Global Journal of Computer Science and Technology
Software & Data Engineering
Volume 12 Issue 15 Version 1.0 Year 2012
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Artificial System for Prediction of Student's Academic Success from Tertiary Level in Bangladesh

By Linkon Chowdhury & Shahana Yeasmin

BGC Trust University Bangladesh

Abstract - Every year a large scale of students in Bangladesh enrol in different Universities in order to pursue higher studies. With the aim to build up a prosperous career these students begin their academic phase at the University with great expectation and enthusiasm. However among all these enthusiastic and hopeful bright students many seem to become successful in their academic career and found to pursue the higher education beyond the undergraduate level. The main purpose of this research is to develop a dynamic academic success prediction model for universities, institutes and colleges. In this work, we first apply chi square test to separate factors such as gender, financial condition and dropping year to classify the successful from unsuccessful students. The main purpose of applying it is feature selection to data. Degree of freedom is used to P-value (Probability value) for best predicators of dependent variable. Then we have classify the data using the latest data mining technique Support Vector Machines(SVM).SVM helped the data set to be properly design and manipulated. After being processed data, we used the MATH LAB for depiction of resultant data into figure. After being separation of factors we have had examined by using data mining techniques Classification and Regression Tree (CART) and Bayes theorem using knowledge base. Proposition logic is used for designing knowledge base. Bayes theorem will perform the prediction by collecting the information from knowledge Base. Here we have considered most important factors to classify the successful students over unsuccessful students are gender, financial condition and dropping year. We also consider the sociodemographic variables such as age, gender, ethnicity, education, work status, and disability and study environment that may inflounce persistence or academic success of students at university level. We have collected real data from Chittagong University Bangladesh from numerous students. Finally, by mining the data, the most important factors for student success and a profile of the typical successful and unsuccessful students are identified.

Keywords : Chi Value, P-value, CART, SVM, Bayes theorem, Degree of freedom.

GJCST-C Classification : I.2.0

ARTIFICIAL SYSTEM FOR PREDICTION OF STUDENTS ACADEMIC SUCCESS FROM TERTIARY LEVEL IN BANGLADESH

Strictly as per the compliance and regulations of:

Artificial System for Prediction of Student's Academic Success from Tertiary Level in Bangladesh

Linkon Chowdhury [α] & Shahana Yeasmin [σ]

Abstract - Every year a large scale of students in Bangladesh enrol in different Universities in order to pursue higher studies. With the aim to build up a prosperous career these students begin their academic phase at the University with great expectation and enthusiasm. However among all these enthusiastic and hopeful bright students many seem to become successful in their academic career and found to pursue the higher education beyond the undergraduate level. The main purpose of this research is to develop a dynamic academic success prediction model for universities, institutes and colleges. In this work, we first apply chi square test to separate factors such as gender, financial condition and dropping year to classify the successful from unsuccessful students. The main purpose of applying it is feature selection to data. Degree of freedom is used to P-value (Probability value) for best predicators of dependent variable. Then we have classify the data using the latest data mining technique Support Vector Machines(SVM).SVM helped the data set to be properly design and manipulated. After being processed data, we used the MATH LAB for depiction of resultant data into figure. After being separation of factors we have had examined by using data mining techniques Classification and Regression Tree (CART) and Bayes theorem using knowledge base. Proposition logic is used for designing knowledge base. Bayes theorem will perform the prediction by collecting the information from knowledge Base. Here we have considered most important factors to classify the successful students over unsuccessful students are gender, financial condition and dropping year. We also consider the sociodemographic variables such as age, gender, ethnicity, education, work status, and disability and study environment that may influence persistence or academic success of students at university level. We have collected real data from Chittagong University Bangladesh from numerous students. Finally, by mining the data, the most important factors for student success and a profile of the typical successful and unsuccessful students are identified.

Keywords : Chi Value, P-value, CART, SVM, Bayes theorem, Degree of freedom.

I. Introduction

Student retention is an indicator of academic performance and enrolment management of the university. Increasing student retention or persistence is a long term goal in all academic institution. High rates of student attrition have been a concern for the past several years at the Chittagong University Institute of Computer Science and Engineering. Level of retention rate goes higher by various reasons and one the most important reasons are government funding, scholarship in the tertiary education environment. There are both academic and administrative staff are under pressure to come up with strategies that could increase retention rates on their courses and programs. The lowest student retention rates at all institutions of higher education are first-year students, who are at greatest risk of dropping out in the first term or semester of study or not completing their programs or degree. Therefore most retention studies address the retention of first-year students. Consequently, the early identification of vulnerable students who are prone to drop their courses is crucial for the success of any retention strategy. This would allow educational institutions to undertake timely and pro-active measures. Once identified, these at risk students can be then targeted with academic and administrative support to increase their chance of staying on the course.

Data from 2004 to 2010, covering over 200 enrolled students stored in the Computer Science and Engineering was used to perform a quantitative analysis of study outcome. There were three main types of objectives conducted in this study. The first approach is descriptive which is concerned with the nature of the dataset such as the frequency table and the relationship between the attributes obtained using cross tabulation analysis (contingency tables). In addition, feature selection is conducted to determine the importance of the prediction variables for modelling study outcome. The third type of data mining approach, i.e. predictive data mining is conducted by using two different types of classification trees. The classification tree models have some advantages. Secondly, the classification tree models are non-parametric and can capture nonlinear relationships and complex interactions between predictors and dependent variable. We decided not to use other data mining techniques such as neural networks and support vector machines even though in some cases they could achieve higher accuracy, because their structure is not transparent and usually described as a black box. It is also difficult to explain their results and how they work to a user who would like

Author α : Chittagong, Bangladesh.
E-mail : linkon_cse_cu@yahoo.com
Author σ : Chittagong, Bangladesh. E-mail : bubly.csecu@gmail.com

to apply them to a new set of data. Finally, a comparison between these models was conducted to determine the best model for the dataset. The population of this study was entering CSE in the first - year from 2004 through 2009. These were full-time degree-seeking students. Entering in the first-year session were excluded from the study for a few reasons. The independent or predictor variables fell into three main categories. These were the students' personal information, high school background. The personal information recorded for each student was:

- Ethnicity(Bengali, Marma, Chakma and Others)
- Gender
- Age
- Financial Status
- Work Status
- Disability

II. Collected Statistical Data

As part of the data-understanding phase we carried out the data on the table 1 and table 2. The Table 2 reports the results. Based on the results shown majority of Information Systems students are female (over 38%). However, percentage of female students who successfully complete the course are higher (41%) which suggests that female students are more likely to[3] pass the course than their male counterpart. When it comes to age over 26% of students are above 24. This age group is also more likely to fail the course because their percentage of students who failed the course in this age group (11.9%) is higher than their overall participation in the student population (26.2%). Statistical data on 42 students:

Table 1 : Total Outcomes

Pass	29
Fail	13

Table 2 : Descriptive statistics (percentage) – Study outcome (42 students)

Variable	Domain Name	Count	Total	Pass	Fail
Gender	Male	26	61.9	58.6	69.2
	Female	16	38.1	41.4	30.8
Age Group	>24	11	26.2	20.7	11.9
	<=24	31	73.8	79.3	61.5
Disabilities	Yes	1	2.4	0.0	7.7
	No	41	97.6	100.0	92.3
Financial Support	Yes	23	54.8	62.1	38.5
	No	19	45.2	37.9	61.5

Students with it are more likely to fail than those without it. There are huge differences in percentage of students who successfully completed [4] the course depending on their ethnic origin. A substantial number of students (over 55%) have financial support more vulnerable than the other two categories in this variable.

III. Support Vector Machines

Support Vector Machine (SVM) is one of the latest clustering techniques which enables machine learning concepts to amplify predictive accuracy in the case of axiomatically diverting data those are not fit properly. It uses inference space of linear functions in a high amplitude feature space, trained with a learning algorithm. It works by finding a hyper plane that linearly separates the training points, in a way such that each resulting subspace contains only points which are very similar. First and foremost idea behind Support Vector Machines (SVMs) is that it constituted by set of similar supervised learning. An unknown tuple is labeled with the group of the points that fall in the same subspace as the tuple. Earlier SVM was used for Natural Image processing System (NIPS) but now it becomes very popular is an active part of the machine learning research around the world. It is also being used for pattern classification and regression based applications. The foundations of Support Vector Machines (SVM) have been developed by V.Vapnik.

SVM is very effective in various data and information classification process. An expert should bear in mind two important factors for implementing SVM, these two factors or techniques are mathematical programming and kernel functions. Kernel methods leads or portrayal data into colossal amplitude margins in the anticipation that in this colossal amplitude margin the data could become more easily separated or better structured. Mathematical Programming refers the conception of the Linear programming for the best fit of Hyper plane. The word programming means to plans or make a time table for regular work. Integer Linear programming (ILP) which is the part of linear programming is very useful analytical and engineering tools to get an optimal solution .The parameters are found by solving a quadratic programming problem with linear equality and inequality constraints; rather than by solving a nonconvex, unconstrained optimization problem. The flexibility of kernel functions allows the SVM to search a wide variety of hypothesis spaces. The for-most reasons of using SVM are to select the proper Support Vectors for the data classification. The figure 1 shows a graphical view of Support Vectors selection of the process. All hypothesis space help to identify the Maximum Margin Hyper plane (MMH) which enables to classify the best and almost correct data the following figure shows the process of SVMs selection from large amount of SVMs.

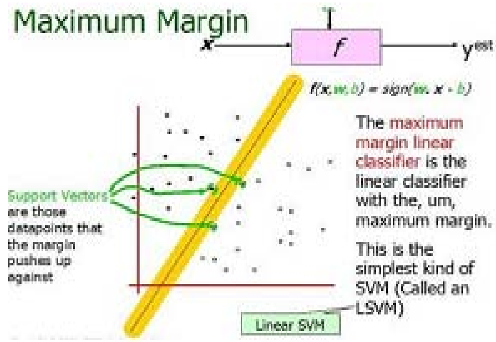

Figure 1 : Representation of Support Vectors

We can calculate the weight boundary maximum margin by using the following equation:

$$\text{margin} \equiv \underset{\mathbf{x}\in D}{\arg\min}\, d(\mathbf{x}) = \underset{\mathbf{x}\in D}{\arg\min} \frac{|\mathbf{x}\cdot\mathbf{w}+b|}{\sqrt{\sum_{i=1}^{d} w_i^2}}$$

Another interesting question is why maximum margin? There are some good explanations which include better empirical performance. Another reason is that even if we've made a small error in the location of the boundary this gives us least chance of causing a misclassification. The other advantage would be avoiding local minima and better classification. The goals of SVM are separating the data with hyper plane and extend this to non-linear boundaries using kernel trick [8] [11]. For calculating the SVM we see that the goal is to correctly classify all the data. For mathematical calculations we have,

[a] If $Y_i = +1$; $\quad wx_i + b \geq 1$

[b] If $Y_i = -1$; $\quad wx_i + b \leq 1$

[c] For all i; $\quad y_i (w_i + b) \geq 1$

In this equation x is a vector point and w is weight and is also a vector. So to separate the data [a] should always be greater than zero. Among all possible hyper planes, SVM selects the one where the distance of hyper plane is as large as possible. If the training data is good and every test vector is located in radius r from training vector. Now if the chosen hyper plane is located at the farthest possible from the data [12]. This desired hyper plane which maximizes the margin also bisects the lines between closest points on convex hull of the two datasets. Thus we have [a], [b] & [c].

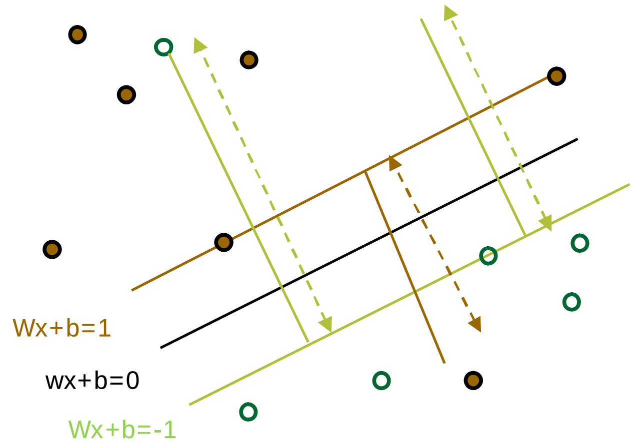

Figure 2 : Representation of Hyper planes

Distance of closest point on hyper plane to origin can be found by maximizing the x as x is on the hyper plane. Similarly for the other side points we have a similar scenario. Thus solving and subtracting the two distances we get the summed distance from the separating hyper plane to nearest points. Maximum Margin = M = 2 / ||w||

IV. Classification and Regression Tree (CART)

Classification and Regression Tree (CART) has many advantages over classification methods. It is naturally non-parametric. CART can handle [5][6] numerical data that are highly skewed. It eliminates analyst time, which would otherwise be spent determining whether variables are normally distributed. CART identifies "splitting" variables based on a complete search of all possibilities. CART is able to search all possible variable splitters, even in problems with many hundreds of possible predictors as well as dealing missing variable. Finally, CART trees are moderately simple and non-statisticians.

The purpose of an analysis based on classification tree is to find out the factors that contribute the separation of successful and unsuccessful students among the all recorded data. When the classification tree is generated it is possible to calculate the probability of each student's outcome. If once the classification tree is formed, it is possible to predict the study outcome for newly enrolled student. In each tree node the percentages for successful and unsuccessful students is given and also absolute size of the node. The variable names above the node are the predictor's criteria that make the split for the node according to classification and regression tree. Each node is split according to predictor criteria. The searching is stops when the split with the largest improvement in goodness of fit. Possible predictor variables in the dataset were included in the classification tree in splits process, which is detected in feature selection criteria. We used two stopping criteria in the training process:

Global Journal of Computer Science and Technology (C) Volume XII Issue XV Version I Year 2012

1. A maximum tree depth has been proceeding into 3 levels for CHAID tree.
2. A maximum tree depth has been proceeding into 4 levels for CART tree.

Lastly for each classification tree we have assigned different costs to the classification outcomes i.e.: classification matrix. It is one of the processes of increasing the correctly classified student's outcomes. We categories the student's outcome according to Pass and Fail in the academic courses

V. Our Contribution

In this research we explore the concepts and technique of SVM to classify the data collected in our experiments. We have approximately collected twelve hundreds (1200) data from University of Chittagong where about thirty (30000) thousands students are studying. To assess these large amounts of data we have found that SVM is very efficient and exact technique in our proceedings. By imposing the SVM, we have mapped the data to meaning full forty two (42) data which are shown in the figure 3 and 4.In figure 3 we have depicts that the reasons for the age where mainly focused on the pivotal age of greater or less than twenty four.

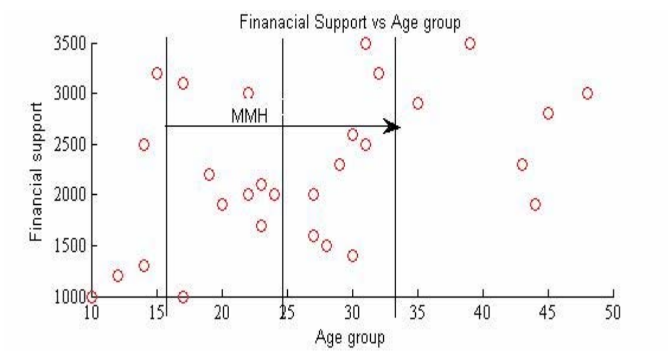

Fig. 3 : Data classification for age group using SVM

From the figure above we can easily measure the Maximum Margin Hyper plane (MMH).At MMH the resultant outcome of age group using SVM is determined. We have also used the MATHLAB to accelerate the accuracy of the implementation. In the same process we have had accomplished our design and implementation for the financial support data using the methodology of SVM.

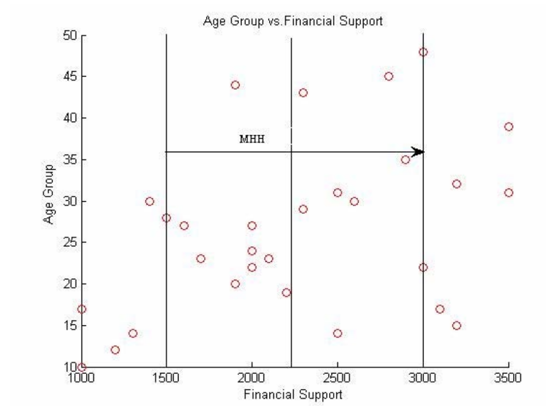

Fig. 4 : SVM to classify the financial support data

Now, we have calculated the chi square values of collected data. The procedures of chi square values are given bellow:

Step1: First insert the observed value in each cell of observable table. Inserted value collected from record.

Domain category	Option1	Option 2	Total
Cotegory1	a	b	a + b
Category	c	d	c + d
Total	a +	b	a + b + c

Step 2: Calculate expected value for every cell of the describing table.

Domain	Option 1	Option 2	Total
Cotegory1	a1=(a+b)*(a+c)/(a+b	b1=(a+b)*(b+d)/(a+b+c	a1 + b1
Category2	c1=(a+c)*(b+d)/(a+b	d1=(c+d)*(b+d)/(a+b+c	c1 + d1
Total	a1 + c1	b1 +d1	a1+

Step 3: calculating chi value for every cell using the following formula:

χ^2 = *(observed value-expected value)²/expected value*

Step 4: calculate total chi –value for domain using the following formula

$$\chi^2 = \sum_{i=1}^{n} (\text{observed value-expected value})^2/\text{expected value}$$

Step 5: calculating degree of freedom using following rule

Degree of freedom,
df = (No.of.rows-1)*(No.of.columns-1)

Step 6: calculate p-value (probability value) using following method in Ms Excel

P-value=CHIDIS (Chi value, df)

VI. Explanation of Chi-Square (χ^2) and P-Value

Step 1: consider the domain is financial support in which category1=Yes

category2=No, Option1=Pass and Option2=Fail. The value of every cell collects from database.

Financial Support	Pass	Fail	Total
Yes	18	5	23
No	11	8	19
Total	29	13	42

Step 2: calculating expected value for each cell using describing formula

Financial Support	Pass	Fail	Total
Yes	23*29/42=15.88	23*13/42=7.12	23
No	19*29/42=13.12	19*13/42=5.88	19
Total	29	13	42

Step 3: calculating chi value for every cell using the describing formula:

Financial Support	Pass	Fail
Yes	$\chi^2=(18-15.88)^2/15.88$ $= 0.283$	$\chi^2=(5-7.12)^2/7.12$ $= 0.631$
No	$\chi^2=(11-13.12)^2/13.88$ $= 0.343$	$\chi^2=(8-5.88)^2/5.88$ $= 0.764$

Step 4: calculate total chi –value for domain

$\chi^2= 0.283+0.631+0.343+0.764=2.021$

Overall chi-value for every domain in Pass and Fail Category:

Table 3 : Individual Chi-value for each category

Domain Name	Possible category	Pass	Fail
Financial Support	Yes	0.283	0.631
	No	0.343	0.764
Age group	Age>24	0.337	0.753
	Age<24	0.120	0.267
Gender	Male	0.050	0.112
	Female	0.082	0.182
Disabilities	Yes	0	1.54
	Female	0.20	0.040

Step 5: calculating degree of freedom using following the rule

Degree of freedom df = (2-1)*(2-1)
= 1

Step 6: calculate p-value (probability value) using in Ms Excels

P-value =CHIDIS (2.021, 1)
= 0.155

VII. Factors Selection

Factors selection is an important process to assess the prediction of dropout of the students. The prediction has relate the variable that determines the rate of success The number of predictor variables is not so large and we don't have to select the subset of variables for further analysis which is the main purpose of applying feature selection to data. However, feature selection could be also used as a pre-processor for predictive data mining to rank predictors according to the strength of their relationship with dependent or outcome variable. During the factors selection process no specific form of relationship, neither linear nor nonlinear, is assumed. The outcome of the factors selection would be a rank list of predictors according to their importance for further analysis of the dependent variable with the other methods for regression and classification. Here the figure below shows the relative outcome of the predictor's value.

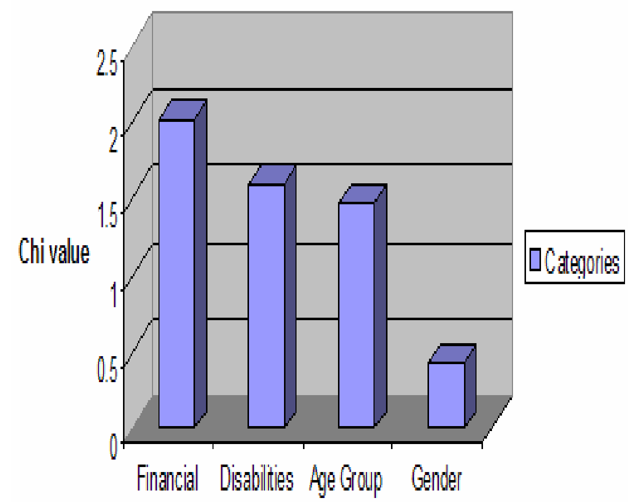

Figure 5 : Importance plot for predictors

Results of factors selection has presented in Figure 3 on the importance plot and in Table 1. The top three predictors for the study outcome are financial support of students, disabilities, age group and gender in which they are study.

Table 4 : Best predictors for dependent variable

Domain Name	Chi-value	P-value
Financial Support	2.021	.155
Disabilities	1.6	.206
Age Group	1.477	.224
Gender	.426	.514

Global Journal of Computer Science and Technology (C) Volume XII Issue XV Version I Year 2012

In all three cases, i.e. for all three definitions of the dependent variable, if the top 4 variables are selected, we get the same list of predictors. Therefore we can conclude that the list of important predictors is quite robust to changes in the study outcome definition. We may proceed into the next step using the top 4 variables:

1. Financial Support
2. Age Group
3. Gender
4. Disabilities

From Table 4, *P*-values from the last column only the first three chi-square values are significant at 10% level. Though the results of the feature selection suggested continuing analysis with only the subset of predictors, which includes Financial Support, Age group and Gender, we have included all available predictors in our classification tree analysis. We follow an advice given in Luan & Zhao (2006) who suggested that even though some variables may have little significance to the overall prediction outcome, they can be essential to a specific record.

a) Contribution of Cart

Figure 6 shows the CART classification tree for study outcome. It shows that only three variables were used to construct the tree: (1) financial support (2) age group and (3) gender.

The largest successful group (i.e. students who successfully completed the course) consists of 18 (78.76% of all participants) students (Node 1). The financial support of students in this group is yes. Students in this group enrolled on the age group in either age>24 or age<=24. The largest successful group (i.e. age group) contains 9 (60% of all participants) students (Node 4).

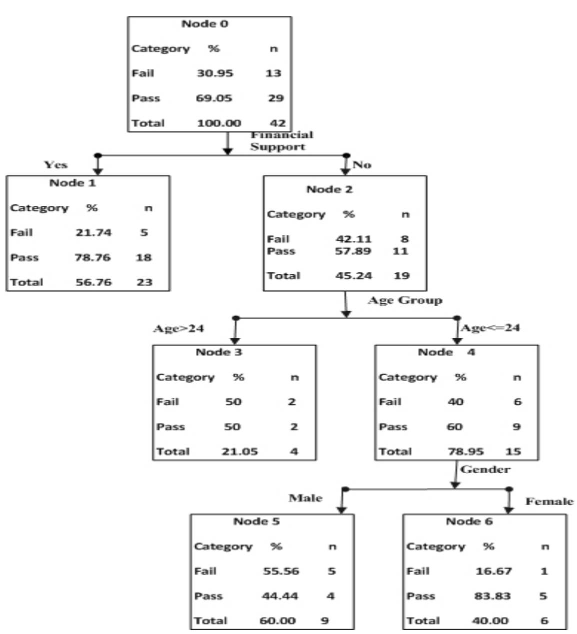

Fig. 6 : CART Classification

VIII. Knowledge Base for Collected Data

A knowledge base in artificial intelligence is a place where information are stored or designed for machine or device by which it will work. In general, a knowledge base is a consolidate stock for information: a library, a database of related information about a particular subject could all be considered to be examples of knowledge bases. The process of building knowledge base is called knowledge engineering. A knowledge base is integrated collection of choosing logic, building a knowledge base, implementing [31] the proof theory, inferring new facts. The main advantage of engineering is that it requires less commitment and thus less work. To help the focus the development of knowledge base and to integrate the designer's thinking the following five step methodology can be used:

1. Decide what to talk about
2. Decide on a vocabulary of predicates, function, and constant.
3. Encode general knowledge about the domain.
4. Encode a description of the specific problem instance.
5. Pose queries to the inference procedure and answers.

In our work we have described a simple method of probabilistic inference that is, the computation from observed evidence of posterior probabilities for query propositions. We have used the joint probability as the knowledge base from which answer to all question may be derived. We have had built the knowledge base by considering two Boolean variables. The table 7 is an example of two valued propositional logic which is the bases of knowledge base representation:

Table 7: Concepts of propositional logic to design a Knowledge Base using the proposition of Boolean events A, B and C.

	B		¬ B	
	C	¬ C	C	¬ C
A	111	110	101	100
¬ A	011	010	001	000

Based on table 7, we have designed the knowledge base (Joint probability distribution) for our research activity. Here we have considered those events which have true (one or 1) Boolean values. Table 8 is an example of knowledge base for events A, B and C:

Table 5 : Fully Joint probability distribution

	B		⌐ B	
	C	⌐ C	C	⌐ C
A	P(A)*P(B) *P(C)	P(A)*P(B)*P(⌐ C)	P(A)*P(⌐ B)*P(C)	P(A)*P(⌐B)*P (C)
⌐ A	P(⌐A)*P(B) *P(C)	P(⌐ A)*P(B) *P(⌐ C)	P(A)*P(⌐ B) *P(C)	P(A)*P(⌐B)*P (C)

By keeping the similarities s with the table 5, we compared our factors as financially good and financially not good, fail and not fail and so on. The designing of knowledge base for the factors which we are considered are given in table 6:

Table 6 : Fully join probability distribution (Knowledge base)

	Financial		⌐ Financial	
	Age<=24	Age>24	Age<=24	Age>24
Fail	23/42*31/42*13/42 =0.125	23/42*11/42*13/42 =0.044	19/42*31/42*13/42 =0.103	11/42*19/42*13/42 =0.037
Pass	23/42*31/42*29/42 =0.279	23/42*11/42*29/42 =0.099	31/42*19/42*29/42 =0.231	11/42*19/42*29/42 =0.082

Where
0.125+0.044+0.103+0.037+0.279+0.099+0.231+0.082=1

IX. Bayes'Theorem and Conditional Probability

Bayes' theorem and conditional probability are opposite to each other. Given two dependent events A and B. The conditional probability of P (A and B) or P (B/A) will be P (A and B)/P (A). Related to this formula a rule is developed by the English Presbyterian minister Thomas Bayes (1702-61).According to the Bayes rule it is possible to determine the various probabilities of the first event given the outcome of the second event in a sequence of two events.

The conditional probability:

$$\mathrm{P(B/A)} = \frac{P(AandB)}{P(A)} \qquad (1)$$

The equation (1) will help to find out the probabilities of B after being occurrences of the A. we get the Bayes' theorem for these two events as follows:

$$\mathrm{P\ (A/B)} = \frac{P(A).P(B/A)}{P(B)} \qquad (2)$$

If there are more events like A1, A2, and B1, B2.In this case the Bayes theorem to determine the probability of A1 based on B1will be as follows:

$$\mathrm{P\ (A1/B1)} = \frac{P(A1).P(B1/A1)}{P(A1).P(B1/A1)+P(A2).P(B2/A2)}$$

X. Experimental Result

Consideration Classification and Regression Tree (CART):

For Node 4:
IF Financial Support="Yes" AND Age<=24 THEN "PASS"= (18/23)*(9/15) =0.49

Now applying the Bayes theorem on table 5 we have got the following outcomes:

If one student pass based on his financial condition is "Yes" and age<=24 then
P (Pass | Financial condition="Yes" ^age<=24) =

P (Pass ^Financial condition="Yes" ^age<=24)/ P (Financial condition="Yes" ^age<=24)

P (Pass ^Financial condition="Yes" age<=24) =0.279

P (Financial condition="Yes" ^ age<=24) =0.125+0.279 = 0.404

P (Pass | Financial condition="Yes" ^age<=24) =0.125/0.404=0.31

The total resultant of Bayes Theorem and CART of all data considering financial condition and age group we have got the following table 7:

Table 7 : Bayes Rules to predict the academic success

Rule	Study Outcome	Probability	
		CART	Bayes Theorem
IF Financial Support="Yes" AND Age<=24 THEN	PASS	0.47	0.69
IF Financial Support="Yes" AND Age>24	PASS	0.39	0.69
IF Financial Support="NO" AND Age<=24	PASS	0.35	0.68
IF Financial Support="NO" AND Age>24	PASS	0.29	0.68

Global Journal of Computer Science and Technology (C) Volume XII Issue XV Version I Year 2012

XI. Conclusion

This study examines the background information from enrolment data that impacts upon the study outcome of Information Systems students at the Department of Computer Science & Engineering. Based on results from feature selection (Figure 5 and Table 4), the CART trees presentation it was found that the most important factors that help separate successful from unsuccessful students are financial support , age group and gender. Demographic data such as gender and age though significantly related to the study outcome, according to the feature selection result.

Based on results from table 5 and 6 by implementing the knowledge of propositional knowledge base and Bayes theorem based on knowledge base to predict the academic success. Better result is found from using Bayes Network.

This study is limited in three main ways that future research can perhaps address. Firstly, this research is based on background information only. Secondly, we used a dichotomous variable for the study outcome with only two categories: pass and fail. Thirdly, from a methodological point of view an alternative to a classification tree should be considered. The prime candidates to be used with this data set are logistic regression and neural networks.

References Références Referencias

1. Al-Radaideh, Q. A., Al-Shawakfa, E. M., & Al-Najjar, M. I. (2006). Mining student data using decision trees. In the Proceedings of the 2006 International Arab Conference on Information Technology (ACIT'2006*)*.
2. Md.Sarwar Kamal, Linkon Chowdhury, Sonia Farhana Nimmy*,* 'New Dropout Prediction for Intelligent System*", International Journal of Computer Applications (0975 – 8887) Volume:42–No.16, March 2012.*
3. Bean, J. P., & Metzner, B. S. (1985). A conceptual model of non traditional undergraduate student attrition. *Review of Educational Research, 55, 485-540.*
4. Boero, G., Laureti, T., & Naylor, R. (2005). An econometric analysis of student withdrawal and progress- sion in post-reform Italian universities. Centro Ricerche Economiche Nord Sud - CRENoS Working Paper 2005/04.
5. Cortez, P., & Silva, A. (2008). Using data mining to predict secondary school student performance. In the Proceedings of 5th Annual Future Business Technology Conference, Porto, Portugal, 5-12. Dekker, G. W., Pechenizkiy, M., & Vleeshouwers, J. M. (2009). Predicting student drop out: A case study.
6. Proceedings of the 2nd International Conference on Educational Data Mining (EDM'09). July 1-3, Cordoba, Spain, 41-50.
7. Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data* mining, infer- ence and prediction (2nd ed.). New York: Springer.
8. Han, J., & Kamber, M. (2006). Data mining: Concepts and techniques (2nd ed.). Amsterdam: Elsevier. Herrera, O. L. (2006). Investigation of the role of pre- and post admission variables in undergraduate insti-Mutational persistence, using a Markov student flow model. PhD Dissertation, North Carolina State Uni-versity, USA.
9. Horstmanshof, L., & Zimitat, C. (2007). Future time orientation predicts academic engagement among first-year university students. British Journal of Educational Psychology, *77* (3): 703-718.
10. Ishitani, T. T. (2003). A longitudinal approach to assessing attrition behavior among first-generation stu- dents: Time-varying effects of precollege characteristics. *Research in Higher Education, 44(4), 433-449.*
11. Ishitani, T. T. (2006). Studying attrition and degree completion behavior among first-generation college students in the United States. Journal of Higher Education, 77(5), 861-885.
12. Jun, J. (2005). Understanding dropout of adult learners in e-learning. PhD Dissertation, the University of Georgia, USA.
13. Kember, D. (1995). *Open learning courses for adults: A model of student progress*. Englewood Cliffs, NJ: Education Technology.
14. Luan, J., & Zhao, C-M. (2006). Practicing data mining for enrolment management and beyond. New Di- reactions for Institutional Research, 31(1), 117-122.
15. Murtaugh, P., Burns, L., & Schuster, J. (1999). Predicting the retention of university students. Research in Higher Education, 40(3), 355- 371.
16. Nandeshwar, A., & Chaudhari, S. (2009). Enrollment prediction models using data mining. Retrieved Janu- ary 10, 2010,
17. Nisbet, R., Elder, J., & Miner, G. (2009). Handbook of statistical analysis and data mining applications. Amsterdam: Elsevier.
18. Noble, K., Flynn, N. T., Lee, J. D., & Hilton, D. (2007). Predicting successful college experiences: Evi- dence from a first year retention program. Journal of College Student Retention: Research, Theory & Practice, 9(1), 39-60.
19. Pascarella, E. T., Duby, P. B., & Iverson, B. K. (1983). A test and reconceptualization of a theoretical model of college withdrawal in a commuter institution setting. Sociology of Education, 56, 88-100.
20. Pratt, P. A., & Skaggs, C. T. (1989). First-generation college students: Are they at greater risk for attrition

than their peers? *Research in Rural Education, 6* (1), 31-34.
21. Reason, R. D. (2003). Student variables that predict retention: Recent research and new developments. *NASPA Journal, 40(4),* 172-191.
22. Rokach, L., & Maimon, O. (2008). Data mining with decision trees – Theory and applications. New Jersey: World Scientific Publishing.
23. Yu, C. H., DiGangi, S., Jannasch-Pennell, A., Lo, W., & Kaprolet, C. (2007). A data-mining approach to differentiate predictors of retention. In the Proceedings of the Educause Southwest Conference, Austin, Texas, USA.
24. Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. Expert Systems with Applications, 33, 135-146.
25. Simpson, O. (2006). Predicting student success in open and distance learning. *Open Learning, 21*(2), 125-138.
26. Siraj, F., & Abdoulha, M. A. (2009). Uncovering hidden information within university's student enrolment data using data mining. MASAUM Journal of Computing, 1(2), 337-342.
27. Strayhorn, T. L. (2009). An examination of the impact of first-year seminars on correlates of college stu- dent retention. Journal of the First- Year Experience & Students in Transition, 21(1), 9-27.
28. Tharp, J. (1998). Predicting persistence of urban commuter campus students utilizing student background characteristics from enrollment data. Community College Journal of Research and Practice, 22, 279-294.
29. Artificial Intelligence A modern Approach by Stuart Russell and Peter Norvig, ISBN 81-297-0041-7
30. Baker, R. S. J. D., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. Journal of Educational Data Mining, 1, 3-17.