Analyzing University Data for Determining Student Profiles and Predicting Performance

D. KABAKCHIEVA
Sofia University "St. Kl. Ohridski", Bulgaria
K. STEFANOVA, V. KISIMOV
University of National and World Economy, Sofia, Bulgaria

Keywords: Educational Data Mining, Student Profiling, Predicting Student Performance, Classification

1. INTRODUCTION AND RELATED RESEARCH

The open access to education within the enlarged Europe, and even outside it, is introducing hard competition among universities and requires them to introduce advanced methods for data analysis, in order to identify their own uniqueness and to select the most appropriate students to reach better performance.

The implementation of data mining is considered a powerful instrument for acquiring new knowledge from existing data to support decision making. Literature reviews of the Educational Data Mining (EDM) research field are provided by Romero and Ventura (Romero et al. 2007), covering the research efforts between 1995 and 2005, and Baker and Yacef (Baker et al. 2009), for the period after 2005. Recent research papers in the EDM field are focused on understanding student types and targeted marketing (Ma et al 2000, Luan 2002, Luan 2004, Antons 2006), using predictive modeling for maximizing student retention (Noel-Levitz 2008, DeLong 2007, Yu et al 2010), developing enrollment prediction models based on admission data (Nandeshwar 2009), predicting student performance and drop-out (Kotsiantis et al. 2004, Vandamme et al. 2007, Cortez and Silva 2008, Dekker et al. 2009, Kovačić 2010, Ramaswami et al. 2010).

This paper presents a data mining research project that is started at a Bulgarian university, with the main goal to reveal the high potential of data mining applications for university management. The specific objective of the research work is to find out interesting patterns in the available data that could contribute to predicting student performance at the university based on their personal and pre-university characteristics. This is a first attempt of applying data mining in the Bulgarian educational sector.

2. THE RESEARCH METHODOLOGY

The data mining project is based on the CRISP-DM (Cross-Industry Standard Process for Data Mining) research approach. The open source software tool WEKA is used for the project implementation. During the ***Business Understanding Phase*** the specific University management needs are identified. In the ***Data Understanding Phase*** the types of data collected from the university applicants during the enrollment campaigns, and stored in electronic format, are studied. During the ***Data Preprocessing Phase***, student data from the two University databases is extracted and organized in a new data mart.

The research sample includes data about 10330 students, described by 20 parameters (gender, birth year and place, living place and country; type, profile, place and total score from previous education, university admittance year, exam and achieved score, current semester, total university score. The provided data is subjected to many transformations – removing parameters that are considered useless (e.g. fields with one value only), replacing fields containing free text with nominal variable (with a number of distinct values), transforming numeric to nominal variables, etc. The data is also being studied for missing values (very few and not important), and obvious mistakes (corrected).

The data mining task is to predict the student university performance based on the student personal and pre-university characteristics. The target variable is the "student class". It is constructed as a categorical variable, based on the numeric values of the "student total university score" attribute, and has five distinct values - "excellent" (5.50-6.00), "very good" (4.50-5.49), "good" (3.50-4.49), "average" (3.00-3.49) and "bad" (below 3.00). The dataset contains 10330 instances (539 classified as excellent, 4336 as very good, 4543 as good, 347 as average, and 564 as bad), each described with 14 attributes (1 output and 13 input variables), nominal and numeric.

During the ***Modeling Phase***, several different classification algorithms are selected and applied. Popular WEKA classifiers (with their default settings unless specified otherwise) are used, including a common decision tree algorithm C4.5 (J48), two Bayesian classifiers (NaiveBayes and BayesNet), a Nearest Neighbour (kNN) algorithm (IBk) and two rule learners (OneR and JRip).

3. THE ACHIEVED RESULTS

The WEKA Explorer application is used at this stage. Each classifier is applied for two testing options - cross validation (using 10 folds) and percentage split (2/3 of the dataset used for training and 1/3 – for testing). The results for the overall accuracy of the applied classifiers, including True Positive Rate and Precision (the average values for the 10-fold cross validation and split options), are presented in Table I. The results for the classifiers' performance on the five classes are presented on Fig.1.

Table I. Results for the accuracy of the applied classifiers

	J48	**NaiveBayes**	**BayesNet**	**k-NN, k=100**	**k-NN, k=250**	**OneR**	**JRip**
TP Rate	0,663	0,586	0,591	0,613	0,593	0,546	0,632
Precision	0,640	0,594	0,597	0,574	0,563	0,480	0,611

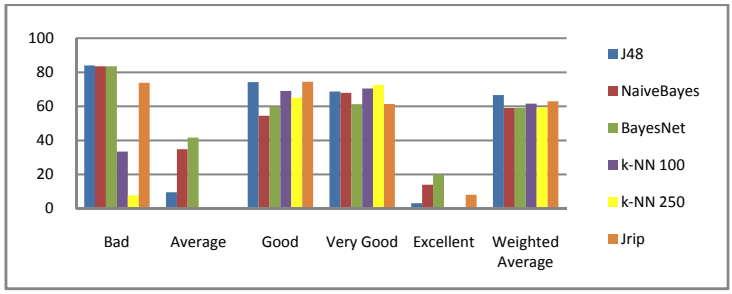

Fig.1. Classification Algorithms Performance Comparison

The achieved results reveal that the decision tree classifier (J48) performs best, followed by the rule learner (JRip) and the k-NN classifier. The Bayes classifiers are less accurate than the others. However, all tested classifiers are performing with an overall accuracy below 70% which means that the error rate is high and the predictions are not very reliable. The predictions are worst for the Excellent class, and bad for the Average class, the k-NN classifier being absolutely unable to predict them. The highest accuracy is achieved for the Bad class (except for the k-NN). The predictions for the Good and Very Good classes are more precise than for the other classes, and all classifiers perform with accuracies around 60-75%.

The selected classifiers perform with similar overall accuracies on the dataset, but they differ with respect to the five classes. All algorithms do not successfully predict the Average and Excellent classes which might be explained with the uneven construction of the target variable. Further research efforts will be directed at achieving higher accuracy of the classifiers' prediction by additional transformations of the dataset, reconstruction of the target variable, tuning of the classification algorithms' parameters, etc.