International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 5, Issue 12, December 2016)

Pre-recognition of Student Academic Failure using Data Mining Techniques

Agrawal Bhawana D

[1]*M. Tech Scholar, Department of Computer Science & Engineering, Maharashtra Institute of Technology (MIT), Aurangabad, Maharashtra, India*

***Abstract*— This paper propounds to apply data mining techniques to pre-recognize failure of engineering students. We use real data on engineering students from Aurangabad, Maharashtra and apply white box classification methods such as induction rules and decision trees. The result of these algorithms are compared and used for predicating which student might fail in future. We have considered all the available attributes of student's first, and then select few best attributes and finally, rebalance data using classification algorithms. The outcomes are compared and the best results are shown. In these ways, use of data mining in the field of education is called education data mining EDM [1].**

***Keywords*— Classification, educational data mining (EDM), prediction, Decision Trees, Induction Rules, Rebalancing Data, Classification Algorithms.**

I. INTRODUCTION

From last few years many countries facing problem of school failure and drop-outs, so Governments of many countries showing their interest in determination of its main contributing factors. Large amount of information that current computer can store in Databases is a ―Gold Mine of valuable information about students. Determination of its main contributing factors from this large amount of data is known as the ―the one hundred factors problem and ton of research has been done on recognizing the variables that influence the low execution of student (failure and dropout) at different educational levels as described by Araque et al.,2009 [2].

The solution to this problem is called Educational Data Mining (EDM). EDM concerned with developing methods that extract knowledge from data come from the educational domain. This new territory of research spotlights on the development of methods to better comprehend students and the settings in which they learn.

In fact, there are great cases of how to apply EDM methods to make models that anticipate student failure [3]. These works have indicated promising results concerning those sociological, monetary, or educational attributes that might be more applicable in the forecast of low academic performance [4].

This study proposes to predict student failure by using Data Mining. Data Mining is a process of extracting useful knowledge and information from data stored in databases and data warehouses. Data Mining is an integral part of KDD (Knowledge Discovery in Database) [1].

KDD can be used to learn the model for the learning process or student modeling. In fact, we want to detect the factors that most affect student failure in young students by using classification techniques and experiments taken to improve the accuracy for predicting which student might fail by first using all available attributes, and then selecting best attribute. Also we use different Data Mining techniques because data have high dimensionality (there are many factors that can influence) and often highly unbalanced (the majority of students pass and too few fail).

This paper concentrates on outlining different techniques that will help the teachers to make sense of the feeble students and enhance their instructive principles and environment in which they learn. We propose the use of data mining procedures, because the complexity of the problem is high that is data to be handled is very large and often highly unbalanced as very few student fail.

The objective of this paper is to detect the failure of student's as early as possible who show these factors so that we can provide some type of assistance for trying to avoid or reduce school failure and to prevent them from dropping out and improve their academic performance.

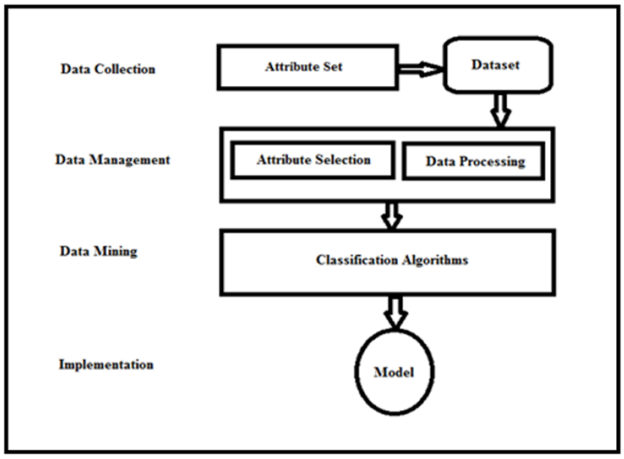
The paper is organized as follows: Section II presents our proposed method for pre-recognizing failure. Section III Describes data used and the information sources from we collected. Section IV describes the data preprocessing step. Section V describes the different experiments carried out and the results obtained. In section VI, we present the interpretation of our results and finally in section VII, summarizes the main conclusions and future research.

II. METHOD PROPOSED

The method proposed in this paper for pre-recognizing the student's academic failure uses the phases like data gathering, data pre-processing, data mining and interpretation of results.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 5, Issue 12, December 2016)

1) Data Collection: This stage comprises in gathering all accessible data on students. To do this, the arrangement of components that can influence the student execution must be distinguished and gathered from the distinct sources of information available. At last, all the data ought to be incorporated into a dataset as appeared in figure 1.



2) Pre-processing: At this stage data mining techniques are applied on the dataset. To do this, pre-processing techniques such as data cleaning, transformation of variables, and data partitioning are applied. Other techniques such as the selection of attributes and the re-balancing of data have also been applied for solving the problems of high dimensionality and imbalanced data that are presented in these datasets.

3) Data Mining: At this stage, DM algorithms like a classification problem are applied to predict student failure. To do this, we used classification algorithms based on rules and decision trees. Decision tree algorithms are like Simple cart, Random Tree etc. These are white box techniques that generate easily interpretable models. Finally, different algorithms have been executed, evaluated and compared in order to determine which one obtains the best results.

4) Implementation: At this stage, the obtained models are analyzed to detect student failure. To achieve this, the factors that appear (in the rules and decision trees) and how they are related are considered and interpreted.

Next, we describe a case of study with real data from engineering students in order to show the utility of the proposed method.

III. Data Collection

Data Collection is a process where information about the student's is collected. This information is nothing but the data that will be useful in predicting the failure of student's in academics.

In this paper, we have used information about engineering student enrolled for Bachelors of Technology from Maharashtra Institute of Technology at Dr. Babasaheb Ambedkar Marathwada University, Aurangabad for the year 2010/14 and 2014/18 academic years. Engineering education system is of four academic years when the students are aged between 18-24 years old. We have used information of 2010/14 year students when they are in Final year and used information of 2014/18 when they are in third year (that is 2016/2017).

The information about student is gathered in three unique classes as shown below in Table I

1) First type of data collected from Specific Information where family and personal information of the student is gathered. For example, StudenName, Age, FamilyIncome, Gender, NoofFamilyMembers, FatherOccupation, MotherOccupation, AnnualIncome, NoofFamilyMembersEarn, SmokingHabits, category etc.

2) The second type of data collection Academic Information of the student's. This data is the data that is required by multiple higher and secondary educational institutions while enrolling the student's in their institutions. For example, StudentName, totalFess, Time_Spent_During_Study, Distance_from_Home_to_college,age, takingNotesInClass, transport_method_used_to_go_college, LevelofAttendenceDuringClass, etc.

3) The third type of data collection is departmental Information where each department's subject wise information of a student is collected. For example, Marksinsem-1, Marksinsem-2, Marksinsem-3, Marksinsem-4, Marksinsem-5, Marksinsem-6 etc.

All this information is then stored in the dataset.

IV. Data Pre-Processing

Before applying DM algorithm it is necessary to do some pre-processing tasks such as cleaning, integration, discretization and variable transformation [5].

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 5, Issue 12, December 2016)

Data pre-preparing is essential undertaking in this work, because of the quality and dependability of accessible data, which straightforwardly influences the outcomes acquired. Actually, some particular pre-processing techniques were applied to create all the previously described data so that the classification could be done effectively. Firstly, all available data were collected into a single dataset. During this process those students without 100% complete information were removed. All students who did not answer our specific survey were excluded. Some modifications were also made to the values of some attributes.

However, our dataset has two issues that typically show up in these sorts of educational data. From one viewpoint, our data set has high dimensionality; meaning, the quantity of attributes becomes very huge. Further, given huge number of attributes, some will not be meaningful for classification and it is likely that few attributes are correlated. From other viewpoint, the data is imbalanced, that is the large set of students passed and minority failed. The problem with imbalanced data appeared because learning algorithms tend to ignore less frequent classes (minority classes) and only focus on the most frequent ones (majority classes). As a result, the classifier obtained will unable to classify data correctly [6].

One of the important techniques of data management is the selection of attributes by applying attribute selection algorithms. The attribute selection algorithm select those attribute of student which have greater impact on their academic status. The attribute selection algorithms are as follows, CfsSubsetEval, Filtered-AttributeEval, FilteredSubsetEval, refer Table II. By using these attribute selection algorithms we can obtain the best attributes as shown in Table III, out of huge number of attributes of student's that affects the student's performance.

Table II:
Attributes Selected By Attribute Selection Algorithm

Attribute Selection Algorithms	Attributes Selected
CfsSubsetEval	StudenName, Age, FamilyIncome, Gender, NoofFamilyMembers, FatherOccupation, MotherOccupation, AnnualIncome, NoofFamilyMembersEarn,NoOfFriendsInClass, Studying_in_group, Marksinsem-1, Marksinsem-2, Marksinsem-3
FilteredAttributeEval	MotherOccupation, AnnualIncome, NoofFamilyMembersEarn, SmokingHabits, statename, localite_Hostelite, HavingPhysicalDisability, doing_part_time_job, having_critical_illness, having_personal_tutor, StudentName, totalFess, Time_Spent_During_Study, LevelofAttendenceDuringClass, takingNotesInClass, NoOfFriendsInClass, Studying_in_group, Marksinsem-1, Marksinsem-2
FilteredSubsetEval	Studying_in_group, Marksinsem-1, Marksinsem-2, Marksinsem-3, Marksinsem-4, Marksinsem-5, Marksinsem-6, Marksinsem-7, Marksinsem-8, totalFess, Time_Spent_During_Study

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 5, Issue 12, December 2016)

These algorithms were chosen because Weka [7] tool has implemented these algorithms that are specifically meant for attribute selection. And these algorithms are implemented for this case study without using Weka[19] tool. Table II shows the result of applying the attribute selection algorithms on all the attributes. To find which attributes to select as best attributes, the results of these algorithms were compared. The attribute that appears more than once (attribute should appear in at least two algorithms) in the result of the algorithms, is considered as the best attribute. For example, "Leve_of_attendance_during_class" attribute is selected by two algorithms. So it is considered to be the best attribute. Next, "Taking_notes_in_class" is also considered to be the best attribute as it is selected by all the three algorithms result; also "Studying_in_group" attribute is selected by two algorithms so it is also considered to be the best attribute, refer Table II. Best attributes selected are shown in Table III.

Table I:
Student's Information Sources (Input to the System)

Personal Information	StudenName, Age, FamilyIncome, Gender, NoofFamilyMembers, FatherOccupation, MotherOccupation, AnnualIncome, NoofFamilyMembersEarn, SmokingHabits, category, statename, localite_Hostelite, HavingPhysicalDisability, doing_part_time_job, having_critical_illness, having_personal_tutor
Academic Information	StudentName, totalFess, Time_Spent_During_Study, Distance_from_Home_to_college, transport_method_used_to_go_college, LevelofAttendenceDuringClass, takingNotesInClass, NoOfFriendsInClass, Studying_in_group
Departmental Information	Marksinsem-1, Marksinsem-2, Marksinsem-3, Marksinsem-4, Marksinsem-5, Marksinsem-6, Marksinsem-7, Marksinsem-8

Table III:
Best Attributes Selected

Best Attributes Selected	Marksinsem1, Marksinsem-2, Marksinsem-3, Marksinsem-4, Marksinsem-5 Marksinsem-6, Time_Spent_During_Study, leve_of_attendance_during_class, Studying_in_group,

V. Data Mining And Experimentation

This section depicts the data mining techniques used for getting the prediction models of students' academic status. We performed several experiments to obtain the highest classification accuracy. In a first experiment we executed 5 classification algorithms using all available information (32 attributes). In a second experiment, we used only the best attributes selected.

In this paper, decision trees and rules induction algorithms are used as "white box" classification techniques; that is, they give an explanation for the classification result and can be used directly for decision making. A decision tree is a set of conditions arranged 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 have a specific-to-general approach, in which obtained rules are generalized (or specialized) until a satisfactory description of each class is obtained.

The classification algorithms that we are going to use are two rules of induction algorithms; NNge (Non-Nested Generalized Exemplars algorithm) which forms a generalization each time a new example is added to the database, by joining to its nearest neighbor of the same class)[8]; OneR [6] (which is a 'One Rule' classification algorithm that generates one rule for each attribute) and two decision tree rules; RandomTree [6], (which randomly chooses N attributes at each node of the tree); SimpleCart [9], (which implements minimal cost-complexity pruning). We are also using another classification algorithm called Naive Bayes Algorithm [10] provided by Microsoft SQL Server Analysis Services. This algorithm is basically used for predictive modeling which is based on Bayesian Techniques.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 5, Issue 12, December 2016)

Finally, the results of all these executed algorithms is evaluated, compared and optimized to determine which one gives the best result. The result of the best algorithm is considered to be our prediction about failure. The decision tree algorithms, induction rules and naive bayes algorithms can be easily implemented in object-oriented programming. 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.

In the first experiment, all the classification algorithms were executed using tenfold cross-validation and all the available information, that is, the original data file with 32 attributes of 150 students. The results with the test files of classification algorithms are shown in Table IV.

Table IV:
Classification Results Using All Attributes

Algorithm	TP rate	Acc	GM
NNGE	81.63	75.21	88.78
OneR	85.23	78.21	86.56
Random Tree	76.34	56.78	67.32
Simple Cart	92.43	59.87	75.55

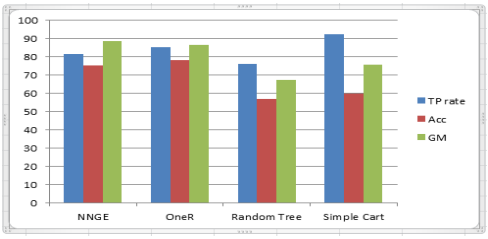


Figure 2. Graphical Representation of TABLE IV

This table shows you the results in the form of three fields i.e. TP Rate, Acc and GM. TP rate are the Passing Rate (It represents the percentage of students that may pass in the academic year), Acc is the (overall Accuracy Rate of each algorithm) and GM is the Geometric Mean. We can see in TABLE IV, the TP rate for SimpleCart algorithm is high and OneR is having the second largest TP rate.

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 III shows the best attributes. The results obtained from this experiment are shown in TABLE V. 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.

Table V:
Classifcation Results Using Best Attributes

Algorithm	TP Rate	Accuracy	GM
Nnge	84.5	76.5	80.6
OneR	87.9	67.8	79.8
Random Tree	76.5	46	65.4
Simple Cart	72.5	60.9	55.9

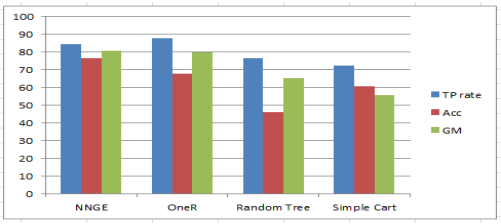


Figure 3. Graphical Representation of TABLE V

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 large number of student's 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 no two attributes of student's 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 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. The highlighted values show the comparison of the values obtained from naive bayes and other algorithms refer TABLE IV and TABLE VII.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 5, Issue 12, December 2016)

Table VI:
Classification Results Of Naive Bayes Using All Attributes

Algorithm	TP rate	TN rate	Acc	GM
NNGE	85.23	65.54	72.65	**84.65**
OneR	89.22	61.62	73.55	70.3
Random Tree	76.34	**81.76**	56.78	67.32
Simple Cart	**92.43**	80.09	59.87	75.55
Navie Bayes	88.96	85.81	**78.89**	81.56

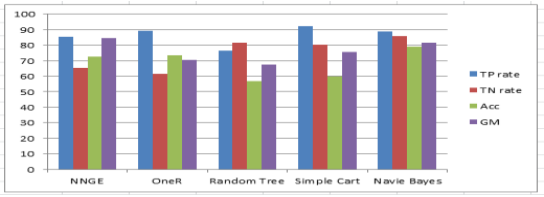


Figure 4. Graphical Representation of TABLE VI

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, Navie Bayes has the highest Accuracy rate, RandomTree has the highest TN rate (It represents the percentage of students that may fail in the academic year) 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 student's is much higher in case of naive bayes algorithm.

Table VII:
Classification Results Of Naive Bayes Using Best Attributes

Algorithm	TP rate	TN rate	Acc	GM
NNGE	84.5	62.3	46	**80.6**
OneR	**87.9**	51.2	67.8	79.8
Random Tree	76.5	46.5	76.5	65.4
Simple Cart	72.5	**73.4**	60.9	59.9
Navie Bayes	**91.2**	**77.9**	**79.56**	**84.3**

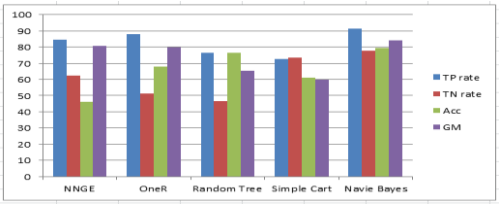


Figure 5. Graphical Representation of TABLE VII

Above TABLE VII shows classification result of all algorithm using best attributes only. Again, we can see that naive bayes algorithm gives all the maximum values for the properties. So the probability of finding the failure of student's is much higher in case of naive bayes algorithm.

VI. INTERPRETATION

Implementation is the last phase where the results obtained from DM techniques are interpreted into a model. For implementation, we have made use of .Net Technology with SQL Server as our backend for storing student's database.

VII. CONCLUSION

Precognition can never provide a fixed or a definite result as we wish for. It will always be uncertain and unpredictable. Through our work, we are trying to find out the probable number of students that might fail in future in their academics. In this paper, we have implemented all the algorithms on our own. We did not outsource the algorithms from Weka tool. In our entire work, we have implemented only two rules of induction, two decision tree algorithms and naive bayes algorithm and compared the results of these algorithms and found that naive bayes gives the best and more accurate result of prediction than the others. Through this paper, we have shown the percentage of students that might get failed in future using the data mining algorithms. The selection of the attributes of the student can be done manually or automatically using algorithms. We made this tool a real-time application which can be used in any educational organization for pre-recognizing the failure of student's.

The future enhancement would be to obtain more accurate and definite results of prediction, as prediction is probabilistic and uncertain. The scope of this project would be to predict the failure of student's and provide the necessary online information and online help and support for those students who are weak in respective subjects.

REFERENCES

[1] 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.

[2] Araque F., Roldan C., Salguero A. Factors Influencing University Drop Out Rates. Computers Education, 53, 563–574, 2009.

[3] S. Kotsiantis, K. Patriarcheas, and M. Xenos, "A combinational incremental ensemble of classifiers as a technique for predicting students' performance in distance education," Knowl. Based Syst., vol. 23, no. 6, pp. 529–535, Aug. 2010.

[4] J. Más-Estellés, R. Alcover-Arándiga, A. Dapena-Janeiro,A. Valderruten-Vidal, R. Satorre-Cuerda, F. Llopis-Pascual, T.Rojo-Guillén, R. Mayo-Gual, M. Bermejo-Llopis, J. Gutiérrez-Serrano, J. García-Almiñana, E. Tovar-Caro, and E. Menasalvas-Ruiz,"Rendimiento académico de los estudios de informática en algunoscentros españoles," in Proc. 15th Jornadas Enseñanza Univ. Inf.,Barcelona, Rep. Conf., 2009, pp. 5–12.

[5] E. Espíndola and A. León, "La deserción escolar en américa latina:Un Tema prioritario para la agenda regional," Revista Iberoamer. Educ., vol. 1, no. 30, pp. 39–62, 2002.

[6] 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.

[7] Sudhir B. Jagtap and Kodge B. G, "Census Data Mining and Data Analysis using WEKA," International Conference in "Emerging Trends in Science, Technology and Management – 2013, Singapore", ICETSTM – 2013, pp. 35-40.

[8] Prof. Nada Lavrac, " Case Study on the use of Data Mining Techniques in Food Science Using Honey Samples, " Jozef Stefan International Post Graduate School, Data Mining and Knowledge Discovery, February 2007, pp. 9.

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

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