International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 3 | ISSN : 2456-3307

Analyzing and Predicting Learning Levels of Students in Higher Education using Machine Learning Approach

Qamar Rayees Khan[1], Parvez Abdulla[2], Majid Bashir Malik[3]
[1]Department of Computer Sciences, Baba Ghulam Shah Badshah University, Rajouri (J&K), India
[2]Department of Management Studies, Baba Ghulam Shah Badshah University, Rajouri (J&K), India
[3]Department of Computer Sciences, Baba Ghulam Shah Badshah University, Rajouri, (J&K), India

ABSTRACT

The genesis of the emerging field that lead to the growth of the analytical observations of the educational data and draw inferences based on the type and pattern of the data is Education Data Mining (EDM). This field has added the power of the decision making in education settings. The role of EDM solves many problems facing the educational institutions by generating the patterns from the data which affect the overall objective of the educational institutions. Various data mining techniques have already been used by the researchers to evaluate the impact of drop out ratio of the institutions using EDM. This paper shall explore the current field of study and identify the parameters that affect the Learning Levels of Students in Higher Education using a Machine Learning Approach. This paper emphasis on the prediction of learning levels of the students so that the institution may evolve a mechanism to bridge the gap for slow learners to perform as per their expectations. The dataset used in this paper is collected from the university students and the weka tool which is an open source tool is used for the experimental analysis. At the end, the model is evaluated using various performance evaluation parameters.

Keywords : Education Data Mining (EDM), Decision making, data mining techniques, Learning levels, Higher education, Machine Learning, slow learners, weka tool.

I. INTRODUCTION

The education data mining (EDM) has evolved a field in itself where the pedagogical data pertaining to the students is being thoroughly analyzed. The probable inferences are drawn so as to explore the possibilities of tapping the weakness of the students in terms of their performance in their chosen domain. The main problem which has been visualized from so many years in the educational institutions is the students learning levels. The learning levels of student comprise of two main categories: the one who are quick and respond and perform better, they are called Fast Learners and the next is the slow learners who perform lesser than their expected performance. The slow learning is not an incurable problem, but it is the ones ability to understand and grasp the concept taking into consideration the various parameters. Learning is a continuous process that evolves a proper mechanism that tunes the parameters of the students and guides them to achieve a particular objective. The compromise of any of the parameter may lead to have an impact on student's performance. The learning level of a student is being analyzed based on the various parameters. As we know, not all the students perform exceptionally

better in a particular domain. There are the instances where some students outperform others in academics. The students who do not perform better than expected always feel inferior and continue to face the downfall.

The techniques used in the Education Data Mining (EDM) have been a breakthrough and various researchers made their contribution in this direction. The ultimate aim is to identify those students in the education setup that does not perform as per the expectations. Data Mining techniques like clustering, Decision trees, Naive Bayes, etc have been used to critically identify the slow learners so that the institutions may press the alarm button well before their semester examinations and arrange the remedial classes for them so that institutional objectives shall not get compromised.

II. Related Work

A number of researchers have contributed to this field of the educational Data Mining for identifying the learning levels of the students in school/college/university level. They have used various techniques to predict the learning levels and devise the appropriate measures/mechanisms to tackle the problem of the dropout ratio that emerge as one of the biggest challenge in this field. The contributions of few of the researchers are listed below for reference:

The work carried out by the researchers has used Data Mining Techniques to evaluate the performance of students in their semester examination so that the students who perform low shall be handled so that their performance may improve. The classification mechanism used in this model is the Decision Tree Algorithm that evaluates the performance of the students at the end of the semester examination so that dropout ratio may be minimized and the appropriate counseling/remedial classes for those students be managed [5]. This study helps both the students and teachers to identify the learning level of the students so that the special attention may be given to those students in their upcoming semesters.

The researchers in the paper have designed an experimental methodology wherein they collect the primary as well as the secondary data from various sources of the education setup. The researchers have collected 1000 instances from the various sources and after necessary preprocessing of the data, they feed the same in the CHAID predication Model and the results so generated were compared with the other models [6]. The Model shows a good accuracy rate. The CHAID model was constructed with seven predictor variables for efficiently predicting the academic performances of the students.

Parneet K. al. [7] in their paper studied and proposed a Prediction Model for the identification of slow learners and tries to evaluate the performance of the students by using the selected list of the variables for the same. The data from high school was collected and nearly 152 instances of the students were trained in the prediction model. Various Classification algorithms like Multilayer Perceptron, SMO, Naïve Bayes, REPTree and J48 were used by the researchers to indentify the learning levels of the students. Out of the five algorithms, Multilayer Perceptron performs better with the accuracy rate of 75% and F-Measure of 82%.

Another researcher in their paper has classified the students based on their demographic features like gender, origin etc and average of attending the particular course [8]. They used K-Means Clustering so as to extract the hidden patterns from the data and divide the students based on four clusters that individually comprises of six variables. They used SPSS to analyze and evaluate the data. In this paper, they used 306 instances of the students. The results were comparatively optimal.

V.Rameshet. al. [9] in their research paper proposed a Model based on selected variables which were collected from the questionnaire after proper consultation with the experts. The variables were

that seem to be influential were used to predict the final grades of the students. The study revealed that type of the institution does not contribute to the performance of the students whereas the occupation of the parents plays a role in their academic performance. Various classifiers were used to predict the grades, but Multilayer Perceptron performs better with the accuracy rate of 72.38%.

Q. Rayees et. al. [10] in their paper proposed the Machine Learning framework to identify the learning levels of the students in higher Education. The researcher have given a proper methodology to solve the problem.

III. PROPOSED METHODOLOGY

The prediction model based on Machine Learning Model is proposed so as to identify the learning levels/ability of the students enrolled in higher education. The main aim focus of this research is to identify slow learners that are not performing as per their expected level. This model shall be able to predict the learning outcomes of the student beforehand so that the institutions may evolve a mechanism to fill their deficiencies in terms of their performance. The Prediction Model comprises of the various phases.

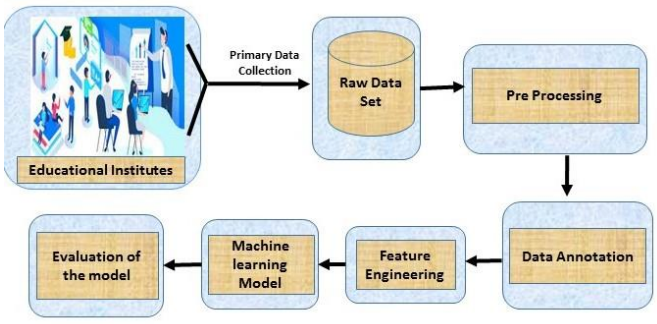

Figure. 1 : Proposed Machine Learning Model

1. Data Collection Phase: Data of computer science discipline is collected from one of the higher education institutions by using the questionnaires. The parameters/ attributes were decided after a threadbare discussion with the university academicians. Various deciding/ important parameters were considered that may influence the learning levels of the students.
2. Pre-Processing Phase: During this phase, discrepancies in the data were removed by cleansing the data so that there shall be no possibilities of null values, outliers and missing values in the dataset.
3. Data Annotations: During this phase, data is annotated by using the appropriate algorithm and class labels are decided. The Academicians were taken into consideration while framing the annotated scheme and the data was annotated as per the approved scheme. The data annotation is an important phase in this model and the output of the same which is the refined data is sending for further necessary step ie: feature Engineering phase.
4. Feature Engineering Phase: The refined data that is complete in all aspects is analyzed and appropriate algorithms are used to select the attributes that contribute to the model and eliminate the other attributes whose contribution is zero or minimum. By doing the feature engineering, we shall be able to reduce the dimensions of the data set by including those attributes who contribute to the model thereby reducing the time for training the model as in phase 5 below.
5. Machine Learning Techniques: The data after proper feature selection and annotation is used by the model for training. Various classifiers are used to classify the data so that the prediction of learning levels is made. The model is then tested by using the testing data. A comparative analysis is to be made for the different machine learning algorithms for better prediction.

IV. EXPERIMENT, RESULTS AND DISCUSSION

The Department of computer science was chosen to identify the learning levels of the students enrolled in computer science course. Various parameters/attributes were taken based on their influence on student's performance and after the discussion with the academic experts. Fourteen (14) attributes are taken for this study and questionnaire was designed and the required data was gathered from the students. The computer science as a discipline comprises of theory courses and lab courses. The lab courses include the programming languages that need a logical/ mathematical/ reasoning background of the student. The data was collected from four batches of the semester first students of PG course offered by the department. A total of 150 instances were recorded in the dataset. The data after necessary preprocessing by eliminating all the discrepancies are annotated so that the appropriate class labels are decided. A requisite feature engineering technique is used to include the attributes that contribute to the model. We have used Information Gain Algorithm to select most contributing parameters as in Figure-2 below.

```
Ranked attributes:
  1 P_QE
  2 P_Interest
  5 Internet connectivity
  8 Aptitude level

Selected attributes: 1,2,5,8
```

Figure-2 Contributing Attributes

It emerged that top four attributes contribute much while identifying the learning levels of the students. After the feature engineering phase, the data is then inputted into the machine learning classifiers. In this model we have used 04 classifiers to predict and identify the learning levels of the students.These are J48, SMO, Multinominial Naive Bayes, Random Forest. The performance of each algorithm is shown in the Table1.

Table-1 (Performance of various Algorithms)

Classifier	Accuracy	Precision	Recall	F1 Score
J48	82.31	81.25	79.50	80.52
Multinominal Naive Bayes	80.50	79.82	78.20	77.52
SMO	89.81	88.15	86.25	85.20
Random Forest	86.52	85.25	81.50	80.35

The corresponding Bar graph for various Algorithms for quick analysis is shown below in Figure 3.

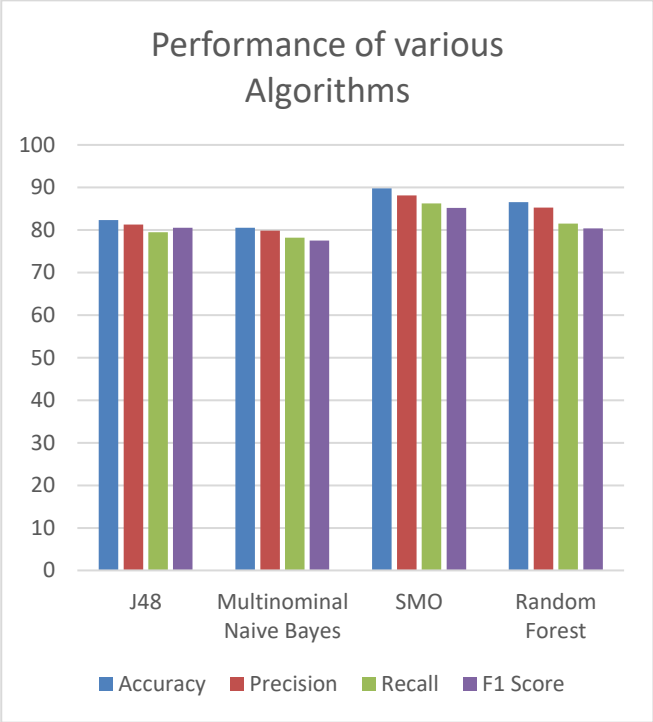

Figure-3. Quick Analysis of Algorithms through Bar Graph

The results clearly signifies that all the Algorithms used in this model perform good but the SMO performs better with an accuracy rate of 89.81 % which shows that we can tap the said problem by identifying the slow learners much better by using SMO technique.

V. RESEARCH OUTCOME

The outcome of this proposed Machine Learning model is to identify the slow learners so that the management in these institutions may evolve a proper mechanism to bridge their learning gap for better performance in the upcoming examinations of a chosen course.

VI. CONCLUSION

In this paper, various algorithms of Machine Learning were used for identifying learning capabilities/ levels of the students in computer science. The main aim focus of this research was to identify slow learners that are not performing as per their expected level. This model shall be able to predict the learning outcomes of the student beforehand so that the institutions may evolve a mechanism to fill their deficiencies in terms of their performance.

VII. REFERENCES

1. Suhas G. Kulkarni, Ganesh C. Rampure, Bhagwat Yadav, —Understanding Educational Data Mining (EDM), International Journal of Electronics and Computer Science Engineering, 2013.
2. Siti Khadijah Mohamad and ZaidatunTasir, "Educational data mining: A review", Procedia Social and Behavioral Sciences, vol. 97, pp. 320-324, November 2013.
3. Cristobal Romero and Sebastian Ventura, "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, pp. 601-618, November 2010.
4. K. Uma maheswari and S. Niraimathi, "A Study on Student Data Analysis Using Data Mining Techniques", International Journal of Advanced Research in Computer Science and Software Engineering, vol. 3, no. 8, August 2013.
5. BK Bhardwaj, S. Pal, " Mining Educational Data to Analyze Students Performance", (IJACSA), Vol. 2, No. 6, 2011
6. M. Ramaswami and R. Bhaskaran (2010), "A CHAID Based Performance Prediction Model in Educational Data Mining", International Journal of Computer Science Issues Vol. 7, Issue 1, pp 10-18.
7. Parneet Kaura, Manpreet Singh ,Gurpreet Singh Josanc "Classification and Prediction based DataMining Algorithms to Predict Slow Learners in Education Sector" Science Direct Procedia Computer Science 57 (2015) 500 – 508 2015 (ICRTC- 2015).
8. Harwatia, Ardita Permata Alfiania, Febriana AyuWulandari," Mapping Student's Performance Based on Data Mining Approach", Science Direct Agriculture and Agricultural Science Procedia3(2015) 173 – 177.
9. V.Ramesh (2013), "Predicting Student Performance: A Statistical and Data Mining Approach", International Journal of Computer Applications (0975 – 8887) Volume 63– No.8.
10. Q. Rayees Khan (2017), "A Machine Learning Framework for Identifying Learning Levels of Students in Higher Education", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Issues Vol. 3, Issue 1, pp 582-584.