Forecast Engineering Students Failure by Using Data Mining Techniques

Komal S. Sahedani [1],Prof. B Supriya Reddy[2]

[1]M.Tech Research Scholar, *C.E. Department, R.K. University,* komalsedani5@gmail.com
[2]Assistant Professor, *C.E. Department, R.K. University,* supriya.byreddy@rku.ac.in

Abstract:-This paper proposes to apply data mining techniques to predict engineering students failure and dropout. We use real data on 951 Engineering students from Rajkot, and employ classification methods, such as regression and decision trees. Experiments attempt to improve their accuracy for predicting which students might fail or dropout by first, using all the available attributes; next, selecting the best attributes.

Keywords: Data Mining, Education data mining (EDM), Knowledge discovery from data (KDD), Decision Tree, Logistic regression, dropout, failure ,prediction.

I. INTRODUCTION

Data mining is the iterative and interactive process of discovering valid, novel, useful, and understandable knowledge (patterns, models, rules etc.) in massive databases

The main term that data mining support for data is

Valid: generalize to the future
Novel: what we don't know
Useful: be able to take some action
Understandable: leading to insight
Iterative: takes multiple passes
Interactive: human in the loop

Many other terms carry a similar or slightly different meaning to data mining, such as knowledge mining from data, knowledge extraction, data/pattern analysis, data archaeology, and data dredging. [1]

Over past few years, many numbers of engineering institutes have opened rapidly in India. This causes a cut throat competition for attracting the student to get them enroll in their campus. Most of the institutes are opened in self-finance mode, so all the time they feel short hand in expenditure. Quality education is one of the most promising responsibilities of any University/ Institutions to their students. Quality education does not mean high level of knowledge produced. But it means that education is produced to students in efficient manner so that they learn without any problem. For this purpose quality education includes features like methodology of teaching, continuous evaluation, categorization of student into similar type, so that students have similar objectives, demographic, educational background etc. [2]

Engineering degrees are mostly offered in different curriculum structures. Engineering students are to fulfill strict requirements in order to graduate and hold a degree in engineering profession. Engineering students' accounts for numbers of departments mainly civil, electrical, and mechanical, computer, electronics, communication, information technology, chemical, mining, metallurgical, textile, and environment etc., Most of the engineering institutes' first five/six major courses.

This education is residential and at the beginning, student affects due to various factors related to their academic path. Most of the core courses are usually same for all the students in first year. They comprise essentially Mathematics, Physics and Chemistry courses. These course are the prerequisites of almost all major courses, students are exposed to the fundamental and basic concepts required to pursue specialized theories on their further studies. Core courses play a decisive role in the student performance and enrolled in this study.

So Due to a greater number of students and institutions, higher education institutions (HEIs) are becoming more oriented to performances and their measurement and are accordingly setting goals and developing strategies for their achievements.[5]

The recent literature related to Educational data mining (EDM) is presented. Educational data mining is an emerging discipline that focuses on applying data mining tools and techniques to educationally related data. Researchers within EDM focus on topics ranging from using data mining to improve institutional effectiveness to applying data mining in improving student learning processes.

The paper is organized as follows: Section II presents our proposed method for predicting students' failure. Section III describes data used and the information sources from we gathered. Section IV describes the data preprocessing step. Section V describes the different experiments carried out and the results obtained. In section VI, summarizes the main conclusions and future research.

II. METHOD

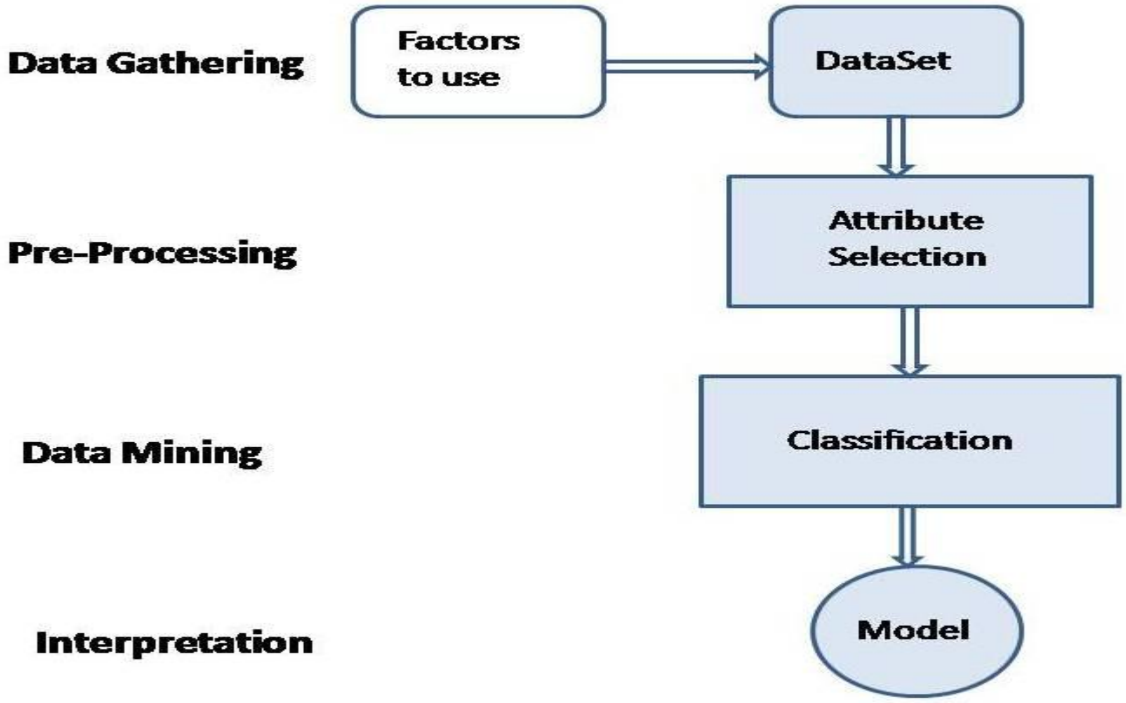

Figure 1: Method proposed for the prediction of student failure

The method proposed in this paper for predicting the academic failure of students belongs to the process of Knowledge Discovery and Data Mining (see Figure. 1). The main stages of the method are:

1) **Data gathering**. This stage consists in gathering all available information on students. To do this, the set of factors that can affect the students' performance must be identified and collected from the different sources of data available. Finally, all the information should be integrated into a dataset.

2) **Pre-processing**. At this stage the dataset is prepared to apply the data mining techniques. To do this, traditional pre-processing methods such as data cleaning, transformation of variables, and data partitioning have to be applied. Other techniques such as the selection of attributes and the re-balancing of data have also been applied in order to solve the problems of high dimensionality and imbalanced data that are typically presented in these datasets.

3) **Data mining**. At this stage, DM algorithms are applied to predict student failure like a classification problem. To do this task, we propose to use classification algorithms based on regression and decision trees. Finally, different algorithms have been executed, evaluated and compared in order to determine which one obtains the best results.

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

III. DATA GATHERING

Institute failure of student is also known as the "one thousand factors problem" [12], due to the large amount of risk factors or characteristics of the students that can influence institute failure, such as demographics, cultural, social, family, or educational background, socioeconomic status, psychological profile, and academic progress.

In this paper, we have used information of students enrolled in Engineering into RKU year 2011/1012 academic year Engineering offers a four year education program. We have only used information about first-year Engineering students, where most students are between the ages of 17 and 18, as this is the year with the highest failure rate. All the information used in this study has been gathered through Google form. And the detail of information is as given in table 1.

Variables	**Description**	**Possible Values**
SIClass	Students in classroom	{30-40 , 40-50 , 50-60 , 60-70 , 70-80 , More Than 80}
Attendence	Attendance in college	{Below 40% , 40%-50% , 50%-60% , 60%-70% , 70%-80% , 80%-90% , 90%-100%}
NoFriends	Number of friends	{ 0 , 1-10 , 10-20 , 20-30 , More Than 30}
hoursInstudy	No of hours spend in study	{0 , 1 , 2 , 3 , 4 , More}
dailyStudy	Daily study habit	{Yes , No}
MethodOfStudy	Method of study used	{Preparing Notes , Reading Books , Group Discussion}
Placedforstudy	Place used for study	{Home , College , With Friends}
ownspace	Own space for study	{Yes , No}
Resources	Resources for study	{ Encyclopedia , Books , Both}
StudyHabits	Study habits	{Daily , Weekly , Last Night Of Exam}
Studygroup	Group study	{Yes , No}
encouragement	Parental	{Yes , No}

	encouragement	
MaritalStatus	Marital status	{Married , Unmarried}
Religion	Religion	{Hindu , Muslim , Crystian , Jain , Sikh , Parsi}
Degree	Type of degree	{C.E. , I.T , E.C , Mechanical , Civil}
influenceofDegree	Influence of degree	{Excellent , Very Good , Good}
Personality	Type of personality	{Contributor , Non Contributor}
physicalDisability	Physical disability	{Yes , No}
SufferingIllness	Suffering from illness	{Critical , Normal Disease}
Tobacco	Tobacco chewing	{Yes , No}
Smoking	Smoking habits	{Yes , No}
Alcohol	Alcohol habits	{Yes , No}
FamilyIncome	Family income level	{ Low , Middle , High}
scholarship	Scholarship	{Yes , No}
HavingJob	Having a Job	{Yes , No}
LivingWithParents	Living WithOne’s parents	{Yes , No}
Mothereducation	Mother’s level of education	{No-education , Elementary , Secondary , Higher Secondary , Graduate , Post-graduate , Professional , Not-applicable}
FatherEducation	Father’s level of education	{No-education , Elementary , Secondary , Higher Secondary , Graduate , Post-graduate , Professional , Not-applicable}
FatherOccupation	Father’s occupation	{Cooley , Private , Govt.Service , Business , Farmer , Professional , Educational-institution , Retired , Not-applicable}
MotherOccupation	Mother’s occupation	{Housewife , Private , Govt.Service , Business , Farmer , Professional , Educational-institution , Retired , Not-applicable , Cooley}
MotherIncome	Mother’s monthly income	{0 .. 0.9k , 1k .. 2.9k , 3k ...4.9k , 5k ..9k , 10k.. 20k , Above 20k , Not-applicable}
FatherIncome	Father’s monthly income	{1.. 2.4k , 2.5k .. 4.9k , 5k .. 9.9k , 10k .. 14k , 15k ..25k , Above 25k , Not-applicable}
Weight	Student’s body mass index	{Underweight , Normal Weight , Over Weight , Obesity}
visualAcuity	Student’s eye visual acuity	{Normal , Defect}
community	Student’s community	{ Open , SEBC , SC , ST}
foodhabit	Student’s food habit	{Vegetarian , Non-vegetarian }
familysize	Student’s family	{ 1 ,2 ,3 ,4 ,5 ,6 ,7 ,8 ,9 , >9}

	size	
livingArea	Student's living area	{Urban , Rural , Semi-urban}
elderBrothers	Number of elder brothers	{0 , 1 , 2 , 3 , 4 , 5}
elderSisters	Number of elder sisters	{0 , 1 , 2 , 3 , 4 , 5}
youngerBrothers	No of younger brothers	{0 , 1 , 2 , 3 , 4 , 5}
youngerSisters	Number of younger sisters	{0 , 1 , 2 , 3 , 4 , 5}
familyStatus	Student's family status	{Individual , Joint}
transportation	Mode of transportation to college	{City-bus , College-bus , Own-vehicle , Hired-vehicle , Others}
vehicle	Own vehicle at home	{No-vehicles , Bicycle , 2-wheeler , 4-wheeler}
secondarymark	Marks/grade obtained at secondary level	{A+ – 90% -100% , A – 80% - 89% , B – 70% - 79% , C – 60% - 69% , D – 50% - 59% , E – 40% - 49% , F - < 40%}
Schoolmedium	Medium of school	{Gujarati , English}
privatetuition	No of subject in private tuitions	{0 , 1 , 2 , 3 , 4 , 5}
schoolType	Type of school	{Co-education , Boys , Girls}
Interestedsports	Interested in sports/athletic	{Yes , No}
homeCare	Care of study at home	{Parents , Grand-parents , Father Only , Mother Only , Self , Others}
Hscmarks	Marks/grade obtained at hsc level	{A+ – 90% - 100% , A – 80% - 89% , B – 70% - 79% , C – 60% - 69% , D – 50% - 59% , E – 40% - 49% , F - < 40%}
LivingInCity	Living in A large city	{Yes , No}
yearsInCity	No of years living in city	{ 0 , 1-5 , 5-10 , 10-15 , 15-20}
DistanceToClg	Distance to the college in km	{0 , 1-5 , 5-10 , 10-15 , 15-20 , 20-25 , 25-30 , >30}
AttendenceLevel	Level of attendance in class	{Low , Medium , High}
boredomLevel	Level of boredom in class	{Low , Medium , High}
InterestInSubjects	Interest In The	{Yes , No}

	Subjects	
LevelOfdifficulty	Level Of Difficulty In Subjects	{ Low , Medium , High}
TakingNotes	Taking Notes In Class	{Yes , No}
teachingMethod	Methods Of Teaching	{Traditional Classroom , Distance Learning , Both}
AssignmentDemand	Heavy Demand Of Assignment	{Yes , No}
TeacherConcern	Level Of Faculty' S Concern	{ Low , Medium , High}
Age	Age	{18 , 19 , 20 }
CPU	Grade in CPU	{A+ , A , B+ , B , C+ , C , D+ , D , F}
CS	Grade in CS	{A+ , A , B+ , B , C+ , C , D+ , D , F}
ES	Grade in ES	{A+ , A , B+ , B , C+ , C , D+ , D , F}
Maths-I	Grade in MathsI	{A+ , A , B+ , B , C+ , C , D+ , D , F}
ECE	Grade in ECE	{A+ , A , B+ , B , C+ , C , D+ , D , F}
EME	Grade in EME	{A+ , A , B+ , B , C+ , C , D+ , D , F}
Physics	Grade in Physics	{A+ , A , B+ , B , C+ , C , D+ , D , F}
EEE	Grade in EEE	{A+ , A , B+ , B , C+ , C , D+ , D , F}
MOS	Grade in MOS	{A+ , A , B+ , B , C+ , C , D+ , D , F}
Maths-II	Grade in MathII	{A+ , A , B+ , B , C+ , C , D+ , D , F}
CPI	Score in CPI	{Excellent,VeryGood,Good,Regular,Sufficient,Poor, VeryPoor}
Study	Study is continue?	{Yes,No}

Table I:Variables Used

The data was collected through the Google form Shared among students and it is filled by the student.

IV. DATA PRE-PROCESSING

Before applying DM algorithm it is necessary to carry out some pre-processing tasks such as cleaning, integration, discretization and variable transformation [13]. It must be pointed out that very important task in this work was data pre-processing, due to the quality and reliability of available information, which directly affects the results obtained. Firstly, all available data were integrated into a single dataset. The continuous variables were changed into discrete variables, which provide a much more understandable view of the data. For example, the numerical values of the scores obtained by students in each subject were changed to categorical values in the following way:
Excellent: score greater than 9.0; Very good: score between 8.0 and 9.0; Good: score between 7.0 and 8.0; Regular: score between 6.0 and 7.0; Sufficient: score between
5.0 And 6.0; Poor: between 4.0 and 5.0; Very poor: less than 4.0.
Then, all the information was integrated in a single dataset and it was saved in the .CSV (comma separated values) format. However, our dataset has two typical problems that normally appear in these

types of educational data. On the one hand, our data set has high dimensionality; that is, the number of attributes or features becomes very large. Further, given a large number of attributes, some will usually not be meaningful for classification and it is likely that some attributes are correlated. On the other hand, the data are imbalanced, that is the majority of students (950) passed and minority (110) failed. The problem with imbalanced data arises because learning algorithms tend 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.

We decide to carry out a study of feature selection to try to identify which feature has the greatest effect on our output variable (academic status). There are a wide range of attribute selection algorithms that can be grouped in different ways. One popular categorization is one in which the algorithms differ in the way they evaluate attributes and are classified as: filters, which select and evaluate features independently of the learning algorithm and wrappers, which use the performance of a classifier (learning algorithm) to determine the desirability of a subset [15]. Weka provides several feature selection algorithms from which we have selected the following ten: CfsSubsetEval, ChiSquaredAttributeEval, Consistency-SubsetEval, FilteredAttributeEval, OneRAttributeEval, FilteredSubsetEval, GainRatioAttributeEval, InfoGain-AttributeEval, ReliefFAttributeEval, SymmetricalUncert-AttributeEval. Table II shows the results of applying 10 algorithms of feature selection.

The results obtained were ranked by these 10 algorithms to select the best attributes from our 76 available attributes. To find the ranking of the attributes, we counted the number of times each attribute was selected by one of the algorithms. Table III shows the frequency of each attribute. From this table only those with a frequency greater than two have been considered by more than two feature selection algorithms. Finally, we selected only the attributes with frequency greater than two (attributes selected by at least two algorithms).In this way, we can reduce the dimensionality of our dataset from the original 76 attributes to only the best 15 attributes.

Finally, we have mentioned that our dataset is imbalanced. One way to solve this problem is to act during the pre-processing of data by carrying out a sampling or balancing of class distribution. There are several data balancing or rebalancing algorithms; one that is widely used and that is available in RGUI as a supervised data filter is SMOTE (Synthetic Minority Oversampling Technique). In the SMOTE algorithm, the minority class is over-sampled by taking each minority class sample and introducing synthetic examples along the line segments joining any or all of the *k* minority class nearest neighbors [16]. In our case, only the training files (with the best 8 attributes) have been rebalanced using the SMOTE algorithm, obtaining 50% Pass students and 50% failed students.

Algorithm	Attribute Selected
Cfssubseteval+GreedyStepwise	dailyStudy; MethodOfStudy; Resources; encouragement;Degree; influenceofDegree;familysize; youngerBrothers; younger Sisters; schoolType; TeacherConcern; MATHS-II;
ChiSquaredAttributeEval+ Ranker	Familysize;MATHS-II;CS;hoursInstudy;FatherOccupation; DistanceToClg;FatherEducation;Degree;Attendence;Physics; Mothereducation;MOS;ECE;EEE;ES;
ConsistencySubsetEval + GreedyStepwise	Familysize
FilteredAttributeEval+ Ranker	familysize; MATHS-II; CS; hoursInstudy; FatherOccupation; DistanceToClg; FatherEducation; Degree; Physics; Attendence;

	MOS; ECE; Mothereducation; EEE; ES;
FilteredSubsetEval + GreedyStepwise	familySize
GainRatioAttributeEval+ Ranker	SufferingIllness; familysize; CS; MATHS-II; hoursInstudy; Degree; DistanceToClg; Religion; FatherOccupation; influenceofDegree; Attendence; FatherEducation; youngerBrothers; Physics; TeacherConcern;
InfoGainAttributeEval+ Ranker	familysize; MATHS-II; CS; hoursInstudy; FatherOccupation; DistanceToClg; FatherEducation; Degree; Physics; Attendence; MOS; ECE; Mothereducation; EEE; ES;
OneRAttributeEval+ Ranker	influenceofDegree; encouragement; Alcohol; LivingInCity; livingArea; Physics; Interestedsports; yearsInCity; Schoolmedium; Resources; HavingJob; MethodOfStudy; dailyStudy; EME; CS;
ReliefFAttributeEval+ Ranker	tobacco; youngerBrothers; HavingJob; schoolType; AssignmentDemand; Degree; MATHS-II; homeCare; familysize; AttendenceLevel; MethodOfStudy; Religion; Schoolmedium; Studygroup; EME;
SymmetricalUncertAttributeEval +Ranker	familysize; MATHS-II; CS; hoursInstudy; FatherOccupation; DistanceToClg; Degree; FatherEducation; Attendence; Physics; MOS; ECE; Mothereducation; EEE; Religion;

Table II: Best Attributes Selected

Attribute	**Frequency**
Degree	10
familysize	9
MATHS-II	7
CS	6
Attendence	6

Attribute	**Frequency**
Physics	6
DistanceToClg	5
hoursInstudy	5
FatherOccup	5
FatherEducation	5

Attribute	**Frequency**
MOS	4
EEE	4
Mothereducation	4
ECE	4
influenceDegree	3

Table III: Most Influential Attributes Ranked By Frequency of Appearance

Then create a MLR (Multiple Logistic Regression) Model into RGUI from these 15 attributes and By using wald test choose the attributes whose probability value is lower because those attributes are affect more in predicting students failure. The Table IV shows those attributes and their Probability value. The attributes whose probability is nearest to Zero are selected.

Attribute	Probability
Degree	0.000141
MathsII	0.000718
Physics	0.009521

Attribute	Probability
FatherEducation	0.000502
EEE	0.009416
hoursInstudy	0.009286

Attribute	Probability
CS	0.000885
DistanceToClg	0.001475

Table IV: Attributes affect more for Prediction

V. DATA MINING AND EXPERIMENTATION

This section describes the experiments and data mining techniques used for obtaining the prediction models of students' academic status at the end of the first year.

We performed several experiments in order to try to obtain the highest classification accuracy. In a first experiment we executed 10 classification algorithms using all available information (76 attributes). In a second experiment, we used only the 15 attributes selected from weka tool. In a third experiment,we use the best 8 attributes selected using wald test from RGUI. In the Final Experiment, we repeated the executions by using re-balanced data files.

In this paper, decision trees and logistic regression are used as they are prediction methods of classification techniques; that is, they provide an explanation for the classification result and can be used directly for decision making. 5 commonly used classical decision tree algorithm and 1 Multiple logistic regression Method of classification algorithms that are available in the RGUI statistical tool have been used:

1. **One Regression Model**: MLE(Multiple Logistic Regression),which generalizes logistic regression to multiclass problems, That is, it is a model that is used to predict the probabilities of the different possible outcomes of a categorically distributed dependent variable, given a set of independent variables.
2. **Five Decision Tree Algorithms**: CTREE(Conditional inference tree),Recursive partitioning for continuous, censored, ordered, nominal and multivariate response variables in a conditional inference framework .RPART(Recursive partitioning),which creates a decision tree that strives to correctly classify members of the population based on several dichotomous dependent variables. C5.0, discovering patterns that delineate categories, assembling them into classifiers ,and using them to make predictions .RF (Random Forest), which operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes output by individual trees.CHAID(Chi-squaredAutomatic Interaction Detector), which is a type of decision tree technique,based upon adjusted significance testing.

Algorithm	**Sensitivity**	**Specificity**	**Accuracy**	**GM**
MLR	**100%**	**100%**	**100%**	**100%**
Rpart	79.21%	93.65%	92.11%	76.27%
Ctree	79.21%	93.65%	92.11%	93.74%
C5.0	75.25%	99.06%	96.53%	93.74%
rF	53.47%	**100%**	95.06%	97.34%
Chaid	84.16%	96.82%	95.48%	86.28%

Table V: Classification Results Using All Attributes

All the classification algorithms were execute and all the available information, that is, the original data file with 76 attributes of 950 students. The results with the test files of classification algorithms are shown in Table V. This table shows the rates or percentages of correct classifications for each of the two classes: Pass (Specificity) and Fail (Sensitivity) and the overall Accuracy rate (Accuracy). It can be seen in Table V that the percentage of accuracy obtained for total accuracy (Acc) and for Pass (Specificity) are high, but not for Fail (Sensitivity). Specifically, the algorithms that obtain the maximum values are: MLR (Sensitivity, Specificity, Accuracy ,GM) and rF (Specificity and GM).

Algorithm	**Sensitivity**	**Specificity**	**Accuracy**	**GM**
MLR	44.55%	97.77%	92.11%	65.0%
Rpart	**61.39%**	96.71%	92.95%	81.10%
Ctree	0%	**100%**	89.38%	81.27%
C5.0	27.73%	98.71%	91.17%	81.27%
rF	52.47%	**100%**	**94.95%**	**97.29%**
Chaid	0%	**100%**	89.38%	89.08%

Table VI: Classification results Using 15 Attributes

In the second experiment, we executed all the classification algorithms using the reduced dataset (with only the best 15 attributes). Table VI shows the results with the test files using only the best 15 attributes. When comparing the results obtained with the previous results obtained using all the attributes, that is, Table V versus Table VI, we can see that in general all the algorithms have improved in some measures (Sensitivity and GM). Furthermore, with regard to the others measures (Specificity and Accuracy) there are some algorithms that obtain a slightly worse or slightly better value, but they are very similar in general to the previous ones. In fact, the maximum values obtained are now better than the previous ones obtained using all attributes. Again the algorithms that obtain these maximum values are rF (Specificity and Accuracy), rpart(Sensitivity and Accuracy) and chaid (Specificity and GM).

Algorithm	**Sensitivity**	**Specificity**	**Accuracy**	**GM**
MLR	29.70%	98.12%	90.85%	77.0%
rpart	**58.42%**	96.82%	92.74%	80.79%
ctree	8.91%	99.53%	89.91%	87.25%
C5.0	46.54%	98.71%	93.17%	87.25%
rF	51.49%	**99.65%**	**94.53%**	**94.54%**
chaid	23.76%	97.65%	89.80%	70.65%

Table VII:Classification Results Using Best Attributes

In the third experiment, we executed all the classification algorithms using the reduced dataset (with only the best 8 attributes). Table VII shows the results with the test files using only the best 8 attributes. When comparing the results obtained with the previous results the algorithms that obtain these maximum values are rF(Specificity and Accuracy) and rpart(Sensitivity and Accuracy)

Algorithm	**Sensitivity**	**Specificity**	**Accuracy**	**GM**
MLR	82.51%	87.38%	85.29%	84.98%
rpart	80.53%	88.12%	84.87%	84.66%
ctree	72.61%	90.84%	83.03%	83.56%
C5.0	**87.79%**	95.30%	**92.08%**	**92.27%**
rF	85.81%	92.33%	89.53%	89.50%
chaid	66.34%	**96.04%**	83.31%	85.64%

Table VIII: Classification Results Using Data Balancing

In the fourth experiment, we again executed all the classification algorithms using the rebalanced training files (using The SMOTE algorithm) with only the best 8 attributes. The results are shown in Table VII, and algorithms that obtain these maximum values are chaid and C5.0.

From the above Table V, VI, VII, VIII we see that using MLR rF, chaid, C5.0 we can get more Accuracy. And by Figure 2 we can say that using 76 attributes and 8 attributes we can get more Accuracy.

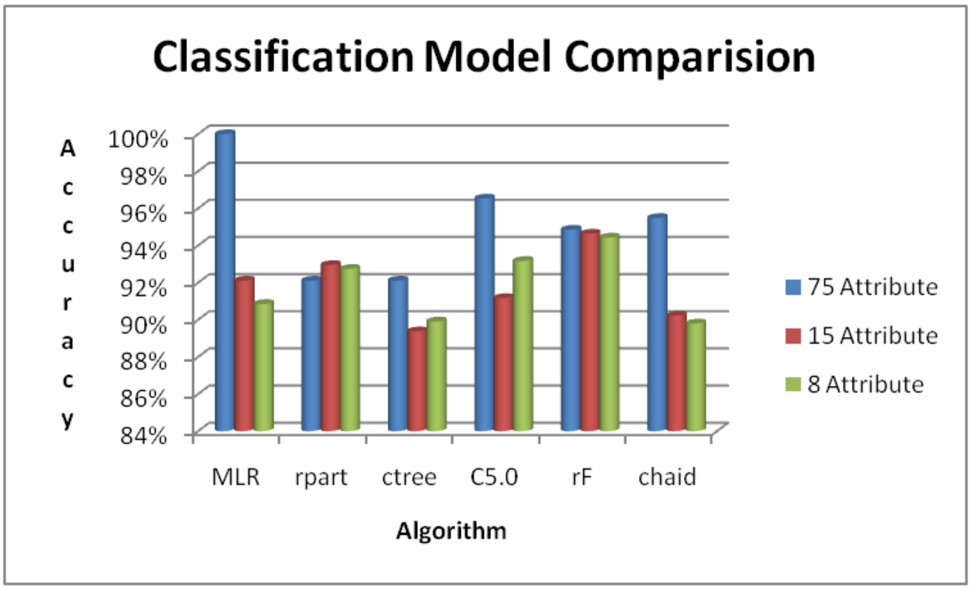

Figure II: Classification Model comparision based on Different Attributes

VI. CONCLUSION

As we have seen, predicting student failure at Institute can be a difficult task not only because it is a multifactor problem (in which there are a lot of personal, family, social, and economic factors that can be influential) but also because the available data are normally imbalanced. To resolve these problems, we have shown the use of different DM algorithms and approaches for predicting student failure. We have carried out several experiments using real data from Engineering Institute students in Rajkot. We have applied different classification approaches for predicting the academic status or final student performance at the end of the course. Furthermore we have shown that some approaches such as selecting the best attributes and data balancing can also be very useful for improving accuracy.

It is important to notice that gathering information and pre-processing data were two very important tasks in this work. In fact, the quality and the reliability of the used information directly affects the results obtained. However, this is an arduous task that involves a lot of time to do.

In general, regarding the DM approaches used and the classification result obtained, the main conclusions are as follows:

1) We have shown that classification algorithms cab be used successfully in order to predict a student’s academic performance and, in particular, to model the difference between Fail and Pass students.

2) We have shown the utility of feature selection techniques when we have a great number of attributes. In our case, we have reduced the number of attributes used from the 76 initially available attributes to the 15 attributes and then 8 best attributes that affect more in predicting students’ dropouts.

3) We have shown way to address the problem of imbalanced data classification by rebalancing the data .

Starting from the previous models (Regression and decision trees)generated by the DM algorithms, a system to alert the faculty and their parents about students who are potentially at risk of failing or drop out can be implemented. As an example of possible action, we propose that once students were found at risk, they would be assigned to a tutor in order to provide them with both academic support and guidance for motivating and trying to prevent student failure.

As future work, we can mention the following:

1) To develop our own algorithm for classification/prediction based on grammar using genetic programming that can be compared versus classic algorithms.

2) To propose actions for helping students identified within the risk group. Then, to check the rate of the times it is possible to prevent the fail or dropout of that student previously detected.

REFERENCES

[1] U. Fayyad and R. Uthurusamy, “Data mining and knowledge discovery in databases,” Commun. ACM, vol. 39, pp. 24–27, 1996.

[2] Shiv Kumar Gupta,Sonal Gupta &Ritu Vijay,” prediction of student success that are going to enroll in the Higher technical education”, IJCSEITR, ISSN 2249-6831, Vol. 3, Issue 1, Mar 2013, pp. 95-108.

[3] Richard A. Huebner,Norwich University,”A survey of educational data- mining research”, Research in Higher Education Journal,2012,pp-1-13

[4] C. Romero *, S. Ventura.(2007)”Educational data mining: A survey from 1995 to 2005”,ScienceDirect Expert Systems with Applications 33 pp. 135–146,2007.

[5] ZeljkoGaraca, MajaCukusic, Mario jadric (2010), “Student Dropout Analysis with application of data mining methods”,Vol 1,pp. 31-46

[6] U. Fayyad and R. Uthurusamy, “Data mining and knowledge discovery in databases,” Commun. ACM, vol. 39, pp. 24–27 , 1996.

[7] Jiawei Han and MichelineKamber, “Data Mining Concepts and Techniques”,2nd Edition, 2000.

[8] J. R. Quinlan, “Introduction of decision tree”, Journal of Machine learning”, pp. 81-106, 1986.

[9] Yoav Freund and Llew Mason, “The Alternating Decision Tree Algorithm”. Proceedings of the 16th International Conference on Machine Learning, pp. 124-133, 1999.

[10] Bernhard Pfahringer, Geoffrey Holmes and Richard Kirkby. “Optimizing the Induction of Alternating Decision Trees”. Proceedings of the Fifth Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, pp. 477-487, 2001.

[11] Saurabh Pal.” Mining Educational Data to Reduce Dropout Rates of Engineering Students”, IJIEEB,April-2012, Vol-2,pp.1-7.

[12] M. M. Hernandez, “Causas del fracaso escolar,” in Proc. 13th Congr. Soc. Espanola Med. Adolescente, 2002, pp. 1–5.

[13] E. Espíndola and A. León, “La desercion escolar en americalatina: Un Temaprioritario para la agenda regional,” RevistaIberoamer. Educ., vol. 1, no. 30, pp. 39–62, 2002.

[14] M. Ramaswami and R. Bhaskaran , ” A CHAID Based Performance Prediction Model in Educational Data Mining” , IJCSI , Vol. 7 , Issue 1 , No. 1 , January 2010 , pp.10-18

[15] M. A. Hall and G. Holmes, “Benchmarking attribute selection techniquesfor data mining,” Dept. Comput. Sci., Univ. Waikato, Hamilton, NewZealand, Tech. Rep. 00/10, Jul. 2002.

[16]N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “Synthetic minority over-sampling technique,” J. Artif. Intell. Res., vol. 16, pp. 321–357, Jun. 2002.

[17]M. Ramaswami and R. Bhaskaran , ” A CHAID Based Performance Prediction Model in Educational Data Mining” , IJCSI , Vol. 7 , Issue 1 , No. 1 , January 2010 , pp.10-18

[18] W. Hamalainen and M. Vinni,” Classifiers for educational data mining”, 2008, pp.1-34

[19] Mohammad Hassan Falakmasir, JafarHabibi,” Using Educational Data Mining Methods to Study the Impact of Virtual Classroom in E-Learning”, 2010, pp.241-248

[20] HyonamJeong,”How Students’ Self-motivation and Learning Strategies Affect Actual Achievement”, Department of Computer Science, Indiana University-Purdue University Fort Wayne

[21] Kun Liu, Yan Xing,” A Lightweight Solution to the Educational Data Mining Challenge”,2010.

[22] A.S. Kavitha,R. Kavitha, J. Viji Gripsy,” Empirical Evaluation of Feature Selection Technique in Educational Data Mining”, ARPN Journal of Science and Technology, VOL. 2, NO. 11, Dec 2012.

[23] Carlos Marquez-Vera , Cristobal Romero Morales , and Sebastián 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.