A REVIEW ON PREDICTING STUDENT PERFORMANCE USING DATA MINING METHOD

Karunendra Verma[1], Arjun Singh[2], Purushottam Verma[3]
[1]PhD-Scholor, CSE Dept., SPSU Udaipur(Raj.)
[2]Asst. Professor, CSE Dept., SPSU, Udaipur(Raj.)
[3]Asst. Professor, CSE Dept., IPS-CTM, Gwlalior (M.P.)
Email: k.verma2006@gmail.com[1], arjun.singh@spsu.ac.in[2], puruverma22@gmail.com[3]

Abstract
Data mining methods are often implemented for analyzing available data and extracting Information and knowledge to support decision-making. This review paper is aimed at revealing the high potential of data mining applications for university management benefits.

Introduction:

Now a days Universities are operating in a very complex and highly competitive environment. The main challenge for modern universities is to deeply analyze their performance, to identify their uniqueness and to build a strategy for further development and future actions.

University management should focus more on the profile of admitted students, getting aware of the different types and specific students' characteristics based on the received data. They should also consider if they have all the data needed to analyze the students at the entry point of the university or they need other data to help the managers support their decisions as how to organize the marketing campaign and approach the promising potential students.

This paper is focused on the implementation of data mining techniques and methods for acquiring new knowledge from data collected by universities. The main goal of the paper is to reveal the high potential of data mining applications for university management.

The specific objective of the proposed research work is to find out if there are any patterns in the available data that could be useful for predicting students' performance at the university based on their personal and pre-university characteristics.

The university management would like to know which features in the currently available data are the strongest predictors of university performance. They would also be interested in the data − is the collected data sufficient for making reliable predictions, is it necessary to make any changes in the data collection process and how to improve it, what other data to collect in order to increase the usability of the analysis results.

Review of literature:

The implementation of data mining methods and tools for analyzing data available at educational institutions, defined as Educational Data Mining (EDM) [15] is a relatively new stream in the data mining research. Extensive literature reviews of the EDM research field are provided by Romero and Ventura [15], covering the research efforts in the area between 1995 and 2005, and by Baker and Yacef [2], for the period after 2005.

The data mining project that is currently implemented at UNWE is focused on finding information in the existing data to support the university management in better knowing their 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. Luan discusses in [9] the potential applications of data mining in higher education and explains how data mining saves resources while maximizing efficiency in academics.
Understanding student types and targeted marketing based on data mining models are the research topics of several papers [1, 9, 10, 11]. The implementation of predictive modeling for maximizing student recruitment and retention is presented in the study of Noel-Levitz [13]. These problems are also discussed by DeLongetal [5]. The development of enrollment prediction models based on student admissions data by applying different data mining methods is the research focus of Nandeshwar and Chaudhari [12]. Dekkeretal. [6] focus on predicting students drop out.

Kovacicin [8] 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 program and course block) that may influence persistence or dropout of students.
Ramaswami and Bhaskaran [14] 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.
Yuetal[18] explore student retention by using classification trees, Multivariate Adaptive Regression Splines (MARS), and neural networks. Cortez and Silva [4] attempt to predict student failure by applying and comparing four data mining algorithms − Decision Tree, Random Forest, Neural Network and Support Vector Machine.
Kotsiantiset al. [7] apply five classification algorithms (Decision Tree, Perceptron-based Learning, Bayesian Net, Instance Based Learning and Rule-learning) to predict the performance of computer science students from distance learning.

Research Methodology:

In Data mining numbers of steps like Data Understanding, Data Preparation, Modeling, Evaluation and Deployment for finding consistent and reliable results.

The software tool that is used for the project implementation is the open source software WEKA, offering a wide range of classification methods for data mining [17].
During the "Data Preprocessing" Phase, student data from the two databases is extracted and organized in a new flat file. The preliminary research sample is provided by the university technical staff responsible for the data collection and maintenance, and includes data about 10330 students, 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 year, admittance exam and achieved score, university specialty/direction, current semester, total university score, etc. The provided data is 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 selected target variable in this case, or the concept to be learned by data mining algorithm, is the "student class". A categorical target variable is constructed based on the original numeric parameter university average score. It has five distinct values (categories) − "excellent", "very good", "good", "average" and "bad". The five categories (classes) of the target (class) variable are determined from the total university score achieved by the students.
During the "Modeling Phase", the methods for building a model that would classify the students into the five classes (categories), depending on their university performance and based on the student pre-university data, are considered and selected. Several different classification algorithms are applied during the performed research work, selected because they have potential to yield good results. Popular WEKA classifiers (with their default settings unless specified otherwise) are used in the experimental study, including a common decision tree algorithm C4.5 (J48), two Bayesian classifiers (NaiveBayes and BayesNet), a Nearest Neighbour algorithm (IBk) and two rule learners (OneR and JRip).

5. Achieved Results & Comparison:

Several different algorithms are applied for building the classification model, each of them

using different classification techniques. The stage WEKA Explorer application is used at this

Class	NaiveBayes				BayesNet			
	10-fold Cross validation		Percentage split		10-fold Cross validation		Percentage split	
	TP Rate	Precision	TP Rate	Precision	TP Rate	Precision	TP Rate	Precision
Bad	0.821	0.791	0.835	0.804	0.817	0.813	0.835	0.819
Average	0.352	0.209	0.348	0.183	0.38	0.237	0.417	0.222
Good	0.521	0.644	0.545	0.649	0.598	0.626	0.597	0.633
V.Good	0.681	0.576	0.679	0.588	0.616	0.599	0.613	0.601
Excellent	0.184	0.277	0.14	0.268	0.237	0.312	0.199	0.264
Weighted Average	0.581	0.59	0.59	0.597	0.591	0.596	0.591	0.598

Class	J48 – 10-fold Cross validation		J48 – Percentage split	
	True Positive Rate	Precision	True Positive Rate	Precision
Bad	**0.83**	**0.851**	**0.84**	**0.892**
Average	**0.081**	**0.384**	**0.096**	**0.344**
Good	**0.729**	**0.665**	**0.742**	**0.667**
Very Good	**0.69**	**0.639**	**0.687**	**0.646**
Excellent	**0.015**	**0.211**	**0.032**	**0.429**
Weighted Average	**0.659**	**0.631**	**0.666**	**0.648**

Table 1. Results for the Bayesian Classifiers

Table 2. Results for the decision tree algorithm (J48)

Class	k-NN Classifier							
	k=100				k=250			
	10-fold Cross validation		Percentage split		10-fold Cross validation		Percentage split	
	TP Rate	Precision	TP Rate	Precision	TP Rate	Precision	TP Rate	Precision
Bad	0.358	0.944	0.335	0.972	0.154	1	0.078	1
Average	0	0	0	0	0	0	0	0
Good	0.662	0.614	0.69	0.617	0.626	0.602	0.651	0.598
Very Good	0.712	0.592	0.705	0.6	0.733	0.576	0.727	0.586

Excellent	0	0	0	0	0	0	0	0
Weighted Average	0.609	0.57	0.616	0.578	0.592	0.561	0.593	0.565

Table 3. Results for the k-NN Classifier

Class	**OneR**				**JRip**			
	10-fold Cross validation		Percentage split		10-fold Cross validation		Percentage split	
	TP Rate	Precision	TP Rate	Precision	TP Rate	Precision	TP Rate	Precision
Bad	0	0	0	0	0.823	0.845	0.738	0.776
Average	0	0	0	0	0.043	0.313	0	0
Good	0.688	0.545	0.689	0.542	0.731	0.618	0.744	0.615
Very Good	0.584	0.555	0.572	0.553	0.625	0.634	0.614	0.636
Excellent	0.006	0.15	0.032	0.214	0.082	0.506	0.081	0.375
Weighted Average	0.548	0.481	0.543	0.479	0.634	0.621	0.63	0.601

Table 4. Results for the rule learners

The results for the performance of the selected classification algorithms (TP rate

percentage split test option) are summarized and presented on Fig. 1.

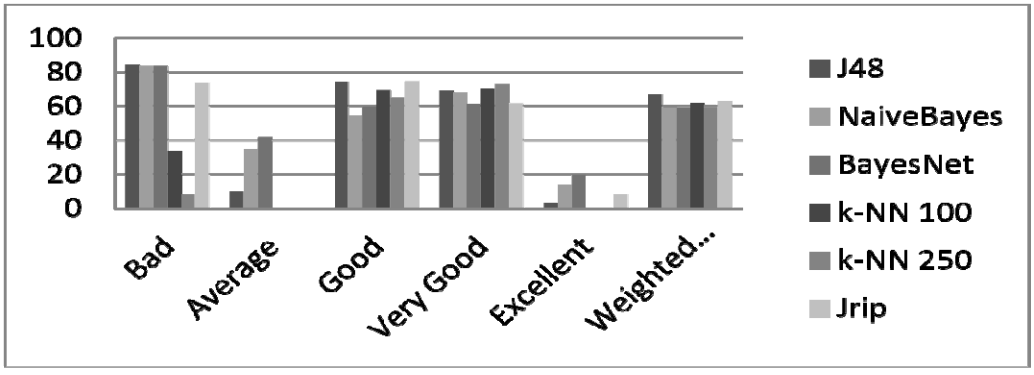

Fig. 1. Classification algorithms performance comparison

,

6. Conclusion:

The results achieved by applying selected data mining algorithms for classification on the university sample data reveal that the prediction rates are not remarkable (vary between 52-67 %). Moreover, the classifiers perform differently for the five classes. The data attributes related to the student's University

Admission Score and Number of Failures at the first-year university exams are among the factors influencing most the classification process.

7. Future Work:

Using Association rule in mining we can find different relationship among the classes and predict the student performance and comparison can be done with the earlier results. To check whether accuracy is increased and prediction rate is up to the mark.

8. References:

1. Antons, C., E. Maltz. Expanding the Role of Institutional Research at Small Private Universities: A Case Study in Enrollment Management Using Data Mining. – New Directions for Institutional Research, Vol. 131, 2006, 69-81.
2. Baker, R., K. Yacef. The State of Educational Data Mining in 2009: A Review and Future Visions. – Journal of Educational Data Mining, Vol. 1, October 2009, Issue 1, 3-17.
3. Chapma n, P., et al. CRISP-DM 1.0: Step-by-Step Data Mining Guide 2000. SPSS Inc. CRISPWP-0800, 2000. http://www.spss.ch/upload/1107356429_CrispDM1.0.pdf
4. Cortez, P., A. Silva. Using Data Mining to Predict Secondary School Student Performance.EUROSIS.A.Brito andJ.Teixeira,Eds.2008,5-12.
5. DeLong, C., P. Radclie, L. Gorny. Recruiting for Retention: Using Data Mining and Machine Learning to Leverage the Admissions Process for Improved Freshman Retention. – In: Proc. of the Nat. Symposium on Student Retention, 2007.
6. Dekker, G., M. Pechenizkiy, J. Vleeshouwer s. Predicting Students Drop Out: A Case Study. – In: Proceedings of 2nd International Conference on Educational Data Mining (EDM'09), 1-3 July 2009, Cordoba, Spain, 41-50.
7. Kotsiantis, S., C. Pierrakeas, P. Pintelas. Prediction of Student's Performance in Distance Learning Using Machine Learning Techniques. Applied Artificial Intelligence,Vol.18,2004,No5,411-426.
8. Kovaicic, Z. Early Prediction of Student Success: Mining Students Enrolment Data. – In: Proceedings of Informing Science & IT Education Conference (InSITE'2010), 2010, 647-665. Luan, J. 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, 2002, Issue 113, 17-36.
9. Luan, J. 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, 2002, Issue 113, 17-36.
10. Luan, J. Data Mining Applications in Higher Education. SPSS Executive Report. SPSS Inc., 2004. http://www.spss.ch/upload/1122641492_Data%20mining%20applications%20in%20higher%20education.pdf
11. Ma, Y., B. Liu, C. K. Wong, P. S. Yu, S. M. Lee. Targeting the Right Students Using Data Mining. – In: Proceedings of 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Boston, 2000, 457-464. 12. Nandeshwar, A., S. Chaudhari. Enrollment Prediction Models Using Data Mining, 2009.
12. http://nandeshwar.info/wpcontent/uploads/2008/11/DMWVU_Project.pdf
13. Noel-Levitz. White Paper. Qualifying Enrollment Success: Maximizing Student Recruitment and Retention Through Predictive Modeling. Noel-Levitz, Inc., 2008. https://www.noellevitz.com/documents/shared/Papers_and_Research/2008/QualifyingEnrollmentSuccess08.pdf
14. Ramaswami, M., R. Bhaskaran. A CHAID Based Performance Prediction Model in Educational Data Mining. – IJCSI International Journal of Computer Science Issues, Vol. 7, January 2010, Issue 1, No 1, 10-18.
15. Romero, C., S. Ventura. Educational Data Mining: A Survey from 1995 to 2005. – Expert Systems with Applications, Vol. 33, 2007, 135-146.
16. Vandamme, J., N. Meskens, J. Superby. Predicting Academic Performance by Data Mining Methods. – Education Economics, Vol. 15, 2007, No 4, 405-419.

17. Witten, I., E. Frank. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann Publishers, Elsevier Inc., 2005.
18. Yu, C., S. DiGangi, A. Jannasch-Pennell, C. Kaprolet. A Data Mining Approach for Identifying Predictors of Student Retention from Sophomore to Junior Year. – Journal of Data Science, Vol. 8, 2010, 307-325.
19. Kabakchieva, D., K. Stefanova, V. Kisimov. Analyzing University Data for Determining Student Profiles and Predicting Performance. – In: Proceedings of 4th International Conference on Educational Data Mining (EDM'2011), 6-8 July 2011, Eindhoven, The Netherlands, 347-348.
20. BULGARIAN ACADEMY OF SCIENCES CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume 13, No 1 Sofia • 2013 Print ISSN: 1311-9702; Online ISSN: 1314-4081DOI: 10.2478/cait-2013-0006