9th International scientific conference
Technics and Informatics in Education – TIE 2022
16-18 September 2022

Session: *IT Education and Practice*
Original research paper
DOI: 10.46793/TIE22.237K

Comparison of regression methods and tools using the example of predicting the success of graduate master's students in different fields of education

Katarina Karić*, Andrijana Gaborović, Marija Blagojević, Danijela Milošević, Katarina Mitrović and Jelena Plašić
University of Kragujevac, Faculty of Technical Sciences, Čačak, Serbia
* katarina.karic@ftn.kg.ac.rs

Abstract: *With the rapid development of ICT, the fields of Artificial Intelligence and Machine Learning and data mining techniques, there is a need for research in which they are applied, in various domains. In this paper, the analysis of the data set was conducted using regression methods, as one of the "Data mining" and prediction techniques, in order to predict further development in the future, ie. number of graduate master's students in all fields of education. The aim of this research is to monitor the current number of students and compare them with the previous one - in academic education of the second degree, in order to predict the number of students annually and possible factors affecting academic university education in the Republic of Serbia. The obtained results related to the number of master's degree students in the field of education in all territorial parts of the Republic of Serbia, may, also indicate the implementation of certain reforms in academic education in the future, adding innovative ideas, student exchange and others.*

Keywords: *regression; data mining; master studies; education*

1. INTRODUCTION

Artificial intelligence represents a way of reasoning and acting on derived conclusions, with the application of logic, whereby reasoning and acting is not carried out by man or any other living organism, but by machines in the broadest sense of the word [1]. Machine learning, as a branch of artificial intelligence, deals with techniques and methods that enable computer systems, ie. machines learn from experience, ie to react to changes in the external environment, without explicit programming [2]. One of the important applications of machine learning is in data research, ie in the field of "Data mining". "Data mining" is a process of "mining" large databases and extracting new and useful information that can contribute to better and more successful business

[3].

Through research in the field of application of regression methods [4], an adequate way of using these methods is presented, as well as the goal of implementing these methods in the future. The techniques used in this study were simple linear regression and multiple linear regression. The authors [4] came to the conclusion that the percentage of reliability using these methods is about 95%, which is extremely important, but that more predictive analyzes should be performed in future work, such as: logistic regression, decision trees, neural networks and dr.

The paper [5] focuses on the implementation of the regression method in a case study where it was shown that regression algorithms for prediction of outgoing traffic and a model based on the "decision tree" algorithm give the best results, while other algorithms have a problem of excessive matching.

The authors [6] conducted research in the field of education, ie research in which linear regression methods, decision trees, SVR ("Support Vector Regression") and "Random Forest" algorithms are applied in order to enable postgraduate students the most reliable and efficient choice of university where he will attend master's studies. The results obtained in this research are given on the basis of student profiles, while students would not decide to conduct similar analyzes, first decide on the basis of consultants' programs and previous admissions, which is not the most reliable and personalized solution.

Similar research [7] applies the linear regression method to predict student academic performance, which aims to help instructors develop a good understanding of how well or poorly students in their classes will adapt and master material from mechanics and dynamics. Based on the results

obtained, instructors can adopt proactive measures to improve student learning.

The results of the previously presented related research indicate the need to apply regression methods and other methods for the purpose of forecasting in order to make safe decisions in the future, but also to point out possible mistakes in business, work and the like.

The aim of this paper is to analyze a set of data on the topic of graduate students of second degree (master students) in the fields of education. Data analysis was conducted using regression methods, as one of the techniques of machine learning and the task of prediction, in order to predict further development in the future, ie. number of graduate master's students in all fields of education.

In the following chapters, data mining techniques will be explained in detail, with an emphasis on regression methods and their application.

2. DATA MINING TECHNIQUES

Data mining is an extremely useful methodology, which aims to obtain information from a multitude of data that is crucial for strategic decision-making. As a systematic, interactive and iterative process of data and information analysis, it enables better business decision-making and management whose main area of application is business [8]. In order to extract the obtained information from a huge amount of data, it is necessary to apply certain techniques [9].

The following is a list of all Data Mining techniques [10]:

- Classification - This technique is used to classify data into different classes according to certain criteria, such as: according to the type of data source, according to the included database, according to the type of knowledge discovered, etc.
- Clustering - Clustering is the division of information into groups of related objects, ie. this technique represents the grouping of data based on their similarities.Regression - Regression analysis is a predictive data mining technique used to identify and analyze the relationships between variables due to the presence of another factor. Used to define the probability of a particular variable. Regression is a method that primarily represents a form of planning and modeling.
- Association Rules - This data search technique helps detect a connection between two or more items.
- Outer detection - A technique that refers to observing data items in a data set that do not match the expected pattern or expected behavior.
- Sequential Patterns - A data mining technique specialized for estimating sequential data to detect sequential patterns, ie. similar patterns.
- Prediction - A technique that uses a combination of other data mining techniques, such as regression, clustering, classification, etc. to analyze past events and phenomena, in order to predict future events and happenings.

3. REGRESSION TECHNIQUE AND REGRESSION METHODS

Predictive analytics includes a number of techniques of statistics and data mining, which analyzes current and historical facts, in order to predict future events. Regression analysis can be used to model the relationship between one or more independent variables and dependent variables. The regression technique consists of methods on the basis of which data analysis can be performed, as follows: Linear regression, Simple linear regression, Multiple linear regression, Nonlinear and multiple nonlinear regression, Logistic regression, Decision tree, etc. [11].

3.1. Linear and simple linear regression

Linear regression is one of the regression methods used to predict results. Model the relationship between an independent variable and a dependent variable. The simplest form of regression is a simple linear regression that contains only one predictor. The relationship between the input and output variables can be mapped to two-dimensional space. Simple linear regression can be applied using several different methods, and one of the chosen ones is the least squares method. This method is a form of mathematical regression analysis that is used to determine the line that best suits the data set, providing a visual representation of the relationship between data points, ie. the relationship between a known independent variable and an unknown dependent variable [6].

3.2. Nonlinear regression

Nonlinear regression models approximate the relationship between dependent and independent variables by a nonlinear function. The data consists of independent variables that do not contain errors - x, and related experimental dependent variables - y. Each value of y is modeled as a random variable with the average given in the form of a nonlinear function f (x, β).

3.3. Multiple linear regression

Multiple linear regression represents the relationship between two or more variable explanations and the response variable by adjusting the linear equation over the observed data. Each value of the independent variable X is related to the value of the dependent variable Y [12]: $y = a1x1 + a2xa + a3x3 + ... + anxn$.

The main goal of multiple regression is to discover as many independent variables as possible that affect the dependent variable. The degree of correlation between variables, ie correlation analysis, is extremely important for this method of regression and provides information such as the relative importance of each independent variable in predicting or influencing the dependent variable and the degree to which all independent variables explain dependent variable variations.

4. RESEARCH METHODOLOGY

The main research problem in this paper is the comparison of regression methods and software using the example of predicting the success of graduate master's students in different fields of education. The number of second-degree graduates can vary by field of education as well as by year. As the regression analysis obtained the prediction and results of future development and flow, it was conducted in three different matre for each of the regression methods, namely:

- Linear and simple linear regression was performed in "SPSS" and "RapidMiner" software,
- Nonlinear regression was performed in the "NCSS" software and
- Multiple linear (complex) in "NCSS" software.

"SPSS" is an IBM product designed for statistical analysis, predictive analysis, text analysis, open source extensibility, big data integration, and offers a vast library of machine learning algorithms [13].

"NCSS" software is intended for statistics, graphics and sample size. It is dedicated to providing services to researchers, researchers, academics, scientists and other professionals [14].

RapidMiner is an open source software platform that provides an integrated machine learning environment, Data Mining, Text Mining, and business analytics. It is used for business and commercial purposes, as well as for research, education, training, and supports all steps of the data mining process, including data preparation, visualization, validation and optimization of results [15].

In the analysis, a set of data named "Number of graduates by fields of education", which was taken from the open data portal [16] was used. The downloaded data set is in .xlsx format. It stores data collected in the period from 2016 to 2019. The pre-transformation data set contains information given in only eight columns: "indicator", "IDTer", "nTer", "mes", "god", "IDISCEDF", "nISCEDF" and "value".

By transforming the data set, the "indicator" column was renamed "id", the "nTer" column was renamed "Territory" and the "year" column was renamed "year". As the values of the "nISCEDF" column for one instance in the table were dates in several rows with certain numerical values, it was necessary to transform the data set so that instead of that one column "nISCEDF" create twelve different columns with date values under the following names"Total", "Generic programs and qualifications", "Education", "Arts and humanities", "Social sciences, journalism and information", "Business, administration and law", "Natural sciences, mathematics and statistics", "Information and Communication Technologies (ICT) ","" Engineering, Production and Construction ","" Agriculture, Forestry, Fisheries and Veterinary Medicine ","" Health and Social Assistance "and the" Services "column. These twelve columns represent the fields of education of students, while the values represent the number of graduates for that period (year) and a certain territory in a given field.

Figure 1. *Demonstration sample from the data set*

id	IDTer	Територија	година	UKUPNO	Генерички програми и квалификације	Образовање	Уметности и хуманистичке науке	Друштвене науке, новинарство и информисање	Пословање, администрација и право	Природне науке, математика и статистика	Информационе и комуникационе технологије (ИКТ)	Инжењерство, производња и грађевинарство	Пољопривреда, шумарство, рибарство и ветерина	Здравство и социјална помоћ	Услуге
1040202021ND01	RS	РЕПУБЛИКА СРБИЈА	2016	14841	0	1352	1525	1086	2583	688	633	2959	399	2545	1071
1040202021ND01	RS1	СРБИЈА – СЕВЕР	2016	11352	0	847	1298	862	1997	536	474	2461	339	1795	743
1040202021ND01	RS11	Београдски регион	2016	7591	0	374	851	591	1386	345	377	1516	253	1396	502
1040202021ND01	RS12	Регион Војводине	2016	3761	0	473	447	271	611	191	97	945	86	399	241
1040202021ND01	RS2	СРБИЈА – ЈУГ	2016	3489	0	505	227	224	586	152	159	498	60	750	328
1040202021ND01	RS21	Регион Шумадије и Западне Србије	2016	1593	0	327	108	33	254	68	109	220	25	377	72
1040202021ND01	RS22	Регион Јужне и Источне Србије	2016	1896	0	178	119	191	332	84	50	278	35	373	256
1040202021ND01	RS	РЕПУБЛИКА СРБИЈА	2017	14122	0	1399	1325	983	2265	877	567	2746	351	2489	1120
1040202021ND01	RS1	СРБИЈА – СЕВЕР	2017	10769	0	915	1067	798	1711	702	426	2129	308	1887	826
1040202021ND01	RS11	Београдски регион	2017	7498	0	487	786	547	1174	491	315	1447	231	1401	619
1040202021ND01	RS12	Регион Војводине	2017	3271	0	428	281	251	537	211	111	682	77	486	207
1040202021ND01	RS2	СРБИЈА – ЈУГ	2017	3353	0	484	258	185	554	175	141	617	43	602	294
1040202021ND01	RS21	Регион Шумадије и Западне Србије	2017	1417	0	271	116	30	218	76	72	252	16	290	76
1040202021ND01	RS22	Регион Јужне и Источне Србије	2017	1936	0	213	142	155	336	99	69	365	27	312	218
1040202021ND01	RS	РЕПУБЛИКА СРБИЈА	2019	11889	0	908	1153	752	2083	756	692	2191	321	2222	811
1040202021ND01	RS1	СРБИЈА – СЕВЕР	2019	9353	0	568	981	604	1662	586	552	1722	311	1742	625
1040202021ND01	RS11	Београдски регион	2019	6541	0	358	696	418	1228	429	329	1240	205	1233	405
1040202021ND01	RS12	Регион Војводине	2019	2812	0	210	285	186	434	157	223	482	106	509	220
1040202021ND01	RS2	СРБИЈА – ЈУГ	2019	2536	0	340	172	148	421	170	140	469	10	480	186
1040202021ND01	RS21	Регион Шумадије и Западне Србије	2019	1047	0	208	82	43	181	72	55	221	10	153	22
1040202021ND01	RS22	Регион Јужне и Источне Србије	2019	1489	0	132	90	105	240	98	85	248	0	327	164

Columns "mes" and "IDISCEDF" were not of great importance for this type of analysis, so they were removed. By creating new columns, the number of rows is reduced, and numeric values are added to empty fields in order to reduce the possibility of problems during analysis.

The created separate columns enable the achievement of better results, because they leave the possibility of analyzing the data according to several dependent and independent variables, ie. variables (X and Y). Figure 1 shows the data prepared for regression analysis. Regression analysis by linear simple, multiple and nonlinear in this paper, were conducted to predict the number of graduate master's students in the different education fields.

5. RESULTS

The results obtained using linear, multiple linear and nonlinear regression methods are presented in this chapter.

5.1. Results of linear regression analysis

The results of linear regression analysis that illustrate the prediction of the number of graduate master's students in the field of Health care and social assistance in relation to years are shown in Figure 2. The graph depicts the movement of the number of graduate master's students in the field of Health care and social assistance in relation to the year from the period 2016 to 2019. The independent variable X is defined as "years", while the dependent variable Y is defined as "Health care and social assistance" which represents the number of graduate master's students in this field. It can be concluded that linear regression predicts a reduction in the number of graduate students of the second degree in the previously mentioned field based on the period from 2016 to 2019.

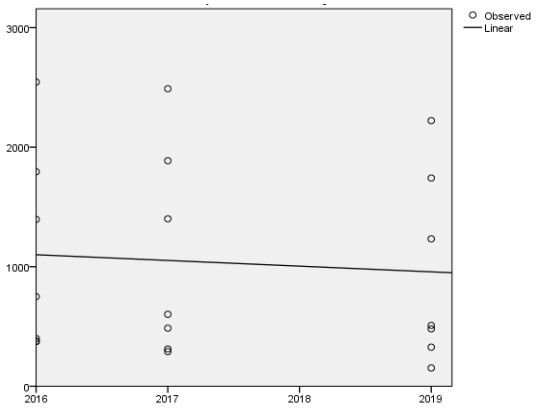

Figure 2. *Number of graduate master's students in the field of Health care and social assistance in relation to the year*

A graph that presents the movement of the number of graduate master's students in the field of ICT in relation to the years is shown in Figure 3. The independent variable X is defined as "years", while the dependent variable Y is defined as "ICT" which represents the number of graduate master's students in this field. It can be concluded that linear regression predicts a reduction in the number of graduate students of the second degree in the ICT field based on the period from 2016 to 2019.

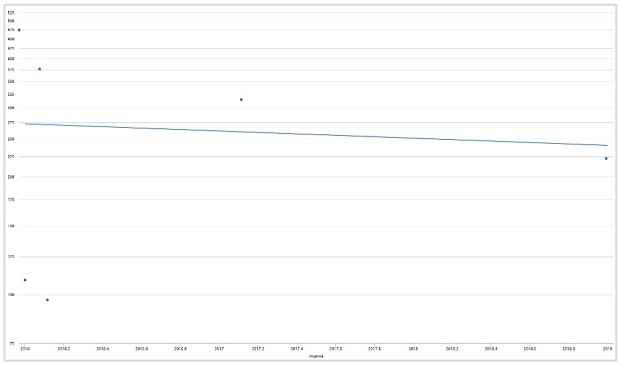

Figure 3. *Number of graduate master's students in the field of ICT in relation to the year*

A graph showing the movement of the total number of graduate master's students in relation to years is presented in Figure 4. The independent variable X is defined as "years", while the dependent variable Y is defined as "total". It can be concluded that linear regression predicts a reduction in the total number of graduate students of the second degree based on the period from 2016 to 2019. It should be noted that the decrease is more pronounced in the field of Health care and social assistance than in the field of ICT. This could be caused by the rapid development of information and telecommunication technologies which led to changes in society and thus affects student choices.

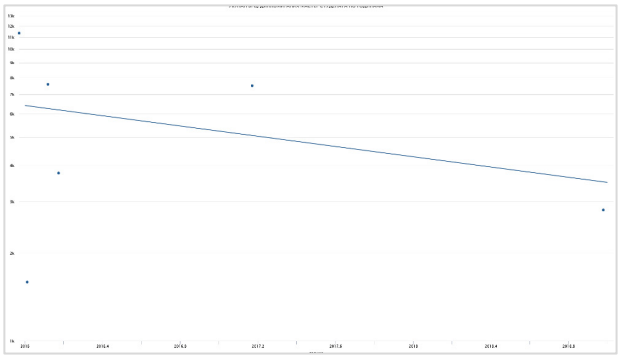

Figure 4. *Number of graduate master's students in relation to the year*

5.2. Results of nonlinear regression analysis

Nonlinear regression analysis was conducted using the NCSS tool. The independent variable X is defined as "ICT" which represents the number of graduate master's students in this field, while the dependent variable Y is defined as "years". This analysis was conducted for 2016 and 2017 separately.

A graphical representation of the movement of the number of graduates in the ICT field during 2016 is shown in Figure 5. Nonlinear regression predicts an

increase in the number of graduate master's students in the ICT field of over 90%, i.e. up to 200 students per year.

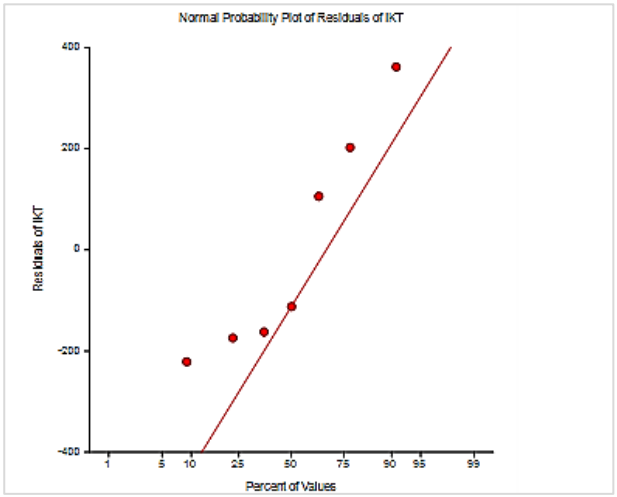

Figure 5. *Number of graduate master's students in the field of ICT during 2016*

A graphical representation of the movement of the number of graduates in the ICT field during 2017 is shown in Figure 6. Nonlinear regression predicts an increase in the number of graduate master's students in the ICT field of over 90%, i.e. up to 100 students per year. The growth trend compared to 2016 has declined which is a result of an overall reduction in the number of graduates.

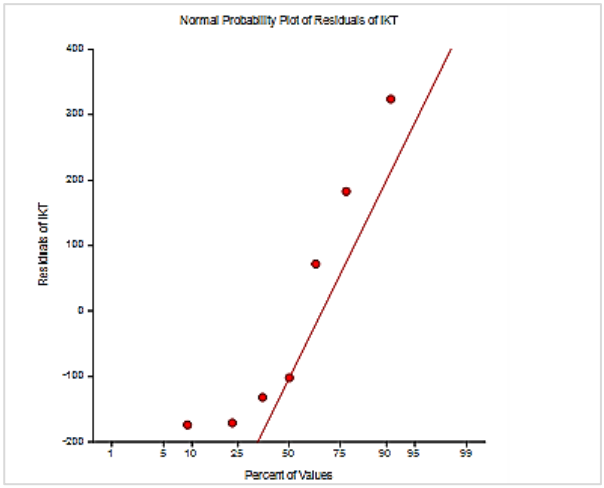

Figure 6. *Number of graduate master's students in the field of ICT during 2017*

5.3. Results of multiple linear regression analysis

The prerequisite for the analysis using multiple linear regression is to select two independent variables X, namely the number of graduates in the ICT field and the number of graduates in the field of Health care and social assistance, as well as one dependent variable Y, i.e. years. Based on the data in the period from 2016 to 2019, multiple linear regression predicts an increase in both variables, as presented in Figure 7. The results show the growth of the number of graduates in the previously mentioned fields with a slight stagnation in 2019 and in future years. As in the introductory part, an overview of related research is given, after the analysis in this paper, there is a need to compare the methods, data, results obtained in this research with related research. According to research [4], the authors concluded that the most reliable results were obtained using multiple regression methods using SPSS software, predicting growth in business profits in 95% of cases, while the results in the example (Figure 7) predict growth in 90% of cases.

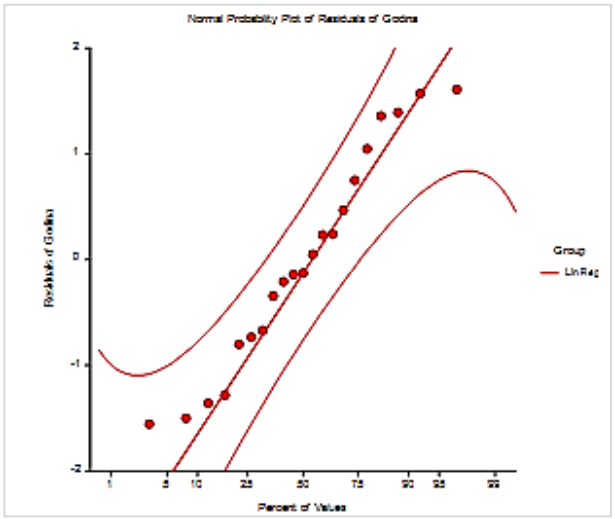

Figure 7. *Number of graduates in the ICT and Health care and social assistance fields in relation to the year*

6. CONCLUSION

With the rapid development of technology, numerous software is being developed that enables the solution of potential problems in various spheres of life. Thanks to the already mentioned: "NCSS", "RapidMiner" and "SPSS" software, which were used for the purpose of this research, and various regression methods, it is possible to predict further trends, such as number of graduate master's students in different fields, according to this research, but also potential problems and risks that can be thus solved (if it is about another field). The results obtained by the method of simple linear regression through "SPSS" software indicate that the prediction is excellent with an error percentage of ±10%. In comparison with the results obtained by the "RapidMiner" software, the same prediction was also achieved, however, the only drawback is that the error percentage is not specified. The prediction made by the non-linear regression method came to an identical prediction - which indicates a general decrease in the number of students in the fields of education. When it comes to the multiple linear regression method using "NCSS" software, the results showed a general growth of the number of graduates in the previously mentioned fields with a slight stagnation in 2019 and in future years. As the authors [7]

state, a good prediction is defined as one with a prediction error of ± 10%, which was the case with them and in this work. The research [5] uses linear regression methods (which proved to be an exceptional prediction method) as well as the decision tree method - which from the perspective of the future course of research in this work, may be a method that will also be implemented.

This research itself found that the number of graduate master's students in the different fields of education is generally decreasing in all territorial parts of the Republic of Serbia, which, unfortunately, does not give a positive outcome, but also indicate to the fact that in the future certain measures could be implemented in the form of possible reforms in academic education, adding innovative ideas, student exchange, etc.

The future flow of research could be reflected in further data collection and monitoring of the current situation and comparison with the previous, in academic education of the second degree (postgraduate education), in order to predict annually what are the factors influencing this area of research, ie. to academic university (postgraduate) education in the Republic of Serbia.

ACKNOWLEDGEMENTS

This study was supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia, and these results are parts of the Grant No. 451-03-68/2022-14/200132 with University of Kragujevac - Faculty of Technical Sciences Čačak.

REFERENCES

[1] Branković, S. (2017). *Artificial intelligence and society*. Serbian political thought. 56. 13-32. 10.22182/spm.5622017.1.

[2] Bell, J. (2014.). *Machine learning: Hands-on for developers and technical professionals*. John Wiley & Sons

[3] Blagojević, M. (2010). *Application of web mining in education.* 3rd International Conference on Technology and Informatics in Education, Čačak, Serbia, retrieved from: http://www.ftn.kg.ac.rs/konferencije/tio2010/PDF/RADOVI/527%20Blagojevic%20-%20Primena%20VEB%20majninga%20u%20obrazovanju.pdf

[4] Halili, F., & Rustemi, A. (2016). *Predictive modeling: data mining regression technique applied in a prototype.* International Journal of Computer Science and Mobile Computing, 5(8), 207-215, retrieved from: https://www.researchgate.net/publication/329018412_Predictive_Modeling_Data_Mining_Regression_Technique_Applied_in_a_Prototype

[5] Janković, S., Kukić, K., Uzelac, A., & Maraš, V. (2019). *Traffic prediction in the local computer network using supervised machine learning*. XXXVII Symposium on New Technologies in Postal Telecommunication Traffic – PosTel 2019, 3 - 4 December 2019., Serbia: Belgrade, retrieved from: http://postel.sf.bg.ac.rs/simpozijumi/POSTEL2019/RADOVI%20PDF/Telekomunikacioni%20saobracaj,%20mreze%20i%20servisi/

[6] Acharya, M. S., Armaan, A., & Antony, A. S. (2019). *A comparison of regression models for prediction of graduate admissions*. In 2019 International Conference on Computational Intelligence in Data Science (ICCIDS), 21 – 23 Feb. 2019., India: Chennai, (pp. 1-5), https://doi.org/10.1109/ICCIDS.2019.8862140

[7] Huang, S., & Fang, N. (2010, June). *Regression models for predicting student academic performance in an engineering dynamics course.* In 2010 Annual Conference & Exposition (pp. 15-1026), retrieved from: https://peer.asee.org/regression-models-for-predicting-student-academic-performance-in-an-engineering-dynamics-course

[8] Janeska, M., & Sotiroski, K. (2005). *Data mining-The road to competitiveness*, retrieved from: https://www.researchgate.net/publication/303934915_Data_mining_-_Put_ka_konkurentnosti

[9] Srivastava, J., Desikan, P., & Kumar, V. (2002, November). Web mining: Accomplishments and future directions. *In National Science Foundation Workshop on Next Generation Data Mining (NGDM'02)* (pp. 1-148), retrieved from: http://ieee.org.ar/downloads/Srivastava-tut-pres.Pdf

[10] Javatpoint: Data Mining Techniques, retrieved from: https://www.javatpoint.com/data-mining-techniques?fbclid=IwAR2kqay3fy_rIQxcbpe0dOHQi1hsJhbmfPQ1w3OpcuZI9X-5b97NyQZSnAY

[11] Datascience foundation, retrieved from: https://datascience.foundation/sciencewhitepaper/data-mining:-models-and-methods?fbclid=IwAR28YFLrwlKohBpkYZOjvlilEsHaSRUpFjeDBImwhTzHLusU7fUrFLrJDDo

[12] A Tutorial on Multiple Linear Regression, retrieved from: http://www.stat.yale.edu/Courses/1997-98/101/linmult.htm

[13] Pallant, J. (2009). *SPSS: survival manual: a step-by-step guide to data analysis using SPSS*. Belgrade: Mikro knjiga.

[14] NCSS Software, retrieved from: https://www.ncss.com

[15] Betterevaluation, Rapidminer, retrieved from: https://www.betterevaluation.org/en/resources/tool/rapidminer

[16] Open Data Portal, Republic of Serbia, retrieved from: https://data.gov.rs/sr/