Predicting Academic Success from Student Enrolment Data using Decision Tree Technique

M Narayana Swamy
Department of Computer Applications, Presidency College
Bangalore ,India

M. Hanumanthappa
Department of Computer Science & Applications, Bangalore University
Bangalore, India

ABSTRACT

The Indian education system has witnessed significant expansion in recent years, both in terms of the number of institutions as well as the student enrollment. There is a massive growth in self financed higher educational institutions in India in the next two decades. This causes a competition among institutions while attracting the student to get admission in these institutions. Therefore, institutions focused on the strength of students not on the quality of student at the time of enrollment. After the enrollment the institution tries to improve the quality of the student.

Like other domain educational domain also produce huge amount of data. To improve the quality of education the data analysis plays an important role for decision support. The data mining is used to extract hidden information from large data set/data warehouse.

In this paper we present the data mining technique to predict the performance of the students based on the enrollment data. It helps the teacher to take remedial measure for slow learners to improve the performance in the university examination.

General Terms

Data mining

Keywords

Educational Data mining, Classification, Decision Tree, Higher Education.

1. INTRODUCTION

India's higher education system is the second largest in the world, after the United States[1] The Indian higher education sector is expected to grow at a 18% Compound Annual Growth Rate(CAGR) till 2020[2]. In next two decade Educational sector will become the main business sector [3].

An educational system has large number of educational data. This data may be students' data, teachers' data, alumni data, resource data etc. EDM focuses on the development of methods for exploring the unique types of data that come from an educational context. These data come from several source, including data from traditional face-to-face class room environment, educational software, online courseware, etc.

The educational data mining (abbreviated as EDM) is defined as: "Educational Data Mining is an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings which they learn in." [4]

In education, teachers and instructors always classify their students based on knowledge, motivation and behavior. Assessing exam answers is also a classification task, where a mark is determined according to certain evaluation criteria. Classifiers can be designed manually, based on expert's knowledge, but nowadays it is more common to learn them from real data.

2. CLASSIFICATION

Classification models describe data relationships and predict values for future observations.

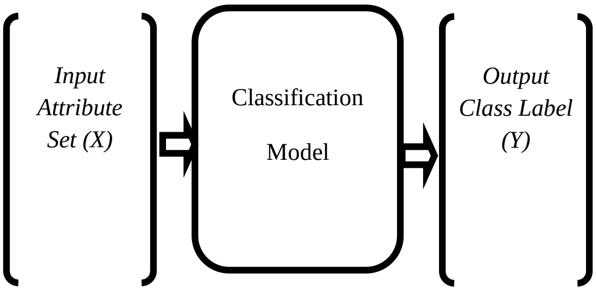

Fig. 1 Classification As The Task Of Mapping An Input Attribute Set X Into Its Class Label Y

Classification is the task of learning a target function that maps each attribute set X to one of the predefined class labels Y. There are different classification techniques, namely Decision Tree based Methods, Rule-based Methods, Memory based reasoning, Neural Networks, Naïve Bayes and Bayesian Belief Networks, Support Vector Machines. But Decision tree is the most popularly used classification method.

2.1 Decision tree

A decision tree is a logical model represented as a binary (two-way split) tree that shows how the value of a *target variable (output)* can be predicted by using the values of a set of *predictor variables (input)*.

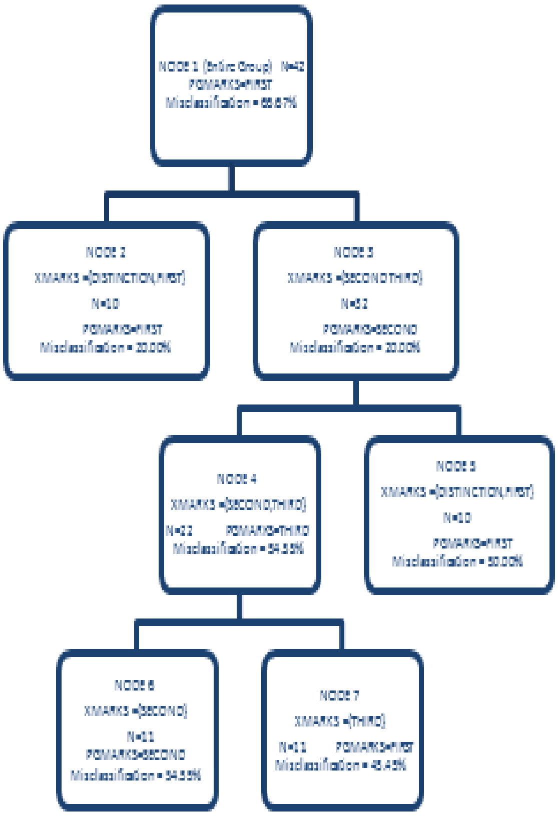
NODE 1 (Entire Group) N=42
PGMARKS=FIRST
Misclassification = 66.67%
NODE 2
XMARKS =(DISTINCTION,FIRST)
N=10
PGMARKS=FIRST
Misclassification = 20.00%
NODE 3
XMARKS =(SECOND THIRD)
N=32
PGMARKS=SECOND
Misclassification = 20.00%
NODE 4
XMARKS =(SECOND,THIRD)
N=22 PGMARKS=THIRD
Misclassification = 54.55%
NODE 5
XMARKS =(DISTINCTION,FIRST)
N=10
PGMARKS=FIRST
Misclassification = 50.00%
NODE 6
XMARKS =(SECOND)
N=11
PGMARKS=SECOND
Misclassification = 54.55%
NODE 7
XMARKS =(THIRD)
N=11 PGMARKS=FIRST
Misclassification = 45.45%

Fig.2 Example Of An Two-Way Split Decision Tree

The rectangular boxes shown in the tree are called "*nodes*". Each node represents a set of records (rows) from the original dataset. Nodes that have child nodes (nodes 1 and 3 in the tree above) are called "*interior*" nodes. Nodes that do not have child nodes (nodes 2, 5,6 and 7 in the tree above) are called "*terminal*" or "*leaf*" nodes. The topmost node (node 1 in the example) is called the "*root*" node. (Unlike a real tree, decision trees are drawn with their root at the top). The root node represents all the rows in the dataset.

2.3 Node Splitting

A decision tree is constructed by a binary split that divides the rows in a node into two groups (child nodes). The same procedure is then used to split the child groups. This process is called "recursive partitioning".

DTREG provides two methods for evaluating the quality of splits when building classification trees, First method is gini Index and second is Entropy index. Experiments have shown that entropy and Gini generally yield similar trees after pruning [5]

2.4 GINI Index

As an impurity measure, it reaches a value of zero when only one class is present at a node. Conversely, it reaches its maximum value when class sizes at the node are equal[6]

Let us denote by $f\ (i,\ j)$ the frequency of occurrence of class j at node i Then, the *GINI index* is given by:

$$I_G(i) = 1 - \sum_{j=1}^{m} f^2(i,j).$$

When a"parent" node is split into p partitions ("children"), the quality of split is given by the *GINI splitting index*:

$$GINI_{split} = \sum_{i=1}^{p} \frac{n_i}{n} GINI(i).$$

The optimal split of a node is that ensuring the lowest GINI splitting index (ideally, zero)

2.5 Entropy index

Entropy index is used to select the optimal value for node splitting based on the information maximization. The splitting point chosen based on this method should maximize the information necessary to classify objects in the resulting partition. Therefore, if all objects have the same class label, then the node entropy (impurity) is zero, otherwise it is a positive value that increases up to a maximum when all classes are equally distributed(

Entropy can be calculated by using the formula:

$$Entropy(i) = I_E(i) = -\sum_{j=1}^{m} f(i,j) \cdot \log_2 [f(i,j)],$$

where, similarly to the GINI index, $f\ (i,\ j)$ is the frequency of occurrence of class j at node i(i.e., the proportion of objects of class j belonging to node i).

When a "parent" node is split into p partitions, the quality of split is given by the *entropy splitting index*:

$$Entropy_{split} = \sum_{i=1}^{p} \frac{n_i}{n} I_E(i)$$

Again, the optimal split of a node is the one that insures the lowest entropy splitting index (ideally, zero).

3. CHALLENGES IN INDIAN HIGHER EDUCATION

One of the major challenges that higher education institutions face today is predicting the path of the student. The institutions want all the enrolled students to complete the Degree with the stipulated time. So identify the which students need assistance and support to graduate

4. OBJECTIVE

This research is conducted in order to fulfil three objectives. The objectives have been identified based on the problem stated above.

- How the student enrolment data can be preprocessed?
- How student enrolment data can be mined?
- How the student enrolment data can be used to predict the academic success?

5. REVIEW OF LITERATURE

The educational data mining is a very recent research area there is an important number of contributions published in journals, international congress, specific workshops and some ongoing books that show it is one new promising area[7]

Delavari and Beikzadeh [8]gives guidelines for using data mining in higher learning institutions and also discusses how various data mining techniques can be applied to the set of educational data. R. R. Kabra and R. S. Bichkar [9] showed 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. Surjeet Kumar Yadav ,Saurabh pal*[10]*discussed how students past academic performance can be used to create the model using ID3 decision tree algorithm that can be used for prediction of student’s enrollment in MCA course. To analyze the students performance, the author of [11] used decision tree method. Here information’s like Attendance, Class test, Seminar and Assignment marks were collected from the student’s previous database, to predict the performance at the end of the semester. Data Mining Techniques (DMT) capabilities provided effective improving tools for student performance. It showed how useful data mining can be in higher education in particularly to predict the final performance of student by analyzing the relationships between students’ behavioral [12]

6. EXPERIMENTAL SETUP

MCA students success highly depends on upon X standard , XII standard and UG performance. But in today’s education system, if a student pass in X, XII and UG of any stream can enroll for MCA Course. So it is the responsibility of the teacher to make sure that all students will complete their MCA with in stipulate time. To predict the performance of students to identify the low performer 2010 and 2011 MCA enrolment data is used

6.1 Data pre-processing

The data set used in this study was obtained from presidency college, Bangalore, Karnataka Computer Applications department of course MCA (Master of Computer Applications). In this step data stored in different tables was joined in a single table. In this step only those fields were selected which were required for data mining. Few sample dataset is shown in the TABLE1 for reference. The TABLE 2 gives the attributes and their domain.

PGMKS	SE	XMKS	XIIMK	GMKS	ATYP
FIRST	M	FIRST	FIRST	FIRST	PGCET
FIRST	M	FIRST	FIRST	SECON	PGCET
FIRST	M	FIRST	SECON	FIRST	PGCET
FIRST	M	DISTIN	SECON	SECON	PGCET
FIRST	F	FIRST	THIRD	SECON	DIREC
SECON	F	DISTIN	THIRD	SECON	DIREC
FIRST	F	DISTIN	SECON	THIRD	PGCET
THIRD	F	DISTIN	THIRD	THIRD	DIREC
FIRST	F	DISTIN	SECON	SECON	PGCET
FIRST	M	SECON	DISTIN	FIRST	PGCET
FIRST	M	SECON	DISTIN	SECON	PGCET
FIRST	F	SECON	SECON	FIRST	PGCET
FIRST	F	SECON	DISTIN	SECON	DIREC
SECON	F	SECON	FIRST	SECON	PGCET
SECON	F	SECON	SECON	SECON	PGCET
SECON	M	SECON	SECON	SECON	DIREC
SECON	M	SECON	THIRD	SECON	PGCET
SECON	F	SECON	THIRD	THIRD	PGCET
THIRD	F	SECON	THIRD	SECON	DIREC
THIRD	M	SECON	THIRD	THIRD	PGCET
THIRD	M	SECON	THIRD	THIRD	DIREC
SECON	M	SECON	THIRD	THIRD	PGCET
THIRD	F	SECON	THIRD	THIRD	DIREC
FAIL	F	SECON	THIRD	THIRD	PGCET
FIRST	F	THIRD	FIRST	DISTIN	DIREC
SECON	M	THIRD	SECON	DISTIN	DIREC
SECON	M	THIRD	FIRST	SECON	PGCET
SECON	M	THIRD	DISTIN	DISTIN	DIREC
SECON	F	THIRD	DISTIN	DISTIN	PGCET
SECON	F	THIRD	SECON	SECON	PGCET
THIRD	F	THIRD	SECON	SECON	DIREC
THIRD	M	THIRD	SECON	DISTIN	PGCET
THIRD	M	THIRD	FIRST	SECON	DIREC
THIRD	M	THIRD	SECON	THIRD	PGCET
THIRD	F	THIRD	THIRD	SECON	PGCET
FAIL	F	THIRD	THIRD	SECON	PGCET
THIRD	F	THIRD	SECON	SECON	DIREC
THIRD	M	THIRD	THIRD	SECON	PGCET
FAIL	M	THIRD	THIRD	THIRD	DIREC
FAIL	M	THIRD	THIRD	THIRD	DIREC

Table 1

Variable	Description	Possible Values
PGMKS	MCA Marks	{ DISTIN,FIRST, SECOND,FAIL}
XMKS	X Standard Marks	{ DISTIN,FIRST, SECOND,THIRD}
XIIMKS	XII Standard Marks	{DISTIN,FIRST, SECOND,THIRD}
GMKS	Under Graduate Marks	{DISTIN,FIRST, SECOND,THIRD }
ATYPE	Admission type	{ DIRECT, PGCET}

TABLE 2

6.2 Model Construction

From this data, student.cvs file was created. This file was loaded into DTREG. DTREG accepts a dataset containing of number of rows with a column for each variable. One of the variables is the "target variable" whose value is to be modeled and predicted as a function of the "predictor variables". DTREG analyzes the data and generates a model showing how best to predict the values of the target variable based on values of the predictor variables.

The process DTREG uses to build and prune a tree is complex and computationally intensive [13]. Here is an outline of the steps:

Build the tree

a) Examine each node and find the best possible split
  i) Examine each predictor variable
  ii) Examine each possible split on each predictor
b) Create two child nodes
c) Determine which child node each row goes into. This may involve using surrogate splitters.
d) Continue the process until a stopping criterion (e.g., minimum node size) is reached

7. RESULT

The accuracy of the model is 65.0 %, shown in the table 3 . That is out of 40 instances 26 instances are correctly classified.

	Actual	Misclassified	
Category	Count	count	Percentage
DISTINCTION	12	0	0.0
FAIL	3	3	100.0
FIRST	12	8	66.7
SECOND	13	3	23.1
Total	40	14	35.0

Table 3

The decision tree generated from student.arff is shown in Figure 3.From the tree it is then easy to generate rules in the form IF condition THEN outcome. Using as a training set the student enrolment data, we can build and use a decision tree that predicts the final examination result.

The rules generated from this tree are

RULE #1 for getting Distinction in PG

If (XMARK=DISTINCTION or XMARK=FIRST) then PGMARKS=DISTINTION

RULE #2 for getting Distinction in PG

If (XMARK=SECOND or THIRD and XIIMARK= DISTINTION or FIRST) then PGMARKS=DISTINCTION

RULE #3 for getting Distinction in PG

If (XMARK=SECOND or THIRD and XIIMARK=SECOND or THIRD and GMARKS=FIRST) then PGMARKS= DISTINTION

RULE #4 for getting First Class in PG

If (XMARK=SECOND or THIRD and XIIMARK= SECOND and GMARKS=DISTINCTIO or SECOND or THIRD) then PGMARKS= FIRST

RULE #5 for getting Second Class in PG

If (XMARK=SECOND and XIIMARK=THIRD and GMARKS=DISTINCTION or SECOND or THIRD) then PGMARKS=SECOND

RULE #6 for identifying the students who are likely to fail in PG

If (XMARK=THIRD and XIIMARK=THIRD and GMARKS=DISTINCTION or SECOND or THIRD) then PGMARKS = FAIL

From the Confusion matrix table 4, it is clear that

There are 12 DISTINCTION, but our model predicted 19 distinctions, so the accuracy is 82.50%.

There are 3 FAIL, but our model predicted as 0 FAIL, so accuracy is 92.50%

There are 12 FIRST, but our model predicted as 6 FIRST, so accuracy is 75%

There are 13 SECOND, But our model predicted as 15 SECOND, so accuracy is 80%

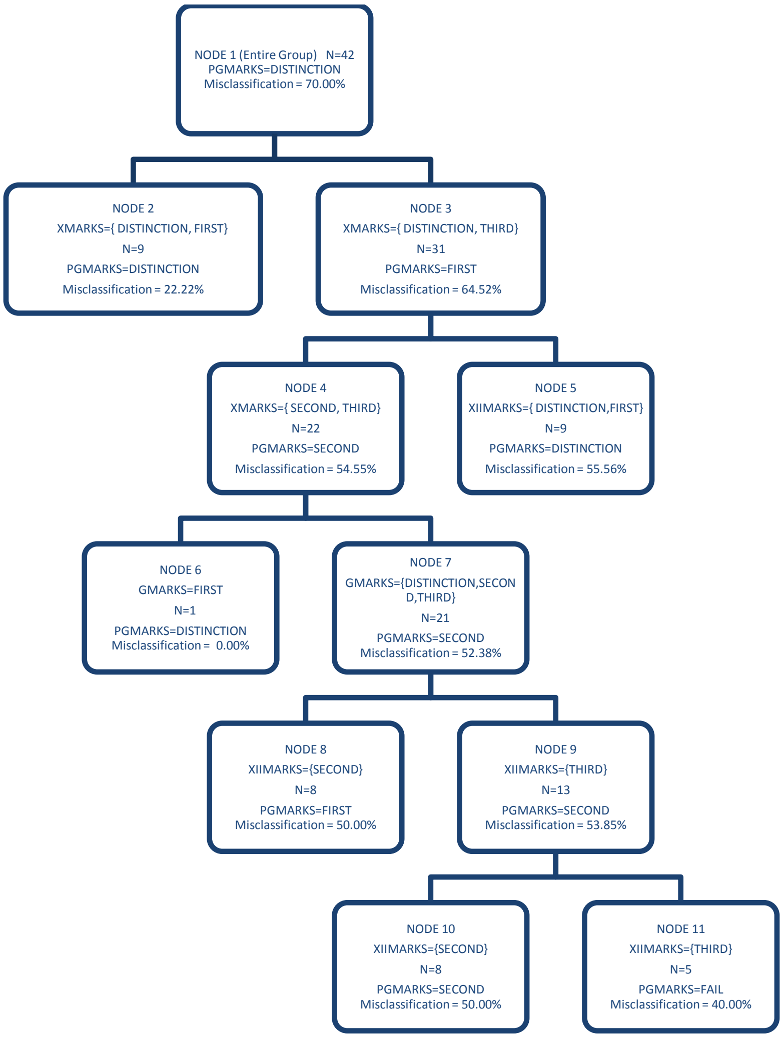
NODE 1 (Entire Group) N=42
PGMARKS=DISTINCTION
Misclassification = 70.00%
NODE 2
XMARKS={ DISTINCTION, FIRST}
N=9
PGMARKS=DISTINCTION
Misclassification = 22.22%
NODE 3
XMARKS={ DISTINCTION, THIRD}
N=31
PGMARKS=FIRST
Misclassification = 64.52%
NODE 4
XMARKS={ SECOND, THIRD}
N=22
PGMARKS=SECOND
Misclassification = 54.55%
NODE 5
XIIMARKS={ DISTINCTION,FIRST}
N=9
PGMARKS=DISTINCTION
Misclassification = 55.56%
NODE 6
GMARKS=FIRST
N=1
PGMARKS=DISTINCTION
Misclassification = 0.00%
NODE 7
GMARKS={DISTINCTION,SECON
D,THIRD}
N=21
PGMARKS=SECOND
Misclassification = 52.38%
NODE 8
XIIMARKS={SECOND}
N=8
PGMARKS=FIRST
Misclassification = 50.00%
NODE 9
XIIMARKS={THIRD}
N=13
PGMARKS=SECOND
Misclassification = 53.85%
NODE 10
XIIMARKS={SECOND}
N=8
PGMARKS=SECOND
Misclassification = 50.00%
NODE 11
XIIMARKS={THIRD}
N=5
PGMARKS=FAIL
Misclassification = 40.00%

Fig 3 Decision Tree

ACTUAL CATEGORY	PREDICTED CATEGORY					
	DISTINCTON	FAIL	FIRST	SECOND	TOTAL	ACCURACY
DISTIN	12	0	0	0	12	82.50%
FAIL	0	0	1	2	3	92.50%
FIRST	5	0	4	3	12	75%
SECOND	2	0	1	10	13	80.00%
TOTAL	19	0	6	15	40	

Table 4

By using split information the relative importance of the predictor variable is shown in fig 4.The most important attribute in predicting student's performance is found to be XMARK.

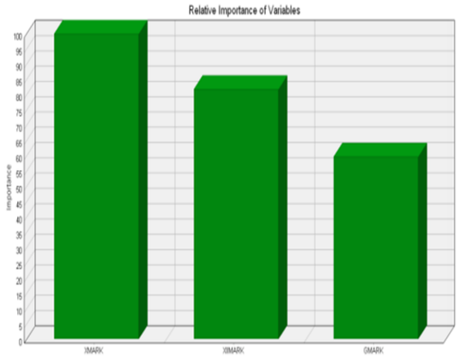

Fig. 4 Relative Importance of Variable

8. CONCLUSION

This study shows that student's enrolment data of MCA course can be used to create the model using decision tree algorithm that can be used for prediction of student's performance in MCA degree.

From the relative importance of the variable, it is clear that X and XII standard performance has great impact in MCA course.

From the confusion matrix it is clear that out of four actual categories, the accuracy of the model for the FAIL class is 92.5% 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.

9. REFERENCE

[1] "India Country Summary of Higher Education". World Bank.

[2] Report of ministry of Human Resource Development annual report 2009-10

[3] 40 million by 2020: Preparing for a new paradigm in Indian Higher Education Ernst & Young - EDGE 2011 report

[4] Refer from www.educationaldatamining.org

[5] Florin Gorunescu "Data Mining Concepts,Models and Techniques" ISBN 978-3-642-19720-8 e-ISBN 978-3-642-19721-5

[6]]Florin Gorunescu "Data Mining Concepts,Models and Techniques" ISBN 978-3-642-19720-8 e-ISBN 978-3-642-19721-5

[7] Romero,C. and Ventura, S. ,"Educational DataMining: A Survey from 1995 to 2005".Expert Systemswith Applications

[8] Delavari N, Beikzadeh M. R. "Data Mining Application in Higher LearningInstitutions ",Informatics in Education, 2008, Vol. 7,No. 1, 31–54

[9] R. R. Kabra and R. S. Bichkar, "Performance Prediction of Engineering Students using Decision Trees "International Journal of Computer Applications (0975 – 8887) Volume 36– No.11, December 2011

[10] Surjeet Kumar Yadav, Saurabh pal ," Data Mining Application in Enrollment Management: A Case Study "International Journal of Computer Applications (0975 – 8887) Volume 41– No.5, March 2012

[11] Brijesh Kumar Baradwaj, Saurabh Pal" Mining Educational Data to Analyze Students' Performance", *(IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6, 2011*

[12] SajadinSembiring, M. Zarlis, DedyHartama, Ramliana S, ElviWani "PREDICTION OF STUDENT ACADEMIC PERFORMANCE BY AN APPLICATION OF DATA MINING TECHNIQUES "2011 International Conference on Management and Artificial Intelligence IPEDR vol.6 (2011) IACSIT Press, Bali, Indonesia

[13] Phillip H. SherrodDTREG Predictive Modeling Software www.dtreg.com