A Decision Tree Algorithm Pertaining to the Student Performance Analysis and Prediction

Mrinal Pandey
Manav Rachna College Of Engineering,
Faridabad, India,

Vivek Kumar Sharma, PhD.
Jagganath University, Jaipur, India,

ABSTRACT

Growth of an educational institute can be measured in terms of successful students of the institute. The analysis related to the prediction of students academic performance in higher education seems an essential requirement for the improvement in quality education. Data mining techniques play an important role in data analysis. For the construction of a classification model which could predict performance of students, particularly for engineering branches, a decision tree algorithm associated with the data mining techniques have been used in the research. A number of factors may affect the performance of students. Here some significant factors have been considered while constructing the decision tree for classifying students according to their attributes (grades). In this paper four different decision tree algorithms J48, NBtree, Reptree and Simple cart were compared and J48 decision tree algorithm is found to be the best suitable algorithm for model construction. Cross validation method and percentage split method were used to evaluate the efficiency of the different algorithms. The traditional KDD process has been used as a methodology. The WEKA (Waikato Environment for Knowledge Analysis) tool was used for analysis and prediction. . Results obtained in the present study may be helpful for identifying the weak students so that management could take appropriate actions, and success rate of students could be increased sufficiently.

General Terms

Data mining

Keywords: Data mining, Decision tree, Classification, Prediction, Classifiers, Cross validation**.**

1. INTRODUCTION

Data mining helps to extract the relevant information from the large and complex databases [1]. Data mining techniques are useful for data analysis and predictions. Classification is an unsupervised learning technique that helps to classify predefined class labels. There are various classification techniques such as Decision tree algorithm, Bayesian network, Nureal network and Genetic algorithm etc.These technique can be use to build the classification model. This classification model helps to predict the future trend based on previous pattern. This paper propose a classification model particularly decision tree algorithm to predict the future grades of the students in their final examinations. WEKA tool kit is used for model construction and evaluation. This is a four class prediction particularly for engineering students.

2. RELATED WORK

Qasem, Emad and Mustafa [2] made an attempt to use the data mining processes, particularly classification. They worked on enhancing the quality of higher educational system by evaluating the student's data which helped in studying of main attributes which may affect the student's performance in C++ courses. The CRISP framework was employed as a methodology. Three classifiers namely ID3, C4.5 decision tree and Naive Bayes were compared and the result showed that the performance of Decision tree C4.5 was better than other classifiers.

Nguyen and Peter [3] conducted a study of two different group of students of undergraduate and postgraduate level to predict the performance of the students and compared the efficiency of two classifiers namely decision tree and Bayesian networks using WEKA tool. In this research the performance of Decision tree was 3-12% more accurate than Bayesian networks. This was useful for identifying the weak students for further guidance and to selecting the good students for scholarship.

A study has been done by Sunita & LOBO L.M.R.J [4] to illustrate that how data mining can be applicable to the educational system. They perform classification using ZeroR algorithm for performance prediction and clustering of student into group using DBSCAN-clustering algorithm.

R. R. Kabra and Bichkar [5] conducted a research for 346 engineering students studying in first year, and developed a classification model based on their past performance. A two class prediction and three class prediction have been compared under the study. The results of two class predictions were better than three class prediction, which helped to identify the students that are likely to be failed.

S. Anupama and Vijayalakshmi [6] applied C4.5 decision tree algorithm to the internal marks of the MCA students and predict their performance in terms of pass or fail in final exam. They compare the predicted results and actual results which indicates, that there was a significant improvement in results as the prediction helped a lot to identify weak and good students and help them to score better marks. They also compared the model with ID 3 decision tree algorithm and prove that the developed model is better in terms of efficiency and time taken to build the decision tree.

Bharadwaj & Pal [7] proposed ID3 decision tree algorithm as a classification model to predict the students division, the previous information such as attendance, class test, seminar and assignment marks were collected from the student's previous databases to predict the performance at the end of semester. All this helped the students and the teachers to improve the division of the students.

Sajadin , Dedy and Elvi [8] investigate a strong correlation between the mantel condition and the final performance of the students .They develop a rule model based on decision tree and implement these rules through SSVM algorithm to predict the final grades of students. They also grouped the students on the basis of their similar characteristics using K-means clustering.

Qasem, Ahmad.and Emad [9] Proposed a classification model using decision tree algorithm to select the suitable academic track for the students ,This model is useful for school management to choose the appropriate academic track for a student based on the previous students data and the similar academic achievements of the students. The efficiency of the model is 87. 9 %.

Surjeet & Pal [10] compared C4.5, ID3 and CART decision tree algorithms to predict the performance of the first year engineering students .It was three class predictions. Students were classified as pass fail and promoted. This model was good to identifying the students that are most likely to fail.

Dorina Kabakchieva [11] attempt to predict student performance by applying and Comparing four data mining algorithms, OneR Rule Learner, Decision Tree, Neural Network and K-Nearest Neighbour, on data set of 10330 Bulgarian students .It was two class prediction problem. The students were classified as strong and weak. The NN model achieved high accuracy for strong class predictions where as other models were good for weak class predictions.

Shovon & Mahfuza [12] proposed a hybrid approach of clustering and classification to improve the student academic performance in their final examinations. Initially students were classified into three categories high, medium and low standards and then applied decision tree algorithm to take appropriate decisions for the students.

3. RESEARCH METHODLOGY

In this section the steps of traditional KDD process would be followed. The process starts from data collection and data preprocessing followed by classification model construction and ends with model evaluation and interpretations.

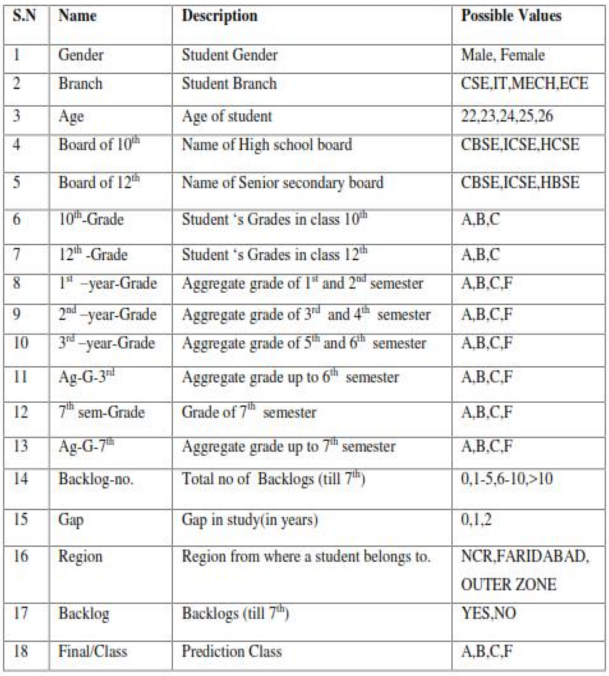
3.1 Data collection and Preprocessing

The Data set for the study has been collected from Manav Rachna College of engineering district Faridabad of Haryana state. This data set consists of 524 instances and each instance consists of 18 different attributes. The study consider the academic performance of the student from high school to the prefinal semester of Engineering (grades up to 7th semester of BTech Program) and predict the final results for the completing the graduate degree in engineering. Initially data is collected in an excel sheet and initial preprocessing is done manually by filling the missing data values by standard data and various inconsistence has been removed. Some irrelevant attributes have been removed manually to maintain the quality of the classifier. The gain ratio measure is selected for ranking the attributes and finally 8 relevant attribute have been selected on the basis of their ranks for the study. Table 1 shows the attributes description and their possible values.

3.2 Model Building

The next step is to build a classification model. In this step, decision tree has been selected as a classifier under the cross validation method.

Table1: Attribute Description



S.N	Name	Description	Possible Values
1	Gender	Student Gender	Male, Female
2	Branch	Student Branch	CSE,IT,MECH,ECE
3	Age	Age of student	22,23,24,25,26
4	Board of 10th	Name of High school board	CBSE,ICSE,HCSE
5	Board of 12th	Name of Senior secondary board	CBSE,ICSE,HBSE
6	10th-Grade	Student 's Grades in class 10th	A,B,C
7	12th -Grade	Student 's Grades in class 12th	A,B,C
8	1st –year-Grade	Aggregate grade of 1st and 2nd semester	A,B,C,F
9	2nd –year-Grade	Aggregate grade of 3rd and 4th semester	A,B,C,F
10	3rd –year-Grade	Aggregate grade of 5th and 6th semester	A,B,C,F
11	Ag-G-3rd	Aggregate grade up to 6th semester	A,B,C,F
12	7th sem-Grade	Grade of 7th semester	A,B,C,F
13	Ag-G-7th	Aggregate grade up to 7th semester	A,B,C,F
14	Backlog-no.	Total no of Backlogs (till 7th)	0,1-5,6-10,>10
15	Gap	Gap in study(in years)	0,1,2
16	Region	Region from where a student belongs to.	NCR,FARIDABAD, OUTER ZONE
17	Backlog	Backlogs (till 7th)	YES,NO
18	Final/Class	Prediction Class	A,B,C,F

Note: Grade Values: A=81-100, B=61-80, C=41-60, F<40

For model construction C4.5 decision tree method has been used, which is based on gain ratio as attribute selection measure. The attribute having maximum gain ratio value is selected for splitting the node. In this study attribute "ag-g-7th "has the highest gain ratio value, therefore it is selected as root node of the decision tree. The attribute "ag-g-3rd "has the next higher value and hence this node has been chosen for further splitting. This process continues till the complete tree is constructed. WEKA tool kit was used to select the attributes and construct the decision tree (J48). Fig 1 shows the decision tree construction. Each leaf node is represented by rectangle and root node/splitting node is represented by an oval.

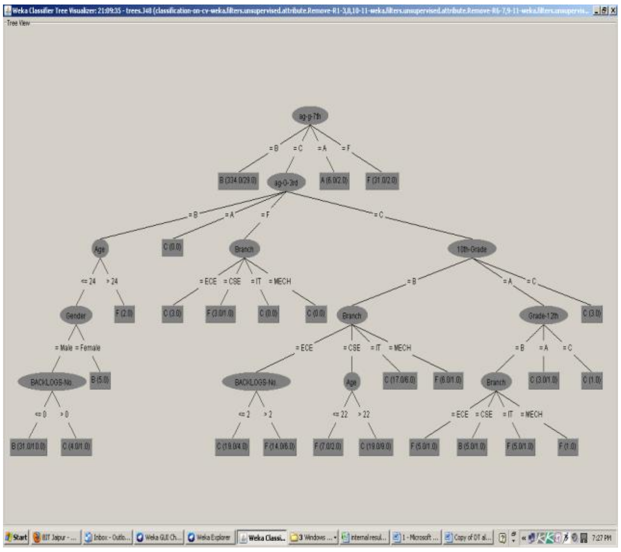


Fig 1: Decision Tree Construction

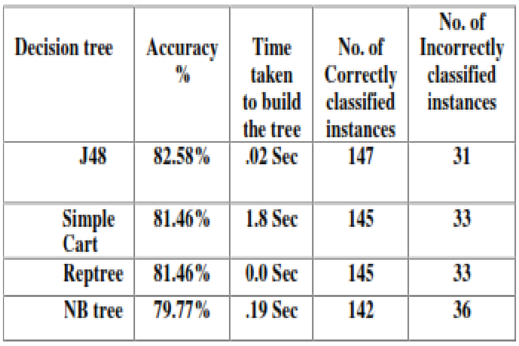
3.1.1 Classification Rules

Once a decision tree is generated, the classification rules can be extracted by tracing the path from root node to each leaf node in the tree. Each splitting node is logically ANDed to form rule antecedent and each leaf node represents the class value for the prediction [1].A set of classification rules can be extracted from the decision tree shown in Fig 1. These rules help to classify students and predict the final grades of the students in BTech Examinations. The class label having four class values A, B, C, and F respectively. Table 2 shows the extracted rules, predicted class and no of correctly and incorrectly classified instances reached to a particular leaf node of the decision tree.

3.3 Model Evaluation and Interpretation

To evaluate the classification model 10 fold cross validation and percentage split methods have been used. In 10 fold cross validation all the data has been divided into 10 disjoint set of approximately equal size. This is an iterative process. Each time 9 disjoints sets acts as a training data and one set is used as a testing data. In percentage split method 66% of the entire data has been used as training data and remaining data as testing data. Four different decision tree algorithms J48, Simple Cart, Reptree and NBtree have been compared .The results of compressions are depicted in Table 3.1 and Table 3.2 for percentage split method and cross validation method respectively. The highest accuracy has been achieved by J48 decision tree algorithm. The over all accuracy of this model is 80.15 % using 10 fold cross validation. It shows that the grades of 420 are correctly classified from the 524 students. Percentage split method determines the accuracy of the model is 82.58%. This model correctly classified the grades of 147 students among 178 students. The confusion matrix shows the accuracy of the predicated classes. Fig 2.1 shows the confusion matrix for cross validation and Fig 2.2 represents the confusion matrix for percentage split method

Table3.1:Comparisons of algorithms using percentage split method



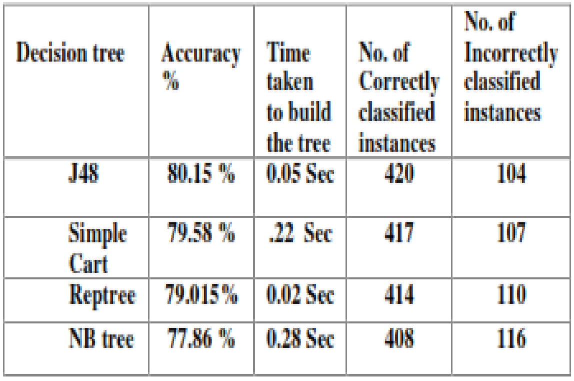
Decision tree	Accuracy %	Time taken to build the tree	No. of Correctly classified instances	No. of Incorrectly classified instances
J48	82.58%	.02 Sec	147	31
Simple Cart	81.46%	1.8 Sec	145	33
Reptree	81.46%	0.0 Sec	145	33
NB tree	79.77%	.19 Sec	142	36

4. CONCLUSION AND FUTURE WORK

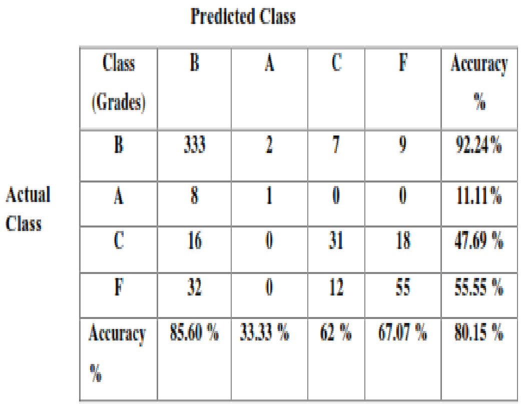
A classification model has been proposed in this study for predicting student's grades particularly for engineering under graduate students. Four decision tree algorithms were compared and J48 decision tree algorithm was selected for model construction, where J48 is a java version of C 4.5.The model obtained accuracy of 80.15% and 82.58% in 10 fold cross validation method and percentage method respectively. It indicates that model is good for forecasting the grades of students. This model helps to the management to identify weak students and can take appropriate decision to prevent them from failure.

This research can be enhanced by comparing various other classifiers and choosing the best of them to obtain the better results. For this purpose data set need to increase in terms of attributes as well as instances. A robust decision support system can be developed based on classifiers as a future work.

Table 3.2 Comparisions of algorithm using cross validation



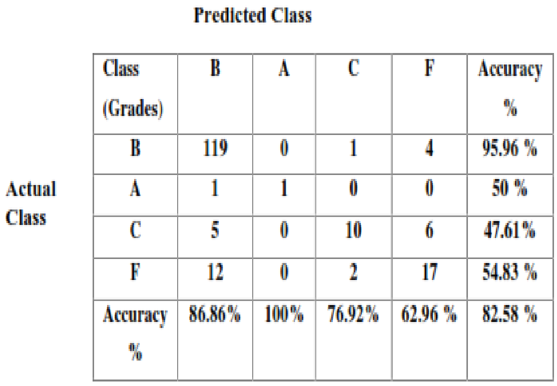
Decision tree	Accuracy %	Time taken to build the tree	No. of Correctly classified instances	No. of Incorrectly classified instances
J48	80.15 %	0.05 Sec	420	104
Simple Cart	79.58 %	.22 Sec	417	107
Reptree	79.015%	0.02 Sec	414	110
NB tree	77.86 %	0.28 Sec	408	116



Predicted Class

	Class (Grades)	B	A	C	F	Accuracy %
	B	333	2	7	9	92.24%
Actual Class	A	8	1	0	0	11.11%
	C	16	0	31	18	47.69 %
	F	32	0	12	55	55.55 %
	Accuracy %	85.60 %	33.33 %	62 %	67.07 %	80.15 %

Fig 2.1 Confusion matrix for J48 algorithm using cross validation



Predicted Class

	Class (Grades)	B	A	C	F	Accuracy %
	B	119	0	1	4	95.96 %
Actual Class	A	1	1	0	0	50 %
	C	5	0	10	6	47.61%
	F	12	0	2	17	54.83 %
	Accuracy %	86.86%	100%	76.92%	62.96 %	82.58 %

Fig 2.2 : Confusion Matrix for J48 algorithm using percentage split method

Table 2: Set of Classification Rules

Rule #	**Rules**	**Predicted class**	**Instances #**
1.	If ag-g-7th =F	Class=F	31/2
2.	If ag-g-7th =A	Class=A	6/2
3.	If ag-g-7th =B	Class=B	334/29
4.	If ag-g-7th =C, ag-g-3rd=B, age<=24,Gender=male ,Backlog No<=0	Class=B	31/10
5.	If ag-g-7th =C, ag-g-3rd=B, age<=24,Gender=male, Backlog No>0	Class=C	4/1
6.	If ag-g-7th =C, ag-g-3rd=B, age<=24,Gender=Female	Class=B	5
7.	If ag-g-7th =C, ag-g-3rd=B, age>24, Backlog No>0	Class=F	2
8.	If ag-g-7th =C, ag-g-3rd=A	Class=A	0
9.	If ag-g-7th =C, ag-g-3rd=F, Branch=ECE	Class=C	3
10.	If ag-g-7th =C, ag-g-3rd=F, Branch=CSE	Class=F	3/1
11.	If ag-g-7th =C, ag-g-3rd=F, Branch=IT	Class=C	0
12.	If ag-g-7th =C, ag-g-3rd=F, Branch=MECH	Class=C	0
13.	If ag-g-7th =C, ag-g-3rd=C, 10th grade=B, Branch=ECE ,Backlog No<=2	Class=C	19/4
14.	If ag-g-7th =C, ag-g-3rd=C, 10th grade=B, Branch=ECE ,Backlog No>2	Class=F	14/6
15.	If ag-g-7th =C, ag-g-3rd=C, 10th grade=B, Branch=CSE ,Age <=22	Class=F	7/2
16.	If ag-g-7th =C, ag-g-3rd=C, 10th grade=B, Branch=CSE ,Age >22	Class=C	19/9
17.	If ag-g-7th =C, ag-g-3rd=C, 10th grade=B, Branch=IT	Class=C	17/6
18.	If ag-g-7th =C, ag-g-3rd=C, 10th grade=B, Branch=MECH	Class=C	6/1
19.	If ag-g-7th =C, ag-g-3rd=C, 10th grade=C	Class=C	3
20.	If ag-g-7th =C, ag-g-3rd=C, 10th grade=A,12th grade=A	Class=C	3/1
21.	If ag-g-7th =C, ag-g-3rd=C, 10th grade=A,12th grade=C	Class=C	1
22.	If ag-g-7th =C, ag-g-3rd=C, 10th grade=A,12th grade=B, Branch=ECE	Class=F	5/1
23.	If ag-g-7th =C, ag-g-3rd=C, 10th grade=A,12th grade=B, Branch=CSE	Class=B	5/1
24.	If ag-g-7th =C, ag-g-3rd=C, 10th grade=A,12th grade=B, Branch=IT	Class=B	5/1
25.	If ag-g-7th =C, ag-g-3rd=C, 10th grade=A,12th grade=B, Branch=MECH	Class=F	1

5. REFERENCES

[1] Han J., Kamber M., and Pie J. (2006). Data Mining Concepts and Techniques. 2nd edition, Morgan Kaufmann Publishers.

[2] Al-Radaideh Q., Al-Shawakfa Emad M., and Al-Najjar Mustafa I. (2006). Mining Student Data Using Decision Trees. The 2006 International Arab Conference on Information Technology.

[3] Nguyen N., Paul J., and Peter H. (2007). A Comparative Analysis of Techniques for Predicting Academic Performance. In Proceedings of the 37th ASEE/IEEE Frontiers in Education Conference. pp. 7-12.

[4] Sunita B Aher, Mr. LOBO L.M.R.J. (2011). Data Mining in Educational System using WEKA, International Conference on Emerging Technology Trends (ICETT). Proceedings Published by International Journal of Computer Applications.

[5] Kabra R. R., Bichkar R.S. (2011). Performance Prediction of Engineering Students using Decision Trees, International Journal of Computer Applications (0975-8887), Vol-36-No.11.

[6] Kumar S. Anupama and Dr. Vijayalakshmi M.N. (2011). Efficiency of Decision Trees in Predicting Students Academic Performance. Computer Science & Information Technology 02, pp. 335–343.

[7] Baradwaj Brijesh Kumar and Pal Saurabh (2011). Mining Educational Data to Analyze Student Performance. International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6.

[8] Sembiring Sajadin,. Zarlis M , Hartama Dedy ,Ramliana S,Wani Elvi (2011). Prediction of Student Academic Performance by an application of data mining techniques.International Conference on Management and Artificial Intelligence IPEDR vol.6 (2011) © (2011) IACSIT Press, Bali, Indonesia

[9] Al-Radaideh Q., Al-Ananbeh Ahmad.,Al-Shawakfa Emad M . (2011). A Classification Model for Predicting the Suitable Study Track for School Students. www.arpapress.com/Volumes/Vol8Issue2/IJRRAS_8_2_15.pdf

[10] Yadav Surjeet Kumar, Pal Saurabh Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification (2012). World of Computer Science and Information Technology Journal (WCSIT) .ISSN: 2221-0741 Vol. 2, No. 2, 51-56, 2012.

[11] Kabakchieva Dorina (2012). Student Performance Prediction by Using Data Mining Classification Algorithms. International Journal of Computer Science and Management Research .ISSN 2278-733X: Vol 1 Issue 4 November

[12] Shovon Md. Hedayetul Islam, Haque Mahfuza (2012).An Approach of Improving Student's Academic Performance by using K-means clustering algorithm and Decision tree . (IJACSA) International Journal of Advanced Computer Science and Applications, Vol.3, No. 8, 2012