International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 4 Issue: 1 01 - 05

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 predicttheir performance in terms of pass or fail in final exam. The predicted results andactual results which indicates, that there was a significant improvement in results asthe 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 andtime 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. Itwill enable to identify the students in advance who are at risk.Giving warning to thestudents 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 beused for prediction. J48 decision tree algorithm is found to be the best suitablealgorithm for model construction. Cross validation method and percentage splitmethod were used to evaluate the efficiency of the different algorithms[6].

Charanjit Bambrah , 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 Saurabhsuggested C4.5, ID3 and CART decision tree algorithms can be used to predict the performanceof the first year engineering students .It was three class predictions. Students wereclassified as pass fail and

promoted. This model was good to identifying the students that are most likely to fail[9].

III. CLASSIFICATION

Classification issupervised 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.

3.1Algorithm:
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 unessentialbranches 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 singlejudgment.

V. METHODOLOGY

The methodology used to predict student's performance isKnowledge 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.

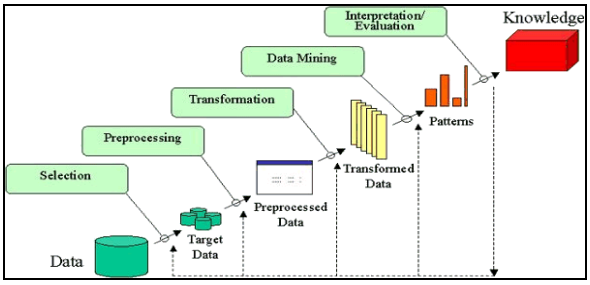

Figure 1. Steps of KDD Process[11]

As shown in the above figure there are different steps of KDD. 1. Information Gathering2.Pre-processing3.Data mining 4.Interpretation

5.1 Information Gathering

For predicting students'performance the datais collected from Shah and Anchor Kutchhi Polytechnic, Chembur,Mumbai. The data is collected using Jotform. The link ofform created by Jotform was kept on facebook page of college. Thestudents entered the information in the form. There were many fields in theform 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 convereted into Excel sheet as shown below.

Table 1: Data Before Pre-Processing

Father's Education :	Father's Occupation:	Mother's	Mother's Occupation:	Branch	Category:	Admission Type:	S.S.C.	S.S.C.	S.S.C.
Engineer	Service	BUMS	Housewife	CM3G	OPEN	Direct Second Year	English	72.18	State
SECOND	WORKER	EIGHTH	HOSE WIFE	CM3G	NT	CAP	Vernacular	81.27	State
BE	Service	Tenth	Housewife	CM3G	OPEN	CAP	English	84	State
DME	Supervisor	BA B ed	Teacher	IF3G	OPEN	CAP	English	61.45	State
BCOM	Government Servant	SSC	Housewife	IF3G	ST	CAP	English	73.2	State
HSC	worker	SSC	housewife	IF3G	OBC	Direct Second Year	Vernacular	78.6	State
TENTH	JOB	SEVENTH	HOUSE WIFE	IF3G	OBC	CAP	Vernacular	78.91	State
TENTH	JOB	SEVENTH	HOUSE WIFE	IF3G	OBC	CAP	Vernacular	78.91	State
graduate	service	graduate	teacher	IF3G	OPEN	Institute Level	English	76.8	State
HSC passed	Private Servant	SSC passed	Housewife	IF3G	NT	CAP	Vernacular	64.6	State
SSC	Business	VII	House Wife	IF3G	OPEN	CAP	English	67.8	State
B A	BUISNESS	H S C	HOUSEWIFE	IF3G	SC	Direct Second Year	English	65	State
SSC	BUSINESS	EIGHTTH	HOUSE WIFE	IF3G	SC	Direct Second Year	Vernacular	84.62	State
BA	service	MA	na	CM6G	OPEN	Institute Level	English	75	State
ssc	business	twelve	house wife	IF4G	OPEN	Minority	Vernacular	62	State
Mcom	Government Servant	HSC	Housewife	IF1G	OPEN	CAP	Vernacular	70.73	State
hsc	goverment service	housewife	ssc	IF2G	OPEN	Institute Level	English	67	State
SSC	worker in RCF	HSC	housewife	IF1G	SC	CAP	English	59	State

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 takenand values arse decided.

Deciding Values for Attributes: It is necessary to decide the valuesfor the attribute because many attributes are having continuous data. Like SSCpercentage can take any valuefrom 35 to 100 with decimal numbers. Same will be the case with all other percentage in every semester. Due tocontinuous values the result may not be accurate. This can be improved byusing discrete data. According to percentage values are decided. E.g. ifpercentage is between 35 to 49 then Pass Class. If between 50 to 59 thensecond class. If between 60 to 74 then First class and if more than 75 thenDistinction. Same way all the other results are considered. In diploma, insemester result if student is failed in two subjects then it is considered tobe ATKT. If more than two subjects then Fail. For Mother's occupationonly 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 2: Details Of Attributes

ATTRIBUTES	POSSIBLE VALUES
FATHER'S EDUCATION(FEducation)	NONSSC,SSC, HSC,GRADUATION
FATHER'S OCCUPATION(Foccupation)	BUSINESS,SERVICE
MOTHER'S EDUCATION(MEducation)	NONSSC,SSC,HSC,GRADUATION
MOTHER'S OCCUPATION(Mocupation)	BUSINESS,SERVICE, HOUSEWIFE
CATEGORY	OPEN,SC,OBC,ST, MINORITY
SSCBOARD (SSCBoard)	STATE,ICSE,CBSE
ADMISSION TYPE (AdmissionType)	MINORITY,CAP,INSTITUTE
SSC MEDIUM (SSCMedium)	ENGLISH,VERNACULAR
SSC CLASS(SSCPercentage)	PASS,SECOND,FIRST, DISTINCTION
FIRST SEMESTER RESULT (FirstSemesterResult)	FAIL,ATKT,PASS,SECOND,FIRST,DISTINCTION
SECOND SEMESTER RESULT (SecondSemesterResult)	FAIL,ATKT,PASS,SECOND,FIRST,DISTINCTION
THIRD SEMESTER RESULT (ThirdSemesterResult)	FAIL,ATKT,PASS,SECOND,FIRST,DISTINCTION
FOURTH SEMESTER RESULT (FourthSemesterResult)	FAIL,ATKT,PASS,SECOND,FIRST,DISTINCTION
FIFTH SEMESTER RESULT (FifthSemesterResult)	FAIL,ATKT,PASS,SECOND,FIRST,DISTINCTION
SIXTH SEMESTER RESULT (SixSemesterResult)	FAIL,ATKT,PASS,SECOND,FIRST,DISTINCTION

After Preprocessing the data will be as shown below.

Table 3:Preprocessed Data

FEducatio	Foccupatic	Meducatic	Moccupati	Category	Admission	SSCMediu	SSCPercen	SSCBoard	FirstSemesterResult
GRADUATI	SERVICE	HSC	HOUSEWI	OPEN	CAP	English	DISTINCTI	State	FIRST
SSC	BUSINESS	GRADUATI	HOUSEWI	OPEN	CAP	English	FIRST	ICSE	FIRST
HSC	BUSINESS	SSC	SERVICE	MINORITY	MINORITY	English	FIRST	State	DISTINCTION
SSC	SERVICE	SSC	HOUSEWI	OPEN	CAP	Vernacula	DISTINCTI	State	ATKT
SSC	BUSINESS	SSC	HOUSEWI	OPEN	CAP	English	SECOND	State	FAIL
GRADUATI	BUSINESS	SSC	HOUSEWI	MINORITY	MINORITY	English	SECOND	State	ATKT
GRADUATI	BUSINESS	GRADUATI	HOUSEWI	OBC	CAP	English	FIRST	State	ATKT
GRADUATI	BUSINESS	SSC	HOUSEWI	OPEN	CAP	Vernacula	DISTINCTI	State	SECOND
SSC	SERVICE	SSC	HOUSEWI	OPEN	CAP	English	FIRST	State	FIRST
HSC	SERVICE	HSC	SERVICE	OPEN	INSTITUTE	Vernacula	DISTINCTI	State	SECOND
HSC	BUSINESS	HSC	HOUSEWI	OPEN	INSTITUTE	English	FIRST	State	FIRST
SSC	SERVICE	SSC	HOUSEWI	MINORITY	MINORITY	English	DISTINCTI	State	DISTINCTION
SSC	BUSINESS	SSC	HOUSEWI	OPEN	CAP	Vernacula	FIRST	State	ATKT
SSC	BUSINESS	SSC	HOUSEWI	OPEN	CAP	Vernacula	FIRST	State	ATKT
HSC	BUSINESS	HSC	HOUSEWI	OPEN	INSTITUTE	English	FIRST	State	DISTINCTION
GRADUATI	BUSINESS	HSC	HOUSEWI	OPEN	CAP	English	DISTINCTI	ICSE	DISTINCTION
GRADUATI	SERVICE	SSC	HOUSEWI	OBC	CAP	English	FIRST	State	FIRST
SSC	BUSINESS	GRADUATI	SERVICE	MINORITY	MINORITY	English	FIRST	State	SECOND
HSC	BUSINESS	SSC	HOUSEWI	OPEN	CAP	English	FIRST	State	ATKT

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.

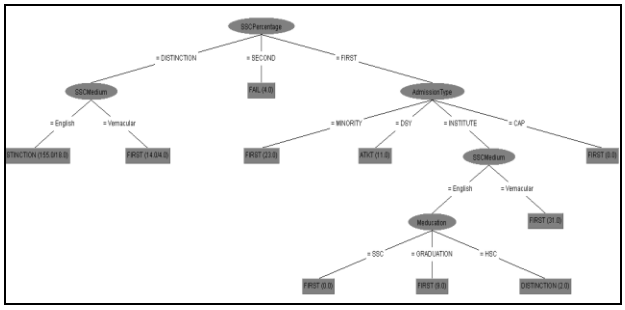

Figure 2. Decision Tree for First Semester

The most important attribute here is SSC marks. Also SSC medium, Admissiontype 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

     I. 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.

5.4.3Confusion Matrix

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

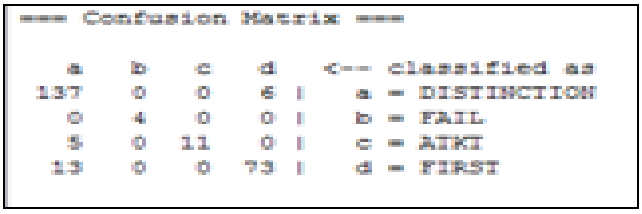

Figure 3. Confusion Matrix

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.

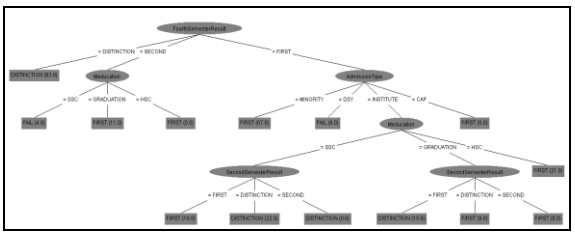

Figure 4.Decision Tree for Fifth Semester

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 treestudents' performance can be predicted. The students, whose performancse 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 beintegrated 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

[1] Crist´Obal Romero And Sebastian Ventura (2010), "Educational Data Mining: A Review Of The State Of The Art",IEEE transactions on systems, man, and cybernetics—part c: applications and reviews, vol. 40, no. 6, November 2010

[2] Kumar S. Anupama and Dr. Vijayalakshmi M.N. (2011). Efficiency of Decision Trees in Predicting Students Academic Performance. Computer Science & Information Technology 02, pp. 335–343.

[3] R. R. Kabra And R. S. Bichkar(2011),"Performance Prediction Of Engineering Students Using Decision Trees", International Journal Of Computer Applications (0975 – 8887) Volume 36– No.11, December 2011

[4] Charanjit Bambrah , Minakshi Bhandari , Nirali Maniar , Prof. Vandana Munde "Mining Association Rules in Student Assessment Data" International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 3, March 2014 Copyright to IJARCCE www.ijarcce.com 5340

[5] D. Magdalene Delighta Angeline "Association Rule Generation for Student Performance Analysis using Apriori Algorithm", The SIJ Transactions on Computer Science Engineering & its Applications (CSEA), Vol. 1, No. 1, March-April 2013

[6] Mrinal Pandey And Vivek Kumar Sharma(2013) , "A Decision Tree Algorithm Pertaining To The Student Performance Analysis and Prediction", International Journal Of Computer Applications (0975 – 8887) Volume 61– No.13, January 2013

[7] Baradwaj Brijesh Kumar and Pal Saurabh (2011). Mining Educational Data to Analyze Student Performance.

[8] International Journal of Advanced Computer Science and

[9] Applications, Vol. 2, No. 6.

[10] A.Dinesh Kumar ,Dr.V.Radhika "A Survey on Predicting Student Performance", A.Dinesh Kumar et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (5) , 2014, 6147-6149

[11] Yadav Surjeet Kumar, Pal Saurabh " Data Mining: A Prediction for PerformanceImprovement of Engineering Students using Classification", World ofComputer Science and Information Technology Journal (WCSIT) .ISSN: 2221-0741Vol. 2, No. 2, 51-56, 2012

[12] http://www.educationaldatamining.org

[13] http://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/1_kdd.html

[14] Clementine_manual.pdf

[15] WEKA_user_manual.pdf