INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Students Performance Prediction using Data Mining Techniques

Ashish Narvekar[1], Valerie Menezes[2], Omkar Pai Angle[3], Manasi Lotlekar[4], Saeel Naik[5], Abhimanyu Purohit[6]
[1, 2]*Assistant Professor,* [3, 4, 5, 6]*Department Of Computer Engineering, Agnel Institute Of Technology &Design, Assagao, Goa*

Abstract: The paper represents the data mining techniques used for predicting student's performance. In today's world the education field is growing, developing widely and becoming one of the most crucial industries. The data available in the educational field can be studied using educational data mining so that the unseen knowledge can be obtained from it. In this paper, various Data Mining approaches like Association mining and classification are used to predict the students' performance in examination in advance, so that necessary measures can be taken to improvise on their performance to score better marks. The results obtained after the implementation may be useful for instructor as well as students. This work will help in taking appropriate decision to improve student's performance.
Keywords: Data Mining, Data Discretization, Classification, Association Mining, Backward Elimination.

I. INTRODUCTION

Data mining is a procedure of analyzing the given data collection from different perspective and discovering the information and knowledge from it, by summarizing it. The outcomes provide the co-relations or patterns among the data present inside the data set. When the data of data set is originated from the education environment, the mining done on such data set is termed as education data mining. Nowadays, education data mining is done on a high scale as education has become remarkable for the personal and economic growth of an individual. The purpose of this work is to predict the future grades of students so that the student can analyse which factors are affecting his/her approach to scoring well in exams and plan in advance in order to do better in the exams.

Using the outcome of the analysis, prediction for the coming final semester grades is made so that the failure rate can be minimized by giving extra attention towards the weak students. This further enhances the overall result of the educational institute.

The scope of the proposed paper is to develop a system that predicts the future performance of students using the domains of Data Mining. This work can be used in schools, colleges and such institutions where they would like to increase the passing percentage of students, decrease dropout ratio or would just like to improve their academic performance. Further this analysis can be effectively utilized to recommend the decision makers of the education institute. Thus, aiming at improving the educational field.

Section II presents the study of the existing systems, their working and limitations. Section III includes the architecture of the proposed system and the various scenarios in which it will function.

II. PROPOSED SYSTEM

Initially, data is collected from the students. The data which is collected is discretized. i.e:- the Attribute values are divided into two categories: High and Low based on range. Now, Discretized data is given to classification algorithm (ID3) to form a decision tree and same data is given to Apriori algorithm to form strong rules. The strong rules are then given to Rule based classifier which performs pruning on it to improve generalised errors. These two algorithms are compared to check which algorithm is more efficient.
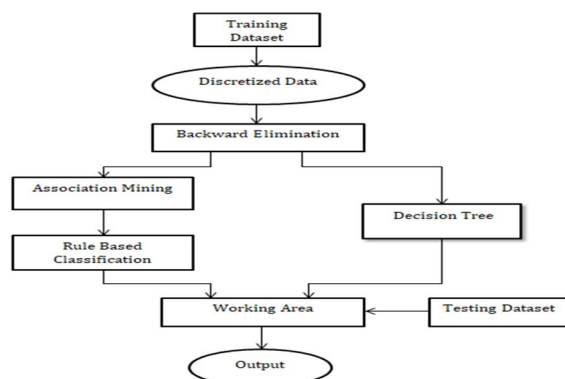
A. Block Diagram

Fig.1 Block Diagram of Proposed Diagram

B. Applications Of Proposed System
1) *Early Prediction:* We proposed a model which predicts the student's performance in any upcoming exam based on their previous marks as well as some lifestyle parameters such as study time, travel time, time spent on internet, etc.
2) *Efficiency:* A cost efficient method that can handle a large number of records in a short time. Being a computer-aided method, it is not affected by fatigue and thus does not lose its precision.

III. METHODOLOGY

Models
The list of methods are as follows:

A. Data Discretization
Data discretization is defined as a process of converting continuous data attribute values into a finite set of intervals with minimal loss of information.
1) *Entropy based Discretization:* Entropy based method uses a split approach. The entropy (or the information content) is calculated based on the class label. It finds the best split so that the bins are as pure as possible that is the majority of the values in a bin correspond to have the same class label. Formally, it finds the split based on maximal information gain.[7]

B. Classification based Decision Tree
Classification is a data mining concept that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. The Decision Tree will have a specific number of Predictors which will lead to a specific single attribute Target. The values at the top of the tree will be the best predictors and the quality of predictors will decrease as we go down the decision tree.[5]
1) *ID3 (Iterative Dichotomiser 3) Algorithm:* ID3 (Iterative Dichotomiser 3) is an algorithm used to generate a decision tree from a dataset where each internal node is an attribute which has the highest entropy gain at that level and branches are split based on the categorical value.

C. Association Mining
Association rule learning is a rule-based method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using some measures of interestingness.[6]
1) *Apriori Algorithm:* Apriori algorithm is used to finding frequent itemsets in a dataset for Boolean association rule. It is called Apriori because it uses prior knowledge of frequent itemset properties. We apply an iterative approach or level-wise search where k-frequent itemsets are used to find k+1 itemsets so as to improve the efficiency of level-wise generation of frequent itemsets.
2) *CBA (Classification Based on Associations):* Classification rule mining and association rule mining are two important data mining techniques. Both classification rule mining and association rule mining are combined to form CBA algorithm .CBA algorithm can find some valuable rules that existing classification system cannot, it can handle both table from data and transaction from data and it does not require whole dataset to fetch into main memory .The CBA algorithm does not focus on entire data set it can focus on specific data set. These algorithms basically focus on CBA-RG(Rule Generator) and CBA-CB(Classifier Builder).[9]
3) *Association Rule Base Classifier:* A rule-based classifier is a technique for classifying record using the collection of if-then rules. The model is generated by a rule base classifier for the vertebrate classification problem. The rule for a model are represent in a disjunctive normal form, R=(r1Vr2V . . . .) where R known as rule set. The two important property of rule-base classifier is Mutually Exclusive Rule and Exhaustive Rule.
4) *Indirect method of a Rule Extraction:* In indirect method rule extraction take the input as from other model and apply rule-pruning to obtain the strong rule set with low pessimistic error. The algorithm focus on:-
a) *Rule Generation:* Let given rule r: A->y, we consider a simplified rule, r': A'->y where A' is obtain by removing one of the conjunction in A. The simplified rule with the lowest pessimistic error rate is retained, provided its error rate is less than that of the original rule. The rule pruning step is repeated until the pessimistic error of the rule cannot be improved. And duplicate rule must be discarded.

b) *Technique for rule Pruning :* A rule is pruned by removing a conjunct (attribute test).We chose to prune rule R. If the pruned version of R has a greater quality, assessed on an independent set of tuples . Pruning can be performed by Foil Pruning given as: FOIL_PRUNE(R)= (pos – neg)/(pos + neg) Where, pos and neg stand for number of positive and negative tuples cover by R.

c) *Rule Ordering:* After generating the rule set, order the extracted rule, the rules that predict the same class are grouped together into the same subset .The total description length from each subset is computed, and the classes are arranged in increasing order of the total description length and classes which have the smallest description length is given the highest priority as it is expected to contain the best rule set.

D. Backward Elimination

Backward elimination is a process which selects all the attributes at the beginning and eliminating the attributes one by one until it reaches the certain threshold which is the stopping criterion in the backward elimination process.

IV. DATASET

The following are the attributes of dataset.

Attribute Name	Description
Travelling_Time	Student's travelling time (in minutes) per day
Study_Time	Student's study time(in minutes) before exam
ExtraEducation	Does the student take any extra education(Yes/No)
Family_Edu_sup	Does the family support for education(Yes/No)
Extra_Activities	Does the student participate In any extra activity(Yes/No)
Internet_Use	Use of internet for education(in minutes) per day
Sleep_Time	Student's sleeping hours (in Minutes)
Relationship_Status	Is the student in relationship(Yes/No)
Health_Issue	Health condition of student (Yes/No)
Family_Problems	Does the student have any family problem (Yes/No)
PartTimeJob	Does the student do any part time job(Yes/No)
Attendance	Student's Attendance(in percentage)
Subject_Interest	Student's interest in the subject(Yes/No)
Fac_Teach_Skill	Faculty teaching skills(Good/Bad)
Previous_score	Previous score of the student(in percentage)
Concentration	How the student focuses during lecture(Good/Bad)
Faculty_Com_Skills	Faculty communication skill(Good/Bad)

Fig.2 Attributes of dataset

V. SUMMARY

Taking 15 attributes as the final number of attributes to generate the model with 160 student data records, we get the accuracy for both the models as shown below:

Sr .no	Algorithm	Accuracy
1	ID3 Algorithm	58.33
2	Apriori + CBA algorithm	79.16

Fig.3 Accuracy of models

VI.   CONCLUSION

In the current educational system, there have been different issues related to student's performance evaluation and assessment. In this project, data mining techniques which are decision tree , association rule mining and rule based classification will be experimented in the next semester. The prediction of the final semester grades of the students is generated before the actual final semester exam held. This can help the education institutes to minimize the dropping of the overall result by giving extra attention to the weak students.The results that we will obtain from our project will also declare which method is more efficient: Classification Using Decision Trees vs Association Rule Mining to get rules and then classifying these rules using Rule Based Classification. The information generated is useful for the further evaluation and prediction and can be vividly applied in many other fields.

REFERENCES

[1] Surjeet Kumar Yadav, Brijesh Bharadwaj, Saurabh Pal: "Data Mining Applications: A comparative study for predicting students performance". INTERNATIONAL JOURNAL OF INNOVATIVE TECHNOLOGY & CREATIVE ENGINEERING (ISSN:2045-711) VOL.1 NO.12 DECEMBER.

[2] Surbhi Agrawal, Santosh K Vishwakarma, Akhilesh K Sharma: "Using Data Mining Classifier for predicting Student's Performance in UG level". International Journal of Computer Applications (0975 – 8887) Volume 172 – No.8, August 2017.

[3] R R Kabra, R S Bichkar: "Performance Prediction of Engineering Students using Decision Trees". International Journal of Computer Applications (0975-8887) Volume 36-No.11, December 2011.

[4] Pang-Ning Tan, Michael Steinbach, Vipin Kumar" Introduction to Data Mining".

[5] www.geeksforgeeks.org/basic-concept-classification-data-mining/

[6] www.geeksforgeeks.org/association-rule/

[7] http://www.saedsayad.com/supervised_binning.htm

[8]]https://www.academia.edu/36887916/Using_Data_Mining_Classifier_for_Predicting_Students_Performance_in_UG_level.

[9] https://www.aaai.org/Papers/KDD/1998/KDD98-012.pdf

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)