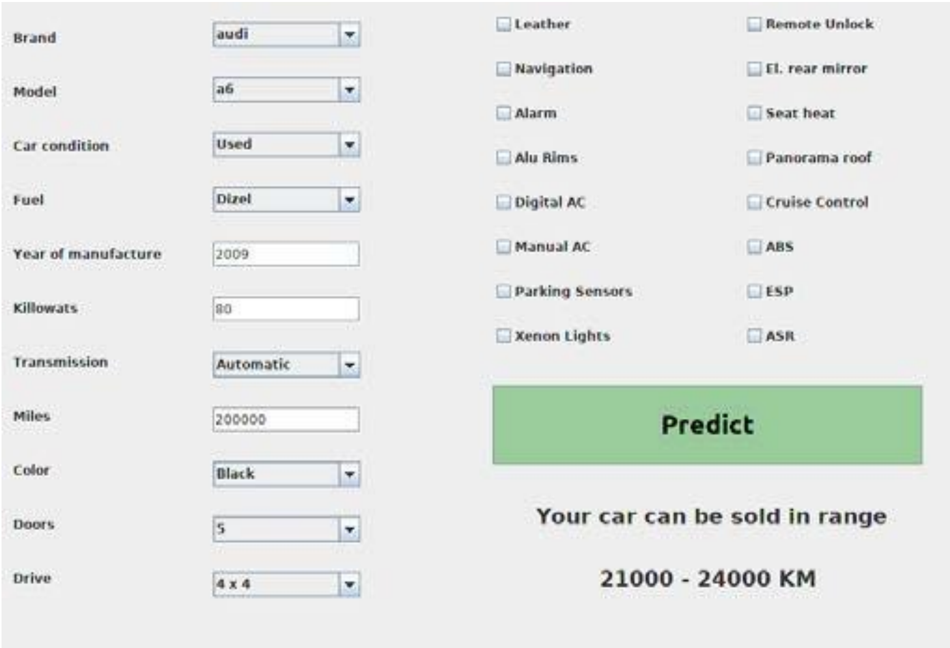


FIGURE 4. Graphical user interface of the Java applicationforcarprice prediction

SVMandANNalgorithmsarefurtherappliedtoExpensivedataset andresultsare obtained.

CONCLUSION

Car price prediction can be a challenging task duetothehighnumberofattributesthatshouldbeconsidered for the accurate prediction. The major stepinthepredictionprocessiscollectionandpreprocessingofthedata.Inthisresearch,PHPscripts were built to normalize, standardize and cleandata to avoid unnecessary noise for machine learningalgorithms.Datacleaningisoneoftheprocessesthatincreases prediction performance, yet insufficient forthe casesofcomplexdatasets as the one in thisresearch. Applying single machine algorithm on thedata set accuracy was less than 50%. Therefore, theensembleofmultiplemachine learning algorithmshasbeenproposedandthiscombinationofMLmethodsgainsaccuracyof92.38%.Thisissignificant improvement compared to single machinelearningmethodapproach.However,thedrawbackoftheproposedsystem isthatitconsumesmuchmore computational resources than single machinelearningalgorithm.Although, this system has achieved astonishingperformance in car price prediction problem our aimfor the future research is to test this system to worksuccessfully with various data sets. We will extendour test data with eBay [16] and OLX [17] used carsdatasets and validatethe proposedapproach.

REFERENCES

[1]. Agencija za statistiku BiH. (n.d.), retrieved from: http://www.bhas.ba . [accessed July 18, 2018.]
[2]. Listiani, M. (2009). Support vector regression analysis for price prediction in a car leasing application (Doctoral dissertation, Master thesis, TU Hamburg-Harburg).
[3]. Richardson, M. S. (2009). Determinants of used car resale value. Retrieved from: https://digitalcc.coloradocollege.edu/islandora/object
[4]. /coccc%3A1346 [accessed: August 1, 2018.]
[5]. Wu, J. D., Hsu, C. C., & Chen, H. C. (2009). An expert system of price forecasting for used cars using adaptive neuro-fuzzy inference. Expert Systems with Applications, 36(4), 7809-7817.
[6]. [5] Du, J., Xie, L., & Schroeder, S. (2009). Practice Prize Paper—PIN Optimal Distribution of Auction Vehicles System: Applying Price Forecasting, Elasticity Estimation, and Genetic Algorithms to Used-Vehicle Distribution. Marketing Science, 28(4), 637-644.
[7]. Gongqi, S., Yansong, W., & Qiang, Z. (2011, January). New Model for Residual Value Prediction of the Used Car Based on BP Neural Network and Nonlinear Curve Fit. In Measuring Technology and Mechatronics Automation (ICMTMA), 2011 Third International Conference on (Vol. 2, pp. 682-685). IEEE.
[8]. Pudaruth, S. (2014). Predicting the price of used cars using machine learning techniques. Int. J. Inf. Comput. Technol, 4(7), 753-764.
[9]. Noor, K., & Jan, S. (2017). Vehicle Price Prediction System using Machine Learning Techniques. International Journal of Computer Applications, 167(9), 27-31.
[10]. Auto pijaca BiH. (n.d.), Retrieved from: https://www.autopijaca.ba. [accessed August 10, 2018].
[11]. Weka 3 - Data Mining with Open Source Machine Learning Software in Java. (n.d.), Retrieved from: https://www.cs.waikato.ac.nz/ml/weka/. [August 04, 2018].
[12]. Ho, T. K. (1995, August). Random decision forests. In Document analysis and recognition, 1995., proceedings of the third international conference on (Vol. 1, pp. 278-282). IEEE.
[13]. Russell, S. (2015). Artificial Intelligence: A Modern Approach (3rd edition). PE.
[14]. Ben-Hur, A., Horn, D., Siegelmann, H. T., &Vapnik,
[15]. V. (2001). Support vector clustering. Journal of machine learning research, 2(Dec), 125-137.
[16]. Aizerman, M. A. (1964). Theoretical foundations of the potential function method in pattern recognition learning. Automation and remote control, 25, 821- 837.
[17]. 3.2.4.3.1. sklearn.ensemble.RandomForestClassifier scikit-learn 0.19.2 documentation. (n.d.). Retrieved from: http://scikit- learn.org/stable/modules/generated/sklearn.ensemble .RandomForestClassifier.html [accessed: August 30, 2018]. Used cars database. (n.d.) Retrieved from: https://www.kaggle.com/orgesleka/used-cars- database. [accessed: June 04, 2018]. OLX. (n.d.), Retrieved from: https://olx.ba. [accessed August 05,2018].