Proceedings of the 9th National Conference on Private Higher Education Institutions (PHEIs) in Ethiopia

Major Theme: The Role of Private Higher Education Institutions in Human Capital Development to achieve the Ethiopian Growth and Transformation Plan

Organized & Sponsored by

St. Mary's University College

August 20, 2011

UNECA Conference Center

Addis Ababa, Ethiopia

Using data mining technique to predict student dropout in St. Mary's University College: Its implication to quality of education

By

Getahun Semeon, St. Mary's University College, P.O.Box 18490, Addis Ababa, Ethiopia.

Abstract

One of the major challenges of PHEIs that affect their performance is the increasing number of dropouts. In order to solve this problem, PHEIs must identify the dropout trends and the major determinants of higher dropout rates. Data mining is becoming a new source of data for higher education institutions that can be used as a means to identify trends of dropout and its possible determinants. Despite the expansion of Private Higher Education Institutions (PHEIs) and enrollment of students in both undergraduate and postgraduate programs, there is high and increasing dropout rate in both private and public HEIs of Ethiopia. The challenge is even more significant in private HEIs. An extensive literature search did not show any study conducted in the areas of application of data mining or other technique to predict dropout within the context of Ethiopian HEIs and other low income countries. Therefore, demonstrating the possibility of applying data mining technique in the areas of student dropout within the context of Ethiopian HEIs is quite relevant and innovative. This study is concerned with applying data mining technique for better and on time prediction of dropout of degree students. The basic research question of the study is: Can the traditional machine learning be applied to rank students by their likelihood to dropout? Classification and feature selection algorithms have been used to build the prediction models. One R, RandomForest and Neural Network (Multi-layerperceptron) demonstrated the highest performance in terms of highest percentage of correct classification. The accuracy of the classifiers ranges between 87% and 94.5%. CGPA is selected as the strongest predictor of dropout which is followed by Term1 and Term2 GPAs. Age and previous college result are in the fourth and fifth place in terms of their predictive power.

Key words: Dropout, Data mining, Classification, Feature Selection, Decision Tree, J48, RandomForest, Neural Network Multilayerperceptron, Higher Education Institutions.

Introduction

Knowledge has become the key of the future prosperity and social well-being of nations and, thus, education becomes one of the key sectors that contribute to the economic and social advancement of a nation. Achievement of sound economic development result is almost impossible without a well-established education system. Due to an increasing number of students and institutions, higher education institutions (HEIs) vision became increasingly oriented to performances setting, their measurement and accordingly developing strategies for better achievements (Jadrić et al., 2010). But, one of the major challenges of HEIs affecting their performance is the increasing number of dropouts. In order to solve this problem, HEIs must identify the dropout trends and the major factors that contribute to higher dropout rates. That will, in fact, enable the institutions to plan, manage and control the education process with the purpose of improving the efficiency of studying (Jadrić et al., 2010). One of the techniques that can be used to identify trends of dropouts and possible determinants is data mining, and this approach is the object of this study.

Data mining is a process of extracting previously unknown, valid, potentially useful and hidden patterns from large data sets (Connolly, 1999 as cited by Ayesha et al., 2010). From last years onwards, data mining has become an important model with wider application in higher education institutions because of the huge and growing educational data available in their databases, since enabling the institutions to extract important but hidden relationships among the data sets. Institutions of higher education use data mining techniques for different purposes, including understanding factors affecting students' performances and their learning behavior, warning students at risk before their final exams and anticipating possible drop-outs. Clustering and decision tree are most widely used techniques for future prediction (Ayesha et al., 2010).

In Ethiopia, there has been an expansion of HEIs for the past five years. In the last four to five years, a total of 13 new public universities were built and made operational, raising the number of public HEIs to 22. On the other hand, about 55 private higher education institutions which offer degree programs have been established in the past ten years. Because of this, an annual increase in enrollment of 22.3% has been registered for undergraduate and 29.5% for postgraduate programs. Despite this significant increase in enrollment, the dropout rate is expected to be high, especially in PHEIs of the country.

Study Background

St. Mary's University College is one of the private HEIs in Ethiopia, established in the year 2000, and it is one of the two private higher education institutions that offer post graduate programs. It also offers degree, diploma and certificate programs in conventional mode (Regular and Extension face-to-face classes) as well as in distance education mode, with a total of 22 departments in both modality offering degree, diploma and certificate programs. The total number of students in both modes is 40,393; of which 16,833 are students in the degree programs, 20,545 diploma students and 3,015 students in the certificate program in both conventional and distance modes. The total number of full time academic staff is 206, and that of administrative staff is 778. The University College is operating in a competitive environment with the expansion of public and private HEIs. Based on the data on the number of students graduated in those programs, the dropout rate is estimated to be between 30 and 40%. Despite this large estimated dropout rate, there is no practice of early anticipation of the problem and no timely corrective action taking. Because of this problem, the University College is thus facing unexpected loss of revenue. This problem is not peculiar only for St. Mary's University College, but it is also common in other HEIs.

Objective of the study

The objective of the study is, therefore, to explore the possibility of applying data mining technique for predicting the likelihood of students to dropout, with the intention of developing possible retention strategy and decreasing the number of dropout students. The research question of the study is: Can the traditional machine learning be applied to rank students by their likelihood to dropout?

Justification for the study

As it is stated in the introduction part, Ethiopian HEIs are characterized by high dropout rate, but there is no system of early identification or prediction of this problem and there is no timely corrective action being taken by the management of these institutions. Because of such failure of early identification of dropout attitude, institutions of higher education are unable to understand the factors affecting students' performance that lead students to dropout and to develop a strategy that improve the efficiency of learning process. This situation is affecting the performance of the HEIs and exposing them to unexpected loss of revenue, especially for those institutions that depend on students' tuition fee. Therefore, demonstrating the possibility of applying data mining technique in the areas of student dropout within the context of the Ethiopian HEIs is quite relevant and important. The good result of the model is measured in terms of its early predictive capacity of the likelihood of students to dropout in a more precise manner. With a better quality predictor and identification of the risk factors, the management can take corrective action towards supporting the risk group and minimizing attrition rate through improving personalized academic support services, designing different forms of tutorial services, strengthening orientation and induction in problem areas and creating ways of improving student engagement. Academic departments can also review their current assessment process.

Novelty

As presented in the section below that deals with related work, data mining models are being widely implemented in higher education institutions in different areas, mainly corresponding to students' performance assessment and analysis. But almost all of the available studies are conducted in economically advanced countries. Based on the extensive literature survey, there is no research conducted in the areas of applying data mining or other technique for predicting students' dropout within the context of economically underdeveloped countries in general and Ethiopian HEIs in particular. Therefore, the application of this technique for predicting student dropout within the context of Ethiopian HEIs with peculiar socio-economic, political, infrastructural, demographic, geographic, etc. features make the study a novel one, hoping to fruitfully contribute to the debate.

The first part will present a review of related literature, followed by the presentation of the methodology where the analysis used tools are detailed, the quantitative techniques as well as the evaluation of the study findings. Next, the results of the data analysis will be presented. Finally, we present discussions on the major findings of the study in order to draw data-based conclusions, and then the study suggests actions to be taken in the future work in terms of developing the technique.

Related Work

Institutions of higher education are increasingly becoming result oriented, and, thus, they are required to measure their performances. This is mainly because they are working in a highly competitive environment and are aiming to get more competitive advantage over the other business competitors (Quadri and Kalyankar, 2010). One of the performance measures in these institutions is the student dropout rate. In their study of student dropout analysis using data mining method, Jadrić et al. (2010) stated that:

> *An indicator of potential weaknesses in the higher education system may be a large number of dropouts in the first years of studies. The strategic goal of Higher Education Institutions should therefore be planning, management and control of education processes with the purpose of improving the efficiency of studying.*

Student retention is, in fact, the key issue that higher education institutions are concerned with, as more students remaining in the university means better academic programs and higher revenue (Zhang, et al., 2010) and early identification of students at risk as well as maintaining intensive continuous intervention is the key to increase student retention (Seidman, 1996 as cited by Zhang, et al., 2010). This would in fact allow educational institutions to undertake timely and pro-active measures. Kovačić (2010) even stated that once identified, these 'at-risk' students can be then targeted with academic and administrative support to increase their chance of staying on the course.

Earlier models like that of Tinto (1982) were even used by institutions to predict student dropout. Such models contributed in identifying factors that determine student's academic success. According to the same author, it is the socio-psychological interplay between the characteristics of the student entering university and the experience at the institute that eventually determine the student attrition. Tinto further argued that, from an academic perspective, performance, personal development, academic self-esteem, enjoyment of subjects, identification with academic norms, and one's role as a student contribute all to a student's overall sense of integration into the university (Tinto, 1995). According to his model, a higher degree of integration is directly related to a higher commitment to the educational institute and to the goal of study completion. Based on this model, other studies identified factors like peer group interactions, interactions with faculty, faculty concern for student development and teaching, academic and intellectual development, and institutional and goal commitments that affect the student's integration (Dekker et al., 2009).

Applying data mining technique in the field of education to students' dropout is a recent phenomenon (Dekker et al., 2009). Data mining, as defined by Quadri and Kalyankar (2010), "is the process of analyzing data from different perspectives and summarizing the results as useful information. It is also defined as the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data". Data mining can find relationships and patterns that exist but are hidden among the vast amount of educational data and variables, and combines machine learning, statistical and visualization techniques to discover and extract knowledge in such a way that humans can easily comprehend (Zhang, et al., 2010).

Actually, there are an increasing number of data mining applications in education, from enrollment management, graduation, academic performance, gifted education, web-based education, retention and other areas (Nandeshwar and Chandhari, 2009 as cited in Kovačić, 2010). By consulting different studies, Zhang, et al. (2010) summarized different areas for the applications of data mining technique within the university context. Some of the areas include: identifying the most effective factors to determine a student's test score and then applying the result to improve student's test score performance in the following year; grouping students to determine which student can easily pile up their courses and which take courses for longer period of time which helps the universities to identify requirements of their students and to decide on how to offer courses and curriculum; predicting students final grades based on their web-use feature so that students at risk can be identified early, thereby allowing the tutor to provide them with appropriate advice in a timely manner.

Data mining provides wide range of benefits to academic institutions. Some of the benefits include: It enables educational institutions to plan their monitoring and support mechanisms (Dekker et al., 2009), to provide personalized education (Zhang, et al., 2010), to attain more accuracy in selecting the kind of training to offer to different kinds of students (Quadri and Kalyankar, 2010), to direct its resources to the students who need it most (Dekker et al., 2009), and to follow the

dropout trend throughout several years in order to check the effectiveness of corrective activities (Jadrić et al., 2010). This process eventually results in a decrease in the dropout rate (Dekker et al., 2009). Zhang, et al. (2010) also stated that as compared to traditional analytical studies, data mining is forward looking and is oriented to individual students. In general, data mining is becoming a new source of data for higher education institutions that can be used as guides for course redesign and as evidence for implementing new assessments and lines of communication between instructors and students (Baepler and Murdoch, 2010).

Regarding the data mining techniques which are used in the higher education context, Baepler and Murdoch (2010) stated that most of the work that has been done in higher education falls into the categories of clustering, classification, visualization, and association analysis. Dekker et al. (2009) also stated that techniques like clustering, classification, associations, Bayesian networks and neural networks are the algorithms that have been used in different studies in educational data mining although the methodology is not yet transparent and it is not clear which data mining algorithms are preferable in this context. The following paragraphs present some of the previous related studies to demonstrate the applications of some of the data mining techniques in educational institutions and the type of determinant factors for academic success.

Dekker et al., (2009) conducted an educational data mining case study aimed at predicting the Electrical Engineering (EE) students drop out after the first semester of their studies or even before they enter the study program as well as identifying success-factors specific to the EE program. They consider data collected over the period 2000 – 2009 that contains information about all the students being involved in the EE program. They selected a target dataset of 648 students who were in their first year phase at the department. They applied decision tree algorithm for their study. Their finding revealed that simple classifiers give a useful result with accuracies between 75 and 80%, difficult to beat with other more sophisticated

models. In addition, they also identified the strongest predictor of success in that specific department which are mainly course based.

Jadrić et al. (2010) conducted analysis of student dropout using data mining method by taking Faculty of Economics in Croatia as a case. By analyzing the existing transaction data on students, they aimed at collecting additional information and defining the crucial processes that have to be adjusted for the purpose of improving studying efficiency. They carried out a detailed analysis of dropout by use of logistic regression, decision trees, and neural networks. The data mining is conducted in SAS 9.1 Enterprise Miner. The attributes used in the study include: ID, Generation, Sex, Date of Birth, Status, Study Program¸ Points obtained from the secondary school, Enrolment Rank, Father Qualifications, Mother Qualifications, Social Status, Housing Indicator, Secondary School, Last year of study, and Last year of enrolment.

Zhang, et al. (2010) conducted a study on improving student retention in higher education by using data mining. The study focused on how data mining can help spot students 'at risk', evaluate the course or module suitability, and tailor the interventions to increase student retention. They considered one year data. The types of students' data include: average mark, online learning systems information, library information, nationality, university entry, certificate, course award, current study level, study mode, postgraduate or undergraduate, current year, age, gender, race and etc.

Kovačić (2010) conducted a study on the prediction of successful and unsuccessful students at the Open Polytechnic of New Zealand. They considered socio-demographic variables (such as age, gender, ethnicity, education, work status, and disability) and study environment (course program and course block) that may influence persistence or dropout of students and examined to what extent these factors, i.e. enrolment data help in pre-identifying successful and unsuccessful students. They used data mining techniques (such as feature selection and

classification trees) to identify the most important factors for students' success and a profile of the typical successful and unsuccessful students. In their study, they found that ethnicity, course program and course block are the most important factors that separate successful from unsuccessful students. Among the techniques, Classification and Regression Tree (CART) was the most successful in growing the tree with an overall percentage of correct classification of 60.5%.

Quadri and Kalyankar (2010) also conducted a study on the work of data mining in predicting the dropout feature of students. The study applied decision tree technique to choose the best prediction and analysis. They identified gender, attendance, previous semester grade, parent education, parent income, scholarship, first child, and part time job as factors that determine students' dropout. After the study had been conducted, the list of students who would be predicted as likely to dropout from college was submitted to teachers and management for direct or indirect intervention.

Having consulting various related studies, Kovačić (2010) summarizes those techniques which can be used by different authors in their research when their main focus is on study outcome. The techniques include:

- Binary logistic regression in order to identify the most significant factors to determine whether or not students passed, failed or dropped out for courses in the mathematics and computing faculty at the Open University in UK (Woodman, 2001).
- Decision trees, artificial neural networks, naïve Bayes classifier, instance-based learning, logistic regression and support vector machines to predict student's performance at the Hellenic Open University (Kotsiantis, Pierrakeas and Pintelas, 2004).
- Decision trees, neural networks and linear discriminator analysis for the early identification of three categories of students: low, medium and high-risk students (Vandamme, Meskens and Superby, 2007)

- Classification tree based on an entropy tree-splitting criterion to differentiate the predictors of retention among freshmen enrolled at Arizona State University (Yu et al., 2007).
- Decision trees, random forests, neural networks and support vector machines to predicted the secondary student grades of two core classes using past school grades, demographics, social and other school related data (Cortez and Silva, 2008).

According to Dekker (2009):

> *Many studies included a wide range of potential predictors, including personality factors, intelligence and aptitude tests, academic achievement, previous college achievements, and demographic data and some of these factors seemed to be stronger than others however there is no consistent agreement among different studies.*

Data Preparation

The data of students is extracted from the student information management system of the St. Mary's University College. The researcher took all the regular and extension degree students who had got registered between the years 2004/05 to 2010 representing the population. The total number of degree students registered in the specified period was about 8,743. These students were from five departments, i.e., from Department of Accounting (3,481), Management (2,293), Law (905), Computer Science (894) and Marketing Management (1,170). The total number of students in the regular program was 2,613; while in the extension program, the number was 6,130. The students that we are looking for to build our model are those who were registered in the years of 2004/05, 2005/06, 2006/07 and 2007/2008. A total of 5,158 students of these batches have already graduated. However, many of the required attributes of the students were missing for batches 2004/05, 2005/06 and 2006/07. Therefore, a new admission form was designed with expanded data entry fields and implemented starting from 2007/08. Thus, most of the required attributes for building the model were generated from the graduates of 2007/08. As

data were available on the graduates of 2007/08, we selected a target dataset of 1,362 students of 2007/08, who already graduated and had the status of "Graduated" and "Dropout" in the dataset.

In this study, the predictors or data types used in building the model include: student demographics (Age, Sex, Employment Status, and Income), pre-college experience (High School Result, Name of Previous College, and Previous College Result) and University College experience (Grade Point Average or GPA of 10 terms and CGPA). A total of 19 attributes were considered and involved in the experiment. These data types are targeted since many studies conducted in the area, as indicated in the review of related work, referred to such factors. All these attributes were considered as independent variables and the dependent or target variable is the dropout rate.

Regarding data preparation, the data were first exported from My-SQL database to EXCEL. Then, we excluded attributes with very significant missing values. All missing values of the selected attributes are replaced with average values and "?". The target is set with nominal values of "pos" representing dropout and "neg" representing non-dropout /graduated. Since some of the attributes are important for the study, inserting a relatively large number of average values and "?" may result in a certain biases on the outcome of the model. It would have been good to include marital status and source of financial support as important input variables in the model, but they couldn't be included since there was no adequate data.

The dependent (output/predicted) variable of the model is dropout, which is represented by "pos". Because of the small number of instances, we couldn't divide the data into training and test set and, thus, the whole 1,362 instances were used as training/test set.

Data Mining and Analysis

Data Mining Techniques

In the experimental study, we used Weka classifiers with their default settings and feature selection algorithm. We compared two algorithms from decision tree (*J48, RandomForest*) and one algorithm from Neural Network (*Multilayerperceptron)*, selected according to suggestions made in similar other studies. Since we have a lot of missing values in our data, we found decision trees as relevant modeling technique. The advantages of decision trees lies in its flexibility in terms of tolerating missing values, which is not the case in Neural Network, and is also important for the classification of attributes regarding the given target variable (Panian and Klepac, 2003 as cited by Jadrić et al., 2010). Decision trees are attractive because they offer, in comparison to neural networks, data models in readable, comprehensible form, in the form of rules (Jadrić et al., 2010). They are used not only for classification but also for prediction (Gamberger and Šmuc, 2001). Neural networks are selected because they are powerful tools in trend prognostics and predictions based on historical data (Jadrić et al., 2010), which is relevant feature for our particular study and they perform very well in more complex classification problems (Jadrić et al., 2010). Their disadvantage, in comparison to simpler methods, is the relatively slow and demanding process of model "learning" (optimization of weight factors) (Gamberger and Šmuc, 2001 as cited by Jadrić et al., 2010).

We also considered the *OneR* classifier as a baseline and as an indicator of the predictive power of particular attributes (Dekker et al., 2009) using 10-fold cross validation for estimating generalization performance.

Analysis – Classification

As stated-above, different classifiers have been used, whose results vary depending on their efficiency and complexity level. We have summarized the accuracy level of each classifier in Table 1 as follows:

Classification Accuracy = (TP+TN)/ m, where m is the number of test instances.

Table 1- Classifier accuracy comparison.

Classifier	Accuracy
OneR	94.5%
J48	90.3
RandomForest	89.06
NN – Multilayerperceptron	87

Source: computed by the author.

As indicated in the above table, considering OneR as a benchmark, the higher performance is demonstrated more by decision trees (*J48 and RandomForest)* than Neural Network. But it is clear, from the table, that no other classifier performs better than the benchmark *OneR*.

Comparison is also made based on area under ROC curve. The findings of the comparison are presented as follows.

Table 2 - Area under the ROC curve.

Classifier	Area under ROC curve
OneR	0.943
J48	0.89
RandomForest	0.93
Multilayerperceptron	0.925

Source: Computed by the researcher.

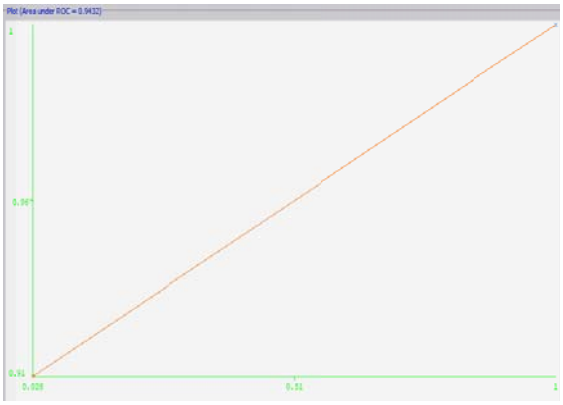


Fig. 1 ROC curve for *OneR*

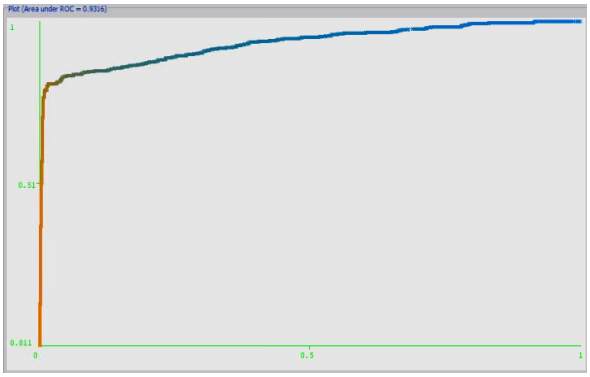


Fig. 2 ROC for *RandomForest*

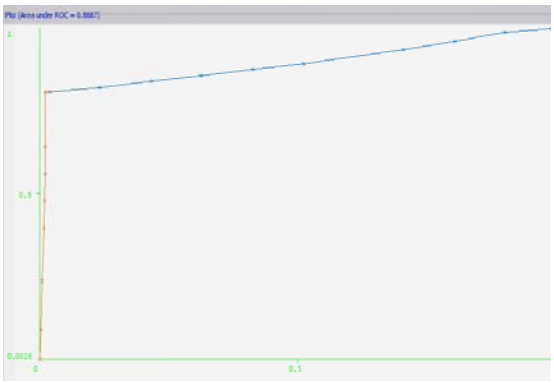


Fig. 3 ROC for *J48*

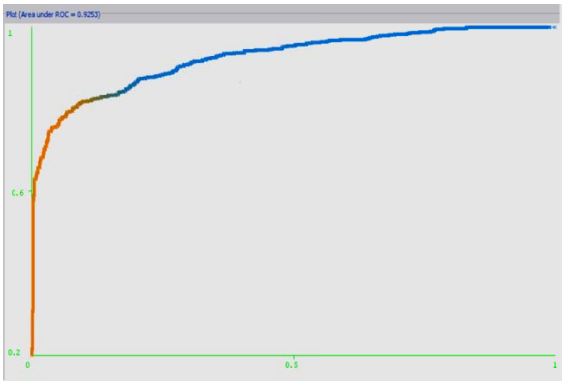


Fig. 4 ROC curve for *Multilayerperceptron*

From those tables and graphs, it is found that *OneR* is the best performer in terms of area under ROC curve which is followed by *RandomForest* and *Multilayerperceptron.*

Based on the confusion matrix we calculated, the true positive rate and the false positive rate for each classification algorithm and the results are summarized as follows:

Table 3 - True Positive Rate and False Positive Rate.

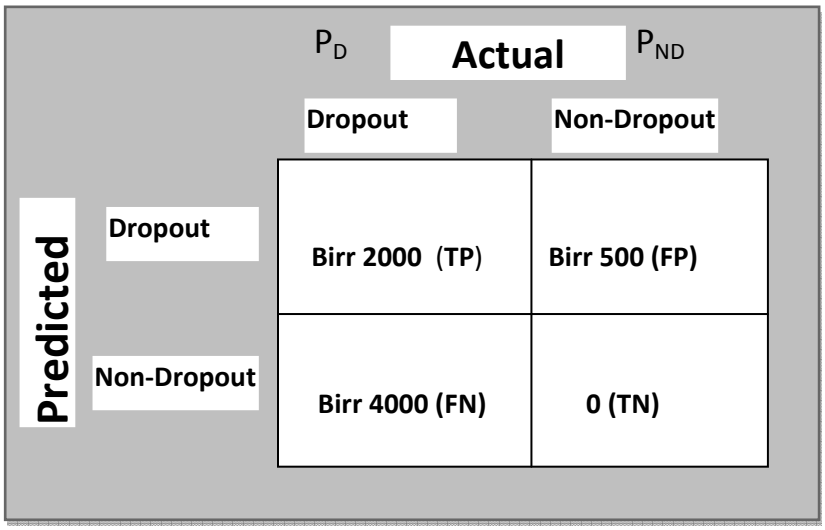
Classifier	True Positive Rate	False Positive Rate
OneR	0.91	0.027
J48	0.81	0.014
RandomForest	0.84	0.057
Multilayerperceptron	0.85	0.11

Source: Calculated by the author.

From the above analysis, it is clearly shown that other than the benchmark (*OneR)* *Multilayerperceptron* and *RandomForest* are statistically better performing classifiers.

Regarding the cost matrix, we estimated the possible loss of the tuition fee to be Eth. Birr 4,000 per student due to dropout, and by correctly predicting it, it is reasonable to assume that the University College minimizes the loss by 50%. There is also a cost that could be incurred by predicting students to dropout but they don't, and that will become an additional investment. The following results illustrate the cost matrix which is built based on this assumption.

Fig. 5 Cost Matrix



Therefore:

- the expected loss of predicting dropout

$= 2000P_D + 500P_{ND}$

$= 2000(1 – P_{ND}) + 500P_{ND}$

$= 2000 – 1500P_{ND}$

- the expected loss of predicting non-dropout

$= 4000P_D + 0\ (P_{ND})$

$= 4000P_D$

- we predict non-dropout when

$Loss_{ND} < Loss_D$

$(2000 – 1500P_{ND}) < 4000P_D$

"Lift" is also used in order to measure and to compare the performance of the three targeting models. In our case, the three models aim at identifying a subgroup (target) from a larger population. The target members selected are those likely to dropout. While comparing the three models using "Lift", the model is performing good if the correctly predicted within the target is much better than average for the population as a whole.

"Lift" is quantified by dividing the population into deciles – ten even groups – into which population members are placed, based on the probability of correctly predicted instances. Predictions with the highest probability are put into decile 1, etc. The following tables and charts show the performances of those three models in terms of "lift".

Table 4 - Lift Table for *J48*

A Decile	B Number of Instances by Decile	C Number of instances correctly predicted by the Model by Decile	D Percentage of correctly classified instances (C/B)	E Lift (D/122.8)	Cumulative instances correctly predicted
1	136	118	87	0.71	118
2	136	127	93	0.76	245
3	136	121	89	0.72	366
4	136	126	93	0.74	492
5	136	118	87	0.71	610
6	136	124	91	0.74	734
7	136	122	90	0.73	856
8	136	123	90	0.74	979
9	136	125	92	0.75	1104
10	136	124	91	0.74	1228
		122.8			

Source: computed by the author.

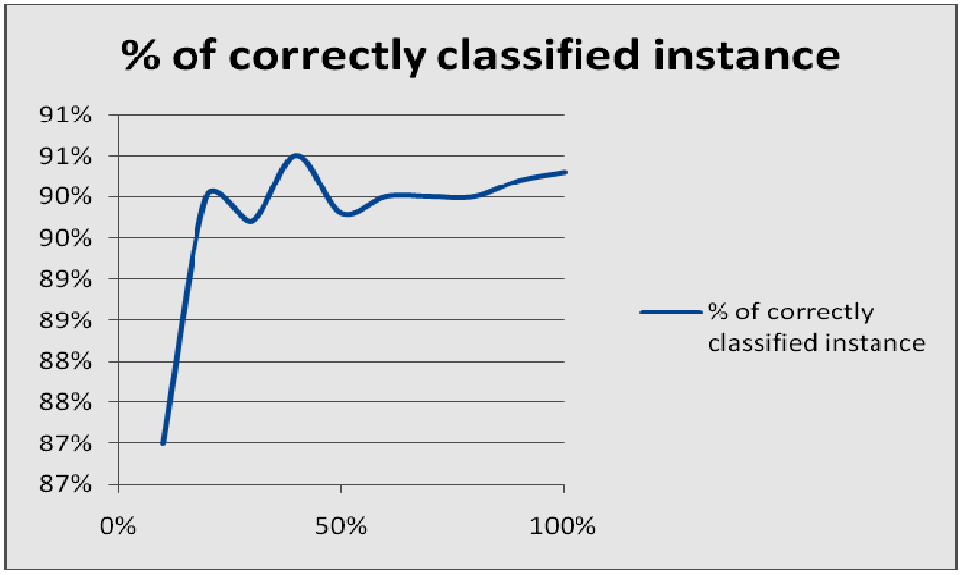


Fig. 6: Lift Chart for J48

Source: Experiment output by the researcher.

Table 5 - Lift table for *Multilayerperceptron*

A Decile	B Number of Instances by Decile	C Number of instances correctly predicted by the Model by Decile	D Percentage of correctly classified instances (C/B)	E Lift (D/117.8)	Cumulative instances correctly predicted
1	136	120	88	0.75	120
2	136	119	88	0.74	239
3	136	114	84	0.71	353
4	136	119	88	0.74	472
5	136	115	85	0.72	587
6	136	118	87	0.74	705
7	136	118	87	0.74	823
8	136	115	85	0.72	938
9	136	123	90	0.77	1061
10	136	117	86	0.73	1178
		117.8			

Source: Experiment data analysis output produced by the author.

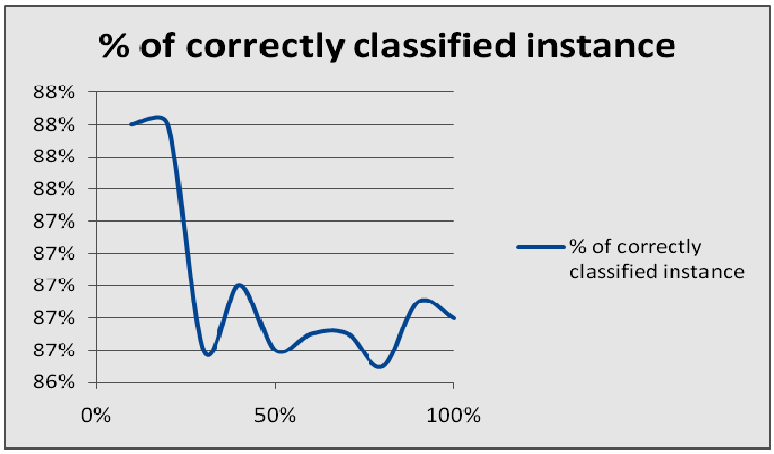


Fig. 7: Lift Chart for *Multilayerperceptron.*

Table 6- Lift table for *RandomForest*

A Decile	B Number of Instances by Decile	C Number of instances correctly predicted by the Model by Decile	D Percentage of correctly classified instances (C/B)	E Lift (D/121.8)	Cumulative instances correctly predicted
1	136	119	88	0.72	119
2	136	121	89	0.73	240
3	136	119	88	0.72	359
4	136	127	93	0.74	486
5	136	120	88	0.72	606
6	136	123	90	0.74	729
7	136	123	90	0.74	852
8	136	120	88	0.72	972
9	136	123	90	0.74	1095
10	136	123	90	0.74	1218
		121.8			

Source: The researcher's own experiment output.

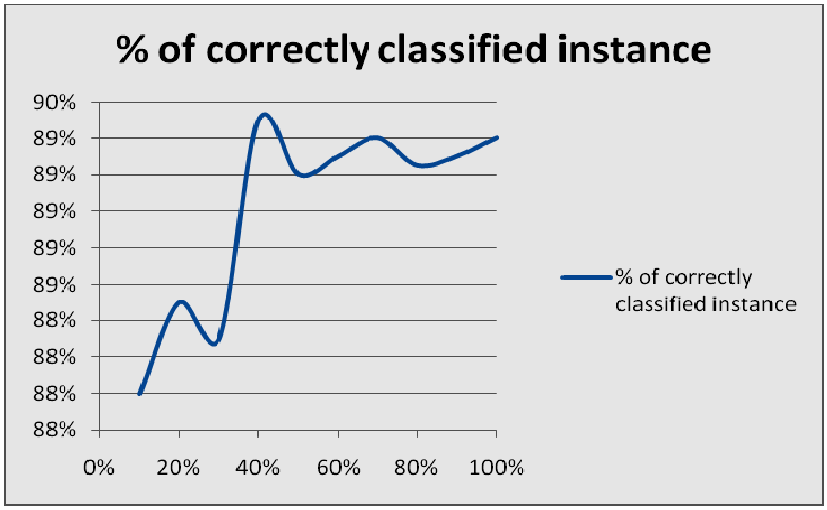


Fig. 8: Lift Curve for *RandomForest*

From the above lift tables and lift charts, we can infer that decision tree models (*J48 and RandomForest)* provide higher lift than a neural network (*Multilayerperceptron).* Such comparison provides a key factor in choosing between the two models, and thus according to the results decision tree model is preferred than neural network.

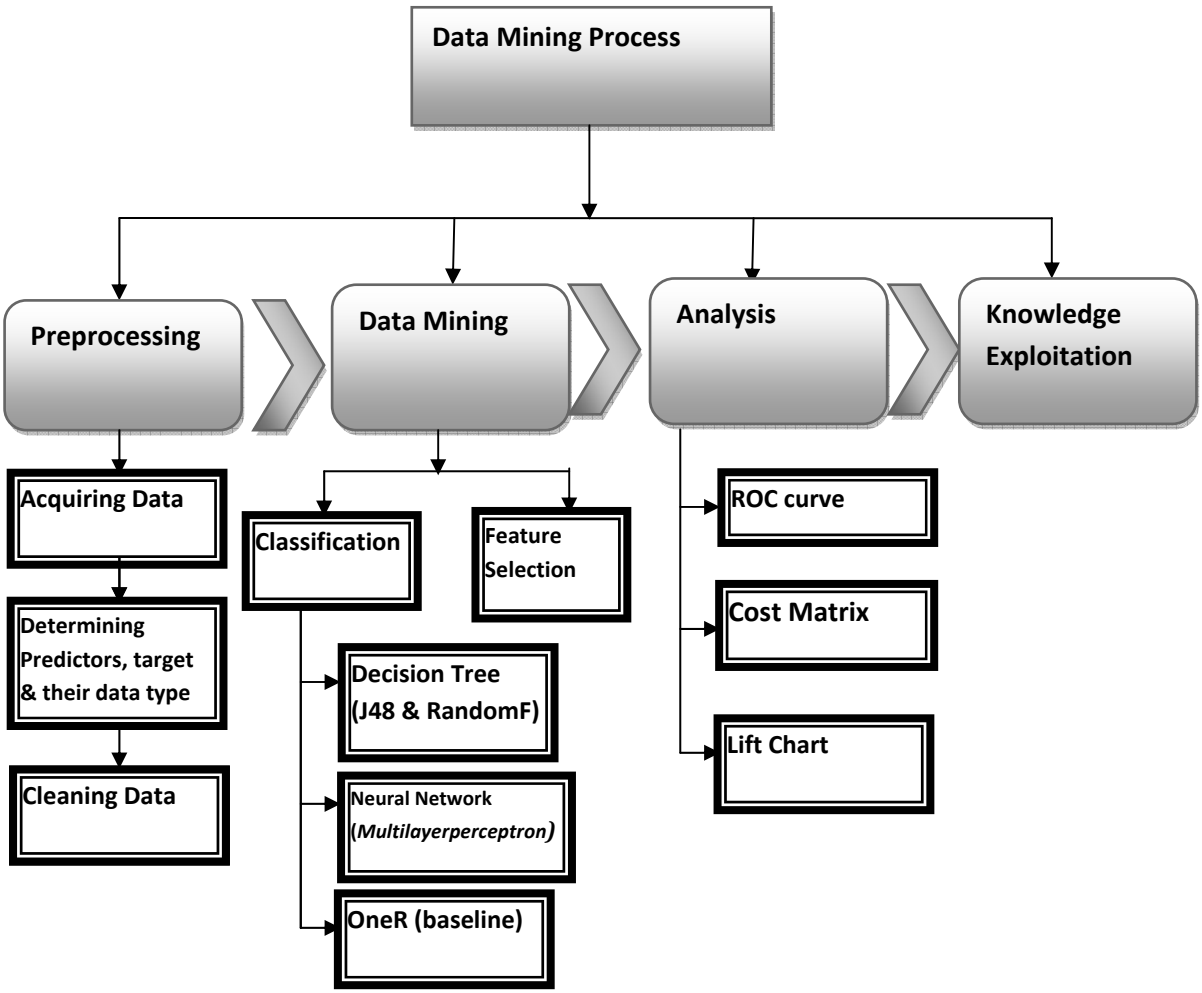
Feature Selection

One important part of the study is feature selection, which is used to rank the predictors according to the strength of their relationship with dependent or outcome variable, which is dropout in our case. Criteria CGPA are selected as the strongest predictor of dropout with a value of 0.62, followed by Term1 GPA and Term2 GPA with a value of 0.036 and 0.026 respectively. Age and Previous College Result are in the fourth and fifth place in terms of their predictive power.

Knowledge Exploitation

The knowledge gained in terms of predicting the likelihood of students to dropout can be exploited by identifying students at risk based on the model built, developing a strategy for improving the performance of students and reducing the attrition rate.

The whole data mining process in this study can be summarized by the following diagram.



Source: Designed by the researcher.

Conclusion and Future Work

For those HEIs which are currently facing strong competition in the sector, predicting student dropout is quite crucial. The study was conducted by taking the data of degree students of the St. Mary's University College as a case and applying

different data mining techniques, with special emphasis on classification and feature selection.

Based on the experiment, it was found that the accuracy level of the classifiers range between 87.0% and 94.5%. In the same manner the values of area under the ROC curve rangesfrom 0.89 to 0.943. In all the comparisons made, the *OneR* classification algorithm demonstrated the highest performance in terms of highest percentage of correct classification. In addition to this, classifiers like *RandomForest and Neural Network (Multilayerperceptron*) are frequently cited as most successful in correct classification. In terms of feature selection, the strongest predictor of dropout was found to be CGPA, followed by GPA of Term1 and Term2 as well as Age and Previous result in college.

On the whole, the findings of the study indicate that students' dropout is more related to performance than other predictors in the study considered. Therefore, the St. Mary's University College should develop mechanisms of providing academic support to those who are students at risk. The support can be in the form of providing tutorial services, strengthening orientation and conducting induction to first year students as well as revisiting assessment processes. These are some of the major corrective actions that should be performed by the University College.

Based on the findings of this study, there is a need to conduct further study that could contribute to increase the validity and predictive power of the models. The experiment was conducted with less rich dataset by taking only graduates of 2007/08. Conducting experiment on a time-serious and more consistent dataset covering all programs of study could result in building a better predictive power.

Annex 1 – List of attributes considered in the experimental study.

Attribute	Type	Remark
Sex	Nominal	M/F
Batch	Numeric	Entry year
Division	Nominal	Regular/Extension
Department	Nominal	Academic department in which the student attends
Program	Nominal	Degree
Age	Numeric	Age of the student
EmpStat	Nominal	Employment Status
Income	Numeric	Income of the student
HiSchRes	Numeric	High School Result / Result of National Exam
PrevCollege	Nominal	Name of Previous College Attended
PrevCollRes	Numeric	Final Result of Previous College Attended
PrevProg	Nominal	Program attended in previous college
Term1 – Term10	Numeric	GPA of 10 terms (Term1 up to Term10)
CGPA	Numeric	Cumulative Grade Point Average

Source: compiled by the researcher.

REFERENCES

Ayesha, S.; Mustafa, T.; Attar, A. R. and Khan, M. I. 2010. Data mining model for higher education system. *European Journal of Scientific Research, 43*(1), pp.24-29.

Baepler, P. and Murdoch, C. J. 2010. Academic analytics and data mining in higher education. *International Journal for the Scholarship of Teaching and Learning, 4*(2).

Crosling, G.; Heagney, M. and Thomas, L. 2009. Improving Student retention in higher education: Improving teaching and learning. *Australian Universities' Review, 52*(2).

Dekker, G. W.; Pechenizkiy, M. and Vleeshouwers, J. M. 2009. Predicting students dropout: A case study. *Educational Data Mining 2009.*

Herzog, S. 2006. Estimating student retention and degree-completion time: Decision trees and neural networks Vis-à-Vis regression. *New Directions for Institutional Research*, *131*.

Jadrić, M. ; Garača, Ž; and Ćukušić, M. 2010. Student dropout analysis with application of data mining methods. *Management, 15*(1), 31-46.

Jun, J. 2005. Understanding dropout of adult learners in e-learning. PhD Dissertation. Georgia, USA: The University of Georgia.

Kovačić, Z. J. 2010. Early prediction of student success: Mining students Enrolment data. *Proceedings of Informing Science & IT Education Conference (InSITE) 2010*.

Quadri, M. N.; and Kalyankar, N. V. 2010. Drop out feature of student data for academic performance using decision tree techniques. *Global Journal of Computer Science and Technology, 10(*2).

Tinto, V. 1982. Limits of theory and practice in student attrition. *Journal of Higher Education, 53*, 687-700.

Zhang, Y.; Oussena, S.; Clark, T.; and Kim, H. 2010. Use data mining to improve student retention in higher education – A case study. In *ICEIS - 12th International Conference on Enterprise Information Systems, 2010*. 8-12 June, Portugal.