Prediction Accuracy on Academic Performance of Students Using Different Data Mining Algorithms with Influencing Factors

Ms.A.Pavithra, Mr.S.Dhanaraj

Department of Computer Science, Sree Saraswathi Thyagaraja College, Pollachi, Tamil Nadu, India

Abstract - Education is the acceptance that education improves the potential of each person. Each person is matchless and elevates the individual academically, substantially and passionately. It is a noteworthy for personal improvement and to be realization in life. Educational Data Mining (EDM) is an upcoming research field on the development of educational psychology and the learning sciences. For an institutional development academic performance records plays a major role for the good will. Though there are many factors like socio-economic, non-academic, some academic things influencing the student performance in an institution. Data mining techniques is the concept of bring out the hidden knowledge from the huge raw datasets. Finally, it brings a Knowledge discovery and Decision making. There are many classification and prediction algorithms in data mining. In this paper it is focused on student performance and prediction of an attribute that, where the student is place able or not. The datasets are original data of a private institution named “Sree Saraswathy Thyagaraja College”, Pollachi which is located in rural area. The research work is done with five different algorithms like Naïve Bayes, Rep tree, J48, Decision tree and Multi-Layer Perception (MLP) along with the verification of student destination survey. This work also shows the different factors which influences student performance.

Keywords— Educational Data Mining (EDM), Multi-Layer Perception (MLP), Classification, Prediction, Knowledge Discovery, Decision Making

I. INTRODUCTION

Data mining is the concept of finding the patterns from huge raw datasets. The data mining provides many algorithms like classification algorithm, association algorithms, and rule based algorithms so on. The datasets are get processed with the help of all these algorithms and gives us a meaningful mining results. That result is fine tuned for some decision making. Thus data mining is fully concentrates on “Knowledge Discovery and Decision Making” (KDD). Educational Data Mining is based on education data which is to be processed for bringing up a meaningful results that is decision making. EDM plays a big role in student performance prediction and analysis. It is an upcoming research area in education. WEKA tool is used in this research work for better understanding. It is very user friendly and provides many visualization effects on datasets.

There are many prediction algorithms in data mining. In this research work it is concentrated on student performance prediction with different algorithms. Totally 127 students data of Bachelor of Computer Science is taken from the college and they are divided into five one third of training datasets and remaining into five different parts of test datasets. All are formatted to WEKA tool accepted ARFF format. Different fields of datasets are student gender, place, father occupation, annual income, caste and finally with their higher secondary marks. They are analyzed deeply and formatted to user need type. First and foremost the dataset is analyzed to which attribute plays a major role and obtained by Info gain attribute selection. Then the dataset is run with five different algorithms like Naïve Bayes, Rep tree, J48, Decision tree and Multi-Layer Perception (MLP) with five different datasets. The average ROC value is calculated. At last with higher average ROC value algorithm the final student place able or not attribute is predicted. This is checked with student destination survey datasets. Finally, the knowledge flow is drawn and all data are visualized with visualize errors.

I.1 About Student Dataset

The dataset consists of 127 students details and they are belong to rural area institution. On working with this dataset it has been know that their parents are mostly working as an agriculturist, belongs to annual income of 50,000 and below. The students need to concentrate little better in their studies according to the observation. The below table explains the different attributes of the dataset along their description and the different values it have.

I.2. Proposed System Focuses On

The proposed system focuses on student performance prediction with different algorithms along with different datasets. There may be many algorithms used on the prediction on academic performance but here we have selected five important and mostly used prediction algorithm and run in student datasets. In most works they may use the test or prepared datasets but here we have the original dataset of an institution which is placed in rural area. A Review on Predicting Student's Performance Using Data Mining
Techniques [1] explains on student performance prediction along with the factors which influence student performance.

### Table 1.1 Student Dataset

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEX</td>
<td>Gender of the student</td>
<td>{M, F}</td>
</tr>
<tr>
<td>PLACE</td>
<td>Place where the students are from divided into two values</td>
<td>{VILLAGE, TOWN}</td>
</tr>
<tr>
<td>OCCUPATION</td>
<td>Occupation of the student father</td>
<td>AGRICULTURE, BUSINESS, OTHERS</td>
</tr>
</tbody>
</table>
| INCOME   | Annual income of the student | A: ABOVE 50000  
B: 50000  
C: BELOW 50000 |
| CATEGORY | Different category of the student | OC: RC, OC UNRESERVED: MBC, BCM SC: SC  
ST: ST |
| HSLC     | Student higher secondary mark | A: ABOVE 1000  
B: 800 - 1000  
C: 700 - 800  
D: 600 - 700  
E: BELOW 600 |
| SEM      | Semester percentage | A: ABOVE 70%  
B: 50% – 69%  
C: 49% - 40%  
D: FAILURE |

There are many factors which influence the student performance and day by day there are new upcoming scenarios which affects the performance. So, it will be not stick on to any single factors which affect the performance of students. The proposed system mainly focuses on student prediction in different algorithms and here we have compared the different ROC values and taken as an average value. The algorithm which has highest average ROC value is taken for the final prediction of the student dataset attribute namely, place able or not place able.

On analysis of this student dataset it has been find out that maximum students are from rural areas, they have only a minimum income of 50000 and below. Most of the students family are belongs to agricultural background and their performance on higher secondary are little poor. And also it is noted that on reviewing many papers it has been cleared that many factors are influencing student performance and we cannot stick into any single factors. This research work shows that semester five performance influence the student performance.

II. FRAMEWORK OF PROPOSED SYSTEM

The 127 students dataset is taken from the database and cleansed that is, preprocessed with Infogain in weka explorer interface. In this step it has been verified that the attribute semester has major importance when compared to other attributes. But in prediction on student performance we cannot stick on to any single factors which influence the student performance, there may be many time being and students places decide the factors which affects their performance.

In article [6] explains the prediction of student performance in a different manner when compared to others. They have explained the different factors which affect student performance. One is parameter which affects student performance, secondly data mining algorithms which we used to predict and finally, data mining tool. The methodology used in this research work is explained by a hybrid algorithm. [1]

Here in this research work it is worked on making the dataset to be divided into five one third of dataset. From that one third individual dataset it has been divided to training and test dataset. So, we will have five different parts of training and test datasets that is runned on five different algorithms that is explained below.

1.1.1 Correctly Classified Accuracy

It shows the correctness ratio of investigation that is correctly classified.

1.1.2 Incorrectly Classified Accuracy

It specifies the accuracy percentage of test that is incorrectly classified.

1.1.3 Mean Absolute Error

It prove the number of errors to study algorithm classification accuracy.

1.1.4 Time

It illustrates how much time necessary to build a model.

1.1.5 ROC Area

Receiver Operating Characteristic show test performance director for classifications accuracy.

Here in this research work ROC plays a major role.

III. LITERATURE REVIEW

The proposed system focuses on student prediction performance with the help of their higher secondary marks and with their semester wise improvements. Here the literature review is made on different student mining papers that gives us more information regarding students influencing factors. The influencing factors may vary to the students’ locality and it mainly affects their psychology. The student prediction is done by many different data mining algorithms and it shows different us different results.

There may be many algorithms used on the prediction on academic performance but here we have selected five important and mostly used prediction algorithm and run in student datasets. In most works they may use the test or prepared datasets but here we have the original dataset of an institution which is placed in rural area.

---

IJSRCAMS

Volume 7, Issue 5 (September 2018)  www.ijsrcsams.com
IV. ALGORITHM USED IN PROPOSED SYSTEM

4.1 Algorithms

In the proposed system it is concentrated mainly on five different prediction algorithms. Prediction on student dataset is more cautious because of the influencing factors. There are many factors influence the student performance but here it has been identified at the end last semester will plays a major role on the student place able or not prediction. Herewith we will discuss the five algorithm which is used in this research work.

A) Naïve Bayes

Naïve Bayes is one of the algorithms that works as a probabilistic classifier of all abilities enclosed in data model discretely and then groups data problems. Running the algorithms Naïve Bayes we inspect the classifier output with some statistics set up output by using 10 cross validation to make an estimate of all instance of the dataset. In most of the research article NB is mainly used in student prediction.

B) J48

J48 classifier is a forthright C4.5 decision tree for classification, which makes a binary tree. It is most useful decision tree method for classification difficulties. This procedure builds a tree to model the classification process. After the tree is built, the algorithm is applied to each tuple in the database and results in classification for that tuple.

Algorithm J48

INPUT: 
P //Training data
OUTPUT
DT //Decision tree
DTBUILD (*P)
{
DT=φ;
DT= Create root node and label with splitting attribute;
DT= Add arc to root node for each split predicate and label;
For each arc do
P= Database created by applying splitting predicate to P;
If stopping point reached for this path, then
DT'= create leaf node and label with appropriate class;
Else
DT'= DTBUILD(P);
DT= add DT' to arc;
}

C) Decision Tree

Decision tree is like a tree or graph model that is used for making some decisions and it is also a decision support tool. The decisions are can be made on any possible consequences, including chance or event outcomes, resource costs and utility. This becomes mainly used to display the conditional control statements that contains. Here it is used to make a decisions on classifying the data given by graph or tree model.

D) Multilayer Perception

- A feedback neural network model
- Extensively used in pattern classification
- Essential issue: training/ learning algorithm

MLP will be worked on neural networks that is by the use of different layers in the algorithm. It is mainly used in the research work for the classification of patterns on student dataset.

E) REP Tree

Given a set S of positive and negative examples of some target concept (a 2-class problem), the entropy of set S relative to this binary classification is

\[ E(S) = - p(P) \log_2 p(P) - p(N) \log_2 p(N) \]

To build a decision tree, we need to calculate two types of entropy using frequency tables as follows:

a) Entropy using the frequency table of one attribute:

\[ E(S) = \sum_{i=1}^{c} - P_i \log_2 P_i \]

b) Entropy using the frequency table of two attributes:

\[ E(T, X) = \sum_{c \in T} p(c)E(c) \]

REP Tree has been used in this paper to form a decision and decreases errors by arranged values of numeric attribute and splits the instances into bits to classify the accuracy. Mostly helped in student dataset.

V. DIFFERENT RUNS

Run 1: In run 1 the first one third dataset is taken and divided into training and test dataset. Then it is load in weka data folder and gets preprocessed and run with five algorithms. The ROC value of the each run will get noted for future analysis. Here we have predicted the semester 1 percentage category. Eg. ‘A’ denotes ‘above 70%’

<table>
<thead>
<tr>
<th>RUN 1</th>
<th>NB</th>
<th>J48</th>
<th>DECISION TREE</th>
<th>MLP</th>
<th>REP TREE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROC VALUE</td>
<td>0.862</td>
<td>0.838</td>
<td>0.827</td>
<td>0.813</td>
<td>0.801</td>
</tr>
</tbody>
</table>

Run 2: In run 2 also another one third dataset will get preprocessed and run with algorithms. Here we have used the run 1 predicted dataset along with one third dataset.
VI. CALCULATING THE AVERAGE ROC VALUE

The research work has been done still on running the dataset with different algorithms. Now it need to calculate the average value of the each algorithm that has been explained above. So, the method is on calculating the average value of each algorithm in different runs brings us an knowledge of which algorithm has worked well on prediction. Below in a table the average value is mentioned.

<table>
<thead>
<tr>
<th>ALGORITHM</th>
<th>AVERAGE ROC VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB</td>
<td>0.7206*</td>
</tr>
<tr>
<td>J48</td>
<td>0.706</td>
</tr>
<tr>
<td>DECISION</td>
<td>0.6946</td>
</tr>
<tr>
<td>MLP</td>
<td>0.6892</td>
</tr>
<tr>
<td>REP</td>
<td>0.5838</td>
</tr>
</tbody>
</table>

In the above table the average value is calculated and it has been showed that the Naïve Bayes algorithm has highest average ROC value and the research work will be carry on next with Naïve Bayes algorithm.

VII. PREDICTION ON STUDENT PERFORMANCE

The final prediction will be carry out with Naïve Bayes algorithm because the average value is higher in it when compared to other algorithms. The next step prediction is whether the student is place able or not, which is predicted by 10 cross validation in naïve bayes. the below flow chart explains the overall flow of the dataset in this research analysis.

After the final prediction of the student performance with the help of the tool we can have a knowledge flow diagram which helps one who needs the flow. It easily explains the different stages we have do so.

Firstly the arff loader is used that is to load the dataset in arff format and in the next stage the class assigner is used to classify the dataset. The cross validation is used next to class assigner, here the data loaded is validate to number of times we have given. Default value in cross validation is 10. The data will get validate for 10 number times to give best results.

This is the final result of the ROC value of different algorithms. With the above detail we can easily find the average value.
VIII. KNOWLEDGE FLOW

The WEKA provides us the knowledge flow of the datasets. There we have different tools for different process. We just click and drag to the working area. Here the have different icons on different usages. The arff icon which shows that we have included the data in arff format. Also this makes the user to understand it very clearly.

IX. RESULTS AND DISCUSSIONS

8.1 Advantages
- Different runs makes the data set preprocessed
- Easily can get the future performance
- Makes the institution to take more care on poor students
- Makes us to know the different category of students
- Easy ABCD category analysis of different students

X. CONCLUSION AND FUTURE WORK

10.1 Conclusion

Educational Data Mining (EDM) is an upcoming field in research development. This research work is primarily dedicated on examining the prediction accurateness of the academic performance of the students using different classification algorithms like Naïve Bayes, MLP, Decision tree, J48 and REP tree. The final prediction is made on the best average value algorithm. Secondly, it helps the institution to find out the students’ performance in future and to know the slow learners who needs extra care on them in academic. It is also concluded that many factors will influence the student performance and it may differ to different locality of students.

10.2 Future Work

The future work of this research is to extend on finding the best algorithms on different runs and also to produce a hybrid algorithm for its accuracy on prediction. Next step of this research work will be go on with the different influencing factors in detail because the locality of the students will affect their mindsets and development towards academic. Also, may include the psychology of the student.

ACKNOWLEDGEMENTS

This research paper is truthfully involvement and guidance who belong to the list of authors. The authors pay morality to Sree Saraswathy Thyagaraja College, Pollachi for continuous encouragement, guidance and support during their work.

REFERENCES