INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH & DEVELOPMENT

ISSN: 2278 – 0211 (Online)

Data Mining: Performance Improvement In Education Sector Using Classification And Clustering Algorithm

Gadde Shravya Sree

IV/IV B.Tech, Dept. of CSE,VVIT, Nambur , Guntur dist, A.P, India.

Ch.Rupa

Professor, Dept. of CSE, VVIT, Nambur , Guntur dist, A.P, India

Abstract:

The ability to predict a student's performance is very important in educationalenvironments. Students' academic performance is based upon diverse factors like personal, social, Psychological and other environmental variables. A very promising tool to attain this objective is the use of Data Mining. Data mining techniques are used to discover hidden information patterns and relationships of large amount of data, which is very much helpful in decision making. A single data contains a lot of information. The type of information is produced by the data and it decides the processing method of data. A lot of data that can produce valuable information, in education sector contains this valuable information. Which helps the education sector to capture and compile low cost information for this information and communication technology is used. Now-a-days educational database is increased rapidly because of the large amount of data stored in it. The loyal students motivate the higher education systems, to know them well; the best way is by using valid management and processing of the students' database. Data mining approach provides valid information from existing student to manage relationships with upcoming students.

1.Introduction

DATA mining sometimes is also called knowledgediscovery in databases (KDD). We can also find the existing relationships and patterns. Data mining combines machine learning, statistics and visualization techniques to discover and extract knowledge. Student retention has become an indication of academic performance and enrollment management. Here, potential problem will be identified as earlier. The raw data was preprocessed in terms of filling up missing values, transforming values in one form into another and relevant attribute/ variable selection. One of the most useful data mining techniques for e-learning learning is classification. Classification maps data into predefined groups of classes. It is often referred to as supervised learning because the classes are determined before examining the data. The prediction of students' performance with high accuracy is more beneficial for identifying low academic achievements students at the beginning. To improve their performance the teacher will monitor the students' performance carefully. Student retention is an indicator of academic performance and enrollment management of university. To assist the low academic achievers in higher education and they are:

- Generation of data source of predictive variables
- Identification of different factors, which effects a student's learning behavior and performance

during academic career

- Construction of a prediction model using classification data mining techniques on the basis of identified predictive variables
- Validation of the developed model for higher education students studying in Indian Universities or Institutions. Predictive classification enhances the quality of higher education system to increase number of loyal students to evaluate students' data to study the main attributes that may affect the enrollment factor.

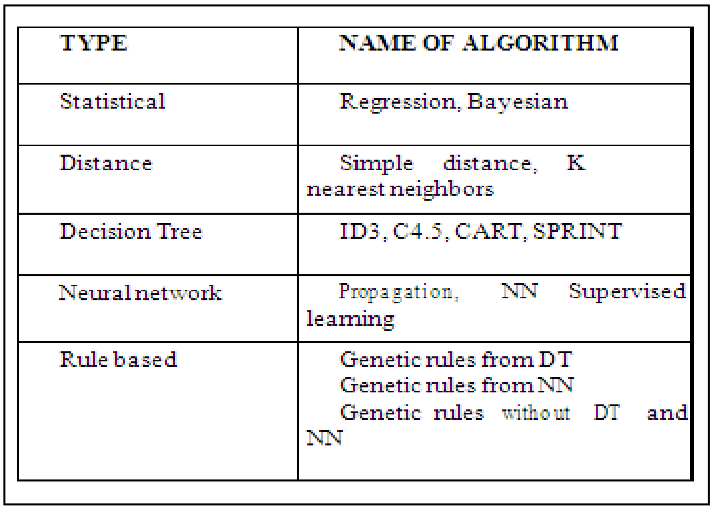
2.Background And Related Work

Data mining software that allow the users to analyze data from different dimensions categorize it and summarize the relationships which are identified during mining process. By means of Bayes Classification on category, language and background qualification, it was found that whether new comer students will performer or not. Data Mining can be used in educational field to enhance our understanding of learning process to focus on identifying, extracting and evaluating variables related to the learning process of students. Mining in educational environment is called Educational Data Mining. Students' attitude towards attendance in class hours spent on daily basis after college, students' family income, students' mother's age and mother's education are significantly related with student performance" was framed. By means of simple linear regression analysis, it was found that the factors like mother’s education and student’s family income were highly correlated with the student academic performance. For universities, data mining techniques could help provide more personalized education, maximize educational system efficiency, and reduce the cost of education processes. It may guide us to increase student’s retention rate, increase educational improvement ratio, and increase student’s learning outcome. Data mining prediction technique to identify the most effective factor to determine a student’s test score, and then adjusting these factors to improve the student’s test score performance. It provides a new way of look into the education which was hidden from humankind. C romero and S vetura made a comprehensive study on the development of this educational data mining since 1995 to 2005. Their paper surveys the application of data mining to traditional education systems, particular web-based course, well known learning content management systems and adaptive and intelligent web-based educational systems. They have concluded from different research papers application of educational system has objective at student. From students orientation its objective is to recommend to learners activities, resources, learning tasks, suggest path pruning etc. From educators point of view its objective is to get more objective feedback, effectiveness on learning process, monitoring, find learners mistake etc. an academic responsible and administrator's objective to use it to improve site, efficiency, better organize institutional resources, enhance educational program etc. their survey work is motivated us to make a study on some research which used data mining technique to find hidden information from educational database.

3.Data Mining Techniques

3.1.Classification

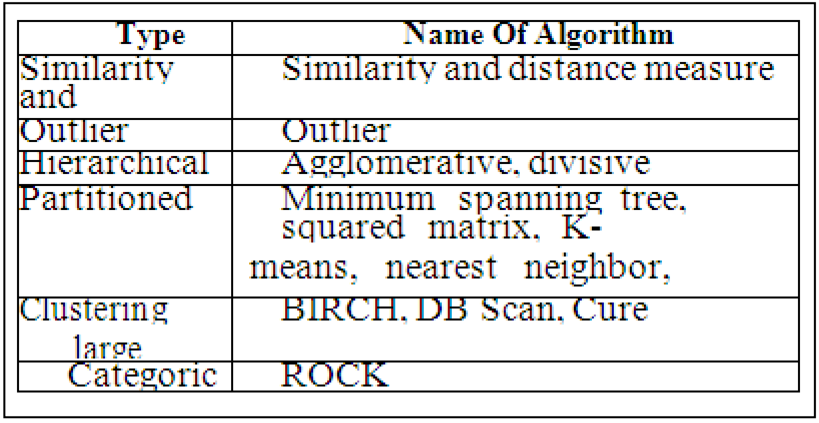
Estimation and prediction are viewed as types of classification. The problem usually is evaluating the training data set and second applied the model developed.



TYPE	NAME OF ALGORITHM
Statistical	Regression, Bayesian
Distance	Simple distance, K nearest neighbors
Decision Tree	ID3, C4.5, CART, SPRINT
Neural network	Propagation, NN Supervised learning
Rule based	Genetic rules from DT Genetic rules from NN Genetic rules without DT and NN

Table 1: Classification Algorithm

Clustering groups the data, this is not predefined. By using this technique we can identify dense and sparse regions in object space. The following table provides the different clustering techniques. Clustering algorithm is best for grouping the data.



Type	Name Of Algorithm
Similarity and	Similarity and distance measure
Outlier	Outlier
Hierarchical	Agglomerative, divisive
Partitioned	Minimum spanning tree, squared matrix, K-means, nearest neighbor,
Clustering large	BIRCH, DB Scan, Cure
Categoric	ROCK

Table 2: Clustering Algorithm

4.Association

The main task of this association rule mining is to find set of binary variables that frequently occurs in the transaction database. The goal of feature selection problem is to identify groups, which is correlated with each one of the target variable. Apriori, CDA, DDA, invitingness measure etc are the association rule mining algorithm.

5.Data Mining Process

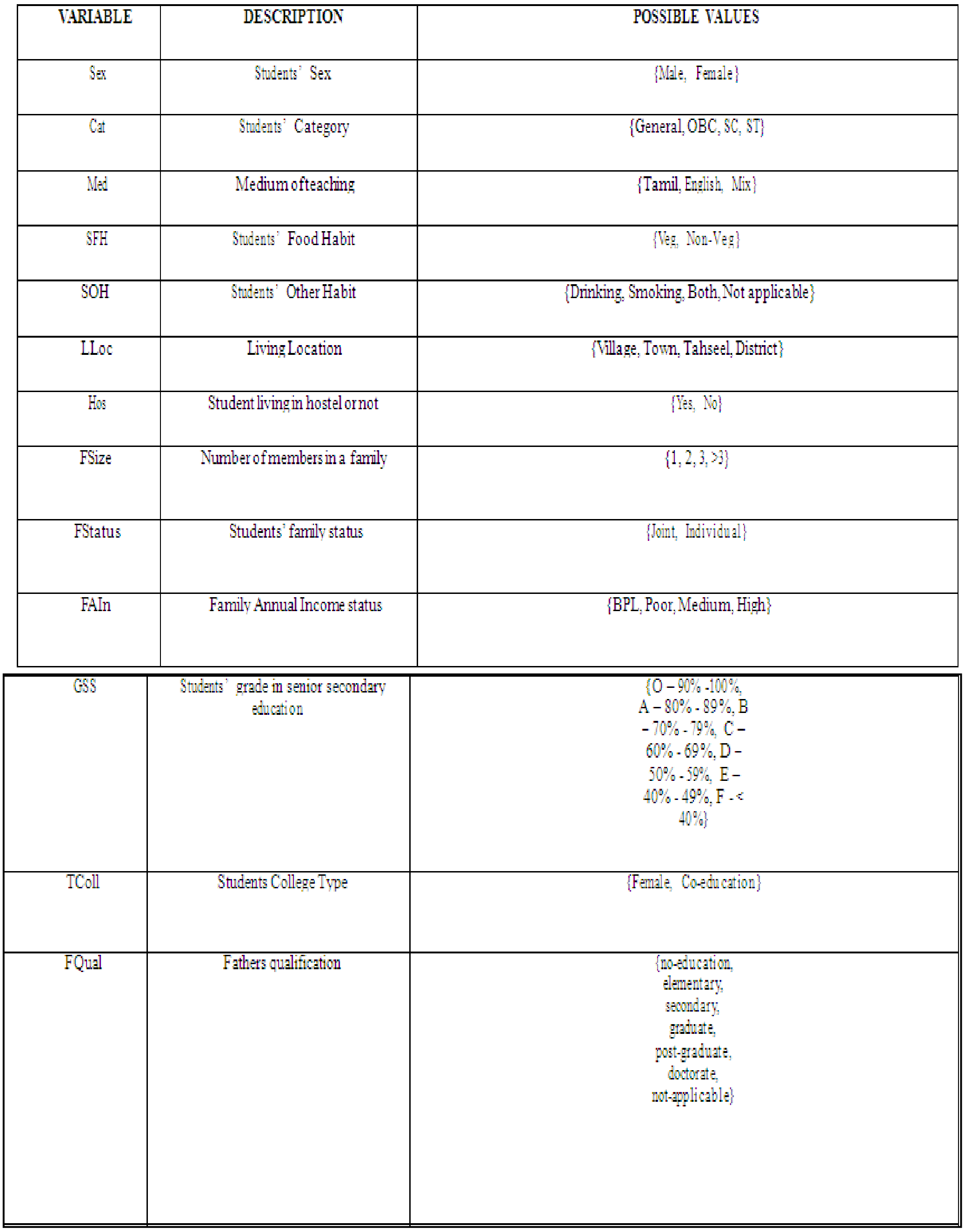
Data are analyzed using classification method to predict the student's performance.

5.1.Data Preperation

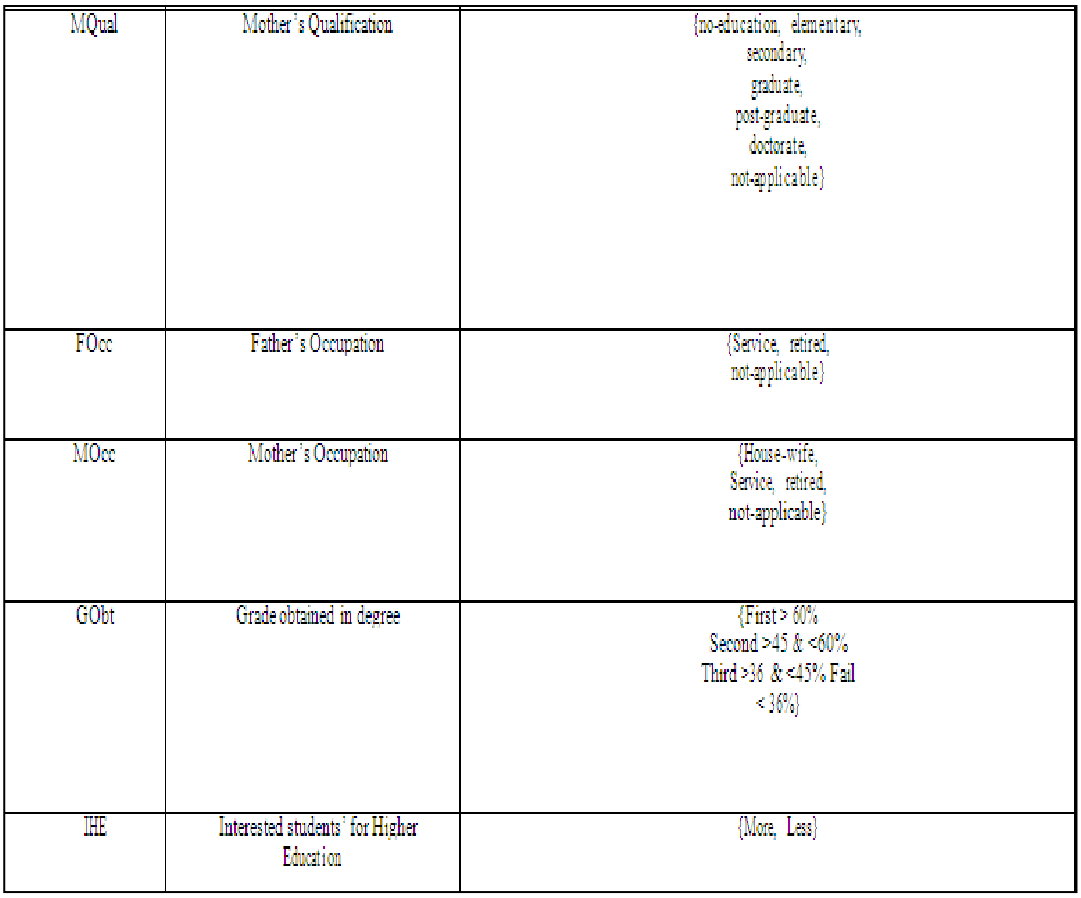
The data set used here, which is obtained from various colleges. Data stored in different tables was joined in a single table after joining process errors were removed.

5.2.Data Selection And Transformation

Fields are selected, that is required for data mining. A few derived variables were selected. While some of the information for the variables was extracted from the database. All the predictor and response variables which were derived from the database are given in Table,



VARIABLE	DESCRIPTION	POSSIBLE VALUES
Sex	Students' Sex	{Male, Female}
Cat	Students' Category	{General, OBC, SC, ST}
Med	Medium ofteaching	{Tamil, English, Mix}
SFH	Students' Food Habit	{Veg, Non-Veg}
SOH	Students' Other Habit	{Drinking, Smoking, Both, Not applicable}
LLoc	Living Location	{Village, Town, Tahseel, District}
Hos	Student living in hostel or not	{Yes, No}
FSize	Number of members in a family	{1, 2, 3, >3}
FStatus	Students' family status	{Joint, Individual}
FAIn	Family Annual Income status	{BPL, Poor, Medium, High}
GSS	Students' grade in senior secondary education	{O – 90% -100%, A – 80% - 89%, B – 70% - 79%, C – 60% - 69%, D – 50% - 59%, E – 40% - 49%, F - < 40%}
TColl	Students College Type	{Female, Co-education}
FQual	Fathers qualification	{no-education, elementary, secondary, graduate, post-graduate, doctorate, not-applicable}



MQual	Mother's Qualification	{no-education, elementary, secondary, graduate, post-graduate, doctorate, not-applicable}
FOcc	Father's Occupation	{Service, retired, not-applicable}
MOcc	Mother's Occupation	{House-wife, Service, retired, not-applicable}
GObt	Grade obtained in degree	{First > 60% Second >45 & <60% Third >36 & <45% Fail < 36%}
IHE	Interested students' for Higher Education	{More, Less}

Table 3: Student Related Variable

The domain values for some of the variables were defined for the present investigation as follows:

- Cat – From ancient time Indians are divided in many categories. These factors play a direct and indirect role in the daily lives including the education of young people. Admission process in India also includes different percentage of seats reserved for different categories. In terms of social status, the Indian population is grouped into four categories: General, Other Backward Class (OBC), Scheduled Castes (SC) and Scheduled Tribes (ST). Possible values are General, OBC, SC and ST.
- Med – This paper study covers the schools, degree colleges and institutions of Tamil Nadu state of India. Here, medium of instructions are Tamil or English or Mix (Both Tamil and English).
- SOH – In modern society bad habits are increasing fast among college students. Here students other habit include Drinking, Smoking, Both or Not-applicable.
- FSize-. According to population statistics of India, the average number of children in a family is 3.1. Therefore, the maximum family size is fixed as 10 and possible range of values is from one to ten
- GSS - Students grade in Senior Secondary education. Students who are in state board appear for five subjects each carry 100 marks. Grade are assigned to all students using following mapping O – 90% to 100%, A – 80%- 89%, B –70% - 79%, C – 60% - 69%, D –50% - 59%, E – 40% - 49%, and F -40%}.
- GObt - Marks/Grade obtained in UG course and it is declared as response variable. It is also split into five class values: First – >60%, Second – >45% and <60%, Third–>36% and < 45%, Fail < 40%.

6.Mining Model

Various algorithms and techniques like Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbor method etc., are used for knowledge discovery from databases. Classification is one of the most frequently studied problems by data mining and machine learning (ML) researchers. It consists of predicting the value of a (categorical) attribute (the class) based on the values of other attributes (the predicting attributes). There are different classification methods. In the present study we use the Bayesian Classification algorithm. Bayes classification has been proposed that is based on Bayes rule of conditional probability. Bayes rule is a technique to estimate the likelihood of a property given the set of data as evidence or input Bayes rule or Bayes theorem is- The approach is called "naïve" because it assumes the independence between the various attribute values. Naïve Bayes classification can be viewed as both a descriptive and a predictive type of algorithm. The probabilities are descriptive and are then used to predict the class membership for a target tuple. The naïve Bayes approach has several advantages: it is easy to use; unlike other classification approaches only one scan of the training data is required; easily handle mining value by simply omitting that probability. An advantage of the naive Bayes classifier is that it requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification. Because independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire covariance matrix. In spite of their naive design and apparently over-simplified assumptions, naive Bayes classifiers have worked quite well in many complex real-world situations.

7.Conclusion

Using this research, the University will have the ability to predict the students' loyalty (numbers of enrolled students) so they can manage and prepare necessary resources for the new enrolled students. This helps the teacher to improve the performance of students' performance and those students needed special attention for reducing falling ratio for taking action at right time. Bayesian classification method is used on student database to predict the students division on the basis of previous year database. Data mining is a powerful analytical tool that enables educational institutions to better allocate resources and staff, and proactively manage student outcomes. The management system can improve their policy, enhance their strategies and thereby improve the quality of that management system. Student performance in university courses is of great concern to the higher education managements where several factors may affect the performance. Data mining extracts hidden information with the help different mining technique. Prediction, results and recommendation are provided by this information, which help the user to take further decision. It also guides the concern person for whom information has been extracted.

8.Reference

1. Umesh Kumar Pandey, Brijesh Kumar Bhardwaj, Saurabh pal "Data Mining as a Torch Bearer in Education Sector" International Journal of Advanced Research in Computer Science and Software Engineering Volume 2, Issue 1, January 2012
2. Brijesh Kumar Bhardwaj, Saurabh Pal "Data Mining: A prediction for performance improvement using classification" (IJCSIS) International Journal of Computer Science and Information Security, Vol. 9, No. 4, April 2011
3. Boumedyen Shannaq, Yusupov Rafael, V. Alexandro "Student Relationship in Higher Education Using Data Mining Techniques" Vol. 10 Issue 11 (Ver. 1.0) October 2010
4. Ying Zhang, Samia Oussena, Tony Clark, Hyeonsook Kim "Use Data Mining to Improve Student Retention in Higher Education – a case study"
5. Qasem A. Al-Radaideh, Emad M. Al-Shawakfa, and Mustafa I. Al-Najjar "Mining Student Data Using Decision Trees" The 2006 International Arab Conference on Information Technology (ACIT'2006)
6. Dr Robert Jones "Student retention and success: a synthesis of Research" April 2008.