ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol.2, Special Issue 5, October 2014

Approach for Predicting Student Performance Using Ensemble Model Method

Shradha Shet[1], Gayathri[2]

Department of software technology, AIMIT, St Aloysius College,Mangalore, India[1]

Department of software technology, AIMI, St Aloysius College, Mangalore, India[2]

ABSTRACT: Educational data mining focuses on developing methods for discovering knowledge from data that come from educational domain. In this paper we used educational data mining to analyze why the post-graduate students' performanceis going down and overcome the problem of low grades. There are many factors affecting student'sperformance. In our study we will focus on the main reasons that affect the students' performance. Some students are very intelligent still they cannot perform up to the mark. Their grades will decrease constantly, so we have to analyses why is that so? Different methods and techniques of data mining were compared during the prediction of Students' performance applying the data collected from the surveys conducted in AIMIT Mangalore.

I. INTRODUCTION

Data mining which is the science of digging into databases for information and knowledge retrieval, has recently developed new axes of applications and engendered an emerging discipline, called Educational Data Mining or EDM. Supervision of the academic performance of the students who are doing their higher education/post-graduation is vital during an early stage of their curricular. Indeed, theirgrades in specific core courses as well asAverage Grade points is very important. The main objective of higher education institutes is to provide a quality education and facilities to its students and to improve the quality of managerial decisions. One way to achieve highest level of quality in higher education system is by discovering knowledge from educational data to study the main factors that may affect the students' performance [5]. The discovered knowledge can be used for helpful and constructive recommendations to the academic planners in higher education institutes to enhance their decision making process, to improve students' academic performance and reduce down failure rate, to better understand students' behavior, to assist instructors, to improve teaching and many other benefits. The success of students studying at higher educationalinstitutions has been investigated for thepurpose of finding the average grades, lengthof study and similar indicators, while factorsaffecting student achievement results in aparticular course have not been sufficientlyinvestigated [5]. In this paper differenttechniques of data mining suitable forclassification have been compared: Bayesianclassifier J48 and decision trees. Their accuracy wascompared with decision trees, decision table, conjunctive Rules and with theBayesian classifier.

II. EXPERIMENTAL DESIGN

The data for the model were collected throughsurvey conducted on students of the our AIMIT College, Mangalore for the academic year 2014-2015, in which, aside from the demographic data, the data about their past success and success in college have been collected. The investigation conducted in our college, during this research period among the IT Department Students such as MCA II year, III year as well as fromMSC (ST) II year Students all together around 150 students and also along with that we conducted IQ test to test Intelligence quotient of each of the student. This analysis was conducted after the training and testing of the algorithms, making it possible to draw conclusions on possible predictors of students' success. In IQ test there were 5 questions along with the options. It included most of numerical reasonsas well as logical questions. In this test, we could analyze about the intelligence of the student.

Based on the attributes given in Table 1, questionnaire was created and response from the students has been taken and collected information regarding these attribute, was taken as data for our research. This data are real-time data. There are some other factors / attributes that may affect the student's performance, but we could not take in our research

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol.2, Special Issue 5, October 2014

because the survey was conducted in our own campus and some factors / attributes was not related to our campus environment, such as physically challenged, age, entrance exam and caste .

The table 1shows some of the common attributes and possible values that is taken as an input for our analysis

Sr.No	Attribute	Possible values	Sr.No	Attribute	Possible values
1.	Institutional support for academics /others	a. Excellent b. Good c. Average d. Not at all	10	Gender	a. Male b. Female
2.	Financial status of family	a. Very good b. Good c. Not Bad d. Bad	11	Parents support for academics/others	a. Excellent b. Good c. Average d. Not at all
3.	Class environment	a. Conducible b. Not conducible c. satisfactory	12	Academic interest	a. Very much interested b. interested c. Average d. Not interested
4.	Coaching classes	a. Yes b. No	13	Medium of instruction	a. English b. Others
5.	Motivation towards academics	a. Yes b. Motivated By others c. None	14	Availability of library/lecturers	a. Very good b. Good c. Not Bad d. Bad
6.	Distance travelled	a. Less 5km b. 5 to 10km c. 10 to 20 km d. more than 20 km	15	Time given for study /other activities	a. minimum 1 hour b. 1 to 2 Hours c. More than 2 Hours d. None
7.	Native place	a. Rural b. Urban	16	Availability of study materials	a. Notes from lectures b. Internet c. others d. None
8.	Material status	a. Yes b. No	17	Health issues	a. Yes b. No

Table 1: Factors affecting Students performance

The table 2 shows that how the collected data attributes was transformed into numerical values, we assigned different numerical values to the each attribute values.

Class	Numerical values
A	1
B	2
C	3
D	4
E	5

Table 2 :Data transformation into grades

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol.2, Special Issue 5, October 2014

III. DATA MINING TECHNIQUES

Extraction of interesting (non-trivial, implicit, Novel, hidden, previously unknown and potentially useful) patterns or knowledge from huge amount of data.Data Mining is an analytic process designed to exploring data to search the consistent patterns and/or systematic relationships between variables, and then to validate by applying the detected patterns to new subsets of data. The ultimate goal of data mining is prediction.**Prediction** involves predicting the unknown or future values of other variables of interest, using some variables or fields in the database. **Description** focuses on finding human-interpretable patterns that may describe the data. The relative importance of prediction and description for particular data mining applications can vary considerably.

3.1 The process of data mining consists of three stages:

Stage 1: Exploration. This stage starts with data preparation which may involve cleaning of data, data transformations, selecting subsets of records and - in case if the data sets with large numbers of then, depending on the nature of the analytic problem, this stage of the process of data mining may involve between a simple choice of straightforward predictors for a regression model .

Stage 2: Model building and validation. This stage involves considering various models and choosing the best one based on their predictive performance. This may sound like a simple operation, but in fact, it sometimes involves a very elaborate process. There are a variety of techniques developed to applying different models to the same data set and then comparing their performance to choose the best.

Stage 3: Deployment. That stage involves selecting the model as best in the previous stage and applying it to new data to obtain predictions or estimates of the expected outcome.

To find the main reasons that affects the students' performance we will not make use of only one method or algorithm. Instead will be using many algorithms together using ensemble model method, so that we can find accurate or exact reason affecting students' performance. Above what we have listed are the favorable or possible factors that can affect students' performance, which may or may not affect. Using algorithms in ensemble model will find the actual factors that effects students' performance.

IV. ENSEMBLE METHODS

Ensemble methods is the most influential development in Data Mining and Machine Learning in the past decade. It includes combining the multiple models into one usually more accurate than the best of its components. Ensembles can provide a critical boost to industrial challenges where predictive accuracy is more vital than model interpretability. Ensembles are useful with most of the modeling algorithms. Ensembles achieve greater accuracy on new data despite their complexity.

4.1 Ensemble Classification

Aggregation of predictions of multiple classifiers with the goal of improving accuracy.

Following are the methods that we will be using for classification:-

4.1.1J48

J48 is an open source Java implementation of the C4.5 algorithm in the weka data mining tool.

Steps

1. Check for base cases.
2. For each attribute ***a*** find the normalized information gain ratio from splitting on ***a.***
3. Let***a_best*** be the attribute with the highest normalized information gain.
4. Create a decision *node* that splits on ***a_best***
5. Recurse on the sublists obtained by splitting on ***a_best***, and add those nodes as children of *node*

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol.2, Special Issue 5, October 2014

4.1.2Decision table

Decision table is a way to decision making that involves considering a variety of conditions and their interrelationships, particular for complex interrelationship. Each decision corresponds to a variable, relation or predicate whose possible values are listed among the condition alternatives. Each action is a procedure or operation to perform, and the entries specify whether (or in what order) the action is to be performed for the set of condition alternatives the entry corresponds to.

4.1.3 Naive Bayes

Steps.

1.Let D be a training set of tuples and their associated class labels, and each tuple is represented by an n-D attribute vector **X** = (x1, x2, …, xn)

2. Suppose there are m classes C1, C2, …, Cm.

3. Classification is to derive the maximum posteriori, i.e., the maximal P(Ci|**X**)

4. This can be derived from Bayes' theorem

$$P(C_i|\mathbf{X}) = \frac{P(\mathbf{X}|C_i)P(C_i)}{P(\mathbf{X})}$$

5. Since $P(C_i|\mathbf{X}) = P(\mathbf{X}|C_i)P(C_i)$ P(X) is constant for all classes, only needs to be maximized.[1]

4.2 ENSEMBLE MODELS

4.2.1Bagging

Bagging is an ensemble model that decreases error by decreasing the variance in the result due to unstable learners, algorithms (like decision tree) whose output can change dramatically when the training data is slightly changed.

Steps

1. Create a set of *m* independent classifiers by randomly resample the training data

2. Given a training set of size *n*, create *m* bootstrap samples of size *n'* by drawing *n'* examples from the original data, ***with replacement***, *n'* usually <*n.*

If *n=n', e*ach ***bootstrap sample*** will on average contain 63.2% of the unique training examples, the rest are duplicates.

3. Combine the *m* resulting models using simple majority vote.

V.RESULTS AND DISCUSSION

5.1 RESULTS

We collected students information by doing investigation in AIMIT Mangalore, by distributing questioner among 150 students and 150 data was Collected data, that data was recorded into excel file and then through online conversion tool excel file was converted into .Arff file which is supported by weka tool. We used weka 3.6 software for our analysis. When the data was compared with various classification techniques following results were obtained.

5.1.1J48

```
Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     whatever
Instances:    150
Attributes:   33

           J48 pruned tree
           ------------------
```

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

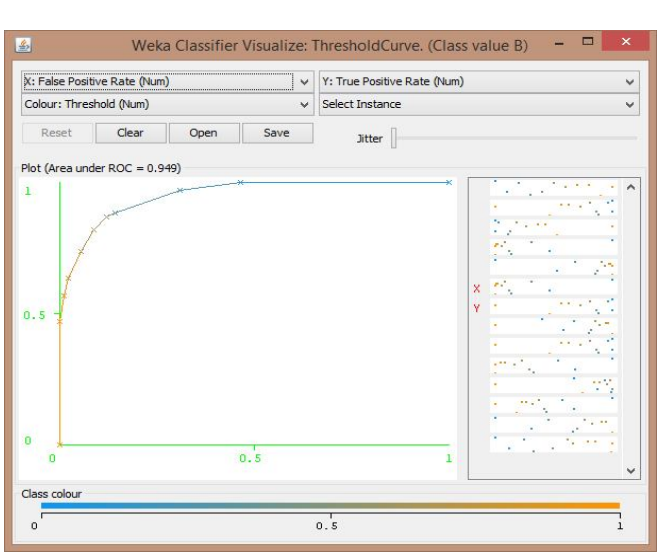
International Journal of Innovative Research in Computer and Communication Engineering

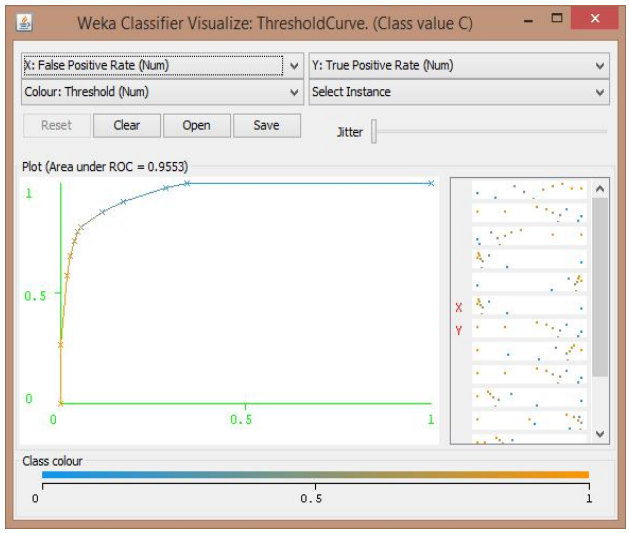
(An ISO 3297: 2007 Certified Organization)

Vol.2, Special Issue 5, October 2014

```
Time taken to build model: 0.22 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances         127               84.6667 %
Incorrectly Classified Instances        23               15.3333 %
Kappa statistic                          0.7679
Mean absolute error                      0.1468
Root mean squared error                  0.2709
Relative absolute error                 33.358  %
Root relative squared error             57.762  %
Total Number of Instances              150

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.933     0.086      0.824     0.933     0.875      0.973    A
                 0.833     0.1        0.847     0.833     0.84       0.949    B
                 0.778     0.048      0.875     0.778     0.824      0.955    C
Weighted Avg.    0.847     0.08       0.849     0.847     0.846      0.958

=== Confusion Matrix ===

  a  b  c   <-- classified as
 42  2  1 |  a = A
  6 50  4 |  b = B
  3  7 35 |  c = C
```

Figure





1:**Samples of threshold curve for some of the grades.**

Figure 1(a):threshold curve for some of the gradeBFigure 1(b): threshold curve for some of the gradeA

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol.2, Special Issue 5, October 2014

5.1.2Decision Table

```
=== Run information ===

Scheme:weka.classifiers.rules.DecisionTable -X 1 -S "weka.attributeSelection.BestFirst -D 1 -N 5"
Relation:     whatever
Instances:    150
Attributes:   33
=== Classifier model (full training set) ===

Decision Table:

Number of training instances: 150
Number of Rules : 1
Non matches covered by Majority class.
        Best first.
        Start set: no attributes
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 156
        Merit of best subset found:   40
Evaluation (for feature selection): CV (leave one out)
Feature set: 33

Time taken to build model: 0.16 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          60               40      %
Incorrectly Classified Instances        90               60      %
Kappa statistic                          0
Mean absolute error                      0.4403
Root mean squared error                  0.4692
Relative absolute error                100.0004 %
Root relative squared error            100.003  %
Total Number of Instances              150
 === Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                   0         0          0         0         0          0.476    A
                   1         1          0.4       1         0.571      0.5      B
                   0         0          0         0         0          0.463    C
 Weighted Avg.     0.4       0.4        0.16      0.4       0.229      0.482

 === Confusion Matrix ===

   a  b  c   <-- classified as
   0 45  0 |  a = A
   0 60  0 |  b = B
   0 45  0 |  c = C
```

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol.2, Special Issue 5, October 2014

5.1.3 Bagging

```
=== Run information ===

Scheme:weka.classifiers.rules.DecisionTable -X 1 -S "weka.attributeSelection.BestFirst -D 1 -N 5"
Relation:     whatever
Instances:    150
Attributes:   33
All the base classifiers:


REPTree
============

A8i < 2.5
|   A16 < 1.5
|   |   A18 < 1.5
|   |   |   A5 < 2.5 : C (4/0) [3/2]
|   |   |   A5 >= 2.5 : A (5/1) [1/0]
|   |   A18 >= 1.5
|   |   |   A11 < 2.5 : B (11/4) [11/4]
|   |   |   A11 >= 2.5 : C (2/0) [1/0]
|   A16 >= 1.5
|   |   A14 < 2.5
|   |   |   A25 < 1.5 : A (4/1) [1/0]
|   |   |   A25 >= 1.5 : B (25/6) [11/5]
|   |   A14 >= 2.5 : B (19/3) [6/0]
A8i >= 2.5
|   A12 < 2.5 : C (16/6) [9/5]
|   A12 >= 2.5
|   |   A17 < 1.5 : A (8/0) [4/2]
|   |   A17 >= 1.5 : B (6/3) [3/1]

Size of the tree : 19
```

5.1.4 Naïve Bayes

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

Naive Bayes Classifier

                        Class
Attribute                   A          B          C            buteSelection.BestFirst -D 1 -N 5"
                         (0.3)      (0.4)      (0.3)
=================================================
studentID
  mean                74.7778    74.9333    76.9778
  std. dev.           43.3394    43.2739    43.2602
  weight sum               45         60         45
  precision                 1          1          1

A1
  mean                 1.3778     1.2333     1.3111
  std. dev.            0.6762     0.4955     0.5506
  weight sum               45         60         45
  precision                 1          1          1

A2
  mean                    1.4     1.5167        1.4
  std. dev.            0.4899     0.4997     0.4899
  weight sum               45         60         45
  precision                 1          1          1

A3
  mean                 1.6444       1.85     1.6444
  std. dev.            1.0145     1.0774     1.0782
  weight sum               45         60         45
  precision                 1          1          1

Time taken to build model: 0.01 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances          87               58       %
Incorrectly Classified Instances        63               42       %
Kappa statistic                          0.3649
Mean absolute error                      0.3428
Root mean squared error                  0.4337
Relative absolute error                 77.897  %
Root relative squared error             92.474  %
Total Number of Instances              150

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.467     0.086       0.7      0.467     0.56       0.712    A
                 0.583     0.256       0.603    0.583     0.593      0.721    B
                 0.689     0.295       0.5      0.689     0.579      0.766    C
Weighted Avg.    0.58      0.217       0.601    0.58      0.579      0.732

=== Confusion Matrix ===

  a  b  c   <-- classified as
 21 11 13 |  a = A
  7 35 18 |  b = B
  2 12 31 |  c = C
```

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol.2, Special Issue 5, October 2014

5.2DICUSSION

Sr.No	Algorithm	Correctly Classified Instances	Incorrectly Classified Instances
1	J48	85%	15%
2	Decision Table	40%	60%
3	Naïve Bayes	58%	42%
Ensemble model			
4	Bagging	82%	18%

Table 3: comparison of algorithms

Table 3 shows comparison details of the algorithms that we have used in our analysis .when we compared we found that **J48** Algorithm have **15%** incorrectly Classified Instances so the classification error is very less compared to other two algorithms that is Decision Table (**60%**Incorrectly Classified Instances)and Naïve Bayes(**42%**Incorrectly Classified Instances).From this we can come to the conclusion that among three classification algorithms that we have used **J48** algorithm best suited for our application.

Along with classification algorithms we have used ensemble model, as we are not depending on only one algorithm for the classification. **Bagging** ensemble model gives **82%** of Correctly Classified Instances.
Table 4 shows the attributes and the values obtained by applying the Karl Pearson Co-efficient Technique.

Sr.No	Attribute	Values
1	A19 (Time spent on studies per day)	1.00
2	A16(study material preferred)	0.772
3.	A30(IQ Ability)	0.763
4.	A25(Self-Motivation for studies)	0.664
5.	A18(interest towards academics)	0.635

Table 4:values obtained by Karl Pearson Co-efficient Technique

The above table shows the 5attributes which will highly affects the performance of the students. These are the attributes we can consider as factors which the institutions must focus on and enforcethe actions against this attributes which will help in improving the student's performance.
There are some of the attributes which may also affects the student's performance, but in our analysis that have got less priority and so we should not consider/ give importance to these attributes as a factor that will affect the performance such as native place of the student(0.00), percentage got in previous academicals studies(0.45), distance traveled by the student (0.00), financial status of their family(0.26) and gender(0.00).

V. CONCLUSION

Many factors may affect the students' performance and if that has been observed properly in advance, ways can be suggested to improve it. To categorize the students' based on the association between performance and attributes, a good classification is needed. Also, rather than depending on the outcome of a single technique, ensemble model could do better. In our analysis, we found that J 48 algorithm is doing better than Naïve Bayesian. Also, bagging technique provides accuracy which is analogues to J 48. Moreover, the correlation between the attributes and performance has been computed and found that some attributes highly affecting the students' performance. Hence, this approach could aid the department or institution to find out means to enhance their students' performance.

REFERENCES

1. Han, J. (n.d.). In M. Kamber, "Data Mining Concepts and Techniques, Morgan Kauphann Publishers.
2. Kumar S. A. &Vijayalakshmi M. N., “Efficiency of Decision Trees in Predicting Student's Academic Performance”, First International Conference on Computer Science, Engineering and Applications(2011).

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol.2, Special Issue 5, October 2014

3. EdinOsmanbegović *, MirzaSuljić **,"Data mining approach for predicting student performance", Journal of Economics and Business, Vol. X, Issue 1, May 2012.
4. M.S. Farooq, A.H. Chaudhry, M. Shafiq, G. Berhanu (2011), "Factors affecting student's quality of academic performance: a case of secondary school level", Journals of quality and Technology Management.
5. Chady El Moucary ,"Data Mining for Engineering Schools Predicting Students' Performance and Enrollment in Masters Programs",International Journal of Advanced Computer Science and Applications.
6. http://slavnik.fe.uni-lj.si/markot/csv2arff/csv2arff.php.
7. http://www.obgyn.cam.ac.uk/cam-only/statsbook/stdatmin.html.
8. Adapted from slides by Todd Hollowayh8p://abeau<fulwww.com/2007/11/23/ ensemble---machine--- learning---tutorial.