Volume 6, Issue 3, March 2016 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Student Performance Prediction in Education Sector Using Data Mining

Shruthi P[*]
M. Tech in Computer Engineering,
PES College of Engineering, Mandya,
Karnataka, India

Chaitra B P
Asst. Professor, Dept of CS&E
PES College of Engineering, Mandya
Karnataka, India

Abstract— ***The data in education sector is increasing periodically which can be used to predict the performance of the students in the upcoming semesters. From this the students who are at the risk of failure can be identified and proper guidance can be given to those students for their better future results. The educational institute will also be benefited by improving their overall percentage of results. The very promising tool for this is data mining. In this paper, the performance of the students is predicted using the behaviours and results of previous passed out students stored in the database and by using the behaviours of the present students. The data mining technique called as classification rule is used to predict the performance of the present students. Many classification algorithms exist like Decision Tree, Neural Networks, K-Nearest Neighbour, Naïve Bayes etc. In this paper Naïve Bayes classification algorithm is used as it has highest accuracy compared to other classification algorithms.***

Keywords — Prediction, Behaviours, Classification rule, Naïve Bayes algorithm.

I. INTRODUCTION

Preliminary education adds to a nation's literacy rate but higher education has a direct impact on the work force being provided to the industry and hence directly affects the economy. Lots of Institutions of higher learning have been set up across India. However the quality of education is judged by the success rate of student's and to what extent an institute is capable of retaining its students. Predicting student's performance can help identify the students who are at risk of failure and thus management can provide timely help and take essential steps to coach the students to improve his performance.

The ability to predict a student's performance is very important in educational environments. 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 student's database. Data mining approach provides valid information from existing student to manage relationships with upcoming students.

The problem statement can be defined as:

- Identification of different factors which affects a student's learning behaviour and performance during academic career.
- Construction of a prediction model using classification data mining technique on the bases of identified predictive variables.

II. LITERATURE SURVEY

Data Mining (sometimes called data or knowledge discovery) is the process of analysing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analysing data. It allows users to analyse data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

Following are the survey papers being studied:

[1] Mining Student's Data for Performance Prediction

Tripti Mishra, Dr. Dharminder Kumar, Dr. Sangeeta Gupta

A country's growth is strongly measured by quality of its education system. Education sector, across the globe has witnessed sea change in its functioning. Today it is recognized as an industry and like any other industry it is facing challenges, the major challenges of higher education being decrease in student's success rate and their leaving a course without completion. An early prediction of student's failure may help the management provide timely counselling as well coaching to increase success rate and student retention. This paper uses different classification techniques to build performance prediction model based on student's social integration, academic integration, and various emotional skills which have not been considered so far.

[2] Performance Prediction of Students Using Distributed Data Mining
Prof. Dineshkumar Vaghela, Dr Priyanka Sharma, Krina Parmar

The performance of students in higher education in India is a turning point in the academics for all students for their brightest career. In today's generation the amount of data stored in educational database increasing at a great rate. These databases contain secret information for improvement of student's performance; these data can be located at different nodes in distributed system. Classification and prediction are among the major techniques in Data mining and widely used in various fields. In this paper classification techniques are used for prediction of student performance in distributed environment. Data mining methods are often implemented at many advance universities today for analysing available data and extracting information and knowledge to support decision making. While it is important to have models at local level, their results makes it difficult to extract knowledge that can be useful at the global level. Therefore, to support decision making at this area, it is important to generalize the information contained in those models, specific classifier method can be used to generalize these rules for global model.

[3] Student's Performance Evaluation in Online Education System Vs Traditional Education System
Udeni Jayasinghe, Anuja Dharmaratne, Ajantha Atukorale

Nowadays most of the education institutes practice online teaching mechanism rather than using the traditional teacher centred teaching mechanism to enhance the learning ability of the students by making a student centred learning mechanism. The teachers have to evaluate the student's performance no matter what their learning mechanism is and there are ways of doing that i.e. formal evaluation and informal evaluation. That means by giving exam papers and giving feedbacks based on the student's behaviour. When it comes to online education mechanism, there are no teachers to monitor the behaviour of the students. Now researchers have come up with various ideas to measure the emotional level and behaviour of the students and to respond according to each and every student through the online education systems. The aim of this article is to review some literature and to do a comparison among the various proposed ideas of researchers whose aim was to evaluate the performance of the students who are engaged in online education systems.

[4] Mining Educational Data to Analyze Student's Performance
Brijesh Kumar Baradwaj, Saurabh Pal

The main objective of higher education institutions is to provide quality education to its students. One way to achieve highest level of quality in higher education system is by discovering knowledge for prediction regarding enrolment of students in a particular course, alienation of traditional classroom teaching model, detection of unfair means used in online examination, detection of abnormal values in the result sheets of the students, prediction about students' performance and so on. The knowledge is hidden among the educational data set and it is extractable through data mining techniques. Present paper is designed to justify the capabilities of data mining techniques in context of higher education by offering a data mining model for higher education system in the university. In this research, the classification task is used to evaluate student's performance and as there are many approaches that are used for data classification, the decision tree method is used here.

By this task knowledge is extracted that describes student's performance in end semester examination. It helps earlier in identifying the dropouts and students who need special attention and allow the teacher to provide appropriate advising/counseling.

[5] Data Mining: A prediction for Student's Performance Using Classification Method
Abeer Badr El Din Ahmed, Ibrahim Sayed Elaraby

Currently the amount huge of data stored in educational database these database contain the useful information for predict of students performance. The most useful data mining techniques in educational database is classification. In this paper, the classification task is used to predict the final grade of students and as there are many approaches that are used for data classification, the decision tree (ID3) method is used here.

III. PHASES OF DATA MINING

Data Mining, also popularly known as Knowledge Discovery in Databases (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process. The following figure Fig 1 shows different phases of data mining.

The Knowledge Discovery in Databases process comprises of a few steps leading from raw data collections to some form of new knowledge. The iterative process consists of the following steps:

- **Data cleaning**: also known as data cleansing, it is a phase in which noise data and irrelevant data are removed from the collection.
- **Data integration**: at this stage, multiple data sources, often heterogeneous, may be combined in a common source.

- **Data selection**: at this step, the data relevant to the analysis is decided on and retrieved from the data collection.
- **Data transformation**: also known as data consolidation, it is a phase in which the selected data is transformed into forms appropriate for the mining procedure.
- **Data mining**: it is the crucial step in which clever techniques are applied to extract patterns potentially useful.
- **Pattern evaluation**: in this step, strictly interesting patterns representing knowledge are identified based on given measures.
- **Knowledge representation**: is the final phase in which the discovered knowledge is visually represented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results.

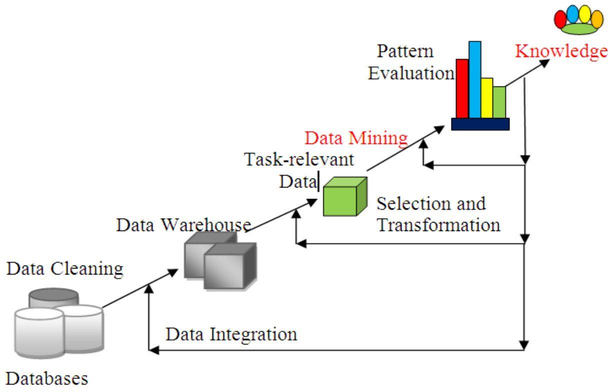


Fig 1 Phases of Data Mining

It is common to combine some of these steps together. For instance, data cleaning and data integration can be performed together as a pre-processing phase to generate a data warehouse. Data selection and data transformation can also be combined where the consolidation of the data is the result of the selection, or, as for the case of data warehouses, the selection is done on transformed data.

The KDD is an iterative process. Once the discovered knowledge is presented to the user, the evaluation measures can be enhanced, the mining can be further refined, new data can be selected or further transformed, or new data sources can be integrated, in order to get different, more appropriate results.

Data mining derives its name from the similarities between searching for valuable information in a large database and mining rocks for a vein of valuable ore. Both imply either sifting through a large amount of material or ingeniously probing the material to exactly pinpoint where the values reside. It is, however, a misnomer, since mining for gold in rocks is usually called "gold mining" and not "rock mining", thus by analogy, data mining should have been called "knowledge mining" instead. Nevertheless, data mining became the accepted customary term, and very rapidly a trend that even overshadowed more general terms such as knowledge discovery in databases (KDD) that describe a more complete process. Other similar terms referring to data mining are: data dredging, knowledge extraction and pattern discovery.

IV. DATA MINING TECHNIQUES

The data mining techniques are:

- Association rule
- Classification rule
- Clustering techniques
- Sequential and Pattern prediction

A. Association rule

Association rule is also called as pattern discovery. The pattern is discovered based on the relationship between a particular item with other item in the same transaction.

Example: Suppose that, the marketing manager at AllElectronics want to know which items are frequently purchased together (i.e., within the same transaction). An example of such a rule, mined from the AllElectronics transactional database, is

buys(X,"computer") => buys(X,"software") [support=1%, confidence=50%]

Where X is a variable representing a customer. A confidence, or certainty, of 50% means that if a customer buys a computer, there is 50% chance that she will buy software as well. A 1% support means that 1% of all the transaction under analysis show that computer and software are purchased together. This association rule involves a single attribute or predicate (i.e., buys) that repeats. Association rules that contain a single predicate are referred to as single-dimensional association rules. Dropping the predicate notation, the rule can be written simply as "computer => software [1%, 50%]."

Suppose, instead, given the AllElectronics relational database releted to purchase. A data mining system may find association rules like.

age(X,"20..29) ^ income(X,"40K..49K) => buys(X,"laptop")
[support=2%, confidence=60%]

The rule indicates that of the AllElectronics customers under study, 2% are 20 to 29 years old wih an income of $40000 to $49000 and have purchased a laptop at AllElectronics. There is a 60% probability that a customer in this age and income group will purchase a laptop. Note that this association rule involving more than one attribute or predicate (i.e., age, income, and buys). This is called ad multidimensional association rule.

B. Classification Rule

Classification is a process of finding a model (or function) that describes and distinguishes data classes or concepts. The model is derived based on the analysis of a set of training data (i.e., data objects for which the class labels are known). The model is used to predict the class label of objects for which the class label is unknown.

Example: Suppose the sales manager of AllElectronics want to classify a large set of items in the store, based on three kinds of responses to sales campaign: good response, mild response and no response. The model for each of these three classes is derived based on the descriptive features of the items, such as price, brand, place_made, type and category. The resulting classification should maximally distinguish each class from the others, presenting an organized picture of the data set

C. Clustering Techniques

The objects are clustered or grouped on the principle of maximizing the intraclass similarity and minimizing the interclass similarity. That is, clusters of objects are formed so that objects within a cluster have high similarity in comparison to one another, but are rather dissimilar to objects in other clusters.

Example: Cluster analysis can be performed on AllElectronics customer data to identify homogeneous subpopulations of customers. These clusters may represent individual target groups for marketing.

D. Sequential and Pattern Prediction

Frequent patters, as the name suggest, are patters that occur frequently in data. There are many kinds of frequent patters, including frequent itemsets, frequent subsequences (also known as sequential patterns), and frequent substructures. A frequent itemset typically refers to a set of items that often appear together in a transactional data set. For example, milk and bread, which are frequently bought together in grocery stores by many customers. A frequently occurring subsequence, such as the pattern that customers, tend to purchase first a laptop, followed by a digital camera and then a memory card, is a sequential pattern.

V. DATA MINING PROCESS

In present day's educational system, a student's performance is determined by the internal assessment and end semester examination. The internal assessment is carried out by the teacher based upon student's performance in educational activities such as class test, seminar, assignments, general proficiency, attendance and lab work. The end semester examination is one that is scored by the student in semester examination. Each student has to get minimum marks to pass a semester in internal as well as end semester examination.

A. Data Preparation

The data set used in this study is obtained from PES College of Engineering, Mandya. The result of previous students of different branches is collected from the college database and their behaviours are collected from their respective faculties and students. It is stored in other database and it is used to predict the performance of the present students in their upcoming semesters.

B. Data Selection and Transformation

In this step only those fields were selected which were 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 I for reference.

In this paper 18 different behaviours of the students are considered which is used to predict the performance of the present students in their upcoming semester. The complete description of the different behaviours with their possible values and variables are as as shown in below Table I

Table I

Variable	**Description**	**Possible Values**
Hrs	Number of hours spent for studies	2, 3, 4, >4
Regular	Regular to class	Regular, irregular
LibraryVisits	Number of visits to library per week	2, 3, 4, >4
BooksType	Type of books borrowed	Technical, Magazine, Novel
Interaction	Interaction in class	Poor, Good, Better, Best
TimeManagement	Time management ability	Poor, Good, Better, Best
Grasping Ability	Grasping Ability	Poor, Good, Better, Best
EXActivities	Participation in extracurricular activities	Poor, Good, Better, Best

StressManagement	Stress management ability	Poor, Good, Better, Best
DecisionMaking	Decision machining ability	Poor, Good, Better, Best
PrevSemResult	Previous Semester Result	S:90-100% A:75-89% B:60-74% C:50-59% D:45-49% E:40-44% F: Less than 40%
SSLC	Result in 10th	Distinction:85-100% First:60-84% Second:45-59% Third:35-44% Fail: Less than 35%
PUC	Result in 12th	Distinction:85-100% First:60-84% Second:45-59% Third:35-44% Fail: Less than 35%
Hostel	Living in hostel	Yes, No
Punctual	Punctual to class	Yes, No
FacultyGuide	Guidance from faculties	Poor, Good, Better, Best
FamilyGuide	Guidance from family	Poor, Good, Better, Best
HIS	Interested in Higher Studies	Yes, No

C. Naïve Bayes Algorithm

The performance of the student is predicted using data mining technique called as classification rules. The Naïve Bayes classification algorithm is used by the administrator to predict the performance of the student in the upcoming semester based on their previous semester result and their behaviour.

A Naïve Bayes classifier is a simple probabilistic classifier based on applying Bayes theorem (from Bayesian statistics) with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be "independent feature model".

In simple terms, a Naïve Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 4" in diameter. Even if these features depend on each other or upon the existence of the other features, a Naïve Bayes classifier considers all of these properties to independently contribute to the probability that this fruit is an apple.

Depending on the precise nature of the probability model, Naïve Bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for Naïve Bayes models uses the method of maximum likelihood; in other words, one can work with the Naïve Bayes model without believing in Bayesian probability or using any Bayesian methods.

In spite of their Naïve design and apparently over-simplified assumptions, Naïve Bayes classifiers have worked quite well in many complex real-world situations. In 2004, analysis of the Bayesian classification problem has shown that there are some theoretical reasons for the apparently unreasonable efficacy of Naïve Bayes classifiers. Still, a comprehensive comparison with other classification methods in 2006 showed that Bayes classification is outperformed by more current approaches, such as boosted trees or random forests.

D. Advantages of Naïve Bayes Algorithm

- Naïve Bayes classifier 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.
- It improves the classification performance by removing the irrelevant features.
- High Performance.
- it is short computational time

E. steps of Naïve Bayes Algorithm

Step 1: Scan the dataset (storage servers)

Step 2: Calculate the probability of each attribute value. [n, n_c, m, p]

Step 3: Apply the formulae

P(attributevalue(ai)/subjectvaluevj)=(n_c + mp)/(n+m)

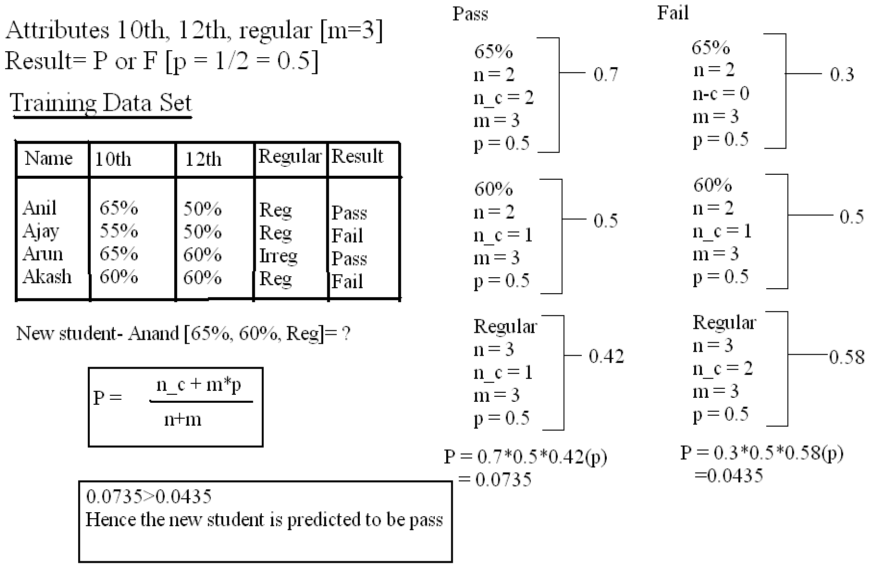
Where:

- n = the number of training examples for which v = vj
- n_c = number of examples for which v = vj and a = ai
- p = 1/number of subject values
- m = the equivalent sample size [number of attributes]

Step 4: Multiply the probabilities by p

Step 5: Compare the values and classify the attribute values to one of the predefined set of class.

The Example of Naïve Bayes algorithm by using only 3 behaviour of the understanding of how the algorithm works is as shown in below Fig 2. The same algorithm will be applied for the set of 18 behaviours that are considered in Table I to predict the performance of the students in the upcoming semesters.



Attributes 10th, 12th, regular [m=3]
Result= P or F [p = 1/2 = 0.5]

Training Data Set

Name	10th	12th	Regular	Result
Anil	65%	50%	Reg	Pass
Ajay	55%	50%	Reg	Fail
Arun	65%	60%	Irreg	Pass
Akash	60%	60%	Reg	Fail

New student- Anand [65%, 60%, Reg]= ?

$$P = \frac{n\_c + m*p}{n+m}$$

Pass

65%
n = 2
n_c = 2
m = 3
p = 0.5
— 0.7

60%
n = 2
n_c = 1
m = 3
p = 0.5
— 0.5

Regular
n = 3
n_c = 1
m = 3
p = 0.5
— 0.42

P = 0.7*0.5*0.42(p)
= 0.0735

Fail

65%
n = 2
n-c = 0
m = 3
p = 0.5
— 0.3

60%
n = 2
n_c = 1
m = 3
p = 0.5
— 0.5

Regular
n = 3
n_c = 2
m = 3
p = 0.5
— 0.58

P = 0.3*0.5*0.58(p)
=0.0435

0.0735>0.0435
Hence the new student is predicted to be pass

Fig 2 Example for Naïve Bayes Algorithm by considering only three Behaviors

VI. CONCLUSION

In this paper, the classification rule is used on student database to predict the student's performance in the upcoming semester on the basis of previous student's database. As there are many approaches that are used for data classification, the Naïve Bayes algorithm is used here. Information's like Attendance, Seminar and Assignment marks were collected from the student's previous database, to predict the performance at the end of the semester. The other attributes are collected by students and their respective faculties who know the behaviour of students.

This study will help to the students and the teachers to improve the result of the students who are at the risk of failure. This study will also work to identify those students who needed special attention to reduce fail ration and taking appropriate action for the next semester examination.

REFERENCE

[1] Tripti Mishra, Dr. Dharminder Kumar, Dr. Sangeeta Gupta, "Mining Students' Data for Performance Prediction", IEEE 2014 Fourth International Conference on Advanced Computing & Communication Technologies.

[2] Udeni Jayasinghe, Anuja Dharmaratne, Ajantha Atukorale, "Students' Performance Evaluation in Online Education System Vs Traditional Education System", IEEE 2015 12th International Conference on Remote Engineering and Virtual Instrumentation (REV).

[3] Krina Parmar, Prof. Dineshkumar Vaghela, Dr Priyanka Sharma, "Performance Prediction of Students Using Distributed Data Mining", IEEE Sponsored 2nd International Conference on Innovations in Information Embedded and Communication Systems.

[4] Brijesh Kumar Baradwaj, Saurabh Pal, "Mining Educational Data to Analyze Student's Performance", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6, 2011.

[5] Abeer Badr El Din Ahmed, Ibrahim Sayed Elaraby, "Data Mining: A prediction for Student's Performance Using Classification Method", World Journal of Computer Application and Technology 2(2): 43-47, 2014.

[6] Behrouz Minaei-Bidgoli, Deborah A. Kashy , Gerd Kortemeyer , William F. Punch , "Predicting Student Performance: An Application of Data Mining Methods with the Educational Web-Based System Lon-Capa", 33rd ASEE/IEEE Frontiers in Education Conference.

[7] Anmol Kumar, Amit Kumar Tyagi, Surendra Kumar Tyagi, "Data Mining: Various Issues and Challenges for Future A Short discussion on Data Mining issues for future work", International Journal of Emerging Technology and Advanced Engineering, Website: www.ijetae.com (ISSN 2250-2459 (Online), Volume 4, Special Issue 1, February 2014).

[8] S.D.Gheware, A.S.Kejkar, S.M.Tondare, "Data Mining: Task, Tools, Techniques and Applications", International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 10, October 2014.

[9] S.Archana, Dr. K.Elangovan, "Survey of Classification Techniques in Data Mining", S.Archana et al, International Journal of Computer Science and Mobile Applications, Vol.2 Issue. 2, February- 2014, pg. 65-71.

[10] Jiawei Han, Micheline Kamber, Jain Pei, "Data Mining Concepts and Techniques", third edition.

[11] Pang-Ning Tan, Vipin Kumar, Michael Steinbach, "Introduction to Data Mining", 2006 by Pearson Education, Inc.