*International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868*
*Foundation of Computer Science FCS, New York, USA*
*International Conference & workshop on Advanced Computing 2014 (ICWAC 2014) – www.ijais.org*

# Novel Approach to Evaluate Student Performance using Data Mining

Rahul Raghavan
TCET,
Mumbai University
Mumbai,India

Sagar Wahal
TCET,
Mumbai University
Mumbai,India

Manas Saxena
TCET,
Mumbai University
Mumbai,India

Anil Vasoya
Assistant Professor,
TCET,
Mumbai, India

## ABSTRACT
Data mining is a process of extracting hidden information from huge volumes of data. The various data mining techniques used are Classification, Clustering and Association mining. All these techniques can be applied to educational data to predict a student's academic performance and also to determine the areas he is currently lacking in.

The student can evaluate his performance and find out area to improve. While calculating a student's performance a student's marks in previous semesters and his term test marks, attendance and other factors.

This paper proposes the use of One R algorithm and Frequency table to predict the "score" which determines how important a particular area is. The accuracy of this algorithm can be measured by comparing the predicted score with the actual score.

Teachers can forward the result of student's report. They can also determine which students are currently lacking based on their marks and other factors. Using this data teacher can motivate a student to improve his performance in a particular area. Also students can view the report themselves and can make improvements based on area which they are lacking in.

## General Terms
Classification, Pattern, Data Mining, Algorithm

## Keywords
One R Algorithm, Score, Decision Tree, Neural Networks, Knowledge Discovery.

## 1. INTRODUCTION
Data mining is a type of sorting technique, which is actually used to extract hidden patterns from databases. The major advantages of using data mining are the fast retrieval of data or information, Knowledge Discovery from databases, detection of hidden patterns, and reduction in the level of complexity, time saving etc. One way to achieve high quality education is by discovering knowledge from educational database and using it to create an environment that helps students grow better. After prediction of the performance of the student, analysis based on different parameters used to make the prediction. Some of the parameters used for prediction include Aggregate Marks, Term Work Marks. Identifying strengths and weaknesses based on the different parameters used to make the prediction.

## 2. LITERATURE SURVEY

### 2.1 Attributes Selection for Predicting Students' Academic Performance using Education Data Mining and Artificial Neural Network
In this paper association rule mining is used to generate rules for evaluation of student performance. The parameters on the basis of which the performance of the student is evaluated is validated by using artificial neural network [1].The artificial neural network selects 5 out of 8 attributes based on the accuracy obtained for correctly classified data. The evaluations are conducted using WEKA tools. In this study, the data considered is of the students who are pursuing Master of Computer Application (MCA) degree from Pune University. The results between the accuracy obtained by Neural Network on all the attributes and accuracy obtained by applying neural network technique on selected attributes have been compared. The results obtained are that if all attributes are considered then the accuracy of correctly classified data is 44.5% while if the first three attributes are removed that is if the attributes considered are graduation%, Attendance%, assignment%, UnitTest%, University result% the accuracy of correctly classified data is 46%.It shows the accuracy increases after removing some of the attributes.

Authors: Suchita Borkar(Asstt Prof., MCA Department, PCCOE, Pune), K.Rajeswari (Assoc. Prof., Department of Computer Engineering, PCCOE, Pune)

### 2.2 Performance Prediction of Engineering Students using Decision Trees
In this paper decision tree algorithm has been used for predicting the performance of students. Decision tree algorithm has been used to generate a model. This model is used to predict the performance of the students. Because the evaluation is based on specific attributes of the student the system provides an insight onto the specific areas the student can work on to improve his performance. This paper describes the model that predicts the academic performance of the engineering students in contact education system. The data is collected from S. G. R. Education Foundation's College of Engineering and Management. Data of 346 students of the institute is collected who appeared for the first year of engineering in the year 2009-10, 2010-11[2]. This paper concludes that past performance of student in various parameters can be used to predict and in turn improve the performance of the student in the future. From the paper it is clear that the true positive rate of the model for the FAIL class is 0.907. It means that the model is successfully identifying the students who are likely to fail.

Authors: R. R. Kabra (S.G.R. Education Foundation's College of Engineering and Management, Ahmednagar, India), R. S. Bichkar (G. H. Raisoni College of Engineering and Management, Pune, India)

## 2.3 Predicting Student Performance: A Statistical and Data Mining Approach

In this paper student performance has been predicted to identify those students who are most likely to fail in their upcoming exams. A survey cum experimental methodology has been used to establish a sample data for the experiment. In this paper a suitable data mining algorithm has been selected for the purpose of identifying all those factors that affect the performance of the student. The study was conducted on the high school students of Tamil Nadu. A sample of 900 students was taken from a group of schools. Students were grouped in a classroom where they were briefed clearly about the questionnaire and it took on average half an hour to fill the questionnaire. Selection of students was at random. Classification of the data has been done using WEKA tools [3]. The algorithms used are Naive Bayes, Multi Layer Perception, SMO, J48, REPTree. It was proven that Multi Layer Perception (MLP) classifier is most appropriate for predicting student performance. MLP gives 72.38% prediction, which is relatively higher than other algorithms [4]. It was also concluded that the type of school does not significantly impact the performance of the student. However the educational qualification of the parents does play a major role in the performance of the students.

Authors: V.Ramesh (Assistant Professor, Department of CSA, SCSVMV University Kanchipuram India), P.Parkavi (Assistant Professor, PG Dept. of Computer Applications, Thirumalai Engineering College Kanchipuram), K.Ramar (Principal, Einstein College of Engineering Tirunelveli, Tamil Nad

# 3. ONE R ALGORITHM

OneR, short for "One Rule", is a simple, yet accurate, classification algorithm that generates one rule for each predictor in the data, and then selects the rule with the smallest total error as its "one rule" [5]. The basic idea behind this algorithm is to test every single attribute for its value.

For this study the attributes or 'parameters' listed in Table 1 are taken into consideration

**Table 1: Input Parameters**

| | |
|---|---|
| Previous Semester(out of 100) | Extra curricular |
| Attendance(out of 100) | Family Income |
| Term Work(out of 100) | Friend Circle |
| $10^{th}$ marks(out of 100) | MH CET score(out of 200) |
| $12^{th}$ marks(out of 100) | Current posting of elder sibling |
| Number of Live KT | Mother's job profile |
| Father's education | Father's job profile |
| Mother's education | Background of elder sibling |
| Travelling Time | Hostel Student |

## 3.1 Algorithm

- Use of One R algorithm to calculate the weight age to be given to each parameter.
- Classifying each parameter in ranges such as high, medium and low.
- Classifying the target attribute in ranges such as high, medium and low.
- Calculating success percentage of each parameter

Now total error for each frequency table is calculated and the frequency table with minimum or low total error is found. A low total error means higher contribution to improve the accuracy of the model.

# 4. IMPLEMENTATION

## 4.1 One R Algorithm on a Chosen Data Set

To illustrate, sample data of 7 students from Thakur College of Engineering and Technology is collected. This sample data is shown in Table 2. Data set contains information of student for those 18 parameters on which his performance will be evaluated. One Rule algorithm is then applied on the data set to establish the rule which will be used to predict student performance for the current semester. For this study information from semester 4 is used to establish the rule. Student performance for semester 5 is predicted using that rule.

**Table 2: Sample Data of Students**

| Name | Rahul Raghavan | Manas Saxena | Sukant Mishra | Sagar Wahal | Pratik Ahuja | Umang Patel | Vignesh R |
|---|---|---|---|---|---|---|---|
| **3rd semester Marks** | 72 | 67 | 74 | 73 | 55 | 63.7 | 59.5 |
| **Attendance** | 85 | 82 | 83 | 92 | 53 | 79 | 55 |
| **Term Work** | 86 | 84 | 88 | 90 | 67 | 82 | 73 |
| **10th Marks** | 84 | 88 | 83.84 | 83.57 | 67 | 74 | 77 |
| **12th Marks** | 76 | 86 | 78 | 70 | 67 | 60 | 50 |
| **No. of Live KT** | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| **Father's education** | PG | PG | Graduate | PG | PG | Graduate | Graduate |
| **Mother's education** | Graduate | PG | Graduate | PG | Graduate | Graduate | PG |
| **Extra curricular** | high | high | high | low | high | high | high |
| **Family income** | 700000 | 1500000 | 1300000 | 1100000 | 600000 | 2500000 | 300000 |
| **Friend circle** | high | high | high | high | low | medium | low |
| **MH CET score** | 140 | 121 | 119 | 113 | NA | 107 | 103 |

| **Current posting of elder sibling** | Singapore | Mumbai | Mumbai | Bangalore | NA | NA | Mumbai |
|---|---|---|---|---|---|---|---|
| **Mother's job profile** | yes | yes | yes | no | no | no | yes |
| **Father's job profile** | no | yes | yes | yes | yes | yes | yes |
| **Background of sibling** | M.CA | B.E. | B.E. | B.E. | NA | 12th | C.A. |
| **Hostel Student** | no | no | no | no | no | no | no |
| **Travelling time** | 0.5hr | 1 hr | 2.5hr | 0.5hr | 3hr | 2hr | 3.5hr |
| **4th semester Marks** | 70 | 72 | 77 | 78 | 54.8 | 64 | 58 |

## 4.2 Classification of Input Parameters

**Table 3: Classification of Parameters (part 1)**

| **Attributes** | **Previous Semester %** | **Attendance** | **Term Work** | **$10^{th}$ Marks** | **$12^{th}$ Marks** | **Number of live KT's** | **Father's Education** | **Mother's Education** | **MH-CET** |
|---|---|---|---|---|---|---|---|---|---|
| **Description** | Marks obtained by student in previous semester | Attendance of a student | Term work scored by a student | Percentage of Marks scored by student in $10^{th}$ standard | Percentage of Marks scored by student in $10^{th}$ standard | Live KT's in current semester | Degree which the father holds | Degree which the mother holds | Marks scored by a student in MH-CET |
| **Range** | Semester %>70=high Semester3 % between 60 and 70=medium | Attendance% >80=high Attendance % between 60 and 80=medium Attendance % <60=low | Term work%> 85=high Term work % between 75 and 85=medium Term Work % <75=lo | $10^{th}$ %>85=high 10th % between 70 and 85=medium 10th % <70=low | $12^{th}$ %>80=high 12th % between 70 and 80=medium 12th % <70=low | KT's=0 high KT's=1 medium KT's>1 low | Post Graduate=high Graduate=medium Below Graduate=low | Post Graduate=high Graduate=medium Below Graduate=low | MHT-CET marks >135=high MHT-CET marks between 115 and 135=medium MHT-CET marks <115=low |

**Table 4: Classification of Parameters (part 2)**

| **Attributes** | **Family Income** | **Posting of Sibling** | **Friend Circle** | **Extra Curricular** | **Background of sibling** | **Travelling time** | **Mother's job profile** | **Father's job** | **Hostel Student** |
|---|---|---|---|---|---|---|---|---|---|
| **Description** | Income of the family | Whether the sibling is posted in Mumbai or not | Marks obtained by a student's friend in the previous semester | Activities participated by a particular student | Whether the sibling is from Technical ,Non Technical background | Time taken for commute | Whether mother is working or not? | Whether mother is working or not? | Whether the student lives in a hostel or not? |
| **Range** | Family Income above>10 lakh Family Income between 5 and 10 lakh Family Income <5 lakh | Posting of sibling in Mumbai-Yes Posting of sibling outside Mumbai-No | Average marks of 2 classmates >70=high Average marks of 2 classmates between 60 and 70=medium Average marks of 2 classmates<60=low | Post holder in college=high Volunteer in college=medium None=low | Elder Sibling with technical background= High Elder Sibling with non technical background and graduate=medium None=low | Travelling Time 0.5hr<=high Travelling Time between 0.5hr and 1.5 medium Travelling Time 1.5hr>low | Mother's Job Yes or No | Father's Job Yes or No | Hostel Student Yes or No |

## 4.3 Prediction of Marks

Now the frequency table for the input parameters is calculated

**Table 5: Frequency table generated all parameters**

| | | Class level attribute: Semester4 | | |
|---|---|---|---|---|
| | | **High** | **Medium** | **Low** |
| **Semester 3** | **High** | 2(Match) | 1 | 0 |
| | **Medium** | 1 | 1(Match) | 0 |
| | **Low** | 0 | 0 | 2(Match) |
| **Attendance** | **High** | 3(Match) | 1 | 0 |
| | **Medium** | 0 | 1(Match) | 0 |
| | **Low** | 0 | 0 | 2(Match) |
| **Term Work** | **High** | 2(Match) | 1 | 0 |
| | **Medium** | 1 | 1(Match) | 0 |
| | **Low** | 0 | 0 | 2(Match) |
| **10th Marks** | **High** | 1(Match) | 0 | 0 |
| | **Medium** | 2 | 2(Match) | 1 |
| | **Low** | 0 | 0 | 1(Match) |
| **12th Marks** | **High** | 1(Match) | 0 | 0 |
| | **Medium** | 1 | 1(Match) | 0 |
| | **Low** | 1 | 1 | 2(Match) |
| **Number of live KTs** | **High** | 3(Match) | 1 | 2 |
| | **Medium** | 0 | 1(Match) | 0 |
| | **Low** | 0 | 0 | 0 |
| **Father's Education** | **High** | 2(Match) | 1 | 1 |
| | **Medium** | 1 | 1(Match) | 1 |
| | **Low** | 0 | 0 | 0 |
| **Mother's Education** | **High** | 2(Match) | 0 | 1 |
| | **Medium** | 1 | 2(Match) | 1 |
| | **Low** | 0 | 0 | 0 |
| **MH-CET** | **High** | 0 | 1 | 0 |
| | **Medium** | 3 | 1(Match) | 0 |
| | **Low** | 0 | 0 | 1(Match) |
| **Family Income** | **High** | 3(Match) | 1 | 0 |
| | **Medium** | 0 | 1(Match) | 1 |
| | **Low** | 0 | 0 | 1(Match) |
| **Friend Circle** | **High** | 3(Match) | 1 | 0 |
| | **Medium** | 0 | 1(Match) | 0 |
| | **Low** | 0 | 0 | 2(Match) |
| **Extra Curricular** | **High** | 2(Match) | 2 | 2 |
| | **Medium** | 0 | 0 | 0 |
| | **Low** | 1 | 0 | 0 |
| **Background of sibling** | **High** | 3(Match) | 1 | 0 |
| | **Medium** | 0 | 0 | 1 |
| | **Low** | 0 | 1 | 0 |
| **Travelling time** | **High** | 1(Match) | 1 | 0 |
| | **Medium** | 1 | 0 | 0 |
| | **Low** | 1 | 1 | 2(Match) |
| | | **High** | | **Low** |
| **Mother's job profile** | **No** | 1(Match) | | 2 |
| | **Yes** | 2 | | 2(Match) |
| **Father's job profile** | **No** | 0 | | 1 |
| | **Yes** | 3 | | 3(Match) |
| **Hostel Student** | **No** | 3(Match) | | 4 |
| | **Yes** | 0 | | 0 |
| **Posting of Sibling** | **Yes** | 2(Match) | | 1 |
| | **No** | 1 | | 1(Match) |

Now the success rate of each parameter is calculated with following formula:

Success Rate= (Number of successful matches for parameter under consideration / Total number of samples)*100

Example success rate for input parameter 3rd semester marks= ((2+1+2)/7)*100=71.42%

Success rate is= 71.42%

Similarly calculation is done for success rate of remaining parameters.

The impact each parameter has on the final aggregate score, also known as Parameter Impact Rate (PIR), of a student is calculated as follows:

Parameter Impact Rate= Success Rate / ∑ Success rate.

For example calculating the Parameter Impact Rate for attendance=85.71/ (1086.11) = 0.07891.

**Table 6: Table for Parameter Impact Rate**

| Input Parameters | Success Rate (%) | Parameter Impact Rate(PIR) |
|---|---|---|
| **Previous Semester marks** | 71.42 | 0.06575 |
| **Attendance** | 85.71 | 0.07891 |
| **Term Work** | 71.42 | 0.06575 |
| **10th Marks** | 57.14 | 0.05260 |
| **12th Marks** | 57.14 | 0.05260 |
| **Number of Live KT's** | 57.14 | 0.05260 |
| **Father's education** | 42.85 | 0.03945 |
| **Mother's education** | 57.14 | 0.05260 |
| **MH-CET** | 33.34 | 0.03069 |
| **Family Income** | 71.42 | 0.06575 |
| **Posting Of Sibling** | 60.00 | 0.05524 |
| **Friend's Circle** | 85.71 | 0.07891 |
| **Extra Curricular** | 28.57 | 0.02630 |
| **Mother's Job profile** | 42.85 | 0.03945 |
| **Father's Job profile** | 42.85 | 0.03945 |
| **Background of Sibling** | 50 | 0.04603 |
| **Hostel Student** | 42.85 | 0.03945 |
| **Travelling Time** | 42.85 | 0.03945 |
| **Total** | 1086.11 | |

Once the PIR for all the parameters is obtained, this information is used to predict the marks of the student in the current semester. To predict current semester marks, first Parameter Impact Score (PIS) for the previous semester is calculated. The following formula is used:

Parameter Impact Score (PIS) = Parameter Impact Rate*(Parameter Value). For example, for the input parameter "10th Marks", PIS will be:

PIS=0.05260*83.57=4.39578

The PIS of Mr. Sagar Wahal for his score in Semester 4 is calculated here:

**Table 7: Parameter Impact Score of Sagar Wahal (Semester 4)**

| Input Parameters | Parameter Impact Score(PIS) |
|---|---|
| **Semester 3** | 4.79975 |
| **Attendance** | 7.25972 |
| **Term Work** | 5.91750 |
| **10th Marks** | 4.39578 |
| **12th Marks** | 3.68200 |
| **Number of Live KT's** | 3.41900 |
| **Father's education** | 2.36700 |
| **Mother's education** | 3.15600 |
| **MH-CET** | 1.73399 |
| **Family Income** | 4.60250 |
| **Posting Of Sibling** | 2.76200 |
| **Friend Circle** | 5.60261 |
| **Extra Curricular** | 0.26300 |
| **Mother's Job profile** | 2.36700 |
| **Father's Job profile** | 1.73580 |
| **Background of Sibling** | 2.30150 |
| **Hostel Student** | 2.36700 |
| **Travelling Time** | 2.36700 |
| **Total** | 61.09915 |

Mean PIS for Semester 4 is = 61.09915/18=3.394397

The PIS of Mr. Sagar Wahal for his score in Semester 5 is calculated here:

**Table 8: Table for Parameter Impact Score (Semester 5)**

| Input Parameters | Parameter Impact Score(PIS) |
|---|---|
| **Semester 4** | 5.12850 |
| **Attendance** | 6.31280 |
| **Term Work** | 5.58875 |
| **10th Marks** | 4.39578 |
| **12th Marks** | 3.68200 |
| **Number of Live KT's** | 3.41900 |
| **Father's education** | 2.36700 |
| **Mother's education** | 3.15600 |
| **MH-CET** | 1.73399 |
| **Family Income** | 4.60250 |
| **Posting Of Sibling** | 2.76200 |
| **Friend Circle** | 5.60261 |
| **Extra Curricular** | 0.26300 |
| **Mother's Job profile** | 2.36700 |
| **Father's Job profile** | 1.73580 |
| **Background of Sibling** | 2.30150 |
| **Hostel Student** | 2.36700 |
| **Travelling Time** | 2.36700 |
| **Total** | 60.15223 |

Mean PIS for 5th semester = 60.15223/18=3.341790

So, when mean PIS of Mr. Sagar Wahal was 3.394397 he obtained 78% marks (Semester 4). Therefore when mean PIS value is 3.341790 the marks he obtains will be:-

Predicted percentage for semester 5 is

= (78/3.394397)*3.341790 =**76.79%**

However his actual 5th semester marks =**73.20%**

Hence an approximation of the marks he will obtain in semester 5 is made.

Using similar calculations a graph was chalked out comparing Mr. Sagar Wahal's actual and predicted percentage for semester 2,3,4,5 and 6.

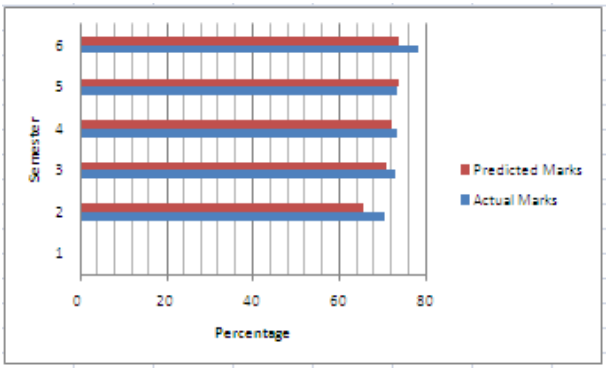


**Figure 1: Comparison between actual and predicted percentage**

## 5. RESULTS AND DISCUSSION

From the calculations performed in above section the impact of each parameter on the semester marks of a student is calculated.

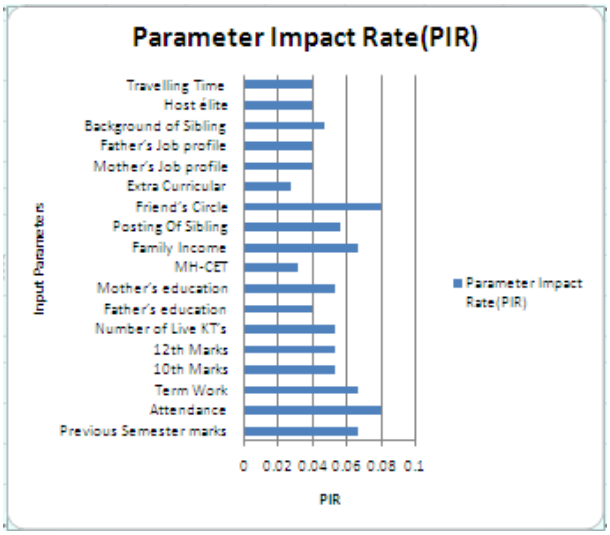


**Figure 2: Comparison of PIR for all parameters**

The above bar chart represents the impact of each parameter on the final predicted score for a data set of 7 students. The bar chart represents Input Parameters on the Y-axis and Parameter Impact Rate (PIR) on the X-axis.

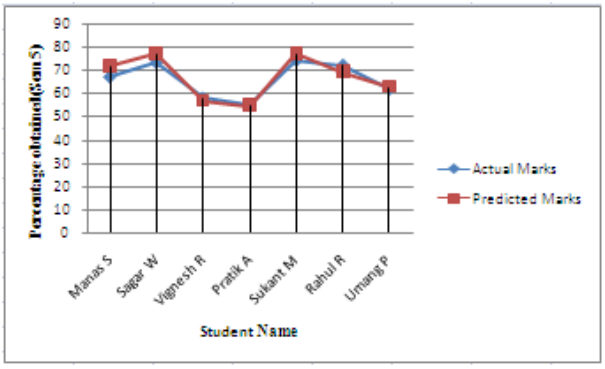


**Figure 3: Comparison of actual and predicted percentage of students in dataset (Semester 5)**

The above graph represents data set of 7 students, the red line depicts the predicted marks and blue line represents the actual marks for the 5th Semester. The maximum difference between actual and predicted marks was found for Mr. Manas Saxena. The difference was found to be 4.644 and the minimum difference was found in for Mr. Umang Patel. The difference was found to be 0.63.

From the above result it can be unambiguously stated that the methodology provided in the paper is accurate enough to predict a student's semester marks.

## 6. CONCLUSION

This study uses One Rule algorithm to evaluate student performance. The tolerance level for the predicted and the actual percentage of the student is ±4%.The study done for 7 students recorded a success rate of 85.71%. In comparison to various other classification algorithms the One Rule algorithm records a higher success rate. However the sample dataset taken in this study is small and must be increased to further validate the methodology adopted. The following graph illustrates accuracy rate of different algorithms [4]:

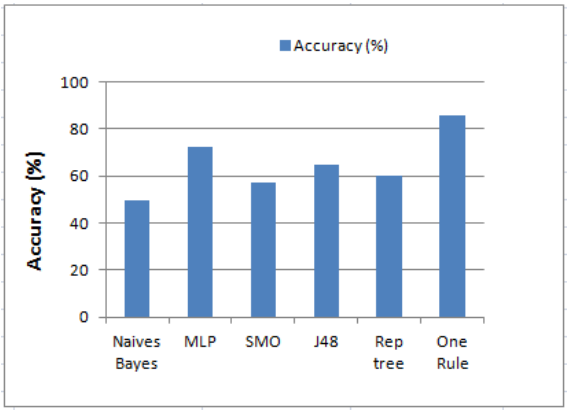


**Figure 4: Accuracy rate of different algorithms**

As the evaluation process adopted in this study is parameter based, student performance can be evaluated for each parameter and his weaknesses zeroed on. This could provide an insight into the weaknesses of the student which could be used to provide suggestions for improvement of student performance.

## 7. ACKNOWLEDGMENTS

We are foremost thankful to the Principal of our college Dr. B.K. Mishra who has taken strenuous efforts in providing us with excellent lab facilities. We are also thankful to our Head of Department Dr.Vinayak Bharadi.

## 8. REFERENCES

[1] Suchita Borkar and K.Rajeswari. 1997 Attributes Selection for Predicting Students' Academic Performance Performance using Education Data Mining and Artificial Neural Network.

[2] R. R. Kabra, R. S. Bichkar 2011,Performance Prediction of Engineering Students using Decision Trees.

[3] http://en.wikipedia.org/wiki/Weka

[4] V.Ramesh,P.Parkavi, K.Ramar ,2013. Predicting Student Performance: A Statistical and Data Mining Approach.

[5] http://www.saedsayad.com/oner.htm

[6] C. Romero, S. Ventura, "Educational data mining: A survey from 1995 to 2005", Expert system with applications 33(2007), 135-146.

[7] Singh, Randhir. An Empirical Study of Applications of Data Mining *Techniques for Predicting Student Performance in Higher Education*, 2013. Tavel, P. 2007 Modeling and Simulation Design. AK Peters Ltd.

[8] Jiawei Han and Micheline Kamber, "datamining Concepts and Techniques", Elsevier Second Edition.

[9] http://www.soc.napier.ac.uk/~peter/vldb/dm/node8.html.