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

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 5, May 2015

Precognition of Students Academic Failure Using Data Mining Techniques

Sadaf Fatima Salim Attar[1], Prof. Y. C. Kulkarni[2]

M.Tech. Scholar, Information Technology, Bharati Vidyapeeth Deemed University College of Engineering,Pune,India[1]

Professor, Information Technology, Bharati Vidyapeeth Deemed University College of Engineering, Pune, India[2]

Abstract: This paper proposes to pre-recognize student's academic failure. Real time data on school or graduating students from an institute is taken and various data mining techniques (classification algorithms), such as induction rules, decision trees and naive bayes are applied on it. The results of these algorithms are being compared and optimized for foretelling which students might fail in future. We first consider all the available attributes of students, then select few best attributes and finally, rebalance the data using classification algorithms. The use of data mining concept in the field of education is called as Educational Data Mining, EDM [2]. This paper focuses on designing various methods that will help the teachers and the principal (Administrator) of the school to figure out the weak students and improve their educational standards and environment in which they learn. I propose the use of data mining procedures, because the complexity of the problem is high, data to be handled is very large and often highly unbalanced. The final objective of this paper is to detect the failure of students as early as possible to prevent them from dropping out and improve their academic performance. The outcomes are compared and the best results are shown.

Keywords: Data Mining, Educational Data Mining, Decision Trees, Induction Rules, Rebalancing Data, Classification Algorithms.

I. INTRODUCTION

Many educational organizations and school administrations today, leave no stone unturned to improve their student's academic performance. They want to increase the number of student's getting passed in their yearly academics. The reason for this is to maintain the brand name of the organization and as well as to educate students in a better way. In order to increase the number of students getting passed, we have to first find out the students that may get failed in that particular year in academics. This projectbasically 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. We need to maintain the record of students each and every activities that he/she does in his/her day to day life. Based on this information, and applying some data mining algorithms on it, we may be able to predict the student's failure.

Data mining is the abstraction of needful data from large databases and ignoring the rest. Data mining tools predict future trends and behaviours, allowing the organizations to make proactive, knowledge-driven decisions [3]. Data mining helps the people to make quick decisions on a situation as compared to statistical analysis. Data mining tools can easily handle large amount of data stored in datasets, they can pre-process the data, and can work on unbalanced data easily. Data mining basically uses more direct approach and does meta-heuristics search on data.

This project makes use of two rules of induction, two decision tree algorithms of data mining and naive bayes algorithm (which is also a classification algorithm used for prediction). Data mining techniques have been under development for decades and are of huge use in research areas like statistics, artificial intelligence and machine learning [3].

This study proposes to foretell the student's academic failure using the algorithms of data mining techniques. The algorithms are applied on huge collection of data on student's activities and the results are obtained, through which the failure can be predicted. This information is more useful for the teachers and 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 academics years. These experiments have shown almost expected results in context with economic, educational or sociological characteristics that may be helpful in foretelling low academic performance.

II. CHALLENGES

The challenges we faced while implementing this project were 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 we need 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 we apply 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

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

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 5, May 2015

the program. This step of attribute selection is only to ease the functionality.

Another challenge we faced is in the prediction result. This is because prediction result depends on the best attributes. So to get a good prediction result we need to focus more on the best attribute selection.

III. RELATED WORK

Carlos Marquez-Vera, Cristobal Romero Morales [1], and Sebastian Ventura Sotohave 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], 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 pre-processing of data is to be 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.

Data mining:In this stage, abstraction of useful 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, etc) [1].

Interpretation of Results: In this stage, the results obtained from the models were analyzed to predict the student's failure.

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. 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 [8] 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.

Dr. Vuda Sreenivasarao, Capt. Genetu Yohannes [9], have made an attempt to improve the engineering system by predicting the student's 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 main advantage of clustering is that it is more adaptable to the changes, and clearly distinguishes between the objects with different kind of behaviours [9]. 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, has ability to deal with noisy data and it is interoperable [9]. For database purpose, they make use of data warehouse where operational data is being transformed into query tools, OLAP tools, and data mining tools.

IV. IMPLEMENTATION

A. Problem Definition

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 my 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. This requires lot of computations and increases the overhead and complexity.

In my paper, i am going to pre-recognize student's academic failure using DM techniques. The classification algorithms that I am going to use are two rules of induction algorithms; NNge (it is a nearest neighbour approach); OneR [1], which uses the minimum-error attribute for class prediction; and two decision tree rules; RandomTree [1], which considers K randomly chosen attributes at each node of the tree; SimpleCart [5], which implements minimal cost-complexity pruning. I am also using another classification algorithm called Naive Bayes Algorithm [6] provided by Microsoft SQL Server Analysis Services. This algorithm is basically used for predictive

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

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 5, May 2015

modelling which is based on Bayesian Techniques. This reduces the complexity of the program and also we obtain precise outcome.

B. System Architecture

The architecture of the system consists of following components:

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

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 so that the teachers and principal can take appropriate measures to improve their performance.

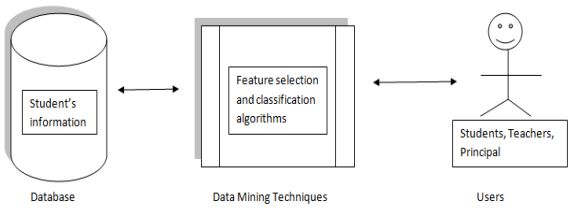

Fig. 1 Components of the Project

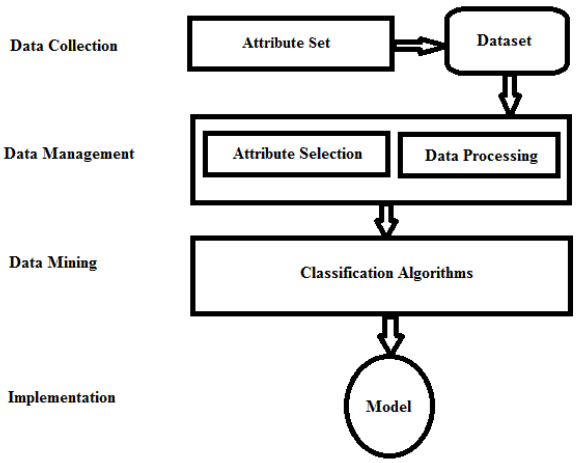

Fig. 2 Architecture of the System

C. Proposed System Details

There are four main modules of the project. They are as follows:

1. Data Collection
2. Data Management
3. Data Mining
4. Implementation

Data Collection is a process where information about the students is collected. This information is nothing but the data that will be useful in predicting the failure of students in academics. The data about students is collected in three different categories; first category is specific survey where personal and family information of the student is collected. For example, number of hours spent studying daily, number of students in each batch, attendance of students in morning/evening tutorials, occupation of father and mother, number of members in a family, studying habits, any illness, etc. second category of data collection is academic information of the students. This data is the information that is required by various higher and secondary education institutions while admitting the students in their institutions. For example, age, gender, previous school information, type of school, marks in math, marks in English, marks in chemistry etc. The third category of data collection is departmental survey where each subject's department wise information of a student is collected. For example marks in math 1, marks in math 2, marks in English 1, and marks in English 2 etc. All this information is then stored in the dataset.

TABLE I
STUDENT'S INFORMATION SOURCES

Specific Survey	Student name, address, cast, age, family name, number of hours spent studying daily, number of students in each batch, attendance of students in morning/evening tutorials, occupation of father and mother, number of members in a family, studying habits, any illness, family income, etc.
Academic Information	age, gender, previous school information, college/school name, university/board name, grade / marks, type of college/school, marks in math, marks in English, marks in chemistry, marks in history, marks in biology, remarks, number of days absent in school, etc.
Departmental Survey	marks in math 1, marks in math 2, marks in English 1, and marks in English 2, marks in chemistry 1, marks in chemistry 2, marks in history 1, and marks in history 2, marks in biology 1, marks in biology 2, etc.

Data Management refers to preparing the data for applying data mining techniques. In data management, we do data processing which involves data cleansing, transformation of variables, data redundancy, spelling mistakes, invalid data, etc. For example. "N'" is to be transformed into "N". Also in a case where the age of a student should be set in the dd/mm/yy format. Another case is that numerical values of the marks obtained by students in each subject are to be changed to categorical values [1]. For e.g. for excellent scoring: score should be between 9.5 and 10, very good scoring: score should be between 8.5 and 9.4 and so on. And at last all the cleaned data is to be integrated into a dataset.

One of the most important techniques of data management is the selection of features (attributes) by applying feature selection algorithms. The attribute selection algorithm tries to select those features of students which have greater

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

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 5, May 2015

impact on their academic status. Few attribute selection algorithms are as follows, CfsSubsetEval, Filtered-AttributeEval, FilteredSubsetEval, etc. Because of these attribute selection algorithms we can select the best attributes out of huge number of attributes of students that affect the student's performance.

TABLE II
BEST ATTRIBUTES SELECTED

Algorithm	Attributes Selected,
CfsSubsetEval	Physical Disability; Age; CPP ; Math 1; Physics 1; Java; Score in Computer
Filtered-AttributeEval	CPP;C#.NET,Math1; Java;Reading Time,Math 2
FilteredSubsetEval	C; ASP.NET;Math1; Java; Testing;CPP

Data Mining consists of certain DM algorithms that help in predicting the student's failure using classification algorithms. The classification algorithms that we are going to use are two rules of induction algorithms; NNge (it is a nearest neighbor approach); OneR [1], which uses the minimum-error attribute for class prediction; and two decision tree rules; RandomTree [1], which considers K randomly chosen attributes at each node of the tree; SimpleCart [5], which implements minimal cost-complexity pruning. I am also using another classification algorithm called Naive Bayes Algorithm [6] 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, given each possible state of the predictable column [6].

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, even a normal user who doesn't have any deep knowledge about data mining, for e.g. teacher and administrator can easily understand the results obtained using these algorithms. Finally, the results of all these executed algorithms evaluated, compared and optimized to determine which one gives the best result.

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.

D. Naive Bayes Algorithm

In proposed system, I am using naive bayes classification algorithm for prediction. It is based on Bayes theorem with strong (naive) independence assumptions [6]. 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 [6].

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,

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

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

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

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

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

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

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

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

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 5, May 2015

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

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

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

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

where Z is the (evidence) scaling factor dependant only on $A_1,A_2,...,\ A_n$, 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 *priorp*(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 [6]. 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 [6].

V. EXPERIMENTATION AND RESULTS

We will be taking approximately 200 student's information form any organization for our project. In the first experiment, all the four classification algorithms (NNge, SimpleCart, OneR and RandomTree) are applied being applied on all the available attributes shown in TABLE I.

The results obtained from this experiment are shown in TABLE III. This table shows you 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.

TABLE III
CLASSIFICATION RESULTS USING ALL ATTRIBUTES

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

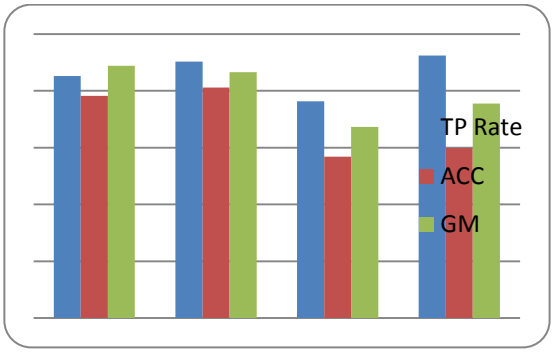

Fig. 3 Graphical Representation of TABLE III

In the second experiment, we are going to apply all four classification algorithm on the best attributes that have selected using attribute selection algorithm. Table II shows the best attributes. The results obtained from this experiment are shown in TABLE IV. The TP rate of OneRalgorithm 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.

TABLE IV
CLASSIFCATION RESULTS USING BEST ATTRIBUTES

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

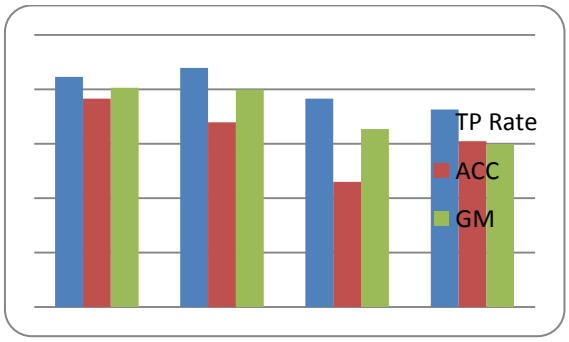

Fig. 4 Graphical Representation of TABLE IV

In the third experiment, we are going to apply the naive bayes classification algorithm on all the attributes from TABLE I and the best attributes from TABLE III, and then compare the results. This experiment shows that, the results obtained by applying the naive bayes algorithms are much better and accurate than those obtained using the four algorithms. This is because the naive bayes algorithm takes into consideration a small number of students data (Attributes) for classification and prediction as compared to the above four classification algorithms. The naive bayes algorithm assumes that every attribute/feature of every student is unique and independent. This means that

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

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 5, May 2015

no two attributes of students are dependent 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 [6] 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.

TABLE V
CLASSIFICATION RESULTS OF NAIVE BAYES USING ALL ATTRIBUTES

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
Random Tree.	76.34	**81.76**	56.78	67.32
Simple Cart	**92.43**	80.09	59.87	75.55
Naive bayes	**94.97**	**85.81**	**87.12**	**89.01**

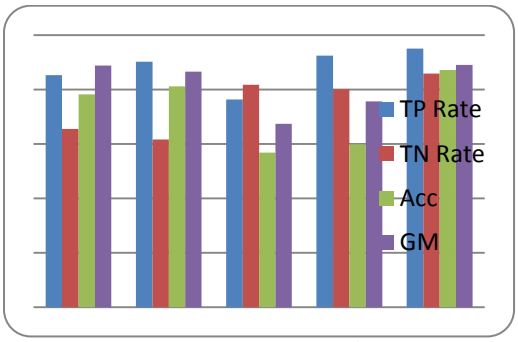

Fig 5. Graphical Representation of TABLE V

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 we can see 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.

TABLE VI
CLASSIFICATION RESULTS OF NAIVE BAYES USING BEST ATTRIBUTES

Algorithm	TP Rate	TN Rate	Acc	GM
NNge	845	623	460	**806**
OneR	**879**	512	678	798
Random Tree.	765	465	**765**	654
Simple Cart	725	**734**	609	599
Naive bayes	**912**	**779**	**800**	**843**

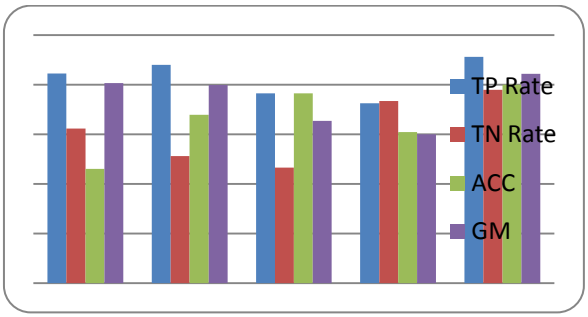

Fig. 6 Graphical Representation of TABLE VI

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 Accuracy rate, RandomTree has the highest TN rate and SimpleCart has the highest TP rate. But we can see 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.

VI. CONCLUSIONAND FUTURE SCOPE

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 rules of induction and five decision tree algorithms which increased the complexity and overhead of the system. In this paper, we implemented the algorithms in the system on our own. We did not outsource the algorithms from Weka tool. Also we implemented only two rules of induction, two decision tree algorithms and naive bayes algorithm which decreased the complexity and overhead of the system. We 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 student can be done manually or automatically using algorithms. We made this project a real-time application which can be used in any educational organization for pre-recognizing the failure of students. The scope of this project is to predict the failure of students and also provide the necessary online information and online help and support for those students who are weak in respective subjects.

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] http://www.thearling.com/text/dmwhite/dmwhite.htm
[4] http://www.anderson.ucla.edu/faculty/jason.frand/teacher/ technologies/palace/datamining.htm
[5] L. Brieman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees.* New York , USA: Chapman & Hall, 1984.
[6] https://msdn.microsoft.com/enus/library/ms174806.aspx
[7] C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," *IEEETrans.Syst., Man, Cybern. C, Appl. Rev.*, vol. 40, no. 6, pp. 601-618, Nov. 2010.
[8] Oyelade, O. J , Oladipupo, O. O, Obagbuwa, I. C, " Application of k-Means Clustering algorithm for prediction of students' academic

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

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 5, May 2015

performance," International Journal of Computer Science and Information Security, vol. 7, no. 1, 2010.

[9] 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.

[10] 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.

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

[12] A Machine Learning Algorithms in Java, Written I, Frank E. WEKA, Morgan Kaufmann Publishers, 2000.

[13] SY. Freund and L. Mason, "The alternating decision tree algorithms," in Proc. 16th Int. Conf. Mach. Learn., 1999, pp. 124-133.