THE INTERNATIONAL JOURNAL OF
SCIENCE & TECHNOLEDGE

# Data Mining Approach for Suggesting Higher Education Courses Based on Student's Performance

**Revathy P.**
Associate Professor, Computer Science Department, Rajalakshmi Engineering College, Thandalam, Tamil Nadu, India
**Kalaiarasi P.**
Student, Computer Science Department, Rajalakshmi Engineering College, Thandalam, Tamil Nadu, India
**Kavitha J.**
Student, Computer Science Department, Rajalakshmi Engineering College, Thandalam, Tamil Nadu, India
**Madhumita D. A.**
Student, Computer Science Department, Rajalakshmi Engineering College, Thandalam, Tamil Nadu, India

***Abstract:***
*The main objective of educational institutions is to provide quality education and analyze the performance of students and help them improve. Higher educational institutions ensure that students have no backlogs and above average students try to go for specialization in particular fields. Educational Data Mining (EDM) has become very important technique where educational institutions extract important details about students easily. Data Mining techniques helps in performing clustering which groups the students according to their performance which in turn shows whether the student is interested in higher studies or not. Using these analysis, classification is done to predict what the student can pursue as higher studies. If the prediction is positive for a given data, then field of interest is determined using classification and a suggestion is made as to which course the student can take up for specialization.*

***Keywords:*** *Clustering, classification, EDM, educational instructions, performance*

## 1. Introduction

Data mining is the process of discovering meaningful patterns using pattern recognition technologies as well as statistical and mathematical techniques. Data mining is also called as data or knowledge discovery.

Educational Data Mining helps in extracting useful patterns from large set of data. Educational Data Mining reduces the time and helps in discovering new relations between the data set provided. Using these analyses, a conclusion can be made on performance of the student. Evaluating the performance of a student is not restricted only to marks but includes interest and active participation. Data Mining actually helps in revealing the complex nature of data set provided and ensures providing proper relation to it. The various stages of data mining involve:

- Data Collection.
- Data Preprocessing.
- Data Mining.
- Pattern Extraction.
- Knowledge Discovery.

## 2. Related Works

Classification method was used to find the Student’s performance and provide them guidelines. They have their own institutions’ mark sheet to evaluate the performance. The accuracy so obtained is 87.9%. The classification rule obtained was used to build a decision tree. A 10-fold cross validation was used to maintain accuracy [1].

A decision tree classifier is one of the most widely used supervised learning methods used for data exploration based on divide & conquer technique. Decision tree algorithms are applied on engineering students’ past performance data and the model was generated. This model was used to predict the students’ performance. It enabled to identify the students in advance who are likely to fail and allow the teacher to provide appropriate inputs. They have found the correlation of the past performance with future performance prediction [2].

The authors have used outlier detection mechanisms for identifying outliers which improve the quality of decision making. They used outlier analysis to detect outliers in the student data. In proposed system, clustering mechanism along with univariant analysis is implemented. While clustering, the large data set was divided into clusters which consisted of outliers. After Clustering, the data points which were outside the clusters were identified and treated as outliers. Identification was done by using univariant analysis which is the simplest form of quantitative (statistical) analysis. Decision making was done based on the outlier‘s performance which results improvement in students educational standard [4].

They have covered all the parameters which have some influence in student’s performance. In this investigation, a survey and experimental methodology was adopted to generate the data store. The authors have also discussed use of decision tree for the prediction. Decision tree algorithms were applied on Post Graduate students who were either pursuing. Academic history and social data were collected and used to design the model. This model was used for the prediction of students’ performance [7].

Using K-Means clustering algorithm, the authors have predicted the pass percentage and fail percentage of the Overall students appeared for a particular examination. The results showed the students’ performance and it seemed to be accurate. The comparison between Naive Bayes algorithm and decision stump tree technique showed that the Naive Bayes techniques produced accurate result than the other and it was measured using confusion matrix. The paper also concluded that for data mining application for effective and faster results prediction, classification and clustering can be done through Weka implementation tool [8].

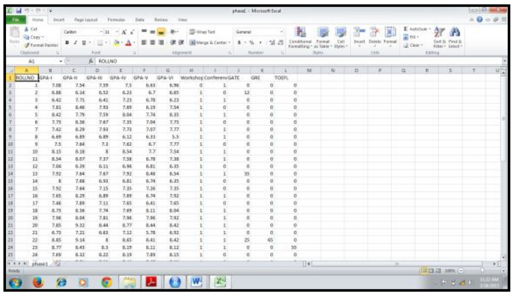
The authors have surveyed the three elements needed to make prediction on Students’ Academic Performances which are parameters, methods and tools. They have also proposed a framework for predicting the performance of first year bachelor students in computer science course. Naïve Bayes Classifier was used to extract patterns using the Data Mining Weka tool. The framework was bused as a basis for the system implementation and prediction of Students’ Academic Performance in Higher Learning Institutions. The proposed framework for predicting SAP based on the selected parameters and NBC is presented [5].

The authors have described the use of data mining techniques to improve the efficiency of academic performance in the educational institutions. Various data mining techniques such as decision tree, association rule, nearest neighbors, neural networks, genetic algorithms, exploratory factor analysis and stepwise regression can be applied to the higher education process, which in turn helps to improve student’s performance. The results show that they produced small but accurate prediction list for the student by applying the predictive models to the records of incoming new students. This study would also work to identify those students which needed special attention [6].

## 3. Data Preprocessing

Real time data of students were collected. At first, 60 student’s data was taken for this process. The following were the parameters taken into consideration:

1. Individual GPA for all the semesters.
2. Workshops attended by the student.
3. Conferences attended by the student.
4. Entrance exam scores. (if attended)



| ROLLNO | GPA-I | GPA-II | GPA-III | GPA-IV | GPA-V | GPA-VI | Workshop | Conferen | GATE | GRE | TOEFL |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 7.08 | 7.54 | 7.59 | 7.5 | 6.83 | 6.98 | 0 | 1 | 0 | 0 | 0 |
| 2 | 6.88 | 6.14 | 6.52 | 6.23 | 6.7 | 6.65 | 1 | 0 | 12 | 0 | 0 |
| 3 | 6.42 | 7.71 | 6.41 | 7.23 | 6.78 | 6.23 | 1 | 1 | 0 | 0 | 0 |
| 4 | 7.81 | 8.46 | 7.93 | 7.89 | 6.19 | 7.54 | 1 | 0 | 0 | 0 | 0 |
| 5 | 8.42 | 7.79 | 7.59 | 8.04 | 7.74 | 8.35 | 1 | 1 | 0 | 0 | 0 |
| 6 | 7.73 | 8.38 | 7.67 | 7.35 | 7.04 | 7.73 | 1 | 0 | 0 | 0 | 0 |
| 7 | 7.42 | 8.29 | 7.93 | 7.73 | 7.07 | 7.77 | 1 | 1 | 0 | 0 | 0 |
| 8 | 6.69 | 6.89 | 6.89 | 6.12 | 6.33 | 5.5 | 1 | 1 | 0 | 0 | 0 |
| 9 | 7.5 | 7.64 | 7.3 | 7.62 | 6.7 | 7.77 | 1 | 0 | 0 | 0 | 0 |
| 10 | 8.15 | 8.18 | 8 | 8.54 | 7.7 | 7.54 | 1 | 1 | 0 | 0 | 0 |
| 11 | 8.54 | 8.07 | 7.37 | 7.58 | 6.78 | 7.38 | 1 | 1 | 0 | 0 | 0 |
| 12 | 7.04 | 6.39 | 6.11 | 6.96 | 6.81 | 6.35 | 1 | 0 | 0 | 0 | 0 |
| 13 | 7.92 | 7.84 | 7.67 | 7.92 | 8.48 | 8.54 | 1 | 1 | 35 | 0 | 0 |
| 14 | 8 | 7.68 | 6.93 | 6.81 | 6.74 | 6.25 | 1 | 0 | 0 | 0 | 0 |
| 15 | 7.92 | 7.64 | 7.15 | 7.35 | 7.26 | 7.35 | 1 | 0 | 0 | 0 | 0 |
| 16 | 7.65 | 8.29 | 6.89 | 7.89 | 6.74 | 7.92 | 1 | 1 | 0 | 0 | 0 |
| 17 | 7.46 | 7.89 | 7.11 | 7.65 | 6.41 | 7.65 | 1 | 0 | 0 | 0 | 0 |
| 18 | 8.73 | 8.36 | 7.74 | 7.89 | 8.11 | 8.04 | 1 | 1 | 0 | 0 | 0 |
| 19 | 7.96 | 8.04 | 7.81 | 7.96 | 7.96 | 7.92 | 1 | 1 | 0 | 0 | 0 |
| 20 | 7.85 | 9.32 | 8.44 | 8.77 | 8.44 | 8.42 | 1 | 1 | 0 | 0 | 0 |
| 21 | 6.73 | 7.21 | 6.63 | 7.12 | 5.78 | 6.92 | 1 | 1 | 0 | 0 | 0 |
| 22 | 8.85 | 9.14 | 8 | 8.65 | 8.41 | 8.42 | 1 | 1 | 25 | 65 | 0 |
| 23 | 8.77 | 8.43 | 8.3 | 8.19 | 8.11 | 8.12 | 1 | 1 | 0 | 0 | 55 |
| 24 | 7.69 | 8.32 | 8.22 | 8.19 | 7.89 | 8.15 | 1 | 0 | 0 | 0 | 0 |

*Figure 1*

The above data for all the students were considered and this data served as an input to clustering algorithm. This data helped in grouping students who were willing to pursue higher studies. The figure 1. is an example for input for clustering algorithm.

For identifying the interest towards specific field, another set of data was taken into consideration. The following were the parameters:

1. Marks scored in college exams in each subject.
2. Attendance for particular subject.
3. Lab performance in particular subject.
4. Semester marks in particular subjects.

From the above parameters, a single parameter, namely ‘Subject Performance’ was derived. Each subject was given a subject code (eg: CS2302 for Computer Networks). For this subject, first the average of college marks were taken. This total mark of 100 was converted to a total of 80. The remaining 20 marks were given based on the attendance and lab performance in the particular subject. The obtained marks were compared with the actual semester marks and a grade was given to denote the “overall performance” in the subject. It was denoted using 5 different grades (S-91 to 100,A-81 to 90,B-71to 80,C-61 to 70,D-55 to 60,E- 50 to 55).Marks below 50 were considered poor and were not taken into account. This grade leads to the derivation of the parameter “subject performance”. Subject Performance had only three values namely excellent, good and average. Poor category was ignored. The grades S and A were given the value as ‘excellent’, B and C as ‘good’, D and E as ‘average’. Thus, each subject was finally a given a single value as illustrated above. This served as an input to classification algorithm.

During Pre-processing, the core identification was done . In this step, the subjects were segregated according to their related field. For e.g., Subjects like Computer Networks (CN), Cryptography and Network Security(CNS) and Mobile and Pervasive Computing (MPC) were combined to form a core called as ‘Networking’. Likewise, all the programming languages were combined under the same group called as ‘Programming Languages’. This enabled us to identify the interest of a student in a particular field of study.

## 4. Data Mining Techniques

### *4.1. Clustering*

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups(cluster).K-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem.

#### 4.1.1. The Algorithm is composed of the Following Steps

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroids.
3. When all objects have been assigned, recalculate the positions of the K centroids.
4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

### *4.2. Classification*

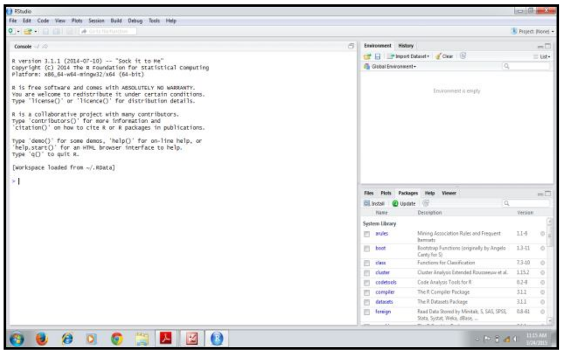
The Naive Bayes classifier works on a simple, but a comparatively intuitive concept. Also, in some cases it is also seen that Naive Bayes outperforms many other comparatively complex algorithms. It makes use of the variables contained in the data sample, by observing them individually, independent of each other.

The Naive Bayes classifier is based on the Bayes rule of conditional probability. It makes use of all the attributes contained in the data, and analyses them individually as though they are equally important and independent of each other. It works under the assumption that one attribute works independently of the other attributes contained by the sample.

P (outcome/evidence) = {P (Likelihood of Evidence) x Prior prob of outcome }/P (Evidence)

## 5. Data Analysis Using RStudio

R (Revolution) is a free software environment for statistical computing and graphics. It provides a wide variety of statistical and graphical techniques. R can be extended easily via packages. As on March 11, there are more than 2800 packages available in the CRAN package repository. The strength of R is the ease with which well-designed plots can be produced.



```
R version 3.1.1 (2014-07-10) -- "Sock it to Me"
Copyright (C) 2014 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Workspace loaded from ~/.RData]

> |
```

| | Name | Description | Version |
|---|---|---|---|
| System Library | | | |
| □ | arules | Mining Association Rules and Frequent Itemsets | 1.1-6 |
| □ | boot | Bootstrap Functions (originally by Angelo Canty for S) | 1.3-11 |
| □ | class | Functions for Classification | 7.3-10 |
| □ | cluster | Cluster Analysis Extended Rousseeuw et al. | 1.15.2 |
| □ | codetools | Code Analysis Tools for R | 0.2-8 |
| □ | compiler | The R Compiler Package | 3.1.1 |
| □ | datasets | The R Datasets Package | 3.1.1 |
| □ | foreign | Read Data Stored by Minitab, S, SAS, SPSS, Stata, Systat, Weka, dBase, ... | 0.8-61 |

*Figure 2*

- Advantages
  - ➢ Very extensive statistical library.
  - ➢ Ability to make a working machine learning program in just 40 lines of code.
  - ➢ Numerical programming is better integrated in R.
  - ➢ R has better graphics.
  - ➢ R is more transparent since the Orange is wrapped C++ classes.
  - ➢ Easy to combine with other statistical calculations.
  - ➢ Import and export of data from spreadsheet is easier in R, spreadsheet are stored in a data frames that the different machine learning algorithms are operating on.
  - ➢ Programming in R really is very different, you are working on a higher abstraction level, but you do lose control over the details.

- Limitation
  - Less specialized towards data mining.
  - There is a steep learning curve, unless you are familiar with array languages.

*5.1. Clustering using RStudio*

Clustering performed in RStudio for the given dataset groups the students according to their performance and active participation. This enables us to effectively identify the students with higher education scope. Figure 3 is an example obtained for clustering sample student data.

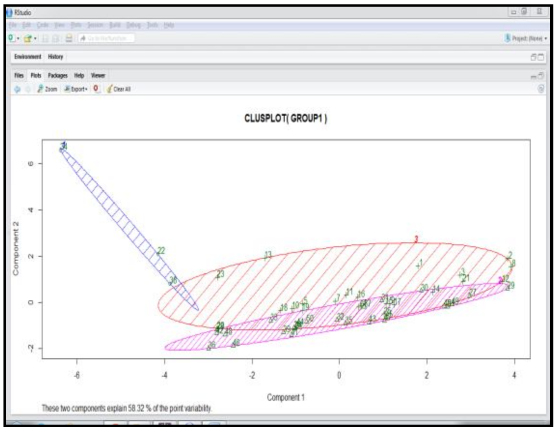


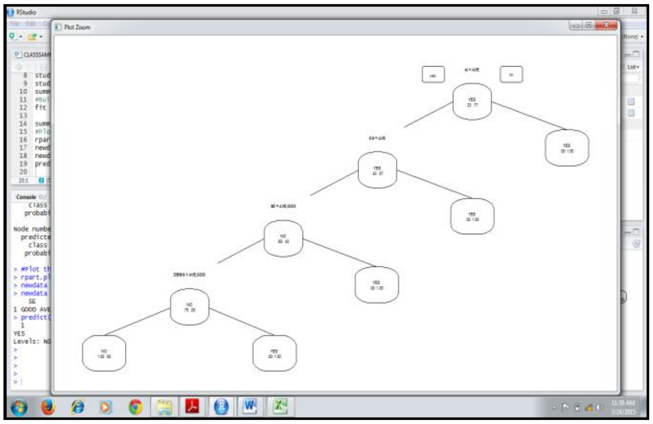
*Figure 3*

The result of the above clustering was given as input to classification algorithm.If the result was positive in clustering i.e. if the student had the criteria for higher studies then classification using Naives Baye Algorithm was done to find the field of interest for the student .

*5.2. Classification using RStudio*

The subjects were grouped according to the related field(eg: all programming subjects were grouped as one). Naives Baye Classifier algorithm was developed for each of such groups, enabling us to identify the student's interest in a particular group. The input given to Naïve Bayes Classifier was a training set. The Classifier predicted whether a student has interest in a particular field or not. Likewise, each classifier predicted for all such groups. The subject performance attribute, which was derived during pre-processing stage was given as input to these Classifiers.The Classifiers predicted the groups in which a student was interested. Using this specialization in a field was suggested.

A sample classifier is Shown in Figure 4.



```
8  stud
9  stud
10 summ
11 #Bui
12 fit
13
14 summ
15 #Plo
16 rpar
17 newd
18 newd
19 pred
20
```

```
  class
  probabi

Node numbe
  predicte
    class
  probabi

> #Plot th
> rpart.pl
> newdata
> newdata
   SE
1 GOOD AVE
> predict(
  1
YES
Levels: NO
>
>
>
>
>
```

*Figure 4*

**6. Result and Discussion**

From the above analysis, the interest in various fields of each and every student can be identified. This is combined with the Clustering output so as to predict the specialization a student can look for in his/her higher studies. This can be effectively extended in predicting the institution a student can pursue his/her higher studies along with the course name.

## 7. Conclusion and Future Work

Using k-means algorithm, we grouped the students who can pursue higher studies in future. Using Naive Bayes Classifier, the various fields a student interested in was identified. By effectively combing the both, the institution and the course a student can pursue his higher studies can be predicted. The main aim of this paper is to guide students with less confusion in selecting their master's program. And the institution can effectively train a student in their interested field, thereby effective engineers can be obtained.

## 8. References

1. Qasem A. Al-Radaideh, Ahmad Al Ananbeh, and Emad M. Al-Shawakfa , "A Classification Model for Predicting the Suitable Study Track for School Students" ,2011.
2. R. R. Kabra, R. S. Bichkar ,"Performance Prediction of Engineering Students using Decision Trees" , International Journal of Computer Applications, Volume 36– No.11, December 2011.
3. Surjeet Kumar Yadav , Saurabh Pal ,"Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification" , (WCSIT) Vol. 2, No. 2, 51-56, 2012.
4. P. Ajith, M.S.S.Sai, B. Tejaswi (IJITEE) ,"Evaluation of Student Performance: An Outlier Detection Perspective" , Volume-2, Issue-2, 2013.
5. Azwa Abdul Aziz, NurHafieza Ismail and Fadhilah Ahmad , "Mining Students' Academic Performance " ,JATIT(2013), Vol.53 No.3.
6. Ajay Kumar Pal, Saurabh Pal , " Data Mining Techniques in EDM for Predicting the Performance of Students", International Journal of Computer and Information Technology (ISSN: 2279 – 0764) Volume 02– Issue 06, November 2013.
7. Jaimin N. Undavia Prashant M. Dolia,Ph.D Nikhil P. Shah , "Prediction of Graduate Students for Master Degree based on Their Past Performance Using Decision Tree in Weka Environment ".
8. M. Durairaj, C. Vijitha , "Educational Data mining for Prediction of Student Performance Using Clustering Algorithms" , International Journal of Computer Science and Information Technologies, Vol. 5 (4) , 2014.
9. J.K. Jothi Kalpana, K. Venkatalakshmi , "Intellectual Performance Analysis of Students by Using Data Mining Techniques",(ICIET'14) Volume 3, Special Issue 3, March 2014.
10. Anupama Kumar and Dr. Vijayalakshmi, " Implication of Classification Techniques in Predicting Student's Recitals",2011.
11. Kalpana Rangra and Dr.K.L.Bansal , "Comparitive Study of Data Mining Tools".
12. SamratSingh , Dr. Vikesh Kumar "Performance Analysis of Engineering Students for Recruitment Using Classification Data Mining Techniques".
13. Dr. Abdullah AL-Malaise, Dr. AreejMalibari and Mona Alkhozae ,"Students' Performance Prediction System using Multiagent Data mining Technique".