ISSN 2278-733X

Student Performance Prediction by Using Data Mining Classification Algorithms

Dorina Kabakchieva

University of National and World Economy
Sofia, Bulgaria

Abstract— **This paper presents the results from data mining research, performed at one of the famous and prestigious Bulgarian universities, with the main goal to reveal the high potential of data mining applications for university management and to contribute to more efficient university enrolment campaigns and to attracting the most desirable students. The research is focused on the development of data mining models for predicting student performance, based on their personal, pre-university and university-performance characteristics. The dataset used for the research purposes includes data about students admitted to the university in three consecutive years. Several well known data mining classification algorithms, including a rule learner, a decision tree classifier, a neural network and a Nearest Neighbour classifier, are applied on the dataset. The performance of these algorithms is analyzed and compared.**

Keywords— **Educational Data Mining, Student Profiling, Predicting Student Performance, Classification**

I. INTRODUCTION

Universities today, similar to business organizations, are operating in a very dynamic and strongly competitive environment. The education globalization leads to more and better opportunities for students to receive high quality education at institutions all over the world. Universities are confronted with a severe competition among each other, trying to attract the most appropriate students who will successfully pass through the university educational process, and making efforts to cope with student retention. University management is very often forced to take quickly important decisions, and therefore timely and high quality information is needed.

Modern universities are collecting large volumes of data referring to their students, the organization and management of the educational process, and other managerial issues. However, the available data is usually used for producing simple queries and traditional reports that are rarely reaching the right people at the right time for making informed decisions. Moreover, much of the data remains unused due to the inability of the university administration to handle it because of the large volumes and the increasing complexity. Advanced information technologies have to be introduced to effectively transform available data into information and knowledge to support decision making.

Data mining, generally defined as the process of discovering meaningful patterns in large quantities of data, offers a great variety of techniques, methods and tools for thorough analysis of available data in various fields. The implementation of data mining in the educational sector, recently defined as "educational data mining" (EDM) [1], is a new stream in the data mining research field. The educational data mining research community is constantly growing, starting by organizing workshops since 2004, then conducting an annual International Conference on EDM beginning since 2008, and now already having a Journal on EDM (the first issue being published in October 2009).

There are already a large number of research papers discussing various problems within the higher education sector and providing examples for successful solutions reached by using data mining. Extensive literature reviews of the EDM research field are provided by Romero and Ventura in 2007 [1], covering the research efforts in the area between 1995 and 2005, and by Baker and Yacef in 2009 [2], for the period after 2005. The problems that are most often attracting the attention of researchers and becoming the reasons for initiating data mining projects at higher education institutions are focused mainly on retention of students (by better knowing their peculiarities and needs, and by providing proper support in advance), more effective targeted marketing, improving institutional efficiency, and alumni management.

The performed research work, presented in this paper, focuses on the development of data mining models for predicting student performance by using four data mining algorithms for classification – a Rule Learner, a Decision tree algorithm, a Neural network, and a K-Nearest Neighbour method. It is a continuation of previous research, carried out with the same dataset and with similar data mining algorithms, but for a different format of the predicted target variable. The achieved results from the performed research, using a target variable with five distinct values – Bad, Average, Good, Very Good, and Excellent, are previously published in [3]. Current research is implemented for a binary target variable.

The rest of the paper is organized in four sections. The research motivation and the state-of-the-art are presented in Section 2. The adopted methodological approach, the experimentation data selection and pre-processing are described in Section 3. The obtained results from the application of the selected data mining algorithms are

presented in Section 4. The paper concludes with a summary of the achievements and discussion of further research.

II. Research Motivation and State of the Art

The rationale behind the research work described in this paper is based on the great potential that is seen in using data mining methods and techniques for effective usage of university data. The discussions with high level managers and administrators of a famous and prestigious Bulgarian university have lead to the identification of existing needs for better knowing the students and performing more effective university marketing policy.

The literature review reveals that these problems have been of interest for various researchers during the last few years. The development of data mining models for predicting student performance at various levels, and comparison of those models, are discussed in a number of research papers. In 2000 the results of a study are described [4] aimed at finding weak students and involving them in additional courses for advanced support by extracting association rules from data. The retention of students is a problem discussed also by Luan, who implemented clustering, neural network and decisions tree methods to predict the students in risk of failure [5], [6]. Data mining methods are implemented for modeling online student grades [7], using three classification approaches used (binary: pass/fail; 3-level: low, middle, high; and 9-level: from 1 - lowest grade to 9 - highest score). Kotsiantis et al. [8] also deal with predicting student performance, recognizing dropout-prone students based on demographic characteristics (e.g. sex, age, marital status) and performance attributes (e.g. mark in a given assignment). Pardos et al. [9] use data from an online tutoring system for teaching Math and implement a regression approach for predicting the math test score based on individual skills. Superby et al. [10] predict students at risk of drop-out, determining factors influencing the achievement of the first-year university students, classifying students into three classes – low-risk, medium-risk and high-risk, using Decision trees, Random forest method, Neural networks and Linear discriminant analysis. Vandamme et al. [11] also deal with early identification of three categories of students: low, medium and high-risk students using Decision trees, Neural networks and Linear discriminant analysis. Cortez and Silva in [12] attempt to predict student failure by applying and comparing four data mining algorithms, Decision Tree, Random Forest, Neural Network and Support Vector Machine. The implementation of predictive modelling for maximizing student recruitment and retention is presented in the study of Noel-Levitz [13]. The development of enrolment prediction models based on student admissions data by applying different data mining methods (Decision trees, Rule induction, Feature subset selection) is the research focus of Nandeshwar [14]. Dekker et al. [15] focus on predicting students drop out. Kovačić in [16] uses data mining techniques (feature selection and classification trees) to explore the socio-demographic variables (age, gender, ethnicity, education, work status, and disability) and study environment (course programme and course block) that may influence persistence or dropout of students, identifying the most important factors for student success and developing a profile of the typical successful and unsuccessful students. Ramaswami et al. in [17] focus on developing predictive data mining model to identify the slow learners and study the influence of the dominant factors on their academic performance, using the popular CHAID decision tree algorithm.

III. Research Approach, Data Selection and Pre-processing

The performed research is based on the CRISP-DM (Cross-Industry Standard Process for Data Mining) model, a non-propriety, freely available, and application-neutral standard for data mining project implementation, widely used by researchers in the data mining field during the last ten years [18]. It is a cyclic approach, including six main phases – Business understanding, Data understanding, Data preparation, Modelling, Evaluation and Deployment, with a number of internal feedback loops between the phases, resulting from the very complex non-linear nature of the data mining process and ensuring the achievement of consistent and reliable results. The open source software WEKA, offering a wide range of classification methods for data mining [19], is used as a data mining tool for the research implementation.

First of all, the business problem is identified – it is the growing need of university management for better knowing the university students and predicting their performance in order to approach in the marketing campaigns exactly those students that will be most successful in the university education process. The stated business problem is transformed into a data mining task – the task for classifying students into two categories – successful and unsuccessful, by analysing the available student data with selected data mining methods for classification.

The next phase in the research implementation includes the data selection and pre-processing, crucial activities within each data mining project, highly influencing the quality of the final results. After studying the application process for student enrollment at the University and reviewing the procedures for collecting and storing data about the academic performance of the university students, it is established that the university data is generally organized in two databases. All the data related to the university admission campaigns is stored in the University Admission database, including personal data of university applicants (names, addresses, secondary education scores, selected admission exams, etc.), data about the organization and performance of the admission exams, scores achieved by the applicants at the admission exams, data related to the final classification of applicants and student admission, etc. All the data concerning student performance at the university is stored in the University Students Performance database, including student personal and administrative data, the grades achieved at the exams on the different subjects, etc. For the purposes of the study, student data from both databases is carefully selected, extracted and combined in a new flat file (in this case Excel file) used for the data mining analysis in the WEKA software tool.

The provided flat file contains data about 10330 students that have been enrolled as university students during the period between 2007 and 2009, described by 20 parameters, including gender, birth year, birth place, living place and country, type of previous education, profile and place of previous education, total score from previous education, university admittance exam and achieved score, total university score at the end of the first year, etc. The data is carefully studied and subjected to many transformations. Some of the parameters are removed, e.g. the "Birth place" and the "Place of living" fields containing data that is of no interest to the research, the "Country" field containing only one value (Bulgaria) because the data concerns only Bulgarian students, the "Type of previous education" field which has only one value as well because concerns only students who have finished secondary education. Some of the variables, containing important data for the research, are text fields where free text is being entered at the data collection stage. Therefore, these variables are processed and turned into nominal variables with a limited number of distinct values. Such a parameter is the "Profile of the secondary education" which is turned into a nominal variable with 9 distinct values (e.g. language, math, natural sciences, economics, technical, sports, arts, etc.). The "Place of secondary education" field is also preprocessed and transformed into a nominal variable with 7 distinct values, corresponding to the capital city and the 6 geographic regions in Bulgaria – North-East, North-Central, North-West, South-East, South-Central, and South-West. A new numeric variable is added – the "Student age at enrollment", calculated by subtracting the values contained in the "Admission year" and "Birth year" fields. Another important operation during the preprocessing phase is also the transformation of some variables from numeric to nominal (e.g. age, admission year, current semester, total university score, etc.) because they are much more informative when interpreted with their nominal values. The data is also being studied for missing values, which are very few and could not affect the results, and obvious mistakes, which are corrected.

Essentially, the challenge in the presented data mining research is to predict the student university performance based on the available student pre-university and university performance data. This is achieved by solving a classification data mining task. A binary categorical target variable is constructed, based on the original numeric parameter "University average score" (the average numeric score achieved by the students at the end of the first year at the University). The predicted variable has two distinct values, corresponding to the two classes in which the students are classified – Weak and Strong. Since a six-level scale is used in the Bulgarian educational system for evaluation of student performance at schools and universities, the students with average university score that is lower than 4.50 are classified as "Weak", and the students with average university score equal or higher than 4.50 are classified as "Strong".

The final dataset, on which the selected classification data mining algorithms are applied, contains 10067 instances and 14 attributes (summarized in Table 1).

TABLE I
FINAL DATASET USED FOR THE DATA MINING ANALYSIS

Type of Data	Attribute Name	Attribute Type	Values
Personal Data	Gender	Nom	Male (49%), Female (51%)
	Age	Nom	29 distinct values (17-43,46,48,53) (18-21–95%, 19–78%)
	BirthYear	Nom	29 distinct values
Pre-University Data	PlacePrevEdu	Nom	7 distinct values (Sofia, NE, NC, NW, SE, SC, SW)
	ProfilePrevEdu	Nom	9 distinct values (Language, Natural_Math, Humanitarian, Economics, Technological, Business_Management, Arts, Sports, General)
	ScorePrevEdu	Num	3.40-6.00
	AdmissionYear	Nom	2007, 2008, 2009
	AdmissionExam	Nom	5 distinct values (BG,Math,Geography, History,Economics)
	AdmissionExamScore	Num	0.00-6.00
	AdmissionScore	Num	0.00-35.98
University Data	UnivSpecialtyName	Nom	10 distinct values
	CurrentSemester	Nom	1-10
	NumFailures	Nom	0-12 (0 – 86,4%)
	StudentClass	**Nom**	**Weak (5340), Strong (4727)**

Most of the attributes (11), including the predicted class variable, are nominal variables accepting a certain number of distinct values, and only 3 of the attributes are numeric variables.

IV. ACHIEVED RESULTS FROM THE DATA MINING ALGORITHMS IMPLEMENTATION

During the Modeling Phase, the algorithms for building models that would classify the students into the two classes – Weak and Strong, depending on their university performance and based on the student pre-university data, are considered and selected. Popular WEKA classifiers (with their default settings unless specified otherwise) are used in the experimental study, including a rule learner (OneR), a common decision tree algorithm C4.5 (J48), a neural network (MultiLayer Perceptron), and a Nearest Neighbour algorithm (IBk). These classification algorithms are selected because they are very often used for research purposes and have potential to yield good results. Moreover, they use different approaches for generating the classification models, which increases the chances for finding a prediction model with high classification accuracy. The OneR Rule Learner algorithm produces a one-level decision tree expressed in the form of a set of rules that all test one particular attribute – the minimum-classification-error attribute. It is a simple, cheap method that often produces good rules with high accuracy. The Decision Tree algorithms generate models in the form of a tree-like structure, which starts from root attributes and ends with leaf nodes, describing the relationship among attributes and the relative importance of attributes. They represent rules which could easily be understood and interpreted by users, do not require complex data preparation, and perform well for numerical and categorical variables. Neural networks produce classification models in the form of a mathematical model, consisting of interconnected computational elements (neurons) and processing information using a connectionist approach to computation. They are used to model complex relationships between inputs and outputs and very often yield very good results. The K-Nearest Neighbor algorithm (k-NN) is a method for classifying instances based on measuring the distance between the classified instance and the closest training examples in the feature space. It is easily understood

by users, often provides good classification results and performs well for large datasets.

The selected data mining algorithms are applied to the dataset using the holdout method (WEKA "Percentage Split" test option, 66%/34%), as shown on Fig.1. The dataset is divided into 3 parts and, each time an algorithm is run, 2/3 of the data is used for training of the classification model and 1/3 of the data is used for testing and evaluation of the model.

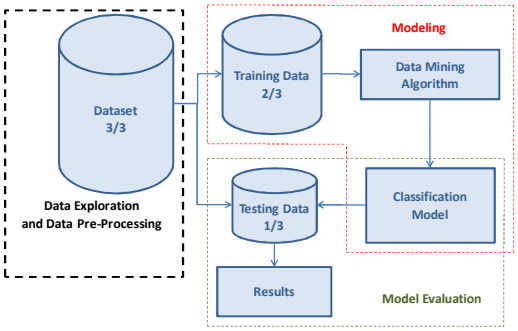

Fig. 1 Data Mining Algorithm Implementation - Classification Model

The results from the evaluation of the classification models generated with the selected four data mining algorithms are presented in Table 2.

TABLE III
ACHIEVED RESULTS FROM THE DATA MINING ALGORITHM IMPLEMENTATION

Data Mining Algorithms	Rule Learner (OneR)			Decision Tree (J48)			Neural Network (MultiLayer Perceptron - 1 hidden layer with 7 neurons)			K-Nearest Neighbour (IBk, k=50)		
Evaluation Parameters	Weak	Strong	Weigh. Av.	Weak	Strong	Weigh. Av.	Weak	Strong	Weigh. Av.	Weak	Strong	Weigh. Av.
Corr. Classified Instances	67.4554%			72.7432%			73.5904%			70.47		
Incorr. Classified Instances	32.5446%			27.2568%			26.4096%			29.53		
Kappa Statistic	0.3439			0.4524			0.4730			0.4085		
TP Rate	0.73	0.61	0.68	0.75	0.70	0.73	0.70	0.77	0.74	0.71	0.70	0.71
FP Rate	0.39	0.27	0.33	0.30	0.25	0.28	0.23	0.30	0.26	0.30	0.29	0.30
Precision	0.68	0.67	0.67	0.74	0.72	0.73	0.78	0.70	0.74	0.73	0.68	0.71
Recall	0.73	0.61	0.68	0.75	0.70	0.73	0.70	0.77	0.74	0.71	0.70	0.71
F-Measure	0.70	0.64	0.67	0.75	0.71	0.73	0.74	0.73	0.74	0.72	0.69	0.71
ROC Area	0.67	0.67	0.67	0.78	0.78	0.78	0.82	0.82	0.82	0.78	0.78	0.78

The four classification models, generated with the selected data mining algorithms, are compared by using the following evaluation measures: % of correctly/incorrectly classified instances, Kappa Statistic, True Positive (TP) and False Positive (FP) Rates, Precision, Recall, F-Measure and ROC Area. These are well known measures for evaluation of data mining models for classification.

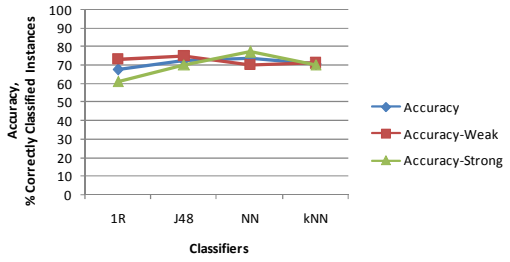

Fig. 2A Classifiers Accuracy Comparison

The results for the classification model comparison are presented on Fig.2. The highest classification accuracy (% of correctly classified instances, Fig.2a) is achieved for the Neural Network algorithm – 73.59%. The Neural Network model is also the only model that predicts the "Strong" class with higher accuracy (TP Rate=77%) than the "Weak" class (TP Rate=70%), which means that this model could most successfully be used to predict the strong students based on their pre-university and university-performance characteristics. The disadvantages of that model are its complexity and the difficult understanding and interpretation by users. The models, generated with the other three algorithms, are predicting the "Weak" class with higher accuracy than the "Strong" class, and they could be used for early identification of students in risk that might need additional support. The Decision Tree classification model also reveals high accuracy of prediction - 72.74%. The advantages of this model are that it is easily interpretable because it produces a set of understandable rules, and that it is working well with both, nominal and numeric variables. The attributes, which appear at the upper part of the decision tree and therefore are considered most informative for the instance distribution into the two classes, are the "Number of Failures" and the "Admission Score". The K-NN model provides 70.5% accuracy of classification, working with similar accuracies for both classes – Weak and Strong. The OneR classifier is the least accurate, performing better for the Weak class, as the Decision Tree classifier. The OneR algorithm uses the "Admission Score" attribute for the classification which once again proves that the Admission Score parameter is very informative for recognizing strong and weak students.

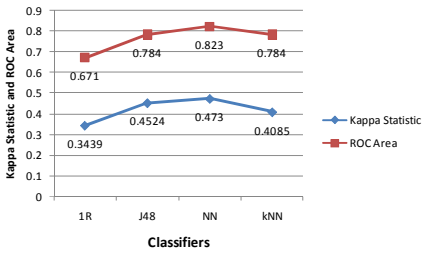

Fig. 3B Classifiers Accuracy Comparison

The results for the Kappa Statistic (Fig.2b), an index that compares correct classifications against chance classifications and taking values in the range from -1 for complete disagreement, to 1 for perfect agreement, also reveal that the Neural Network model outperforms the other three classification models with the maximum achieved value of 0.473. The ROC curve plots the true positives against the false positives and the area under the curve represents the accuracy of the model – the larger the area, the more accurate the model. The achieved results for three of the generated classification models (Fig.2b), the Neural Network, the Decision Tree and the k-NN, reveal values of the ROC Area above 0.7, highest for the Neural Network Model - 0.82, which means that all the models are reliable for prediction (if the ROC area is below 0.5, random guesses outperform the model).

V. CONCLUSIONS

The classification models, generated by applying the selected four data mining algorithms – OneR Rule Learner, Decision Tree, Neural Network and K-Nearest Neighbour, on the available and carefully pre-processed student data, reveal classification accuracy between 67.46% and 73.59%. The highest accuracy is achieved for the Neural Network model (73.59%), followed by the Decision Tree model (72.74%) and the k-NN model (70.49%). The Neural Network model predicts with higher accuracy the "Strong" class, while the other three models perform better for the "Weak" class. The data attributes related to the students' University Admission Score and Number of Failures at the first-year university exams are among the factors influencing most the classification process.

The presented results will be compared to previous results, achieved for the same dataset but for a different format of the predicted target variable (a nominal variable with five distinct values – Bad, Average, Good, Very Good, and Excellent). The results and conclusions will also help to define the further steps and directions for continuing the university data mining research, including possible transformations of the dataset, adding new data, tuning the classification algorithms' parameters, etc., in order to achieve better prediction. Recommendations will also be provided to the university management, concerning the sufficiency and availability of university data, and related to the improvement of the data collection process.

REFERENCES

[1] Romero, C., Ventura, S. (2007). *Educational Data Mining: A Survey from 1995 to 2005*. Expert Systems with Applications 33, 2007, pp.135-146.

[2] Baker, R., Yacef, K. (2009). *The State of Educational Data mining in 2009: A Review and Future Visions*. Journal of Educational Data Mining, Vol.1, Issue 1, Oct. 2009, pp.3-17.

[3] Kabakchieva, D., Stefanova, K., Kisimov, V. (2011). *Analyzing University Data for Determining Student Profiles and Predicting Performance*. Conference Proceedings of the 4th International Conference on Educational Data Mining (EDM 2011), 6-8 July 2011, Eindhoven, The Netherlands, pp.347-348.

[4] Ma, Y., Liu, B., Wong, C. K., Yu, P. S., Lee, S. M. (2000). *Targeting the right students using data mining*. Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, Boston, pp 457-464.

[5] Luan, J. (2002). *Data Mining and Its Applications in Higher Education*. New Directions for Institutional Research, Special Issue titled Knowledge Management: Building a Competitive Advantage in Higher Education, Vol. 2002, Iss.113, pp.17–36.

[6] Luan, J. (2004). *Data Mining Applications in Higher Education*. SPSS Executive Report, SPSS Inc.

[7] Minaeli-Bidgoli, B., Kashy, D., Kortemeyer, G., Punch, W. (2003). *Predicting Student Performance: An Application of Data Mining Methods with the Educational Web-Based System LON-CAPA*. 33rd ASEE/IEEE Frontiers in Education Conference, 5-8 Nov 2003, Boulder, CO.

[8] Kotsiantis, S., Pierrakeas, C., Pintelas, P. (2004). *Prediction of Student's Performance in Distance Learning Using Machine Learning Techniques*. Applied Artificial Intelligence, Vol. 18, No. 5, 2004, pp. 411-426.

[9] Pardos Z., Heffernan N., Anderson B., and Heffernan C. (2006). *Using Fine-Grained Skill Models to Fit Student Performance with Bayesian Networks*. In Proceedings of the Workshop in Educational Data Mining held at the 8th International Conference on Intelligent Tutoring Systems (ITS2006), June 26, 2006, Taiwan.

[10] Superby, J. Vandamme, J., Meskens, N. (2006). *Determination of factors influencing the achievement of the first-year university students using data mining methods*. Proceedings of the Workshop on Educational Data Mining at the 8th International Conference on Intelligent Tutoring Systems (ITS 2006). Jhongli, Taiwan, pp37-44.

[11] Vandamme, J., Meskens, N., Superby, J. (2007). *Predicting Academic Performance by Data Mining Methods*. Education Economics, 15(4), pp405-419.

[12] Cortez, P., Silva, A. (2008). *Using Data Mining to Predict Secondary School Student Performance*. EUROSIS, A. Brito and J. Teixeira (Eds.), 2008, pp.5-12.

[13] Noel-Levitz White Paper (2008). *Qualifying Enrollment Success: Maximizing Student Recruitment and Retention Through Predictive Modeling*. Noel-Levitz, Inc., 2008.

[14] Nandeshwar, A., Chaudhari, S. (2009). *Enrollment Prediction Models Using Data Mining*. Available at: http://nandeshwar.info/wp-content/uploads/2008/11/DMWVU_Project.pdf

[15] Dekker, G., Pechenizkiy, M., Vleeshouwers, J. (2009). *Predicting Students Drop Out: A Case Study*. Conference Proceedings of the 2nd International Conference on Educational Data Mining (EDM'09), 1-3 July 2009, Cordoba, Spain, pp.41-50.

[16] Kovačić, Z. (2010). *Early Prediction of Student Success: Mining Students Enrolment Data*. Proceedings of Informing Science & IT Education Conference (InSITE) 2010, pp.647-665.

[17] Ramaswami, M., Bhaskaran, R. (2010). *A CHAID Based Performance Prediction Model in Educational Data Mining*. IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 1, No.1, January 2010, pp.10-18.

[18] Chapman, P., et al. (2000). *CRISP-DM 1.0: Step-by-step data mining guide*. 2000 SPSS Inc. CRISPWP-0800. Available at: http://www.crisp-dm.org/CRISPWP-0800.pdf

[19] Witten, I., Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers, Elsevier Inc. 2005.