Journal of Advances in Science and Technology
Vol. 15, Issue No. 2, September-2018, ISSN 2230-9659

An Investigation of Dropout Assessment Methods for College Students in Madhya Pradesh

Shivendra Kumar Dwivedi[1]* Dr. Prabhat Pandey[2]

[1] Research Scholar, Department of Computer Science, APS University, Rewa

[2] OSD Additional, Directorate of Higher Education Rewa Division, (M.P.), India – 486001

Abstract – Student dropout prediction is an indispensable for numerous intelligent systems to measure the education system and success rate of any colleges well. Therefore, it becomes essential to develop efficient methods for prediction of the students at risk of dropping out, enabling the adoption of proactive process to minimize the situation. Thus, this research paper propose a prototype machine learning tool which can automatically recognize whether the student will continue their study or drop their study using classification technique based on decision tree and extract hidden information from large data about what factors are responsible for dropout student. Further the contribution of factors responsible for dropout risk was studied using discriminate analysis and to extract interesting correlations, frequent patterns, associations.

In this study, the descriptive statistics analysis was carried out to measure the quality of data using SPSS 24.0 statistical software.

The main reason recorded for dropout of students at this college were dropout factor (illness & homesickness, poor economic condition), Educational factors (learning problems & difficult courses, change of Institution with present goal and low placement rate) and institutional factors (campus environment, too many rules in hostel life and poor entertainment facilities).

Keywords: Data Mining, Dropout, Prediction, Machine Learning Algorithm, Classification, Decision Tree, Discriminate Analysis, Association Rule.

----------♦----------

1. INTRODUCTION

We live in the information- era, accumulating data is easy and storing it inexpensive. Today the amount of increases, the ability to understand and make use of it does not keep pace with its growth. Stored data can be used to generate useful information for decision making. The data mining can be applied in various real life applications like market analysis, education, and scientific exploration, etc.

If data is rich in quality and quantity then reliable information hidden in the data can be generated. To generate relevant information from the data which have stored in the repositories over the years, Data mining tools and techniques is used. Data mining is an automatic data analysis process that helps users and administrators to discover and extract patterns from stored data. The use of data mining technique to analyze an educational database is absolutely expected to be great benefit to the higher educational institutions.

Education is one of the social factors whereby gender disparity is reflected. The number and proportion of educated females is very low. As the grade level of education increases, the number of female students starts to decline. Consequently, higher education remains the level of learning where females are less represented both as students and staff. The very few women that are fortunate enough to join higher learning institutions can be characterized by lower academic performance and higher forced withdrawal. Female education in India has got momentum after independence. There has been slow development in technical fields, but during the recent past, the continuation of students in technical as well as in almost every field has envisaged higher status, still due to several factors there has been comparatively lesser percentage of student education as per national statistics conducted by Govt. of India. .

Students higher studies being undertaken from students coming from different parts of the Madhya Pradesh as per need of the present education and

employment scenario, the college has restructured its traditional educational system and introduced professional courses in the field of commerce and management

The purpose of this research is to study the student's dropout risk assessment and causes of dropout at undergraduate level using data mining tool and techniques to assist the student dropout program on campus. Information like factor affecting of dropout were collected from the student's residing in college campus, to predict the students drop out rate who need special attention.

1.1 Data Mining

"Data mining", often known as Knowledge Discovery in Databases (KDD), refers to mining knowledge from immense amount of data [1]. Data mining techniques are used to operate on huge amount of data to discover hidden patterns and relationships helpful for decision making. While data mining and KDD are frequently treated as synonyms, actually data mining is a part of the knowledge discovery process. The data mining defined as "the non-trivial process of identifying valid, novel, previously unknown, potentially useful information, and ultimately understandable patterns from data in database" [3]. Finding a useful patterns in data are known by different names in different communities (e.g. Knowledge extraction, Information discovery, data dredging, Information harvesting, data/pattern analysis and business intelligence) [3].

1.1.1 Data Mining Task

Description: describe the dataset in a concise and summary manner and presents intere Sting data into human interpretable or understandable format .e.g. Clustering and Association Rule etc.

Prediction: constructs one or a set of models, perform inference on the dataset and attempt to predict unknown or future values of other variables of interest .e.g. Classification and Regression Analysis etc.

1.1.2 Data Mining Functionalities

Functionalities of data mining are as follows-

Characterization

Data characterization is a summarization of the general characteristics or features of objects in a target class of data**.** Relevant data to a user- specified class are collected by a database query and run through a summarization module to extract the essence of the data at different level of abstractions.

Discrimination

Data discrimination is a comparison of the general features of target class data objects with the general features of objects from one or a set of contrasting class.

Association Analysis

Association analysis is the discovery of association rules. It studies the frequency of items take place together in transactional databases based on threshold called support and confidence. Association analysis is commonly used for market basket analysis. The discovery of association rule can help retailers to develop marketing strategies by gaining into which items are frequently purchase together.

Classification and Prediction

Classification analyzes a set of training data and builds a model for each class based on the features in the data. This model is used to classify new objects and also known as supervised learning. Derived model may be represented in various forms, such as classification (IF- THEN) rules, decision trees and neural network etc.

A decision tree is a flow –chart like tree structure where each node denotes a test on an attribute value and each branch represents an outcome of the test. A neural network is a collection of linear threshold units that can be trained to distinguish objects of different classes. Classification can be used to predicting the class label of data objects. Classification predicts categorical labels (or discrete values), prediction models continuous- valued function.

Clustering

Clustering is an unsupervised learning, in which the class labels of the training samples are not known. Cluster is a collection of data objects that are similar to on one another. Similarity can be expressed by distance functions, specified by use experts. A good clustering method produces high quality clusters to ensure that inter- cluster (object of different class) similarity is low and the intra-cluster (object in a same class) similarity is high.

1.2 Educational Data Mining

Educational organizations are one of the important parts of our society and playing a vital role for growth and development of any nation. Educational data mining is the application of data mining. It is an emerging interdisciplinary research area that deals with the development of methods to explore data originating in an educational context. Educational data mining is an emerging trend, designed for automatically exploring the unique types of data from large repositories of educationally related data.

The Educational Data Mining community defines EDM as follows: Educational data mining is an

emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings in which they learn.

EDM include analysis (evaluation/ exploration) of educational processes including admission, alumni relations, course selections, predicting drop out student, student's success rate, course success rate, performance evaluation of student, learning behavior of students, list of course taken by the student, when the student selected or changed his or her academic major, finding which tasks, courses etc are difficult/ easy for which student's, finding elective courses often taken by student's.

1.3 Need of Current Research

The earlier prediction of dropout student is challenging task in the higher education. Data analysis is one way to scale down the rate of dropout students and increase the enrollment rate of students in the college. It is fact that the number of student dropout quite often in the first year of graduation. The rate of student's dropout in the college depends on the educational system adopted by the college. The needs of current research are as follows:

- Predicting dropout students at an early stage of the degree program help management 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.
- The generated knowledge will be quite useful for understanding the problem in better way and to have a proper planning or decision to scale down the dropout rate.
- This study is quite useful for better planning and implementation of education program and infrastructure to increase the enrollment rate of students in the college.

1.4 Social Implication of this Research

The larger number of dropout in the higher education has serious consequences for our society. Dropouts experiences high level of unemployment and receive lower earning than graduates [4]. Dropouts are also more likely than graduates to become dependent on welfare engage in illegal activities, and experience health and affective problems .Finally, high rate of dropping out of higher education create a negative momentum for youths in a society, particularly during difficult economic periods.

1.5 Objective of research

1. To study the strength of relationship between attributes (correlation).
2. Analyze the effects of independent variables influencing graduation and dropout rates in higher education and indicates which variable are important in explaining a dropout student (multiple regression).
3. Finding association of various factor leading to students dropout at higher education in college, where discovering of pattern or association helps in effective decision making.

2. REVIEW OF LITERATURE

The several authors have been worked out in the area of educational data mining at national and international level. Some of the important studies are as follows:

Al-Radaideh et al. [], applied classification data mining techniques to improve the quality of the higher education by evaluating the main attributes of students that affect the their performance. This study was used to predict the student's final grade in a course.

Ayesha et.al. [], performed study on student learning behavior. For this factors like class quizzes mid and final exam assignment are studied. This study will help the tutors to reduce the ratio of drop out and improve the performance level of students.

Bharadwaj and Pal [], used the decision tree method for classification to evaluate performance of student's. The objective of their study is to discover knowledge that describes students' performance in end semester examination. This study was quite useful for identifying the dropout's student in earlier stage and students who need special attention and allow the teacher to provide appropriate advising.

Boero, Laureti & Naylor [], they found that gender (males have a higher probability of dropping out relative to the reference group of females) is one of the principal determinants of the probability of dropping out and age has a significant positive effect.

Bray [], in his study on private tutoring and its implications, observed that the percentage of students receiving private tutoring in India was relatively higher than in Malaysia, Singapore, Japan, China and Sri Lanka. It was also observed that there was an enhancement of academic performance with the intensity of private tutoring and this variation of intensity of private tutoring depends on the collective factor namely socio-economic conditions.

El-Halees [], proposed a case study that used educational data mining to analyze students' learning behavior. The objective of his study is to show how useful data mining can be used in higher education to improve student's performance.

They applied data mining techniques to discover relevant information from large database such as association rules and classification rules using decision tree, clustering and outlier analysis.

3. RESEARCH METHODOLOGY

Success percentage rate of any institute can be improved by knowing the reasons for dropout student. In present research, information on various parameters was collected through a structured questionnaire on personal interview basis from a composite sample of 1353 students of college Predicting the students dropout status whether they continue to their study or not, needs lots of parameters such as personal, academic record, social, environmental, etc. Variables are necessitated for the effective prediction.

Since the present study is in relation to classify the various quantitative and qualitative factors to study the causes of dropout which belongs to the process of knowledge discovery and data mining. This information will be helpful for the management to reduce the dropout rate in campus. In order to achieve the above mentioned objectives the following steps were followed

3.1 Data Preparation

The data used in this study was prepared from the higher educational institute of Madhya Pradesh through structured questionnaire. The questionnaire has been constructed based on theoretical and empirical grounds about factor affecting student's performance and causes of dropout. The questionnaire included socio-demographic indicators; Educational factors Parental Attitudes, Causes of dropout, and Institutional factors, etc. Data were collected with consulted of various institute of Madhya Pradesh

Table 1

Name of college	District	Total Dropout	Dropout factor							
			No proper faculty in college	Long illness	Ignorance of guardian	Agriculture work	Poor economic condition	Long distance	No friendly environment of college	Lack of education facilities
TRS College	Rewa	167	34	12	19	23	16	21	22	20
Govt PG college	Satna	157	24	19	15	21	23	14	19	22
Govt college Jaitpur	Shahdol	124	13	10	19	21	14	13	16	18
MK College	Anooppur	175	32	16	22	17	23	22	28	15
Govt College Niwas	Mandala	203	35	21	30	22	41	21	12	21
GTB College	Jabalpur	123	19	14	12	13	21	16	11	17
MB College	Indore	98	19	7	16	13	10	11	9	13
LSA College	Dhar	102	21	6	18	16	9	11	12	9
Govt college Harsud	Khandwa	115	18	9	11	16	14	19	13	15
Govt college Patan	Narsingpur	89	13	8	11	14	11	12	10	7

Before the initial visit to review the records, a coding system was created for each variable to be documented. It was not important to document dropout status but also all withdrawal reasons for the students.

3.2 Data Selection and Transformation

After collection of data, the dataset was prepared to apply the data mining techniques. Before data preprocessing was applied to measure the quality and suitability of data. In this step only those attributes were selected which were needed for data mining. For this, remove missing values; selection of relevant attribute from database or removing irrelevant attributes, identifying or remove outlier values from data set, and resolving inconsistencies of data. Some of the irrelevant parameters was removed from database such as ID, age, date of birth, category, marital status, state of domicile, mother tongue, religion, etc. A categorical variable is constructed based on the numeric parameter .A grade scale is used for evaluation of student performance at college. "Dropout status" is constructed based on the view of respondents; it has two possible values- "Yes" (students who are completely decided to withdraw from their course) and "No" (students who are want to continue their study).

The final dataset used for the study contains 1353 instances (119 in the "No" category and 1234 are in "Yes" category) each described with 8 attributes (1 output and 7 input variables), nominal and numeric. The study is limited to the student data for undergraduate. Finally, the pre-processed data were transformed into a suitable format to apply data mining techniques.

3.3 Data Analysis Techniques

In this study, one quantitative and the other qualitative data analysis techniques have to be employed using statistical methods and data mining methods.

- **Statistical Methods**

The data collected were analyzed using SPSS statistical software to measure the quality of data based on descriptive statistics, crosstab and statistical test. Examine each variable by using cross tab. The entire student was examined to understand the reasons of withdrawal.

- **Data Mining Methods**

After proper collection, scrutiny and transformation of data using appropriate measures, Data mining classification and decision tree approach were applied to predict student dropout rates and dropout causes in early stage of their study either before or after completion of their first year of their study program.

4. RESULTS AND DISCUSSION

The methodology described in the previous chapter provides the baseline for data gathering; however the

present chapter focuses on analysis and interpretation of data using statistical tools and data mining techniques. The data collected was compiled in Microsoft Excel-2007 software. The SPSS software was used for frequency distribution and descriptive statistics analysis for each response variable. After measuring the quality of data, data mining classification techniques were applied to study the factors affecting student's dropout in higher education at college level and subsequently causes for dropout was also analyzed.

4.1 Description of the sample

As stated from the first chapter, the aim of the study was to analyze the factors influencing dropout student and its reasons. For this study, data was collected by using questionnaires which were delivered to the respondents by hand (N=1353) of college students along with approval letter of university authority and instructions of research problem. After collection of datasets, about 119 datasets were deleted due to incomplete information. The remaining (N=1234) datasets were analyzed which includes information like socio- demographic factors (age, category, marital status, residential status, family status, religion), Parental status (annual income, parent's education, parent's occupation), educational profile of respondents (grade in secondary school, grade in higher secondary, location of school, medium of education, stream in higher secondary, course admitted, admission criteria, source of education expenses, study duration) and attitude of respondent Long illness , ignorance of guardian, poor economic condition, non-friendly environment of college etc. The findings of the research problems are as follows.

4.2 Descriptive statistics

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Dropout rate	9	17.8454	20.8943	18.971285	.8802368
No proper faculty available in college	9	12.6233	15.2504	14.108897	.9600758
Long illness	9	20.1156	22.7606	21.354460	.7764938
Ignorance of guardian	9	19.4992	23.0138	20.664486	1.1056729
Agriculture work	9	.0000	23.0990	7.307711	10.9748948
Poor economic condition	9	29.2450	34.4377	31.672553	1.6188069
Long distance	9	16.4752	18.9492	17.442767	.7255850
Non friendly environment of college	9	32.2846	35.3155	33.710773	.9958006
Lack of educational faculty	9	15.4815	18.3892	16.720540	1.0926251
Valid N (list wise)	9				

From the table we found that there are nine factors which are affecting student dropout in higher education of selected district of Madhya Pradesh. the table also reviles that the most important factor with highest mean value 33.7 is non-friendly environment of college which lead to maximum dropout in the research area where as the factor which is least responsible for dropout is student involve in agriculture work(mean = 7.31) the majorities of factor found significant towards measuring dropout of students are Long illness , ignorance of guardian, poor economic condition, non-friendly environment of college etc. the deviation from mean is found to be maximum for agriculture work and minimum for long distance.

Correlations

		No proper faculty available in college	Long illness	Ignorance of guardian	Agriculture work	Poor economic condition	Long distance	Non friendly environment of college	Lack of educational faculty	Dropout rate
No proper faculty available in college	Pearson Correlation	1	.730*	.659	.650	-.193	.712*	.367	-.303	.709*
	Sig. (2-tailed)		.025	.054	.058	.619	.031	.331	.428	.033
Long illness	Pearson Correlation	.730*	1	.782*	.681*	-.176	.827**	.374	-.232	.809**
	Sig. (2tailed)	.025		.013	.044	.651	.006	.322	.548	.008
Ignorance of guardian	Pearson Correlation	.659	.782*	1	.828**	-.372	.956**	.571	-.086	.971**
	Sig. (2tailed)	.054	.013		.006	.324	.000	.108	.826	.000
Agriculture work	Pearson Correlation	.650	.681*	.828**	1	-.391	.698*	.319	-.275	.725*
	Sig. (2-tailed)	.058	.044	.006		.297	.037	.402	.474	.027
Poor economic condition	Pearson Correlation	-.193	-.176	-.372	-.391	1	-.191	.074	.403	-.246
	Sig. (2-tailed)	.619	.651	.324	.297		.623	.849	.282	.523
Long distance	Pearson Correlation	.712*	.827**	.956**	.698*	-.191	1	.629	-.073	.991**
	Sig. (2-tailed)	.031	.006	.000	.037	.623		.070	.853	.000
Non friendly environment of college	Peason Correlation	.367	.374	.571	.319	.074	.629	1	.607	.669*
	Sig. (2-tailed)	.331	.322	.108	.402	.849	.070		.083	.049
Lack of educational faculty	Pearson Correlation	-.303	-.232	-.086	-.275	.403	-.073	.607	1	-.023
	Sig. (2tailed)	.428	.548	.826	.474	.282	.853	.083		.953
Dropout rate	Pearson Correlation	.709*	.809**	.971**	.725*	-.246	.991**	.669*	-.023	1
	Sig. (2-tailed)	.033	.008	.000	.027	.523	.000	.049	.953	

*. Correlation is significant at the 0.05 level (2-tailed).
**. Correlation is significant at the 0.01 level (2-tailed).
c. Listwise N=9

From the above table it was observed that the correlation between non proper faculties available in college is 70.9 it means dropout is explain by 70.9 % due to non-proper faculty of higher education. Of Madhya Pradesh. In case of second factor the drop out was explain by long illness is 80.9 % drop out of ignorance of guardian is reported as 97 % which was the highest correlation

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-19.271	.000		.	.	-19.271	-19.271
	No proper faculty available in college	.218	.000	.238	.	.	.218	.218
	Long illness	.366	.000	.323	.	.	.366	.366
	Ignorance of guardian	3.101	.000	3.895	.	.	3.101	3.101
	Agriculture work	-.093	.000	-1.160	.	.	-.093	-.093
	Poor economic condition	.311	.000	.573	.	.	.311	.311
	Long-distance	-3.433	.000	-2.830	.	.	-3.433	-3.433
	Non-friendly environment of college	.734	.000	.831	.	.	.734	.734
	Lack of educational faculty	-.645	.000	-.801	.	.	-.645	-.645

a. Dependent Variable: Dropout rate

The ignorance of guardian is seen to be maximum contributory factor for measuring dropout where as poor economic condition is lowest among of them. There are various factors such as nonfriendly env. Lack of education faculty which are significant contributor to dropout. other factor are seems to be in significant current study.

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	17.845442	20.894281	18.971285	.8802368	9
Residual	0E-7	0E-7	0E-7	0E-7	9
Std. Predicted Value	-1.279	2.185	.000	1.000	9
Std. Residual	.	.	.	.	0

a. Dependent Variable: Dropout rate

From the residual table it is found that the model is able to predict 21 % of the entire factor which is significant for our study with the help of all variables under study

5. CONCLUSION AND FUTURE WORK

The main purpose of the study was to investigate the major factors causing the dropout of students in

undergraduate Courses at Madhya Pradesh. It has been also assessed the pros cons of the affirmative action provided for the students at the college while analyzing the reaction of respondents about college environment and college infrastructure etc.

Taking the objectives into account, an extensive review of the available literature was made. Based on the review of the related literature, basic questions were to indicate the nature of assumed relationships among various parameters considered in this study. To verify the stated assumptions, the study had employed different procedures and techniques. In particular, the study was conducted taking samples of 10 colleges from Madhya Pradesh. Data were collected in pre-scheduled format which was handed over to the college along with instructions.

Apart from its significance in providing information about the factors for students' dropout either it is low or high in higher education at college level, important measure could be taken to tackle it. The present study provides valuable information to upgrade the college education system Madhya Pradesh.

The findings of student questionnaire were analyzed and interpreted. The computer software called SPSS 24.0 was used for the treatment of the collected data. Statistical techniques such as Frequency distribution for single variable, Cross Tabulation, , Discriminates analysis and association rule have been used to study the causing factors for dropping out the students.

- Based on the Correlation based feature selection result, it was found that a student's dropout was positively correlated with 8 response variables No proper faculty in college, Long illness, Ignorance of guardian, Agriculture work, Poor economic condition, Long distance, No friendly environment of college, Lack of education facilities.

The generated information will be quite useful for management of college to develop policies and strategies for better planning and implementation of educational program and infrastructure under measurable condition to increase the enrolment rate in College to take effective decision to reduce student dropout.

5.1 Suggestion

- Though the students' dropout rate is very low (<10%) which is good indicator for the college, even though the college should give due emphasis to minimize the dropout rate.
- Thus, the college authorities have to organize workshops, seminars and conferences about the issues equity, affirmative action and multiculturalism to strengthen the already available positive beliefs towards affirmative action. Such type of concept will certainly reduce the student's dropout rate.
- Retention and college success are affected by many factors in which the institutional environment takes the lions share.

RFERENCES

1. Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules. In *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, J. B. Bocca, M. Jarke and C. Zaniolo, Eds. Morgan Kaufmann, pp. 487-499.
2. Alaa el-Halees (2009) Mining Students Data to analyze e-Learning Behavior: A Case Study.
3. 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)*.
4. Bharadwaj B.K. and Pal S. (2011). "Data Mining: A prediction for performance improvement using classification", International Journal of Computer Science and Information Security (IJCSIS), Vol. 9, No. 4, pp. 136- 140.
5. Bharadwaj B.K. and Pal S.: "Mining Educational Data to Analyze Students" Performance", International Journal of Advance Computer Science and Applications (IJACSA), Vol. 2, No. 6, pp. 63-69.
6. Boero, G., Laureti, T., & Naylor, R. (2005). An econometric analysis of student withdrawal and progression in post-reform Italian universities. Centro Ricerche Economiche Nord Sud - *CRENoS Working Paper 2005/04*.
7. Cortez P., and Silva A. (2008), "Using Data Mining to Predict Secondary School Student Performance", In EUROSIS, A. Brito and J. Teixeira (Eds.), pp.5-12.
8. El-Halees, A. (2008) 'Mining Students Data to Analyze Learning Behavior: A Case Study', The International Arab Conference of Information Technology (ACIT2008) – Conference Proceedings, University of Sfax, Tunisia, Dec 15- 18.
9. Garson G. D. (2008). Discriminant Function Analysis, Statnotes: Topics in Multivariate Analysis.

10. Han, J & Kamber, M. (2001). Data Mining: Concepts and Techniques. San Francisco (CA, USA): Morgan Kaufmann Publishers, Academic Press. 550 p. ISBN: 1-55860-489-8.

11. Herrera, O. L. (2006). *Investigation of the role of pre- and post-admission variables in undergraduate institutional persistence, using a Markov student flow model*. PhD Dissertation, North Carolina State University, USA.

Corresponding Author

Shivendra Kumar Dwivedi*

Research Scholar, Department of Computer Science, APS University, Rewa