International Journal of Latest Trends in Engineering and Technology
Special Issue SACAIM 2017, pp. 415-418
e-ISSN:2278-621X

A SURVEY ON DROPOUT OF STUDENTS USING PREDICTIVE ANALYTICS

Carol Monteiro[1], Jeevitha Dsouza[2] & Mr. Suman Antony Lasrado[3]

Abstract: Dropouts of students in correspondence and open courses are increasing in academic year. This requires factors analysis that can lead to an increase in the dropout rate. Data Mining Technology from the Educational Data System Quest for hidden knowledge from an effective process It may lead to analyzing factors that affect the student, A good educational plan for students to evaluate less to get out of the course and Can lead to maintenance, Plus valuable information can be generated Take the decision of the shareholders to improve the quality of the higher education system.For analysis and predictionData mining technique can be used.Various levels of mining are applied to reduce useful effects. Different scenarios are compared and accuracy is calculated there. This study provides data mining work to predict the student's drop out feature. In this paper data mining algorithms is used to analysis of data set.After analysis the output is resulting factor is that the dropping rate of students in open courses is more than others. These analysis and information about prediction is more helpful for college management and teachers to improve education in a better way and also make changes if necessary.
Key Words: Data Mining, Education, Accuracy, dropping rate of students, prediction.

1. INTRODUCTION

Prediction - The method or actions to predict some aspect of the method or the various predictions of the data (predictor).Prediction is used for-

• Sometimes the prediction is used to predict

• Sometimes used to make inferences about the present Prediction can be done by using data mining techniques on large data sets. Data mining is a broad concept that involves a series of steps. Data is pre-processed first and then mining techniques are applied. The results of mining techniques are evaluated and understood. In recent years, many countries focus on factorsthis will affect the student's ability to get out of schoolor university or factors affecting low performance of students, etc. One of the major issues is the increase in dropout rates of students in higher educationin allinstitutions. Analyse the factors that affect student data hiding from the effective process of educational data systemExploration of knowledge is a good educational project and can lead to maintenance; reduce dropout rates for students and so on Valuable information can be made for determination Manufacturer of steak holders to improve the quality of higher education system. Study data, research on many issues in education Analysing and obtaining useful knowledge Educational Data Mining is called. EDM Uses data mining methods, some of them are predictive applications such as classification, and others such as clustering are considered descriptive in the field of education.The main goal of this seminar is to use the predictive method of educational data mining.The choice of education domain for priority analysis is the availability of information. For many courses such as MCA, MSC I have collected data for some students enrolled. Many students dropped out of their graduate education or did not join the master programs Found. The combination that causes the dropout rate to search for factors that affect and decrease or decrease factors dropout rate the main purpose to guess. We have considered three cases -

1. Take all the features identified by applying data algorithms and algorithms on them.
2. Based on their attributes based on the attributes of the attribute option, and based on their events, I have the first 10 properties andapplied classification algorithms have been taken.
3.Dueto the imbalance of the data we apply the data balance algorithm in the selected attribute and this sample for classification we use data.

2. LITERATURE REVIEW

[1]In recent years it has been observed that there is increase in rate of dropouts in correspondence courses and part-time courses. Many factors could be behind this increase in rate of dropout. Researchers are working towards this field to identify the factors so that teachers and management could be informed. If we know the root cause of the problem it will be easier to find solutions for problems. Researchers can identify factors and also predict which factors are dominating the dropout rate. This could be achieved by using data mining. Data can be taken from databases of universities and colleges and by application

[1] III Sem M.Sc, Department of Software Technology, St. Aloysius College AIMIT, Mangalore
[2] III Sem M.Sc, Department of Software Technology, St. Aloysius College AIMIT, Mangalore
[3] Asst. Professor M.Sc.ST, St. Aloysius College AIMIT Mangalore

of various techniques we can perform educational data mining to predict useful knowledge. Data mining methods are considered better then statistical methods as the data may be huge in size and it is difficult to process large datasets using statistical methods. Data mining process involves many steps like data gathering, data preprocessing, applying mining techniques, result interpretation. [3] Educational data mining also referred to as EDM use mining strategies and techniques to answer few important educational questions. It is based on computational approaches that analyze data collected from educational institutions. This can also be helpful in improving e-learning experience for students. EDM is dependent on distribution of data collected hence slight variations can cause change in results. [2] Students academic performance is the key focus area that plays important role in analysis process. [4]Educational institutions can use the outcome of analysis in strategic planning to help students improve performance. Dropouts can be due to various risk factors like financial conditions, parental education, marital status, etc. Relationship between dropout performance and these risk factors must be understood before devising any strategy. For mining of data many tools is available in market. One of the tools is weka. It is a powerful mining tool .It provides options of many techniques and there algorithms.

3. METHODOLOGY

Methods to follow:

1. Data collection. At this point, all the information available to the students will be collected. To do this,A group of factors that affect students' performance is identified and stored from various sources of information such as a college database, questionnaire, etc. Lastly, all information is organized in a dataset.
2. Data Pre-Processing. At this point, data mining techniques are placed in a database location to apply. To do this, traditional pre-processing techniques such as variable transforms, and data breakage, Data cleaning can be applied. Some other techniques, such as selection and re-balance of symptoms Data is applied to deal with problems with high dimension and imbalance data in these Data Collections.
3. Data Mining. At this stage, data mining algorithms are applied to predict student failure such as classification Techniques. The output will be loaded in the form of mining equipment and determination of wood and mathematical values to make this data Will be calculated. In addition, the cost-effective micro-classification method is also used unbalanced data to solve the problem. Various algorithms are compared to implement, evaluate and decide It gets better results.
4. Interpretation. At this point, the results obtained will be analyzed. Factors found (Rules and decision trees) and how they relate to them and are understood.

A. Data collection-

Student dropout rates are affected by thousands of factors. These factors include population, family, economy, educational background, educational, etc. Some examples of factors include family income, poor performance in school, distance from center, marital status, financial status, gender etc. The first step in data storage is to store raw data in it. I have collected data from St Aloysius College Mangalore. Students' data collected on BCA, MCA, and B.com are collected from the University database and survey form. The survey form is shown below. Several factors identified by this data in table 3.1.1

Table1. Identified factors

FEATURE	Description
PROGRAM	BCA, MCA, PGDCA,MSc
MEDIUM	English
DOB	Mm/dd/yyy
AGE	Numeric value
SEX	Male,female
CATEGORY	GEN,SC,OTHR,ST
EMPLOYED	Unemployed, Employed
YEARS_OF_EXP	Numeric value
LOAD OF WORK IN A WEEK	Numeric value
TYPE OF SHIFT	Rotational shifts with weekend off, Rotational shifts with weekend off, Fixed shifts with weekend off, Fixed shifts with weekend off
SALARY	Numeric value
RELIGION	Sikh, Hindu, Muslim, Christian,
NO_OF_SIBLINGS	Numeric value
LIVE_WITH_PARENTS	Yes,no
MAR_STATUS	Married/Unmarried
CHILDREN	Yes,no
BPL	Yes,no
NATIONALITY	Indian,others

AREA	Urban, Rural, Tribal
DISTANCE_FRM_HOME	NUMERIC VALUE
MODE_OF_TRANSPORT	LOCAL TRANSPORT, OWN VEHICLE, WALK TO STUDY CENTER
Time for Travel	NUMERIC VALUE
YEARPASS	NUMERIC VALUE
PERCENTAGE	NUMERIC VALUE
SCH_HOLDER	YES, NO

Data collected was in raw form with some missing values, incorrect values. Data was ensemble in excel sheet and data preprocessing step was used to make data ready for further mining process.

B. Data pre-processing-

Data is pre-prepared after raw data is collected. We propose to make data better in data mining. It may include data cleanup, variable conversion, selection of features, and data balance. When data is collected it may contain incomplete records and incompatible values. Pre-processing helps us get data in form that can effectively minimize data mining and analysis. It is important to maintain this step because data quality can be affected as a result. Using pre-processed data mining techniques such as classification give a good result. Data collected and available in a form that is expected for mining accuracy and high quality results is processed and represented. Initially stored data in one place in the Excel Sheet format is stored. Documents with complete data have been removed and entire lines are only considered. Furthermore, some data was added from a survey filled by students. Some features or properties are summarized for the simplicity of representation and understanding. In some cases, the data collected can be paralyzed, meaning that the number of documents belonging to one category is less than the other category. Since we have collected data through a survey form, it is possible that the data does not explicitly represent the entire situation. The algorithm's performance will fall if the dataset is not equilibrium. We can use the balance algorithm to deal with it. weka Tool provides a filter option called SMOTE. This is the highest-sample method. This category is less than instances. After applying the algorithms of the data balanced feature selection. These algorithms are used to identify those characteristics that affect most algorithms. Some examples of feature selection algorithms provided in the weka tool are the Principal Component, Consensancy Subset Eval, FilteredAttributeEval, FilteredSubsetEval, etc. We apply these algorithms in the dataset and get a range of features or features. The results of these algorithms are shown in the table below.

Table2. Feature Selected

Algorithm used for selecting feature	Feature selected
cfsSubsetEval	ENRNO,FEE_AMOUNT,MAR_STATUS,SEX,AREA
InfoGainAttributeEval	DOB, PROGRAM, CITY, Age, AREA, DIVISION
GainRatioAttributeEval	AREA, PROGRAM, DOB, SEX
FilteredSubsetEval	AREA, DIVISION
Principal components	PROGRAM, MEDIUM, DOB, Age, SEX

Table3. 2ranking of attributes selected by attribute selection algorithm

Feature selected	Frequency of occurrence
PROGRAM	3
AREA	4
SEX	3
DOB	3
MEDIUM	1

C. Data mining -

Data mining applies some mining techniques to the knowledge of understanding knowledge. There are several mining techniques applicable to data for analysis. These are classification, association, clustering. I applied classification mining technique in a dataset. Obtaining a prediction model is the main goal based on identified features. There are various classification algorithms that apply in the datasets in the vea tool. I applied the classification technique in three cases and compared them to their accuracy. These three cases are as follows:

1. Take all the features identified by applying data algorithms and algorithms on them.
2. After adopting the attribute optional algorithm and based on their features, I have taken the first 10 properties and Applied classification algorithms.
3. Due to the imbalance of data, we apply data balance algorithms on select attributes and use this sample data for classification

We calculate the accuracy of all three cases and in which case the price becomes more effective and gives better results. We have classification algorithms in Week.

1. Rule based classification algorithm - Rule-based classification uses IF-THEN rules for classification. In the following we can express the rule and the situation is finalized later

a.Part of the rule is known as a predetermined or precedent.

b. The latter part of the rule is known as the result of the rule.

c.The situation involves one or more attribute tests in the previous section, and these tests are rational.

d.The latter will have a class future. In Week we have various options for rule-based classification algorithms such as Jrip, conjuctive rule, DTNB, PART, etc.

2. Decision Tree Algorithm - A crucial way to represent information on classification algorithms. In Weka we have many algorithms for the decision tree like J48, NBTree, REPTree, SimpleCart. In accordance with these classification algorithms we obtain access rules and decision tree that will help us understand the future model and compare the exact value in different situations.

D. Interpretation -

By analyzing the decision tree and access rules we can identify the characteristics that influence the most classification and may not be noticeable in the guess. The accuracy of calculating in three cases is to determine what is the most cost effective way to classify our dataset. This step test data can be applied to model structure and desired output, which means class information can be obtained.

4. CONCLUSION

Students may encounter decisive action by predicting the dropout reasons. Moreover, the whole procedure will take a long and long time. The data collected is from different sources and data pre-processing needs to be made before the data is cleaned and balanced. Preprocessing step feature is used to identify the characteristics that affect the prediction process in algorithms. Many factors, such as demographic factors, socio-economic factors, family factors, etc. may be responsible for the fact that students come out of there. We can help us find useful knowledge of the predictive model that is tested in test data through the mining classification algorithms data and analysis of decision tree and access rules. The result obtained from such models helps teachers and management to identify problems and problems. In all three cases, we consider three cases and accuracy compared to the classification, which helps to understand the performance of students and the most effective way to help detect causes for drop-outs.Google Translate for Business: Translator Toolkit

5. FORECAST TO THE FUTURE

1. Identification of reasons for loss of efficiency during classification in all three cases.
2. Using other mining techniques and comparing outcomes
3. Data can be tuned and we can try to improve efficiency of classification algorithms.

6. REFERENCES

[1] C. Márquez-Vera, C.R.Morales, and S.V.Soto,"Predicting School Failure and Dropout by Using Data Mining Techniques", IEEE journal of Latin-American learning technologies, vol. 8, no. 1, February 2013, pp.7-14.

[2] M. N. Quadri1, N.V. Kalyankar, "Drop Out Feature of Student Data for Academic Performance Using Decision Tree Techniques", Global Journal of Computer Science and Technology, Vol. 10 Issue 2 (Ver 1.0), April 2010, pp. 2-5.

[3] M. Nasiri, B. Minaei, F. Vafaei, "Predicting GPA and Academic Dismissal in LMS Using Educational Data Mining: A Case Mining" ,IEEE, 6th National and 3rd International conference of e-Learning and eTeaching(ICELET),2012,pp.53-58.

[4] W. Yathongchai, C. Yathongchai, K. Kerdprasop, N. Kerdprasop, "Factor Analysis with Data Mining Technique in Higher Educational Student Drop Out", Latest Advances in Educational Technologies, 2003.

[5] P. S. Kumar, A. K. Panda, D. Jena, "Mining the factors affecting the high school dropouts in rural areas", International Journal of Advanced Computer Engineering and Communication Technology (IJACECT), Volume-2, Issue – 3, 2013, pp.1-6.

[6] E. Yom-Tov, G.F. Inbar,"Feature Selection for the Classification of Movements From Single MovementRelated Potentials", IEEE Transactions on neural systems and rehabilitation engineering, vol. 10, no. 3, September 2002,pp. 170-177.

[7] M. S. Chen, J. Han, P. S, Yu, "Data Mining: An Overview from a Database Perspective", IEEE Transactions on Knowledge and Data Engineering, Vol. 8, No. 6, December 1996

[8] Y. Kurniawan, E. Halim, "Use Data Warehouse and Data Mining to Predict Student Academic Performance in Schools: A Case Study (Perspective Application and Benefits)", IEEE International Conference on Teaching, Assessment and Learning for Engineering (TALE), August 2013,pp.98-103.

[9] M. Wook, Y. Hani Yahaya, N. Wahab, "Predicting NDUM Student's Academic Performance Using Data Mining Techniques", IEEE, Second International Conference on Computer and Electrical Engineering,2009,pp.357-361 [10] E. Gharavi, M. J. Tarokh, "Predicting customers' future demand using data mining analysis: A case study of wireless communication customer", IEEE,5th Conference on Information and Knowledge Technology,2013,pp.338343