Performance Evaluation and Improvement in Academic Abilities Using Prediction Analysis

Ravi Kumar V G[#1], Vinay M G[*2]
Department of CSE
GSSSIETW, Mysuru
[#1]ravikumarvg@gsss.edu.in
*[*2]vinaymg@gsss.edu.in*

***Abstract*—The education performance is a turning point for all the students in academics. The data stored in educational database contain hidden information for evaluation and improvement of students' performance. The ability to predict a students' performance is very important in educational environments. A very promising tool to achieve this prediction is Data Mining. The academic performance of a student is based upon diverse factors like personal, social, psychological and other environmental variables. The University/Institution will have the ability to predict the students' performance, so that, they can manage and prepare necessary resources for the new students. This helps the teacher to improve the student's performance and decides on those students who need special attention. This helps in identifying the slow learners and to take actions at the right time. Bayesian classification method is used on student database to predict the present student's performance.**

***Keywords* —Data Mining, Educational Data Mining, Predictive Model, Bayesian algorithm.**

I. INTRODUCTION

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations. It 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. Student retention has become an indication of academic performance and enrollment management. One of the most useful data mining techniques for e-learning is classification. 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. 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. The procedures to assist the low academic achievers in higher education are:

(a) Generation of data source of predictive variables.
(b) Identification of different factors, which affects a student's learning behavior and performance during academic career.
(c) Construction of a prediction model using classification data mining techniques on the basis of identified predictive variables.
(d) Validation of the developed model for higher education students studying in Indian Universities or Institutions.

II. LITERATURE SURVEY

Brijesh Kumar Bhardwaj and Saurabh Pal [1], in their paper Data Mining: A prediction for performance improvement using classification, discussed an experimental methodology those are used to generate a database, where the raw data was preprocessed in terms of filling up missing values, transforming values in one form into another and relevant attribute selection. They only proved that the academic performances of the students are not always depending on their own effort.

Umesh Kumar Pandey, Brijesh Kumar Bhardwaj and Saurabh pal [2], proposed that, Data Mining as a Torch Bearer in Education Sector and discussed the different type of researches used in the education sector using data mining. Students, educators and academic responsible person can use these findings to improve the quality of education.

Boumedyen Shannaq, Yusupov Rafael and V. Alexandro [3], in their paper Student Relationship in Higher Education Using Data Mining Techniques, aimed to improve the current trends in the higher education systems to understand from the system which factors might create loyal students. Their research work concentrated only on the number of students enrolling in the upcoming years.

M. Sukanya, S. Biruntha, Dr.S. Karthik and T. Kalaikumaran [4], proposed that Performance improvement in Education Sector, using Classification and Clustering Algorithm to discover hidden patterns and relationships of large amount of data, which is very much helpful in decision making.

A decision tree is a classifier expressed as a recursive partition of the instance space. The decision tree consists of nodes that

form a rooted tree, meaning it is a directed tree with a node called "root" that has no incoming edges. All other nodes have exactly one incoming edge. A node with outgoing edges is called an internal or test node. All other nodes are called leaves which are also known as terminal or decision nodes. In a decision tree, each internal node splits the instance space into two or more sub-spaces according to a certain discrete function of the input attributes values.

The outcome of the decision tree predicted the number of students who are likely to pass, fail or what will be the marks one can score.

Here are some of the interesting things about the decision tree.
(a)It divides up the data on each branch point without losing any of the data (the number of total records in a given parent node is equal to the sum of the records contained in its two children).
(b) It is pretty easy to understand how the model is being built (in contrast to the models from neural networks or from standard statistics).
(c) It can be constructed relatively fast compared to other methods of classification.
(b) Trees can be easily converted into SQL statements that can be used to access database efficiently.

Iterative Dichotomiser3, it is a decision tree algorithm introduced in 1986 by Quinlan Ross. It is based on Hunts algorithm. The tree is constructed in two phases. The two phases are tree building and pruning.

Pruning means to change the model by deleting the child nodes of a branch node. Pruned node is regarded as a leaf node, leaf node cannot be pruned. It only accepts categorical attributes in building a tree model. It does not give accurate result when there is noise. To remove the noise pre-processing technique has to be used.

To build decision tree [5], information gain is calculated for each and every attribute and select the attribute with the highest information gain to designate as a root node. Label the attribute as a root node and the possible values of the attribute are represented as arcs. Then all possible outcome instances are tested to check whether they are falling under the same class or not. If all the instances are falling under the same class, the node is represented with single class name, otherwise choose the splitting attribute to classify the instances. Continuous attributes can be handled using the ID3 algorithm by discretizing or directly, by considering the values to find the best split point by taking a threshold on the attribute values. ID3 does not support pruning.

C4.5 is a successor to ID3 developed by Quinlan Ross. It is also based on Hunt"s algorithm. C4.5 handles both categorical and continuous attributes to build a decision tree. In order to handle continuous attributes,C4.5 splits the attribute values into two partitions based on the selected threshold such that all the values bove the threshold as one child and the remaining as another child. It also handles missing attribute values. C4.5 uses Gain Ratio as an attribute selection measure to build a decision tree. It removes the biasness of information gain when there are many outcome values of an attribute. At first, calculate the gain ratio of each attribute. The root node will be the attribute whose gain ratio is maximum. C4.5 uses pessimistic pruning to remove unnecessary branches in the decision tree to improve the accuracy of classification.
Some of the Limitations of C4.5 Algorithm are as follows:
(a) Empty branches: Constructing tree with meaningful value is one of the crucial steps for rule generation by C4.5 algorithm.Raj Kumar and DR. Rajesh Verma [6] in their experiment, they have found many nodes with zero values or close to zero values. These values neither contribute to generate rules nor help to construct any class for classification task. Rather it makes the tree bigger and more complex.
(b) **Insignificant branches**: Numbers of selected discrete attributes create equal number of potential branches to build a decision tree. But all of them are not significant for classification task. These insignificant branches not only reduce the usability of decision
(b) **Over fitting**: Over fitting happens when algorithm model picks up data with uncommon characteristics. This cause many fragmentations is the process distribution. Statistically insignificant nodes with very few samples are known as fragmentations [7]. Generally C4.5 algorithm constructs trees and grows it branches „just deep enough to perfectly classify the training examples". This strategy performs well with noise free data. But most of the time this approach over fits the training examples with noisy data.

III. METHODOLOGY

A. System architecture for student performance evaluation

System architecture for the proposed system is shown in figure 1. The data is collected from the database and analyzed. The necessary preprocessing steps are applied on the data. Then the Bayesian Classification Algorithm is applied for the preprocessed data. The performance of the result obtained is evaluated and the pattern is extracted to predict the student performance in the further examination.

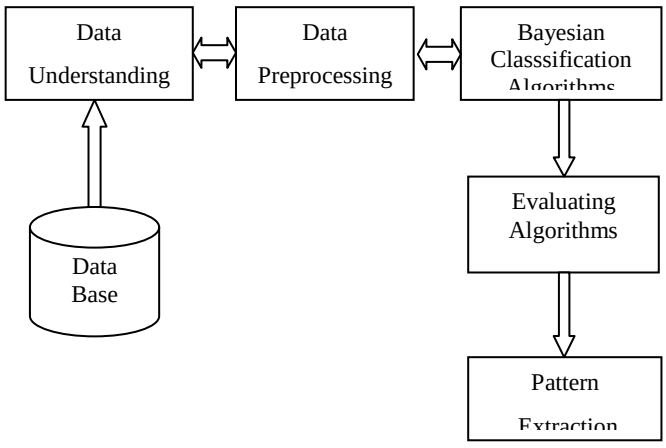


Figure.1.System architecture for Performance Evaluation.

A Naive Bayes algorithm is one of the most effective methods in the field of text classification, but only in the large training sample set can it get a more accurate result. The requirement of a large number of samples not only brings heavy work for previous manual classification, but also puts forward a higher request for storage and computing resources during the computer post-processing [8]. Naive 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 naive 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.

The Classifier

The Bayes Naive classifier selects the most likely classification V_{nb} given the attribute values a1,a2,...,n
This results in:

$$V_{nb} = \text{argmax}_{v_j \epsilon V} P(v_j) \prod P(a_i \mid v_j) \quad (1)$$

We generally estimate $P(a_i \mid v_j)$ using m-estimates:

$$P(a_i \mid v_j) = \frac{n\_c + mp}{n + m} \quad (2)$$

n is the number of training examples for which $v=v_j$
n_c is number of examples for which $v=v_j$ and $a=a_i$
p is a priori estimate for $P(a_i \mid v_j)$
m is the equivalent sample size

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 1 for reference.

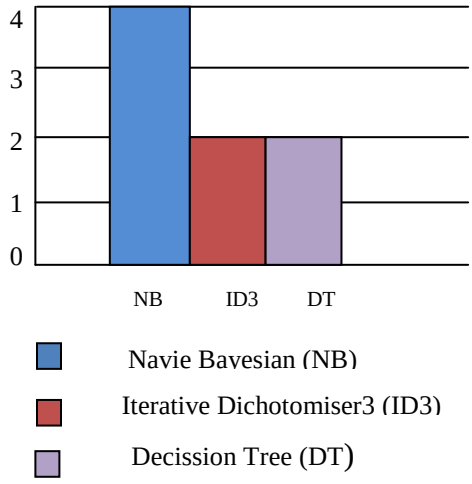
TABLE I
Student related variables

VARIABLE	DESCRIPTION	POSSIBLE VALUES
Gender	Students‟ gender	Male, female
Cat	Students‟ category	Gm, Obc, Sc, St
Med	Medium of teaching	Kannada, English, Hindi
SFH	Students food habit	Veg, Non-veg
SOH	Students other habit	Smoking, Drinking
LLoc	Living location	Village, Taluk, District
Hos	Where do u stay	Hostel, Room, Pg
FSize	Number of members in a family	2, 3, >3
FStatus	Students family status	Joint, Individual
FAIn	Family annual income status	Bpl, Poor, Medium, High
GSSLC	Students grade in 10th	<40,40-59, 60-80,>80
GPUC	Students grade in 12th	<40,40-59, 60-80,>80
TColl	Students college type	Boys, Girls, Co-ed
FQual	Fathers qualification	No-ed, Elementary, Secondary, Graduate, PG,Doctarate
MQual	Mothers qualification	No-ed, Elementry, Secondary,Graduate, PG,Doctarate
FOcc	Fathers occupation	Farmer, Business, Service, Retired
MOcc	Mothers occupation	Farmer, Business, Service, Retired
IHE	Student interested in higher education	Yes, No
UOM	Do u use mobile?	Yes, No
UOI	Do u use internet?	Yes, No
UOSN	Do u use social network?	Yes, No
SQ	How many siblings & their qualification?	Yes, No
RH	Reading habit	Night, Early morning
UOV	Do u use vehicle	Yes, No

The attributes that are required as the input to the system are shown in the table 1. Information pertaining to 25 attributes is collected from individual student, which include the diverse factors that affect the student behavior. These include personal, social and academic details. The Naïve Bayesian Classification Algorithm is applied to these attributes to predict the performance of the student.

C. Result

The comparison analysis of number of outputs of ID3 technique, Decision Tree with the Naïve Bayesian for the Education Data Mining system. The ID3 algorithm and Decision Tree predicts the result in terms of only 2 parameters, i.e., pass and Fail. Whereas, the proposed Naïve Bayesian predicts the result in terms of four parameters i.e. Bad, Average, Good and Excellent.



D. Conclusion

In this paper, Bayesian classification method is used on student database to predict the students division on the basis of previous year database. This study will help to the students and the teachers to improve the division of the student. This study will also work to identify those students which needed special attention to reduce failing ration and taking appropriate action at right time. Present study shows that academic performances of the students are not always depending on their own effort. Our investigation shows that other factors have got significant influence over students' performance. This proposal will improve the insights over existing methods.

IV. REFERENCES

[1] 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.

[2] 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.

[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] M. Sukanya, S. Biruntha, Dr.S. Karthik and T. Kalaikumaran, "Data Mining: Performance Improvement in Education Sector using Classification and Clustering Algorithm" International Conference on Computing and Control Engineering (ICCCE 2012), 12 & 13 April, 2012.

[5] D.Lavanya, Dr.K.Usha Rani, "Performance Evaluation of Decision Tree Classifiers on Medical Datasets", International Journal of Computer Applications (0975 –8887),Volume 26–No.4, July 2011.

[6] Raj Kumar. Dr. Rajesh Verma, "Classification Algorithm for Data Mining: A Survey". International Journal of Innovations in Engineering and Technology (IJIET) Vol. 1 Issue 2 August 2012.

[7] J. Han and M. Kamber, "Data Mining: Concepts and Techniques", Morgan Kaufmann Publish, 2001

[8] Yuguang Huang Beijing Univ. of Posts &, Telecommunication., Beijing, china Lei Li "Naïve Bayes classification algorithm based on small sample set", Cloud Computing and Intelligence Systems (CCIS), 2011 IEEE International Conference.

