ISSN 1512-5785 **TRANZICIJA / *TRANSITION*** UDK 330-142
Godina XVI Vitez-Tuzla-Zagreb-Beograd-Bukarest, juli - decembar 2014. Broj 34
Vol. XVII Vitez-Tuzla-Zagreb-Beograd-Bucharest, July - December 2014. No. 34

Original scientific paper
Izvorni znanstveni rad

Edin Osmanbegović[*]
Mirza Suljić[*] [1]
Hariz Agić[2]

DETERMINING DOMINANT FACTOR FOR STUDENTS PERFORMANCE PREDICTION BY USING DATA MINING CLASSIFICATION ALGORITHMS

Abstract

The central problem in the process of a discovering knowledge from data, in the field of educational data mining, is to identify a representative set of data, on whose basis a classification model will be constructed. This paper presents the research results in reduction of data dimensionality, in the classification problems of prediction of student's performances on the example from high schools, in Canton Tuzla. In this paper are shown different algorithms that are used to reduce the dimensionality of data and to development of a data mining model for predictions of performances of students, on the basis of their personal demographic and societal features. It was found that algorithms Random Forest and J48 generate classification model with and accuracy higher than 71%.
Keywords: *data mining, educational data mining, predicting performance, student success, secondary schools*

Introduction

For educational institutions, whose aim is to contribute to improvement of quality of education, the success of creating human capital is the subject of a constant analysis whose effects are long-term investment in an educational process that will result in other necessary products. Therefore, the quality of education at the level of primary and secondary schools represents one of the most important factors of forming a future full

Primljeno: 20.09.2014; Prihvaćeno: 16.10.2014
Recceived: 20-09-2014; Accepted: 16-10-2014

[*] **Ph. D. Edin Osmanbegović, associate professor, Faculty of Economics in Tuzla**
[**] **Mirza Suljić, mag. oec., Faculty of Economics in Tuzla**
[***] **Ph.D. Hariz Agić, Prosvjetni zavod, Tuzla**

member of the community. Regarding, the quality of education is one of the most important carriers of continuity and intergenerational exchange of basic principles and social values, but also actuator of necessary review and adaptation of those principles and values to the new circumstances. Because of this, relevant data and information are gathered regularly, discussed on the certain bodies and certain measures are taken in order to maintain quality, The quality of educational institutions involves providing services, which mostly meets the needs of students, teaching stuff and other participants of educational system.

Current qualification statistics of people capable for work in Bosnia and Herzegovina, is not very promising. Almost half of the working population in Bosnia, still has finished only primary school or less (46,9%), 46,4% has finished only secondary school, and only 6,7 has a higher level of education. In this moment, there are no reliable data about the literacy of the population in Bosnia. There is no reliable data even on the educational structure of the population. According to results of the questionnaire MICS (UNDP, 2006), in pre-school education 9% of the population is included, and net rate of inscription in the primary education in Bosnia is 97,2%. Net rate of inscription into secondary school is 76,2% and 54% of the students finished secondary school in regular term. Net generational rate of inscription of students is only 24% (Branković & Arapović, 2010), while the European average of leaving school early is only 12,7% (Group, 2013).

On the other hand, it is clear that thanks to today technology, gathering data stops being a problem, and in the focus of interest comes their analysis and obtaining valuable information from the data (knowledge). Central for that problem is a process of knowledge discovering from database (KDD). As the process of knowledge discovery represents a computer processing data from different perspectives in order to extract implicit and interesting patterns, trends or information from data, it can help every participant of educational process in order to improve decision making and optimize performing of the students. By determining dominant factors of student performances, it is important for all participants of educational system, because it can help with better understanding of learning process and it focuses on discovering, recognizing and explaining educational phenomenon and ultimately improving them. So KDD, in educational systems builds a cycle which consists of forming hypothesis, testing and training, i.e. the application can be focused on different actors of the educational process in accordance to specific needs of students, teachers, administration, supporting administration and community (Romero & Ventura, 2007).

In recent years, an increased interest in using data mining to search for answers to scientific questions for educational purposes, an area of inquiry recognized as educational data mining (also named as "EDM"). A first literature review of data mining in education was provided by Romero and Ventura (2007) (Romero & Ventura, 2007), covering the research efforts in the area between 1995 and 2005, followed by Baker and Yacef (2009), in the period between 2005 and 2009. A very comprehensive literature review of EDM research can be found in (Romero & Ventura, 2010).

One of the problems in education which are solved by applying data mining is prediction of student performances, whose aim is predicting values of an unknown variable (result or grade) which describes a student. This is a difficult problem to solve due to the large number of factors that can bear influence on students' performance, such as demographic, cultural, social, or family factors, socio-economic status, psychological profile, previous schooling, ets. According to Bakare (1975) summarized the factors and variables affecting students' performance into the intellective and non-intellective factors, emphasizing that the intellectual abilities were the best measure. He categorized causes of poor student performance into four major classes namely: Causes resident in society, Causes resident in school, Causes resident in the family and Causes resident in the student. It is important to notice that most of the current research on the application of EDM for predicting student performance has been applied primarily on data coming from of higher education or university students, while secondary education does not apply so much importance.

There are several studies oriented toward use data mining techniques on data that coming from secondary school students such as: 264 students from the gifted education program focuses on both the elementary and secondary school levels in Singapore is used in real-life data mining application for select the targeted students much more precisely for remedial classes (Yimin, Bing, Wong, Yu, & Lee, 2000); 222 students from 12 different schools to assess the relationship between achievement and involvement in additional after-school activities of the secondary school students in Spain (Moriana, Also, Alcala, Pino, Herruzo, & Ruiz, 2006); sample of 300 students (225 males, 75 females) from the colleges located in Province Punjab of Pakistan used to find the factors that highly correlated with the student performance (like mother's education and student's family income) (Hijazi & Naqvi, 2006); real data coming from secondary school students in the Czech Republic is used to present a novel visualization technique that allows the user to interactively explore and analyze differences in mean values of analyzing attributes (Zoubek & Burda, 2009); 400 students (200 boys and 200 girls) selected from senior secondary school, with a main objective to establish the prognostic value of different measures of cognition, personality and demographic variables for success at higher secondary level in science stream (Zebun, 2005); Cortez and Silva (2008) attempt to predict student failure by applying and four data mining algorithms on two datasets related to Mathematics (with 395 examples) and the Portuguese language classes (649 records) from two secondary schools of Alentejo region (Portugal); 772 student records from five different schools in three different districts of Tamilnadu (India), were used to developing predictive data mining model to identify the slow learners and study the influence of the dominant factors on studenst performance, using the popular CHAID decision tree algorithm (Ramaswami & Bhaskaran, 2010); using real data about 670 high school students from Zacatecas (Mexico) different data mining approaches are proposed for predicting student failure at school (Márquez-Vera, Cano, Romero, & Ventura, 2013); sample of 907 students from secondary schools located in Tuzla Canton (Bosnia and Herzegovina) used to predict

students final grade by applying and comparing four data mining algorithms, Decision Tree (J48), Random Forest, Naive Bayes and Multilayer Perceptron (Osmanbegović, Agić, & Suljić, 2014).

In this paper different techniques of data mining suitable for classification have been compared: Rules-based, Trees-based, Functions-based and Bayes-based algorithms. The study is grounded along the survey conducted during the first half of the secondary schools in Tuzla Canton, school year 2011/12 and 2012/13. This analysis was conducted after the training and testing of the algorithms, making it possible to draw conclusions on possible predictors of students' success. The objective of our study is to find a reply to the following research questions:

- What attributes are a dominant factor for students performance prediction ?
- What DM algorithm are best for predicting student performance ?

Knowledge Discovery from Data

Conducted research is based on CRISP-DM (Cross-Industry Process for Data Mining) model, which is neutral regarding the used tools and industry in which it is applied, which is considered to be standard in the area of knowledge discovery from data (Chapman, i dr., 2000). This methodology represents cyclical approach, which consists of six basic phases: Business understanding, Data understanding, Data preparation, Modeling, Evaluation and Deployment. There is a number of internal feed-back between phases, which are a result of a complex nonlinear nature of the process, so they ensure the achievement of consistent and reliable results. For the realization of the project, open source software Weka is used, which offers wide specters of different algorithms suitable for use in the process of discovering knowledge from data (Witten, Frank, & Hall, 2011).

Business understanding – first of all, in this initial phase, available literature was reviewed in order to study the existing problems in educational institutions which are solved by applying algorithms of data mining in similar researches. With decision maker on all levels (from school to the competent ministries) a formal interview was conducted, with the aim of learning about specific problems which are not solved yet, but are considered to be very important for the improvement of educational system and efficient management of the same. The initial problem is transformed into KKD task, i.e. the task for determining dominant factors for prediction of student performances by using data mining algorithms for classification. Analyzing available data with selected data mining methods for classification, students are classified into two classes. The main aim of the project is to discover a high potential of KDD, which refers to the optimal use of data methods and data mining technique in the analysis of historical data.

Data understanding – this phase in the realization of research begins with initial data collection, and continues with activities of data selection and pre-processing, and it is a key activity within each data mining project, and significantly affects the quality of final results. After revising and reviewing procedure for collecting and storing data about the success of the students, it is confirmed that data are mostly organized in a database

information system EMIS (Education Management Information System). The specified database was not available to the researchers, so the data were collected through a questionnaire survey conducted during the second semester of secondary schools in the Tuzla Canton, year 2011/12 and 2012/13.

Data preparation – This is a very intense process and it can be done only by the persons who understand well the aims of discovering knowledge, which have access to relevant data and understand their meaning and conditions under which they were collected. Data collected through questionnaire survey were entered into a database, which was transformed into a suitable format for the analysis in the WEKA software package. In year 2011/12 and 2012/13. During the investigation, secondary schools in Tuzla Canton attended about 20.000 students, and the survey method included 7,05% of students. After eliminating incomplete data, the sample consisted of 1210 students who, at the research time, attended classes. The sample was described with 19 variables as an input to the model whose names and codes are shown in the Table 1.

Table 1 Input attribute

Attribute	**Coding**
sex (SP)	Nominal: M - male or Z – female
age (ST)	Numeric: from 14 to 18
type of school (TS)	Nominal: G, MS or O
address (A)	Binary: 1-urban or 0-rural
parent's cohabitation status (SR)	Numeric: coding form 1 to 4
mother's education (OM)	Numeric: coding form 1 to 5
mother's job(ZM)	Nominal: A, B or C
father's education (OO)	Numeric: coding form 1 to 5
father's job (ZO)	Nominal: A, B or C
family size (F)	Numeric: coding form 1 to 5
reason to choose this school (RI)	Numeric: coding form 1 to 5
home to school travel time (US)	Numeric: coding form 1 to 5
type of travel from home to school (DS)	Nominal: A, B or C
monthly scholarship (S)	Binary: 1-yes or 0-no
weekly study time (V)	Numeric: coding form 1 to 6
internet access at home (I)	Binary: 1-yes or 0-no
importance of grades obtained (VO)	Numeric: coding form 1 to 3
years of schooling (GS)	Numeric: coding form 1 to 4
average income of the parent's (PP)	Numeric: coding form 1 to 5
Performance score (OU)	Nominal: A, B

The challenge in the presented KDD research is to build a model by which it would be possible to predict the category of the student, for an unknown sample, through attributes

about student's characteristics. The selected, targeted variable in this case, i.e. concept for learning algorithms of data mining "class of the student". Aimed variable is built on the basis of original numerical parameters i.e. average score. Estimated variable has two different values, which are in correspondence with two classes in which the student are classified - weak and strong. Since, in the educational system of Bosnia and Herzegovina, the scale of five levels is used for the evaluation of student's performances in secondary schools, students with average grade which is lower than 4.00 are classified as "Weak", and students with equal or higher than 4.00 are considered as "Strong". It is noticeable that the mistake of prediction in the earlier research (Osmanbegović, Agić, & Suljić, 2014) was much higher because of the unbalanced input data set, so the classes with small number of instances trained harder. Gained results are presented in the following chapter.

Modeling – algorithms for building a model will classify students into two classes (categories), depending on their performances and on the basis of data collected through the questionnaire. Methods of classifying data represent the process of learning the function that maps the data into one of the few predefined classes. To every algorithm for classification, which is based on inductive learning, input data set was given, which consists of vector value of the attribute, and the corresponding class. Several different algorithms for classification were applied during this research. They were chosen because they have a potential to give good results. Popular WEKA classifiers (with its default settings were used, unless it was stated differently) were used in this research, including common Decision trees-based algorithms, two Bayes-based algorithms, four Rules-based algorithms and four Functions-based algorithms. Research results are presented in the next chapter.

Evaluation – Evaluation of results acquired by conducted algorithms of data mining is primarily based on the evaluation of experimental results. In order to create and evaluate classification model, we measure his efficiency i.e. the possibility of classifier to correctly classify a large number of samples from the test data set. Because of the specificity in the data sets for studying, or testing, which are not characteristic of class but the defect caused by the sets choice, evaluation by using a test set can result in imprecise evaluation of frequency of mistakes. The main way to avoid these anomalies is multiple repetitions of evaluation processes on the test sets by using different, randomly selected sets, for studying and testing, and averaging acquired estimation of frequency of mistakes (Kohavi, 1995). For the evaluation of classification accuracy, the cross validation method is used. The process of studying and evaluation is repeated k time, every time by using one subset as a test set. Estimation of accuracy of prediction by cross validation is a random number which depends on the distribution of samples to subsets. Several different metrics were used in this research to evaluate the performances of the algorithms. The performance of classifiers involves Accuracy, Error rate, Precision, Recall and F-Measure.

Deployment – This step is important in the case of everyday use of the model in the educational system. The acquired models can be used in the purposes of: process

monitoring, data evaluation, support in decisions. In every case, it is important that contractor is fully aware of the models limits and all the actions that are prerequisite for its successful implementation.

Experiment results and discussions

For the purposes of this research a software package WEKA was used, and previously described data set. To test the accuracy of acquired classification models we used a method 10-fold cross validation. The next experiments were done:

- selection of attributes and
- classification and evaluation.

Selection of attributes

Once the data is collected and separated into a unique relation, it is necessary to gain insight into their structure and informative value, so they would be prepared well for the appliance of data mining algorithms and methods. In the preparatory phase, the evaluation of input attributes was done in relation to the output attributes. The objective of the evaluation process and selection of the attributes is detaching unimportant and redundant attributes from the data set for studying. Filter methods include techniques for evaluation attributes value relying on heuristics based on the general data characteristics. For data mining, filter methods are more practical solution for certain reasons: choice and evaluation of attributes is shorter, independency from algorithm of machine studying enables the use in combination with and technique of modelling data.

Filter methods InfoGain and GainRatio with Ranke search method were applied. InfoGain represents an estimation of attribute value by measuring his informativeness in relation to the class. Gain Ratio represents an assessment of attribute value, by measuring his relative informativeness in relation to the class. Attributes, with the assessment smaller than 0.01 should be excluded from analyzed data set. The results of assessment and ranking attributes on the basis of their individual values are shown in the Table 2.

Table 2. Estimation results and attribute ranking

ATTRIBUTE	GainRatio	InfoGain	
V	**0,0748947**	**0,095567**	+
GS	**0,0471055**	**0,074099**	+
ST	0,0258588	**0,0392569**	+
OO	0,0270311	**0,0368014**	+
ZO	0,0172055	0,0204491	+
TS	0,0139914	0,0187145	+
I	0,0238647	0,0156845	+
VO	0,0260715	0,0156051	+
OM	0,0120987	0,0112837	+

DS	0,0082165	0,0085921	-
PP	0,0116932	0,0077145	+/-
S	0,0179593	0,0076635	+/-
ZM	0,0001008	0,0001188	-
SP	0,0000592	0,0000592	-
A	0	0	-
SR	0	0	-
F	0	0	-
RI	0	0	-
US	0	0	-

It can be seen from Table 3 that attributes with the highest rank are V, GS, ST and OO, and with the lowest A, SR, F, RI and US. Finally, we selected as the best attributes the first eleven and nine attributes in the ranking, because that attributes with the estimate of less than 0.01 excluded from further analysis.

Classification and evaluation

In the second experiment, we have carried out some experiments in order to evaluate the performance and usefulness of different classification algorithms by executing a representative number of classifications of different types:

- Rules-based algorithms: JRip, NNge, PART and Ridor
- Trees-based algorithms: ADTree, J48, LAD Tree, and RandomForest
- Functions-based algorithms: Logistic, MultilayerPerceptron, RBFNetwork and SMO
- Bayes-based algorithms: BayesNet and NaiveBayes

Table 3 shows the accuracy obtained by the previous classification algorithms using all the attributes (A), the eleven selected attributes (B) and the nine selected attributes (C).

Table 3 Accuracy of classification algorithms

	Algorithm		**(A)**	**(B)**	**(C)**
Rules-based algorithms	JRip		0,717	0,701	0,712
	NNge	+	0,687	0,706	0,700
	PART	+	0,690	0,699	0,720
	Riod		0,687	0,679	0,682
Trees-based algorithms	ADTree	+	0,701	0,712	0,712
	J48	+	0,711	0,741	**0,740**
	LAD Tree	+	0,719	0,724	0,730
	Random Forest		**0,732**	0,718	0,716
Functions-based algorithms	Logistic	+	0,700	0,705	0,707
	MultilayerPerceptron	+	0,693	0,730	0,699
	RBFNetwork	+	0,666	0,706	0,681

	SMO	+	0,696	0,701	0,688
Bayes-based	BayesNet	+	0,680	0,682	0,682
algorithms	NaiveBayes	+	0,697	0,708	0,684

All the algorithms obtained a solid accuracy with more similar values (65%-75%). The results indicate that most algorithms improve when using only eleven and nine attributes. The highest results are obtained by J48 when using only nine attributes and Random Forest when using all the attributes. J48 classificatory has generated a model with 73,98% correctly classificated examples, accuracy of 74% (0.74) and classification above the ROC curve area (0,762>0.5). It has been generated a confusion matrix for J48 classificatory (Table 5). Two cases of nominal class attributes are marked with the letters A- Strong and B-weak. The number of exactly classified examples is set on the matrix diagonal, and other elements of the matrix indicate the number of incorrectly classified examples as some of the other classes.

Table 4 Confusion matrix of J48 classification

	Predicted class	
Observed class	A	B
A	447	111
B	125	224
%	80,11	64,18

On the other hand, in education it is very important that a classification model obtained to be user friendly (Osmanbegović & Suljić, 2012), so that decision makers at all levels (from school to the relevant ministries) can make decisions to improve the quality of education and student performance. J48 classificatory are considered easily understood models because a reasoning process can be given for each conclusion. Knowledge models under this paradigm can be directly transformed into a set of IF-THEN rules that are one of the most popular forms of knowledge representation, due to their simplicity and comprehensibility decision makers at all levels can easy understand and interpret (Figure 1.).

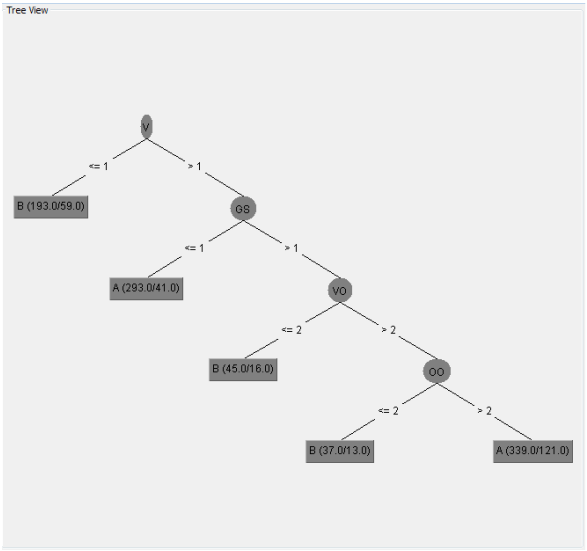

Figure 1 Obtained decision tree model

The model (see Figure 1.) is easy to be read and understood. By applying these methodologies in education, we have considerable improved making efficient, useful and, in practice, conductible decisions in order to improve learning results. The results got could represent the basis for some future research, so with bigger number of input attributes and samples we could create more successful model that would be base for building of a support decision system at the secondary education level.

Conclusion

Conducted research on data mining in education is done with the aim of emphasizing the data mining possibilities that can be of significant help during monitoring, decision-making and management in education. Results gained by the use of chosen data mining algorithms for classification of performances of students in secondary school, indicate that prediction rate varies between 65-75%. Results from the published study represent the first steps in applying data mining on the educational system. Experience gathered in this research underscores the importance of cooperation with the experts in the domain issue, making it easier to come to know the structure and importance of the data which need to be examined, and it accelerates and directs the process of KDD.

Data mining of all input attributes showed that the studying time, years of education, the student's age and father's education among the most significant predictors of grade for successfulness of a student. From the professional point of view, the aim of this analysis is to present a method for reducing the dimensional complexity in data sets, which are often present in the analysis, so it could be shown to the school management and creators of politics in education, the significance of certain attributes. An example of applying this methodology on prediction of performances of students, could be used in

certain domains of education, i.e. selecting the most important research characteristics, recognition of certain characteristics in other groups, which can facilitate interpretation to some extent, the degree of understanding and recognition of the most important attributes with the aim of improvement in future research.

References

1. Bakare, C. (1975). Some Psychological Correlates of Academic Success and Failure. African Journal of Educational Research.
2. Baker, R., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. JEDM-Journal of Educational Data Mining, I(1), 3-17.
3. Branković, N., & Arapović, A. (2010). OBAVEZNO SREDNJE OBRAZOVANJE U BiH: AMBIJENT I PERSPEKTIVE. Tuzla: Centri civilnih inicijativa.
4. Chapman, P., Clinton, J., Randy, K., Khabaza, T., Reinartz, T., Shearer, C., et al. (2000). CRISP-DM 1.0 Step-by-step data mining guide. SPSS Inc.
5. Cortez, P., & Silva, A. (2008). Using Data Mining to Predict Secondary School Student Performance. 5th Annual Future Business Technology Conference, (pp. 5-12). Porto.
6. Group, T. W. (2013). Reducing early school leaving: Key messages and policy support. Final Report, European Commission.
7. Hijazi, S., & Naqvi, R. (2006, January). Factors Affecting Students' Performance A Case Of Private Colleges. Bangladesh e-journal of sociology, III(1), 90-100.
8. Márquez-Vera, C., Cano, A., Romero, C., & Ventura, S. (2013). Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data. Applied Intelligence, 38(3), 315-330.
9. Moriana, J., Also, F., Alcala, R., Pino, M., Herruzo, J., & Ruiz, R. (2006). Extra Curricular Activities and Academic Performance in Secondary Students. Electronic Journal of Research in Educational Psychology, 4(1), 35-46.
10. Osmanbegović, E., & Suljić, M. (2012, May). Data mining approach for predicting student performance. Economic Review, X(1), 3-12.
11. Osmanbegović, E., Agić, H., & Suljić, M. (2014, March). Prediction of Students' Success by Applying Data Mining Algorithams. Journal of Theoretical and Applied Information Technology, 61(2), 378 - 388.
12. Ramaswami, M., & Bhaskaran, R. (2010, January). A CHAID based performance prediction model in educational data mining. IJCSI International Journal of Computer Science Issues, VII(1), 10-18.
13. Romero, C., & Sebastián, V. (2010). Educational data mining: a review of the state of the art. Systems, Man, and Cybernetics, Part C: Applications and Reviews (pp. 601-618). IEEE Transactions on 40.

14. Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. Expert Systems with Applications, 33(1), pp. 135-146.
15. Witten, I. H., Frank, E., & Hall, M. A. (2011). Data Mining: Practical Machine Learning Tools and Techniques (3rd ed.). Amsterdam: Morgan Kaufmann.
16. Yimin, M., Bing, L., Wong, C., Yu, P., & Lee, S. (2000). Targeting the Right Students Using Data Mining. KDD '00 The Second Annual International Conference on Knowledge Discovery in Data, (pp. 457–464). Boston.
17. Zebun, N. K. (2005). Scholastic Achievement of Higher Secondary Students in Science Stream. Journal of Social Sciences, I(2), 84-87.
18. Zoubek, L., & Burda, M. (2009). Visualization of differences in data measuring mathematical skills. International Conference on Education Data Mining, (pp. 315-324.). Cordoba.