International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-2, July 2019

Students' Performance Prediction Modelling using Classification Technique in R

Thingbaijam Lenin, N. Chandrasekaran

Abstract*: Among several important tasks an academic institution performs, the most fundamental focus still remains very much on graduating best quality students. It then becomes of paramount importance to identify students whose performance is below par in order to help them to make them better learners. This study makes an earnest attempt to develop an automated system to tackle such a problem using a classification technique of Data Mining implemented with R programming language. Data pertaining to students' demographic features, their previous academic records and personality traits were analyzed employing Random Forest, Naïve Bayes and K-Nearest Neighbors algorithms. The study shows that Personality, as defined by Myers-Briggs type indicator, influences the student's performance. Random Forest is found to be the most promising algorithm for developing the students' performance prediction system.

KEYWORDS*: Classification Technique, Educational Data Mining, KNN, Naïve Bayes, Random Forest, R Programming.

I. INTRODUCTION

It is a known fact that the primary goal of an educational institution is to open all possible avenues to every student to enable him or her to accomplish as high a level of academic achievement as possible. In reality, the performance varies widely, even when they learn under the same educational setup and are provided the same environment and tools. Several factors contribute to sustain these disparities. An intelligent data mining algorithm might be able to identify the exact nature of the problem to enable one to rectify any lacuna that may arise due to issues that can be tackled by the concerned teachers and the University. It is in this context that we attempt to gain greater experience in the use of appropriate data mining tools that will provide a deeper insight. Several researchers and data analysts have made significant contributions to this field. Some of the classification algorithms are listed in Ref. [1], while [2] provides a comparison of 8 different machine learning algorithms. Further, Refs. [3] and [4] discuss some relevant applications.

Revised Manuscript Received on 30 July 2019.
***** Correspondence Author
Thingbaijam Lenin*, Research Scholar and Assistant Professor, Department of Computer Science, Martin Luther Christian University, Meghalaya, India
N. Chandrasekaran, Visiting Professor, Martin Luther Christian University, Meghalaya, India

The main objective of the current study is to find a solution to the difficulty that many students face while studying in a University situated at a remote North Eastern Region of India like MLCU. Also of concern is the preparedness level of students at the time of entry into a higher education institution.

The objectives of the study are twofold:

1. To determine if personality and/or aptitude of the student determines the academic performance of the student.
2. To find out the classifier that can be used to develop an appropriate model to better understand the issues.

We have used Random Forest, Naïve Bayes and K-Nearest Neighbors (KNN) classification algorithms and these have been implemented using R programming language employing the free version of RStudio integrated development environment. The students are categorized as "Good" and "Fair" and the students predicted "Fair" are considered to be the one who may be weak and are likely to fail or drop out of the University.

II. RELATED WORKS

Baradwaj and Pal [4] collected data of 50 MCA students from the Computer Applications Department VBS Purvanchal University, Jaunpur, UP and constructed a decision tree using ID3 algorithm in order to identify those students who might need special attention during their course of study. The attributes used are Marks of Previous Semester, Class Test Grades, Attendance, Assignments, Lab Work, General Proficiency, Seminar Marks and End Semester Marks.

Sing and Kumar [5] also used 50 instances of student data collected from different branches of an Engineering College in an effort to analyze the performance of students. The attributes used for the study are student's Name, Branch, passing percentage (%) of 10th class, 12th class and B.Tech. They used WEKA to perform analysis of six classification methods namely BayesNet, Naïve Bayes, Multilayer Perceptron, IB1, Decision Table and PART Classification method. For this particular study, IB1 Classifier was found to be the most suitable method. In Ref. [6], 182 records of the students of Coimbatore Institute of Technology were used to develop a prediction model for predicting the performance using C. Seven factors, namely, Marks of SSLC, HSC, Subject Difficulty, Family Income, Stay, Medium of Instruction and Staff Approach, are identified to study the influence on performance. Bayes classifier showed highest accuracy with a success score of 82.4%. Ref. [7]

considered 13 attributes, viz, 'Types of High School', 'Types of Education board ', 'Medium of Instruction',' Type of School', 'Gender of student', 'Private Tuition taken', 'Location of the school area',' Internal Grade of student','Mobile Phone',' Computer at Home ', 'Internet access to student', 'Attendance in the school ','Class'. They analysed the data in the classifiers - Naive Bayes, J48, RepTree, ZeroR, OneR, SMO and Multilayer Perceptron.

From these classifiers, Multilayer Perceptron algorithms gave an accuracy of 87.44% and proved to be very effective for the identification of slow learners.

Ref.[8] makes an attempt to analyze the academic performance of the MCA students to link it to the probability of placement at the time of admission itself. The study used numerical data of features related to previous academic scores. They constructed a classification tree, which predicted the final year results and success of placement with an accuracy of 38.46% and 45.38% respectively.

In Ref.[9], the authors used ID3, C4.5, CHAID classification algorithms for designing a model which can predict the placement of the students. They used marks obtained at 10th, 12th or Diploma, Degree levels, Semester Performance, Communication Skills, Work on Projects, Internship, education gap, backlog as the attributes. ID3 algorithm gave the best results with 95.33% accuracy.

III. PREDICTION MODELLING

A. Data Collection and Preparation

Data for the study is collected from the Department of Computer Science, Martin Luther Christian University (MLCU), Shillong obtained from the Administration Office of the University. The data consists of information collected at the time of admission and the students' transcripts. The University has also collected data related to personality of the students as defined by MBTI [10]. This MBTI information is also collected for the study.

A total of 106 observations with 11 attributes were collected from the batch from 2014 - 2016 of the department. MBTI and other relevant information of 27 students were found to be missing. As the MBTI and the other missing values are intrinsic in nature and for achieving the goal of the study, the missing values were not substituted by any computing technique and they are discarded from the study. Finally a total of 79 observation were selected. Eleven attributes as provided below are used for the purpose of the study.

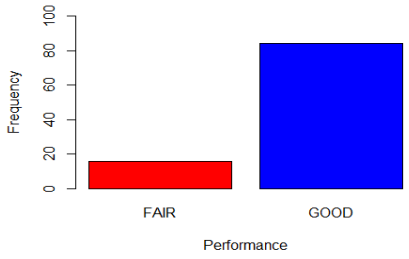
All the attributes are categorical in nature and appropriate values have been assigned for developing the model as in table 1.

Table 1: Features used in the study

Sl. No.	Attribute	Description	Values
1	Gender	Gender of the student	"M" denoting Male and "F" denoting Female
2	Category	Whether the student belongs to General, Other backward class, schedule caste or schedule tribe as categories at India	"GEN" denoting General, "ST" for Schedule Tribe, SC for schedule caste and OBC for other backward class
3	PL	Permanent Location where the student was brought up	Rural, Urban
4	Sinsti	Location of the institute where the student studied or appeared for the Matriculation or appeared as a private candidate	Rural, Urban, Private
5	HInsti	Location of the institute where the student studied or appeared for the Higher Secondary Examination/12th Std. or appeared as a private candidate	Rural, Urban, Private
6	Poccu	Parents Occupation	Govt, Business, Pvtservice, Housewife
7	MPerform	Performance of the student at the Matriculation as provided by the Matriculation Exam	"Good" for >=60%; "Medium" >=45% else " Fair"
8	Hsub	Subjects studied at the Higher Secondary/12thStd	"Com" if Computer is studied as a subject at 12th else "others"
9	Hperform	Performance of the student at the higher secondary as provided by the Board Examination at 12thStd	"Good" for >=60%; "Medium" >=45% else " Fair"
10	MBTI	Myers–Briggs Type Indicator constructed by Katharine Cook Briggs and her daughter Isabel Briggs Myers.	ISTJ,ISFP,ISFJ, ISTP,ESFP,INFP, ENTJ,ESTP,INTJ, ENTP, ESTJ,INTP,ENFJ, INFJ, ESFJ,ENFP
11	Perform	Performance of the student as defined by the CGPA	If CGPA>1.6, Good else Fair

Using the sample() of R, a random sample of 57 observations is taken as training data and the remaining 22 as testing data. The distribution of the target class "Perform" in the training data is found to be 48 of GOOD and 9 of FAIR as shown in Fig. 1.

Figure 1: Distribution of the Target Class, Perform



This study focuses more on predicting and understanding the students who are in "Fair" group. This group is a minority class in this dataset. This minority class may be considered as noise while performing the modelling leading to misclassification and bias.

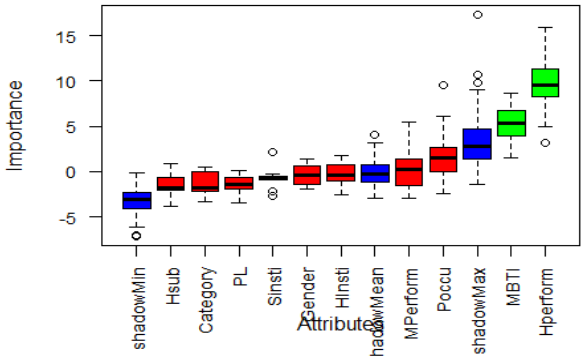
In order to avoid the above problem, the training data is over sample and balanced using ovun.sample() of the Rose library available for R.

This function uses Synthetic Minority Over-sampling Technique (SMOTE) where the minority class is over-sampled by creating synthetic examples rather than by over-sampling with replacement [11]. As a result the size of the training data is increased from 57 to 96.

B. Feature Selection

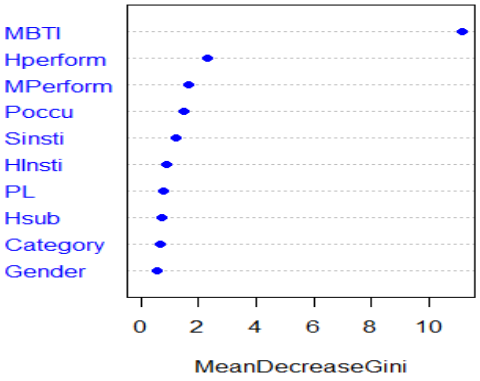
Not all attributes in any dataset may be contributing to developing a predictive model. In order to identify which attributes will be contributing to models, feature selection process is performed. In this study, Boruta Library of R has been used for this purpose, with parameter, maxRuns of set at 1000. 81 iterations are executed in 3.39 secs and attributes "Hperform" and "MBTI" are confirmed to be important attributes and others as unimportant, see Fig. 2.

Figure 2: Most influencing features



In order to further ensure that we are on the right track, this feature has also been investigated using Random Forest Algorithm. It uses Gini index to assign a score and the features are ranked based on it. It is found that the attribute "MBTI" shows the high test followed by "MPerform" and "Hperform", see Fig. 3.

Figure 3: Feature Selection using Random Forest



From the above analysis, we can conclude that "MBTI" and "Hperform" are the most important features in developing the model for this dataset.

C. Modelling

Three algorithms are explored for the purpose of this study. They are Random Forest, Naïve Bayes and K-Nearest Neighbors

i. Random Forest:

Random Forest is a supervised learning algorithm. It is defined as classifier consisting of a collection of tree-structured classifiers {h(x,Θk),k=1,..} where the {Θk} are independent, identically distributed random vectors. Each tree casts a unit vote for the most popular class at input x. It provides better accuracy in prediction [12].

Using randomForest(), the balanced training data is used for training the random forest model and to validate the model using the testing data.

Three variables are used at every split and a total of 500 trees were generated. The confusion matrix obtained in the process is given below:

```
## Confusion Matrix and Statistics
##                 Reference
## Prediction      FAIR   GOOD
##      FAIR        4       0
##      GOOD        1       17
##
##                Accuracy          : 0.9545
##                  95% CI          : (0.7716, 0.9988)
##     No Information Rate          : 0.7727
##     P-Value [Acc> NIR]           : 0.0257
##                   Kappa          : 0.8608
##  Mcnemar's Test P-Value          : 1.0000
##             Sensitivity          : 0.8000
##             Specificity          : 1.0000
##          PosPred Value           : 1.0000
##          NegPred Value           : 0.9444
##              Prevalence          : 0.2273
##          Detection Rate          : 0.1818
##    Detection Prevalence          : 0.1818
##       Balanced Accuracy          : 0.9000
##       'Positive' Class           : FAIR
```

The confusion matrix shows an accuracy of 95% with sensitivity 0.8, specificity of 1.0 and kappa value 0.86.

ii. Naïve Bayes:

Classification decision provided by Naïve Bayes is found to be correct even if the probability estimates are inaccurate [13]. This algorithm is considered for the study as it is considered to provide a better classifier than more powerful alternatives when the sample size is small [14]. We used naïve_bayes function to obtain a model of this technique. The result obtained is given below:

```
## Confusion Matrix and Statistics
##                 Reference
## Prediction      FAIR   GOOD
##    FAIR         5       2
##    GOOD         0       15
##                          Accuracy          : 0.9091
##                            95% CI          : (0.7084 ,0.9888)
##     No Information Rate                    : 0.7727
##     P-Value [Acc> NIR]                     : 0.09444
##                   Kappa                    : 0.7732
##  Mcnemar's Test P-Value                    : 0.47950
##             Sensitivity                    : 1.0000
##             Specificity
          : 0.8824
##          PosPred Value
          : 0.7143
```

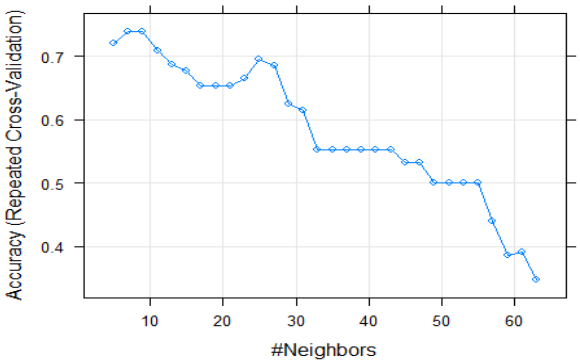
```
##          NegPred Value : 1.0000
##              Prevalence : 0.2273
##          Detection Rate : 0.2273
##    Detection Prevalence : 0.3182
##       Balanced Accuracy : 0.9412
##       'Positive' Class : FAIR
```

The result shows an accuracy of 90.91%. The Kappa value is reasonably good with 0.77. Sensitivity and specificity are found to be 1.00 and 0.88 respectively.

iii. K-Nearest Neighbors:

The "train" function with the method "knn" of R is used for the purpose of modelling using K-Nearest Neighbors. The model was constructed using the balanced data and test using the test data.

Figure 4: Accuracy of K-NN and K value



The model has the highest accuracy at K=9. The confusion matrix produced for the prediction using the testing data is as follows:

```
## Confusion Matrix and Statistics
##                 Reference
## Prediction      FAIR  GOOD
##      FAIR          4     7
##      GOOD          1    10
##
##               Accuracy : 0.7442
##                 95% CI : (0.416, 0.826)
##    No Information Rate : 0.7727
##    P-Value [Acc> NIR] : 0.8432
##                  Kappa : 0.2727
##  Mcnemar's Test P-Value : 0.0771
##             Sensitivity : 0.8000
##             Specificity : 0.5882
##          PosPred Value : 0.3636
##          NegPred Value : 0.9091
##              Prevalence : 0.2273
##          Detection Rate : 0.1818
##    Detection Prevalence : 0.5000
##       Balanced Accuracy : 0.6941
##       'Positive' Class : FAIR
```

The model predicts with an accuracy of 74.42% and a sensitivity of 0.80 and specificity of 0.58. The kappa value is found to be 0.27.

IV. RESULT AND CONCLUSION

The results of modelling for the purpose of predicting the students' performance and identifying the student who may need special attention can be seen in table 2:

Table 2: Comparison of the models

Algorithm	Accuracy	Sensitivity	Specificity	Kappa
Random Forest	95.45%	0.8	1.0	0.86
KNN	74.42%	0.8	0.58	0.27
Naïve Bayes	90.91%	1.0	0.8	0.77

Out of the three widely used algorithms, Random Forest is found to provide better accuracy with high specificity and Kappa value. This further proves that Random Forest is one of the best classifiers for classification [15]. Based on this study, we can conclude that Random Forest model should be deployed to provide a predicting system for the student performance. This study also shows that personality of the student as provided by Myers–Briggs type indicator is an influencing factor in determining the performance of the students and thus merits further investigation using larger dataset and features. The performance of the students at Senior Secondary level also plays an important role in their academic performance in the University. These powerful algorithmic tools and the findings can be a boon to the administration while devising appropriate strategies to improve overall performance.

ACKNOWLEDGMENT

We are thankful to the administration of Martin Luther Christian University, Meghalaya for providing the necessary data without which this study would have been impossible.

REFERENCES

1. Suraj V. Vidyadaran, "Implementation of 17 classification algorithms in R", https://www.datasciencecentral.com/profiles/blogs/implemetation-of-17-classification-algorithms-in-r, March 13, 2016.
2. K. Markham, "Comparing supervised learning algorithms", https://www.dataschool.io/comparing-supervised-learning-algorithms/, February 27, 2015.
3. Olawunmi George, "Implementing a simple prediction model in R", https://codeburst.io/implementing-a-simple-prediction-model-in-r-ab1dafa6b954, Apr 9, 2018.
4. B. Baradwaj and S. Pal, "Mining educational data to analyze student's performance," *Int. J. Adv. Comput. Sci. Appl.*, vol. 2, no. 6, pp. 63–69, 2012.
5. S. Singh and V. Kumar, "Performance Analysis of Engineering Students for Recruitment Using Classification Data Mining Techniques," vol. 3, no. 2, pp. 31–37, 2013.
6. J. Shana and T. Venkatachalam, "Identifying Key Performance Indicators and Predicting the Result from Student Data," *Int. J. Comput. Appl.*, vol. 25, no. 9, pp. 45–48, 2011.
7. M. Kumar, S. Shambhu, and P. Aggarwal, "Recognition of Slow Learners Using Classification Data Mining Techniques," no. 12, pp. 741–747, 2016.
8. N. Naik and S. Purohit, "Prediction of Final Result and Placement of Students using Classification Algorithm," *Int. J. Comput. Appl.*, vol. 56, no. 12, pp. 35–40, 2012.
9. N. Puri, D. Khot, P. Shinde, K. Bhoite, and P. D. Maste, "Student Placement Prediction Using ID3," vol. 3, no. Iii, pp. 81–84, 2015.

10. "The Myers-Briggs Foundation - MBTI® Basics." [Online]. Available: https://www.myersbriggs.org/my-mbti-personality-type/mbti-basics/home.htm?bhcp=1. [Accessed: 29-Apr-2019].
11. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, no. Sept. 28, pp. 321–357, 2002.
12. L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
13. I. Rish, "An empirical study of the naive Bayes classifier," *J. Mach. Learn. Res.*, pp. 41–46, 1999.
14. P. Domingos, M. Pazzani, "On the Optimality of the Simple Bayesian Classifier under Zero-One Loss," *Mach. Learn.*, vol. 29, pp. 103–130, 1997.
15. M. Fernández-Delgado, E. Cernadas, S.Barro, D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?," *J. Mach. Learn.*, vol. 15, pp. 3133–3181, 2014.

AUTHORS PROFILE

Thingbaijam Lenin is a Research Scholar and also an Assistant Professor of Department of Computer Science, Martin Luther Christian University, Meghalaya. He received MCA degree from University of Madras and is a strong supporter of Free and Open Source Software. He is keen on working on the implementation of Open source systems in various fields. His area of interest includes Data mining and its applications, Data Science and Analytics, Machine Learning, and Software development. He is also a member of Computer Society of India.

Dr. N. Chandrasekaran obtained his Ph.D. degree from McMaster University, Canada and M.Tech. from the Indian Institute of Technology, Madras. He is the proud recipient of the Sir Rajendra Nath Mookerjee Memorial Gold Medal, the Institution Award of IE (India) and the Outstanding Contribution Award from IBM. He has served as a Scientist at National Aerospace Lab., Deputy Director at Defence R&D Organization, Hyderabad, Principal Scientist at Ontario Centre for Advanced Manufacturing, Cambridge, Forming Technology Inc., Oakville and McMaster University. He has been instrumental in setting up two World Class Centers (Competency Center and IBM Solution Partnership Center) for IBM in India at Bangalore. He has over 25 International publications and several industrial reports to his credit. He has delivered close to 26 Key Note Speeches in important Conferences in India. He has served as the Chairman of the Karnataka IT Task Force on Higher Education and also as the Chairman of the Karnataka State Computarium Project, Director of C-DAC, a Society formed under Ministry of Information Technology, Government of India,Senior Vice President of Manipal Education & Medical Group, Pro-Vice Chancellor of Manipal University, Director at SRM University, Chennai. He has established Middle East College of IT at Muscat. He has served as an Adjunct Faculty at Martin Luther Christian University, Shillong. He had been an active Panel Member in FKCII, CII, etc. and one of the three members of the Software Development Fund of Electronics Corporation of Tamil Nadu Limited. He had served as a Member in the Governing Council of R.V College of Engineering and SSIT, Tumkur and as Director at SRM University, Chennai.