Application Of Data Mining Techniques For Student Success And Failure Prediction (The Case Of Debre_Markos University)

Muluken Alemu Yehuala

Abstract: This research work has investigated the potential applicability of data mining technology to predict student success and failure cases on University students' datasets. CRISP-DM (Cross Industry Standard Process for Data mining) is a data mining methodology to be used by the research. Classification and prediction data mining functionalities are used to extract hidden patterns from students' data. These patterns can be seen in relation to different variables in the students' records. The classification rule generation process is based on the decision tree and Bayes as a classification technique and the generated rules were studied and evaluated. Data collected from MS_EXCEL files, and it has been preprocessed for model building. Models were built and tested by using a sample dataset of 11,873 regular undergraduate students. Analysis is done by using WEKA 3.7 application software. The research results offer a helpful and constructive recommendations to the academic planners in universities of learning to enhance their decision making process. This will also aid in the curriculum structure and modification in order to improve students' academic performance. Students able to decide about their field of study before they are enrolled in specific field of study based on the previous experience taken from the research-findings. The research findings indicated that EHEECE (Ethiopian Higher Education Entrance Certificate Examination) result, Sex, Number of students in a class, number of courses given in a semester, and field of study are the major factors affecting the student performances. So, on the bases of the research findings the level of student success will increase and it is possible to prevent educational institutions from serious financial strains.

Index Terms: Data mining, failure, performance, predict, student, success, universities

1 INTRODUCTION

Data mining is the extraction of implicit, previously unknown, and potentially useful information from data. It is about solving problems by analyzing data already present in large quantities of data in order to discover meaningful patterns and rules [1]. Universities in their lifelong teaching process have a large amount of student data stored in a database. But the problem is not storing the data, rather Data handling, extraction of meaningful patterns, and discovery of knowledge buried in the huge database is very difficult. Deploying data mining tools is a mechanism to analyze and manage large volume of data so as to discover new patterns that are helpful for problem solving and decision making. Data mining tools can extract important things unknown to the user or what is going to happen in the future. This research aims to prove how different factors affect a student success and failure rate in relation to other variables in students' data set by applying data mining techniques. As studied by different researchers, there are many factors affecting student performance in the University like gender, enrolment status, faculty of study, campus residence (the physical environment), financial assistance, health condition, communication language, family background, and so on. This study is to investigate the potential applicability of Data Mining Techniques in developing a model that predict student performance on student data, discover patterns that identify the major determinant factors on students' attrition at Universities, so as possible to minimize failure rate and aid to identify the number of potential students in lifelong learning Programs or those that need additional motivation. It is possible to follow the failure and academic dismissal trends throughout several years in order to check the effectiveness of corrective activities. In this paper, the overview of data mining and its applications in educational areas, specifically, the potential applicability of data mining techniques in university student data set and the result patterns will be discussed. The conclusion and recommendations are also included. This paper is organized as follows. Section 2 is about data mining for student performance analysis. Section 3 explains Materials and methods used by the researcher. Section 4 discusses on different result patterns and Section 5 includes the conclusion and recommendations.

2 DATA MINING FOR STUDENT PERFORMANCE PREDICTION

Here in this section the researcher shows the potential applicability of data mining techniques on university student data sets to predict factors that affect students' performance. [7] Use three sets of factors to predict the success or failure of students. Those factors are:

- **i)** history of student (his identity, socio-family past, academic past, age, and gender),
- **ii)** Student's involvement in studies (participation in optional activities, meeting with lecturers
- **iii)** Student's perception (views on academic context, professors, course to determine their influence on the students' performances in academics. The study used students' data survey to collect data.

The questionnaire (comprised of 42 questions) was distributed to first year students and based on it a database was created in which each student is described according to attributes (explanatory variables X). Each student is also assigned with a risk-of –failure category (high, medium, low risk of failures) and so created dependent variable Y. Several data mining techniques like decision tree, Neural network, Linear Discriminant Analysis were used. 20% variables showed significant correlations with academic success and 80% rate of correct classification in predicting the success or failure of students. Students' performance were categorized in five groups: "Very good", with a high probability of succeeding; "Good" students, who are above average and with a little more effort can succeed with good grades ; "Satisfactory" students, who may succeed; "Below Satisfactory" students, who require more efforts to succeed; and "Fail" , who have a high probability of dropping out. [3] used data sets of students' past education results and university grades and related data collected from Eindhoven University of Technology to test their impact on students' performance and determine whether they

can help in predicting performance using data mining techniques. The study was aimed to differentiate successful and unsuccessful students as well as those who are at risk of dropping out as early as possible in the first degree program. Various simple and sophisticated data mining techniques were used and their results were compared. The overall result shows that simple classifier gives result with accuracy between 75 to 80%. [6] Explored the use of data mining techniques in predicting students' performance. He used the decision tree approach and Bayesian network and also compared the accuracy of two data mining techniques algorithms applied on the students of two different academic institutes. The Asian Institute of Technology data sets included students' records and GPA at the end of the second year to predict the students' rate of performance in the third year. The other set of CanTho University in Vietnam included students admission data to predict GPA at the end of the first year. The conclusion was that Decision tree algorithm performed better than the Bayesian network algorithm. [4] Used 1500 records of engineering students' data in Defence University College, Debrezeit, Ethiopia, to predict students' performance in a specific course. The students' evaluation factors like: class quizzes, mid and final exam, assignment, course type, lecturer qualification, and grade obtained are studied. They used k-Means clustering, to group students in to homogenous groups and Decision tree Data mining techniques, to evaluate student data of main attributes that may affect the performance of student in courses. Students in a class were grouped in to three and evaluated as: LOW, MEDIUM, & HIGH in relation to the grade obtained, enrolled course and the professor that provided the course. The accuracy of the decision tree classifier on test data was 100%. All related works above show that application of data mining techniques to predict student performance. The classification data mining techniques they have used are decision tree, neural network, and Bayesian network and they have used University students' data in the junior years of the teaching program. The researcher of this paper considers the performance of a student before he/she joins the university is very important. Hence the unique feature of this research is that, it includes combination of datasets from the students' secondary education, the students' related data in the university, and other demographic data. Some of the attributes that differ this research work than other related works include: sex, age, region, Higher Education Entrance Certificate Examination result, field of study, College, Number of courses given in a semester, Total credit hours given in a semester, Number of students in a class, semester Cumulative GPA of a student. In fact further research is needed to get more important attributes in relation to student performance like family back ground in education, economical background, health related conditions, environmental factors and so on.

3 RESEARCH METHODS AND TECHNIQUES

3.1. RESEARCH METHODS

The data mining process must be reliable and repeatable by business people with little knowledge or no data mining background. So following a standard data mining process guides users. The CRoss-Industry Standard Process for Data Mining (CRISP-DM) is a standard data mining process which was developed in 1997 by [2] along with two industrial partners. CRISP-DM consists of six phases intended as a cyclical process. Being cyclical model provides a chance for researchers to revise their work iteratively. This model begins with business understanding, then captures and understands data, Data preparation and applies data mining techniques for model building, evaluates the model results, and deploys the knowledge gained in operations. The CRISP-DM offers a uniform framework for experience documentation and guidelines and it can apply in different industry with different type of data.

3.2. Research Tools

The Waikato Environment for Knowledge Analysis (WEKA) is a widely used tool (Software) for data mining research. It is open source software which supports several standard data mining tasks. The input data for WEKA can be loaded from various sources. The researcher used ARFF (Attribute Relation File Format) format for this study. WEKA data mining tool has J48 implementation with tree pruning method. When a decision tree is built, many of the branches will reflect anomalies in the training data due to noise or outliers. Tree pruning methods use statistical measures to remove the least reliable branches [5]. For this particular research, decision tree implementation with J48 and NaiveBayes available in WEKA are selected to be used for classification purpose. The nature of data used for this research can best manipulated with these classifiers.

3.3 MODEL EVALUATION

Taking the results of the data mining models for granted, without any evaluation process it could be very risky and lead to wrong decision making [5]. There are different evaluation of the performance of the model used by this research work like Confusion matrices, number of correctly classified instances, number of leaves in a tree, and size of the trees, execution time, and ROC area. The method of validation for decision tree is decided to be full training set splitting at 80% of the dataset for training and the rest 20% for testing dataset. The following learning curve shows highest performance i.e. 81 at 80% sample dataset.

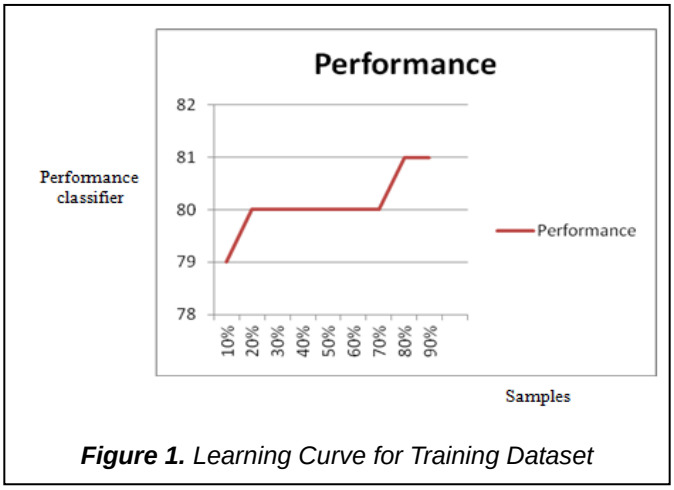

Figure 1. *Learning Curve for Training Dataset*

4 RESEARCH RESULTS

The result of attribute selection in WEKA indicates the top six determining attributes of the dataset for predicting success/failure status of a student, which are: Sex, Higher Education Entrance Certificate Examination result, number of courses given in a semester, College, Department, and

Number of students in a Class. By conceptual hierarchy Department is represented in terms of College so as to minimize complexity of experiment results. The researcher has set ten testing options with two categories (using 10 folds cross validation, and percentage 80% split) to build all classification models with reduced attributes. Experiment was done on the Original Dataset as well as on the Resampling Dataset. Resampling is a sampling technique available in WEKA which produces a random subsample of a dataset. It increases performance in case of unbalanced dataset. The following tables show the summary of results of the experiment that is done on the unbalanced and balanced dataset.

Classifiers	TP Rate	Precision	ROC area
J48 without pruning	0.814	0.787	0.787
J48 with pruning	0.813	0.782	0.657
J48 with parameter tuning	0.811	0.777	0.654
NaïveBayes (10-fold cross validation)	0.802	0.772	0.754
NaïveBayes (80% split)	0.80	0.767	0.739

***TABLE 1**. RESULTS FOR THE ACCURACY OF THE APPLIED CLASSIFIERS ON THE UNBALANCED DATA SET*

Classifiers	TP Rate	Precision	ROC area
J48 without pruning	0.923	0.922	0.958
J48 with pruning	0.922	0.92	0.950
J48 with parameter tuning	0.916	0.914	0.938
NaïveBayes (10-fold cross validation)	0.874	0.865	0.887
NaïveBayes (80% split)	0.863	0.865	0.877

***TABLE 2.** RESULTS FOR THE ACCURACY OF THE APPLIED CLASSIFIERS ON THE BALANCED DATA SET*

From the above tables we can see that the accuracy of the classifiers on the balanced data set is greater than the accuracy of the classifiers on the unbalanced data set. The decision tree classifier is also performing more than the Naïvebayes classifier. The accuracy of Naivebayes (10-fold cross validation) classifier on the balanced dataset is greater than the accuracy of the same classifier with 80% split. On the other hand we can compare the decision tree classifiers on the balanced dataset by the number of leaves, size of the tree, and the time required generating rules. Here consider the advantages of tree pruning and parameter tuning in WEKA. Pruned trees tend to be smaller and less complex and, thus, easier to comprehend. They are usually faster and better at correctly classifying independent test data than unpruned trees [5].

Classifiers	number of leaves	size of the tree	Correctly classified instances	Incorrectly classified instances
J48 without pruning	81	120	92.33%	7.66%
J48 with pruning	73	106	92.16%	7.83%
J48 with parameter tuning	32	46	91.62 %	8.37%

***TABLE 3.** RESULTS FOR THE ACCURACY OF THE APPLIED CLASSIFIERS ON THE BALANCED DATA SET*

5 CONCLUSIONS AND RECOMMENDATION

This study has shown that data mining techniques can be applied by higher education institutions or universities in determining student failure/success rate so that managing students' enrollment at the beginning of the year, assist students before they reached risk of failure, effective resource utilization and cost minimization, helping and guiding administrative officers to be successful in management and decision making. The objective of this research undertaking was to investigate the possible application of data mining technology in the Ethiopian higher education context, particularly at Debre Markos university students data, by developing a predictive model that could help higher education institutions to identify university students at risk of failure so that they can be treated before the condition escalate into students academic dismissal and wastage of resources. A data set of totaling 11,873 records of students was used to build and test both decision tree and NaiveBayes models. In case of identifying those students who are not performing well and those who are at the risk of failing (CGPA below 2.00) with highest prediction accuracy is that of decision tree at 92.34% as compared to naïveBayes classifier. The balanced datasets were used to minimize the impact of majority class data sets on degree of accuracy. Hence the performance of decision tree classifier on the resampling data set was increased than the unbalanced dataset. In general, the results obtained from this research have shown the potential applicability of data mining technology to classify university students' academic performance as failure/success. It was possible to identify the main determining attributes/variables and their values for the failure or success of students in a specific college; number of students in a class, number of courses given in a semester, Higher Education Entrance Certificate Examination result of a student, and sex were the main determining attributes obtained from this research result. The generalized decision tree with pruning by some parameter tuning found to be the best relevant technique on the dataset to get meaningful patterns from the decision tree experiments. Based on the findings of this research work, the researcher would like to make the following recommendations: In this research work, an attempt has been made to assess the applicability of data mining technology to predict the likelihood of student success/failure in the university by using some set of variables/attributes that were considered important by different literatures. For a number of other variables, in education area of Ethiopia; especially student performance versus health related problems, financial resource problems, family

background, academic schedule and assessment method, qualification of lecturers and much more, it remains to investigate further the effect of those variables to build models with better accuracy and performance than the models built in this research work. Even though there are many data mining techniques, the researcher done experiment by classification only. The number of the scenarios for experimentation were also very few. But the other data mining techniques which were not tested by the researcher might reveal important patterns in relation to factors affecting student success/failure. Therefore future research works can be explored by other data mining techniques.

REFERENCES

[1] Berry, M.,& Linoff, S. (2000). Mastering Data Mining: The Art and Science of Customer Relationship Management. New York: Wiley.

[2] Cross Industry Standard Process for Data Mining (CRISP-DM). August 29, 2004, Available at: www.crisp-dm.org

[3] Dekker, G., Pechenizkiy, M., and Vleeshouwers, J. (2009). Predicting students drop out: A case study. Proceedings of the 2nd International Conference on Educational Data Mining, EDM'09, 41-50.

[4] Dr.Vuda, S. & Capt. Genetu, Y. (2012). Improving Academic Performance of Students of Defence University Based on Data warehousing and Data mining. August 08/2004, Available at: http://www.dmu.edu.et

[5] Han,J. & Kamber, M. (2006). Data Mining: Concepts and Techniques(2nd edition).

[6] Thai, N. (2007). A Comparative Analysis Of Techniques For Predicting Academic Performance. In Proceedings of 37th conf. on ASEE/IEE Frontiers in Education.

[7] Vandamme, J.P., Meskens, N., & Superby, J.F. (2007). Predicting Academic Performance by Data Mining Methods.