International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 08 | Aug 2020 www.irjet.net p-ISSN: 2395-0072

PREDICTION OF STUDENT PERFORMANCE USING RANDOM FOREST CLASSIFICATION TECHNIQUE

P.Ajay[1], M.Pranati[2], M.Ajay[3], P.Reena[4], T. BalaKrishna[5]

[1,2,3,4]UG Scholar, Dept. of Computer Science Engineering, Gudlavalleru Engineering College, Andhra Pradesh, India
[5]Assistant Professor, Dept. of Computer Science Engineering, Gudlavalleru Engineering College, Andhra Pradesh, India

---

Abstract - *Abstract: Data mining is the process of analyzing data from different perspectives and summarizing it into important information so as to identify hidden patterns from a large data set. Educational data processing (EDM) is an emerging discipline that is concerned with data from different academic fields to develop various methods and to spot unique patterns which can help for exploring student's academic performance. EDM are often considered as learning science, also as an feature of knowledge mining. Assessing students learning process may be a very complex issue .Education data processing helps in predicting students' performance so as to recommend improvements in academics. The past several decades have witnessed a rapid climb within the use of knowledge and knowledge mining as a tool by which academic institutions extract useful unknown information in the student result repositories in order to improve students' learning processes .The main objective of this project is prediction of student's performance based on random forest classification technique using tools such as WEKA , ORANGE and scikit-learn libraries in python.*

Key Words: Data Mining, EDM, Classifiers, WEKA, Random Forest, Decision Tree etc.

1. INTRODUCTION

The student's performance prediction is an important part in education system. In recent years due to the rapid development of technology the amount of data has been growing tremendously in all areas. The need of discovering novel and most useful information from these large amounts of data has also grown. With the advent of data mining, different mining techniques have been applied in different application domains, such as, Education, banking, retail sales, bioinformatics, and Telecommunications. To extract useful information to fulfill the needs of the industry. With the enormous amount of data stored in files, databases, and other repositories, it is increasingly important, though not necessary, to develop a powerful means for analysis as well as interpretation of such data and for the extraction of interesting knowledge that could help in decision-making. It is intended to obtain meaningful and valuable information that is not previously known from these data by applying data mining techniques. One of the significant facts in higher learning institutions is the explosive growth of educational data. These data are increasing rapidly without any benefit to the management. Good prediction of student's success in higher learning institution is one way to reach the highest level of quality in the higher education system. Many prediction models available with a difference in approach to student performance were reported by the researcher, but there is no certainty that there are any predictors who can accurately determine whether a student will be an academic genius, a drop out, or an average performer. The higher education institutions use automated computer programs/tools developed with different technologies to predict the trades in the college. With the potential techniques in Data Mining and with the growth of technologies to handle huge databases, the predictive technologies have started growing tremendously. The academic research in Data Mining also contributed a lot to predictive technologies. The prediction of academic performance is regarded as a challenging task of temporal data prediction. Data analysis is one way of predicting increase or decrease of future academic performance.

Dataset characteristics	Multivariate
AttributeCharacteristics	Multivariate
Associated Task	Classification
Number of instances	480
Number of Attributes	16

2. LITERATURE REVIEW

The application of knowledge mining widely spreaded in education system. This are in Education domain there are many the researchers and authors are explored and discussed various applications of knowledge mining in education. The authors had skilled the survey of the literature to know the importance of knowledge mining applications in education , the utilization of knowledge mining to research scientific questions within educational research for the standard improvements during this area. For predicting students performance V. Ramesh et al applied Naïve Bayes Simple, Multilayer Perception, SMO, J48, REP Tree techniques. From the results obtained Multilayer Perception algorithm is outstanding algorithm for predicting student performance. MLP gives 87% prediction which is comparatively above other algorithms. This study is an

attempt to use classification algorithms for predicting the student performance and comparing the performance of Naïve Bayes Simple, Multilayer Perception, SMO, J48, and REP Tree [1].

Cortez and Silva [2] attempted to predict failure within the 2 core courses that is namely Mathematics and Portuguese of two lyceum students from the Alentejo region of Portugal by using 29 predictive variables. Four data processing algorithms such as, Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM) and Neural Network (NN) were applied on a knowledge set of 788 students, who sat for 2006 examination. DT algorithm got 93% of accuracy and NN algorithm got 91% of accuracy. It was also reported that both DT and NN algorithms had the predictive accuracy of 72% for a four-class dataset.

Erdogan and Timor 2005 et al used educational data processing to spot and enhance educational process which will improve their deciding process. Finally Henrik, 2001 et al observed that clustering was effective find hidden relationships and associations between different categories of students[3].

Kotsiantis [4] applied five classification algorithms namely, Decision Trees, Perceptron-based Learning, Bayesian Nets, Instance-Based Learning and Rule learning, to predict the performance of computing students from distance learning stream of Hellenic Open University, Greece. A total of 365 student records comprising several demographic variables like sex, age and legal status were used. In addition, the performance attribute, namely the marks during a given assignment was used as input to a binary (pass/fail) classifier. Filter based variable selection technique was wont to select highly influencing variables and every one the above five classification models were constructed. It was noticed that the Naïve-Bayes algorithm got the highest accuracy of 74% for two-class (pass/fail) dataset.

Khan [5] conducted a performance study on 400 students comprising 200 boys and 200 girls selected from the senior lyceum of Aligarh Muslim University, Aligarh, India, with the target of establishing the prognostic value of various measures of cognition, personality and demographic variables for fulfillment at higher secondary level within the science stream. The selection was supported cluster sampling technique during which the whole population with interest was divided into groups, or clusters, and a random sample of those clusters was selected for further analyses. It was found that the girls with high socio-economic status had relatively higher academic achievement in the science stream and boys with low socioeconomic status had relatively higher academic achievement in general. Cristóbal Romero[6] compared different data mining methods and techniques for classifying students based on their Moodle usage data and therefore the final marks obtained in their respective courses, and developed a selected mining tool for creating the configuration and execution of knowledge mining techniques easier for instructors. A classifier model appropriate for educational use has got to be both accurate and comprehensible for instructors so as to be of use for deciding.

By using the CGPA grading system M.N. Quadri [7] have predicted student's academic performance where the data set consists the parents educational details, his financial background and the students gender.

In [8] the author explored the various variables to predict the students who are at the risk of failing in the exam. The solution strongly suggests that the previous academic result strongly plays a serious role in predicting their current outcome.

To find the relationships between the students behaviour and their success, Sajadin Sembiring [9] et al applied processing techniques and predicted the model on student performance. This is done by using kernel k-means clustering and Smooth Support Vector Machine classification techniques. Recent advances in data collection and storage technology have made it possible to gather vast amounts of knowledge everyday in many areas of business and science. Examples are climate measures, stock exchanges, web logs, recordings of sales of products and so on. One major area of knowledge mining from these data is association pattern analysis. Association rules discover interrelationships among various data items in transactional data.

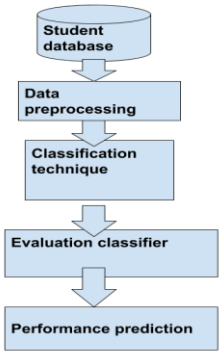
T.BalaKrishna, [2017] [10], "Diagnosis of chronic disease using Random Forest Classification Technique" has discussed about the classification panel enables the user to apply classification algorithms to the resulting dataset, to estimate the accuracy of the resulting predictive model.

3. PROBLEM STATEMENT

Educational organizations are one among the important parts of our society and playing an important role for growth and development of any nation. Educational data mining is the application of data mining. It is an emerging interdisciplinary research area that deals with the development of methods to explore data originating in an educational context.

Predicting student performance becomes more challenging due to the large volume of data in educational databases. This is due to main two reasons. First, the study of existing prediction methods remains insufficient to spot the foremost suitable methods for predicting the performance of scholars within the University. Second is due to the lack of investigations on the factors affecting student's achievements in particular courses. Therefore, a necessary literature reviews on predicting student performance by using data technique.

4. PROPOSED METHODOLOGY



This model uses classification algorithms. The idea of Classification Algorithms is pretty simple. By analyzing the training dataset, you predict the target class. This is one of the most, if not the most essential concept you study when you learn Data Science. Our model uses the training dataset to get better boundary conditions which could be used to determine each target class. Once the boundary conditions are determined, subsequent task is to predict the target class. The whole process is known as classification.

The classification algorithms which are considering for our model are

• Decision Tree

• Random Forest

• Naïve Bays Classifier

By using WEKA tool predicted the accuracy for our model. Therefore Random Forest got the highest accuracy.

Random Forest:

Random forest is a supervised learning algorithm. It are often used both for classification and regression.

How Random Forest algorithm works: There are two stages in Random Forest algorithm, one is random forest creation, the opposite is to form a prediction from the random forest classifier created in the first stage. The whole process is shown below, and it's easy to understand using the figure.

1. Here the author firstly shows the Random Forest creation pseudocode: Randomly select "K" features from total "m" features where k << m.

2. Among the "K" features, calculate the node "d" using the best split point.

3. Split the node into daughter nodes using the best split.

4. Repeat the a to c steps until "l" number of nodes has been reached.

5. Build forest by repeating steps a to d for "n" number times to create "n" number of trees.

In the next stage, with the random forest classifier created, we'll make the prediction. The pseudo code for random forest prediction is shown below:

1. Takes the test features and it uses the rules of each randomly created decision tree to predict the outcome and later stores the predicted outcome (target).

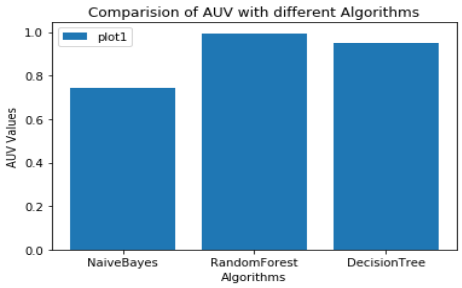
2. Calculate the votes for each predicted target.

3. The final prediction comes from high voted predicted target from the random forest algorithm.

The process is easy to understand, but it's somehow efficient.

5. IMPLEMENTATION AND RESULTS

To accurately recommend students with an appropriate choice for their study at Master level, an intelligent predictive model is proposed. The tool WEKA 3.8.3 is used in this work to implement data mining approach and later implemented in Jupiter notebook. By considering the few attributes and records the model predicted the accuracy. Naïve-Bayes classifier algorithm got 0.74 percentage of accuracy. Decision tree got 0.94 percentage of accuracy and the Random Forest technique got 0.99 percentage of accuracy.



And at last random forest technique is applied on the overall dataset and predicted the accuracy as 0.81 percentage.

```
from sklearn.metrics import classification_report, accuracy_score
y_true = test["Class"]
y_pred = my_prediction
target_names = ['class 0(L)', 'class 1(M)', 'class 2(H)']
print(classification_report(y_true, y_pred, target_names=target_names))
```

```
              precision    recall  f1-score   support

  class 0(L)       0.81      0.89      0.85        19
  class 1(M)       0.80      0.80      0.80        45
  class 2(H)       0.83      0.78      0.81        32

   micro avg       0.81      0.81      0.81        96
   macro avg       0.81      0.83      0.82        96
weighted avg       0.81      0.81      0.81        96
```

```
print("Accuracy is:")
print(accuracy_score(y_true,y_pred))
```

```
Accuracy is:
0.8125
```

6. DISCUSSIONS

The ambition of project is to compare classification methodology and to forecast the student performance. It enhances the moderate students in the semester tests those are likely to be poor in examination to upgrade their performance by the end semester. The assignment might be prepared dependent on the few attributes or properties to forecast the student performance individually. In this exploration, the paper centered the enhancement of Prediction/arrangement strategies which are utilized to break down the aptitude ability dependent on their scholarly execution by the extent of learning. Additionally the model uses WEKA 3.8.3 to demonstrate the relative execution of Decision tree, Naïve Bayesian classifier calculation, Random tree calculation to analyze the performance of students. Out of the three classifier techniques random forest predicts the more accuracy and this prediction mechanism is implemented in the Jupiter notebook.

7. CONCLUSION

In this research, an effort is made to find the impact of our proposed features on student performance prediction with the help of classification models. This model uses the data mining techniques such as Decision Tree, Random Forest and Naïve-Bayes classifier to predict the accuracy of our model. Firstly by considering few attributes the model predicted the accuracy. Out of three algorithms Random Forest got the highest accuracy. Hence random forest technique is applied on the entire dataset which results the 0.81 percentage of accuracy.

8. FUTURE SCOPE

In future, accuracy rate can be calculated on actual data for different – different organizations by modifying or changing attributes. Anyone can apply different – different methods to know the best method suited for education domain.

9. ACKNOWLEDGEMENT

We thank the coordinator of Smart bridge academy Mahankali Surya Tej for useful discussions on machine learning techniques. And we thank Tilakachuri Bala Krishna(Assistant professor) for the support.

REFERENCES

[1] Dunham, M.H., (2003) Data Mining: Introductory and Advanced Topics, Pearson Education Inc.

[2] Witten, I.H. & Frank E. (2000), data processing – Practical Machine Learning Tools and Techniques, Second edition, Morgan Kaufmann, San Francisco .

[3] W. H¨am¨al¨ainen, M. Vinni, Comparison of machine learning methods for intellige nt tutoring systems, in: Intelligent Tutoring Systems, Springer, 2006, pp. 525–534.

[4] S. Sembiring, M. Zarlis, D. Hartama, S. Ramliana, E. Wani, Prediction of student academic performance by an application of data mining techniques, in: International Conference on Management and AI IPEDR, Vol. 6, 2011, pp. 110–114.

[5] G. Gray, C. McGuinness, P. Owende, An application of classification models to predict learner progression in tertiary education, in: Advance Computing Conference (IACC), 2014 IEEE International, IEEE, 2014, pp. 549–554.

[6] Kabakchieva, D., Stefanova, K., Kisimov, V. (2011). Analyzing University Data for Determining Student Profiles and Predicting Performance. Conference Proceedings of the 4th International Conference on Educational data processing (EDM 2011), 6-8 July 2011, Eindhoven, Netherlands , pp.347-348.

[7] Baradwaj, B.K. and Pal, S., 2011. Mining Educational Data to Analyze Students' Performance. (IJACSA) International Journal of Advanced computing and Applications, Vol. 2, No. 6, 2011.

[8] Ahmed, A.B.E.D. and Elaraby, I.S., 2014. Data Mining: A prediction for Student's Performance Using Classification Method. World Journal of Computer Application and Technology, 2(2), pp.43-47.

[9] Amjad Abu Saa, 2016. Educational Data Mining & Student's Performance Prediction. International Journal of Advanced computing and Applications, Vol. 7, No. 5, 2016.

[10] T.BalaKrishna, B. Narendra, M H Reddy, D.Jayasri, 2017. Diagnosis of Chronic renal disorder Using Random Forest Classification Technique. HELIX Journal, 7(1),pp,873-877