Predicting Learning Behavior of Students using Classification Techniques

K. Prasada Rao
Research Scholar
Dept.of CSE
AITAM, Tekkali

M.V.P. Chandra Sekhara Rao, PhD
Professor, Dept.of.CSE
R.V.R.&J.C College of Engg.,

B. Ramesh
Assistant Professor
Dept.of CSE
AITAM, Tekkali

ABSTRACT
The main objective of any educational organization is to provide quality education and improve the overall performance of an institution by looking at individual performances. One way to analyze learners' performances is to identify the areas of weakness and guide their students to a better future. Although data mining has been successful in many areas, its use in student performance analysis is still relatively new, i.e. the knowledge is hidden in educational data set and it is extracted using data mining techniques. This paper discusses about a learning model for predicting student performance using classification techniques. Also the paper shows the comparative performance analysis of J48, Naïve Bayesian classifier and Random forest algorithm.

Keywords
Educational Data Mining, Random forest, Classification

1. INTRODUCTION
Measuring of academic performance of students is a challenging task because student's performance is based on different factors such as their understanding levels, capacity to learn, ability to perform well in exams, psychological factors, socio-demographic variables and so on. So the scope of the research is always there to find out what are the factors that affect the performance of the students [1].
This study focuses on investigating the factors that are affecting the performance of Computer Science Engineering students of rural-based. Educational institutes admit students under various courses from different locations, educational background, and with different board of examinations (i.e. different subjects with different level of depths). Analyzing the past performance of the students would provide a better perspective of student performance in the future. This can be very well achieved with data mining methods.

Data mining is very promising area for decision making process. It is also known as Knowledge Discovery in Databases (KDD) which discovers novel and potential useful information from large amount of data [2]. In recent years, there has been an increasing interest within educational research, termed as Educational Data Mining. Many techniques such as Naïve Bayes, Neural network, Fuzzy logic, Genetic algorithm etc., are used in Educational data mining system.

2. LITERATURE SURVEY
2.1 Data Mining
Data mining is the process of analyzing data from different perspectives and summarizing it into useful information. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. While large scale information technology has been evolving separate transaction and analytical systems, data mining provides the link between the two. Data mining deals with finding relationships and novel patterns in the data. These applications are found in fields such as statistics, machine learning, Artificial Intelligence and neural networks.

Merceron, A et al. gave a case study on educational data mining to identify the behavior of failing students and to warn students who are at risk before final exams [4]. Al-Radaideh, Al-Shawakfa and Al-Najjar (2006) applied a decision tree model to predict the final grade of students who studied the C++ course in Yarmouk University. Jordan.Romero et. al [7] have discussed about applicability of data mining techniques for the moodle course management and data mining techniques have been used extensively for mining e-learning data. Also, educational data mining was used by Minaei Bid goli et. al [8] to predict students' final grade using data collected from Web based system. Beikzadeh et. al [5] Used educational data mining to identify and enhance educational process in higher educational system. It has been observed that there is a improvement in their decision making process. Waiyamai et. al [10] used data mining to assist in development of new curricula, and to help engineering students to select an appropriate major.

2.1.1 Classification
Classification is a data mining technique used to predict group membership for data instances where the target attribute for the prediction must be discrete. Popular classification techniques include decision trees, Naïve Bayes and Random forest. The data classification process involves learning and classification of data. In Learning, the training data are analyzed by classification algorithm. In classification, Test Data are used to estimate the accuracy of the classification rules. If the accuracy is acceptable, the rules can be applied to new data records.

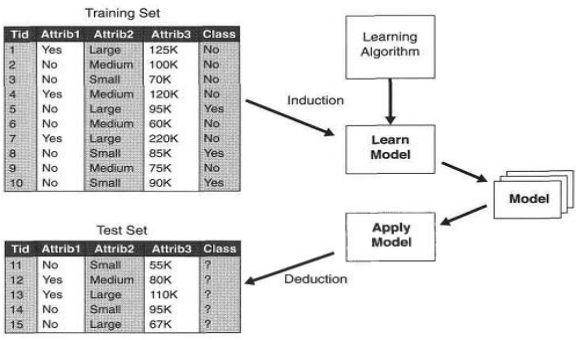

Figure 1: Classification model

2.2 J48 Algorithm

J48 is an extension of ID3 developed by the WEKA project team [3] invented by Ross Quinlan. The additional features of J48 are accounting for missing values, decision trees pruning, continuous attribute value ranges, derivation of rules, etc. Decision tree is constructed by using a fixed set of examples. The resulting tree is used to classify future samples. The example has several attributes and belongs to a class (like Yes or No). The leaf nodes of the decision tree assigned with the class label value whereas a non-leaf node is a decision node. The decision node is an attribute to test with each branch being a possible value of the attribute. Decision tree uses Entropy and Information gain measures to decide, which attribute to be selected as decision node. It selects the attribute which has the smallest entropy or largest information gain value [9]. The measures used in Decision tree are

2.2.1 Entropy

Entropy measures the impurity of t, given a collection of records with c outcomes,

$\boldsymbol{Entropy(t) = -\sum_{i=0}^{c-1} p(i|t) \log_2 p(i|t)}$ Where p(i|t) is the fraction of records belonging to class i at node t.

2.2.2 Information Gain

Information gain measures how well a given attribute separates training examples into targeted classes. The one with the highest information is selected.

$$Gain = I(parent) - \sum_{j=1}^{k} \frac{N(V_j)}{N} I(V_j)$$

Where, I(.) is the impurity of a given node, N is the total no of records at parent node, k is the no of attribute values, and $N(V_j)$ is the no of records associated with the child node, v_j.

2.3 Naïve Bayes Classifier

A Naive Bayes classifier estimates the class-conditional probability by assuming that the attributes are conditionally independent, given the class label y. The conditional independence assumption can be formally stated as follows

$$P(X|Y = y) = \prod_{i=1}^{d} P(X_i|Y = y)$$

Where each attribute set X = $\{X_1, X_2 \ldots X_d\}$ consists of d attributes [2].

2.4 Random Forest

In general, in decision tree model, one tree will be constructed from which the class label is predicted. Whereas Random Forest constructs multiple decision trees for the given data and for the test sample it predicts the class label by taking majority votes of the decision trees.

3. PROPOSED 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. The Data set used in this study contain 200 records were collected from rural-based computer science & Engineering students. The data set was used to predict and improve the performance or skills of students by using different classification techniques.

This paper mainly focuses on two parts, the first part deals with predicting the performance of Computer Science & Engineering students using classification techniques to classify them as **Excellent, Good, Average and Slow Learner.** The second part of the paper deals with analysis of accuracy and model building time of three different classification techniques such as J48, Naïve Bayes and Random Forest algorithm by continually increasing size of the data set.

4. IMPLEMENTATION

This work is carried out in three stages. In first Stage, information about students was collected. In the second stage, extraneous information was removed from the collected data and relevant information was fed into database. The third stage includes applying classification techniques on the training data to obtain decision tree.

4.1 Data Collection

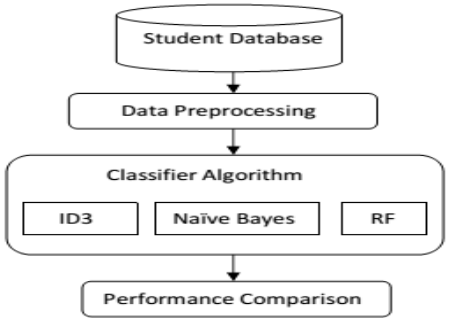

Figure 2: Paradigm of proposed system

Viewer

Relation: student3

No.	1: Gender Nominal	2: SSC_Medium Nominal	3: SSC_Percentage Nominal	4: SSC_Maths Nominal	5: Inter_Medium Nominal	6: Inter_Percentage Nominal	7: Inter_Maths Nominal	8: EamcetRank Nominal
1	m	eng	average	poor	eng	average	poor	poor
2	m	eng	good	good	eng	average	poor	poor
3	m	eng	excellent	excellent	eng	excellent	poor	poor
4	f	eng	good	excellent	eng	excellent	excellent	poor
5	m	eng	good	poor	eng	good	poor	poor
6	m	eng	good	poor	eng	good	poor	poor
7	m	eng	good	poor	eng	excellent	poor	average
8	f	native	good	excellent	eng	excellent	excellent	poor
9	f	eng	good	excellent	eng	excellent	excellent	poor
10	f	eng	good	poor	eng	good	excellent	poor
11	m	eng	good	good	eng	good	poor	poor
12	m	native	excellent	excellent	native	good	poor	poor
13	f	eng	good	excellent	eng	good	poor	poor
14	f	eng	excellent	excellent	eng	excellent	poor	average
15	f	native	excellent	excellent	eng	good	average	poor
16	f	native	average	good	native	good	poor	poor
17	f	eng	excellent	excellent	eng	excellent	excellent	average
18	f	eng	good	excellent	eng	good	good	poor
19	f	eng	excellent	excellent	eng	excellent	excellent	average
20	f	eng	excellent	excellent	eng	excellent	excellent	average

Figure 3: Sample set of Student Database

Some portion of the training dataset used for this study is shown in Figure 3. The details of each student includes: Student ID, gender, SSC and Inter medium of instruction, percentage and Maths marks, EAMCET Rank, Admission Type, Parental status (occupation and income), Lab hours spent, Assignments, Attendance, etc.

Table I. Student related Variables

Attributes	**Possible values**
Student ID	ID of the student
Gender	Male/Female
SSC Medium	{English, Native}
SSC Grade	SSC Grade {>9 and ≤10=Excellent} {>7 and ≤9=Good} {>5 and ≤7=Average} {<5=Poor}
INTER Medium	Medium of study, Other than English consider as Native. {English, Native}
INTER Grade	INTER Grade {>8=Excellent} {>7 and ≤8 =Good} {>6 and ≤7=Average} {≤6=Poor}
EAMCET Rank	Entrance rank for Engineering {≤ 10000 =Very Good} {>10000 and ≤25000= Good} {>25000 and ≤50000= Average} {>50000 = Poor}
Admission Type	Management/convener {Mgmt,Conv}
Parental Income status	Includes parents occupation and income {High,Medium,Low}
Hostler	Yes/No
Material	Resources or Material used for preparation. {T=Text book, O= Online/ Internet, R=Lecture Notes}
Assignment Submission	Done on his/her own or copied from others {Own, Others}
Hours spend to study	Hours sped to study in Library and at Home per Week. {≤3 hrs =Low} {>3 and ≤ 10 hrs =Medium} {≥ 10 hrs = High}
Attendance	{Poor, Average, Good}
Lab work	{Excellent, good, average, poor}
Communication Skills	{Good, Average, Needs to Improve}
Learning behavior (Class Variable)	{Excellent, Good, Average, slow}

4.2 Data Preprocessing

Data was pre-processed by performing the following three operations:

- Converting all attributes to categorical.
- Feature are eliminated as well as combined so as to reduce the dimensionality.
- Missing values in the database are appropriately handled by replacing them with the most commonly occurring value in that feature.

4.3 Classification Techniques for prediction

4.3.1 Decision Tree Algorithm

Step 1: If all the records in D_t belong to the same class y_t, then t is a leaf node labeled as y.

Step 2: If D_t contains records that belong to more than one class, an attribute test condition is selected to partition the records into smaller subsets. A child node is created for each outcome of the test condition and the records in D_t are distributed to the child nodes based on the outcomes. The algorithm is then recursively applied to each child node.

4.3.2 Naïve Bayes Algorithm:

The learning algorithm:

Training: Estimate the probabilities P(Y) and $P(X_i|Y)$ based on their frequencies over the training data. The learned hypothesis consists of the set of estimates

Test: Use below formula to classify new instances

$$P(X|Y=y) = \prod_{i=1}^{d} P(X_i|Y=y)$$

4.3.3 Random Forest Algorithm:

Let N $_{trees}$ be the number of trees to build for each of N_{trees} iterations

1) Select a new bootstrap sample from training set

   and grow an un-pruned tree on this bootstrap.

2) At each internal node, randomly select *m try* predictors and determine the best split using only thesepredictors.

3) Outputs overall prediction as the average response (regression) or majority vote (classification) from all individually trained trees.

5. EXPERIMENTATION AND RESULTS ANALYSIS

For the purpose of this study WEKA software package was used, which was developed at the University of Waikato in New Zealand [6]. This package has been implemented in the software language java and today stands out as probably the most competent and comprehensive package with machine learning algorithms in academic and nonprofit world.

From the collected data, 200 samples were taken for this experiment and stored in MS Excel and converted into .arff (Attribute-Relation File Format). This file was given as input to WEKA 3.7.5 tool to obtain results.

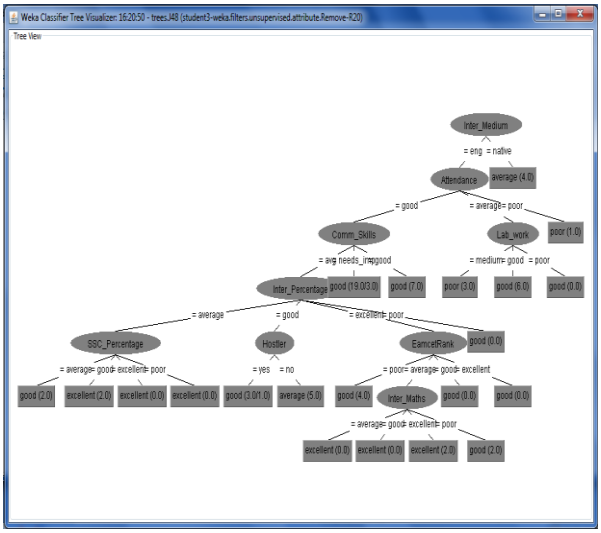

Figure 4: A sample Decision tree.

Based on the derived model, some sample rules so obtained are as follows:

If Eamcet_Rank=*poor* and Inter_percentage=good and attendance=good and Study material=*Online content* then ***Excellent Learner***.

If Eamcet_Rank=*poor* and Inter_percentage=*good* and Inter_Medium=*English* and attendance=*poor* and Admission type=*Mgmt* then ***Poor Learner***.

If Eamcet_Rank=*poor* and Inter_ percentage=*Excellent* and communication skills=*needs_improve* then ***Good Learner***.

If Eamcet_Rank=*poor* and Inter percentage=*Excellent* and Ssc_Medium=*English* and Lab_work=*good* then ***Good Learner.***

If Eamcet_Rank=*poor* and Inter percentage=*Excellent* and Ssc_Medium=*English* and Lab_work=*Average* and Study material=*Running notes* then ***Average Learner.***

The below figure shows the comparison of correctly classified instances among the three algorithms:

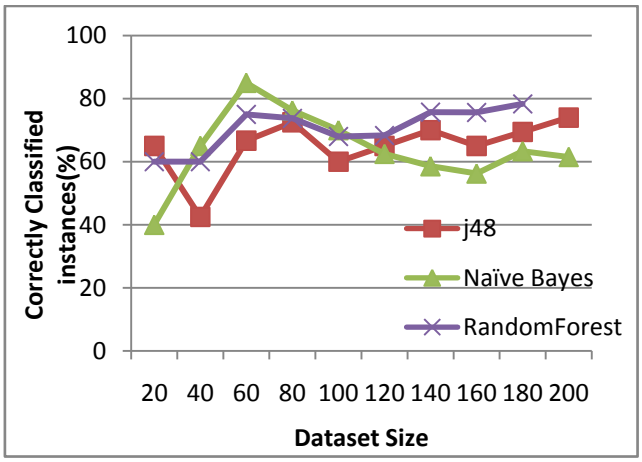

Figure 5: Prediction Accuracy

The below figure shows the comparison of Learning time for model building of the three algorithms:

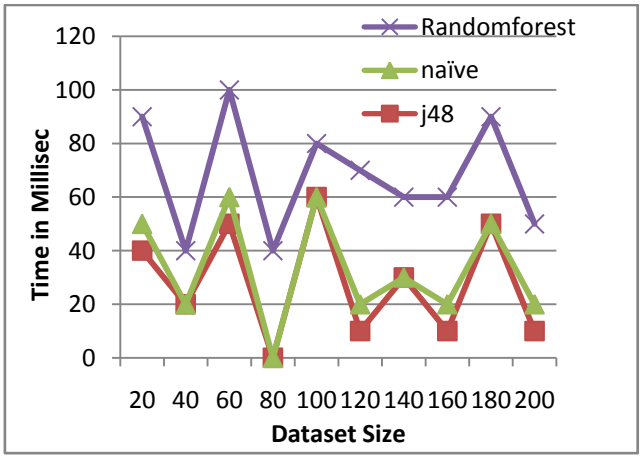

Figure 6: Learning time for Model building

6. CONCLUSION

In this paper, the classification techniques are used on student database to predict the learning behavior. This study helps to identify the slow learner, rectify the failures early and take appropriate action to improve the weaker section students in perfect manner. This paper also compares the performance of J48, Naïve Bayes and Random forest algorithm. Experimentation results concluded that as the data set size goes on increasing Random forest algorithm shows better accuracy.

After the careful study of the performance of the students, one can observe that even if the student is poor in the entrance exam, after teacher providing a good study material, they can become excellent learners. With the best practices in lab sessions, poor students also turned into good performers. With respect to performance analysis, we observed that though Random Forest algorithm took much time to construct the

model, the classification accuracy was better when compared to other algorithms.

Even though this study includes 20 features which are related to the social and demographical, In future, we want to extend this work by considering additional features such as social media and internet access which may have an impact on student performance.

7. REFERENCES

[1] Hijazi and Naive, "Factors Affecting Students' Performance" Bangladesh e-Journal of Sociology, Volume 3. Number 1. January 2006.

[2] Tan and Vipin Kumar, "Introduction to Data Mining" Pearson, 2013.

[3] http://datamining.businessintelligence.uoc.edu/home/j48-decision-tree.

[4] Merceron, A. and Yacef, K.,"Educational Data Mining: a Case Study" In Proceedings of the 12th International Conference on Artificial Intelligence in Education AIED 2005, Amsterdam, The Netherlands, IOS Press. 2005.

[5] Beikzadeh,M. and Delavari, N., "A New Analysis Model for Data Mining Processes in Higher Educational Systems". On the proceedings of the 6th Information Technology Based Higher Education and Training 7-9 July 2005.

[6] Weka, University of Waikato, New Zealand, http://www.cs.waikato.ac.nz/ml/weka/

[7] Romero, C., Ventura, S. and Garcia, E., "Data mining in course management systems: Moodle case study and tutorial". Computers & Education, Vol. 51, No. 1. pp. 368-384. 2008.

[8] Minaei-Bidgoli B., Kashy, D. Kortemeyer G., Punch W., "Predicting Student Performance: An Application of Data Mining Methods with an Educational Web-Based System". In the Processing of 33rd ASEE/IEEE conference of Frontiers in Education. 2003.

[9] Ch.Ravi Kishore, K.Prasada Rao et.al, "Performance Evaluation of Entropy and Gini using Threaded and Non Threaded ID3 on Anaemia Dataset" 2015 IEEE DOI 10.1109/CSNT.2015.112

[10] Waiyamai,K. "Improving Quality of Graduate Students by Data Mining" Department of Computer Engineering, Faculty of Engineering, Kasetsart University, Bangkok, Thailand. 2003.