# Prediction of Graduate Students for Master Degree based on Their Past Performance using Decision Tree in Weka Environment

Jaimin N. Undavia
Assistant Professor, CMPICA,
Charusat, Changa, Gujarat.

Prashant M. Dolia,Ph.D
Maharaja Krushnakumar
Sinhji University,
Bhavnagar, Gujarat.

Nikhil P. Shah
Assistant Professor,
MCA Department,
Dharamsinh Desai
University,Nadiad,Gujarat

## ABSTRACT

For generating comprehensive and precise analysis, Decision Tree technique is found as most adequate technique. Usually decision trees are used in data mining to study historical data and on the basis of the data analysis and its rules, one can predict the result. Most of the higher education institutions are suffering from low percentage of result, placement and interest of the students. To address this issue, we have suggested one Decision Support System using decision tree which predicts the post graduation stream for the students on the basis of their past academic performance. Prediction of students' performance is a great concern to the higher education institutions. So, this paper covers all the parameters which have some influence in student's performance. In this investigation, a survey cum experimental methodology is adopted to generate the data store. Paper also discusses use of decision tree for the prediction. Decision tree algorithms are applied on Post Graduate students who are either pursuing or have completed. Academic history and social data are collected and used to design the model. This model is used for the prediction of students' performance.

## General Terms

Data Mining, Classification

## Keywords

Education Data Mining, Decision Support System (DSS), Education Decision Support System (EDSS), Decision Tree, Information and Communication Technology (ICT)

## 1. INTRODUCTION

Enormous amount of data are collected and maintained by an education institution. These data can be used by an education institute for the prediction, classification, clustering, etc., for enhancement of overall systems. In this competitive environment, the higher education institutions will only be able to solve the major challenges that they have long faced by making creative use of ICT resources [5]. A Key aspect of any higher education system is to improve placement ratio and result of the students. The selection of algorithm or technique within the education system depends on categories of education system. The education systems are of two types [1]:

Traditional Education System: This is the system in which students are in direct contact with their teachers. Students' records are kept manually or digitally.

Web based learning System: Here no direct contact is required to be established between students and teachers. It is most popular learning system now a days and known as e-learning. Here a student can learn from any place without any time boundary or constraints. Students' data are collected automatically through logs.

For both the types of educational system, measuring of student performance is most vital. As performance of a student depends on many diverse factors like personal, socio-economic, psychological and other environmental variables, it is required to consider each of that before any prediction can be made for students' performance [2]. The scope of the paper is dealing very little with actual prediction of marks in their higher education. Here we are predicting their performance to guide them whether the particular branch is suitable or not on the basis of historical data of other students and past performance of the student under prediction. Possible factors also have been devised that may have some influence on the performance of the student.

The importance of DSS in education system is also narrated by human resource development [3]. As per the department of HR development, education plays always a very important role in development of any country. Education decisions are very important and have a strong impact on students, educators and society too. A wrong decision in education environment would be the wastage of efforts and money of students, educators and bad outcome would be produced and exist for a long time. Thus, it is always a right choice to investigate a decision support system (DSS) in the education domain which is called Education Decision Support System (EDSS).

A DSS must be simple, robust, easy to control, adaptive, comprehensive on important issue and easy to communicate with and to provide users with a flexible set of tools and capabilities for analyzing important blocks of data [4].

In this paper, we have planned an EDSS for the higher education institutions to assist the enrolling students for their better future. For this model, we have selected the features based on students' past performance. Past performance of a student is always indicative of his future performance.

For this study, data from 5 various colleges are of MBA and MCA of Gujarat, India are collected and conducted the study to devise the model for the prediction.

## 2. Review of Literature

A significant amount of work is found in this type of research. Here we have reviewed and used following references for this article.

R. Kabra & R. S. Cichkar have published their article for performance prediction of engineering students using decision tree. They have used decision tree technique with C4.5 algorithm[1].

V. Ramesh, P. Parkavi & K. Ramar of India have shown use of statistics in prediction of data mining. They have demonstrated prediction of students with 29 different parameters [2].

Vo Thi Ngoc Chau & Nguyen Hua Phung narrated EDSS and its importance not only for the student and the educator but also for the society [3].

Chen Z has shown the characteristics of DSS and its impact in decision making process[4].

Vasile Paul Bresfelean has published an article to analyze and predict students' behavior using decision trees in WEKA environment [5].

Naeimeh Delavari, Dr. Mohammad Reza Beikzadeh & Dr. Somnuk Phon-Amnuaisuk have demonstrated a very effective model that allows decision makers to evaluate better and enhance the higher educational organizations [6].

M.Ramaswami and R.Bhaskaran have used CHAID prediction model to analyze the interrelation between variables that are used to predict the outcome of the performance at higher secondary school education. They have used various parameters like medium of instruction, marks obtained in secondary education, location of school, living area and type of secondary education which are the strongest indicators for the student performance in higher secondary education. They have constructed CHAID prediction model [7].

Surjeet Kumar Yadav & Saurabh Pal have applied C4.5, ID3 and CART decision tree algorithm on engineering student to predict their performance and their improvement [8].

Kobus Ehlers, Malan Joubert, Johann Kinghorn & Arnold van Zyl have concluded that data mining can be effectively utilized in a DSS analyzing research outputs at universities and enable the identification of research focus, intensity and synergy. They have generated various reports that have been identified as particularly useful to institutional research management staff [9].

Behrouz Minaei-Bidgoli, DeborahA. Kashy, Gerd Kortemeyer & William F. Punch have studied the approaches how students learn and what approaches to learning lead to success. They have presented an approach to classify students in order to predict their final grade based on features extracted from logged to in an education web-based system. They have designed, implemented, and evaluated a series of pattern classifiers and compare their performance on an online course data set. They concluded that combination of multiple classifiers leads to a significant improvement in classification performance and by learning an appropriate weighting of the features used via a genetic algorithm for further improvement in prediction accuracy [10].

## 3. Classification & Decision Tree

Classification is a learning process in which past decisions are used for reference to take new decisions [11]. Basically classification is a two step processes. The training step is considered as a first step and data set with all attributes are analyzed. Classification algorithm is applied on training data to create the model. In second step, the model is used to classify the unknown data tuples for which class label is not associated. In the field of data mining, techniques such as decision tree induction, Bayesian classification, Bayesian belief networks, and neural networks are available for classifying the data [1].

A Decision Tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences. Decision Tree is the most powerful and popular approach to discover knowledge. In data mining, decision trees have great importance as it is useful in modeling and knowledge extraction from the abundance of available data.

There are five decision tree algorithms used most widely. C4.5 (J48 in weka) is used to build pruned or un-pruned decision tree. Simple Cart is another decision tree algorithm which implements pruning with minimal cost complexity. AdTree is also available to build the decision tree. RandomTree is another variation in decision tree algorithms. It chooses attributes randomly to consider in tree. PERTree is a fast decision tree learner option.

Mostly, decision tree offers following benefits in the field of data mining.

1. Prompt in prediction with relatively small computational effort.
2. Available in many data mining packages over a variety of platforms.
3. Wide spectrum of usage in classification, regression, clustering and feature selection.
4. Self explanatory and easy to follow when compacted.
5. Can process the database even in presence of error and missing values.
6. Can work with various types of data inputs like nominal, numeric and textual.

## 4. Data Collection & Pre-processing

For the construction of the model, we have collected data from various MBA/MCA colleges of Gujarat. As this model is primary model, it will reflect the percentage of the students who have selected their post graduation course correctly.

As we have used classification technique, we will use the current post graduating students' data for the training phase of the model. Then based on the current data, in the model we predict the right post graduation stream for the record to be classified.

Out of available data, we have included only these data which has influence in their study. We have considered only these data which reflect their past performance. Past performance data is easily available and their past performance gives their tendency towards their study. For this model, we have only considered the MCA and MBA PG courses and later on it can be implemented for variety of PG courses. The model is constructed to assist the students once after they complete their graduation and struggling to find their PG course. In Gujarat, the general scenario is like the commerce graduating students are likely to take admission in MBA and others are possibly admitted in MCA. The key point here to consider is that graduation is not only factor which has impact on selection of PG course. Before the student can finalize their PG course, one has to go through the academic history of student and then student should finalize their PG course [1]. For example, if one commerce graduate is good in Math or Statistics then still MCA is a good option for him/her or a science graduate with good communication skill and personality then he would have prefer MBA with marketing also.

We have decided following list of attributes to be considered for prediction of PG Course for students.

**Table – 1 List of Attributes**

| Sr. No. | Name of Attribute | Possible Value |
|---|---|---|
| 1 | HSC_Per | Distinction(above 70%),<br>Fclass(60%-70%),<br>Sclass(50%-59.99%),<br>Tclass(less than 50%) |
| 2 | HSC _Stream | {SC, CM & AT} |
| 3 | Graduation | {B.Com., B.Sc., BCA, BA} |
| 4 | Grad_Mjr_Sub | {Sc, Math, Comp, Litr} |
| 5 | Grad_Per | Distinction (above 70%),<br>Fclass(60%-70%),<br>Sclass(50%-59.99%),<br>Tclass(less than 50%) |
| 6 | Gender | {Male, Female} |
| 7 | Category | {Res, Op} |
| 8 | PG_Course | {MBA,MCA} |
| 9 | PG_Grade | Distinction (above 70%),<br>Fclass(60%-70%),<br>Sclass(50%-59.99%),<br>Tclass(less than 50%) |
| 10 | PG_Blg | Number of backlog during their PG |
| 11 | Predicted_PG_Course (Target Variable) | MCA, MBA |

Data with these attributes have been collected and then refined in order to feed up into the system. Here, meaning and purpose of each attribute has been explained.

- HSC_Per – The percentage class obtained by the students in their higher secondary education.
- HSC_Stream - The stream opted by the students in their higher secondary education. Streams are Science (SC), Commerce (CM) and Arts (AT).
- Graduation - This is the graduating degree obtained by the students. Here we have considered Bachelor of Commerce (B.Com.), Bachelor of Science (B.Sc.), Bachelor of Computer Applications (BCA) and Bachelor of Arts (BA).
- Grad_Mjr_sub – This parameter reflects the major subject which was opted by the student during graduation.

- Grad_Per – This is the percentage obtained by the students in their graduation degree. The possible values are Distinction (above 70%), First class(60%-70%), Second class(50%-59.99%), Third class(less than 50%). We are not considering the students who have failed in their graduation.
- Gender – This specifies the gender of student.
- Category – This reflects the category of students. It is required to consider category because student may avail with reservation benefits when they have taken admission in their graduation.
- PG_Course – As we are designing primary model, we have only considered two PG courses i.e. Master of Business Administration (MBA) and Master of Computer Applications (MCA).
- PG_Grad – This is the percentage obtained by the students in their post graduation degree. The possible values are Distinction (above 70%), First class(60%-70%), Second class(50%-59.99%), Third class(less tn 50%). We are not considering the students who have failed in their post graduation.
- PG_Blg – This is the number of backlogs students got during their post graduation degree.
- Predicted_PG_Course – Upon successful implementation of model, the history of students will be analyzed and then the model will predict the most suitable PG course for the students.

The prerequisite of any data mining algorithm is to clean the data before we apply any test on the data. This phase is known as Data pre-processing [2]. Data are required to clean and prepare before they undergo any test. This process includes ETL (Extraction, Transformation and Loading). The collected data are transformed into arff( Attribute Relationship File Format) as WEKA accepts data only in .arff.

## 5. RULE GENERATION FOR DECISION TREE

We have collected data from different MBA and MCA colleges of Gujarat State. These data we have used for the training and once the model is trained then we can use the same model to predict the most suitable post graduation stream for the new student.

Out of 10 parameters, the major parameters are Graduation, Grad_Mjr_Sub and Grad_Per. Category and gender has not that much impact on overall result.

Parameters like PG_Course, PG_Grade and PG_Blg will be used only for the training purpose. We will analyze these parameters in order to predict for the new students. And finally Predicted_PG_Course will predict the PG course for the students.

We have generated following basic rules.

1. The 12th Science student has only two possible graduations, BCA & B.Sc. If students have completed B.Sc. with computer with Distinction or First Class and if MCA is without any backlog and with good performance in PG then for such instance MCA is the most prominent option.
2. If students have completed B.Sc. with science with Distinction or First Class and if MCA is without any backlog and with good performance in MCA then for such instance MCA is the most prominent option. MBA is equally a preferred choice for such a candidate because actually such candidate should prefer M.Sc. but it is not in the scope of the paper.
3. If students have completed B.Sc. with science with Second Class or Third Class then again we have to check their performance in PG course and then we can suggest their PG course.
4. Students who have completed BCA with Distinction / First Class or Second Class then MCA is preferred except very poor performance in MCA.
5. If students have completed B.Sc. and gender is female then performance in MBA is required to be analyzed. For such instance MBA is also a prominent option.
6. A 12th commerce student has only two possible graduations, BCA & B.Com. If students have completed B.Com. with computer with Distinction or First Class and if MCA is without any backlog and with good performance in PG then for such instance MCA is the most prominent option.
7. If students have completed B.Com. With commerce (regardless of result) the student must proceed with MBA only as for MCA interested student he would have adopted BCA if he is interested in MCA. M.Com may be other choice but it is out of the scope of the paper.
8. According to admission criteria for MBA and MCA, Arts graduate are only eligible for MCA if they have statistics or math either in 12th or in graduation. So for all arts students, MCA is the only choice.

## 6. IMPLEMENTATION & METHODS

Above rules have been implemented through Weka 3.2 using j48 algorithm. The data are collected and converted into Attribute Relational File Format (ARFF) which is supported by Weka.

The data are arranged in ARFF format as shown below.

```
@relation student



@attribute hsc { Dist,FClass,SClass, TClass }

@attribute hscStream { SC,CM,AT }

@attribute graduation {Bcom, BSc, Bca, BA}
```

```
@attribute gradMjrSub {Science, Cm, Maths,Computer, Lit}

@attribute graduatePer { Dist,FClass,SClass, TClass }

@attribute gender { M,F }

@attribute pgCourse{ MCA, MBA }

@attribute pgGrade {Dist,FClass,SClass, TClass}

@attribute pgBacklog Real


@data

Dist, SC, BSc, Computer, SClass, M, MBA, FClass, 0

FClass,CM, BSc, Computer, FClass, M, MCA, FClass, 0

TClass,AT, BA, Computer, FClass, M, MCA, SClass, 0

SClass,CM, Bcom, Computer, FClass, M, MBA, FClass, 0

FClass,AT, BA, Lit, FClass, M, MCA, TClass, 2

TClass,SC, Bca, Computer, FClass, F, MBA, SClass, 1

TClass, SC, BSc, Science, SClass, F, MCA, SClass, 1

SClass, CM, BA, Lit, FClass, F, MCA, SClass, 2

FClass, CM, Bcom, Computer, FClass, F, MBA, SClass, 0

SClass, SC, BSc, Science, FClass, M, MCA, FClass, 0

SClass, AT, BA, Lit, FClass, M, MCA, FClass, 0

FClass, SC, Bca, Computer, SClass, F, MCA, SClass, 0

FClass, SC, Bca, Computer, FClass, M, MCA, Dist, 0

SClass, CM, Bcom, Cm, FClass, M, MBA, FClass, 0
```

The final attribute is the target class lable, i.e. predicted PG Course, MBA or MCA.

As we have limited number of records, we have used cross-validation method for classification and we have implemented the algorithm with 6 folds and 4 folds. The accuracy of the result varies from 66% to 68% respectively.

The cross validation method creates the group of records according to supplied number of folds and uses all the records for training and testing as well. The fact is explained as below.

If folds are 6 and total records are 200 then the training-testing will proceed as 200/6 i.e. approximately 34 groups. So, around 34 groups will be created consisted of 6 records. In each pass, out of 34 groups one group is considered for testing and remaining group is considered for training. So, each record is utilize in training as well as in testing. So in case of less number of records this technique proves an efficient technique for the classification accuracy.

## 7. RESULT & CONCLUSION

J48 algorithm applied on data set and we have given decision tree as shown below.

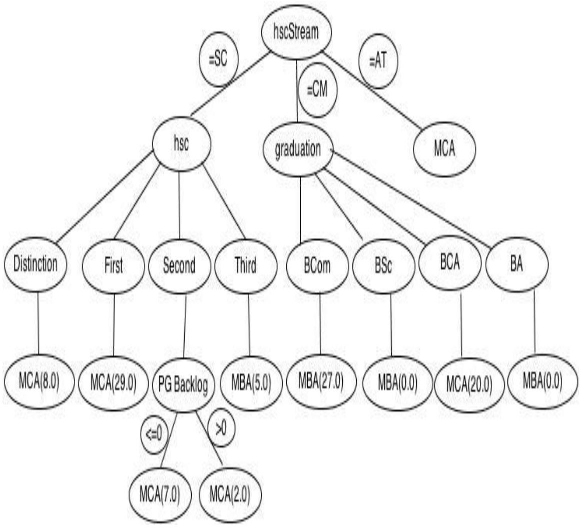

**Fig – 1. J48 Pruned Tree**

The tree can be visualized with the following generated result set.

```
Test mode:    6-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree

------------------

hscStream = SC

|   hsc = Dist: MCA (8.0/4.0)

|   hsc = FClass: MCA (39.0/10.0)

|   hsc = SClass

|   |   pgBacklog <= 0: MCA (7.0)

|   |   pgBacklog > 0: MBA (2.0)

|   hsc = TClass: MBA (5.0/1.0)

hscStream = CM

|   graduation = Bcom: MBA (27.0/8.0)
|   graduation = BSc: MBA (0.0)

|   graduation = Bca: MCA (20.0/6.0)

|   graduation = BA: MBA (0.0)

hscStream = AT: MCA (20.0)


Number of Leaves  :             10


Size of the tree :      14
```

Each record verified by the below visualize. It is sample output.

```
=== Predictions on test data ===


 inst#,    actual, predicted, error, probability distribution

     1      1:MCA      1:MCA         *1      0

     2      1:MCA      1:MCA         *1      0

     3      1:MCA      2:MBA      +   0.133 *0.867

     4      1:MCA      1:MCA         *1      0

     5      1:MCA      1:MCA         *0.667  0.333

     6      1:MCA      1:MCA         *1      0

     7      1:MCA      1:MCA         *0.889  0.111

     8      1:MCA      1:MCA         *1      0

     9      1:MCA      1:MCA         *0.667  0.333

    10      1:MCA      2:MBA      +   0.2   *0.8

    11      1:MCA      1:MCA         *1      0

    12      1:MCA      1:MCA         *1      0

    13      1:MCA      1:MCA         *0.909  0.091
```

The accuracy of result is 67.1875% and Kappa we have obtained is 0.1896. Each cell value of the confusion matrix collectively gives total number of instances. The confusion matrix reflects that 86 (73+13) records are classified correctly for MBA & MCA both whereas 42 (32+10) records are classified incorrectly for both the PG Courses.

As shown in figure, record numbered 3 and 10 are classified incorrectly and remaining are classified correctly. Correctly classified records are denoted by “*” sign, otherwise “+”.

```
=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances         86              67.1875 %

Incorrectly Classified Instances       42              32.8125 %

Kappa statistic                      0.1896

Mean absolute error                    0.393

Root mean squared error                0.4866

Relative absolute error               86.0133 %

Root relative squared error          101.8822 %

Total Number of Instances             128

=== Detailed Accuracy By Class ===

              TP Rate   FP Rate   Precision   Recall  F-Measure
ROC Area  Class

                 0.88       0.711       0.695      0.88        0.777
0.612    MCA

                0.289       0.12        0.565      0.289       0.382
0.612    MBA

Weighted Avg.    0.672     0.503      0.65      0.672     0.638
0.612


=== Confusion Matrix ===
 a  b   <-- classified as
 73 10 |  a = MCA
 32 13 |  b = MBA
```

In this paper, two class classifier problems have been discussed and the problem can be enhanced for the multiple class labels. Next version will be incorporated with multiple class labels which will predict multiple PG Courses.

## 8. REFERENCES

[1] R. R. Kabra, R. S. Cichkar – “Performance Prediction of Engineering Students using Decision Tree”, International Journal of Computer Applications(0975-8887), Vol. 36 – No. 11, December 2011.

[2] V. Ramesh, P. Parkavi, K. Ramar – “ Predicting Student Performance: A Statistical and Data Mining Approach”, International Journal of Computer Applications (0975-8887), Vol. 63 – No. *, February 2013.

[3] Vo Thi Ngoc Chau & Nguyen Hua Phung, “ A knowledge-Driven Educational Decision Support System” , 978-1-4673-0309-5/12/$31.00©2012 IEEE.

[4] Chen Z., “Computational Intelligence for Decision Support”, CRC Press LLC: 2000.

[5] Vasile Paul Bresfelean – “ Analysis and Predictions on Student’s Behavior Using Decision Tree in WekaEnvironment”, Prodeeding of the ITI 2007 29th International Conference on Information Technology Interfaces, June 25-28, 2001, Cavtat, Croatia.

[6] 6. Naeimeh Delavari, Dr. Mohammad Reza Beikzadeh & Dr. Somnuk Phon-Amnuaisuk – “ Application of Enhanced Analysis Model for Data Mining Process in Higher Education System”, ITHET 6th International Conference, 2005.

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

[8] Surjeet Kumar Yadav & Saurabh Pal – “Data Mining: A Prediction for Performance Improvements of Engineering Students using Classification”, Worlds of Computer Science and Information Technology Journal (2221-0741), Vol. 2, No. 2,51-56, 2012.

[9] Kobus Ehlers, Malan Joubert, Johann Kinghorn & Arnold van Zyl - “ A Decision Support System for Institutional Research Management in Higher Education”, 2009 International Conference on Computational Science & Engineering, IEEE Computer Society.

[10] Behrouz Minaei-Bidgoli, DeborahA. Kashy, Gerd Kortemeyer & William F. Punch – “PREDICTING STUDENT PERFORMANCE: AN APPLICATION OF DATA MINIG METHODS WITH AN EDUCATIONAL WEB-BASED SYSTEM” , 33'd ASEEIIEEE Frontiers in Education Conference T2A-13, November 5 – 8, 2003.

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