Performance Prediction of Engineering Students using Decision Trees

R. R. Kabra
S.G.R. Education Foundation's College of Engineering and Management, Ahmednagar, India.

R. S. Bichkar
G. H. Raisoni College of Engineering and Management, Pune, India.

ABSTRACT
Data mining can be used for decision making in educational system. A decision tree classifier is one of the most widely used supervised learning methods used for data exploration based on divide & conquer technique. This paper discusses use of decision trees in educational data mining. Decision tree algorithms are applied on engineering students' past performance data to generate the model and this model can be used to predict the students' performance. It will enable to identify the students in advance who are likely to fail and allow the teacher to provide appropriate inputs.

General Terms
Data Mining.

Keywords
Classification, Decision trees, Educational Data Mining.

1. INTRODUCTION
An educational system has large number of educational data. This data may be students' data, teachers' data, alumni data, resource data etc. Educational data mining is used to find out the patterns in this data for decision-making.There are two types of education system:

1) Traditional Education system: In this system there is direct contact between the students and the teacher. Students' record including the information such as attendance, grades may be kept manually or digitally. Students' performance is the measure of this information.

2) Web based learning system: It is also known as e-learning. It is becoming more popular as the students can learn from any place without any time constraint. In a web based system, various data about the students are automatically collected through logs.

Educational data mining can answer number of questions from the patterns obtained from student data such as

1) Who are the students at risk?
2) What are the chances of placement of student?
3) Who are the students likely to drop the course?
4) What is the quality of student participation?
5) Which courses the institute should offer to attract more students?

Results of educational data mining can be used by different members of education system [1], [2]. Students can use them to identify the activities, resources and learning tasks to improve their learning. Teachers can use them to get more objective feedback, to identify students at risk and guide them to help them succeed, to identify the most commonly made mistakes and to organize the contents of site in efficient way. On the other hand, administrators can use them to decide which courses to offer, which alumni are likely to donate more to the institution etc.

This paper describes the model that predicts the academic performance of the engineering students in contact education system. The following scenario will help us to understand the importance of this model. As the number of engineering seats and colleges are increasing in India, the inferior students are also enrolled in engineering courses. So the results of the universities for engineering courses are going down. If we know 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 will finally help in improving placements. Good placement is one of the key factors that will help the college to attract students.

Most of the features selected for creating the model are based on students' past performance, as we feel that the past performance of a student is indicative of his future performance in most of the cases. Also, it is difficult to obtain the social data. For example, students are reluctant to reveal the information like parents' income and may provide incorrect data. In this paper, we would like to find the correlation of the past performance with future performance prediction.

Different data mining techniques can be applied in education systems like clustering, classification, outlier detection, association rule mining and sequential pattern mining. This paper discusses classification of engineering student data using decision trees.

2. CLASSIFICATION
Classification is to build structures from examples of past decisions that can be used to make decisions from unseen cases [3]. Data classification is a two step process. In the first step, a model is built by analyzing the data tuples from training data having a set of attributes. For each tuple in the training data, the value of class label attribute is known. Classification algorithm is applied on training data to create the model. In the second step of classification, test data is used to check the accuracy of the model. If the accuracy of the model is acceptable then the model can be used to classify the unknown data tuples (i.e. for which the class label is not unknown). Basic techniques for classification are decision tree induction, Bayesian classification, Bayesian belief networks and neural networks. Other approaches like genetic algorithms, rough sets, fuzzy logic, case based reasoning can also be used for classification.

3. DECISION TREE INDUCTION

3.1 Decision trees

A decision tree is a flow-chart-like tree structure, where each internal node is denoted by rectangles, and leaf nodes are denoted by ovals. All internal nodes have two or more child nodes. All internal nodes contain splits, which test the value of an expression of the attributes. Arcs from an internal node to its children are labeled with distinct outcomes of the test. Each leaf node has a class label associated with it.

A decision tree is constructed from a training set, which consists of data tuples. Each tuple is completely described by a set of attributes and a class label. Attributes can have discrete or continuous values. Decision trees are used to classify the data tuples whose class label is unknown. Based on the attribute values of the tuple, the path from root to a leaf can be followed. The class of the leaf is the class predicted by decision tree for that tuple.

3.2 Classification by Decision Tree Induction

The task of constructing a tree from the training set has been called tree induction or tree building. Most existing tree induction systems adopt a greedy (i.e. non-backtracking) top-down divide and conquer manner. Starting with an empty tree and the entire training set, following algorithm is applied on the training data (where each tuple is associated with a class label) until no more splits are possible [3].

Algorithm:

1) Create a node N.
2) If all the tuples in the partition are of the same class then return N as a leaf node labeled with that class.
3) If attributes list is empty then return N as a leaf node labeled with the most common class in samples.
4) Identify the splitting attribute so that resulting partitions at each branch are as pure as possible.
5) Label node N with splitting criterion which serves as test at that node.
6) If splitting attribute is discrete valued then remove splitting attribute from attribute list.
7) Let *Pi* be the partitions created based on the *i* outcomes on splitting criterion.
8) If any *Pi* is empty then attach a leaf with the majority class in the partition to node N.
9) Else recursively apply the complete process on each partition.
10) Return N.

3.3 Decision Tree Algorithms

ID3 algorithm introduced by J. R. Quinlan [4] is a greedy algorithm that selects the next attributes based on the information gain associated with the attributes. The attribute with the highest information gain or greatest entropy reduction is chosen as the test attribute for the current node.

C4.5, the most popular algorithm, is a successor of ID3. C4.5 made a number of improvements to ID3. C4.5 uses Gain ratio [5] as an attribute selection measure. Also C4.5 can handle both discrete and continuous attribute.

CART algorithm, which was proposed by Breiman, is conceptually is same as that of ID3. The impurity measure used in selecting the variable in CART is Gini index [5]. If the target variable is nominal it generates classification tree and for continuous-valued numerical target variable it generates regression tree.

CHAID uses Chi square contingency test for tree construction in two ways [6]. First, it determines whether levels in the predictor can be merged together. Once all predictor level are compressed to their smallest significant form, it determines most significant predictor in distinguishing among the dependent variable levels.

4. RELATED WORK

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 to predict 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 results were obtained using different classification techniques. It was found that the tree based algorithms outperformed the methods like Neural Networks and SVM.

Z. J. Kovacic presented a case study on educational data mining in [9] to identify up to what extent the enrolment data can be used to predict student's success. The algorithms CHAID and CART were applied on student enrolment data of information system students of open polytechnic of New Zealand to get two decision trees classifying successful and unsuccessful students. The accuracy obtained with CHAID and CART was 59.4 and 60.5 respectively.

M. Ramaswami and R. Bhaskaran [10] used the CHAID prediction model to analyze the interrelation between variables that are used to predict the outcome on the performance at higher secondary school education in India. The features like medium of instruction, marks obtained in secondary education, location of school, living area and type of secondary education were the strongest indicators for the student performance in higher secondary education. This CHAID prediction model of student performance was constructed with seven class predictor variables with accuracy 44.69%.

Thai-Nghe, Drumond, Krohn-Grimberghe, Schmidt-Thieme [11] have used recommender system technique in educational data mining to predict student performance.

In India, after higher secondary education students have to take crucial decision which branch to choose so that there will be good chances of placement. Elayidom, Idikkula, J. Alexander, A. Ojha [12] created the decision tree which helps admission seekers to choose a branch with high industrial placement. The data was supplied by National Technical Manpower Information System (NTMIS) via Nodal center. Data was compiled by them from feedback by graduates, post graduates, diploma holders in engineering from various engineering colleges and polytechnics located within the state during the year 2000-2003. The standard database is processed to get a table, in which corresponding to each input combination, the percentage placement is computed.

Nghe, Janecek, and Haddawy [13] compared the accuracy of decision tree and Bayesian network algorithms for predicting the academic performance of undergraduate and postgraduate students at two very different academic institutes: Can Tho University (CTU), a large

national university in Viet Nam, and the Asian Institute of Technology (AIT), international university in Thailand. It was found that decision trees are 3-12% more accurate than Bayesian Networks.

V. P. Bresfelean, M. Bresfelean and N. Ghisoiu [14] 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.

A. Merceron and K. Yacef [15] presented how pedagogically relevant knowledge can be discovered from web-based educational system. The authors built the decision trees from the student data of Logic-ITA web based tutoring tool used at Sydney university to generate *if then rules* which predict student marks he is likely to achieve.

Baradwaj and Pal [16] obtained the university students data like attendance, class test, seminar and assignment marks from the students' previous database, to predict the performance at the end of the semester.

5. CASE STUDY

Institute's success highly depends upon students' success in that institute. Knowing the reasons of failure of student can help the teachers and administrators to take necessary actions so that the success percentage can be improved. The data is collected from S. G. R. Education Foundation's College of Engineering and Management. The institute has been started in the year 2008 and is affiliated to University of Pune in Maharashtra, India. It is very important for a newly started institute to improve the students success rate every year to attract superior students.

5.1 Data Selection and Preprocessing

Data of 346 students of the institute is collected who appeared for the first year of engineering in the year 2009-10, 2010-11. The data was collected through the enrolment form filled by the student at the time of admission. The student enter their demographic data (category, gender etc), past performance data (SSC or 10^{th} marks, HSC or 10 + 2 exam marks etc.), address and contact number. From these the attributes that possibly influence their result are selected as shown in Table 1. Most of the attributes reveal the past performance of the students. Reason behind concentrating on the past performance data is

1. Data is easily available in the administrative department of the institute.
2. If student has performed well in the past, it is most likely that he will perform better in subsequent exams as well.

The atteributes are described below.

Branch - The courses offered by institute Computer Engineering (COMP), IT engineering (IT), Electronics and Telecommunication (ETC), Electronics (ELX), Mechanical Engineering (MECH).

HSCPercent - The percentage of marks obtained by student in Higher secondary class.

HSCMaths - Marks in HSC Mathematics

Table 1: Attributes and their domain

Sr No.	Name	Possible values
1	Branch	COMP, IT, ELX, ETC, MECH
2	HSCPercent	Distinction(above 75%), Firstclass(60%-75%), HSecondclass(50%-60%), Secondclass(less than 50%)
3	HSCMaths	Real
4	HSCPCM	Real
5	HSCCET	Real
6	SSCpercent	Distinction(above 75%), Firstclass(60%-75%), HSecondclass(50%-60%), Secondclass(less than 50%)
7	SSCMaths	Real
8	SSCScience	Real
9	Category	Open, OBC, SC, ST, Others
10	Gender	Male, Female
11	Livinglocation	Rural, Urban
12	SSCBoard	State, CBSE
13	Atype	CAP, MGMT
14	Father_Occupation	PublicSectorjob, Privatejob, business, Farmer, Teacher, Other
15	Mother_Occupation	PublicSectorjob, Privatejob, business, Farmer, Teacher, Other
16	SSCMedium	English, Regional
17	Feresult (target variable)	PASS, FAIL, ATKT

HSCPCM - Sum of Physics, Chemistry and Mathematics marks in HSC exam (i.e. out of 300).

HSCCET - Marks obtained in common entrance test. The entrance test is compulsory for the student to get admission in engineering course. The Maharashtra state CET is out of 200. The All India Engineering Entrance Examination (AIEEE) is held on all India basis and its marks are converted to be out of 200 accordingly.

SSCpercent - The percentage of marks obtained by student in Higher secondary class.

SSCMaths – The Percentage of marks in SSC Mathematics.

SSCSci - The Parentage of marks obtained in SSC Science.

Atype - The admission type which may be through central process or through Management of institute.

SSCMedium - Medium of the student at secondary school level.

FEresult - Result of student in First Year of Engineering. This can take the values PASS, FAIL, or ATKT**.** In general, if a student fails in up to three theory and two practical subjects of an academic year or vice versa, he/she is awarded ATKT and promoted to next class, provided they do not have backlog of previous year.

5.2 Model Construction

From this data, student.arff file was created. This file was loaded into WEKA explorer. The WEKA workbench contains a collection of

visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to this functionality [17]. It is portable because it is fully implemented in the Java programming language and thus runs on almost any modern computing platform. It has various panels to perform data mining task. The *classify* panel enables the user to apply classification and regression algorithms to the resulting dataset, to estimate the accuracy of the resulting predictive model, and to visualize erroneous predictions, or the model itself. There are 16 decision tree algorithms like ID3, J48, SimpleCART etc. implemented in WEKA. The algorithm used for classification is J48 (java implementation of C4.5 algorithm). J48 does not require discretization of numeric attributes, in contrast to the ID3 algorithm from which C4.5 has evolved. So we don't need to discretize all the attributes. Under the "Test options", the 10-fold cross-validation is selected as our evaluation approach. Since there is no separate evaluation data set, this is necessary to get a reasonable idea of accuracy of the generated model. The model is generated in the form of decision tree.

5.3 Results Obtained

The decision tree generated from student.arff is shown in Figure 1.The accuracy of the model is 60.46 %. That is out of 346 instances 209 instances are correctly classified. The most important attribute in predicting student's performance is found to be HSCCET. The social attributes like category, parents' occupation, living location and other attributes like gender, medium at secondary level are not appearing in the decision tree indicating less relevance of the prediction with such attributes. However (logically) past performance encompasses all these attributes. For example, a student is more likely to get good schooling if his/her parents are well educated and/or having good income and he/she is likely to perform well. (Of course, there are some exceptions as well.)

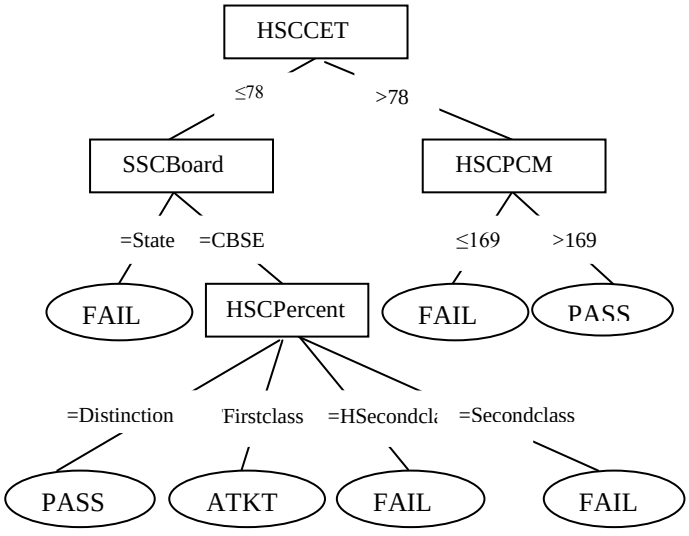

Figure 1: Decision tree for three class prediction.

The rules generated from this tree are

1. If HSCCET <=78 and SSCBoard = State then FEresult = FAIL
2. If HSCCET <=78 and SSCBoard =CBSE and HSCPercent = Distinction then FEresult = PASS
3. If HSCCET <=78 and SSCBoard =CBSE and HSCPercent = Firstclass then FEresult = ATKT
4. If HSCCET <=78 and SSCBoard =CBSE and HSCPercent = Secondclass then FEresult = FAIL
5. If HSCCET <=78 and SSCBoard =CBSE and HSCPercent = Distinction then FEresult = FAIL
6. If HSCCET > 78 and HSCPCM <=169 then FEresult = FAIL
7. If HSCCET > 78 and HSCPCM > 169 then FEresult = PASS

It is clear from the confusion matrix in Table 2 that out of 205 fail students, 186 students are classified as FAIL. So the true positive rate of FAIL is found to be 0.907, but the false positive rate is o.617. Out of 63 pass students 16 are classified as pass, 32 as FAIL and 15 as ATKT. Out of 78 ATKT student, are classified as PASS, 55 as FAIL and 7 as ATKT.

Table 2: Confusion matrix for three class prediction

		Predicted		
		PASS	FAIL	ATKT
Actual	PASS	16	32	15
	FAIL	13	186	6
	ATKT	16	55	7

The class wise accuracy is shown in Table 3.

Table 3: Class wise accuracy for three class prediction

CLASS LABEL	TP Rate	FP Rate
PASS	0.254	0.102
FAIL	0.907	0.617
ATKT	0.09	0.078

Another interesting prediction is about which students will be promoted to next class leading to two class prediction that is Promoted and FAIL. For that, we have considered all ATKT students as Promoted students since they are allowed to attend classes and appear for exams of next year. The decision tree generated is shown in Figure 2.

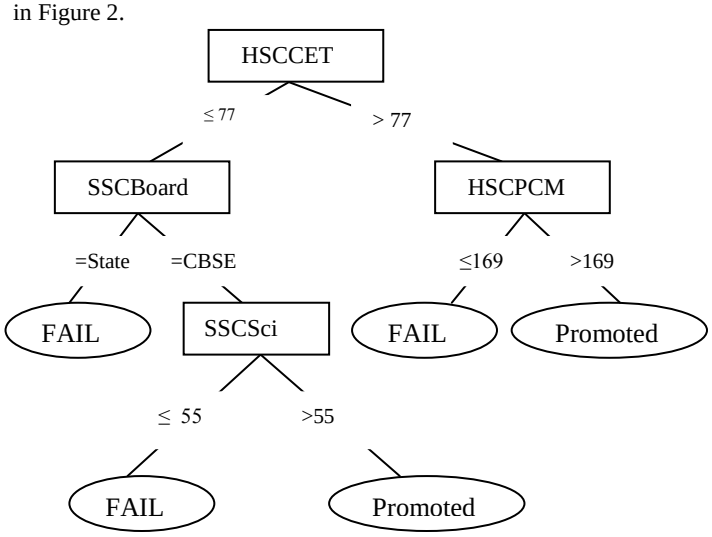

Figure 2: Decision tree for two-class prediction

The accuracy of this model is 69.94 %, that is out of 346 instances 242 are correctly classified. The confusion matrix in Table 4 shows

that out of 205 failed students 170 are correctly classified as FAIL but 35 are classified as Promoted. And Out of 141 promoted students 72 are correctly classified as Promoted but 69 are classified as FAIL.

Table 4: Confusion matrix for two class prediction

		Predicted	
		Promoted	FAIL
Actual	Promoted	72	69
	FAIL	35	170

From Figure 1 and Figure 2 it can be observed that the attributes HSCET, HSCBoard, HSCPCM play a major role in predicting the students performance in engineering exam. Root of the tree in both the trees is HSCCET marks. HSCCET is a continuous attribute, C4.5 converts it in two range, less than or equal to 78 and the second one greater than 78 (In two class prediction the value is 77). The leftmost partition created for HSCCET ≤ 78, is split according to SSCBoard. The State board students getting less than marks in ≤ 78 HSCCET are likely to fail. But CBSE Board students' HSCPercent (SSCSci in case of two class prediction) will be deciding the success. The rightmost partition for HSCCET >78, is further divided based on HSCPCM. If student's PCM total is less than or equal to 169 then student is likely to fail otherwise pass.

6. CONCLUSION

This study shows that students past academic performance can be used to create the model using decision tree algorithm that can be used for prediction of student's performance in First Year of engineering exam. From the confusion matrix it is clear that the true positive rate of the model for the FAIL class is 0.907, that means model is successfully identifying the students who are likely to fail. These students can be considered for proper counseling so as to improve their result. The accuracy of the model is likely to improve if we add the attributes that reveal the current performance (e.g. attendance, test marks etc) and consider more instances.

7. REFERENCES

[1] C. Romero, S. Ventura, "Educational data mining: A survey from 1995 to 2005", Expert system with applications 33(2007), 135-146.

[2] C. Romero, S. Ventura, "Educational Data Mining: A Review of the State of the Art", IEEE transactions on Systems, Man, and Cybernetics-Part C: applications and Reviews, Vol.40, No. 6, November 2010.

[3] J. Han, M. Kamber, Data Mining Concepts and Techniques, Second edition, Morgan Kaufmann, SanFrancisco, ISBN: 978-81-312.

[4] J. R. Quinlan,"Induction of decision trees", Machine Learning, Volume 1, Morgan Kaufmann, 1986, 81-106.

[5] R. Kohavi, R. Quinlan, "Decision Tree Disovery", In Handbook of Data Mining and Knowledge Discovery, University Press,1999.

[6] K. P. Soman, S. Diwakar, V. Ajay, Insight into Data Mining-Theory and Practice, Prentice Hall of India, New Delhi, ISBN: 81-203- 2897-3.

[7] V. P. Bresfelean, "Analysis and Predictions on Students' Behavior 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. Teixeira (Eds.), 2008, pp.5-12.

[9] Z. J. Kovacic, "Early prediction of student success: Mining student enrollment data", Proceedings of Informing Science & IT Education Conference (InSITE) 2010.

[10] M. Ramaswami and R. Bhaskaran, "A CHAID based performance prediction model in educational data mining", IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 1, No. 1, January 2010.

[11] N. Thai-Nghe, L. Drumond, A. Krohn- Grimberghe, L. Schmidt-Thieme, "Recommender System for Predicting Student Performance", Elsvier B.V., 2010.

[12] S. Elayidom, Dr. S. M. Idikkula, J. Alexander, A. Ojha, "Applying Data mining techniques for Placement chance prediction", International Conference on Advances in Computing, Control, and Telecommunication Technologies, 2009.

[13] N. Thai Nghe, P. Janecek, and P. Haddawy, "A Comparative Analysis of Techniques for Predicting Academic Performance", 37th ASEE/IEEE Frontiers in Education Conference, October 2007.

[14] 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.

[15] A. Merceron and K. Yacef "Educational data mining: A case study", In Proceedings AIED, 2005, pp.467-474.

[16] B. K. Baradwaj, S. Pal, "Mining Educational Data to Analyze Students' Performance", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6, 2011.

[17] I. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, 2nd Edition, Morgan Kaufmann, San Francisco, ISBN: 0-12-088407-0.