Mining Educational Data to Predict Student’s Academic Performance

Jyoti Bansode
Shah And Anchor Kutchhi Polytechnic
Mumbai, India
jyoti.bansode@sakp.ac.in

Abstract: Data Mining Technique can be used in different fields to extract knowledge from large data. Data Mining is very useful in educational field to find important pattern from the data. The educational institutes are always insists on giving quality education. By using prediction method a model can be developed which can be used to predict students’ performance. Prediction can be done by using students’ academic background and family background. Different Data Mining Techniques like Classification, Clustering, Association Rule Mining, Regression etc can be used for this purpose. The prediction will help the teachers to identify the weak students and help them to improve their performance.

Keywords: Educational Data Mining, Classification, Decision Tree

I. INTRODUCTION

Educational Data Mining (EDM) is an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings and using those methods to better understand students and the settings which they learn in[10]. There are different types of educational environments:

(i) Off-Line Education (Traditional Class room): It is used to deliver knowledge and skills based on face-to-face contact. In this we have to consider students’ behavior, performance, curriculum, etc. that was gathered in classroom environment.

(ii) E-Learning and Learning Management System (LMS): E-learning provides online instruction. LMS also provides communication, collaboration, administration and reporting tools. Web Mining (WM) techniques have been applied to students’ data stored by these systems in log files and databases.

(iii) Intelligent Tutoring System (ITS) and Adaptive Educational Hypermedia System (AEHS): It adapts teaching to the needs of each particular student. DM has been applied to data picked up by these systems, such as log files, user models, etc.

In this paper Traditional Class Room Educational System is considered. The performance of the polytechnic students depends on many factors such as their family background like father’s education, father’s occupation, mother’s education, mother’s occupation. Also admission type (CAP, institute level, direct second year), category (SC, OBC, OPEN) and finally most important is the academic background i.e. percentage of SSC (Secondary School Certificate) and all previous semesters percentage has to be considered.

Predicting students’ academic performance, those students are on risk can be find out and some remedial action can be taken to improve their performance. Due to this results of the college can improve. The educational data mining can be used to get the feedback for the teachers so the teacher can improve teaching method. Also the information can be useful for those who are designing the course contents.

There are many data mining techniques can be used for this like classification, clustering, regression, association rule mining.

II. LITERATURE SURVEY:

According to Crist´Obal Romero And Sebasti´an ventura EDM can be used for Predicting Students Performance, to get feedback to the teachers, to improve teaching learning method. Different Data Mining Techniques can be used for this like Clustering, Classification, Association Rule Mining[1].

Kumar S. Anupama, Dr. Vijayalakshmi M.N suggested C4.5 decision tree algorithm can be used on marks of the students and predict their performance in terms of pass or fail in final exam. The predicted results and actual results which indicates, that there was a significant improvement in results as the prediction helped a lot to identify weak and good students and help them to score better marks. The ID3 decision tree algorithm is better in terms of efficiency and time taken to build the decision tree[2].

R. R. Kabra And R. S. Bichkar suggested that Decision tree can be used on engineering students’ past performance data to generate the model and this model can be used to predict the students’ performance. It will enable to identify the students in advance who are at risk. Giving warning to the students those are on risk of failing the students can improve their performance[3].

According to Mrinal Pandey and Vivek Kumar Sharma, different decision tree algorithms J48, NBtree, Reptree and Simple CART can be used for prediction. J48 decision tree algorithm is found to be the best suitable algorithm for model construction. Cross validation method and percentage split method were used to evaluate the efficiency of the different algorithms[6].

Charanjit Bambrath, Minakshi Bhandari, Nirali Maniar, Prof. Vandana Munde suggested Data mining is used to extract meaningful information and to develop relationships among variables stored in large data set. In this case, Apriori algorithm is used which extracts the set of rules, specific to each class and analyzes the given data to classify the student based on their performance [4].

Yadav Surjeet Kumar and Pal Saurabh suggested C4.5, ID3 and CART decision tree algorithms can be used to predict the performance of the first year engineering students. It was three class predictions. Students were classified as pass fail and...
promoted. This model was good to identifying the students that are most likely to fail[9].

III. CLASSIFICATION

Classification is a supervised learning method. It consists of two steps: 1. Model is built by analysing the data tuples from training data. 2. Test data is used to check accuracy. There are various classification techniques such as Decision Tree algorithm, Bayesian Network, Neural Network and Genetic algorithm etc can be used. These techniques can be used to build the classification model.

3.1 Decision Tree

The decision tree systems adopt a greedy (i.e. non-backtracking) top-down divide and conquer manner. Following algorithm is used to create decision tree.

3.1 Algorithm:

1) Create a node N.
2) If all the tuples in the partition are of the same class then return N as a leaf node labeled with that class.
3) If attributes list is empty then return N as a leaf node labeled with the most common class in samples.
4) Identify the splitting attribute so that resulting partitions at each branch are as pure as possible.
5) Label node N with splitting criterion which serves as test at that node.
6) If splitting attribute is discrete valued then remove splitting attribute from attribute list.
7) Let Pi be the partitions created based on the i outcomes on splitting criterion.
8) If any Pi is empty then attach a leaf with the majority class in the partition to node N.
9) Else recursively apply the complete process on each partition. 10) Return N. [3]

IV. DECISION TREE ALGORITHMS

There are many decision tree algorithms such as CART, ID3, C4.5 can be used for predicting students’ performance. In this paper C4.5 has been used to create decision tree.

1.1 C4.5 (J48) algorithm

This algorithm is a successor to ID3 developed by Quinlan Ross. The Hunt’s algorithm. C4.5 handles each categorical and continuous attributes to create a decision tree, so as to handle continuous attributes. C4.5 splits the attribute values into two partitions based on the chosen threshold. It also handles missing attribute values. C4.5 has the concept of Gain Ratio as an attribute selection measure to create a decision tree. At first, the gain ratio of every attribute is calculated. The root nodeis the attribute which has maximum gain ratio. C4.5 uses pessimistic pruning to get rid of unessential branches within the decision tree to enhance the accuracy of classification.

1.2 Fold cross-validation

When the data is less then fold cross validation can be used. In this the original sample is randomly partitioned into k subsamples. Of the k subsamples, a single subsample is taken as the validation data for testing the model, and the remaining k – 1 subsamples are used as training data. The cross-validation process is then repeated k times (in this 10 folds are used), with each of the k subsamples used exactly once as the validation data. The k results from the folds then can be averaged (or combined) to produce a single judgment.

V. METHODOLOGY

The methodology used to predict student’s performance is Knowledge Discovery in Databases (KDD). It is the process of searching for hidden knowledge in the large amount of data. The steps of KDD are shown in the following diagram.

As shown in the above figure there are different steps of KDD. 1. Information Gathering 2. Pre-processing 3. Data mining 4. Interpretation

5.1 Information Gathering

For predicting students’ performance the dataset collected from Shah and Anchor Kutchhi Polytechnic, Chembur, Mumbai. The data is collected using Jotform. The link of form created by Jotform was kept on Facebook page of college. The students entered the information in the form. There were many fields in the form like name of students, address, contact number, father’s education, mother’s education, SSC board, medium of SSC i.e. English or vernacular, admission type, previous semester result. The data collected from Jotform is converted into Excel sheet as shown below.

Table 1: Data Before Pre-Processing

5.2 Pre-processing

Selection Of Attributes: As mentioned above there are many
attributes. But the attributes such as phone number, registration number, name are not useful for prediction. So only selected attributes are taken and values are decided.

**Deciding Values for Attributes:** It is necessary to decide the values for the attribute because many attributes are having continuous data. Like SSC percentage can take any value from 35 to 100 with decimal numbers. Same will be the case with all other percentage in every semester. Due to continuous values the result may not be accurate. This can be improved by using discrete data. According to percentage values are decided. E.g. if percentage is between 35 to 49 then Pass Class. If between 50 to 59 then second class. If between 60 to 74 then First class and if more than 75 then Distinction. Same way all the other results are considered. In diploma, in semester result if student is failed in two subjects then it is considered to be ATKT. If more than two subjects then Fail. For Mother’s occupation only three values are considered such as Business, Service and Housewife. Like this all the attributes as shown below are considered with discrete values. The following table shows the attributes and the possible values.

<table>
<thead>
<tr>
<th>ATTRIBUTES</th>
<th>POSSIBLE VALUES</th>
</tr>
</thead>
<tbody>
<tr>
<td>FATHER’S EDUCATION (FEducation)</td>
<td>NONSSC, SSC, HSC, GRADUATION</td>
</tr>
<tr>
<td>FATHER’S OCCUPATION (FOccupation)</td>
<td>BUSINESS, SERVICE</td>
</tr>
<tr>
<td>MOTHER’S EDUCATION (MEducation)</td>
<td>NONSSC, SSC, HSC, GRADUATION</td>
</tr>
<tr>
<td>MOTHER’S OCCUPATION (MOccupation)</td>
<td>BUSINESS, SERVICE, HOUSEWIFE</td>
</tr>
<tr>
<td>CATEGORY</td>
<td>OPEN, SC, OBC, ST, MINORITY</td>
</tr>
<tr>
<td>SSCBOARD (SSCBoard)</td>
<td>STATE, ICSE, CBSE</td>
</tr>
<tr>
<td>ADMISSION TYPE (AdmissionType)</td>
<td>MINORITY, CAP, INSTITUTE</td>
</tr>
<tr>
<td>SSC MEDIUM (SSCMedium)</td>
<td>ENGLISH, VERNACULAR</td>
</tr>
<tr>
<td>SSC CLASS (SSCPercentage)</td>
<td>PASS, SECOND, FIRST, DISTINCTION</td>
</tr>
<tr>
<td>FIRST SEMESTER RESULT (FirstSemesterResult)</td>
<td>FAIL, ATKT, PASS, SECOND, FIRST, DISTINCTION</td>
</tr>
<tr>
<td>SECOND SEMESTER RESULT</td>
<td>FAIL, ATKT, PASS, SECOND, FIRST, DISTINCTION</td>
</tr>
<tr>
<td>THIRD SEMESTER RESULT (ThirdSemesterResult)</td>
<td>FAIL, ATKT, PASS, SECOND, FIRST, DISTINCTION</td>
</tr>
<tr>
<td>FOURTH SEMESTER RESULT (FourthSemesterResult)</td>
<td>FAIL, ATKT, PASS, SECOND, FIRST, DISTINCTION</td>
</tr>
<tr>
<td>FIFTH SEMESTER RESULT (FifthSemesterResult)</td>
<td>FAIL, ATKT, PASS, SECOND, FIRST, DISTINCTION</td>
</tr>
<tr>
<td>SIXTH SEMESTER RESULT (SixSemesterResult)</td>
<td>FAIL, ATKT, PASS, SECOND, FIRST, DISTINCTION</td>
</tr>
</tbody>
</table>

After Preprocessing the data will be as shown below.

### 5.3 Data Mining

Different data mining techniques can be used to predict students’ performance such as classification, clustering, association rule mining. In this paper the classification technique is used with decision tree. Using the above data, student.arff file was created. This file was loaded into WEKA explorer. The WEKA is a tool which contains a collection of visualization tools and algorithms for data analysis and predictive modeling. There are different techniques which can be used such as Classification, Clustering, and Association Rule Mining. Decision tree algorithms like ID3, J48, Simple
CART etc. implemented in WEKA. The algorithm used for classification is J48 (java implementation of C4.5 algorithm). The “Test options” has the 10-fold cross-validation which is selected as evaluation approach due to which a reasonable idea of accuracy can be maintained for the generated model. The model is generated in the form of graph which is known as decision tree.

5.4 Interpretation

The prediction of result can be done for all six semesters.

5.4.1 Prediction for First Semester

The result shown here is for first semester. Taking into consideration the SSC marks of the student, his admission type, category and family background the prediction of the student performance in semester one can be predicted.

The decision tree generated from student.arff is shown in Figure 2.

![Decision Tree for First Semester](image)

Figure 2. Decision Tree for First Semester

The most important attribute here is SSC marks. Also SSC medium, Admission type and mother’s occupation these attributes are taken into consideration. The rules are generated considering this, are as given below.

5.4.2 Rules generated for First Semester:

- If SSC result is Distinction and

  1. If SSC Medium is English

     Then Probable result of First semester is Distinction

  2. Else If SSC Medium is Vernacular

     Then Probable result of First semester is First Class

- Else If SSC result is First Class and

  1. If Admission Type is Minority

     Then Probable result of First semester is ATKT

  2. Else If Admission Type is Institute and

     1. If SSC Medium is English and

        i. If Mother’s Education is SSC

     Then Probable result of First semester is First Class

ii. Else If Mother’s Education is Graduation

     Then Probable result of First semester is First Class

iii. Else If Mother’s Education is HSC

     Then Probable result of First semester is Distinction

II. Else If SSC Medium is Vernacular

Then Probable result of First semester is First Class

3. Else If Admission Type is CAP

Then Probable result of First semester is First Class

- Else If SSC result is Second Class

Then Probable result of First semester is Fail

According to this rules if any student is from Vernacular medium and having Distinction in SSC then he may get First Class in First Semester diploma.

5.4.3 Confusion Matrix

The confusion matrix shown below indicates how the records are classified. It shows that 137 are classified correctly as Distinction. 18 are classified wrong as Distinction and 6 are classified wrong as First Class.

![Confusion Matrix](image)

Figure 3. Confusion Matrix

As shown above the prediction of First semester result is done. Likewise the prediction for all semesters can be done.

5.4.4 Prediction for Fifth Semester

The decision tree shown below is for Prediction of Fifth Semester result.

![Decision Tree for Fifth Semester](image)

Figure 4. Decision Tree for Fifth Semester
As can be seen from above decision tree, the root attribute is Fourth Semester result. If the Fourth Semester result is Distinction then the student may get Distinction in Fifth Semester. If he gets Second Class in Fourth Semester then his Mother’s Education is considered. If he gets First Class in Fourth Semester then Admission type is considered. According to this Decision Tree, if a student is having First Class in Fourth Semester and his Admission type is Direct Second Year (DSY) then he may Fail in Fifth Semester.

VI. CONCLUSION
The main objective of Educational Data Mining (EDM) is to improve teaching-learning process. Predicting students’ performance is one of the major applications of EDM. So using decision trees students’ performance can be predicted. The students, whose performance is poor, can be warned. The management can take necessary action to improve their performance by giving more attention, taking extra lectures etc. Due to such measures student performance can be improved. The number of failures can be reduced. Ultimately college results also get improved.

VII. FUTURE SCOPE
In this paper only traditional learning is considered but in future it can be integrated with the e-learning system. Here only improvement of students’ performance has been considered but EDM can be used to provide feedback to teachers so that teachers can make necessary changes in their teaching method. EDM also can be used to detect undesirable students’ behavior. Using EDM grouping of students according the same interest is possible due to which they can be given some work/projects according to their interest. Course designers and Universities can use the information for course designing, planning and scheduling.

REFERENCES