Data Mining: Prediction for Performance Improvement of Graduate Students using Classification

Kamal Bunkar
Institute of Comp. Science,
Vikram University,
Ujjain, India

Rajesh Bunkar
Institute of Comp. Science,
Dr.H.S.Gaur University,
Sagar, India

Umesh Kumar Singh
Institute of Comp. Science,
Vikram University,
Ujjain, India

Bhupendra Pandya
Institute of Comp. Science,
Vikram University,
Ujjain, India

Abstract— **Student performance in university courses is of great concern to the higher education where several factors may affect the performance. This paper is an attempt to apply the data mining processes, particularly classification, to help in enhancing the quality of the higher educational system by evaluating student data to study the main attributes that may affect the student performance in courses. For this purpose, we have used data obtained from Vikram University, Ujjain of course B.A. first year student. The classification rule generation process is based on the decision tree as a classification method where the generated rules are studied and evaluated. A system that facilitates the use of the generated rules is built which allows students to predict the final grade in a course under study.**

Key Words— **Data Mining, Classification, Decision Trees, Student Data, Higher Education**

I. INTRODUCTION

Data mining concepts and methods can be applied in various fields like marketing, medicine, real estate, customer relationship management, engineering, web mining etc. Educational data mining is a new emerging technique of data mining that can be applied on the data related to the field of education. There are increasing research interests in using data mining in education. This new emerging field, called Educational Data Mining. Educational Data Mining uses many techniques such as Decision Trees, Neural Networks, Naïve Bayes, K- Nearest neighbor, and many others.

Using these techniques many kinds of knowledge can be discovered such as association rules, classifications and clustering. The discovered knowledge can be used for prediction regarding enrollment of students in a particular course, prediction about student's performance and so on.
In a University results overall performance of a student is determined by internal assessment as well as external exam. Internal assessment is made on the bases of a student's assignment marks, class quiz, lab work, attendance previous year grade and his/her involvement in extra curriculum activities. While at the same time external assessment of a student based on marks scored in final exam. In this paper we make prediction about fail and pass ratio of students based on final exam.

Examination plays a vital role in any student's life. The marks obtained by the student in the examination decide his future. Therefore it becomes essential to predict whether the student will pass or fail in the examination. If the prediction says that a student tends to fail in the examination prior to the examination then extra efforts can be taken to improve his studies and help him to pass the examination.

II. DECISION TREE

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.

Decision tree are commonly used for gaining information for the purpose of decision -making. Decision tree starts with a root node on which it is for users to take actions. From this node, users split each node recursively according to decision tree learning algorithm. The final result is a decision tree in which each branch represents a possible scenario of decision and its outcome.
The three widely used decision tree learning algorithms are: ID3, C4.5 and CART.

A. *ID3 (Iterative Dichotomiser 3)*

This is a decision tree algorithm introduced in 1986 by Quinlan Ross [1]. It is based on Hunts algorithm. The tree is constructed in two phases. The two phases are tree building and pruning.
ID3 uses information gain measure to choose the splitting attribute. It only accepts categorical attributes in building a tree model. It does not give accurate result when there is noise. To remove the noise pre-processing technique has to be used.
To build decision tree, information gain is calculated for each and every attribute and select the attribute with the highest information gain to designate as a root node. Label the attribute as a root node and the possible values of the attribute are represented as arcs. Then all possible outcome instances are tested to check whether they are falling under the same class or not. If all the instances are falling under the same class, the node is represented with single class name, otherwise choose the splitting attribute to classify the instances.

Continuous attributes can be handled using the ID3 algorithm by discretizing or directly, by considering the values to find the best split point by taking a threshold on the attribute values. ID3 does not support pruning.

B. *C4.5*

This algorithm is a successor to ID3 developed by Quinlan Ross [2]. It is also based on Hunt's algorithm.C4.5 handles both categorical and continuous attributes to build a decision tree. In order to handle continuous attributes, C4.5 splits the attribute values into two partitions based on the selected threshold such

978-1-4673-1989-8/12/$31.00 ©2012 IEEE

that all the values above the threshold as one child and the remaining as another child. It also handles missing attribute values. C4.5 uses Gain Ratio as an attribute selection measure to build a decision tree. It removes the biasness of information gain when there are many outcome values of an attribute.

At first, calculate the gain ratio of each attribute. The root node will be the attribute whose gain ratio is maximum. C4.5 uses pessimistic pruning to remove unnecessary branches in the decision tree to improve the accuracy of classification.

C. *CART*

CART [1] stands for Classification and Regression Trees introduced by Breiman. It is also based on Hunt's algorithm. CART handles both categorical and continuous attributes to build a decision tree. It handles missing values.

CART uses Gini Index as an attribute selection measure to build a decision tree .Unlike ID3 and C4.5 algorithms, CART produces binary splits. Hence, it produces binary trees. Gini Index measure does not use probabilistic assumptions like ID3, C4.5. CART uses cost complexity pruning to remove the unreliable branches from the decision tree to improve the accuracy.

III. BACKGROUND AND RELATED WORK

Nguyen *et al*. [4] 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. These predictions are most useful for identifying and assisting failing students, and better determine scholarships. As a result, the decision tree classifier provided better accuracy in comparison with the Bayesian network classifier.

Al-Radaideh *et al.* [5] proposed to use data mining classification techniques to enhance the quality of the higher educational system by evaluating students' data that may affect the students' performance in courses. They used the CRISP framework for data mining to mine students' related academic data. A classification model was built using the decision tree method. They used three different classification methods ID3, C4.5 and the NaïveBayes. The results indicated that the decision tree model had better prediction accuracy than the other models. As a result, a system was built to facilitate the usage of the generated rules that students need to predict the final grade in the C++ undergraduate course.

Cesar *et al.* [6] proposed the use of a recommendation system based on data mining techniques to help students to make decisions related to their academic track. The system provided support for students to better choose how many and which courses to enroll on. As a result, the authors developed a system that is capable to predict the failure or success of a student in any course using a classifier obtained from the analysis of a set of historical data related to the academic field of other students who took the same course in the past.

Muslihan *et al.* [7] have compared two data mining techniques which are: Artificial Neural Network and the combination of clustering and decision tree classification techniques for predicting and classifying student's academic performance. Students' data were collected from the data of the National Defence University of Malaysia (NDUM). As a result, the technique that gives accurate prediction and classification was chosen as the best model. Using the proposed model, the pattern that influences the student's academic performance was identified.

Han and Kamber [8] describes data mining software that allow the users to analyze data from different dimensions, categorize it and summarize the relationships which are identified during the mining process.

Bharadwaj and Pal [9] conducted study on the student performance based by selecting 300 students from 5 different degree college conducting BCA (Bachelor of Computer Application) course of Dr. R. M. L. Awadh University, Faizabad, India. By means of Bayesian classification method on 17 attributes, it was found that the factors like students' grade in senior secondary exam, living location, medium of teaching, mother's qualification, students other habit, family annual income and student's family status were highly correlated with the student academic performance.

Pandey and Pal [10] conducted study on the student performance based by selecting 600 students from different colleges of Dr. R. M. L. Awadh University, Faizabad, India. By means of Bayes Classification on category, language and background qualification, it was found that whether new comer students will performer or not.

Ramaswami and Bhaskaran [11] have constructed a predictive model called CHAID with 7-class response variable by using highly influencing predictive variables obtained through feature selection so as to evaluate the academic achievement of students at higher secondary schools in India. Data were collected from different schools of Tamilnada, 772 students' records were used for CHAID prediction model construction. As a result, set of rules were extracted from the CHAID prediction model and the efficiency was found. The accuracy of the present model was compared with other models and it has been found to be satisfactory.

Shannaq et al. [12], applied the classification as data mining technique to predict the numbers of enrolled students by evaluating academic data from enrolled students to study the main attributes that may affect the students' loyalty (number of enrolled students). The extracted classification rules are based on the decision tree as a classification method, the extracted classification rules are studied and evaluated using different evaluation methods. It allows the University management to prepare necessary resources for the new enrolled students and indicates at an early stage which type of students will potentially be enrolled and what areas to concentrate upon in higher education systems for support.

IV. BUILDING THE CLASSIFICATION MODEL

This section describes the building of the classification model. In general, data classification is a two-step process. In the first step, which is called the learning step, a model that describes a predetermined set of classes or concepts is built by analyzing a set of training database instances. Each instance is assumed to belong to a predefined class. In the second step, the model is tested using a different data set that is used to estimate the classification accuracy of the model. If the accuracy of the model is considered acceptable, the model can be used to classify future data instances for which the class label is not known. At the end, the model acts as a classifier in the decision making process. There are several techniques that can be used for classification such as decision tree, Bayesian methods, rule based algorithms, and Neural Networks.

Decision tree classifiers are quite popular techniques because the construction of tree does not require any domain expert knowledge or parameter setting, and is appropriate for exploratory knowledge discovery. Decision tree can produce a model with rules that are human-readable and interpretable. Decision Tree has the advantages of easy interpretation and understanding for decision makers to compare with their domain knowledge for validation and justify their decision. Some of decision tree classifiers are C4.5/ID3/CART, NBTree, and others.

V. DATA COLLECTION

In this step only those fields were selected which were required for data mining. The data are collected from the regular students who had studied in BA first year. The data set used in this study was obtained from Vikram University, Ujjain of course B.A 2009. Our objective is to use the Examination data of the student. The data is stored in a database: MS Excel, but it can also be used with Oracle, Access, Interbase, any database supporting ODBC connections and others. We have used MS Excel because is the world's most popular database. Since the data mining software used to generate association rules accepts data only in arff format, we have first converted the data on Ms Excel file into comma separated text format and then to arff format.

Predictor and response variables which are in the database are given in Table 1 for reference.

Table 1 Examination data of BA first year student

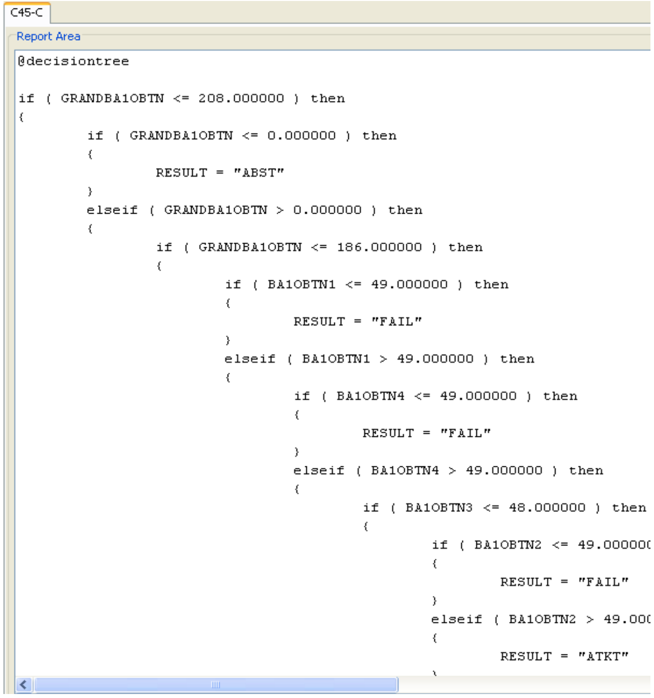
```
@relation BA1
@attribute BA1OBTN1 numeric
@attribute BA1OBTN2 numeric
@attribute BA1OBTN3 numeric
@attribute BA1OBTN4 numeric
@attribute GRANDBA1OBTN numeric
@attributeRESULT{FAIL,PASS,ATKT,ABST,
W.H.}
@data
42,44,28,40,154,FAIL
58,52,69,70,249,PASS
56,40,65,45,206,FAIL
50,36,41,40,167,FAIL
43,38,50,52,183,FAIL
55,58,56,67,236,PASS
36,35,42,38,151,FAIL
36,40,29,43,148,FAIL
34,35,31,45,145,FAIL
53,78,54,62,247,PASS
44,67,62,64,237,ATKT
```

In Table relation BA1 specify the class BA and attribute BAO1OBTN1,BAO1OBTN2, AO1OBTN2, AO1OBTN3, RANDBAO1OBTN respectively show marks obtained in all the subject of BA first year and grand obtain marks. The last attribute in Table 1 (RESULT) is the class label to be predicted.

VI. RESULTS AND DISCUSSION

The three decision trees as examples of predictive models obtained from the student data set by three machine learning algorithms: the C4.5 decision tree algorithm, the CART algorithm, and ID3 decision tree algorithm.

Figure 1, 2 and 3 shows the rules generated by C4.5, CART and ID3respectively.



```
@decisiontree

if ( GRANDBA1OBTN <= 208.000000 ) then
{
        if ( GRANDBA1OBTN <= 0.000000 ) then
        {
                RESULT = "ABST"
        }
        elseif ( GRANDBA1OBTN > 0.000000 ) then
        {
                if ( GRANDBA1OBTN <= 186.000000 ) then
                {
                        if ( BA1OBTN1 <= 49.000000 ) then
                        {
                                RESULT = "FAIL"
                        }
                        elseif ( BA1OBTN1 > 49.000000 ) then
                        {
                                if ( BA1OBTN4 <= 49.000000 ) then
                                {
                                        RESULT = "FAIL"
                                }
                                elseif ( BA1OBTN4 > 49.000000 ) then
                                {
                                        if ( BA1OBTN3 <= 48.000000 ) then
                                        {
                                                if ( BA1OBTN2 <= 49.00000(
                                                {
                                                        RESULT = "FAIL"
                                                }
                                                elseif ( BA1OBTN2 > 49.000
                                                {
                                                        RESULT = "ATKT"
                                                }
```

Figure 1: C 4.5 rules

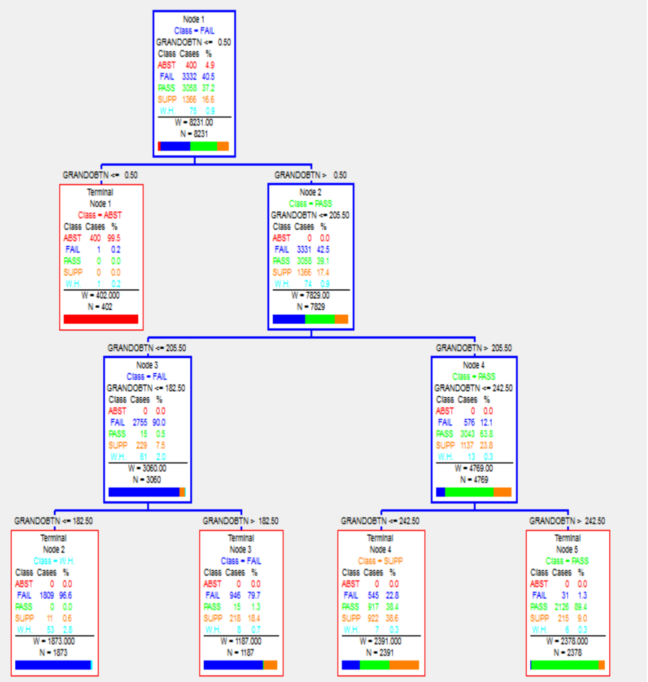
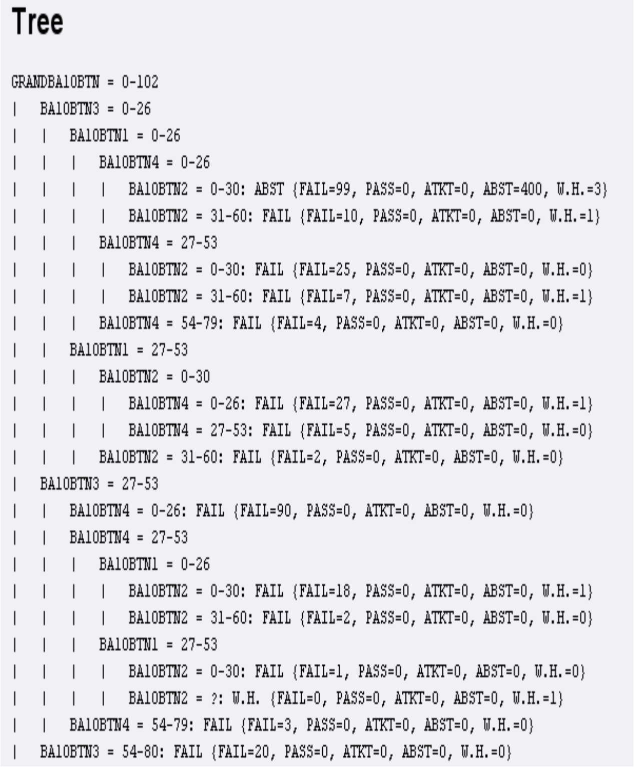


Figure 2: CART rules



Tree

```
GRANDBA1OBTN = 0-102
|  BA1OBTN3 = 0-26
|  |  BA1OBTN1 = 0-26
|  |  |  BA1OBTN4 = 0-26
|  |  |  |  BA1OBTN2 = 0-30: ABST {FAIL=99, PASS=0, ATKT=0, ABST=400, W.H.=3}
|  |  |  |  BA1OBTN2 = 31-60: FAIL {FAIL=10, PASS=0, ATKT=0, ABST=0, W.H.=1}
|  |  |  BA1OBTN4 = 27-53
|  |  |  |  BA1OBTN2 = 0-30: FAIL {FAIL=25, PASS=0, ATKT=0, ABST=0, W.H.=0}
|  |  |  |  BA1OBTN2 = 31-60: FAIL {FAIL=7, PASS=0, ATKT=0, ABST=0, W.H.=1}
|  |  |  BA1OBTN4 = 54-79: FAIL {FAIL=4, PASS=0, ATKT=0, ABST=0, W.H.=0}
|  |  BA1OBTN1 = 27-53
|  |  |  BA1OBTN2 = 0-30
|  |  |  |  BA1OBTN4 = 0-26: FAIL {FAIL=27, PASS=0, ATKT=0, ABST=0, W.H.=1}
|  |  |  |  BA1OBTN4 = 27-53: FAIL {FAIL=5, PASS=0, ATKT=0, ABST=0, W.H.=0}
|  |  |  BA1OBTN2 = 31-60: FAIL {FAIL=2, PASS=0, ATKT=0, ABST=0, W.H.=0}
|  BA1OBTN3 = 27-53
|  |  BA1OBTN4 = 0-26: FAIL {FAIL=90, PASS=0, ATKT=0, ABST=0, W.H.=0}
|  |  BA1OBTN4 = 27-53
|  |  |  BA1OBTN1 = 0-26
|  |  |  |  BA1OBTN2 = 0-30: FAIL {FAIL=18, PASS=0, ATKT=0, ABST=0, W.H.=1}
|  |  |  |  BA1OBTN2 = 31-60: FAIL {FAIL=2, PASS=0, ATKT=0, ABST=0, W.H.=0}
|  |  |  BA1OBTN1 = 27-53
|  |  |  |  BA1OBTN2 = 0-30: FAIL {FAIL=1, PASS=0, ATKT=0, ABST=0, W.H.=0}
|  |  |  |  BA1OBTN2 = ?: W.H. {FAIL=0, PASS=0, ATKT=0, ABST=0, W.H.=1}
|  |  BA1OBTN4 = 54-79: FAIL {FAIL=3, PASS=0, ATKT=0, ABST=0, W.H.=0}
|  BA1OBTN3 = 54-80: FAIL {FAIL=20, PASS=0, ATKT=0, ABST=0, W.H.=0}
```

Figure 3: ID3 rules

Table 2: Confusion Matrix

=== Confusion Matrix ===

a	b	c	d	e	<-- classified as
2852	371	10	99	0 |	a = FAIL
108	2949	1	0	0 |	b = PASS
354	982	30	0	0 |	c = ATKT
0	0	0	400	0 |	d = ABST
60	12	0	3	0 |	e = W.H.

The performance of the classifier is customarily evaluated by a confusion matrix as illustrated in above Table 2. The rows of the table are the actual class label of an instance, and the columns of the table are the predicted class label of an instance. Typically, the class label of a minority class set as positive and that of a majority class set as negative. TP, FN, FP, and TN are True Positive, False Positive, False Negative and True Negative, respectively. From Table 2, the six performance measures on classification; accuracy, precision, recall, F-value, TP rate, and FP rate, are defined by formulae in (1)-(6).

- Accuracy=(TP+TN)/(TP+FN+FP+TN) (1)
- Recall = TP/(TP+FN). (2)
- Precision = TP/(TP+FP). (3)
- F-value= ((1+)^2.Recall.Precision)/(^{2}.Recall+ Precision). (4)
- True Positive Rate = TP/(TP+FN). (5)
- FP Rate = FP/(TN+FP) (6)

=== Detailed Accuracy By Class ===						
TP Rate Class	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.856	0.107	0.845	0.856	0.851	0.921	FAIL
0.964	0.264	0.684	0.964	0.8	0.922	PASS
0.022	0.002	0.732	0.022	0.043	0.749	ATKT
1	0.013	0.797	1	0.887	0.992	ABST
0	0	0	0	0	.679	W.H.

These results are conducted to find the best classifier for prediction of student's performance in First Year of BA examination. From the classifier accuracy it is clear that the true positive rate of the model for the FAIL class is .856 and ATKT class is .022 that means model is successfully identifying the student for proper counseling so as to improve result.

VII. CONCLUSION

One of the data mining techniques i.e., classification is an interesting topic to the researchers as it is accurately and efficiently classifies the data for knowledge discovery. Decision trees are so popular because they produce classification rules that are easy to interpret than other classification methods. Frequently used decision tree classifiers are studied and the experiments are conducted to find the best classifier for prediction of student's performance in First Year of BA exam. Decision trees 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. This finding is a preliminary research in this area and we think it is a good starting point for researchers in the region to establish a research track related to using data mining to enhance college/university education. This research should be further enhanced as a future work by considering data from several other university including private university in other cities in India and collect more instances to build the model. Other attributes could also be added to the data set for further enhancing the generated model. Furthermore, some other classification models could be tested in this domain.

REFERENCES

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

[2] J. R. Quinlan, “C4.5: Programs for Machine Learning”, Morgan Kaufmann Publishers, Inc, 1992.

[3] Alaa el-Halees, “Mining students data to analyze e-Learning behavior: A Case Study”, 2009.

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

[5] Al-Radaideh Q., Al-Shawakfa E., and AI-Najjar M., Mining Student Data using Decision Trees, In Proceedings of the International Arab Conference on Information Technology (ACIT'2006), Yarmouk University, Jordan, 2006

[6] Cesar V., Javier B., liela S., and Alvaro O., Recommendation in Higher Education Using Data Mining Techniques, In Proceedings of the Educational Data Mining Conference, 2009.

[7] Muslihah W., Yuhanim Y., Norshahriah W., Mohd Rizal M., Nor Fatimah A., and Hoo Y. S., Predicting NDUM Student's Academic Performance Using Data Mining Techniques, In Proceedings of the Second International Conference on Computer and Electrical Engineering, IEEE computer society, 2009

[8] J. Han and M. Kamber, “Data Mining: Concepts and Techniques,” Morgan Kaufmann, 2000.

[9] B.K. Bharadwaj and S. Pal. “Data Mining: A prediction for performance improvement using classification”, International Journal of Computer Science and Information Security (IJCSIS), Vol. 9, No. 4, pp. 136-140, 2011.

[10] U . K. Pandey, and S. Pal, “Data Mining: A prediction of performer or underperformer using classification”, (IJCSIT) International Journal of Computer Science and Information Technology, Vol. 2(2), pp.686-690, ISSN:0975-9646, 2011.

[11] Ramaswami M., and Bhaskaran R., CHAID Based Performance Prediction Model in Educational Data Mining, IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 1, No. 1, 2010.

[12] Shannaq, B. , Rafael, Y. and Alexandro, V. (2010) ‘Student Relationship in Higher Education Using Data Mining Techniques’, Global Journal of Computer Science and Technology, vol. 10, no. 11, pp. 54-59.