Research Article

Association Rule Mining In Student's Dropout Risk Assessment: A Case Study

Dr. NandiniCahudhari[1], Dr. PradnyaVikhar[2], Dr. AvaniVasant[3]

[1]Professor, Babaria Institute of Technology, Vadodara
[2]Assistant Professor, KCES's COEM, Jalgaon
[3]Professor, Babaria Institute of Technology, Vadodara

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 23 May 2021

ABSTRACT
Student dropout risk assessment is essential for numerous intelligent systems to improve the performance and success rate of an institute. Therefore, efficient methods for prediction of the students at risk of dropping out, is the need of today's education system which enables the adoption of proactive process to minimize the situation. This paper propose a prototype machine learning tool which can automatically recognize the causes for whether the student will continue their study or drop their study using association rule mining. It also extracts hidden information from large data about the factors that are responsible for dropout student.

In this case study, the association rule analysis is carried out to find whether the student is having a dropout risk or not so that some preventive measures can be done to avoid it and improve the performance of the student. The analysis further used to predict the students drop out risk using five major problems such as family problem, health related problem , personal problem, financial problem and institutional problem.

Keywords— Data mining, Education Data Mining (EDM), Dropout, Prediction, Association Rule, Apriori algorithm

I. INTRODUCTION

All educational institutions aim to produce good results in their academic examinations. The data mining techniques can plays vital role in improving the results by predicting the performance of the students and prove beneficial to impart the quality of education in the educational institutions. Education Data mining (EDM)is a very promising discipline which has an imperative impact on prediction of students' performance.

In this paper students' drop out risk is evaluated and some attributes are selected which generate rules by means of association rule mining. 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 especially in the first semester. The rate of student's dropout in the college depends on the educational system adopted by the University. The needs of current research are as follows:

- Predicting the reasons for 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.
- Finding association of various factor leading to students dropout at education in engineering college, where discovering of pattern or association helps in effective decision making
- To study the dropout rate and causes of students in engineering education in North Maharashtra region.

II. RELATED WORK

Educational Data Mining (EDM) has been applied in various studies for exploring hidden pattern to improve students' academic performance.

Ali and Kerem studied the dataset of students of Istanbul EyupI. M.K. B. Vocational Commerce High School and found the relationship between the student performance and course. In their finding they have generated a rule that shows if a candidate is unsuccessful in numerical course in 9th class then those students are likely to be unsuccessful in 10th class. Such results were generated for different courses. This study can facilitate students to choose their appropriate profession by revealing the relation between their concern fields. [1]

R singh and Tiwari et al., conducted a study on engineering students to evaluate their performance by applying data mining techniques to assist them in decision making. They used K-Means algorithm to cluster students. The result predicted that if students are poor in attendance and assignment then there is 75% probability that their grades are poor. [2]

Sen and Ucar analyzed the achievements of students of Computer Engineering Department in Karabük University by means of various factors such as age, gender, type of high school graduation and the students studying in distance education or regular education through data mining techniques. They have taken the dataset of 3047 records. In their study they have used NN architecture called multilayer perceptron (MLP) with back propagation type supervised-learning algorithm to produce both classification and regression type prediction models and decision tree for achieving the highest possible prediction accuracy. The results revealed that as the age of the student increases the success score decreases and students success rate is much better in distance than in formal education, students coming from vocational high school are more successful in cultural lessons than those taking vocational lesson. [3]

Baradwaj and Pal have discussed methods to achieve high quality in higher education. They used various data mining algorithms including different classification algorithm to estimate the accuracy of data. Clustering algorithm was used to cluster the objects which are used as preprocessing approach for attributes. An association rule identifies the correlation between frequent item set with confidence value less than one. Neural Network was used to derive patterns from complicated or imprecise data. The case study identifies the weak students which needsmore attention than others[4].

Ramaswami and Bhaskaran developed a predictive data mining model to identify academically weak students and attributes that affect their performance using CHAID prediction model. The attributes were selected on the basis of chi-square values. If chi-square values of attributes are greater than 100 they are given due considerations and consider the highly influencing variables with high chi-square values. [5]

The SAP prediction of Introductory Engineering Course is done to understand and identify the students' level of performance. For example, if the result of the prediction shows there are some students that will perform poorly in the course, so the lecturers can take appropriate action to help those students. The additional exercise, assignment, or lesson given by lecturers may help the students to improve their understanding in subject taken [6].

The study is also conducted in Malaysia using students' data taken from University Malaysia Pahang (UMP) database management system. The 1000 student records with three courses in the faculty of Computer System and Software Engineering, UMP are considered which contained students' personal, academic, and course information. The students' grade is selected as a predictor parameter and was divided into five categories which are excellent, very good, good, average, and poor. The result indicated that the proposed model is suitable to be used as an SAP prediction [7].

The students' information such as exam scores, grades of team work, attendance, and practical exams are used for profiling and grouping the SAP using selected DM algorithms. The output from analysis process will help the institution to predict academic trends and patterns by categorizing the students into good,

satisfactory, or poor group. It allows the lecturers to get a better understanding about students' learning styles and behaviors [8].

The study involving first year students of school engineering at the National Autonomous University of Mexico (UNAM) is conducted using students' socio-demographic and previous academic information. The data were divided into three categories; students who passed none or up to two courses (low group), students who passed three or four courses (middle group), and students who passed all five courses (high group). The extract patterns from the experiment will allow the IHL to predict academic performance of the new students so that the lecturers will know the level of the new students' preparedness at admission [9].

In our research we have studied the dataset of 600 engineering students to predict their dropout risk. In our work we have proposed that some selected attribute are more influencing for student's academic performance and generated association rules.

III. ASSOCIATION RULE MINING

A formal statement of the association rule problem is [4]:

Definition 1: Let I ={I1, I2, … , Im} be a set of m distinct attributes, also called literals. Let D be a database, where each record (tuple) T has a unique identifier, and contains a set of items such that $T \subseteq I$ An association rule is an implication of the form $X \Rightarrow Y$, where $X, Y \subseteq I$, are sets of items called item sets, and $X \cap Y = \phi$. Here, X is called antecedent, and Y consequent.

Support (s) and confidence (α) are used as two important measures for association rules. They can be defined as follows.

Definition 2: The support (s) of an association rule is the ratio (in percent) of the records that contain XUY to the total number of records in the database.

Definition 3: For a given number of records, confidence (α) is the ratio (in percent) of the number of records that contain XUY to the number of records that contain X.

The problem of mining association rules can be decomposed into two sub problems [5] as stated in the following algorithm.

Algorithm Basic:

Input: I, D, s, α

Output: Association rules satisfying s and α

Algorithm:

1. Find all sets of items which occur with a frequency that is greater than or equal to the user-specified threshold support, s.

2. Generate the desired rules using the large item sets, which have user-specified threshold confidence, α.

The first step in above algorithm is to find large or frequent item sets. Item sets other than those are referred as small item sets. Here an item set is a subset of the total set of items of interest from the database. In association mining, an interesting (and useful) observation about large item sets is, if an item set X is small, any superset of X is also small. Of course the contrapositive of this statement (If X is a large item set than so is any subset of X) is also important to remember. Here L is used to designate the set of large item sets. The second step in above algorithm finds association rules using large item sets obtained in the first step.

A. Apriori Algorithm

Apriori algorithm [10] is used for finding the frequent itemsets and association rule mining. There are two major steps in Apriori algorithm; join and prune.

The new candidate set is generated in the join step. Depending on the support count, the candidate set can be defined as frequent or infrequent. Higher level candidate item sets (Ci) are generated from previous level frequent item sets Li-1 by the method join. The pruning step filtered out the infrequent candidate item sets. This step guarantees that every subset of a frequent item set is also frequent. Hence, if the candidate item set contains more infrequent item sets, will be removed from the process of frequent item set and association mining. [11] This process is called pruning.

Apriori Algorithm Basics

Input D, a database of transactions

Min_sup, the minimum threshold support

Output Lk Maximal frequent item sets in D

Ck Set of Candidate k-item sets.

Method:

1. L1 =Frequent items of length 1.

2. For(k=1;Lk!=ϕ;k++) do.

3. Ck+1=candidates generated from Lk.

4. For each transaction t in database D do.

5. Increment the count of all candidates in Ck+1 that are contained in t.

6. Lk+1 =candidates in Ck+1 with minimum support

7. end

8. Return the set Lk as the set of all possible frequent itemsets

The main notation for association rule mining that is used in Apriori algorithm is the following.

1) A_k –item set is a set of k items.

2) The set Ck is a set of candidate k-itemsets that are potentially frequent.

3) The set Lk is a subset of Ck and is the set of k-itemsets that are frequent.

B. FP Growth Algorithm

The FP-Growth Algorithm, proposed by Han in [8], is an efficient and scalable method for mining the complete set of frequent patterns by pattern fragment growth, using an extended prefix-tree structure. It stores compressed and crucial information about frequent patterns named frequent-pattern tree (FP-tree). The FP-Growth Algorithm is an alternative way to find frequent itemsets without using candidate generations, thus improving performance. For so much it uses a divide-and-conquer strategy. The core of this method is the usage of a special data structure named frequent-pattern tree (FP-tree), which retains the itemset association information.

In simple words, this algorithm works as follows: first it compresses the input database creating an FP-tree instance to represent frequent items. After this first step the compressed database is divided into a set of conditional databases, where each one is associated with one frequent pattern. Finally, each such database is mined separately. Using this strategy, the FP-Growth reduces the search costs looking for short patterns recursively and then concatenating them in the long frequent patterns, offering good selectivity.

In large databases, it’s not possible to hold the FP-tree in the main memory. The problem can be solved by first partitioning the database into a set of smaller databases (called projected databases), and then construct an FP-tree from each of these smaller databases.

IV. METHODOLOGY

Any institute can improve success percentage rate by identifying the reasons for dropout student. In present study, information on various parameters was collected through a structured questionnaire using Google form from first year engineering students of North Maharashtra region. 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. In order to achieve the abovementioned objectives the following steps were followed

Data Collection

The data used in this study was prepared through a structured questionnaire using Google form (https://forms.gle/CiuqptPZQSFsxeGe7). The questionnaire has been constructed by considering theoretical and empirical grounds about factors affecting student's performance and causes of dropout. The questionnaire included socio-demographic indicators (Age, Date of birth, Geographical location, Marital status, Parents education, Parents occupation and Annual income), Educational factors (Performance in High school, Senior Secondary School , Location of Schooling, Type of Examination Board, Medium of Study etc.), Parental Attitudes, and Institutional factors, etc. Data of 600 students was collected from first year engineering students. The data format is presented in Table 1.

TABLE I: STUDENT RELATED VARIABLES

VARIABLES	DESCRIPTION	POSSIBLE VALUES
AGE	AGE	{<18, 18-20, >20}
RES	RESIDENCE	{RURAL, URBAN}
FTYPE	FAMILY TYPE	{NUCLEAR, JOINT}
ANN	ANNUAL INCOME	{LOW, MEDIUM, HIGH, VHIGH}
FEDU	FATHER'S EDUCATION	{ILLITERATE, SEC, HSEC, UG,PG}
MEDU	MOTHER'S EDUCATION	{ILLITERATE, SEC, HSEC, UG,PG}
FOCC	FATHER'S OCCUPATION	{GOVT. SERVICE, PVT. SERVICE, BUSINESS, AGRICULTURE, RETRIED, NA}
MOCC	MOTHER'S OCCUPATION	{GOVT. SERVICE, PVT. SERVICE, BUSINESS, AGRICULTURE, RETRIED, NA}
S_LOC	LOCATION OF SCHOOL	{VILLAGE, TOWN CITY}
HSCG	12 TH GRADE	{A=90-100%, B=80-89%, C=70-79%, D=60-69%, E=LESS THAN 60%}
SSCG	10TH GRADE	{A=90-100%, B=80-89%, C=70-79%, D=60-69%, E=LESS THAN 60%}
PAR_CURR	PARTICIPATION IN EXTRACURRICULAR ACTIVITY	{YES, NO}
SELF STUDY	TIME SPARE FOR STUDY	{<1 HR, 2-3 HRS, 4-5 HRS, >6 HRS}

ENTER	AVAILABILITY OF ENTERTAINMENT IN CAMPUS	{EXCELLENT, V.GOOD, GOOD, POOR, V.POOR}
CURR	EXTRACURRICULAR IN COLLEGE	{EXCELLENT, V.GOOD, GOOD, POOR, V.POOR}
CINF	COLLEGE INFRASTRUCTURE	{EXCELLENT, V.GOOD, GOOD, POOR, V.POOR}
CES	COLLEGE EDUCATION SYS.	{EXCELLENT, V.GOOD, GOOD, POOR, V.POOR}
CLIK	LIKE COLLEGE	{YES, NO}
STRESS	ANY TYPE OF STRESS IN FAMILY	{NO, FINANCIAL, ILLNESS, OTHER}
COL_EXPENSES	EXPENSES IN COLLEGE	{OWN_INCOME, LOAN, BOTH}
C_SYLL	SYLLABUS OF COURSE	{V.SATISFACTORY, SATISFACTORY, BALANCED, DIFFICULT, V.DIFFICULT, LENGTHY}
SAT_LEVEL	STUDENT'S SATISFACTION WITH COURSE	{V.SATISFIED, SATISFIED, NOT V.SATISFIED, NOT SATISFIED}
C_ADMITTED	ENROLLED IN COURSE	{COMPUTER,E&TC,MECHANICAL,CIVI}
A_TYPE	ADMISSION TYPE	{ENTRANCE, MERIT, MANAGEMENT}
MED	MEDIUM OF SCHOOLING	{HINDI, ENGLISH}
PLAC	PLACEMENT STATUS	{BELOW AVG, AVG, GOOD, V.GOOD, EXCELLENT}

Before the initial visit to review the records, a coding system was created for eachvariable to be documented (e.g., rural=0, urban=1).

A. Data Preprocessing

Before application of prescribed model, data preprocessing was applied to measure the quality and suitability of data. In this step, only required and needed attributes for mining of data are chosen. For this, remove missing values; smoothing noisy data, selection of relevant attribute from database or removing irrelevant attributes, identifying or remove outlier values from data set, and resolving inconsistencies of data. The final dataset used for the study contains 60 instances. The study is restricted to the engineering undergraduate students. Finally, the pre-processed data were transformed into a suitable format to apply data mining techniques. Under data set, the reason provided by the students for dropping out of the engineering courses at institute level were divided into five factors such as family problem, health related, personal problem, financial problem and institutional problem.

B. Association Rule Analysis

In EDM, association rule learning is a conventional and well researched method for determining interesting relations between attributes in large databases. Association rule Mining aims to identify strong rules from databases using support and confidence measures.

In this study data was accumulated from the dropout students. These data are analyzed using Association Rule Mining to find out the causes or factors behind dropout. Under data set, the reason provided by the students for dropping out of the engineering courses at institute level were divided into five factors such as family problem, health related, personal problem, financial problem and institutional problem. Overall association rule mining technique was applied to find out the relationship between two different factors affecting the student dropout at the college. From the above analysis it can be conclude that the students who have Personal problem are more prone to dropouts in comparison to Family problem and Institutional problem.

V. EXPERIMENTAL RESULTS AND DISCUSSIONS

From the experimental result, it is found that using Apriori algorithm minimal rules are obtained. The pattern extracted using Apriori algorithm are found more effective in predicting the student drop out risk under three categories: high, medium, low. The parameters used for Apriori algorithm are minimum support, minimum confidence. The importance of the rule is measured using the support value.

In order to conduct the experiment, dataset of 600 students of engineering colleges from North Maharashtra region is collected. For experimentation interesting attributes are selected using number of association rules for different confidence values.

The analysis for generated association rules is as follows:

Final Rules:

Rule #1: 3 --> 1

Support = 0.36667

Confidence = 0.78571

Lift = 1.0476

Rule #2: [3 4] --> 1

Support = 0.11667

Confidence = 0.77778

Lift = 1.037

RULES GENERATED FOR 78% CONFIDENCE AND 0.3 SUPPORTS ARE:
IF Personal Problems THEN Student Dropout Risk = HIGH

RULES FOR CONFIDENCE 77% CONFIDENCE AND 0.1 SUPPORTS ARE:
IF Personal Problems AND Financial Problems THEN Student Dropout Risk = HIGH

The association rules for different confidence values can be interpreted in a way that the students' drop out risk will be high in unit test if either their attendance is poor or assignment is poor or both. So we can interpret that the student's dropout risk is high with personal problems and financial problems as compared to other factors.

The result shows that if a student is suffering from personal problems and financial problems then there are chances that he/she will perform low examinations. This will result in poor performance in institute result. So to improve the student's performance the factors can be considered.

VI. CONCLUSION

Education Data Mining is a promising area of research which has an imperative influence on prediction of students' academic performance. In this paper, student's drop out risk is evaluated using association rule mining algorithm. Research has been done on predicting students drop out risk based on various attributes. In our study important rules are generated using apriori algorithm to measure the correlation among various attributes which will help to predict students drop out risk earlier and improve the student's academic performance. The collected data is categorized in five major factors for association rule analysis. The results showed that the personal problems are more prone to dropouts in comparison to others. This study will help the student's to reduce students dropout risk, to identify those students which needed special attention to reduce failing ration and taking appropriate action at right time. The study can further motivate and help institute to perform data mining tasks on their students' data regularly to find out interesting results and patterns which can help both the university as well as the students in many ways.

REFERENCES

[1] Ali Buldua, Kerem Üçgün,. Data mining application on students' data. Procedia Social and Behavioral Sciences 2 5251–5259, 2010.

[2] Singh, Randhir. An Empirical Study of Applications of Data Mining Techniques for Predicting Student Performance in Higher Education, 2013.

[3] BahaSen, Emine Ucar. Evaluating the achievements of computer engineering department of distance education students with data mining methods. Procedia Technology 1 262 – 267, 2012.

[4] Baradwaj, Brijesh Kumar, and Saurabh Pal. Mining Educational Data to Analyze Students' Performance. Arxiv preprint arxiv: 1201.3417, 2012.

[5] Ramaswami, M., and R. Bhaskaran. A CHAID based performance prediction model in educational data mining. Arxiv preprint arxiv: 1002.1144, 2010.

[6] [6] S. Huang, & N. Fang, Work in Progress - Prediction of Students' Academic Performance in an Introductory Engineering Course, In 41st ASEE/IEEE Frontiers in Education Conference, (2011), 11–13.http://dx.doi.org/10.1109/fie.2011.6142729

[7] S. Sembiring, M. Zarlis, D. Hartama, & E. Wani, Prediction of student academic performance by an application of data mining techniques, 2011 International Conference on Management and Artificial Intelligence, 6 (2011). 110–114.

[8] S. Parack, Z. Zahid, & F. Merchant, Application of data mining in educational databases for predicting academic trends and patterns, 2012 IEEE International Conference on Technology Enhanced Education (ICTEE), (2012), 1–4. http://dx.doi.org/10.1109/ictee.2012.6208617

[9] E. P. I. García, & P. M. Mora, Model Prediction of Academic Performance for First Year Students, 2011 10th Mexican International Conference on Artificial Intelligence, (2011), 169–174. http://dx.doi.org/10

[10] Mohammed Al-Maolegi, Bassam Arkok, An Improved Apriori Algorithm for Association Rules, International Journal on Natural Language Computing (IJNLC) Vol. 3, No.1, February 2014.

[11] J. Han, H. Pei, and Y. Yin, Mining Frequent Patterns without Candidate Generation. In: Proc. Conf. on the Management of Data (SIGMOD'00, Dallas, TX). ACM Press, New York, NY, USA 2000.