Croatian Journal of Education
Vol.16; 2/2014 pages: 491-523
Original research paper
Paper submitted: 7th December 2012
Paper accepted: 3rd January 2013

Using Data Mining to Predict Success in Studying

Vlado Simeunović[1] and Ljubiša Preradović[2]
[1]*Faculty of Education in Bijeljina*
[2]*Faculty of Architecture and Civil Engineering in Banja Luka*

Abstract

This paper deals with the creation of a model for predicting the performance of students during their studies using data mining, as well as with the analysis of factors which affect the achieved level of success. The model that is created on the basis of students' socio-demographic data, data on their behaviour, personality characteristics, attitudes towards learning and the entire teaching process organization tends to classify students into one of two categories of success. Performance is measured by students' grade point average achieved over the period of studies. We tested three methods of data mining: logistic regression, decision trees and neural networks. We believe that the presented model would serve as a test for the creation of a broader base of updated data by using some of the information tools and that, based on this model, a number of attributes that would relatively reliably predict the performance in studying will be defined.

Key words: *backward stepwise analysis; CART algorithm; decision trees; logistic regression; neural networks.*

Introduction

Future prediction (forecast) is the crown of every science. Education is of strategic importance for economic and social development, i.e. for the development of society founded on knowledge. In the process of European integration it is necessary to harmonize the education system with the criteria and recommendations of the European Union or other European organisations and processes, with particular attention being paid to the indicators of success of the education system, defined by

[1] More details at:
http://europa.eu/legislation_summaries/education_training_youth/general_framework/index_en.htm

the EU[1]. The analysis of success in studying is very important for higher education institutions, since strategic planning of study programme implies extension or reduction of the scope or depth of studying material as well as modification of the structure of pedagogical and educational process, depending on the students' success. Research into success in studying at college has been carried out so far mainly with the aim to determine the grade point average, the duration of studying and similar indicators, while factors which influence the achievement of success have not been explored enough. There are models developed for prediction of success that could help in making decisions regarding the admission of the applicant to studies and those models mostly include demographic data about students, even though the importance of inclusion of other information in the applicant is emphasised (Hardgrave, Wilson, & Kent, 1994; Lin, Imbrie, & Reid, 2009).

The aim of this paper is to discover the important factors that influence students' success, which is represented by the grade point average. We used three methods of data mining for this purpose, adequate for the classification: logistic regression, decision trees and neural networks. Logistic regression is a statistical method based on probability distribution that proved to be effective in many fields of prediction. Its accuracy is improved with the decision trees in order to identify the model which gives accurate classification of students. Research results are based on the survey conducted among students of the Faculty of Education in Bijeljina throughout the academic years 2009, 2010, and 2011. The collected data was classified into three groups: the first group of data is related to intellectual ability (Raven's Matrices) and motivation (Vallerand's Academic Motivation Scale, designed in 1992); the second group is composed of social and demographic data and the third group contains data on students' opinion on work organisation at the university and the way in which the teaching process is conducted. Based on this data a causal model was made, including demographic and other students' characteristics as input variables, and grade point average in the previous academic year as an output variable.

The analysis of the importance of the output variables resulting from logistic regression, decision trees and neural networks indicates the strength of influence of each individual input variable on the success of students, which makes it possible to draw conclusions regarding the presumptive predictors of success.

This paper consists of the presentation of the methodology used, an overview of the previous body of research carried out in this area and of the presented results and the conclusion with the guidelines for the future research.

Methodological Approach

Over the last 15 years, many applications using data mining have been developed for the purposes of education. Romero and Ventura (2007) published the study "Educational data mining: a survey from 1995 to 2005", which includes all the most important achievements in this area. The above study was amended by Baker

and Yacef (2009) in their study "The State of Educational Data Mining in 2009: A Review and Future Visions". Baker and Yacef discuss the methodological profiles of early years of the educational data mining (the main basis for the discussion was the mentioned study of Romero and Ventura from 2007) and compare them with approaches from 2008 and 2009, adding a new category which was not mentioned in the study from 2007 – the so-called "discovery with models". Apart from this, a comparison of educational data mining methods in the period 1999-2005 in relation to 2009 was performed by a number of papers discussing the data in those periods. This comparison was made with the purpose of seeing what the trends and shifts in the research related to the educational data mining are. Both those studies address almost all issues in the field of education, from enrolment into school (university) to web-based education. However, only research with the main focus on educational results will be discussed in this paper. The body of research in the area of utilisation of the intelligent methods for the prediction of students' success is mainly oriented towards the development of the models that are going to be used in decision making regarding admission of students to studies (Witten & Frank, 2000; Romero, Ventura, Espejo, & Hervás, 2008). As criteria, such models take into consideration the available information about the applicant, such as the completed secondary education, grade point average in secondary education, social status and other information regarding the period before enrolment into studies, and using the statistical methods or methods of artificial intelligence they strive to find the model which would produce the best possible accuracy in prediction. Prediction of success in studying by means of data mining methods to a great extent depends on the quality of the acquired data. It seems, though, that it is not possible to form a unique predictive model that would be applicable in all education systems, but the results of the previous body of research indicate that certain groups of variables should always be taken into account as they participate in the prediction of success with the high degree of probability.

In this paper we applied so far most frequently used methods of data mining for prediction of students' success (neural networks, decision trees and logistic regression).

Methodology of Logistic Regression Application

Logistic regression or logistic model or logit model is used for the prediction of the likelihood of events by adjusting the data to logistic curve. Logistic regression is the type of regression analysis where dependent (criterion) variable is dichotomous, i.e. binary, and is coded with 0 or 1, and there is at least one independent (predictor) variable. As events in education are often dichotomous, the logistic regression is often used in prediction of these events. Some earlier research showed a very high level of success of this method. Woodman (2001) carried out a survey at the Open University in the UK using 25 predictors and by using logistic regression tried to reach a model that would, in the best possible way, predict whether a student is going to pass an exam. The author states that the problem in applying this method to large samples

is that wrong conclusions are being drawn regarding confirmation of importance of certain hypotheses even though it is not the case. Kotsiantis, Pierrakeas, and Pintelas (2004) applied several methods (decision trees, artificial neural networks, naive Bayes classifier, instance-based learning, logistic regression and support vector machines) for prediction of student success. The results showed that when a large number of predictors is included, the best results will be achieved by using naive Bayes classifier.

Interpretation of Logistic Regression Model

Statistical modelling of binary variables assumes measuring of choice which can be successful or unsuccessful for each subject. Binary data is probably the most common form of categorical data. The most widely used model of binary data is *logistic regression.*

For binary choice Y and quantitative explanatory variable X, let π(x) represents the probability of success when X has value of x. This probability is the parameter for binomial distribution. Model of logistic regression is linear for logit of this probability.

$$\log it\left[\pi(x)\right] = \log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta x \qquad \text{(Equation 1)}$$

This formula shows that π(x) rises or falls with S - function of *x*.

The second formula for logistic regression refers directly to the probability of success. This formula uses the exponential function exp(x) = e^x in the form shown below:

$$\pi(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)} \qquad \text{(Equation 2)}$$

Interpretation of Linear Approximation

Parameter β determines the rate of rise or fall of S-curve. Designator β indicates whether the curve is falling or rising, as well as the rise rate of the variable as |β| is rising. When a model has the value β = 0, the right-hand side of the Equation 3 is simplified into the constant. Then, π(x) is identical with all *x*s; therefore the curve turns into a straight horizontal line. Binary choice Y then turns into the constant X.

Interpretation of the Likelihood Ratio Test

The following interpretation of the logical regression model uses the likelihood and the likelihood ratio. As a model of the likelihood of choice (i.e. the odds for success), the following equation is used:

$$\frac{\pi(x)}{1-\pi(x)} = \exp(\alpha + \beta x) = e^{\alpha}\left(e^{\beta}\right)^{x} \qquad \text{(Equation 3)}$$

Exponential ratio gives the interpretation for β: the odds are being increased multiplicatively for e^{β} for every one-unit increase in *x*. In other words, the likelihood at x+1 level is equal to the likelihood at *x* multiplied by e^{β}. When β = 0, e^{β} = 1 then the likelihood does not change with the change of *x*.

The logarithm of the likelihood, which represents logit transformation π(x), has a linear ratio. This is a logit model expression, which means that the logit increases with β unit for every unit of change in *x*. Majority does not consider the logit scale as something natural, and therefore it has a limited use.

Test of Significance

When speaking of logistic regression, the null hypothesis $H_0 : \beta = 0$ means that the probability is independent of X.

The statistics of the test for larger samples

$$z = \frac{\beta'}{ASE}$$

has a standard, normal distribution when β = 0. Additionally, *z* could be added to the standard table in order to obtain one-sided or double-sided P-value. Similarly, for double-sided alternative $\beta \neq 0$, $(\beta' / ASE)^2$ Wald statistic is valid, where chi-squared distribution of the large sample with *df = 1* is valid.

Even though the Wald test works well with large samples, the test of credibility ratio is more effective and more reliable for sizes of samples used in practice. The statistics of the test compares the minimal L_0 of log credibility function when β = 0 (i.e. when π(x) has to be identical to all values of *x*) to a maximum L_1 of log credibility function for unrestrictive β. The statistics of the test, $-2(L_0 - L_1)$ also has a chi-square distribution of the large sample with *df = 1*. Majority of software for logistic regression gives data for maximal log-credibility L_0 and L_1, and statistics of credibility ratio is obtained from these maxims.

Distribution of Likelihood Calculation

The likelihood estimate that Y = 1 for fixed set x, out of X is:

$$\pi(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)} \qquad \text{(Equation 4)}$$

The majority of software for logistic regression can give estimation as well as confidence intervals for actual probabilities.

Methodology of the Decision Tree Application

Aiming to create as successful model as possible, one of the nonparametric methods of data mining was tested on the observed sample - the decision tree, or to be more precise, its subgroup classification and regression trees (CART). This method gives a graphic description of the model of impact of input variables on the output, which is expressed in a form of classes and categories. Every node of the graphic tree represents one input variable with "children nodes" marked on its edges for every possible value of some input variable. Every leaf on the tree represents the value of targeted (output) variable if the given values of the input variables are represented from the root to the leaf of that tree. The tree is obtained by "learning" on data, splitting the source set of

data into subsets by testing variable values. The process is repeated on each derived subset by recursive partitioning. The recursion is finished when the subset at a certain node has all the values equal to those of the output variable, or when further splitting does not contribute to further improvement of the results (Witten et al., 2000).

To create the tree we used the CART algorithm (Breiman, Friedman, Olshen, & Stone, 1984) which creates a binary tree, using the available data on input and output variables, by splitting syllables in every node according to the function determined for each input variable. Evaluation function used for splitting is the Gini index (IG), defined by Equation 5:

$$I_G(t) = 1 - \sum_{i=1}^{m} p_i^2 \qquad \text{(Equation 5)}$$

where t is the current node, p_i is the likelihood of the class i in the node t, and m is the number of classes in a model (in our case $m = 2$).

The tree in this paper is created based on 16 input categorical variables. The decision trees are a common tool used in the prediction of success of students in studying with the success rate higher than 65% (Romero et al., 2008). The success rate depends on the sample, the chosen variables and evaluation criteria. In the research carried out in Indonesia (Sembiring, Zarlis, Hartama, Ramliana, & Elvi, 2011) four groups of variables were used as predictors, and those were: interest, study behaviour, time of studying and family support. Categorical variables were: excellent, very good and good. The decision tree predicted all cases with the average success rate of 69.33%. The research conducted in New Zealand (Kovačić, 2010) using the CART algorithm shows that the success of students can successfully be predicted in 62.3% cases. Using twelve input variables which describe a student's status (gender, ethnicity, study mode, mode of financing, type of studies, etc.), Lourens (Lourens & Smit, 2003) conducted research at the University in South Africa, predicting the first-year success at the higher education level using logistic regression and achieved the result of 74.68%. He also confirmed this result using the decision tree.

Vandamme, Meskens and Superby (2007) used decision trees, neural networks and linear discriminant analysis for the early identification of three categories of students: low, medium and high-risk students. Some of the background information (demographics and academic history) of the first-year students in Belgian French-speaking universities were significantly related to academic success. Those were: previous education, number of hours of mathematics, financial independence and age, while gender, the parents' level of education and occupation and marital status were not significantly related to academic success. However, neither of the three methods used to predict academic success performed well. The overall correct classification rate was 40.63% using decision trees, 51.88% using neural networks and the best result was obtained with discriminant analysis with overall classification accuracy of 57.35% (Kovačić, 2010, p. 4).

Methodology of Neural Networks Application

Neural networks simulate the functioning of human brain while performing given task or some function. Neural network is massively parallelised distributed processor with a natural capability to memorise *a posteriori* knowledge and enable its usage. Artificial neural networks are similar to the human brain in two ways:

- neural network acquires knowledge through training process,
- the weight of the inter-neural connections (strength of synaptic connections) is used for memorising the knowledge (Milosavljević, 2005).

Artificial neural networks belong to the intelligent methods of data mining, the aim of which is to find the hidden relationships among the data. Artificial neural network is an interconnected group of simple elements of processing, units or nodes, the functioning of which is based on a mode of functioning of neurons in the living creatures. The network's capability of processing is the consequence of strength of links between those units, and is achieved through process of adaptation or by learning on the set of examples created for that purpose (Russell & Norvig, 2002). In other words, neural networks are programmes or hardware circuits which, mostly by iterative procedure from previous data, tend to find relationships between input and output variables of the model, in order to get the output value for the new input variable. Artificial neuron is the unit for data (variable) processing which receives weighted input values from other variables, transforms the received value by a formula and sends the output to the other variables. Learning is performed by altering the value of "weight" of variables (weights w_{ji} are coefficients that are multiplied by the input variables of some "neuron"). Taking into consideration the number of layers, type of learning, type of links between neurons, relationship between input and output data, the input and transfer function and purpose, numerous algorithms of neural networks can be distinguished. Because of its general purpose (since it is convenient for problems of prediction and classification), and frequent utilisation in vast research, multilayer perceptron algorithm is used for modelling. Multilayer perceptron belongs to supervised feed forward algorithms, where layers of network are linked in a way that signals travel only in one direction, from entrances towards the exits of the network. The best known and most frequently used algorithm, applied for learning and training of multilayer perception networks, is the so-called network "backpropagation". The algorithm of the "backpropagation" network was crucial for the broad commercial application of this methodology, and therefore made neural networks a broadly used and popular method in various areas. The standard algorithm of the "backpropagation" network includes error optimisation using deterministic algorithm of gradient descent (Milosavljević, 2005). The major deficiency of this algorithm is the problem of frequent finding of local instead of global minimum of error, therefore recent research includes its improvement by some other deterministic (e.g. methods of second order) or stochastic methods (e.g. simulated annealing). The

structure of network includes the input layer, output layer and at least one hidden layer with forward feed.

The architecture of the tested neural network in our work consisted of three layers. The number of units (neurons) in the hidden layer and the duration of learning are obtained by cross-validation procedure. Modelling is performed through three phases: (a) data preparation and modelling, (b) training and testing of neural networks, (c) interpretation of the results and the selection of the best model. Training of the network was done on a training sample (70 % of the entire sample), duration of learning of the network was obtained by cross-validation procedure in which the network learns on a training sample in an iterative procedure using various parameters (e.g. various number of hidden neurons), and every combination is tested on the validation sample (30 % of the entire sample). The aim is to find the duration of studying and network structure which give the best result on the validation sample. Finally, the network obtained in such way is tested on the testing sample (30 % of the entire sample), and the result of the testing phase is used as a measure of network success. As for the output functions, sigmoid function was tested (because it is also used in logistic regression) and the used rule of learning was delta learning rule with the momentum 0.7 and the dynamic coefficient of learning from 0.1 to 0.9. The mean square error (MSE) and the root of the mean square error (RMSE) are commonly used for calculation of neural network errors in the training stage.

Since the classification problem is observed in this work, after the stage of neural network testing, the rate of classification is calculated, for each class individually, and the average rate of classification, which is considered as the measure for the evaluation of neural network model success. The rate of classification of every class is the percentage of events that the network correctly placed into it.

Among the first authors in this area using neural networks are Hardgrave, Wilson and Kent (1994), who were comparing neural networks with the traditional statistics techniques in the prediction of university students' success. Further work of the same authors stresses that decision whether to admit a student into the university is based on numerous factors, and it is necessary to develop predictive models that are going to enable a university to accept students who have high potential to study successfully. Their research shows the following: (a) classification techniques are more suitable for the prediction of student success in comparison to predictive methods; (b) prediction of success or failure of students at university is not accurate enough if only the typical data describing a student is used, and (c) non-parametric procedures, such as neural networks, have at least equally accurate results as traditional methods and may be considered as valuable potential for further research in that area.

The same problem of decision making regarding the admission of applicants into the studies was addressed by Naik and Regotman (2004), who had carried out research into students' success at MBA studies. They used neural networks, logit and probit models for the prediction of success of the students enrolled in MBA studies.

Neural networks classified students into successful and unsuccessful based on their grade point average in their undergraduate studies, results of GMAT test, module at undergraduate studies, age and other variables. The results show that neural networks are as successful as other techniques, but their utilisation is recommended in this area because of their numerous advantages.

The research by Sulaiman and Mohezar (2006) deals with the same topic, but goes one step further in identifying the key factors of success. Their model showed that the previous grade point average of the student is the most important predictor of their future success, while variables such as age, ethnicity, gender and years of work experience are not significant for success in studies.

Authors from Malaysia (Zaidah & Daliela, 2007) compared neural networks, decision trees and linear regression analysis in prediction of success of the students. They measured success by cumulative grade point average over the studies, while as the input variables they used demographic profile of the students and grade point average for the first term/semester of undergraduate studies. Results showed that all three methods produced accuracy higher than 80 %, while neural networks had better accuracy than the other two methods.

A group of authors from the Faculty of organisational studies in Belgrade (Vukićević, Išljamović, Jovanović, Delibašić & Suknović, 2011) used neural networks for the prediction of success of students using the data about students, which also included personal data and data about success in the first year of studies. Altogether there were 14 different variables. As an output of the neural networks, it created a variable that represents the predicted grade point average for each student at the end of studies. Six algorithms were tested, out of which, according to all criteria, the best algorithm was *Exhaustive Prune.* Karamouzis and Vrettos (2008) used twelve input parameters to predict the success of 307 students. The prediction success was at the level of 70.27 % for successful and 66.29% for unsuccessful.

Methodological Framework of Research

Research Objectives and Tasks

The objective of this work is to find important factors that have influence on the students' success, represented with the grade point average. For this purpose we used two methods of data mining, appropriate for the classification: logistic regression and decision trees, and we intended to test the quality of each one of them.

Research Hypothesis and Variables

The methods of data mining enable us to relatively precisely predict the success of students at the Faculty of Education in Bijeljina based on the estimation of the likelihood of individual variables participation.

Variables that are input to the model:

1. Criteria variable: achieved grade point average (up to 7.5 – less successful; from 7.51 to 10.00 - successful)

2. Independent variable: general intellectual capability, motivation for studying, gender, place of study, information on scholarships, time devoted to studying, literature, sources and means used for studying, attendance at lectures, attendance at tutorials, attendance at tests, attitude towards the importance of the grade that the student will get at the exam, quality of lectures, quality of tutorials, quality of the course curriculum, quality of lecturers, quality of the process of evaluation of knowledge.

The scope of research:

– Population (all students from the Faculty of Education in Bijeljina).
– Sample of 354 students in the second, third and fourth year.

Data processing: Logistic regression, Decision tree, CART, Neuron networks

Results and Discussion

The Application of Logistic Regression in the Prediction of Students' Success in Their Studies

There are several methods of estimation in logistic regression but the most common and perhaps with the smallest risk, in a sense of hypothesis confirmation, is the METHOD = BSTEP(LR) for *stepwise* analysis backwards. The method consists of the possibility for testing of "log-likelihood" (probability) with a given variable, obtained from the equation. The total statistics of the tested events is shown in Table 1.

Table 1.

Summarised overview of the processed data

Unweighted events		N	Percentage
	Total number of events	234	100.0
	Lost events	0	.0
	Total	234	100.0
Unclassified events		0	.0
Total		234	100.0

The entire test of the model is given in Table 2. Omnibus test of the model. In our logistic regression BSTEP(LR), all variables entered the equation at the beginning and then the model was tested in ten steps. As it can be seen, all values were given at the beginning as "step", "model" and "result" equal with the level of importance .00. At the initial step, all variables are in the model. In the second step, one variable that did not have statistical importance was removed (.95). Similarly, several more variables which did not have statistical importance were removed from the model. Through the ten-step procedure, chi-square test was gradually reduced, which is also the imperative of the model, so that after starting with the value 68.79 we reached the reduced value 64.22.

Table 2.
Omnibus test of the model

		χ^2	df	Sig.
Step 1	Step	69.70	16	.00
	Result	69.70	16	.00
	Model	69.70	16	.00
Step	……	……..	…..	……
Step 10	Step	-2.46	1	.12
	Result	64.22	6	.00
	Model	64.22	6	.00

From the previous table we could not conclude which variables were removed from the model. Only with insight into Table 3 it is possible to notice the regularities in the removal of each variable from the equation. The first variable that was removed from the model was "gender". In the next step the variable "quality of tutorials" was removed. In the third step the variable "attendance at lectures" was removed. In the fourth step, the variable "quality of lectures" did not significantly contribute to the improvement of the entire model. In the following steps, variables were eliminated in the following order: quality of curriculum, study mode and place of study.

Table 3.
Variables that are not input of the equation

			Result	df	Degree of importance
Step 2	Variable	Quality of tutorials	.00	1	.96
	Total statistic		.00	1	.96
Step	……		……..……	……..……	……..……
Step 11	Variable	Gender	.05	1	.82
		Place of study	2.45	1	.12
		Study mode	.58	1	.45
		Attendance at lectures	.14	1	.71
		Quality of lectures	.03	1	.87
		Quality of tutorials	.00	1	.96
		Quality of curriculum	.44	1	.51
		Quality of lecturer	.14	1	.71
		Evaluation	.83	1	.36
		Motivation	.90	1	.22
	Total statistic		6.23	10	.80

The following variables were included in the final model of likelihood estimation: the importance of mark (0.00), attendance at tests (0.00), intellectual capabilities (0.01), scholarship (0.04), attendance at tutorials (0.05), and duration of studies (0.09). In the next table pseudo R – squares are shown. Cox & Snell indices range from 0 to .75, while the correction is only performed with the Nagelkerke index and brings the level into the range from 0 to 1. Of course, here R could not be considered as a coefficient of determination in linear regression, as this is about proportional

participation of individual variables in total probability. With every new stage of step-by-step regression the total result of the encompassed variance was increased. In the final model Cox & Snell index amounts to 0.24, and adjusted by the Nagelkerke index it becomes 0.32, which could be considered as a satisfactory outcome.

Table 4.
Coefficient of determination

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	247.12	.26	.35
2	247.12	.26	.35
3	247.13	.26	.35
4	247.20	.26	.35
5	247.26	.26	.35
6	247.62	.26	.35
7	247.92	.26	.34
8	248.93	.25	.34
9	250.13	.25	.33
10	252.59	.24	.32

The predicted values of dependent variables are shown in Table 5 and are based on the model of the entire logistic regression. This table shows how many cases were accurately predicted, and how many were not. The objective of regression in ten steps was to increase the percentage of successful prediction. In the first step there were 50 cases that had been expected to have value 1, but they obtained value 2. There were also 42 cases, out of 170 cases, that had been expected to have value, 2, but they obtained value 1. Therefore, the total variance of accurate prediction is 74.8 %. In the final model, out of 184 cases that had been expected to have value 1, only 49 cases had value 2, and out of 170 cases that had been expected to have value 2, 40 cases had value 1.

In order not to remain at the surface of the quantitative data, we have performed a detailed statistical and mathematical analysis of every bit of data. Using the following equation we calculated the probability of each variable that has significant influence on the level of success in studies.

In the tenth step we obtained the value of the constant b_0 (b_0 = - 3.69). The constant is input to the exponential function as the first element. We calculated the probability of each input variable to the equation. The highest probability, P(x) = 0.88, has the predictor – the importance of grade, the *b* coefficient of which is 0.94, then follows the attendance at tests (P(x) = 0.85; b_1 = 1.00); intellectual capability has the following probability (P(x) = 0.81; b_1 = 0.93); scholarship is, according to its weight, at the fourth place in the equation (P(x) = 0.70 b_1 = 0.49), then follows the attendance at tutorials (P(x) = 0.61 b_1 = 0.63), and in the end the duration of studying (P(x) = 0.59 b_1 = -0.38).

If we look at the contribution of each variable to the total degree of prediction probability, it could be noted that the regular "attendance at tests" increases the probability for success at studies by five times (Exp. B 5.28). If the "achieved grade" is important to a student, then the degree of success is increased by four times (Exp. B 4.22). A high level of "intellectual capabilities" increases the probability of success by four times (Exp. B 4.12), while regular "attendance at tutorials" increases the level of

success by three and a half times (Exp. B 3.25). If students receive "scholarship", the degree of success is increased by two and a half times (Exp. B 2.59), and if they study at least two hours a day, they will be successful by one time (Exp. B 1.06).

Table 5.

Percentile accuracy of successful prediction

Observed			Prediction		
			Success		Percentile accuracy
			1.00	2.00	
Step 1	Success	1.00	134	50	72.8
		2.00	42	128	75.3
	Total accuracy				**74.02**
Step 2	Success	1.00	133	51	72.2
		2.00	41	129	75.8
	Total accuracy				**74.0**
Step 3	Success	1.00	133	41	72.2
		2.00	41	129	75.8
	Total accuracy				**74.0**
Step 4	Success	1.00	133	51	72.2
		2.00	41	129	75.8
	Total accuracy				**74.0**
Step 5	Success	1.00	133	51	69.8
		2.00	40	130	76.4
	Total accuracy				**73.1**
Step 6	Success	1.00	132	52	71.7
		2.00	39	131	77.0
	Total accuracy				**74.3**
Step 7	Success	1.00	131	53	71.1
		2.00	38	132	77.6
	Total accuracy				**74.3**
Step 8	Success	1.00	135	49	73.3
		2.00	40	130	76.4
	Total accuracy				**74.8**
Step 9	Success	1.00	136	48	73.9
		2.00	41	129	75.8
	Total accuracy				**74.8**
Step 10	Success	1.00	135	49	73.3
		2.00	40	130	76.4
	Total accuracy				**74.8**

In order to support the results obtained by regression analysis, it could be noted that 85 % of successful students regularly attend tests, compared to 53 % of attendance of less successful students. Successful students attend the tutorials in 88 % cases, compared to 63 % of students from the opposite group. 79 % of them stated that they consider it important which grade they would get, compared to 45 % students from the group of the less successful students. 85 % of them study two to five hours a day on average, compared to 60 % students from the opposite group, while 27 % of successful students receive the scholarship, compared to 11 % from the opposite group.

The results obtained by logistic regression in our research show 74.8% successful prediction of students' success at the Faculty of Education. The variables that predominantly contribute to a total probability of prediction could be classified into the following three groups: didactic (attendance at tests, attendance at tutorials, duration of studying, and importance of achieved grade), personal characteristics (intellectual capability) and social group (receiving a scholarship).

The Application of Decision Tree in the Prediction of Students' Success over the Studies

For the purpose of prediction of success in studying of the students at the Faculty of Education in Bijeljina, we used the CART decision tree. In some of the previous body of research the CART tree was proved to be an adequate tool in the prediction of students' success.

```
*******************************************************
* 2nd algorithm CART
*******************************************************
CART Decision Tree
importance of mark < 1.5
| test < 1.5: 1(87.0/38.0)
| test >= 1.5: 2(31.0/14.0)
importance of mark >= 1.5: 2(134.0/50.0)
Number of Leaf Nodes: 3
Size of the Tree: 5
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances         252       71.250 %
Incorrectly Classified Instances       102       28.759 %
Kappa statistic                          0.40
Mean absolute error                      0.39
Root mean squared error                  0.46
Relative absolute error                 80.00 %
Root relative squared error             93.53 %
Total Number of Instances              354
=== Detailed Accuracy By Class ===
TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
 0.68    0.25      0.64      0.66      0.65      0.68          1
 0.74    0.34      0.76      0.75      0.75      0.68          2
=== Confusion Matrix ===
  a   b   <-- classified as
 126  58 |   a = 1
 44  126 |   b = 2
*******************************************************
```

From the given preview, it is noticeable that the average rate of classification obtained from the decision tree is 71.2 %, which is less than the average rate obtained by logistic regression (74.8 %). The tree is especially accurate in recognising the "weaker" students with the grade point average lower than 7.5, where the rate of classification is 74.1 %. A somewhat lower rate is obtained for class 1 – "better" students (68.4 %).

The results of student classification by employing the decision tree at the subsample for testing could also be illustrated by the confusion matrix, which shows in its columns the actual number of students who belong to the category with a lower (2) or higher (1) grade point average, while its rows show the number of students whom the model of decision tree classified in either Category 2 or 1. The diagonal of the confusion matrix shows the number of students whom the model classified correctly. It is visible from the table that a total of 44 students with the lower grade point average (Category 2) were not classified correctly by the decision tree, while 126 were classified correctly. The situation is different for Category 1 where 126 of students were correctly classified, while 58 were classified into a wrong category. The entire model correctly classified 252 out of 354 students. The tree is somewhat more precise than the regression analysis and extracts only two variables which have the influence at the total probability, and those are: attendance at tests and the importance of the grade.

The Application of Neural Networks in the Prediction of Students' Success in Their Studies

A multilayer neural network was trained in this work. The rate of correct classification for each class on the training and testing sample was obtained for the mentioned architecture. The average rate of classification on testing sample, already described in Chapter two, was used as a measure of success of the model.

Table 6.

Classification table for prediction using neural networks

SAMPLE	ATTRIBUTE	PREDICTION		
		2	1	Percentage
TRAINING	2	98	21	82.8 %
	1	99	27	78.3 %
	TOTAL	56.8 %	43.2 %	80.9 %
TESTING	2	43	7	84.4 %
	1	36	22	63.0 %
	TOTAL	66.7 %	33.3 %	76.4 %

In order to solve the above problem in neural network, 33 hidden neurons were used, as well as logistic transfer function, delta-bar-delta rule of learning and network learned at the maximum 1,000 epochs. The number of hidden neurons that give the smallest error in a phase of cross validation was 33. The obtained results show that in the testing sample there were 76.4 % correctly classified cases, while 23.6 % cases were put into a wrong class. If the particular rates of classification of each individual

class are looked at, then it is noticeable that the rate of classification for Category 2 ("weaker" students) is 84.4 %, while the rate of classification for Category 1, i.e. for "better" students is 63.0 %. Better accuracy of classification for "weaker" students indicates that students with the grade point average lower than 7.5 have common characteristics which the neural network model succeeded in recognising and finding relationships, compared to the case of students with the grade point average higher than 7.5. In order to interpret the results obtained by neural networks it is essential to have an explanation of each variable in the procedure of success prediction. The probability of every independent variable in the entire defined model is shown in Table 7. Similarly to the model obtained by logistic regression, the variable "the importance of grade" is input to a prediction model with the probability of hundred percent. Variables "duration of studying", "attendance at tests" and "intellectual capability" with the probability greater than 80% are important elements of the model. Other variables have weaker impact on the model for prediction. Regardless of the size of the sample, it should be noted that the obtained results point to the weaknesses of the system of the education process since the variables such as "quality of lecturer", "attending the lectures", "quality of tutorial" and "quality of curriculum" have a very low predictive value in prediction of students' success. A further analysis of these results would give more detailed answers to these open questions.

Table 7.

Participation of each variable in the model of success prediction

Name of the variable	Importance	Normalised importance
Importance of mark	.15	100 %
Duration of studying	.13	87.3 %
Attendance at tests	.12	80.3 %
Intellectual capability	.12	80.1 %
Attendance at tutorials	.08	52.9 %
Motivation	.08	52.9 %
Quality of lectures	.08	52.9 %
Type of learning	.07	46.5 %
Scholarship	.07	45.6 %
Mode of evaluation	.07	43.4 %
Place of study	.06	41.2 %
Gender	.05	31.5 %
Quality of lectures	.03	21.8 %
Attendance at lectures	.03	19.6 %
Quality of tutorials	.03	19.5 %
Quality of the course curriculum	.02	15.5 %

Conclusion

All three models of data mining offer a possibility to predict the students' success accurately. The results of prediction obtained for all three applied approaches are of approximately equal value. The best results were achieved using neural networks (76.4%), then follows logistic regression (74.8%), while somewhat weaker results of prediction were obtained using the decision tree (71.2%). In the previous body

of research, when only social and demographic data about students were used, the model had had a higher level of accuracy compared to the level that we obtained in our research. The results that we obtained reflect the state of our system of education in the area of higher education. It is very clear that the reform of the system caused the change in students' behaviour. The concerning part is the fact that a large number of students do not care what grade they would get at the exam, do not consider it important to attend lectures, as they, obviously, do not have a high opinion of the possibilities of studying during the pedagogical and educational process.

It has been shown that more successful students pay more attention to studying, take exams (which indicates the applicability of the studies) and that they care about the grade that they will get. This is also the area of opportunity for improvement in the process of modelling the work at the university. Of course, these factors should be linked with the possibility of receiving scholarship.

It is necessary to perform the analysis of all of the obtained results, enrich the system of prediction with new variables and then make corrections in the implementation of the educational process. Some pieces of advice for improvement of the educational process stemming from this research could be classified into the following groups:

- the learning process should be intensified;
- work in smaller groups should be organized more frequently;
- there should be a better connection of the theoretical and practical content;
- there should be regular follow-up of students' success through written tests and other models of knowledge assessment;
- students should be supported in adaptation to studies;
- the scope and content of literature used in studying should be updated, thus making it adjusted to the objectives and mode of assessment;
- lecturers should be motivated in order to raise the level of their interest in lectures and in the students;
- the level and the quality of knowledge of lecturers regarding methodology of teaching and evaluation of success should be raised;
- communication skills and commitment of the lecturer to the methodology of teaching should be improved;
- the system of evaluation of teaching and the lecturer should be modernized, as one of the ways of monitoring the quality of teaching and initiating changes.

How to improve the model for prediction of students' success at studies?

- introduce new variables, i.e. outcome of studies;
- extend the sample;
- include other lecturers and faculties of education into the sample;
- create an intelligent sample of support to the educational system at universities.

It could be stated that data mining has the potential for monitoring the educational achievements at universities. It offers practical solutions in the area of prediction of the future students' behaviour and provides the basis for changing certain components of the education system in order to improve the overall success.

References

Baker, R. S. J. d. & Yacef, K. (2009). The State of Educational Data Mining in 2009: A Review and Future Visions. *Journal of Educational Data Mining*, *1*(1), 3-17.

Breiman, L., Friedman, J.H., Olshen, R.A., & Stone, C.J. (1984). *Classification and Regression Trees*. New York: Chapman & Hall.

Hardgrave, B.C., Wilson, R.L., & Kent, K.A. (1994). Predicting Graduate Student Success: A Comparison of Neural Networks and Traditional Techniques. *Computers* & *Operations Research*, *21*, 249-263.

Han, J., & Kamber, M. (2001). *Data Mining – Concepts and Techniques.* San Francisco: Morgan Kaufman Publishers.

Karamouzis, S.T., & Vrettos, A. (2008). *An Artificial Neural Network for Predicting Student Graduation Outcomes,* proceedings of the World Congress on Engineering and Computer Science 2008. San Francisco, USA.

Kirkby, R. (2002). *WEKA Explorer User Guide for Version 3-3-4*, University of Waikato.

Kotsiantis, S., Pierrakeas, C., & Pintelas, P. (2004). Predicting students' performance in distance learning using machine learning techniques. *Applied Artificial Intelligence*, *18*, 411-426.

Kovačić, J. Z. (2010). *Early Prediction of Student Success: Mining Student Enrolment Data*, Proceedings of Informing Science & IT Education Conference (InSITE) 2010. Cassino, Italy, 647-665.

Lin, J. J., Imbrie, P. K. & Reid, K. J., (2009). *Student Retention Modelling: An Evaluation of Different Methods and their Impact on Prediction Results*. Paper presented at the Research in Engineering Education Symposium 2009. Palm Cove, Australia.

Lourens, A. & Smit, I.P.J. (2003). Retention: predicting first-year success. *South African Journal of Higher Education, 17*(2), 169-170.

Masters, T. (1995). *Advanced Algorithms for Neural Networks: A C++ Sourcebook.* New York: John Wiley & Sons.

Milosavljević, M. (2005). *Neuronske mreže*. Belgrade: Faculty of Electrical Engineering.

Naik, B. & Ragothaman, S. (2004). Using Neural Networks to Predict MBA Student Success. *College Student Journal*, *38*(1), 143-150.

Oladokun, V.O., Adebanjo, A. T., & Charles-Owaba, O.E. (2008). Predicting Students' Academic Performance using Artificial Neural Network, A Case Study of an Engineering Course. *The Pacific Journal of Science and Technology*, *9* (1), 72-79.

Ramaswami, M., & R. Bhaskaran, (2010). A CHAID based performance prediction model in educational data mining. *IJCSI International Journal of Computer Science Issues*, 7(1), 10-18.

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, *33*(1), 135-146, doi:10.1016/j.eswa.2006.04.005.

Romero, C., Ventura, S., Espejo, P. G., & Hervás, C. (2008). *Data Mining Algorithms to Classify Students*. Proceedings of the Educational Data Mining 2008, 1st International Conference on Educational Data Mining. Montreal-Quebec, Canada.

Romero, C., & Ventura, S. (2011). Educational data mining: a review of the state-of-the-art. *IEEE Trans. Syst. Man Cybernet. C Appl. Rev*., *40*(6), 601–618.

Reason, R. D. (2003). Student variables that predict retention: Recent research and new developments. *NASPA Journal*, *40*(4), 172-191.

Russell, S.J, & Norvig, P. (2002). *Artificial Intelligence*: A Modern Approach. New York: Prentice Hall.

Shulruf, B., Hattie, J., & Tumen, S. (2008). The Predictability of Enrolment and First-Year University Results from Secondary School Performance: The New Zealand National Certificate of Educational Achievement. *Studies in Higher Education*, *33*(6), 685-698.

Sembiring, S., Zarlis, M., Hartama, D., Ramliana, S., & Elvi, W. (2011). *Prediction of student academic performance by an application of data mining techniques.* International Conference on Management and Artificial Intelligence, IPEDR vol.6 (2011) © (2011) IACSIT Press. Bali, Indonesia, 110-114.

Sulaiman, A., & Mohezar, S. (2006). Student Success Factors: Identifying Key Predictors. *Journal of Education for Business*, *81*(6), 328-333.

Vallerand, R. J., Pelletier, L. G., Blais, M. R., Briere, N. M., Senecal, C., & Vallieres, E. F. (1992). The Academic Motivation Scale: A measure of intrinsic, extrinsic, and motivation in education. *Educational and Psychological Measurement*, *52*, 1003-1017.

Vandamme, J.-P., Meskens, N., & Superby, J.-F. (2007). Predicting academic performance by data mining methods. *Education Economics*, *15*(4), 405-419.

Vukićević, M., Išljamović, S., Jovanović, M., Delibašić, B. & Suknović, M. (2011). *Primena neuronskih mreža za predviđanje uspeha studenata.* Belgrade: Faculty of Organisational Sciences.

Witten, I.H., & Frank, E. (2000). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementation*. San Francisco: Morgan Kaufman Publishers.

Woodman, R. (2001). *Investigation of factors that influence student retention and success rate on Open University courses in the East Anglia region*. M.A. Dissertation. Sheffield Hallam University, UK.

Zaidah, I., & Daliela, R. (2007). *Predicting students' academic performance, comparing artificial neural network, decision tree and linear regression. P*roceedings of 21st Annual SAS Malaysia Forum, Kuala Lumpur, 1-6.

Vlado Simeunović
Pedagogical Faculty in Bijeljina, Semberskih ratara bb,
76 300 Bijeljina, Bosnia and Herzegovina
vlado.simeunovic@gmail.com

Ljubiša Preradović
Faculty of Architecture and Civil Engineering in Banja Luka,
Bulevar vojvode Petra Bojovića 1A
78000 Banja Luka, Bosnia and Herzegovina
ljpreradovic@agtbl.org

Primjena rudarenja podataka u predviđanju uspješnosti studiranja

Sažetak

Rad se bavi stvaranjem modela za predviđanje uspješnosti studenata tijekom studiranja primjenom rudarenja podataka (engl. data mining) i analizom čimbenika koji utječu na postignuti stupanj uspješnosti. Model koji je stvoren na temelju socio-demografskih podataka o studentima, podataka o njihovu ponašanju, osobnim karakteristikama, stavovima prema učenju i organizaciji cjelokupnog nastavnog procesa svrstava studente u jednu od dviju kategorija uspješnosti. Uspješnost u studiranju mjeri se srednjom prosječnom ocjenom koju studenti stječu tijekom studiranja. Ispitali smo tri metode rudarenja podataka: logističku regresiju, drvo odlučivanja i neuronske mreže. Smatramo da bi prikazani model mogao poslužiti kao test za stvaranje šire baze ažuriranih podataka korištenjem nekih informacijskih alata i da bi se na temelju toga modela mogli definirati brojni atributi koji bi relativno pouzdano predviđali uspješnost studenata u studiranju.

Ključne riječi: *CART algoritam; logistička regresija; neuronske mreže; stablo odlučivanja; stupnjevita analiza unatrag.*

Uvod

Predviđanje budućnosti je vrhunac svake znanosti. Obrazovanje je od strateške važnosti za ekonomski i društveni razvoj, tj. za razvoj društva koje se temelji na znanju. Tijekom procesa europske integracije neophodno je uskladiti obrazovni sustav s kriterijima i preporukama Europske unije ili drugih europskih organizacija i procesa, poklanjajući posebnu pažnju pokazateljima uspješnosti obrazovnog sustava koje je definirala Europska unija.[2] Analiza uspješnosti u studiranju iznimno je bitna za institucije koje provode visoko obrazovanje, budući da strateško planiranje studijskih programa implicira proširenje ili smanjenje opsega i dubine materijala koji se koristi u studiranju, kao i promjene u strukturi odgojnog i obrazovnog procesa, ovisno o uspješnosti studenata. Do sada je provedeno mnogo istraživanja o uspjehu

[2] Više na: http://europa.eu/legislation_summaries/education_training_youth/general_framework/index_en.htm

u studiranju, uglavnom s ciljem određivanja prosječne ocjene, duljine studiranja i sličnih pokazatelja, dok čimbenici koji utječu na postizanje uspjeha nisu dovoljno istraživani. Postoje modeli koji su izrađeni da bi se predvidio uspjeh u studiranju, a koji bi mogli pomoći u donošenju odluka koje se tiču upisivanja kandidata na studij. Ti modeli uglavnom uključuju demografske podatke o studentima, iako se naglašava važnost razmatranja i ostalih podataka o kandidatima (Hardgrave, Wilson, i Kent, 1994; Lin, Imbrie, i Reid, 2009).

Cilj ovoga rada je otkriti važne čimbenike koji utječu na uspjeh studenata koji se izražava srednjom prosječnom ocjenom. Za potrebe ovog rada koristili smo tri metode rudarenja podataka koje su bile pogodne za klasificiranje: logističku regresiju, stablo odlučivanja i neuronske mreže. Logistička regresija je statistička metoda koja se temelji na distribuciji vjerojatnosti, a koja se pokazala učinkovitom u mnogim poljima predviđanja. Njezina je točnost povećana primjenom stabla odlučivanja da bi se utvrdio model koji omogućava točnu klasifikaciju studenata. Rezultati istraživanja temelje se na anketi provedenoj među studentima Pedagoškog fakulteta u Bijeljini tijekom akademskih godina od 2009. do 2011. Prikupljeni podaci klasificirani su u tri skupine: u prvoj skupini podaci su povezani s intelektualnim sposobnostima (Ravenova matrica) i motivacijom (Vallerandova skala akademske motivacije, kreirana 1992). Druga skupina podataka sastoji se od društvenih i demografskih podataka, a treća skupina sadrži podatke o mišljenju studenata o organizaciji rada na sveučilištu i o načinu na koji se provodi nastavni proces. Na temelju tih podataka stvoren je kauzalni model, koji uključuje demografske i druge karakteristike studenata kao ulazne varijable, i srednju prosječnu ocjenu tijekom prethodne akademske godine kao izlaznu varijablu.

Analiza važnosti izlaznih varijabli koja se dobije logističkom regresijom, stablom odlučivanja i neuronskim mrežama ukazuje na intenzitet utjecaja svake pojedinačne ulazne varijable na uspjeh studenata. To omogućava donošenje zaključaka o mogućim prediktorima uspjeha.

Ovaj rad sastoji se od prezentacije metodologije koja je korištena, od pregleda prijašnjih istraživanja provedenih u ovom području, od prikazanih rezultata i zaključka sa smjernicama za provođenje daljnjih istraživanja.

Metodološki pristup

Tijekom proteklih 15 godina razvijene su brojne aplikacije za korištenje rudarenja podataka u obrazovnim istraživanjima. Romero i Ventura (2007) objavili su svoje istraživanje „Rudarenje podataka u obrazovanju: istraživanje od 1995. do 2005.", koje uključuje sva najvažnija dostignuća u tom području. Spomenuto istraživanje dopunili su Baker i Yacef (2009) svojim istraživanjem „Stanje rudarenja podataka u obrazovanju u 2009: Pregled i vizije za budućnost." Baker i Yacef raspravljaju o metodološkim profilima iz ranih godina rudarenja podataka u obrazovanju (glavna osnova za raspravu bila je spomenuto istraživanje Romera i Venture iz 2007. godine) i uspoređuju

ih s pristupima korištenima 2008. i 2009. godine, dodajući im novu kategoriju koja nije bila spomenuta u istraživanju iz 2007. godine – tzv. „otkrivanje s modelima". Osim toga, usporedba metoda rudarenja podataka u obrazovanju od 1999. do 2005. s onima do 2009. godine provedena je u brojnim radovima koji su analizirali podatke u tim razdobljima. Ta usporedba provedena je s namjerom uočavanja trendova i pomaka u istraživanjima koja se bave rudarenjem podataka u području obrazovanja. Oba istraživanja bave se gotovo svim pitanjima u području obrazovanja, od upisa u školu (fakultet) do obrazovanja putem interneta. Međutim, u ovom radu raspravljat će se samo o istraživanjima čija su glavna tema obrazovni rezultati. Istraživanja o korištenju inteligentnih metoda za predviđanje uspjeha studenata uglavnom su orijentirana na razvoj modela koji će se koristiti u postupku donošenja odluka o upisu studenata na studije (Written i Frank, 2000; Romero, Ventura, Espejo, i Hervás, 2008). Takvi modeli kao kriterije uzimaju dostupne podatke o kandidatima, kao što su: završena srednja škola, srednja prosječna ocjena u srednjoj školi, društveni status, kao i druge podatke koji se tiču razdoblja prije upisa na fakultet. Korištenjem statističkih metoda ili metoda umjetne inteligencije pokušava se pronaći model koji bi doveo do najveće moguće točnosti u predviđanju uspjeha studenata. Predviđanje uspjeha u studiranju korištenjem rudarenja podataka u velikoj mjeri ovisi o kvaliteti prikupljenih podataka. Ipak, čini se da nije moguće stvoriti jedinstveni model predviđanja koji bi bio primjenjiv u svim obrazovnim sustavima. No, rezultati prijašnjih istraživanja ukazuju na to da bi se neke skupine varijabli uvijek trebale uzeti u obzir jer se koriste u predviđanju uspjeha s visokim stupnjem vjerojatnosti.

U ovom radu primijenili smo do sada najčešće korištene metode rudarenja podataka pri predviđanju uspjeha studenata (neuronske mreže, stabla odlučivanja i logističku regresiju).

Metodologija primjene logističke regresije

Logistička regresija/logistički model/logit model koristi se za predviđanje vjerojatnosti događaja prilagođavanjem podataka logističkoj krivulji. Logistička regresija je tip regresijske analize u kojoj je zavisna varijabla (kriterij) dvojaka, tj. binarna, pa je kodirana brojevima 0 ili 1. Postoji uvijek barem jedna nezavisna varijabla (prediktor). Kako su događaji u obrazovanju često dvojaki, logistička regresija se često koristi u predviđanju tih događaja. Neka prijašnja istraživanja pokazala su vrlo visok stupanj uspješnosti te metode. Woodman (2001) je proveo istraživanje na Otvorenom sveučilištu u Velikoj Britaniji koristeći 25 prediktora. Logističkom regresijom pokušao je doći do modela koji bi na najbolji način mogao predvidjeti hoće li student položiti ispit. Autor navodi da je problem primjene te metode na velikom uzorku taj što se izvode pogrešni zaključci vezani uz potvrđivanje važnosti određenih hipoteza, iako to baš i nije uvijek slučaj. Kotsiantis, Pirrakeas i Pintelas (2004) primijenili su nekoliko metoda (stablo odlučivanja, umjetne neuronske mreže, naivni Bayesov klasifikator, učenje na temelju primjera, logističku regresiju i stroj s

potpornim vektorima) za predviđanje uspjeha studenata. Rezultati su pokazali da će se prilikom uključivanja velikog broja prediktora najbolji rezultati dobiti korištenjem naivnog Bayesova klasifikatora.

Interpretacija modela logističke regresije

Statističko modeliranje binarnih varijabli podrazumijeva mjerenje izbora koji za svaku pojedinu stavku može biti uspješan ili neuspješan. Binarni su podaci vjerojatno najčešći oblik kategoričkih podataka. Najčešće korišten model binarnih podataka je *logistička regresija*.

Za binarni izbor Y i kvantitativnu eksploratornu varijablu X, let π(x) predstavlja vjerojatnost uspjeha kada X ima vrijednost x. Ta je vjerojatnost parametar za binomijalnu distribuciju. Model logističke regresije je linearan za logit ove vjerojatnosti.

$$\log it[\pi(x)] = \log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta x \qquad \text{(Jednadžba 1)}$$

Ova formula pokazuje da π(x) raste ili pada sa S - funkcijom od *x*.

Druga formula logističke regresije odnosi se izravno na vjerojatnost uspjeha. Ta formula koristi eksponencijalnu funkciju exp(x) = e^x u prikazanom obliku:

$$\pi(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)} \qquad \text{(Jednadžba 2)}$$

Interpretacija linearne aproksimacije

Parametar β određuje stopu porasta ili pada S-krivulje. Oznaka β pokazuje da li krivulja pada ili raste, kao i stopu porasta varijable kada je |β| u porastu. Kada model ima vrijednost β = 0, desna strana jednadžbe 3 je pojednostavljena u konstantu. Tada je π(x) identičan svim x vrijednostima, pa se stoga krivulja pretvara u ravnu horizontalnu liniju. Binarni izbor Y se tada također pretvara u konstantu X.

Interpretacija testa omjera vjerojatnosti

Sljedeća interpretacija modela logističke regresije koristi vjerojatnost i omjer vjerojatnosti. Kao model vjerojatnosti izbora (tj. izgleda za uspjeh), korištena je sljedeća jednadžba:

$$\frac{\pi(x)}{1-\pi(x)} = \exp(\alpha + \beta x) = e^{\alpha}\left(e^{\beta}\right)^{x} \qquad \text{(Jednadžba 3)}$$

Eksponencijalni omjer daje interpretaciju za β: izgledi su povećani umnoženo sa e^{β} za svako pojedinačno povećanje za x. Drugim riječima, vjerojatnost na stupnju x+1 nije jednaka vjerojatnosti na stupnju x pomnoženo sa e^{β}. Kada je β = 0, e^{β} = 1, tada se vjerojatnost ne mijenja kada se mijenja x.

Logaritam vjerojatnosti, koji predstavlja logit transformaciju π(x) ima linearan omjer. To je izraz logit modela, što znači da se logit povećava s jedinicom β za svaku

pojedinačnu promjenu u x. Većina ne smatra da je logit skala nešto prirodno, pa stoga ona ima ograničenu uporabu.

Test signifikantnosti

Kada se govori o logističkoj regresiji, nulta hipoteza $H_0 : \beta = 0$ znači da je vjerojatnost neovisna o X.

Statistika testa za veće uzorke

$$z = \frac{\beta'}{ASE}$$

ima standardnu, normalnu distribuciju kada je $\beta = 0$. K tomu, *z* bi se mogao dodati standardnoj tablici da bi se dobila jednostrana ili dvostrana vrijednost P. Slično, za dvostrani alternativni $\beta \neq 0$, $(\beta' / ASE)^2$ vrijedi Waldova statistika u kojoj vrijedi distribucija chi-kvadrata velikog uzorka *df = 1*.

Iako se pokazalo da Waldov test dobro funkcionira s velikim uzorcima, test omjera vjerodostojnosti je učinkovitiji i pouzdaniji za veličine uzorka u praksi. Statistika testa uspoređuje minimalni L_0 log funkcije vjerodostojnosti gdje je $\beta = 0$ (tj. kada $\pi(x)$ mora biti jednak svim vrijednostima *x*) s maksimalnim L_1 log funkcije vjerodostojnosti za neograničeni β.

Statistika testa $-2(L_0 - L_1)$ također ima distribuciju chi-kvadrata velikog uzorka *df = 1*s. Većina računalnih programa za logističku regresiju daje podatke za maksimalnu log-vjerodostojnost L_0 i L_1, pa se statistika omjera vjerodostojnosti dobiva iz ovih načela.

Distribucija izračunavanja vjerojatnosti

Procjena vjerojatnosti da je Y = 1 za određeni skup x, izvan X je:

Većina računalnih programa za logističku regresiju može dati procjenu, kao i

$$\pi(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)} \qquad \text{(Jednadžba 4)}$$

intervale pouzdanosti za stvarne vjerojatnosti.

Metodologija primjene stabla odlučivanja

S ciljem stvaranja što je moguće uspješnijeg modela, jedna od neparametrijskih metoda rudarenja podataka testirana je na promatranom uzorku. Ta je metoda stablo odlučivanja ili, preciznije, njegova podkategorija i regresijska stabla (CART). Ta metoda daje grafički opis modela utjecaja ulaznih varijabli na izlazne, što se izražava u obliku vrsta i kategorija. Svaki čvor grafičkog stabla predstavlja jednu ulaznu varijablu s „izdancima" označenima na njezinim rubovima za svaku moguću vrijednost neke ulazne varijable. Svaki list na stablu predstavlja vrijednost ciljne (izlazne) varijable ako su dane vrijednosti ulaznih varijabli prikazane od korijena do lista toga stabla. Stablo se dobije „učenjem" podataka, dijeljenjem izvornog skupa podataka na podskupove

testiranjem vrijednosti varijabli. Proces se ponavlja na svakom dobivenom podskupu rekurzivnim particioniranjem. Rekurzija je završena kada podskup na određenom čvoru ima sve vrijednosti jednake vrijednostima izlazne varijable ili kada daljnja razdioba ne doprinosi poboljšanju rezultata (Witten i sur., 2000). Da bismo izradili stablo, koristili smo CART algoritam (Breiman, Friedman, Olshen i Stone, 1984) koji stvara binarno stablo koristeći se dostupnim podacima o ulaznim i izlaznim varijablama, tako što dijeli slogove u svakom čvoru prema funkciji određenoj za svaku ulaznu varijablu. Funkcija ocjenjivanja koja se koristi za razdiobu jest Gini indeks (IG), koji je definiran Jednadžbom 5:

$$I_G(t) = 1 - \sum_{i=1}^{m} p_i^2 \qquad \text{(Jednadžba 5)}$$

Ovdje je t trenutni čvor, p_i je vjerojatnost vrste i u čvoru t, a m je broj vrsta unutar modela (u našem slučaju m=2).

Stablo je u ovom radu izrađeno na temelju 16 ulaznih kategoričkih varijabli. Stabla odlučivanja su često korišten alat u predviđanju uspjeha studenata u studiranju i imaju stopu uspješnosti veću od 65 % (Romero i sur., 2008). Stopa uspješnosti ovisi o uzorku, odabranim varijablama i kriterijima ocjenjivanja. U jednom istraživanju u Indoneziji (Sembiring, Zarlis, Hartama, Ramliana i Elvi, 2011) četiri grupe varijabli korištene su kao prediktori, a one su: interes, ponašanje tijekom studiranja, vrijeme studiranja i obiteljska potpora. Kategoričke varijable bile su: izvrstan, vrlo dobar i dobar. Stablo odlučivanja predvidjelo je sve slučajeva s prosječnom stopom uspješnosti od 69,33 %. Istraživanje provedeno na Novom Zelandu (Kovačić, 2010), u kojem je korišten CART algoritam, pokazuje da se uspjeh studenata može uspješno predvidjeti u 62,3 % slučajeva. Koristeći dvanaest ulaznih varijabli koje opisuju status studenta (spol, etnička pripadnost, način studiranja, način financiranja, vrsta studija itd.), Lourens (Lourens i Smit, 2003) je proveo istraživanja na Sveučilištu u Južnoj Africi i predvidio uspjeh na prvoj godini na stupnju visokog obrazovanja koristeći se logističkom regresijom, te postigao rezultat od 74,68 %. Također je potvrdio taj rezultat koristeći se stablom odlučivanja.

Vandamme, Meskens i Superby (2007) koristili su se stablom odlučivanja, neuronskom mrežom i linearnom diskriminativnom analizom za rano prepoznavanje triju kategorija studenata: onih niskog, srednjeg i visokog rizika. Neke informacije o studentima prve godine (kao što je njihova demografska i akademska povijest) na sveučilištima na kojima se govori belgijska varijanta francuskog jezika bile su značajno povezane s akademskim uspjehom. Te informacije odnosile su se na: prethodno obrazovanje, broj sati matematike, financijsku neovisnost i dob, dok spol, stupanj obrazovanja roditelja i njihovo zanimanje, kao i bračno stanje, nisu značajno utjecali na akademski uspjeh. Međutim, sve tri metode koje su se koristile za predviđanje akademskog uspjeha nisu dale dobre rezultate. Ukupna stopa točnog klasificiranja bila je 40,63 % korištenjem stabla odlučivanja, 51,88 % korištenjem neuronskih mreža, dok je najbolji rezultat dobiven diskriminativnom analizom s ukupnom stopom točnog klasificiranja od 57,35 %. (Kovačić, 2010, str. 4)

Metodologija primjene neuronskih mreža

Neuronske mreže oponašaju način rada ljudskog mozga dok izvode zadani zadatak ili neku funkciju. Neuronska je mreža uvelike paralelno distribuiran procesor s prirodnom sposobnošću memoriranja a posteriori znanja, koji ujedno i omogućava korištenje tog istog znanja. Umjetne neuronske mreže slične su ljudskom mozgu u dva aspekta:

– neuronska mreža stječe znanje putem procesa vježbi
– opterećenost međuneuronskih veza (snaga sinaptičkih veza) koristi se za memoriranje znanja (Milosavljević, 2005.).

Umjetne neuronske mreže ubrajaju se u inteligentne metode rudarenja podataka, čiji je cilj pronaći skrivene veze između podataka. Umjetna neuronska mreža je skupina međusobno povezanih jednostavnih elemenata procesiranja, jedinica ili grananja, čija se funkcija temelji na načinu funkcioniranja neurona u živim bićima. Sposobnost mreže da obrađuje podatke posljedica je jačine veza između tih jedinica, a postiže se procesom prilagodbe ili učenjem na nizu primjera stvorenih za tu svrhu (Russel i Norvig, 2002). Drugim riječima, neuronske mreže su programi ili hardverski krugovi koji, većinom iteracijskim, ponavljajućim postupkom iz prethodnih podataka pokušavaju pronaći veze između ulaznih i izlaznih varijabli modela, da bi dobile izlaznu vrijednost za novu ulaznu varijablu. Umjetni neuroni su jedinice za obradu podataka (varijabli) koje dobivaju opterećene ulazne vrijednosti od drugih varijabli, pretvaraju primljene vrijednosti preko formula i šalju izlazne varijable ostalim varijablama. Učenje se odvija tako što se mijenja vrijednost opterećenosti varijabli (opterećenja w_{ji} su koeficijenti koji su pomnoženi ulaznim varijablama nekih „neurona"). Uzimajući u obzir broj slojeva, načine učenja, vrste veza između neurona, vezu između ulaznih i izlaznih podataka, funkciju ulaza i prijenosa i svrhu, možemo razlikovati brojne algoritme neuronskih mreža. Zbog svoje općenite namjene (budući da je pogodan za probleme predviđanja i klasificiranja) i čestog korištenja u brojnim istraživanjima, algoritam višeslojnog perceptrona koristio se za modeliranje. Višeslojni perceptron ubraja se u *feedforward* algoritme u kojima su slojevi mreže povezani tako da signal putuje samo u jednom smjeru, od ulaza prema izlazu iz mreže. Najpoznatiji i najčešće korišteni algoritam koji se koristi za učenje i uvježbavanje višeslojnih perceptivnih mreža je tzv. „backpropagation" mreža. Algoritam takve mreže bio je neophodan za širu komercijalnu primjenu te metodologije, pa je stoga zaslužan za to što se neuronske mreže često koriste i što su popularna metoda u različitim područjima. Standardni algoritam „backpropagation" mreže uključuje optimizaciju pogreške koristeći se deterministčkim algoritmom gradijenta pada (Milosavljević, 2005). Glavni nedostatak tog algoritma jest taj što se često pronalazi lokalni umjesto globalnog minimuma pogreške. Stoga novija istraživanja nude poboljšanja korištenjem nekih drugih deterministčkih (npr. metoda drugog reda) ili stohastičkih metoda (npr. simulirano popuštanje – *engl. simulated annealing*). Struktura mreže sastoji se od ulaznog sloja, izlaznog sloja i barem jednog skrivenog *feedforward* sloja.

Arhitektura testirane neuronske mreže u našem radu sastojala se od tri sloja. Broj jedinica (neurona) u skrivenom sloju i trajanje procesa učenja dobiveni su postupkom kros-validacije. Modeliranje je provedeno u tri faze: (a) priprema i modeliranje podataka, (b) uhodavanje i testiranje neuronskih mreža i (c) interpretacija rezultata i odabir najboljeg modela. Uhodavanje mreže provedeno je na testnom uzorku (70% ukupnog uzorka), duljina učenja mreže dobiva se kros-validacijskim postupkom u kojemu mreža uči na testnom uzorku iterativnim postupkom, te koristeći se raznolikim parametrima (npr. različitim brojem skrivenih neurona), a svaka kombinacija testira se na validacijskom uzorku (30 % ukupnog uzorka). Cilj je utvrditi duljinu procesa učenja i strukturu mreže koja daje najbolji rezultat na validacijskom uzorku. Na kraju, mreža dobivena na takav način testira se na testnom uzorku (30% ukupnog uzorka), a rezultat testne faze koristi se kao mjera uspješnosti mreže. Što se tiče izlaznih funkcija, testirana je sigmoidna funkcija (jer se ona također koristi i u logističkoj regresiji), a kod učenja se koristilo delta pravilo s momentom od 0,7 i dinamičkim koeficijentom učenja od 0,1 do 0,9. Srednja kvadratna pogreška (MSE) i korijen srednje kvadratne pogreške (RMSE) često se koriste za izračunavanje pogrešaka neuronskih mreža u fazi uhodavanja.

Budući da se u ovom radu analizira problem klasificiranja, stopa klasificiranja se izračunava nakon faze testiranja neuronske mreže, i to za svaku grupu posebno. Prosječna stopa klasificiranja se također izračunava i ujedno se smatra mjerom za procjenu uspješnosti modela neuronske mreže. Stopa klasificiranja svake grupe je postotak događaja koje je mreža točno smjestila u tu grupu.

Među prvim autorima koji su u ovom području koristili neuronske mreže su Hardgrave, Wilson i Kent (1994). Oni su uspoređivali neuronske mreže s tradicionalnim statističkim tehnikama korištenima za predviđanje uspjeha studenata na sveučilištu. Daljnji rad istih autora naglašava činjenicu da se odluka o upisivanju studenata na fakultet temelji na brojnim čimbenicima, te je potrebno razviti modele predviđanja koji će fakultetima i sveučilištima omogućiti upisivanje studenata koji imaju visok potencijal za uspješno studiranje. Njihovo istraživanje pokazuje sljedeće: (a) tehnike klasificiranja pogodnije su za predviđanje uspjeha studenata u usporedbi s metodama predviđanja; (b) predviđanje uspjeha ili neuspjeha studenata na fakultetu nije dovoljno precizno ako se koriste samo uobičajeni podaci koji opisuju studenta; (c) neparametrijski postupci, kao što su neuronske mreže, ostvaruju barem jednako precizne podatke kao i tradicionalne metode, pa se mogu smatrati vrijednim potencijalom za buduća istraživanja koja će se provoditi u tom području.

Istim problemom donošenja odluka o upisivanju kandidata na fakultetske studije bavili su se i Naik i Regotman (2004). Oni su proveli istraživanja o uspjehu studenata u MBA studijima. Koristili su neuronske mreže, logit i probit modele za predviđanje uspjeha studenata koji su se upisali na MBA studij. Neuronske mreže klasificirale su studente u uspješne i neuspješne na temelju prosječne ocjene koju su studenti imali na dodiplomskim studijima, rezultata GMAT testa, modula na dodiplomskim studijima, dobi i drugih varijabli. Rezultati pokazuju da su neuronske mreže jednako uspješne

kao i druge tehnike, ali da se njihovo korištenje preporučuje baš u ovom području zbog njihovih brojnih prednosti.

Istraživanje koje su proveli Sulaiman i Mohezar (2006) bavi se istom temom, ali ujedno ide i korak dalje u određivanju čimbenika koji su ključni za uspjeh. Njihov model pokazao je da je ranije stečena prosječna ocjena studenata najvažniji prediktor njihova uspjeha u budućnosti, dok varijable poput dobi, etničke pripadnosti, spola i godina radnog iskustva nisu važni za uspjeh u studiranju.

Autori iz Malezije (Zaidah i Daliela, 2007) uspoređivali su neuronske mreže, stabla odlučivanja i linearnu regresijsku analizu u predviđanju uspjeha studenata. Oni su mjerili uspjeh ukupnom prosječnom ocjenom stečenom tijekom studiranja, dok su kao ulazne varijable koristili demografski profil studenata i prosječnu ocjenu stečenu u prvom semestru dodiplomskog studija. Rezultati su pokazali da su sve tri metode postigle preciznost veću od 80%, dok su neuronske mreže postigle bolju preciznost od ostalih dviju metoda.

Autori s Fakulteta organizacijskih nauka u Beogradu (Vukićević, Išljamović, Jovanović, Delibašić i Suknović, 2011) koristili su neuronske mreže za predviđanje uspjeha studenata, služeći se podacima o studentima koji su obuhvatili osobne podatke i podatke o uspjehu u prvoj godini studija. Ukupno je korišteno 14 varijabli. Kao rezultat, neuronska mreža dobila je varijablu koja predstavlja predviđenu srednju ocjenu za svakog studenta na kraju studiranja. Testirano je šest algoritama od kojih je, prema svim kriterijima, najbolji bio *Exhaustive Prune*. Karamouzis i Vrettos (2008) koristili su dvanaest ulaznih parametara za predviđanje uspjeha 307 studenata. Uspjeh predviđenog uspjeha bio je na stupnju od 70,27% za uspješne studente i 66,29 % za neuspješne.

Metodologijski okvir istraživanja

Ciljevi i zadaci istraživanja

Cilj ovoga rada je odrediti važne čimbenike koji imaju utjecaj na uspjeh studenata, iskazan prosječnom ocjenom. Za tu smo se svrhu koristili dvjema metodama rudarenja podataka koje su bile pogodne za klasificiranje: logističkom regresijom i stablom odlučivanja. Namjeravali smo ispitati kvalitetu svake od tih dviju metoda.

Hipoteza i varijable istraživanja

Metode rudarenja podataka omogućavaju nam relativno precizno predviđanje uspjeha studenata na Pedagoškom fakultetu u Bijeljini na temelju procjene vjerojatnosti važnosti pojedinačnih varijabli.

Varijable koje su bile ulazne varijable u modelu:

1. Kriterijske varijable: stečena prosječna ocjena (do 7,5 – manje uspješno; od 7,51 do 10 – uspješno)
2. Nezavisne varijable: opća inteligencija, motivacija za studiranje, spol, mjesto studiranja, informacije o stipendiji, vrijeme uloženo u studiranje, literatura,

izvori i sredstva korištena u studiranju, prisutnost na predavanjima, prisutnost na tutorijalima, pristupanje ispitima, stavovi o važnosti ocjene koju će student dobiti na ispitu, kvaliteta predavanja, kvaliteta tutorijala, kvaliteta predmetnog kurikula, kvaliteta predavača, kvaliteta procesa ocjenjivanja znanja.

Opseg istraživanja:

– Populacija (svi studenti Pedagoškog fakulteta u Bijeljini)
– Uzorak od 354 studenta na drugoj, trećoj i četvrtoj godini

Obrada podataka: logistička regresija, stablo odlučivanja, CART, neuronske mreže

Rezultati i rasprava

Primjena logističke regresije u predviđanju uspjeha studenata na studijima

Postoji nekoliko metoda procjenjivanja u logističkoj regresiji, no najčešće korištena metoda s možda i najnižim stupnjem rizika u smislu potvrđivanja hipoteze je METHOD = BSTEP(LR) za regresijsku stupnjevitu analizu *(stepwise analysis backward).* Ta metoda sastoji se od mogućnosti testiranja „log-vjerojatnosti" zadanom varijablom, dobivenom jednadžbom. Ukupna statistika testiranih događaja pokazana je u Tablici 1.

Tablica 4.

Cijeli test modela prikazan je u Tablici 2 – Ukupan test modela. U našoj logističkoj regresiji BSTEP(LR) sve varijable ušle su u jednadžbu na početku, a potom je model testiran u deset koraka. Kao što se može vidjeti, sve vrijednosti bile su dane na početku kao „korak", „model" i „rezultat" s jednakim stupnjem važnosti od 0,00. Tijekom početnog koraka sve varijable su prisutne u modelu. U drugom koraku jedna varijabla koja nije imala statističku važnost uklonjena je iz modela (0,95). Slično tome, još nekoliko varijabli koje nisu imale statističku važnost bile su uklonjene iz modela. Tijekom procesa koji se sastoji od deset koraka, chi-kvadrat test bio je postupno smanjen, što taj model i zahtijeva. Stoga, nakon početne vrijednosti od 68,79, dosegli smo smanjenu vrijednost od 64,22.

Tablica 2.

Iz gornje tablice nismo mogli zaključiti koje su varijable bile uklonjene iz modela. Tek nam je uvid u Tablicu 3 pomogao kako bismo mogli uočiti pravilnosti u uklanjanju svake pojedinačne varijable iz jednadžbe. Prva varijabla koja je uklonjena iz modela bila je „spol". U sljedećem je koraku uklonjena varijabla „kvaliteta tutorijala". U trećem koraku uklonjena je varijabla „prisutnost na predavanjima". U četvrtom koraku varijabla „kvaliteta predavanja" nije značajno utjecala na poboljšanja ukupnog modela. U sljedećim koracima varijable su bile uklonjene ovim redom: kvaliteta kurikula, način studiranja i mjesto studiranja.

Tablica 3.

Sljedeće varijable bile su uključene u finalni model procjene vjerojatnosti: važnost ocjene (0,00), polaganje ispita (0,00), intelektualne sposobnosti (0,01), stipendija (0,04), prisutnost na tutorijalima (0,05) i trajanje studija (0,09). U sljedećoj su tablici prikazani pseudo R-kvadrati. Cox i Snell indeksi variraju od 0 do 0,75, dok se korekcija obavlja jedino Nagelkerke indeksom i dovodi stupanj unutar raspona vrijednosti od 0 do 1. Naravno, ovdje se R ne može smatrati koeficijentom determinacije u linearnoj regresiji, jer se radi o proporcionalnom sudjelovanju pojedinačnih varijabli u ukupnoj vjerojatnosti. Sa svakim novim stupnjem postupne regresije povećan je ukupan rezultat obuhvaćene varijance. U krajnjem modelu Cox i Snell indeksi imaju vrijednost do (0,24), dok ta vrijednost nakon korekcije Nagelkerke indeksom iznosi (0,32). To bi se moglo smatrati zadovoljavajućim rezultatom.

Tablica 4.

Predviđene vrijednosti zavisnih varijabli prikazane su u Tablici 5. i utemeljene su na modelu potpune logističke regresije. Ta tablica pokazuje koliko je slučajeva točno predviđeno, a koliko nije. Cilj te regresije od deset koraka bio je povećati postotak uspješnog predviđanja. U prvom koraku bilo je 50 slučajeva za koje se očekivalo da će imati vrijednost 1, ali su ostvarili vrijednost 2. Bila su 42 slučaja, od ukupno 170, za koje se očekivalo da će imati vrijednost 2, no ostvarili su vrijednost 1. Stoga, ukupna varijanca točnog predviđanja je 74,8 %. U finalnom modelu, od 184 slučaja koji trebaju imati vrijednost 1, samo je 49 slučajeva imalo vrijednost 2. Od 170 slučajeva koji trebaju imati vrijednost 2, 40 slučajeva je imalo vrijednost 1.

Tablica 5.

Da ne bismo ostali na razini kvantitativnih podataka, proveli smo detaljnu statističku i matematičku analizu svakog pojedinog podatka. Koristeći se sljedećom jednadžbom, izračunali smo vjerojatnost svake varijable koja ima značajan utjecaj na stupanj uspješnosti u studiranju.

U desetom koraku dobili smo vrijednost konstante b_0 (b_0 = - 3,69). Konstanta je prvi ulazni element u eksponencijalnoj funkciji. Izračunali smo vjerojatnost svake ulazne varijable u jednadžbi. Najveća vjerojatnost, P(x) = 0,88, ima prediktor – važnost ocjene, čiji je koeficijent b 0,94. Zatim slijedi polaganje ispita (P(x) = 0,85; b_1 = 1,00) i intelektualna sposobnost koja ima sljedeću vjerojatnost (P(x) = 0,81; b_1 = 0,93). Nakon toga slijedi stipendija, koja je po svojoj težini na četvrtom mjestu u jednadžbi (P(x) = 0,70 b_1 = 0,49). Zatim slijedi prisutnost na tutorijalima (P(x) = 0,61 b_1 = 0,63), a na kraju duljina studiranja (P(x) = 0,59 b_1 = -0,38).

Ako pogledamo važnost svake varijable za ukupan stupanj vjerojatnosti predviđanja, možemo primijetiti da redovito „polaganje ispita“ povećava vjerojatnost uspjeha u studiju za pet puta (Exp. B 5,28). Ako je „dobivena ocjena” važna studentu, tada se stupanj uspješnosti povećava četiri puta (Exp. B 4,22). Visok stupanj intelektualnih sposobnosti povećava vjerojatnost uspjeha za četiri puta (Exp. B 4,12), dok redovito

„pohađanje tutorijala" povećava stupanj uspješnosti za tri i pol puta (Exp. B 3,25). Ako student ima „stipendiju", stupanj uspješnosti povećava se za dva i pol puta (Exp. B 2,59), a ako student uči barem dva sata dnevno, vjerojatnost njegova uspjeha povećava se za jedan put (Exp. B 1,06).

Da bi se potvrdili rezultati dobiveni regresijskom analizom, moglo bi se reći da 85% uspješnih studenata redovito polaže ispite, u usporedbi sa stopom od 53 % polaganja ispita studenata koji su manje uspješni. Uspješni studenti pohađaju tutorijale u 88 % slučajeva, u usporedbi sa 63 % studenata iz suprotne grupe. 79 % studenata navelo je da smatraju bitnim koju će ocjenu dobiti, u usporedbi s 45 % studenata iz grupe manje uspješnih studenata. 85 % njih uči u prosjeku dva do pet sati dnevno, u usporedbi sa 60% studenata iz suprotne grupe. 27 % uspješnih studenata ima stipendiju, u usporedbi s 11% studenata iz suprotne grupe.

Rezultati dobiveni regresijskom analizom u našem istraživanju pokazuju 74,8 % uspješnih predviđanja uspješnosti studenata na Pedagoškom fakultetu. Varijable koje pretežno doprinose ukupnoj vjerojatnosti predviđanja mogle bi se klasificirati u tri skupine: didaktičke (polaganje testova i ispita, prisutnost na tutorijalima, duljina studiranja, važnost postignute ocjene), osobne karakteristike (intelektualna sposobnost) i društvene (dobitak stipendije).

Primjena stabla odlučivanja u predviđanju uspjeha studenata tijekom studiranja

S ciljem predviđanja uspjeha u studiranju studenata na Pedagoškom fakultetu u Bijeljini koristili smo se CART stablom odlučivanja. U nekim prijašnjim istraživanjima CART stablo pokazalo se prikladnim alatom u predviđanju uspjeha studenata u studiranju.

Iz priloženog pregleda vidljivo je da je prosječna stopa klasificiranja dobivena stablom odlučivanja 71,2 %, što je manje od prosječne stope dobivene logističkom regresijom (74,8). Stablo je posebno točno u prepoznavanju „slabijih" studenata čija je prosječna ocjena niža od 7,5, pri čemu je stopa klasificiranja 74,1 %. Nešto niža stopa dobivena je za grupu 1 – „bolje studente" (68,4 %).

Rezultati klasificiranja studenata stablom odlučivanja u poduzorku za testiranje također bi se mogli ilustrirati matricom grešaka, koja u svojim stupcima pokazuje stvaran broj studenata koji pripadaju kategoriji s nižom (2) ili višom (1) prosječnom ocjenom, dok njezini redovi pokazuju broj studenata koje je model stabla odlučivanja klasificirao ili u Kategoriju 2 ili u Kategoriju 1. Dijagonala matrice grešaka pokazuje broj studenata koje je model ispravno klasificirao. Iz te tablice vidljivo je da ukupno 44 studenta s nižom prosječnom ocjenom (Kategorija 2) nije bilo ispravno klasificirano stablom odlučivanja, dok je 126 studenata bilo ispravno klasificirano. Situacija je bila drugačija za Kategoriju 1 u kojoj je 126 studenata bilo ispravno klasificirano, dok je 58 studenata bilo svrstano u pogrešnu kategoriju. Cijeli model ispravno je klasificirao 252 od 354 studenata. Stablo je nešto preciznije od regresijske analize i izdvaja samo

dvije varijable koje imaju utjecaja na ukupnu vjerojatnost, a to su: polaganje ispita i važnost ocjene.

Primjena neuronskih mreža u predviđanju uspjeha studenata na studijima

U ovom je radu isprobana višeslojna neuronska mreža. Za spomenutu strukturu bila je dobivena stopa uspješne klasifikacije za svaku grupu u probnom i testnom uzorku. Prosječna stopa klasifikacije testnog uzorka, opisana u drugom poglavlju, korištena je kao mjera uspješnosti modela.

Tablica 6.

Da bi se riješio navedeni problem u neuronskoj mreži, korištena su 33 skrivena neurona, kao i funkcija logističkog transfera, delta-bar-pravilo učenja i mreža naučena na maksimalnih 1000 epoha. Broj skrivenih neurona koji daju najmanje pogreške u fazi kros-validacije bio je 33. Dobiveni rezultati pokazuju da je u testnom uzorku bilo 76,4 % točno klasificiranih slučajeva, dok je 23,6 % slučajeva svrstano u pogrešnu kategoriju. Ako se pogledaju određene stope klasifikacije svake pojedinačne kategorije, tada se može primijetiti da je stopa klasifikacije za Kategoriju 2 („slabiji" studenti) 84,4 %, dok je stopa klasifikacije za Kategoriju 1, tj. za „bolje" studente 63 %. Veća preciznost klasifikacije za „slabije" studente ukazuje na to da studenti s prosječnom ocjenom nižom od 7,5 imaju zajedničke karakteristike koje je model neuronske mreže uspio prepoznati i u kojima je taj model uspio pronaći veze, u usporedbi sa slučajem studenata s prosječnom ocjenom višom od 7,5. Da bi se interpretirali rezultati dobiveni neuronskim mrežama, neophodno je imati objašnjenje za svaku varijablu u procesu predviđanja uspjeha. Vjerojatnost svake nezavisne varijable u cijelom definiranom modelu prikazana je u Tablici 7. Slično modelu koji je bio dobiven logističkom regresijom, varijabla „važnost ocjene" je ulazna varijabla za model predviđanja s vjerojatnošću od 100 %. Varijable „duljina studiranja", „polaganje ispita" i „intelektualna sposobnost", s vjerojatnošću većom od 80 %, važi su elementi u modelu. Druge varijable imaju slabiji utjecaj na model predviđanja. Bez obzira na veličinu uzorka, trebalo bi istaknuti da dobiveni rezultati ukazuju na nedostatke u sustavu obrazovnog procesa jer varijable kao što su „kvaliteta predavača", „prisutnost na predavanjima", „kvaliteta tutorijala" i „kvaliteta kurikula" imaju jako nisku prediktivnu vrijednost u predviđanju uspjeha studenata. Daljnje analize tih rezultata mogle bi dati detaljnije odgovore na ta otvorena pitanja.

Tablica 7.

Zaključak

Sva tri modela rudarenja podataka predstavljaju mogućnost točnog predviđanja uspjeha studenata. Rezultati predviđanja koji su dobiveni za sva tri primijenjena pristupa otprilike su iste vrijednosti. Najbolji rezultati postignuti su korištenjem neuronskih mreža (76,4 %), zatim slijede rezultati dobiveni logističkom regresijom (74,8 %), dok su nešto slabiji rezultati predviđanja dobiveni primjenom stabla

odlučivanja (71,2 %). U prijašnjim istraživanjima, kada su se koristili samo opći i demografski podaci o studentima, model je imao viši stupanj preciznosti u usporedbi sa stupnjem preciznosti koji smo mi postigli u našem istraživanju. Rezultati koje smo mi dobili odražavaju stanja našeg obrazovnog sustava u području visokog obrazovanja. Vrlo je jasno da je reforma obrazovnog sustava uzrokovala promjene u ponašanju studenata. Ono što zabrinjava jest činjenica da se velik broj studenata ne brine za ocjenu koju će dobiti na ispitu, ne smatra važnim prisustvovati predavanjima jer, očito, nemaju visoko mišljenje o mogućnostima učenja tijekom odgojno-obrazovnog procesa.

Pokazalo se da uspješniji studenti posvećuju više pažnje učenju, polažu ispite (što ukazuje na primjenjivost studija) i da im je važna ocjena koju će dobiti. To je također područje koje ima priliku za poboljšanja u procesu planiranja rada na sveučilištu. Naravno, ti čimbenici trebali bi biti povezani s mogućnošću dobivanja stipendije.

Neophodno je provesti analizu svih dobivenih rezultata, obogatiti sustav predviđanja novim varijablama i tada korigirati provedbu obrazovnog procesa. Neki savjeti za poboljšanje obrazovnog procesa koji su se nametnuli u ovom istraživanju mogli bi se svrstati u sljedeće skupine:

- proces učenja trebao bi biti intenzivniji
- češće bi trebalo organizirati rad u manjim skupinama
- trebalo bi pronaći bolju povezanost teorijskih i praktičnih sadržaja
- trebao bi postojati sustav redovnog praćenja uspjeha učenika putem pisanih testova i drugih modela ocjenjivanja znanja
- studenti bi trebali imati podršku u prilagodbi studiju
- opseg i sadržaj literature koja se koristi u studiranju i učenju trebao bi biti dopunjen i tako prilagođen ciljevima i načinima ocjenjivanja
- predavači bi trebali biti motivirani da bi mogli povećati svoj interes za predavanja i za svoje studente
- stupanj i kvaliteta znanja predavača u području metodike nastave i ocjenjivanja trebali bi se povećati
- vještine komuniciranja i predanost predavača metodici nastave trebali bi se povećati
- sustav ocjenjivanja nastavnog procesa i predavača trebao bi se modernizirati, kao jedan od načina praćenja kvalitete nastave i uvođenja promjena.

Kako poboljšati model predviđanja uspjeha studenata na studijima?

- uvesti nove varijable, tj. ishode studija
- proširiti uzorak
- uključiti i druge predavača i pedagoške fakultete u uzorak
- stvoriti inteligentan način potpore obrazovnog sustava na sveučilištima.

Moglo bi se reći da rudarenje podataka ima potencijal za praćenje obrazovnih postignuća na sveučilištima. Ono nudi praktična rješenja u području predviđanja ponašanja studenata u budućnosti i stvara temelj za promjenu određenih komponenti obrazovnog sustava da bi se popravio opći uspjeh.