PREDICTING FACTORS THAT AFFECT STUDENTS' ACADEMIC PERFORMANCE BY USING DATA MINING TECHNIQUES

Najmus Saher Shah
College of Computer Science & Information Systems
Institute of Business Management, Karachi

Abstract

The study attempts to investigate the influence of factors on students' academic performance by comparing the accuracy of different classifiers. The result will be useful in identifying good as well as weak students who may perform poorly and will be potential dropouts. Students are categorized in five groups according to their performance: "Very Good", who have a high probability of succeeding; "Good" students, who are above average and with little efforts, may 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. The data set comprised of 231 students out of 637 from a leading degree institute of Karachi: registered in the graduate programs of year 2009. Weka, an open source data mining software is used to explore the influence of factors on predicting students' academic performance. Dataset is applied to different classifiers of Weka : J48,RandomForest, RepTree and BFTree of Decision Tree, Bayes and NaiveBayes of BayesNetwroks, Logistic and RBFNetwork functions and JRip rule. The study also shows that re-sampling of data was a critical step which is the reason of the success of the study.

Keywords: Academic performance, data mining, Decision Tree, Bayesin Netwroks.
JEL Classification: Z000

I. Introduction

To identify potential drop outs of the institute's graduate program is a complex process mostly due to the fact that students coming from different backgrounds have certain characteristics as well as perceptions and apprehensions of the world of the university and their actions embody them.

Students' failure to integrate and acquire good grades are considered to be one of the main factors but many researchers have also suggested that there are various other factors that may affect students progress at the university level.

Predicting successful and unsuccessful students at an early stage of the degree program help academia not only to concentrate more on the bright students but also to apply more efforts in developing programs for the weaker ones in order to improve their progress while attempting to avoid student dropouts.

II. Literature Review:

Abdul Mannan(2007) explores the relation between academic and social integration and also the impact of student's integration in academics and social activities on their academic performance. A sample of 2400 full time undergraduate students enrolled in 3rd term of 2002 at the University of Papua New Guinea was considered. Social integration was studied on the basis of i) informal social contact with academic staff ii) extracurricular activities and iii) peer group interaction. Academic integration consisted of two criteria: i) academic staff concern for students' development and learning and ii) students informal contact with academic staff on academic matters. Data analysis was done by using the SPSS package. A strong negative relationship was found between academic and social integration which indicates that less integration in the social domain of the university was compensated by higher academic integration leading to students' persistence in continuing their studies. It was found that students' persistence differs according to their subject area of studies.

Ali (2010) conducted a study based on 324 students in different cities of Pakistan and concluded that student motivation is an element that determines students' attitude towards the learning process. Data was collected with the help of a questionnaire. Intrinsic motivation and extrinsic motivation i.e. motivation derived from external rewards were selected as independent variables and academic performance was selected as the dependent variable. Ali (2010) found that academic performance is positively influenced by intrinsic motivation and negatively influenced by extrinsic motivation. The aspiration within and various external factors help in motivating students to work hard. Highly motivated students' success in academics is stronger as compared to others.

Nattavudh, A.Vignoles(2009) showed that there is a significant gap in the dropout rate between students who have a strong family background and those who have a weak background. Various factors like income, occupation, education, social status can determine a background. It was found in the study that Students from higher socio-economic and more educated backgrounds have lower rates of dropouts.

Similarly, Ishitani (2003) has shown that first generation students (whose parents did not graduate from college) tend to dropout more as compared to others. He investigated the impact of independent variables such as race, gender, family income and high school GPA on the behavior of a dropout student. He also explored whether their influence vary at different points of a student's academic career. Five types of outcomes were considered; continue stopout, dropout, transfer, and graduate. Exogenous (race, gender, high school GPA, family income etc) and time dependent variables (college GPA, financial aid, and athletic status) are assumed to affect an individual student's outcome in each time period. A sample of college students in fall of 1995 with enrollment status for 5 academic years was collected. Information of student characteristics used is based on a freshman survey conducted during the 1995 freshman orientation. An event history model (survival analysis) was used which is a powerful statistical tool used to study events and their causes (Blossfeld, 2002). Survival function is estimated by product-limit-estimation and exponential models. Product line estimation model considered three groups of students with different parental educational backgrounds. The result showed that students whose parents both of whom were college graduate achieved the highest survival rates. Group of students who had one college-educated parent had a slightly lower rate as compared to those whose both of whose parents are college graduate in 1st and 2nd semester but this gap widened through 3rd to 6th semesters. The exponential model assumes that the effects of explanatory variables are constant and change proportionally over time (Ishitani, 2003). The exponential model 's result showed that the coefficient estimate for first-generation student was 0.253 indicating that first-generation students had attrition rates higher than those who had two college–educated parents. Thus, first generation students had 29% higher rate of departure than the reference group.

Another survival analysis modeled was the piecewise exponential model. Piecewise exponential model with period-specific effects model assumes that the effects of explanatory variables are constant within each period but vary across different periods (Ishitani, 2003). The risk of departure for first generation students were 71% higher. Minority students have lower attrition rate rather than their counterparts in first and second years. Students belonging to lower income had 49%higher risk of leaving in the first year but decrease to 26% higher as compared to high income group students in the second year. Overall outcomes in the study were consistent with the previous studies. First –generation students were more likely to depart than were their peers and risk of departure among first generation students varied over time.

Data mining techniques relate to the field of database and can be useful in predicting or forecasting various hidden trends in the data. It is an emerging field originally applied to the commercial transactions. It is being used in addition to traditional statistical methods in higher education for finding causes for students' behavior, their decision of transferring or quitting and identifying students at risk of dropping out. Data mining techniques can be very helpful in determining students' performance.

Data mining is comprehensively defined by the Gartner group(2000) as" "the process of discovering meaningful new correlations, patterns, and trends by sifting through large amounts of data stored in repositories and by using pattern recognition technologies as well as statistical and mathematical techniques"

Rubenking (2001) view data mining as "the process of automatically extracting useful information and relationships from immense quantities of data. In its purest form, data mining doesn't involve looking for specific information. Rather than starting from a question or a hypothesis, data mining simply finds patterns that are already present in the data."

Data mining techniques like decision tree, neural network, Bayesian network were used in the studies in predicting performance to come up with results that can more accurately predict students' performance as compared to traditional statistical methods.

Decision trees are a collection of nodes, branches, and leaves. Each node represents an attribute; this node then split into branches and leaves until the data are classified to meet a stopping condition.

Bayesian network is based on decision theory. It is a branch of probability and statistics which investigates how to minimize risk and loss when making decisions based on uncertain information. It is a graphical model that encodes relationships among variables that it models.(Rahel Bekele et al(2005).

Neural network is a set of interconnected nodes. Nodes at one layer are connected to all nodes at the next layer. The layer which receives input is called input layer and the final layer is called output layer. In between these two layers are hidden layers. The variables of the problem are acted upon and weighted by the entry layer, which then transmits this information to the hidden layers; these combine all the information into hidden layers; these combine all the information into a single value which is passed to the exit node and which acts as a kind of estimated value for the decision variable.(Vandamme et al ,2007)

Logistic regression is a type of predictive model based on linear regression technique that associates a conditional probability score with each data instance. (Richard J. Roiger)

Gerben W. Dekker (2009) used data sets of students' past education as well as university grades and related data collected from the electrical engineering department of Eindhoven

University of Technology to test their impact on students' performance and determine whether they can help in predicting performance. 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 bachelors program. Various simple and sophisticated data techniques were used and their results were compared. The overall result shows that simple classifier gives result with accuracy between 75 to 80%.

Thai Nghe et al (2007) 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 (AIT) 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 Can Tho University (CTU) in Vietnam included students admission data to predict GPA at the end of the first year. Using the data mining tool i.e Weka, it was found that predictions for the CTU data set were more accurate as compared to AIT due to larger number of records for CTU. Decision tree algorithm consistently performed better than the Bayesian network algorithm .Other findings were that accuracy in smaller classes was much lower than in larger classes. This was overcome by using resample function in Weka. It significantly improved the accuracy of prediction where the original data had a much smaller sample size.

Some students take more years than usual to complete their degree program. Universities have to anticipate this type of behavior in students and have to come up with policies and rules to discourage this type of behavior but before this they have to know the reasons or factors influencing students conduct.

To estimate student retention and degree-completion time, Herzog (2006) used data mining techniques which included decision tree, neural networks and logistic regression. He compared the prediction accuracy of these methods by generating data from three different sources. i)institutional student information for student demographics, academic, residential and financial aid information, ii) American college test's students profile section for parent income data, National student clearing house for identifying transfer-year enrollment of new full time freshmen who started in fall semester of 2000 to 2003. Multitopolgy neural network; creates several networks in parallel based on a specified number of hidden layers and nodes in each layer. It performed significantly better in identifying dropout but perform poorly in estimating who is likely to transfer.

Neural network with hidden layers achieved 25% improvement in predicting when analysis included only new students. This is also true for the decision tree and pruned neural network models. Data mining algorithms worked better with a large set of exploratory predictors used to estimate degree completion time. Where time to graduate is estimated for new and transfer student simultaneously, pruned neural network and C5. Decision tree performed better.

Pruned neural network starts with a large network of layers and nodes as specified and removes (prunes) weakest nodes in input and hidden layers during training.

C5.0 uses the 5.0 algorithm to generate a decision tree or rule set based on the predictor that provides maximum information gain. The split process continues until the sample is exhausted. Lowest –level splits are removed if they fail to contribute to model significance.

Mario Jadric et al (2010) attempted to find the dropout rate among students. Transaction data on students was collected through the faculty of economics Information system within the autonomous subsystem student service. Data mining was conducted in SAS 9.1 Enterprise Miner (software) by using techniques of decision tree, neural network, and logistic regression. The model developed was based on SEMMA (sampling, exploring, modifying, modeling and assessment) methodology. It was found that women dropout comparatively less than men and students with better entrance ranking dropout less. The analysis separated the causes of students' dropout and it also determined the typical profile of the student inclined to dropout.

William R. Vetch's (2004) study was conducted to identify the relationship between independent variables such as academic performance, age, gender, ethnic group; socio-economic status, grade point average on the dropout behavior in high schools. He used a decision tree model to test his hypotheses. He utilized extant data resources and all variables were extracted from district electronic databases. High school students recorded as "dropped" (no transfer record) over the course of 2001-2002 academic year were matched with a random sample of non- dropouts. The study revealed that academic performance is the most related variable to dropping out behaviors. Older students are more likely to dropout as compared to younger students. The tree exhibits a certain ability to predict which students may drop out of school.

Vandamme et al (2007) uses three sets of factors i)history of student (his identity, socio-family past, academic past, age, gender), ii)Student's involvement in studies (participation in optional activities, meeting with lecturers and 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 (based on Shannon's entropy and ID3 algorithms), Neural network (multi layer), Linear Discriminant Analysis were used and SAS /Enterprise Miner software was used in applying these techniques. 20% variables showed significant correlations with academic success and 80% rate of correct classification in predicting the success or failure of students.

The objective is to not only to identify the factors that have an effect on the students' academic performance but to compare the accuracy of different classifiers in predicting students' performance. The goal is to categorize students' performance 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. Thai Nghe et al (2007) in his study also grouped students in different classes according to their performance and compared and evaluated the accuracy of Decision Tree and Bayes algorithms for predicting performances.

III. Data:

The attributes selected are identified and tested by different studies reviewed above (section II). Factors that are part of the study are given in the Appendix Table 1along with the variables that determine these factors.

Students registered in Fall 2009 belonging to BBA, BS programs of a private business university in Karachi are targetted. For the study, the collection of data has been done through online questionnaire .Complete questionnaire of 231 students out of 637 were received. Data including list of students enrolled and current GPA i.e. GPA of first year was collected from the Institute.

Online questionnaire was developed (Appendix 3)which is based on the findings of the literature review .The Likert scale rating 5 steps has been used for questions 21-84. 84 items are included to ascertain the affect of these variables on students' GPA. Some of the questions used in the questionnaire are part of other studies also (Appendix 4&5).Items representing a particular factor are explained in (Appendix 2)

Weka is selected as a data mining tool as not only is it an open source software but also because it is used by many researchers. Weka (Waikato Environment for Knowledge Analysis) is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand. Weka, is a free software available under the GNU General Public License and it supports several standard data mining tasks, more specifically, data preprocessing, clustering, classification, regression, visualization, and feature selection.

Students' GPA data collected from the Institute was in continuous form which has to be transformed in to discrete categories. Therefore, students' GPA was categorized in five groups as shown in the Table 1.

Table 1: ***Categories of GPA***

GPA	GROUP
3.78-4	Very Good
3.33-3.77	Good
2.67-3.32	Satisfactory
1.5-2.66	Below Satisfactory
0-1.49	Fail

After categorizing the data, I checked for missing data, fortunately there was none. First generation student was one of the variables of factor: however family background was not considered in data analysis as majority of the students have misunderstood the question resulting in incorrect answers. Since dataset consists of 87 items, therefore, data was further analyzed by selecting relevant items defining a particular factor by using Weka's feature of selecting attributes. Weka is comprised of many attribute selection selectors. InfoGain Attribute Evaluator was selected to distinguish relevant attributes by ranking them according to their significance in determining the dependent variable as well as to reduce the size of prediction.

The information gained, with respect to a set of examples is the expected reduction in entropy that results from splitting a set of examples using the values of that attribute. This measure is used for identifying those attributes that have the greatest influence on classification. Thai Nghe et al (2007)

Before the data is subjected to attribute selector, unsupervised discretization on numerical attributes is applied as some of the classifiers like tree and rule learners work well with discretized attributes.

Best five variables which have the highest significance in determining a particular factor were selected and the rest discarded. This practice was applied for all factors. All together 8 factors were used in the research for determining their significance in predicting students' performance.

Weka provides various algorithms grouped in different classifying methods. On the basis of studies reviewed, most commonly used and good predictor algorithm of Weka classifiers (with their default settings) were selected as shown below in Table 2. The aim is to compare these algorithms in predicting students' performance.

Table 2 : Data mining techniques used in the study

Decision Trees:	J48(C4.5),RandomForest,BFTree, RepTree
Functions:	Logistic, RBFNetwork
Rule:	JRip
Bayesian Network:	BayesNet, NaiveBayes

J48 decision tree and Random Forest are used in the study by Vandamme et al (2007). Thai Nghe et al (2007), Al-Radaideh et al (2006), and Othman et al (2007) used J48 decision tree and BayesNetwork classifiers. Affendy et al (2010) also used j48, BayesNetwork along with Naïve BayesNetwork, RepTree, BFTree, RBF Network and Logistic function. Gerben W.Dekker(2009) used rule learning JRip along with J48 decision tree and Random Forest, Bayes Network along with Naïve Bayes Network, and Logistic function.

All the classifiers are run on the dataset using 10-fold cross validation for estimating generalization performance. It was used to evaluate the prediction accuracy.

Cross-validation divides the data in to fixed number of folders. For example, if a data is divided into *n* folders $f_1,f_2,\ldots,f_n$, then cross validation method asks f_1 to train($f_2,f_3,\ldots, f_n$). In second stage f_2 train (f_3, $f_{4,}$ …, f_n) and so on until the folders are trained to prepare the certainty.

The classification accuracy of each classifier is summarized in Table 3 to compare their relative performance.

Table 3: Simulation result of each algorithm

	Original Data				**Resample Data**		
Predicted GPA Classes	**Algorithm (Total Instances,231)**	**Correctly Classified Instances %(value)**	**Incorrectly Classified Instances %(value)**	**Kappa Statistics**	**Correctly Classified Instances %(value)**	**Incorrectly Classified Instances %(value)**	**Kappa Statistics**
5 classes:(Very Good, Good,Satisfactory,Below Satisfactory,Fail)	J48(C4.5)	49.3506 (114)	50.6494 (117)	0.0467	84.632 (391)	15.368 (71)	0.7429
	RandomForest	47.619 (110)	52.381 (121)	0.0734	92.4242 (427)	7.5758 (35)	0.8742
	BFTree	47.619 (110)	52.381 (121)	0.0203	79.4372 (367)	20.5628 (95)	0.6514
	RepTree	49.7835 (115)	50.2165 (116)	0.0623	74.4589 (344)	25.5411 (118)	0.5633
	JRip	46.3203 (107)	53.6797 (124)	0	74.8918 (346)	25.1082 (116)	0.5704
	RBFNetwork	45.8874 (106)	54.1126 (125)	0.0667	66.0173 (305)	33.9827 (157)	0.4215
	Logistic	38.0952 (88)	61.9048 (143)	0.0405	66.4502 (307)	33.5498 (155)	0.4443
	BayesNet	51.0823 (118)	48.9177 (113)	0.1307	58.658 (271)	41.342 (191)	0.2857
	NaïveBayes	49.3506 (114)	50.6494 (117)	0.0879	58.0087 (268)	41.9913 (194)	0.2658

IV. Results:

From the obtained results, it was found that the classification accuracy for all the classification algorithms is not high. This indicated that the data set was small and not sufficient to generate a classification model of high quality in terms of accuracy. On inspection of confusion matrices from tables 4-12(original data column), it was found that accuracy for smaller classes was much smaller as compared to larger classes in every classifier output. For example, in Table 4: classifier J48 decision tree, there was no prediction of the number of students failing but 84% students have been predicted as satisfactory performers. Even the class of "Very Good" was not considered in confusion matrices of every classifier, the reason being that those students who participated in the survey, their first year CGPA did not fall in this group. Therefore, confusion matrices for all classifiers only represented four classes.

To resolve the problem of imbalances in the dataset, data was resampled by using Weka resample function. This function oversamples the minority class and under samples the majority class in order to create a more balanced distribution for training the algorithms. By looking at the table, it is clearly observed that predictions using the re-sampled data set are significantly more accurate for all classifiers.

On comparing classifiers, Table 3 (Resmapled Data column), RandomForest turned out to be the most effective classifier in predicting students' performances with accuracy 92% followed by J48 decision tree 84%, BFTree 79%, RepTree 74% and JRip rule 74%. The least effective is the BayesNet and NaïveBayes with 59%and 58% accuracy respectively. Even though before resampling, BayesNet had the highest accuracy rate of 51% in predicting students' performance and and NaiveBayes was same as J48 tree 49%. This shows that resampling does not affect these algorithms' ability to predict as much as it affects the others algorithms on the dataset.

Table 4: J48 Tree

Actual Class	Original Data				Re-Sampled Data			
	Satisfactory	Below Satisfactory	Good	Fail	Satisfactory	Below Satisfactory	Good	Fail
Satisfactory	**97**	16	2	0	**194**	22	1	1
Below Satisfactory	65	**16**	0	0	26	**156**	3	0
Good	23	3	**1**	0	7	7	**30**	2
Fail	5	3	0	**0**	1	1	0	**11**
%Hit	84%	20%	4%	0%	89%	84%	65%	85%

Table 5: RandomForest

Actual Class	Original Data				Re-Sampled Data			
	Satisfactory	Below Satisfactory	Good	Fail	Satisfactory	Below Satisfactory	Good	Fail
Satisfactory	**76**	36	3	0	**203**	13	1	1
Below Satisfactory	46	**33**	2	0	8	**176**	1	0
Good	18	8	**1**	0	3	6	**37**	0
Fail	3	4	1	**0**	0	1	1	**11**
%Hit	66%	41%	4%	0%	93%	95%	80%	85%

RandomForest's correctly predicted students in classes "Good" and "Fail" as 80% and 85% (Table 5) respectively on resampled data. It showed that resampling of data have a significant impact on predicting capability of a classifier. Before that, the classifier predicted poorly for these classes. Even the predictions for "Satisfactory" and "Below Satisfactory" classes have also improved considerably.

RepTree correct prediction for "Satisfactory" class (Table 6) decreases by 1% on resampled data as compared to the prediction on original data but its overall performance is much better.

The accuracy in the predictions of BFTree, JRIP rule and RBF function, performances increases as shown in tables 7,8 and 9. Logistics Function (Table 9) prediction performance increases more for "Good" and "Fail" classes as compared to "Satisfactory" and "Below Satisfactory" classes.

Table 6: RepTree

Actual Class	Original Data				Re-Sampled Data			
	Satisfactory	Below Satisfactory	Good	Fail	Satisfactory	Below Satisfactory	Good	Fail
Satisfactory	**95**	18	2	0	**179**	33	5	1
Below Satisfactory	61	**19**	1	0	41	**142**	2	0
Good	22	4	**1**	0	14	13	**19**	0
Fail	7	1	0	**0**	5	2	2	**4**
%Hit	83%	23%	4%	0%	82%	77%	41%	31%

Internal Factors

Table 7: BF Tree

Actual Class	Original Data				Re-Sampled Data			
	Satisfactory	Below Satisfactory	Good	Fail	Satisfactory	Below Satisfactory	Good	Fail
Satisfactory	**90**	25	0	0	**184**	28	5	1
Below Satisfactory	61	**20**	0	0	34	**150**	1	0
Good	24	3	**0**	0	10	5	**31**	0
Fail	5	3	0	**0**	6	5	0	**2**
%Hit	78%	25%	0%	0%	84%	81%	67%	15%

Table 8: JRip

Actual Class	Original Data				Re-Sampled Data			
	Satisfactory	Below Satisfactory	Good	Fail	Satisfactory	Below Satisfactory	Good	Fail
Satisfactory	**89**	24	1	1	**178**	36	1	3
Below Satisfactory	62	**18**	1	0	42	**141**	2	0
Good	25	2	**0**	0	20	3	**23**	0
Fail	6	2	0	**0**	8	0	1	**4**
%Hit	77%	22%	0%	0%	82%	76%	50%	31%

Table 9: Logistic Function

Actual Class	Original Data				Re-Sampled Data			
	Satisfactory	Below Satisfactory	Good	Fail	Satisfactory	Below Satisfactory	Good	Fail
Satisfactory	**46**	42	16	11	**151**	57	7	3
Below Satisfactory	26	**39**	8	8	71	**112**	1	1
Good	17	6	**3**	1	6	7	**33**	0
Fail	5	2	1	**0**	0	1	1	**11**
%Hit	40%	48%	7%	0%	69%	60%	72%	85%

Table 10: RBFNetwork

Actual Class	Original Data				Re-Sampled Data			
	Satisfactory	Below Satisfactory	Good	Fail	Satisfactory	Below Satisfactory	Good	Fail
Satisfactory	**74**	30	10	1	**159**	53	4	2
Below Satisfactory	46	**30**	5	0	60	**121**	4	0
Good	17	7	**2**	1	13	13	**20**	0
Fail	5	2	1	**0**	5	2	1	**5**
%Hit	64%	37%	7%	0%	73%	65%	43%	45%

Table 12: BayesNet

Actual Class	Original Data				Re-Sampled Data			
	Satisfactory	Below Satisfactory	Good	Fail	Satisfactory	Below Satisfactory	Good	Fail
Satisfactory	**80**	32	2	1	**152**	61	2	3
Below Satisfactory	42	**37**	2	0	72	**106**	6	1
Good	18	8	**1**	0	28	9	**9**	0
Fail	6	2	0	**0**	4	3	2	**4**
%Hit	70%	46%	7%	0%	70%	57%	20%	31%

Kappa Statistics which is used to compare different classifiers predicting performance is quiet low for every classifier while predicting on the original dataset (Table 3). In fact, for JRip, it was zero. But after resampling, Kappa Statistics for all classifiers have improved significantly. RandomForest has the highest 0.8742 followed by J48 decision tree 0.7429. Bayesian networks have the lowest, again, confirming the fact that resampling does not affect the predicting capabilities of Bayesian networks on the data set.

Kappa statistics is used to assess the accuracy of any particular measuring cases, it is usual to distinguish between the reliability of the data collected and their validity.

Factors that are strong and significant predictors of the students' performances are academic integration, family background and social integration.

Academic integration i.e. students' involvement in their studies by spending more hours on study and its related issues, students contentment with their choice of study program as well as the pressure of workload that they can cope with easily, make their interaction with studies more constructive.

Family background with mother's education, father's income and occupation along with the size of family ensure students' contentment at home, which reflects in their achievements.

At the academia, interaction with faculty and peers, involvement in sports and other extracurricular activities help students in their performances. In fact, students' motivation within one self as well as friendly surroundings boosts their inspirations to succeed. This all, support students to work hard and achieve their goals.

Individual characteristics and extrinsic motivation have the least effect on student's performance which shows that these are not the significant predictor.

IV Conclusion:

This study shows that datamining techniques applied can be used by universities and higher education in determining best students so that resources available can be affectively be utilized in helping and guiding these students to achieve success, especially in selecting students for scholarships and other means of financial assistance. This will also help these academic bodies and board of studies in developing meaningful programs that motivates and encourage those students who got the potential to excel but need assistance to progress. Academia can also develop programs that build up close relationship between teachers and the students as shown in the study that interaction between these two factors help the latter one immensely in their success.

In case of identifying best students (good) and those (Satisfactory), who with little help can fall in good students' class, highest prediction accuracy is that of Random Forest at 80% and 93% respectively. Over all, Decision tree classifiers predict these students more accurately as compared to other classifiers. To identify those students who are not performing well and those who are at the risk of failing (below satisfactory) and eventually be droppedout (fail), again RandomForest's accuracy is the highest with accuracy of 95% and 85% respectively. Looking at the performance of all the classifiers on the dataset of students, it is evident that Decision tree classifiers are better, in terms of accuracy, in predicting students' academic performance.

Acknowledgments:

I thank Rizwan Ahmed (Senior Lecturer, Institute of Business Management) for his time, valuable input and assistance in doing this research.

References:

Afzal, H.,Imran,A., Khan,M.A. and Kashif,H. (2010) A study of university students' motivation and its relationship with their academic performance, International Journal of Business and Management,, Vol. 5, 4.

Affendey, L.S., Paris,I.H.M., Mustapha,N., Sulaiman, Md.N., and Muda,Z.(2010) Ranking of Influencing Factors in Predicting Students' Academic Performance.Information Technology Journal Vol. 9, 4.

Al-Radaideh, Q. A., Al-Shawakfa,E.M. and Al-Najjar, M.I(2006) Mining Student Using Decision Trees, The International Arab Conference on Information Technology.

Bekele, R. and Menzel,W.(2005) A Bayesian Approach to Predict Performance of A Student(BAPPS): A Case with Ethiopian Students.Artificial Intelligence and Applications: IASTED International Conference Proceedings.

Bellaachia, A.,Guven, E.(2005). Predicting Breast Cancer Survivability Using Data Mining Techniques. Department of Computer Science The George Washington University.

Bouckaert, Remco R.(2004) Bayesian Network Classifiers in Weka, 1st September

Dekkar,W.G., Pechenizky,M. and Vleeshouwers, M.J.(2009) Predicting Students Drop Out: A case study. Submitted to the 2nd Int. Conf. on Educational Data Mining (EDM '09).

Hall, M. and Holmes,G.(2002) Benchmarking Attribute Selection Techniques for Discrete Class Data Mining. Working paper Series, 18 April.

Herzog, S. (2006) Estimating student retention and degree-completion time: Decision trees and neural networks vis—vis regression. New Directions for Institutional Research, Vol.2006,Issue 131

Ishitani, T.T. 2003. A Longitudinal Approach To Assessing Attrition Behavior Among First-Generation Students: Time-Varying Effects Of Pre-College Characteristics.Research In Higher Education. 2003, Vol. 44, No.4,pp.433-449.

Jadric, M., Garaca,Z. and Cukusic,M.(2010) Student Dropout Analysis With Application Of Data Mining Methods.Management,Vol.15, pp.31-46

Karegowda, A.G., Manjunaath,A.S. and Jayaram, M.A.(2010) Comparative study of attribute selection using Gain Ratio and Correlation Base Feature Selection, Internationl Journal of Information Technology And Knowledge Management, July-Dec,Vol. 2, pp. 271-277.

Kirkby,R., Frank, E. and Reutemann, P. (2007)WEKA Explorer User Guide for Version 3-5-5,26 January.

Kokkodis,M. and Akabay,M. CS 235 Project Report: UCSD Data Minning ContestMannan, MD.A.(2007) "Student Attrition And Academic And Social Integration: Application of Tinto's Model at the University of Papua New Guinea.", Higher Education, Vol 53, pp. 147-165

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

Ogor, E.N.(2007) Student academic performance monitoring and evaluation using datamining techniques. Electronics, Robotics and Automotive Mechanics Conference, CERMA 2007, pp. 354–359

Othman,M.F. and Moh Shan Yau, T.(2007) Comparison of Different Classification Techniques Using WEKA for Breast Cancer. IFMBE in university drop out.

Rubenking, N. "Hidden Messages." *PC Magazine*, May 22, 2001, *20*(10), 86–88

Sahay, A., Mehta, K. and QMS,LLP(2010) Assisting Higher Education in Assessing, Predicting, and Managing Issues Related to Student success: A web-based software using data ,mining and quality function deployment, Academic and

Business Research Institute Conference, las Vegas Vandamme, J.P. and Superby,J.F(2007) "Predicting Academic Performance by Data Mining Methods."Education Economics, Vol.15 No(4), pp.405–419.

Veitch, W.R. (2004) "Identifying characteristics of high school dropouts: Data mining with a decision tree model", Presented at the Annual Meeting of the American Educational research Association held on April at San Diego, CA.

Kappa at http://www.dmi.columbia.edu/homepages/chuangi/Kappa

Gartner Group. "The GartnerGroup CRM Glossary." [http://www.gartnerweb.com/public/static/hotc/hc00086148.html].

Appendix Table 1: Factors of the Study

Factors that are part of the model are given in the table below along with the variables that determine these factors.

Factors	Variables of the study	Variables used in other studies
Family background	Father's education	Nattavudh, A.Vignoles(2009), Ishitani (2003), Vandamme et al (2007), Mario Jadric et al (2010), Gerben W.Dekker(2009)
	Mother's education	Nattavudh, A.Vignoles(2009), Ishitani (2003), Mario Jadric et al (2010), Vandamme et al (2007), Gerben W.Dekker(2009)
	Father's occupation	Nattavudh, A.Vignoles(2009), Thai Nghe et al (2007), Vandamme et al (2007), Ishitani (2003), Mario Jadric et al (2010), William R. Veitch (2004)
	Family income	Nattavudh, A.Vignoles(2009), Ishitani (2003), Herzog (2006)
	First generation	Ishitani (2003), Mario Jadric et al (2010)
	Parental martial stability	Vandamme et al (2007)
	Family size	Vandamme et al (2007)
	Availability of computer and net at home	None as added as a new variable to test
social integration	Quality and quantity of student's relationships with peers	Abdul Mannan(2007)
	Interaction with faculty	Abdul Mannan(2007)
	Extracurricular involvements	Abdul Mannan(2007)
Academic Integration	GPA	Abdul Mannan(2007), William R. Veitch (2004)
	Hours spend on academic or extra-curricular activities	Vandamme et al (2007)
	Entrance test score	Thai Nghe et al (2007), Mario Jadric et al (2010)
	Enrolled in preferred course	Thai Nghe et al (2007), Mario Jadric et al (2010), Gerben W.Dekker(2009), Herzog(2006)
	Field of study	Thai Nghe et al (2007), Mario Jadric et al (2010), Gerben W.Dekker(2009), Herzog(2006)
	Missed classes	Vandamme et al (2007)
Individual Characteristics	Gender	Ishitani (2003), Thai Nghe et al (2007), William R. Veitch (2004), Vandamme et al (2007), Mario Jadric et al (2010), Herzog(2006), Nattavudh, A.Vignoles(2009),
	age	Vandamme et al(2007),Gerben W.Dekker(2009),Herzog(2006),Thai Nghe et al (2007),William R. Veitch (2004)
Career expectations and Goal Commitment	Educational expectations or career expectations	Herzog(2006)
Satisfaction with Institutional Characteristics	Student's satisfaction with the college experience i.e Resources, Facilities (computer,net,library) and Structural arrangement	Herzog(2006)
Individual Motivation	**Intrinsic** Self-exploration, Altruism	Imran Ali et al, 2010
	Extrinsic Rejection of Alternative Options, Career and Qualifications, Social Enjoyment, Social Pressure	

Appendix 2:

Theory support that relates the questionnaire to questionnaires of reviewed studies.

Age

Question 3 is framed to find out the age of a student. Impact of age on student's performance was tested by Vandamme et al(2007),Gerben W.Dekker(2009),Herzog(2006),Thai Nghe et al (2007),William R. Veitch (2004).

Gender

Question 4 is asked to find whether student's persistence to continue education differ with gender. This factor was included in the studies of Nattavudh, A.Vignoles(2009), William R. Veitch (2004), Terry T. Ishitani (2003), Vandamme et al(2007), Thai Nghe et al (2007), Herzog(2006)

Family back ground

This dimension is added as it has been investigated in many studies. Questions 10-16 are related to this dimension.

Parents' education

Question 5and 6 are included to gather information about the qualification of father and mother to see whether their education has an impact on students' performance. Terry T. Ishitani (2003), Gerben W.Dekker(2009), Vandamme et al(2007), Mario Jadric et al (2010) and Nattavudh, A.Vignoles(2009) all included questions in their questionnaire related to parents education.

First generation

Question 7 is included because Terry T. Ishitani (2003)and Mario Jadric et al (2010) used this measure to find its impact on students' persistence to perform well.

Father's occupation and Family Income

Question 8 and 9 are included as Nattavudh, A.Vignoles(2009) included this aspect along with the annual income of the family in his studies. Annual income of the family is used as a measure by Terry T. Ishitani (2003), Thai Nghe et al (2007), Vandamme et al (2007)Mario Jadric et al (2010) and William R. Veitch (2004) in their studies.

Parental marital status

Question 10 is asked to see the impact of parents' relation on students' performance. It was investigated by Vandamme et al (2007) in his study.

Family size

Question 11 is asked to explore this factor influence on student's performance. Vandamme et al (2007) used these measures in their studies.

Academic Integration

This dimension is included in almost all the studies reviewed.

Field of study and preferred program

Question 12 and 13 are included to find out the impact of choice of study and enrollment in an unwanted study program on a student's performance. Gerben W.Dekker(2009),Herzog(2006),Thai Nghe et al (2007), Mario Jadric et al (2010) included this factor in their studies.

Missed classes and hours spend on academic activites outside college

Question 14 is added to find out how much a student try hard to acquire knowledge. Vandamme et al(2007) in their studies used this factor to investigate its level of impact on performance.

Question 15 is added to explore the difference in student's performance who spend less or more time on academic activities.

Computer and Internet

Since computer and net has become so much part of a student's life , therefore, Question 16and 17 are added to see the impact of facilities of computer and net on the academic progress of a student.

Work load

Questions 46-49 are included to see the impact of workload of assignments and quizzes on student's performance.

Social Integration

This dimension is generally used as a measure to predict students' performance.

Interaction with faculty

Questions 21-30 are asked to find the importance and influence of teaching method as well as teacher student relation on students' performance. These questions have been adopted by MD.Abdul mannan(2007) in his questionnaire.

Extracurricular activities

Questions 31-35 are framed to find out how well a student is integrated in his social life. These questions were adopted by MD.Abdul mannan(2007) and Mike taffe in their questionnaires . Influence of extracurricular activities on students' performance was also tested by Vandamme et al(2007) and Terry T. Ishitani (2003).

Peer group interaction

Questions 36 -40 are included to find whether students relation and interaction with his peers has an impact or not on his academic performance. Abdul mannan(2007) did include questions related to this measure in his questionnaire.

Goal commitment and career expectations

Questions 41-43 are framed to find whether better career prospects motivates student to work hard in completing their degree. Herzog (2006) included this factor to estimate student retention.
Question 44 and 45 are formulated to find whether student's determination let a student to persist and aspire for degree completion.

Student satisfaction with his selected institute

Questions 50-64 are included to test whether students satisfaction with the facilities and services provided by the institute and institute's atmosphere have an impact on students performance. This factor was again included in the study carried out by Herzog (2006) in estimating student retention.

Motivation

This dimension is included to explore the influence of student's own motivation on his performance.
Questions 65- 84 will investigate the impact of this factor. Imran Ali et al(2010) in their study tried to find out to what degree this measure has an influence on students academic progress.

Current GPA

This factor is included and will be collected from the institute. Gerben W.Dekker(2009), Thai Nghe et al (2007),William R. Veitch (2004) and Herzog(2006) included this factor to find student retention and dropout rates.

Appendix 3: **QUESSTIONNAIRE**

Predicting factors that affect the student's academic performance by using data mining techniques

1. Name : ___________________
2. Student ID: ___________________

Individual Characteristics:

3. Age: a. Below 18 b. Between 18 and 20 c. Between 21 and 25

4. Gender: a. Male b. Female

Family Background:

5. Father's level of education
   a. Middle school b. Intermediate c. Graduate d. Masters e. Doctorate
6. Mother's level of education
   a. Middle school b. Intermediate c. Graduate d. Masters e. Doctorate
7. Are you the first one in your family to attend bachelors level institute a. YES b. NO
8. Father's occupation
   a. Business b. services c. Government officer d. others
9. Annual Family Income
   a. <2,40,000 b. 2,40,000- 4,80,000 c. 4,80,00 0- 7,20,000 d. 7,20,000-9,60,000 e. Above 9,60,000
10. Parental marital status
    a. Married b. Divorce c. Widower
11. How many members in your family?
    a. 3 b. 4 c. 5 d. 6 e. Others__________

Academic Integration:

12. Field of study:
    a. Computer science b. Finance
    c. Marketing d. Economics
    e. Human Resource Management f. Information Technology

13. Are you enrolled in your preference program a. YES b. NO
14. How often do u miss your classes? a. Daily b. Once a week c. Twice a week
15. Hours spend on academic activities
    a. Less than 2 hr b. Between 2 to 4 hrs c. More than 4 hrs
16. Do you have access to computer at home? a. YES b. NO
17. Do you have access to internet at home? a. YES b. NO

Social Pressure (Motivation)

18. Your family expects you to graduate? a. Yes. b. No
19. Do you try to live of others expectations a. Yes. b. No
20. Is it expected of you to enroll for an advance degree when, or if, you complete your graduate degree?
    a. Yes. b. No

Social Integration:

Select the most appropriate one (Strongly disagree, Disagree, Neutral ,Agree, Strongly Agree)

Interaction with faculty

21. Faculty are genuinely concerned in my academic work
22. Faculty are genuinely interested in teaching
23. Faculty are interested in alleviate my academic weakness
24. Faculty are always available for obtaining information
25. Faculty are accessible to discuss matters of intellect
26. Faculty are accessible to discuss career goals
27. Faculty feedback make you work harder
28. Faculty have positive influence on personal growth
29. Faculty promote good relationship
30. I have regular contact with teachers outside of class

Extracurricular activities

31. I participate in clubs and organizations
32. I participate in sports and cultural events
33. I attended a meeting of a club, organization
34. Interpersonal relationship enhance personal growth
35. Interpersonal relationship expand intellectual growth

Peer group interactions

36. Students help in personal problems
37. I get on well with other students
38. I sat around in the student center talking with other students
39. I prefer to study with other students i.e. in group
40. Studies bodies and groups promote friendship

Career Expectations And Goal Commitment:

41. The job prospects for the major are promising
42. The major has well paid jobs
43. There is value of the university education I am receiving
44. I give up easily on difficult projects/assignments.
45. I have goals in life that I try to achieve

Academic Integration:

Work load

46. The workload on my degree courses is manageable and do not apply unnecessary pressure.
47. Degree courses do not apply unnecessary pressure on me as a student.
48. The volume of work on my courses means I can always complete it to my satisfaction.
49. I am generally given enough time to understand the things I have to learn

Students’ Satisfaction:

Campus:

50. Transport facilities provided are sufficient
51. Facilities (stationary, photocopier, printing) provided by tuck-shop is sufficient.
52. Friendly atmosphere and pleasant learning environment
53. Clean and nice campus grounds
54. Class room size and number of students in a class is satisfactory.
55. Class environment as a whole interest you in studying

Library:

56. The library is a good, quiet place for studying and the study rooms are great for working on group assignments.
57. Vast array of books in library, and usually have available what I’m looking for

Computer and net facilities:

58. Number of computers available in the library is sufficient.
59. Internet facility is good and is always available at the institute

Staff:

60. Friendly and approachable staff

Security:

61. You feel secure and safe at the campus.

Institutional commitment

62. I feel a sense of pride about my campus.
63. I am able to experience intellectual growth here.
64. There is a commitment to academic excellence on this campus.

Motivation

I attend university…

65. because I don’t know what else to do.
66. to understand myself better.
67. to gain valuable skills for my career.
68. because its fun place to be.
69. because I genuinely want to help others.
70. because it’s a better alternative than working.
71. because I want to explore new ideas.
72. because I enjoy the social life.
73. because other people have told me I should.
74. because I want to contribute to society.
75. to avoid being unemployed.
76. because I want to challenge myself.
77. because it gives me something to do.
78. because it will help set up my future career.
79. because of the social opportunities.
80. because I want to improve the world situation.
81. because I don’t have any better options.
82. because I love learning.
83. so I can get a better job.
84. because its a great place to develop friendships.

Appendix 4: Similar Questions used in survey of study by Mannan,A.(2007)

(A) Academic staff concern for students development and teaching

1. Attended departmental meetings
2. Genuinely concerned in my acad. work
3. Willing to spend time outside class
4. Genuinely interested in teaching
5. Interested in alleviate my academic weakness

(B) Informal contact with academic staff on academic matters

6. Always available for obtaining information
7. Accessible to discuss matters of Intellect
8. Positive influence on personal growth
9. Accessible to discuss career goals
10. I am satisfied with opportunities

(C) Informal social contact with academic staff

11. Accessible to discuss campus etc issue
12. Interested for socialization
13. Accessible to solve personal problems
14. Involve promoting good relationship

(D) Extracurricular activities

15. Participation in clubs and organizations
16. Participate in sports and cultural events
17. Participate in public lecturers seminars

(D) Peer group interactions

18. Student bodies and groups promote friendship
19. Interpersonal relationship for personal growth
20. Interpersonal relationship for intellectual growth

21. Students helped in personal problems
22. Alcohol consumption helpful for socialization
23. Regional groups successful in socialization
23. Satisfied with socialization with peer groups

Appendix 5: Similar Questions used in survey of study by Afzal, H.,Imran,A., Khan,M.A. and Kashif,H. (2010)

Closed Items and Codes
The 30 items in The University Student Motivation and Satisfaction Questionnaire Version 2 (TUSMSQ version
2) are displayed in the table, along with the item numbers, variable codes (for use in data analysis), the target
factor, and the response scale.

No.	**Factor**	**I attend university…**
1	RA	because I don't know what else to do.
2	SE	to understand myself better.
3	QC	to gain valuable skills for my career.
4	SO	because its fun place to be.
5	SP	because others expect me to get a degree.
6	AL	because I genuinely want to help others.
7	RA	because it's a better alternative than working.
8	SE	because I want to explore new ideas.
9	QC	to enhance my job prospects.
10	SO	because I enjoy the social life.
11	SP	because other people have told me I should.
12	AL	because I want to contribute to society.
13	RA	to avoid being unemployed.
14	SE	because I want to challenge myself.
15	QC	in order to get the qualification.
16	SO	because I enjoy the social environment.
17	SP	because it would disappoint other people if I didn't.
18	AL	because I want to help solve society's problems.

19	RA	because it gives me something to do.
20	SE	for my personal growth and development.
21	QC	because it will help set up my future career.
22	SO	because of the social opportunities.
23	SP	it seems to be the recommended thing to do.
24	AL	because I want to improve the world situation.
25	RA	because I don’t have any better options.
26	SE	because I love learning.
27	QC	so I can get a better job.
28	SO	because its a great place to develop friendships.
29	SP	of social expectations from those around me.
30	AL	because I want to be more useful to society.

International Journal of Business and Management Vol. 5, No. 4; April 2010

The following items contain the key to sources of motivation mentioned in the survey:-

- Rejection of Alternative options (Extrinsic) 1, 7,13,19,25
- Self-exploration (Intrinsic) 2, 8,14,20,26
- Career and Qualifications (Extrinsic) 3, 9,15,21,27
- Social enjoyment (Extrinsic) 4, 10,16,22,28
- Social Pressure (Extrinsic) 5, 11, 17, 23, 29
- Altruism (Intrinsic) 6, 12,18,24,30

Table 1. Sources of Motivation

Intrinsic Self-exploration	Student with full motivation and is actually interested in learning and exploring ideas for its own sake. • Believed to do well and show good result. • e.g. 'I attend university because I have a genuine interest in the subject I am studying'.
Altruism	Wants to learn for own satisfaction, becoming useful to the society, helping others and solving their problems. • Believe to be genuinely motivated and show better results. • e.g. 'I attend university because I want to be more useful to society.'
Extrinsic Rejection of Alternative Options	Concerned with their careers and more inclined towards completing the degree only to get good jobs in future. • Continuing studies just to avoid working or because does not know what else to do. • Believed to be de-motivated and can not keep a consistent performance academically. • e.g. 'I attend university because I don't know what else I would do'.
Career and Qualifications	Concerned with their careers and inclined towards completing the degree only to get good jobs in future. • Concerned with getting the degree but not for learning sake. • Believe to show good performance but not for long-term or where the reward does not exist anymore. • e.g. 'I attend university to enhance my job prospects
Social Enjoyment	View University as a socializing place where they can have fun and make friends. • De-motivated students who are thought to show very little academic performance. • e.g. 'I attend university because I enjoy the social life'.
Social Pressure	Surrounded by social pressures, peers, parents, etc. • Try to live up to other's expectations. • Do not have genuine interest in studies with no consistency. • e.g. 'I attend university because others expect me to get a degree