IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 11, November 2015

Students' Academic Failure Prediction Using Data Mining

Lumbini P. Khobragade[1], Prof. Pravin Mahadik[2]

ME Student, Department of Computer Science, DIEMS, Aurangabad, India[1]

Assistant Professor, Department of Computer Science, DIEMS, Aurangabad, India[2]

Abstract: This paper proposes to apply Data Mining Techniques to predict the students' failure on real time data of school or graduating students. Experiment attempts the detection of students' failure to improve their academic performance and to prevent them dropping out. Research has been done on assessing students' failure based on various attributes. In this experiment, 11 best attributes has been selected. Different approaches have been applied to resolve the problem of high dimensionality and using classification algorithm on Engineering students' previous and present education information to generate the model and this model can be used to detect students' academic failure. The results are compared and presented.

Keywords: Educational Data Mining (EDM), Academic failure, Classification, Prediction, Decision tree, Induction Rule.

I. INTRODUCTION

Data mining is the process of analyzing data from different perspectives and summarizing the results as useful information. A solution to achieve this goal is to use the knowledge discovery in databases techniques or data mining in education, called educational data mining, EDM [2]. The EDM process converts raw data coming from educational systems into useful information that could potentially have a great impact on educational research. Data mining has been recognized as promising area for database research. This area can be defined as efficiently discovering interesting rules from large database. This new area of research focuses on the development of methods to better understand students and the settings in which they learn. In fact, there are good examples of how to apply EDM techniques to create models that predict student performance specifically. This proposed system will show promising results with respect to those sociological, economic or educational characteristics that may be more relevant in the prediction of low academic performance.

Many educational organizations and school administrations today, leave no stone unturned to improve their student's academic performance. In which the marks obtained by the student in the examination decide his/her future. They want to increase the number of student's getting passed in the yearly academics. The reason for this is to develop the best quality of the education process in their institute, to maintain the brand name of the organization and to educate students in a better way. In order to increase the number of students getting passed, the students that may get failed in that particular year in academics need to find firstly. This project basically aims to foretell the student's failure beforehand, so that some measures can be taken to avoid the student's failure in future.

To predict the failure of students is a complex task, as it requires large number of the data to be handled. For which the record of students, their each and every activities, academic related information need to maintain. Based on this information, it will be easier to predict the student's failure by applying data mining algorithms on it.

This paper describes the model that predicts the academic performance of the engineering students; specifically their failure has become one of the most important needs in order to improve the quality of the teaching process of that institution. It actually proposes to foretell the student's academic failure using the algorithms of Data Mining techniques. Most of the features selected for creating the model are based on real-world data of student information (e.g. student marks, family background, social and academic related features) and also their past performance, as the past performance of a student is indicative of his present/future performance. In most of the cases, these data can be collected by using college reports and surveys. The DM algorithms are applied on huge collection of Dataset and the results are obtained on it through which the failure can be predicted. If properly analyzed, these data could be used by the academic organizations to understand the behavior of students. This information is more useful for the teachers and the principal of the organization, so that they can make proper arrangements and facilities to increase the capability of students and reduce / prevent the failure of students in academic year.

The following scenario will help us to understand the importance of this model. The main goal of any educational institutions is to develop the best quality of the education process. Examination plays a vital role in any student's life. The marks obtained by the student in the examination decide his future. In higher education, computer literacy is the most important and basic skill in the knowledge-based economy. Nowadays, in order to promote basic computer skills of graduate students, some universities have developed specific certification programs to examine students' computer proficiency. As a number

of engineering seats and colleges are increasing in India, the inferior students are also enrolled in engineering courses. So the results of the colleges/universities for engineering courses are going down. If it is known in advance which students are likely to fail, the colleges or the teachers can take the necessary actions (like increasing tuition hours per week) to improve the results. This result can help the college to identify a number of student groups who need to pay much more attention from the teachers. This will finally help in improving even the placements. Good placement is one of the key factors that will help the college to attract students. These experiments can show almost expected results in context with economic, educational or sociological characteristics that may be helpful in foretelling low academic performance.

The paper is organized as follows: Section II presents Related Work. Section III, describes the Dataset Classification. Section IV, present the Experimentation and results obtained. And finally section V, summarizes the Main conclusions and future research.

II. RELATED WORK

Carlos Marquez-Vera, Cristobal Romero Morales [1], and Sebastian Ventura So to have tried to attempt to solve this problem of predicting student's academic failure using either clustering algorithms, induction rules or decision trees algorithms of data mining techniques. They used five methods which are followed in manner shown below:

Data Gathering: In this stage, they gathered huge amount of data related to the students. The set of factors that may affect the performance of the student were gathered in this stage. There are three sources of information from which the data is collected. Firstly a specific survey (personal, family information), second is CENEVAL [1], a general survey which collects information when students register themselves in National Evaluation Center and third is Departmental survey (collects information from respective departments of student's courses). After the information is collected, all the information is transformed into a dataset.

Data pre-processing**:** The information about the students gathered in the dataset above is large and also it is not in proper format. So this information / data is to be processed before it is used for further evaluation. In this stage, pre-processing of data is done. Pre-processing involves data cleansing, transformation of variables, integration, discretization and data partitioning [1]. In this stage selection of best attributes and re-balancing of data is also done. The existing system used a tool named Weka tool for feature selection. This tool is usually being used for attributes classification, data balancing and rebalancing algorithms. The data filter used in this system is SMOTE(Synthetic Minority Over-Sampling Technique) [1].

Data mining: In this stage, abstraction of useful data is done using various data mining techniques. They applied five rules of induction and five decision tree algorithms on the dataset for developing predicting models of student's academic failure. The five induction rules are (JRip, NNge, OneR, Prism, Ridor) and the five decision tree algorithms are (J48, ADTree, Random Tree, REP Tree, SimpleCart). For handling imbalanced data, cost – sensitive classification [1] approach is also used. The data obtained from data mining techniques will also be useful for quick decision making.

Interpretation of Results: In this stage, the results obtained from the models were analyzed to predict the student's failure. Then the results were compared, the factors were identified and then interpreted accordingly.

Disadvantages: The existing system makes use of readymade DM software called Weka tool for applying data mining techniques. It uses 10 classification algorithms, i.e. five rules of induction and five decision tree algorithms. This increases the overhead and complexity of the problem. Only three algorithms are more than sufficient for classification of attributes of students.

C. Romero and S. Ventura [2] discussed all about educational data mining and its use. According to them educational data mining has become a very popular research community because of the increase in the interest of people in data mining techniques and educational systems. In their paper they discussed about the application of data mining in the field of education systems. The education systems may be various educational universities, distance learning courses, training institutes, information management systems, etc. After the pre-processing of data is done, they apply the data mining procedures on this data, for example clustering; association rule mining, classification algorithms, statistics and visualization [2] etc. In brief it gives you the idea of using data mining techniques in the field of educational systems.

Oyelade, O. J , Oladipupo, O. O, Obagbuwa, I. C [5] predicts the student's academic performance using cluster analysis and statistical algorithms. They implemented k-means clustering algorithm to analyze the students' data and predict their results. In this paper they have implemented the model of k-means clustering algorithm on a private institute of Nigeria. The clustering algorithm divides the students in homogenous groups according to their capabilities and characteristics. This information can be helpful for both the instructor and the students to improve their academic performance.

There are two types of cluster analysis techniques. One is hierarchical clustering technique and the other one is non-hierarchical clustering techniques. Single linkage, complete linkage, average linkage and median are the hierarchical clustering techniques [5] and k-means, adaptive k-means, and fuzzy clustering are non-hierarchical clustering techniques [5]. A clustering algorithm is said to be good if it produces groups with distinct non-overlapping boundaries. They have overcome over the other models of prediction like fuzzy logic decision making method for prediction. The problem with fuzzy logic was that it could be applied on smaller set of datasets i.e. 2 or 3.

Dr. Vuda Sreenivasarao, Capt. Genetu Yohannes [6], have made an attempt to improve the engineering system by predicting the students academic performance. They have made use of data warehousing and data mining techniques. The data mining techniques used by them are k-means clustering algorithm, and decision trees. The whole paper focuses mainly on k-means clustering algorithm. Clustering is a process of grouping similar objects together. The objects that have same characteristics and behavior are organized in one group. One cluster forms one group. In the process of cluster analysis the objects with similar behavior are grouped together and then particular labels are assigned to each group. The main advantage of clustering is that it is more adaptable to the changes, and clearly distinguishes between the objects with different kind of behaviors [6].

This is more preferable than classification technique of data mining. Clustering is required in data mining because it supports scalability, has the ability to deal with different kind of attributes, discovery of clusters with attribute shape, has high dimensionality and ability to deal with noisy data and it is interoperable [6]. In the projects related with predicting the student's academic failure, the most commonly used clustering method is the k-means clustering algorithm. For database purpose, they make use of data warehouse where operational data is being transformed into query tools, OLAP tools, and data mining tools. The focus is mainly to increase the number of students that would have good performance in their academics.

Bresfelean worked on the data collected through the surveys from senior undergraduate students at the faculty of economics & business administration in Cluj-Napoca [7]. Decision tree algorithms in the WEKA tool, ID3 and J48 were applied which students are likely to continue their education with the postgraduate degree. The model was applied on two different specializations students' data and an accuracy of 88.68% and 71.74% was achieved with C4.5.

P. Cortez and A.Silva [8] worked on secondary students' data to predict their grade in contact education system. Past performance as well as socio-economic information was collected and the results were obtained using different classification techniques. It was found that the tree based algorithm outperformed the methods like Neural Networks and SVM.

V.P. Bresfelean, M. Bresfelean and N. Ghisoiu [9] found that students' success depends on students choice in continuing their education with post university studies or other specialization attribute, students admittance grade and the fulfillment of their prior expectation regarding their present specialization.

S.Anupama Kumar and Vijayalakshmi M.N concluded that Decision rule and One R rule algorithms can be used to predict the result of the fifth semester of student in higher education based on the marks obtained by the students in the previous four semesters [10]. Rule based algorithm can provide efficiency in predicting the student's performance in higher education using the previous historical data.

CHALLENGES

The challenges I have faced while implementing this project was in the field of attribute selection algorithms and prediction result. Attribute Selection basically deals with selecting the best attributes out of huge collection of attributes based on which the results can be calculated. To obtain set of best attributes I needed to apply three attribute selection algorithms which are very complex. Those are CfsSubsetEval, Filtered-AttributeEval, FilteredSubsetEval, etc. The first two algorithms when applied on the set of attributes give you a result (X) which contains the attributes that are randomly occurred. Then on this set of (X) attributes I used the third algorithm i.e FilteredSubsetEval that gives you the best attributes. The best attribute is nothing but the subset of (X) attributes. This simplifies the complexity of the programmer and also the program. This step of attribute selection is only to ease the functionality. Another challenge I faced is in the prediction result. This is because prediction result depends on the best attributes. So to get a good prediction result I need to focus more on the best attribute selection.

III. DATASET CLASSIFICATION

A. PROBLEM DEFINATION

In the research papers discussed above, the authors have implemented various models to predict the students academic failure using either k-means clustering algorithm, decision tree rules, fuzzy logic or by statistical analysis.

In the base paper, the authors have made use of 10 algorithms i.e. five decision tree algorithms and five induction rules for prediction. For each of these algorithms you need to do attribute selection and compute the results. Then in addition to these, they have used two different approaches like re-balancing the data and using Cost sensitive classification. This requires lot of computations and increases the overhead and complexity.

In this paper, I am going to generate a model that predicts students' academic failure using DM techniques. The classification algorithms that I am going to use are two rule induction algorithms, two decision tree rules and Naive Bayes Algorithm [15] provided by Microsoft SQL Server Analysis Services. This algorithm is basically used for predictive modeling which is based on Bayesian Techniques. This reduces the complexity of the program and also obtains precise outcome. The model is easy to be read and understood. This model can give professor interesting information about student and provides guidance to teacher to choose a suitable track, by analyzing experiences of students with similar academic achievements.

B. PROPOSED SYSTEM DETAILS

The method proposed in this paper for predicting the students' academic failure belongs to the process of knowledge Discovery and Data Mining (shown in Fig. 1).

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 11, November 2015

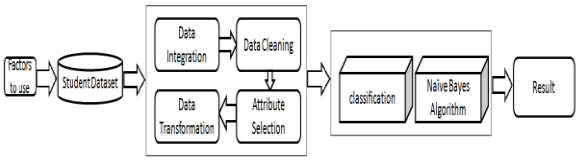

Fig. 1 Method used for prediction of students academic failure.

The main stages of the method are:

Data collection refers to information gathering. The information about the students is collected through the survey, through the enrolment form filled by the students at the time of admission and also through the department survey. This information is nothing but the data that will be useful in predicting the academic failure of students. It must be pointed out that a very important task in this work was information gathering and data preprocessing. All the information used in this study has been gathered in three different categories.

a.) First category is specific survey: where personal and family related information of the student is collected.

b.) Second category is general survey: where previous education information of the students is obtained. The data is the information that is required by various higher and secondary education institutions while admitting the students in their institutions.

c.) Third category is departmental survey: Where the academic related factor of students i.e. marks obtained by the students in different class at the end of semester.

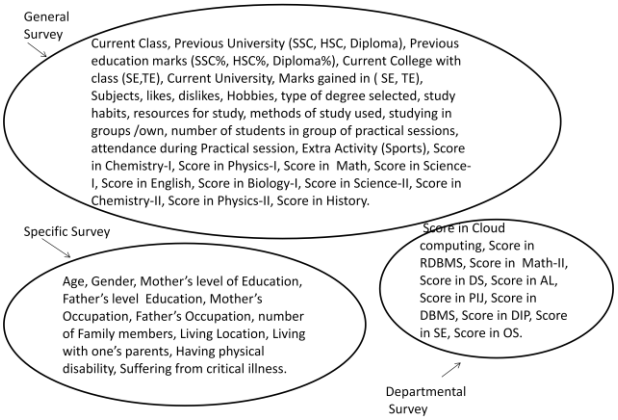

Fig. 2 Data Gathered in three different categories are shown grouped as above.

Data Management refers to preparing the data for applying data mining techniques. In data management; data cleaning, transformation of variables, data reduction and data partitioning is carried out. Firstly, all available data were integrated into a single dataset. During this process those students without 100% complete information were eliminated. All the students who did not answer the specific survey or the general survey were excluded. Some modifications were also made to the values of some attributes. For example, "N" is to be transformed into "N". Also in case of student's age, it should be set in the dd/mm/yy format. Another case is that numerical values of the marks obtained by students in each class should be changed to categorical values [1]. For e.g. if marks obtained between 70-100% then "Topper", if marks obtained between 55-69% then "Average" and if marks obtained between 40-54% then "Below Average" and so on. And lastly, all the cleaned data is to be integrated into a dataset. It will enable to identify the students in advance who are likely to fail and allow the teacher to provide appropriate inputs.

Data Mining consists of certain algorithms that help in predicting the student's failure using classification algorithms. For doing this task, the classification algorithms based on two rule induction algorithms, and two decision tree algorithms is proposed to use. Those are two rule induction algorithms; NNge (it is a nearest neighbor approach); OneR [1], which uses the minimum-error attribute for class prediction; and two decision tree algorithms; Random tree [1], which considers K randomly chosen attributes at each node of the tree; SimpleCart [3], which implements minimal cost-complexity pruning. And Naive Bayes Algorithm is also used to resolve the problem of high dimensionality [15] provided by Microsoft SQL Server Analysis Services.

This algorithm is basically used for predictive modeling which is based on Bayesian Techniques. It calculates the probability of every state of each input column; it gives each possible state of the predictable column [15]. The decision tree, Rule induction and naïve bayes algorithms can be easily implemented in the form of IF-THEN rules of object- oriented programming, which can be easily understood. In this way, even a normal user who doesn't have any deep knowledge about data mining for eg. Teacher and the administrator can easily understand the results obtained using these algorithms. Finally, all the algorithms have been executed, evaluated and compared in order to determine which one obtains the best results.

Implementation is the last phase of the project where the results obtained from DM techniques are interpreted into a model. For implementation, I am going to make use of .Net Technology. It will enable to identify the students in advance who are likely to fail and allow the teacher to provide appropriate inputs.

C. SYSTEM ARCHITECTURE

The architecture of the system consists of following components:

1. Users (Students, Teacher, Principal/ Admin)
2. Data mining techniques (Feature selection algorithms, Classification algorithms)
3. Database (Student's information)

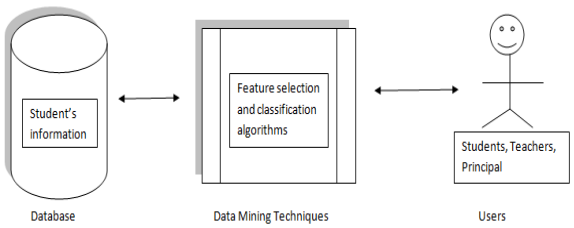

Fig. 3 Components of the Project

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 11, November 2015

The system architecture shows how the three components interact with each other. The student's information is stored in the database on which data mining techniques are applied for prediction. The prediction result is then made visible to the users of the system i.e. students, teachers and principal /administrator so that the teachers and principal can take appropriate measures to improve their performance.

1. USERS

Student is the basic component of the project. The project mainly focuses on predicting the students' academic failure so that proper guidance can be provided to those students who may fail in future and help them from dropping out. Each student registers itself on the site, and can fill its information details. The information can be his/her personal, academic and department wise information. The students only have the authority to see their results and notices arranged for them by their teacher. Students can use them to identify their learning tasks, activities and resources to improve their learning.

Teacher has a very important role in this project. The teacher is the only person who has the rights and authentication to access the prediction tool. The teacher can view the results calculated by the prediction tool and take appropriate decisions regarding that particular student. The teacher can view the details of all the students, manage the lecture, manage the practical batches of the students, add/update other skill-sets of students, short-list the students, arrange exam schedule for students, arrange notices regarding test, exams, results or any other departmental activity for the students and prepare a report of it. Teachers can use them to get more feedback, to identify the students at risk of failure and guide them to help them succeed, to identify most commonly made mistakes and to organize the contents of site in an efficient way.

Administrator can view the final class wise result of students and accordingly arranges the notices for the teachers. He can also use them to decide which course to offer. Basically here all the users have different rights and authentication to access the information.

2. DATA MINING TECHNIQUES

Data Collection Techniques basically deal with gathering student related information that will be useful in predicting their failure in future. The information is provided by the student itself. There are three categories in which the data is collected as shown previously. All the data for the model is collected through the questionary survey, from college reports which include the personal and family related information of the students like number of members in a family, occupation of father and mother, living with one's parents, their living location, whether suffering from critical illness, etc. Then students' information of about their Past education institutions that is required while admitting the students in the next school/ institutions are also used. For example, age, gender, previous school information, type of school, marks obtained in previous class, extra activity, sports etc as past performance of a student is indicative of his present/future performance in most of the cases. And lastly the final score obtained in the present education institution is also collected. Finally all this information is then stored in the dataset. Classification algorithm based on rule induction and decision tree are applied in the obtained attribute set. In addition to these, Naïve bayes algorithm is also used for prediction purpose as explained previously.

Attribute Selection basically deals with selecting the best attributes out of huge collection of data based on which the results can be calculated. Practically the information provided by each student is more than sufficient for the prediction. Instead of making use of each and every information, here we can select few best attributes out of the huge collection of data for prediction and precede the further process of prediction. This simplifies the complexity of the programmer and also the program. There will not be much difference in the results obtained. This step of attribute selection is only to ease the functionality.

3. STUDENTS DATABASE

College academic failure of student is also known as the "one thousand factors problem", due to the large amount of risk factors or characteristics of the students that can influence academic failure, such as demographics, cultural, social, family, or educational background, socioeconomic status, psychological profile, and academic progress. In this paper, the information about students enrolled on Academic Program of the Department of computer science and engineering for the academic year 2014/15 is used. In the educational system, college offers a four-year education program that provides the students of age between 19-24 years old with Technical skills, computer literacy, and general knowledge to continue studying at university. In any education system, as 1st year is considered to be highest failure rate, but here instead of using the 1^{st} year students' information, here second and third year college student information is used, because not only 1^{st} year but all class i.e. $1^{st} – 4^{th}$ year the failure rate increases in any technical field. So the results of colleges/ universities for engineering courses are going down. Institute's success highly depends upon students' success in that institute. Knowing the reason of failure of student can help the teachers and administrators to take necessary actions so that the success percentage can be improved. It is very important for newly started institute to improve the student's success rate every year to attract superior students. Data of approximately 150 students with 49 attributes /variables is collected from department of Computer Science of the institute during the period of 2014-15 year as explained previously. The data is collected through the survey, through college report and also through the department survey.

Next, the whole dataset was divided randomly and classification algorithm can be evaluated using available information (variables). So after preprocessing, dataset with 49 variables/attributes of 150 students is used.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 11, November 2015

However, the dataset has two typical problems that normally appear in these types of educational data. On the one hand, data set has high dimensionality; that is, the number of attributes or features becomes very large. On the other hand, the data are imbalanced, that is the majority of students (119) are passed and minority (31) failed. The problem with imbalanced data arises because learning algorithms tends to overlook less frequent classes (minority classes) and only pay attention to the most frequent ones (majority classes). As a result, the classifier obtained will not be able to correctly classify data instances corresponding to poorly represented classes.

Algorithm	Attributes Selected
NNGe	Score in Previous class, Current Class, Score in DS, Score in AL, Score in Math-II, Likes, Dislikes, Department, Score in OS, Current University, Age, Subjects, Score in DIP, Score in RDBMS, Living location.
OneR	Studying in groups/ own, Score in DBMS, Score in English, Father's Occupation, Score in PIJ, Current University, Age, Score in Cloud Computing, Score in Math-II.
SimpleCart	Annual Income, Score in DBMS, Number of Family members, Score in Physics-I, Current University, Score in DIP, Gender, Score in English, Score in OS.
RandomTree	Score in Cloud Computing, Score in Science-I, Living Location, Score in OS, Score in RDBMS, Score in DS, Score in English, Hobbies, score in Math-II.

Table 1 Attribute Selected.

Then I decided to carry out the study of feature selection to identify which feature has the greatest effect on the output variable (academic status). Predicting tool provides several feature selection algorithms from which I have selected the following feature selection algorithms and those are CfsSubsetEval, Filtered-AttributeEval, FilteredSubsetEval, etc. Because of these attribute selection algorithms, the best attributes out of huge number of attributes of students that affect the student's performance can be selected.

CfsSubsetEval	Score in Previous College, Score in Math, Score in English, Score in Physic-II, Score in Science-I, Studying in Groups / Own, Age, Methods of study used, Gender, Mother's level of Education, Father's level Education, Mother's Occupation, Father's Occupation, number of Family members, Living Location, Score in DS, Score in PIJ, Score in RDBMS, Score in OS, Score in SE, Likes, dislikes, Hobbies, Score in Science II, Extra activity - Sports, Department, Current class.
Filtered-AttributeEval	
FilteredSubsetEval	Current batch, Current university, Pre Marks Pre University, Father Occupation, Annual income, AL, Math-II, DBMS, Cloud Computing, DIP.

Table 2 Best Attribute Selected.

D. NAÏVE BAYES ALGORITHM

Naïve bayes algorithm is a simple method for classification based on the theory of probability, i.e. the Bayesian theorem. It is called Naïve because it simplifies problems relying on two important assumptions: it assumes that the prognostic attributes are conditionally independent with familiar classification and it supposes that there are no hidden attributes that could affect the process of prediction. This classifier represents the promising approach to the probabilistic discovery of knowledge, and it provides a very efficient algorithm for data classification.

In proposed system, I am using naive bayes classification algorithm for predictionpurpose . It is a simple probabilistic classifier based on applying Bayes theorem (from Bayesian statistics) with strong (naive) independence assumptions [15]. In simple terms, naive bayes classifier assumes that the presence (or absence) of a particular attribute of a student is unrelated to the presence (or absence) of any other attribute [15].

The probability model for a classifier is a conditional model

$$p(S|A_{1,\ldots,}A_n)$$

over a dependent student variable S with small number of *students*, conditional on several attributes A_1 through A_n.

If the number of attributes is large i.e. if n is large, then designing such a model on probability tables is impractical. And therefore the model is then reformulated to make in practical.

Using Bayes' theorem, we get

$$p\,(S|A_{1,\ldots,}A_n) = \frac{p(S)\;p(A_{1,\ldots,}A_n|S)}{p(A_{1,\ldots,}A_n)}$$

The above equation can be written as,

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}$$

We will only consider the numerator part of the equation, because the denominator does not depend on S, and also the attributes A_i are given. Thus the denominator is constant.

Now the numerator is equivalent to the joint probability model

$$p(S,A_{1,\ldots,}A_n)$$

The above equation can be re-written again in the form of conditional probability as,

$p(S,A_{1,\ldots,}A_n)$

$= p(S)\;p(A_{1,\ldots,}A_n|S)$

$=p(S)\;p(A_1|S)\;p(A_{2,\ldots,}A_n|S,A_1)$

$=p(S)\;p(A_1|S)\;p(A_2|S,A_1)\;p(A3,\ldots,A_n|S,A_1,\quad A_2)$

$= p(S)\;p(A_1|S)\;p(A_2|S,A_1)\;p(A_3|S,A_1,A_2)\;p(A_4,\ldots,A_n|S,A_1,A_2,A_3)$

$= p(S)\;p(A_1|S)\;p(A_2|S,A_1)\;p(A_3|S,A_1,A_2)\ldots\ldots p(A_n|S,A_1,A_2,A_3,\ldots,A_{n-1})$.

Now the "naive" conditional assumptions are to be taken into consideration:

Assume that each attribute A_i is conditionally independent of every other attribute A_j for $j\neq i$. This means that

$p(A_i|S,A_j) = p(A_i|S)$

for $i\neq j$, and so joint model can be equated as

$$p(S,A_1,\ldots.A_n)=p(S)p(A_1|S)\;p(A_2|S)\;p(A_3|S)\ldots.$$

$$= p(S)\prod_{i-1}^{n} p(A_i|S).$$

Considering the above independence assumptions, the conditional distribution over the students S can be expressed as

$$p(S,A_1,\ldots.A_n)= 1/Z\;\; p(S)\prod_{i-1}^{n} p(A_i|S).$$

DOI 10.17148/IJARCCE.2015.41165

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 11, November 2015

where Z is the (evidence) scaling factor dependant only on $A_1, A_2, \ldots, A_n$, i.e. it is constant if the value of the attribute is known.

Such models can be easily managed because they can be factorized into *prior* $p(S)$ and independent probability distributions $p(A_i|S)$. If there are k students and if a model for each $p(A_i|S=s)$ can be expressed in terms of r parameters, then corresponding naive bayes model has (k-1) + n r k parameters [15]. It is often considered, $k=2$ (binary classification) and $r = 1$ (Bernoulli variables as attributes) are common, and so the total number of parameters of the naive Bayes model is 2n+1, where n is the number of binary features used for classification and prediction [15].

IV. EXPERIMENTATION & RESULT

It describes the DM techniques used for obtaining the prediction models of students' academic status at the end of semester. I have performed different experiments to obtain the highest classification accuracy. Starting from gathering information then creating the dataset and applying classification algorithms to reduce the dimensionality of the dataset and then using naïve bayes algorithm. And lastly outcomes are compared and shown. In first experiment I executed four classification algorithms using all available information (49 variables). In second experiment, I used only the best attribute selected (11 attributes). In third experiment, I applied the naive bayes algorithm on all the attributes and the best attributes, and then compare the results. In final experiment I considered different costs in the classification.

In this paper, Classification algorithm based on induction rule and decision tree are used. They are selected because they are considered as a "White Box" Classification techniques; that is, they easily generate the interpretable model and can be directly used for decision making. A decision tree is a set of conditions organized in a hierarchical structure. An instance is classified by following the path of satisfied conditions from the root of the tree until a leaf is reached, which will correspond with a class label. Rule induction algorithms usually employ a specific-to-general approach, in which obtained rules are generated until a satisfactory description of each class is obtained. Then Naïve bayes algorithms that are available in the well- known DM predicting tool have also been used.

Prediction Tool is a tool that calculates the number of students that may fail in future. The tool is basically based on data mining concept and consists of classification algorithm that calculates the failure of students. The classification algorithm is composed of two rule induction algorithm, two decision tree algorithms and naive bayes algorithm. Naive Bayes classification algorithm calculates the probability of every state of each input column, given each possible state of the predictable column [15]. The decision tree algorithms, induction rules and naive bayes algorithms can be easily implemented in the form of IF-THEN rules of object-oriented programming, which can be easily understood. In this way a non-expert user of DM such as teacher or instructor can directly use the output obtained by these algorithms to detect students' failure and to make decisions about how to help them and prevent their possible failure.

I have taken approximately 150 students' information with 49 variables/attributes from department of Computer Science of a well known organization for this project. In the first experiment, all the four classification algorithms (NNge, SimpleCart, OneR and RandomTree) are applied on all the available attributes shown in TABLE I. The results obtained from this experiment are shown in TABLE III. Different algorithms provide different results, i.e. each of them accounts the relevance of attributes in a different way. Table III shows the results in the form of three fields i.e. TP Rate, Acc and GM. TP is the Passing rate, Acc is the overall Accuracy rate and GM is the Geometric Mean. We can see in TABLE III, the TP rate for SimpleCart algorithm is high and OneR is having the second largest TP rate.

Algorithm	TP Rate	Acc	GM
NNge	85.23	78.21	88.78
OneR	90.22	81.15	86.56
Random Tree.	76.34	56.78	67.32
Simple Cart	92.43	59.87	75.55

Table III Classification Result using Attributes

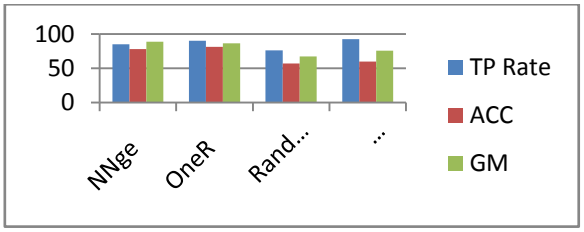

Fig. 4 Graphical Representation for table III

In the second experiment, I executed the classification algorithms on the reduced dataset (with only the best 11 attributes) that have selected using attribute selection algorithm. Table II shows the selected best attributes. The results obtained from this experiment are shown in TABLE IV. The TP rate of OneR algorithm is the highest of all and NNge has the second highest TP rate. So we are going to implement OneR algorithm and NNge algorithm which give the best results.

Algorithm	TP Rate	Acc	GM
NNge	845	765	806
OneR	879	678	798
Random Tree.	765	460	654
Simple Cart	725	609	599

Table IV Classification Result using Best Attributes

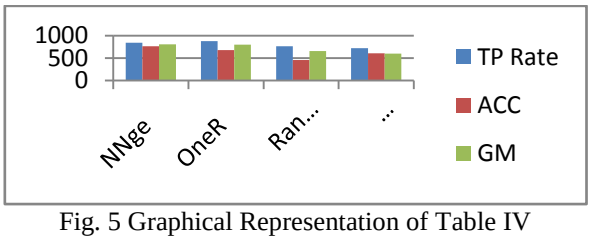

Fig. 5 Graphical Representation of Table IV

DOI 10.17148/IJARCCE.2015.41165

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 11, November 2015

In the third experiment, I executed the naive bayes classification algorithm on all the attributes from TABLE I and the best attributes from TABLE II and then compared the results. This experiment shows that the results obtained by applying the naïve bayes algorithms are much better and accurate than those obtained using the four algorithms. This is because the naïve bayes algorithm takes into consideration a small number of students data (Attributes) for classification and prediction as compared to the above four classification algorithms (OneR, NNge, SimpleCart and RandomTree). The naive bayes algorithm assumes that every attribute/feature of every student is unique and independent. This means that no two attributes of students are dependant on each other. For e.g. if a student is studying for more number of hours, has good occupation of parents and also from good school and has good marks in almost every subject then the probability of that student getting passed in the academic year is more and positive. Even if the other student has the same features/attributes, naive bayes considers all of these attributes to independently contribute to the probability [15] that the first student is going to pass. The naive bayes algorithm has a factor called *posterior* which is probability factor.

The results of the third experiment are shown below in the form of tables and graphs.

Algorithm	TP rate	TN rate	Acc	GM
NNge	85.23	65.54	78.21	**88.78**
OneR	90.22	61.62	**81.15**	86.56
RandomTree	76.34	**81.76**	56.78	67.32
SimpleCart	**92.43**	80.09	59.87	75.55
Naïve Bayes	**94.97**	**85.81**	**87.12**	**89.01**

Table V Classification result of Naïve Bayes using All Attributes

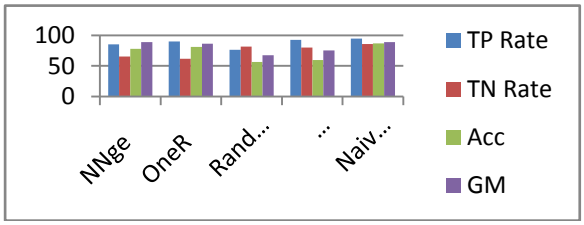

Fig. 6 Graphical Representation of Table V

In the above TABLE V, We can see that each of these algorithms are best in any one of the properties for e.g. NNge has the highest GM rate, OneR has the highest Accuracy rate, RandomTree has the highest TN rate and SimpleCart has the highest TP rate. But it seems that naive bayes algorithm gives all the maximum values for the properties. So the probability of finding the failure of students is much higher in case of naive bayes algorithm.

Algorithm	TP rate	TN rate	Acc	GM
NNge	854	623	460	**806**
OneR	**879**	512	678	798
RandomTree	765	465	**765**	654
SimpleCart	725	**734**	609	599
Naïve Bayes	**912**	**779**	**800**	**843**

Table VI Classification Result of Naïve Bayes using Best Attributes

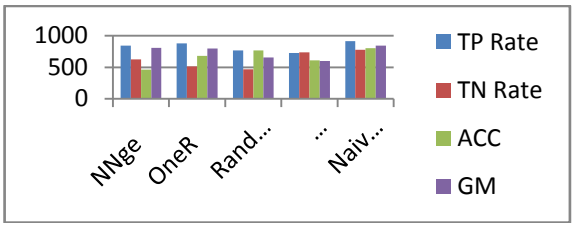

Fig. 7 Graphical Representation of TABLE VI

In the above TABLE VI, We can see that each of these algorithms are best in any one of the properties for e.g. NNge has the highest GM rate, OneR has the highest TP rate, RandomTree has the highest Accuracy rate and SimpleCart has the highest TN rate. But it seems that naive bayes algorithm gives all the maximum values for the properties. So the probability of finding the failure of students is much higher in case of naive bayes algorithm.

V. CONCLUSION & FUTURE WORK

Prior work on predicting student's academic failure was based on Weka tool. All the algorithms required for obtaining results were just outsourced by the previous system. Also the existing system implement five rule induction and five decision tree algorithms which increased the complexity and overhead of the system. In this paper, I implemented the algorithms in the system on my own. I did not outsource the algorithms from Weka tool. Also I implemented only two rule induction, two decision tree algorithms and naive bayes algorithm which decreased the complexity and overhead of the system. I have compared the results of these algorithms and found that naive bayes gives the best and accurate result of prediction. The selection of the features attributes of the students can be done manually or automatically using algorithms. I made this project on a real-time application which can be used in any educational organization for predicting the failure of students. The scope of this project is to predict the failure of students and also provide the necessary information and help and support for those students who are weak in respective subjects. However, it can be concluded that this methodology can be used to help students and teachers to improve student's performance; reduce failing ratio by taking appropriate steps at right time to improve the quality of learning. As learning is an active process, interactivity is a basic elements in this process that affects students' satisfaction and performance. It is important to answers these questions:

- How to obtain that predicting models are user friendly for professors or non-expert users?
- How to integrate data collection system of university and data mining tool?

For future work, the experiment can be extended with more distinctive attributes to get more accurate results, useful to improve the students learning outcomes. Also, experiments could be done using other data mining algorithms to get a broader approach, and more valuable and accurate outputs. Some different software may be utilized while at the same time various factors will be used.

DOI 10.17148/IJARCCE.2015.41165

ACKNOWLEDGMENT

The author wish to acknowledge B.E. Department of Computer Science of Deogiri institute of Engineering and Management Studies, Aurangabad for their support in providing the necessary data.

REFERENCES

[1] Carlos Marquez-Vera, Cristobal Romero Morales, and Sebastian Ventura Soto, "Predicting school failure and dropout by using data mining techniques". IEEE Journal of Latin-American Learning Technologies, Vol. 8, No. 1, February 2013.

[2] C. Romero and S. Ventura, "Educational data mining: A survey from 1995 to 2005," Expert Syst. Appl., vol. 33, no. 1, pp. 135-146, 2007.

[3] L. Brieman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. New York , USA: Chapman & Hall, 1984.

[4] C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," *IEEE Trans.Syst., Man, Cybern. C, Appl. Rev.*, vol. 40, no. 6, pp. 601-618, Nov. 2010.

[5] Oyelade, O. J , Oladipupo, O. O, Obagbuwa, I. C, " Application of k-Means Clustering algorithm for prediction of students' academic performance," International Journal of Computer Science and Information Security, vol. 7, no. 1, 2010.

[6] Dr. Vuda Sreenivasarao, Capt. Genetu Yohannes, " Improving Academic Performance of Students of Defence University Based on Data Warehousing and Data Mining," Global Journal of Computer Science and Technology, Vol 12, Issue 2, Version 1.0, January 2012.

[7] V.P. Bresfelean, "Analysis and Predictions on Students' Behaviour using Decision trees in WEKA Environment", Proceedings of the ITI 2007 29th Int Conf. on Information Technology Interfaces, June 25-28, 2007.

[8] P.Cortez and A.Silva, "Using Data Mining To Predict Secondary School Student Performance", In EUROSIS, A. Brito and J. Teixeria (Eds.), pp.5-12.

[9] P. Bresfelean , M. Bresfelean, N. Ghisoiu, " Determining Students' Academic Failure Profile Founded on Data Mining Methods", Proceedings of the ITI 2008 30th International Conference on Information Technology Interfaces, June 23-26 2008.

[10] S.Anupama Kumar, Vijayalakshmi M.N., ―Mining of Student Academic Evaluation Records in Higher Educationǁ, International Conference on Recent Advances in Computing and Software Systems (RACSS), IEEE conference publication, pp 67–70, Apr,2012.

[11] M. N. Quadril and N. V. Kalyankar, "Drop out feature of student data for academic performance using decision tree techniques," *Global J. Comput. Sci. Technol*., vol. 10, pp.2-5, Feb.2010.

[12] A. Parker, "A study of variables that predict dropout from distance education," *Int. J. Educ. Technol*., vol. 1, no. 2, pp. 1-11, 1999.

[13] http://www.thearling.com/text/dmwhite/dmwhite.htm

[14]http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm

[15] https://msdn.microsoft.com/enus/library/ms174806.aspx

BIOGRAPHIES

Lumbini P. Khobragade, received the B.E. degree in Information Technology from P.E.S. College of Engineering, Dr. BAMU University, in 2011. She has worked as a Lecturer, in Department of Computer Science, of P.E.S College of Engineering, Dr. BAMU University. She is currently pursuing the M.E. degree in Department of Computer Science from Deogiri institute of Engineering & Management Studies, Dr. BAMU University, Aurangabad. Her research interest includes Data Mining and Software Testing.

Pravin Mahadik, Assistant Professor at Deogiri institute of Engineering & Management Studies, Dr. BAMU University, Aurangabad. He has received the B.E. degree and M Tech. degree in Computer Science from JNEC College of Engineering, Aurangabad and Dr. Babasaheb Ambedkar Technological University, Raigad in 2010 and 2013, respectively. His area of research includes Network Programming, Java, and Software Engineering. He has published two papers in Siber Times Internal Journals.