

PREDICTION OF STUDENTS ACADEMIC FAILURE AND DROP OUT FROM COLLEGE USING DATAMINING TECHNIQUES

Suganya S¹, Dr.V.Narayani²

¹Research Scholar, Bharathiar University, Coimbatore.(India)

²Director i/c- MCA, Karpagam College of Engineering, Coimbatore.(India)

ABSTRACT

In recent years, there has been increasing interest in the use of data mining to investigate scientific questions within educational research, an area of inquiry termed educational data mining. Educational Data Mining (also referred to as "EDM") is defined as the area of scientific inquiry centered on the development of methods for making discoveries within the unique kinds of data that come from educational settings, and using those methods to better understand students. The inclination that all students should finish college has grown throughout the last decades and continues to be an important goal for all educational levels in this new century. Non-completion has been related to all sorts of social, financial, and psychological. Many studies have attempted to put together a process that will identify students at risk of dropping out by using various research methodologies. The purpose of this study is to investigate college dropping out through the use of data mining of existing data sources with decision trees. Data mining techniques are applied to predict students dropping out from college. This paper presents an architecture that uses educational data mining techniques to predict and identify those who face the risk of dropping out.

Keywords: Data Mining Techniques, Dropouts, Education Data Mining, Research Methodology, Students At Risk,

I. INTRODUCTION

College education is the essential element of knowledge based economy: universities are nowadays the most important source of basic research and are therefore crucial for the development of new technologies (Enders, 2001; Garača, 2007). Past years have shown a growing interest and concern in several countries about the problem of college failure and the determination of its main contributing factors [1]. The great deal of analysis [2] has been done on characteristic the factors that affect the low performance of students (school failure and dropout) at totally different instructional levels. To identify and find useful information hidden in massive databases is a trouble task [3]. A very promising solution to reach this goal is that the use of information discovery in databases techniques or data mining in education, referred to as instructional data processing, EDM [4]. This new area analysis focuses on the event of methods to better understand students and therefore the

settings in which they learn [5]. In fact, there are good samples of the way to apply EDM techniques to make models that predict dropping out and student failure specifically [6]. These works have shown promising results with respect to those social science, economic, or instructional characteristics that may be additional relevant in the prediction of low educational performance [7].

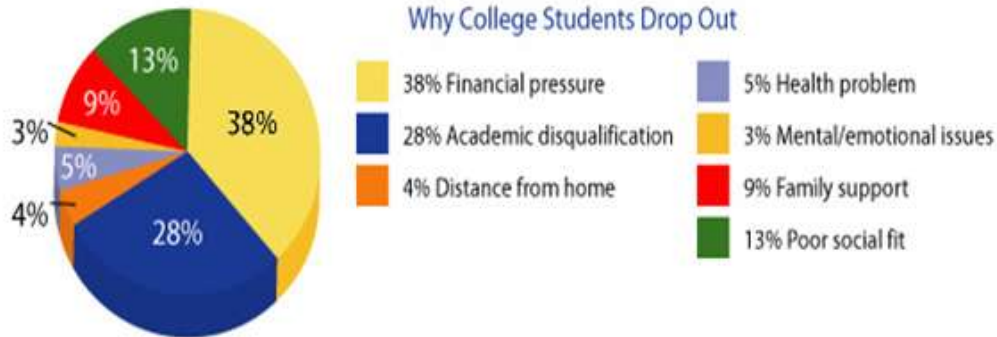
II. BACKGROUND AND RELATED WORK

The topic of explanation and prediction of academic performance is widely researched. In the earlier studies, the model of Tinto [12] was the predominant theoretical framework for considering factors in academic success. Tinto considers the process of student attrition as a socio-psychological interplay between the characteristics of the student entering college and the experience at the institute. This interaction between the student's past and the academic environment leads to a degree of integration of the student into this new environment. A higher degree of integration is directly related to a higher commitment to the educational institute and to the goal of study completion. Later studies tried to operationalize this model identifying the factors like peer group interactions, interactions with faculty, faculty concern for student development and teaching, academic and intellectual development, and institutional and goal commitments that affect the student's integration [11]. These factors proved to have a predictive capacity across different institutions, and showed therefore to be a potential tool in identifying students who might drop out. Many studies included a wide range of potential predictors, including personality factors, intelligence and aptitude tests, academic achievement, previous college achievements, and demographic data and some of these factors seemed to be stronger than others, however there is no consistent agreement among different studies ([8], [9], [10], and [13]). One of the recent European studies [8] has confirmed that sex (only in technical schools), age at enrollment, score on pre-university examination, type of pre-university education, type of financial support, father's level of education and whether or not living at the university town may all have an impact on the drop out. All studies show that academic success is dependent on many factors, where grades and achievements, personality and expectations, as well as sociological background all play a role.

III. A SAMPLE SURVEY ON INDIAN STUDENTS DROPPING OUT FROM COLLEGE

NAME OF INSTITUTE	NO. OF DROPOUTS DURING 2012-2015		
	2012-13	2013-14	2014-15
IIT Bhubaneswar	4	15	16
IIT Bombay	74	66	72
IIT Mandi	Nil	Nil	Nil
IIT Delhi	130	249	169
IIT Indore	4	Nil	3
IIT Kharagpur	197	129	209
IIT Hyderabad	32	15	39
IIT Jodhpur	Nil	Nil	Nil
IIT Kanpur	Nil	Nil	Nil
IIT Madras	Nil	8	Nil
IIT Gandhinagar	1	4	1
IIT Patna	1	12	10
IIT Roorkee	159	188	228
IIT Ropar	1	Nil	Nil
IIT-BHU, Varanasi	3	2	2
IIT Guwahati	Nil	9	8
Total	606	697	757

IV. PROBLEM OF COLLEGE DROPOUTS



According to the National Center for Education Statistics, only 59% of first-time students that began seeking bachelor's degrees or the equivalent in 2005 graduated within six years. Only 34% of males graduated in four years, compared to 42% of females. Similarly, only 22.5% of American Indians and Alaska natives graduated within four years. The study collected over twenty possible reasons or 'shocks' for students dropping out. These include events occurring at the school, such as an assault, conflict with a faculty member or departure of a close friend from the institution. Reasons for dropping out also include events occurring outside of the college or university, such as marriage and other personal issues. According to the study, six critical events are the most common reasons for withdrawals: Students are recruited by a job or other institution, receive an unanticipated bad grade, have conflicts with a roommate, lose financial aid, become clinically depressed or have a substantial increase in tuition or living costs. While other events, such as becoming pregnant or developing a substance abuse problem, are very likely to cause a dropout, they occur in low numbers and are therefore limited as major causes.

V. RULES FOR DROPOUT PREDICTION

Data mining techniques and soft computing techniques are much more effective in the field of learning and prediction. The extracted features are given to the training phase, in this phase the data samples are learned their operation and actions to perform effectively. Once the training method is completed, the models itself learns the state and achieve the accurate result. Based on the trained samples the new data samples are given to the model to test the student performance.

DROP OUT CONDITION	STATUS
If(parent occupation = average) and (change of goal = yes) and (history of arrear = above 5)	then dropout = yes
If(parent occupation = high)) and (history of arrear = below5) and (stress = yes) and (health problem = yes)	then dropout = yes

if(enrolled in other = no) and (placement facility = yes) and (bad habits = no) and (satisfaction = poor) and (social impacts = low)	then dropouts = no
if(participation in other activities = yes) and (history of arrear = below 5) and (stress = no) and (placement = yes)	then dropouts = no
if(social impacts = high) and (satisfaction = poor) and (change of goal = yes) and (participation in other activities = yes)	then dropout = yes

VI. DATA MINING TECHNIQUES

5.1 k-Nearest Neighbour classifier (k-NN)

k-NN algorithm is one of the well-known classification methods. It is based on learning by comparing a given test tuple with training tuples that are similar to it. When a new instance is introduced, k-NN finds the k-nearest neighbors of this new instance and determines the label of the new instance by using these k instances.

In this study, closeness is defined in terms of a distance metric called Euclidean distance. Although our data set is mostly consisted of categorical variables, each category has a numerical counterpart; thus used the Euclidean distance. To assign a particular class to the test sample, the most common class among the k nearest neighbours is used and unclassified test sample is classified by a majority vote of its neighbours. A good value for k, the number of neighbours, was determined experimentally. Starting with k=1, 10-fold cross validation technique is used to estimate the error rate of the classifier. This process was repeated for k=10 times and in each iteration by incrementing k to allow for one more neighbour. The value k was selected as '3' that gave the minimum error rate. The accuracy rate of classification can be very satisfying in some cases, there are just two parameters to learn, and the classification is very robust to missing values. However, the selection of distance function d might be difficult particularly for educational data sets.

5.2 Decision Tree classifier (DT)

DT is a powerful classification and prediction technique). There are several popular decision tree algorithms such as ID3, C4.5, and CART (classification and regression trees). DT is in the form of a tree structure, where each node is either a leaf node (indicating the value of the target class of examples) or a decision node (specifying a test to be carried out on a single attribute value, with one branch and sub-tree for each possible outcome of the test) .DTs have many advantages such as very fast classification of unknown records, easy interpretation of small-sized trees, robust structure to the outliers' effects, and a clear indication of most important fields for prediction but DTs are very sensitive to over-fitting particularly in small data-sets.

In this study, to generate a decision tree, the C4.5 algorithm was used, which is an extension of earlier ID3 algorithm. To construct the tree, entropy measure was used in the determination of nodes. Since the attributes with higher the entropy cause more uncertainty in outcome, they were selected in order of increasing entropy.

5.3 Naive Bayes classifier (NB)

A simple probabilistic classifier called as Naive Bayes classifier was also used in student dropout classification.

Naive Bayes algorithm as the simplest form of Bayesian network is one of the easiest algorithms to perform and has very satisfactory accuracy and sensitivity rates. The posterior probability of each class, C_i , is obtained by the Naive Bayes classifier using Bayes rule.

The classifier makes the simplifying assumption that the attributes, A , are independent given the class, so the likelihood can be obtained by the product of the individual conditional probabilities of each attribute given the class. Thus, the posterior probability $P(C_i|A_1, \dots, A_n)$ can be given by the following equation/assumption:

$$P(C_i|A_1, \dots, A_n) = P(C_i)P(A_1|C_i) \dots P(A_n|C_i)/P(A)$$

This assumption is usually called the Naive Bayes assumption, an a Bayesian classifier using this assumption is called the Naive Bayesian classifier, often abbreviated to 'Naive Bayes'. Effectively, it means that we are ignoring interactions between attributes within individuals of the same class.

5.4 Neural Network classifier (NN)

The prediction of the student dropouts was also performed by feed-forward NN. It is another inductive learning method grounded on computational models of neurons and their networks as in humans' central nervous system. NN is a set of connected input/output units where each connection has a distinct weight associated with each other. During the learning phase, the network learns by adjusting the weights so as to be able to predict the correct class of the input samples.

In this study, the back propagation algorithm was performed for learning on a multilayer feed-forward neural network. The input layer of the network consisted of nine variables of the students. The hidden layer included 50 neurons and the output layer had one neuron, which was determined by our experimental studies.

5.5 Genetic Algorithm (GA)

In this study GA is used as a feature selection method after the classification algorithms. GA is a paradigm developed based on natural selection and Darwin's theory about evolution, and used for solving optimization problems. Algorithm starts with a random initial set of solutions for the feature subset selection problem. This initial set is called as chromosome. The chromosome was represented by a binary vector of dimension nine, the total number of features. After the initial population is selected randomly, roulette wheel selection, mutation and uniform crossover operations continue until a new generation is replaced with the original generation. It is expected that the new generation would give a better solution than the old one. Solutions are selected according to their fitness to form new solutions.

A lower fitness value refers a more optimized solution to the problem, which is minimizing the classification error after 10-fold cross-validation in this study. These steps were repeated until the number of populations, which is 200, is satisfied. Due to the high number of the solution space, we have used the number of population throughout the analyses. Over successive generations, the population evolved toward an optimal solution.

VI. CONCLUSION

This study continues with deeper mining of academic failure. In this paper we studied newly emerging field of

Educational Data Mining (EDM). Predicting student drop outs at college can be a difficult task because it is a multifactor problem and also because the available data are normally imbalanced. Several strategies can be developed and implemented enabling the educational institutions to transform a wealth of information into a wealth of predictability, stability and profits. In this article, we are presenting the various data mining techniques related to the students' failure prediction and we can improve the performance of the students' failure and dropout prediction.

REFERENCES

- [1]. L. A. Alvares Aldaco, "Comportamiento de la deserción y reprobación en el colegio de bachilleres del estado de baja california: Caso plantel ensenada," in Proc. 10th Congr. Nat. Invest. Educ., 2009, pp. 1–12.
- [2]. F. Araque, C. Roldán, and A. Salguero, "Factors influencing university drop out rates," *Compuwet. Educ.*, vol. 53, no. 3, pp. 563–574, 2009.
- [3]. M. N. Quadril and N. V. Kalyankar, "Drop out feature of student data for academic performance using decision tree techniques," *Global J. Comput. Sci. Technol.*, vol. 10, pp. 2–5, Feb. 2010.
- [4]. C. Romero and S. Ventura, "Educational data mining: A survey from 1995 to 2005," *Expert Syst. Appl.*, vol. 33, no. 1, pp. 135–146, 2007.
- [5]. C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 40, no. 6, pp. 601–618, Nov. 2010.
- [6]. S. Kotsiantis, K. Patriarcheas, and M. Xenos, "A combinational incremental ensemble of classifiers as a technique for predicting students' performance in distance education," *Knowl. Based Syst.*, vol. 23, no. 6, pp. 529–535, Aug. 2010.
- [7]. J. Más-Estellés, R. Alcover-Arándiga, A. Dapena-Janeiro, A. Valderruten-Vidal, R. Satorre-Cuerda, F. Llopis-Pascual, T. Rojo-Guillén, R. Mayo-Gual, M. Bermejo-Llopis, J. Gutiérrez- Serrano, J. García-Almiñana, E. Tovar-Caro, and E. Menasalvas-Ruiz, "Rendimiento académico de los estudios de informática en algunos centros españoles," in Proc. 15th Jornadas Enseñanza Univ. Inf., Barcelona, Rep. Conf., 2009, pp. 5–12.
- [8]. Lassibille, G., Gomez, L. N. Why do higher education students drop out? Evidence from Spain, *Education Economics* 16(1), p. 89-105, 2007.
- [9]. References [1] Herzog, S. Measuring determinants of student return vs. dropout/stopout vs. transfer: A first-to-second year analysis of new freshmen. In Proc. of 44th Annual Forum of the Association for Institutional Research (AIR), 2004.
- [10] Terenzini, P. T., Lorang, W. G., Pascarella, E. T. Predicting freshman persistence and voluntary dropout decisions: a replication, *Research in Higher Education* 15(2), p. 109-127, 1981.
- [11] Thai Nge, N., Janecek, P., Haddawy, P. A comparative analysis of techniques for predicting academic performance. In Proc. of 37th Conf. on ASEE/IEEE Frontiers in Education, 2007.
- [12] Tinto, V. Limits of theory and practice in student attrition, *Journal of Higher Education* 53, p. 687-700, 1982.
- [13] Touron, J. The determination of factors related to academic achievement in the university: implications for

4th International Conference on Science, Technology and Management

India International Centre, New Delhi

(ICSTM-16)

15th May 2016, www.conferenceworld.in

ISBN: 978-81-932074-8-2

the selection and counseling of students, Higher Education 12, p. 399-410, 1983.