New Dropout Prediction for Intelligent System

Md.Sarwar kamal
Lecturer
Computer Science and Engineering
BGC Trust University, Bangladesh
Chondanaish, Chittagong.

Linkon Chowdhury
Lecturer Computer Science and Engineering
BGC Trust University
Bangladesh
Chandanaish, Chittagong.

Sonia Farhana Nimmy
Lecturer Computer Science and Engineering
BGC Trust University
Bangladesh
Chandanaish, Chittagong.

ABSTRACT
The main purpose of this research is to develop a dynamic dropout prediction model for universities, institutes and colleges. In this work, we first identify dependent and independent variables and dropping year to classify the successful from unsuccessful students. Then we have classify the data using Support Vector Machines(SVM).SVM helped the data set to be properly design and manipulated . The main purpose of applying this identification is to design a Knowledge Base which is sometimes known as joint probability distribution .The concepts of propositional logic helped to build the 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 socio-demographic variables such as age, gender, ethnicity, education, work status, and disability and study environment that may in-flounce persistence or dropout of students at university level

Keywords: Intelligent System, Dynamic dropout Prediction, Joint Probability Distribution, Bayes Theorem Dependent ad Independent variables, Propositional Logic, Knowledge Base,MATHLAB,SVM.

1. INRODUCTION
Increasing student retention or persistence is a long-term goal in all academic institutions. The consequences of student attrition are significant for students, academic and administrative staff. The importance of this issue for students is obvious: school leavers are more likely to earn less than those who graduated. Since one of the criteria for government funding in the tertiary education environment in Chittagong University is the level of retention rate and academic are under pressure to come up with most vulnerable students to low student retention at all institutions of higher education are the first-year students, who are at greatest risk of dropping out in the first term or semester of study or not completing their program /degree on time. 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 [1] and pro-active measures. Once identified, these 'at-risk' students can be targeted with academic and administrative support to increase their chance of staying on the course.

The background characteristics such as academic and socio-demographic variables [2] (age, sex, ethnic and financial aid) have been identified in retention literature as potential predictor variables of dropout. At the time of enrolment in the Computer Science and Engineering (CSE), University of Chittagong, the only information. i.e. variables we have about students are those contained in their enrolment forms. The question we are trying to address in this paper is whether we can use the enrolment data alone to predict study outcome for newly enrolled student.

The main objective of this work[2] is to explore factors that may impact the study outcome in the Technical course at the Computer science & Engineering. The Technical course is a core course for those majoring in IT and for most students an entry point, i.e. the first choice they are taking with the CSE. This issue have not been examined so far for CSE and this paper attempts to fill the gap. More specifically the enrolment data have used to achieve the following objectives:

1. Build a knowledge Base for Student information.
2. Build models for early prediction of study outcome using the student enrolment data.
3. Present results, which can be easily understood by the users (students and academic staff).

2. COLLECTED STATISTICALDATA
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.

3. 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 hyperplane 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 Hyperplane. 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 non-convex, 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 Hyperplane (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.

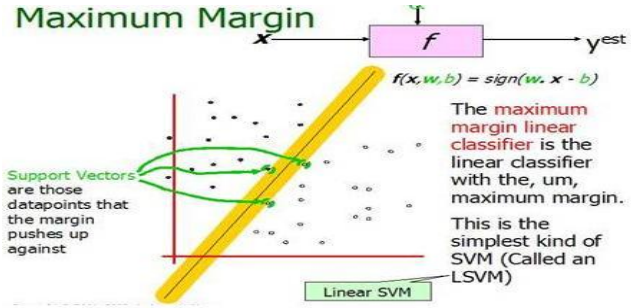

Figure 1: Representation of Support Vectors.

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

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

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, $wx+b \geq 1$

[a] If $Y_i = +1$;

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

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

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]

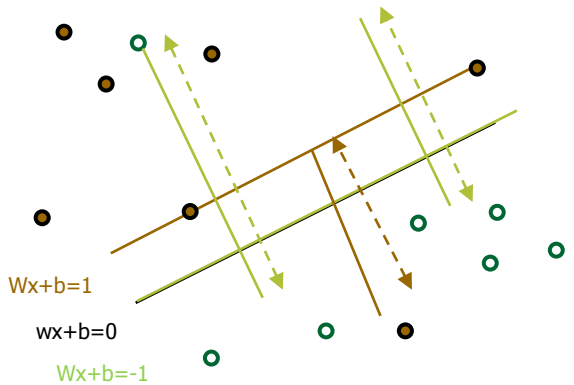

Figure 2: Representation of Hyper planes

Distance of closest point on hyperplane 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 hyperplane to nearest points. Maximum Margin = M = 2 / ||w||

3.1 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.

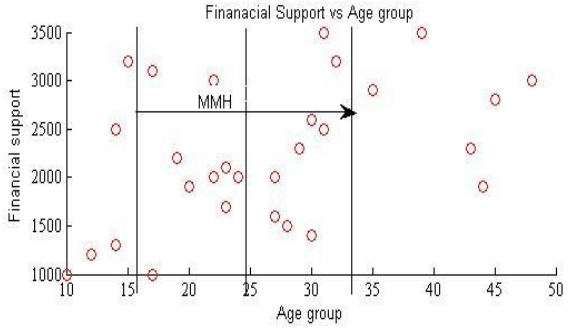

Fig 3: Data classification for age group using SVM

From the figure above we can easily measure the Maximum Margin Hyperplane (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.

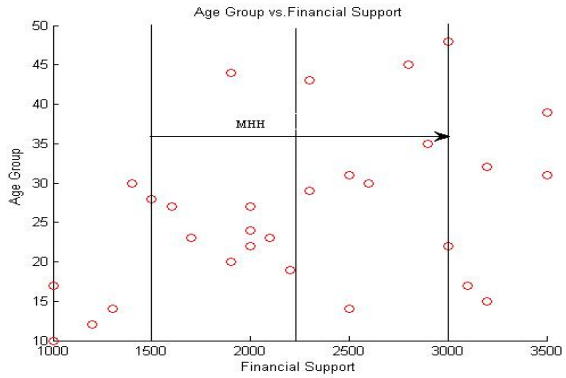

Fig 4: SVM to classify the financial support data.

4. 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:

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 3 is an example of two valued propositional logic which is the bases of knowledge base representation:

Table 3: 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 3, 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 4 is an example of knowledge base for events A, B and C:

	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)

Table 4: Fully Joint probability distribution

By keeping the similarities s with the table 4, 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 5:

	Financial support(Good)		⌐ Financial support(Good)	
	Male	⌐ Male	Male	⌐ Male
Fail	23/42*26/42*13/42 =0.105	23/42*16/42*13/42 =0.065	19/42*26/42*13/42 =0.087	16/42*19/42*13/42 =0.053
⌐ Fail	23/42*26/42*29/42 =0.234	23/42*16/42*29/42 =0.144	26/42*19/42*29/42 =0.193	16/42*19/42*29/42 =0.119

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

5.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:

$$P(B/A)= \frac{P(AandB)}{P(A)} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

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:

$$P (A/B) = \frac{P(A).P(B/A)}{P(B)} \quad \ldots\ldots\ldots\ldots\ldots\ldots (2)$$

If there are more events like A1, A2, and B1, B2.In this case the Bayes theorem to determine the probability of A_1 based on B1will be as follows:

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

Now applying the Bayes theorem on table 5 we have got the following outcomes:
If one student fail based on his financial condition is "Good" and age<=24 then

P (Fail | Financial condition="Good" ^age<=24) =

$$\frac{P(Fail \wedge finacialcondition = good \wedge age <= 24)}{P(financialcondition = good \wedge age < 24)}$$

P (Fail ^ Financial condition="Good" ^ age<=24) = 0.125
P (Financial condition="Good" ^ age<=24) =0.125+0.279 = 0.404
P (Fail | Financial condition="Good"^ age<=24) = 0.125/0.404=0.309

The total resultant of Bayes Theorem of all data considering financial condition we have got the following table 6:

	Financial support(Good)		⌐ Financial support(Good)	
	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
⌐ Fail	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

Table 7: Fully join probability distribution (Knowledge base)

Rule	**Outcome**
P (Fail | Financial condition="Good" ^age<=24)	30.9 %
P (Fail | Financial condition="Good" ^age>24)	30.8 %
P (Fail | ⌐Financial condition="Good" ^age<=24)	30.8 %
P (Fail | ⌐Financial condition="Good" ^age>24)	31.1 %
P (⌐Fail | Financial condition="Good" ^age<=24)	69.1 %
P(|⌐Fail | Financial condition="Good" ^age>24)	24.2 %
P(⌐Fail | ⌐Financial condition="Good" ^ age<=24)	69.2 %
P (⌐Fail | ⌐Financial condition="Good" ^age>24)	68.9 %

Table 8: Bayes Rules to predict the dropout based on gender:

Rule	Outcome
P (Fail | Financial condition="Good" ^Gender=Male)	30.9 %
P (Fail | Financial condition="Good" ^ Gender= ⌐ Male)	31.1 %
P (Fail | ⌐Financial condition="Good" ^ Gender=Male)	31.7 %
P (Fail | ⌐Financial condition="Good" ^ Gender= ⌐ Male)	30.8 %
P (⌐Fail | Financial condition="Good" ^ Gender=Male)	69.0 %
P (|⌐Fail | Financial condition="Good" ^ Gender= ⌐ Male)	31.1 %
P (⌐Fail | ⌐Financial condition="Good" ^ Gender=Male)	68.9 %
P (⌐Fail | ⌐Financial condition="Good" ^ Gender= ⌐ Male))Male)	69.2 %

6.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 table 6 and 8 by implementing the knowledge of propositional knowledge base and Bayes theorem based on knowledge base to predict the dropout 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. This would suggest that the background information (disabilities) gathered during the enrolment process, does not contain sufficient information for an accurately separation of successful and unsuccessful students.

This study is limited in three main ways that future research can perhaps address. Firstly, this re- search 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.

7.REFERENCES

[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].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.

[3].Bean, J. P., & Metzner, B. S. (1985). A conceptual model of nontraditional 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 progres- 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

[6].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.

[7].Proceedings of the 2nd International Conference on Educational Data Mining (EDM'09). July 1-3, Cordoba, Spain, 41-50.

[8].Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data* mining, infer- ence and prediction (2nd ed.). New York: Springer.

[9].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-

[10].mutational persistence, using a Markov student flow model. PhD Dissertation, North Carolina State Uni- versity, USA.

[11].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.

[12].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, 44(4), 433-449.*

[13].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.

[14].Jun, J. (2005). Understanding dropout of adult learners in e-learning. PhD Dissertation, The University of Georgia, USA.

[15].Kember, D. (1995). *Open learning courses for adults: A model of student progress*. Englewood Cliffs, NJ: Education Technology.

[16].Luan, J., & Zhao, C-M. (2006). Practicing data mining for enrollment management and beyond. New Di- rections for Institutional Research, 31(1), 117-122.

[17].Murtaugh, P., Burns, L., & Schuster, J. (1999). Predicting the retention of university students. Research inHigher Education, 40(3), 355-371.

[18].Nandeshwar, A., & Chaudhari, S. (2009). Enrollment prediction models using data mining. Retrieved Janu- ary 10, 2010,

[19].Nisbet, R., Elder, J., & Miner, G. (2009). Handbook of statistical analysis and data mining applications.Amsterdam: Elsevier.

[20].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.

[21].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.

[22].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.

[23].Reason, R. D. (2003). Student variables that predict retention: Recent research and new developments.*NASPA Journal, 40(4),* 172-191.

[24].Rokach, L., & Maimon, O. (2008). Data mining with decision trees – Theory and applications. New Jersey: World Scientific Publishing.

[25].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.

[26].Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. Expert Systems with Applications, 33, 135-146.

[27].Simpson, O. (2006). Predicting student success in open and distance learning. *Open Learning*, *21*(2), 125-138.

[28].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.

[29].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.

[30].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.

[31].Artificial Intelligence A modern Approach by Stuart Russell and Peter Norvig, ISBN 81-297-0041-7