Data Mining & Predictive Analysis

(ICT Department)

Manipal Institute of Technology

**Random Forest Classifier to analyze Student Performance**

A Project report by:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sl#** | **Full Name** | **Reg#** | **Roll#** | **CCE section name** |
| **1** | Priyansh Jain | 190953116 | 39 | B (Batch 4) |
| **2** | Avanish Shinde | 190953098 | 35 | B (Batch 4) |

**Abstract**

With the increasing amount of data being generated, it is important that the data is properly maintained. In educational Institutes, a lot of decisions are needed to be made based on various factors that can affect the student’s academic performance and overall health and well-being. Decision trees provide an effective method of Decision Making because they clearly lay out the problem so that all options can be challenged and allow us to analyse fully the possible consequences of a decision.

In our project, we will be using random forest classifier which combines several decision trees and try to predict factor’s related to a student. We then analyse the results obtained and discuss the benefits and the shortcomings of the same.

**Introduction**

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 to improve students’ learning processes.Decision tree is one of the most popular and powerful approach to classification and prediction. It is a hierarchical structure that is built using the features (or the independent variables) of a data set. These have been implemented on an Institutions student record but its variation, random forest algorithm’s effectiveness is still being thought in this area. The key principle underlying the random forest approach comprises the construction of many “simple” decision trees in the training stage and the majority vote (mode) across them in the classification stage.

The random forest classifier bootstraps random samples where the prediction with the highest vote from all trees is selected. The individuality of the trees is important in the entire process. The individuality of each tree is guaranteed due to the following qualities. First, every tree training in the sample uses random subsets from the initial training samples. Secondly, the optimal split is chosen from the unpruned tree nodes’ randomly selected features. Thirdly, every tree grows without limits and should not be pruned whatsoever.

Random Forest is based on the bagging algorithm and uses Ensemble Learning technique which reduces overfitting problem in decision trees and reduces the variance and therefore improves the accuracy. This study will also help to identify those students who needed special attention to reduce fail ratio and taking appropriate action for the next semester examination. Thus, if such an algorithm is implemented on an Institutions student record, it could help in predicting how any decision taken can influence a student more accurately.

**Literature Survey**

The application of data mining has spread widely in education system. In Education domain there are many researchers and authors who have explored and discussed various applications of data mining in education. The authors did comprehensive survey of the literature to know the importance of data mining applications in education, the utilization of data mining to research scientific questions within this area for further improvements.

By using the CGPA grading system [1] have predicted student’s academic performance where the data set consists of the parents’ educational details, his financial background and the student’s gender.

In [2] 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 student’s behaviour and their success, [3] 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

**Methodology**

**About the Dataset**

1] school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)

2] sex - student's sex (binary: 'F' - female or 'M' - male)

3] age - student's age (numeric: from 15 to 22)

4] address - student's home address type (binary: 'U' - urban or 'R' - rural)

5] famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)

6] Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)

7] Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2- 5th to 9th grade, 3- secondary education or 4- higher education)

8] Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)

9] Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at\_home' or 'other')

10] Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at\_home' or 'other')

11] reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')

12] guardian - student's guardian (nominal: 'mother', 'father' or 'other')

13] traveltime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)

14] studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)

15] failures - number of past class failures (numeric: n if 1<=n<3, else 4)

16] schoolsup - extra educational support (binary: yes or no)

17] famsup - family educational support (binary: yes or no)

18] paid-extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)

19] activities - extra-curricular activities (binary: yes or no)

20] nursery - attended nursery school (binary: yes or no)

21] higher - wants to take higher education (binary: yes or no)

22] internet - Internet access at home (binary: yes or no)

23] romantic - with a romantic relationship (binary: yes or no)

24] famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)

25] freetime - free time after school (numeric: from 1 - very low to 5 - very high)

26] goout - going out with friends (numeric: from 1 - very low to 5 - very high)

27] Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)

28] Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)

29] health - current health status (numeric: from 1 - very bad to 5 - very good)

30] absences - number of school absences (numeric: from 0 to 93)

31] G1 – grade1 (numeric: from 0 to 20)

31] G2 – grade2 (numeric: from 0 to 20)

32] G3 – grade3 (numeric: from 0 to 20)

33] Stu\_per - 1-6 Representing A-F grades

**Data Pre-processing**

We are using RapidMiner to pre process our data so that it is fit to be passed into our random forest algorithm. To do so we have mapped all Nominal data to corresponding numerical data. We are also using RapidMiner’s correlated operator to get a good grasp on the relationship between attributes.

Graphical user interface

Description automatically generated

Converting Nominal to Numerical Data

Chart

Description automatically generated

Correlation Matrix Visualisation using RapidMiner

**Random Forest Classification**

After the pre-processing is done, we use the sci-kit learn library to train and implement our random forest algorithm. After importing the data, we split the data into two parts, one for training and the other for testing in the ratio 0.7:0.3. Labels are added for target attributes.

Then we are importing Random Forest Classifier and training the model. We are making 100 decision trees and based on their votes the final prediction will be made. We are also importing the metrics to get the accuracy of our model. To visualize how the results are being processed we have visualized feature importance, heatmap and 2 decision trees. Random Forests allow us to look at feature importance’s, which is the how much the Gini Index for a feature decrease at each split. The more the Gini Index decreases for a feature, the more important it is.

A screenshot of a computer

Description automatically generated with medium confidence

One of the decision trees formed

**Results**

The Random Forest Algorithm gives an accuracy of around 0.87 whereas the decision tree gives an accuracy of 0.80. We can also predict any student’s performance by passing the required parameters. We also get the feature importance list which is highly beneficial.

We also learn that a student’s performance is highly correlated with his performance in previous semester because when the same dataset is classified using random forest algorithm removing semester grades the accuracy drops to around 0.30 and with decision tree the accuracy is 0.24.

We have also predicted the student health using Random Forest algorithm as well as decision trees. With Random Forest the accuracy comes as 0.31 and with Decision tree the accuracy comes out to be 0.27. We can see that the attributes are not strongly attribute to health hence the low accuracy.

**Conclusion**

Random Forest Algorithm has a good potential to be used in such classification problems and can be helpful in making decision. It is less prone to overfitting than Decision trees and other algorithms. It also outputs the importance of features which is very useful. No feature scaling is required since it uses rule-based approach. It is robust to outliers and is a very stable algorithm.

On the other hand, since random forest algorithm creates multiple decision trees, the overall complexity increases, and it also has longer training period. Despite this, Random Forest Algorithm has good performance and can be a huge help in classification problems and assist in better decision making and analyzation of a problem.

**References**

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

[2] 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.

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

[4] <https://www.kaggle.com/impapan/student-performance-data-set>