**BEST: International Journal of Management, Information Technology and Engineering (BEST: IJMITE)**
**ISSN 2348-0513**
**Vol. 2, Issue 5, May 2014, 51-56**


# MINING STUDENTS DATA TO PREDICT STUDENTS PERFORMANCE IN UNIVERSITY EXAM

**KAMLESH D KHOPKAR**

Department of Computer Applications, RMDSTIC, Maharashtra, India

## ABSTRACT

In today's competitive world the education plays very important role. There are various streams which are selected by the students to guide their career towards success. Student's success in any examination is based on his knowledge, study, behavior and many other things. The students can perform better if they are guided in proper manner. So it is important to know the different things on which students result is based on.

The main aim of any education institute is to provide good facilities, infrastructure and Knowledge to students. The students can utilize the facilities provided by the institute to get good grades in examination. Still in some cases the students face failure. In this paper the classification, decision tree technique is used to analyze the students' performance. The original data of all the students were collected. By using these decision trees the faculty can decide the proper plan for the students who need special attention.

**KEYWORDS:** Educational Data Mining, Classification, J48 Algorithm

## INTRODUCTION

In the era of Information Technology each Institute stores its data in different databases. There are various applications and different types of databases used for this purpose. The tables that are the combination of rows and columns are used to store the data related to students. The collection of tables called as database. The data related to each and every student in the institute is stored in the database. The data in databases is stored by using the different applications.

In the technique of data mining the data stored in the databases can be used for knowledge Extraction purpose. The knowledge Extraction refers to extracting knowledge from the previously stored databases. Educational Data Mining (EDM) describes a research field concerned with the application of data mining, machine learning and statistics to information generated from educational settings.

The main objective of this paper is to use data mining methodologies on the education data of the students. Data mining provided various mythologies which can be used for this purpose. In this research the classification methodology is used to evaluate the students, the decision method is used in this research.

## DATA MINING DEFINITION

The data mining, the extraction of hidden predictive information from the large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses.

The term Data mining is defined as,

**The nontrivial process of identifying valid, Novel, potentially useful, and ultimately understandable patterns in data stored in structured databases**. - Fayyad et al., (1996).

The sequences of steps identified in extracting knowledge from data are shown in Figure 1.

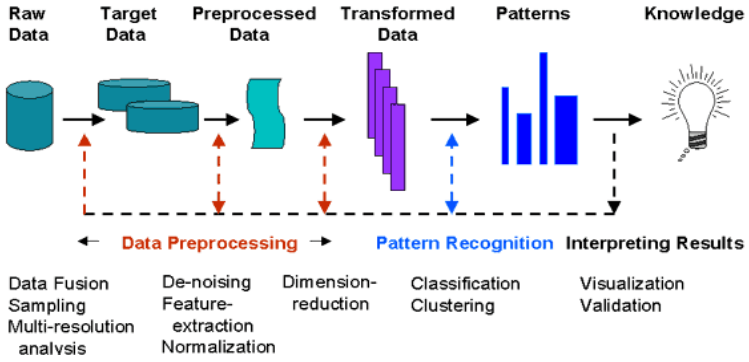


**Figure 1: The Steps of Extracting Knowledge from Data**

## CLASSIFICATION: THE DATA MINING TECHNIQUE

Classification deals with predicting the class of test instances by using attributes of the test instances, based on attributes of training instances, and the actual class of training instances. There are several types of classifiers, such as:

- Decision-Tree Classifiers: Decision-Tree classifiers perform classification by constructing a tree based on training with leaves having class labels. The tree is traversed for each test instance to find a leaf, and the class of the leaf is predicted class. Several techniques are available to construct decision trees, most of them based on greedy heuristics.
- Bayesian classifiers:- Bayesian classifiers are simpler to construct than the decision-tree classifiers, and they work better in the case of missing/null attribute values.
- The support vector machine is another widely used classification technique.

## RELATED WORK

Data mining in higher education is a recent research field and this area of research is gaining popularity because of its potentials to educational institutes.

Pandey and Pal[1] conducted the study of the 600 students to find out the new student will perform good in academics or not. The research was conducted in Awadh University, Fazabad, India.

Hijazi and Naqvi[2] conducted the research on 300 students from different colleges, in this research the study about effect of students family background on his performance was discussed.

Abeer Badr El Din Ahmed1, Ibrahim Sayed Elaraby2[3] performed research using data mining techniques for prediction of Student's Performance Using Classification Method.

Dekker et al.(2009)[4] conduct research to find out students drop-out reasons by using several classification techniques on 500 instances with different attributes.

## DATA MINING PROCESS USED IN RESEARCH

### Preparation of Data

The data set as a result set was obtained from RMD Sinhgad School of Computer Studies, Pune (Maharashtra).

The data was collected from Computer Application Department (MCA) from session 2012 to 2013. All the data is stored in a table and all errors are removed from the data, data is arranges according to different attributes.

**Data Selection**

The required data is selected from the tables; some of the attributes which are important according to research are collected. Other information of the student like address, contact number and other personal details are ignored. All the selected data is collected in the different table with the format shown below.

**Table 1: Table Structure with Variables Names**

| SNO | PREM | TG | PS | AS | P | AT | L | ENDT | MIDT | UR |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | |
| | | | | | | | | | | |

**Table 2: Variables and Descrisption**

| Variable | Description | Possible Values |
|---|---|---|
| PREM | Marks of Previous Semester | First> 60%, Second>50% & < 60%, Third >40% & <50%, Fail< 40% |
| TG | Test Grade | Poor, Average, Good |
| PS | Presentation Skills | Poor, Average, Good |
| AS | Assignments | Yes or No |
| P | Students Proficiency | Yes or No |
| AT | Attendance of Student | Poor, Average, Good |
| L | Lab Work | Yes or No |
| ENDT | Students marks in End Term Test | First> 60%, Second>50% & < 60%, Third >40% & <50%, Fail< 40% |
| MIDT | Students marks in Mid Term Test | First> 60%, Second>50% & < 60%, Third >40% & <50%, Fail< 40% |
| UR | Students University Result | First> 60%,Second>50% & < 60%, Third >40% & <50%, Fail< 40% |

**Decision Tree Induction**

A decision tree is a decision support tool that uses a tree-like graph or model of decision and their possible consequences, including chance event outcomes, resource costs, and unity.

Decision trees are commonly used in operational research, specifically in decision analysis, to help identify a strategy most likely to reach a goal. Another use of decision tree is as a descriptive means for calculating conditional probabilities. In data mining and machine learning, a decision tree is a predictive model; that is, a mapping from observations about an item to conclusions about its target value. More descriptive names for such tree models are classification tree or regression tree. In these tree structures, leaves represents classification and branches represents conjunctions of features that lead to those classifications. The machine learning technique for including a decision tree from data is called decision tree learning, or decision trees.

The popular rules are applied I automatic creation of classification trees. The Gini rule splits off a single group of as large a size as possible, whereas the entropy rule find multiple groups comprising as close to half the samples as possible.

I have use all the algorithms for classification, and best algorithm was selected based on accuracy.

The following is the table (TABLE 3) describing different algorithms and accuracy provided by those algorithms.

**Table 3: Accuracy of Algorithms**

| Algorithm | Correctly Classify Instances | Incorrectly Classified Instances |
|---|---|---|
| Simple Cart | 95% | 5% |
| REPTree | 93.3333% | 6.6667% |
| J48 | 96.6667% | 3.3333% |

The accuracy of J48 algorithm is more than other algorithms. So J48 algorithm is selected for further study.

**Decision Tree Classifier**

C4.5 is an algorithm used to generate a decision tree developed by Ross Quinlan. C4.5 is an extension of Quinlan's earlier ID3 algorithm. The decision trees generated by C4.5 can be used for classification, and for this reason, C4.5 is often referred to as a statistical classifier. C4.5 builds decision trees from a set of training data in the same way as ID3, using the concept of information entropy. At each node of the tree, C4.5 chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other.

The splitting criterion is the normalized information gain (difference in entropy). The attribute with the highest normalized information gain is chosen to make the decision. The C4.5 algorithm then recourses on the smaller sublists. J48 is an open source Java implementation of the C4.5 algorithm in the WEKA data mining tool.

**Data Mining Tool Selection**

The data mining tool selection is performed after defining the problem which is to be solved. However, more Appropriate tool should be selected for better results. Selection of data mining tool is depends on task of data mining process. In this paper we have used WEKA software for extracting rules and built a decision tree.

**Results on Training Data Set**

After applying j48 algorithm on the training data set we got following results. The data of 60 students is provided and classified by using j48 algorithm.

**A.** ***Attributes Selected***

```
Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:   mainstud
Instances:  60
Attributes:  11
        SNO
        PREM
        TG
        PS
        AS
        P
        AT
        L
        ENDT
        MIDT
        UR
Test mode: evaluate on training data
```

**B.** ***J48 pruned tree***

```
MIDT = Fail
|  PREM = Third: Third (4.0/1.0)
|  PREM = First: Third (0.0)
|  PREM = Second: Third (0.0)
|  PREM = Fail: Fail (2.0)
MIDT = First: First (16.0)
MIDT = Second: Second (21.0)
MIDT = Third
|  PS = Poor: Third (13.0)
|  PS = Good: Third (0.0)
|  PS = Average: Second (4.0/1.0)

Number of Leaves : 9
Size of the tree :  12
Time taken to build model: 0 seconds
```

**Evaluation on Training Set**

=== Summary ===

| | | |
|---|---|---|
| Correctly Classified Instances | 58 | 96.6667 % |
| Incorrectly Classified Instances | 2 | 3.3333 % |
| Kappa statistic | 0.951 | |
| K&B Relative Info Score | 5541.1531 % | |
| K&B Information Score | 99.8776 bits | 1.6646 bits/instance |
| Class complexity \| order 0 | 106.2593 bits | 1.771 bits/instance |
| Class complexity \| scheme | 6.4902 bits | 0.1082 bits/instance |
| Complexity improvement (Sf) | 99.7691 bits | 1.6628 bits/instance |
| Mean absolute error | 0.025 | |
| Root mean squared error | 0.1118 | |
| Relative absolute error | 7.2453 % | |
| Root relative squared error | 26.9904 % | |
| Total Number of Instances | 60 | |

**Decision Tree**

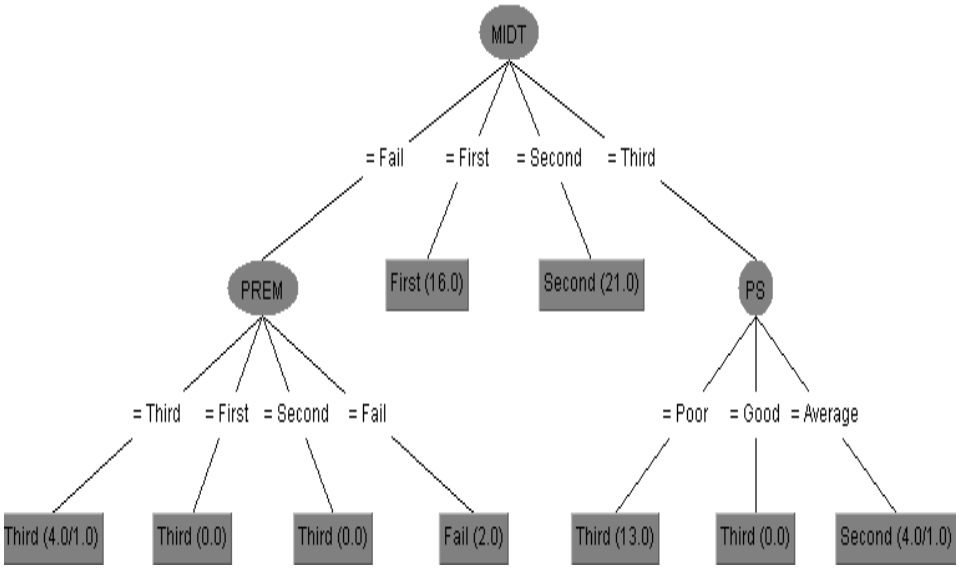


**Figure 2**

**Result on Test Data Set**

=== Predictions on test split ===

inst#, actual, predicted, error, prob distribution

1 ? 2:First + 0 *1 0 0

2 ? 3:Second + 0 0 *1 0

3 ? 3:Second + 0 0 *1 0

4 ? 3:Second + 0 0 *1 0

5 ? 1:Third + *1 0 0 0

6 ? 3:Second + 0 0 *1 0

7 ? 2:First + 0 *1 0 0

8 ? 2:First + 0 *1 0 0

9 ? 3:Second + 0 0 *1 0

10 ? 3:Second + 0 0 *1 0

## CONCLUSIONS

In this paper the information of students were collected to predict students' performance in university examination. Among various approaches of classification the decision tree method is used. The information about students' performance in the college internal exams like Mid Term Exam and End Term Exam and information like attendance. etc. were collected. The simple tool in java is created for faculty members by using which they can easily check the results.

This study is will be useful for students as well as faculty members to predict the performance before examination and plan accordingly. This study will help the institute to take decision to give special training to the students.

## REFERENCES

1. 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.
2. ST. Hijazi, and R. S. M. M. Naqvi, "Factors affecting student's performance: A Case of Private Colleges", Bangladesh e-Journal of Sociology, Vol. 3, No. 1, 20.
3. Abeer Badr El Din Ahmed1, Ibrahim Sayed Elaraby2, World Journal of Computer Application and Technology 2(2): 43-47, 2014.
4. G. W. Dekker, M. Pechenizkiy, and J.M. Vleeshouwers. Predicting students drop out: a case study. In T. Barnes, M. Desmarais, C. Romero, and S. Ventura, editors, Proceedings of the 2nd International Conference on Educational Data Mining, pages 41{50, 2009}.
5. Bhise R.B, Thorat S.S, Supekar A.K, Importance of Data Mining in Higher Education System, 2013.