Analysing Student Performance Through Data Mining Model

Raghav Mehta
School of Computing (MSc in Data Analytics)
National College of Ireland
Dublin, Ireland
x17155151@student.ncirl.ie

***Abstract*—With constant innovation in the field of technology and data science, it is not surprising that education institutions are interested in understanding the performances' of their students. The best indicator to measure a student's performance is through their grades, but institutions are more interested in the factors that affect these scores. These institutions are looking to develop tools to enhance the quality of education and ensure a high success rate amongst students, to be facilitated by business intelligence and data mining techniques. In this paper we will assess the right technique to be applied to our problem dataset and optimize this model. Through this research, we are able to model a student's final grade in a particular subject and link it directly to certain relevant features that influence the outcome. We use the C5.0 decision tree technique to model the data.**

***Keywords*—data mining, C5.0, decision tree, feature selection, classification, exploratory data analysis**

I. Introduction

Economic progress is correlated to the education level of its youth and is a driving factor for the growth of a nation. However recent statistics show high failing rate amongst students in the 16 to 22 year age group (Cortez et al., 2008). Schools are interested in figuring out the students that may have difficulties in clearing a particular course ie. Mathematics or Portuguese language. The purpose is to transform the raw data to actionable information for the educational institutions stakeholders. The data of the students from a Portuguese school has been collected by merging two sources of data – past evaluation performance of the student and a questionnaire for demographic and social inference. The model created will be used to answer our underlying hypothesis - can students' performance be predicted and what factors are likely to affect these student's at risk ?

The grade outcome has been transformed to fall in one of four classes – fail, average, satisfactory or excellent. To ensure the best model with the most important features are retained, we perform exploratory data analyses and reduce the dimensionality of the data to form an appropriate model, quick in its run-time.

The paper is arranged in eight sections. In section 2 we describe our dataset, followed by preparing and preprocessing the data in section 3. In section 4 we conduct a literature review on already employed techniques in this field with section 5 describing the used approach and evaluating the model. In section 6 & 7, we interpret our results and conclude our research in section 8.

II. Dataset

A. Introduction

The dataset is a public dataset available on the UCI Machine Learning repository (Archive.ics.uci.edu., 2018). The dataset contains 1043 instances of student data for the two courses – Mathematics and Portuguese. Our target variable to be analyzed is the categorical variable final grade – G3.

B. Metadata

The dataset contains 34 features as described –

- school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
- sex - student's sex (binary: 'F' - female or 'M' - male)
- age - student's age (numeric: from 15 to 22)
- address - student's home address type (binary: 'U' - urban or 'R' - rural)
- famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
- Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
- Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)
- Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)
- Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
- Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')

- reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
- guardian - student's guardian (nominal: 'mother', 'father' or 'other')
- 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)
- studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
- failures - number of past class failures (numeric: n if 1<=n<3, else 4)
- schoolsup - extra educational support (binary: yes or no)
- famsup - family educational support (binary: yes or no)
- paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
- activities - extra-curricular activities (binary: yes or no)
- nursery - attended nursery school (binary: yes or no)
- higher - wants to take higher education (binary: yes or no)
- internet - Internet access at home (binary: yes or no)
- romantic - with a romantic relationship (binary: yes or no)
- famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
- freetime - free time after school (numeric: from 1 - very low to 5 - very high)
- goout - going out with friends (numeric: from 1 - very low to 5 - very high)
- Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
- Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
- health - current health status (numeric: from 1 - very bad to 5 - very good)
- absences - number of school absences (numeric: from 0 to 93)
- G1 - first period grade (numeric: from 0 to 20)
- G2 - second period grade (numeric: from 0 to 20)
- *Target variable*
  G3.grade – final grade (categorical with 4 classes – fail, average, satisfactory and excellent)

III. Data Preparation

Before we begin our research, it is imperative to understand the trends in our data to select a correct model. We also wish to remove the unnecessary features so that our final model only consists of the relevant factors that influence our dependent variable – the final grade of the student.

A. Data types

We start by ensuring that all categorical variables that appear as numeric data are converted to factors. Our dependent variable is a continuous variable with values ranging from 0 to 20. We convert this variable in to categories as we are interested in the weak students ie. fail. The classification has been done by assigning the following class based on the student's marks : 0-8 Fail, 9-12 Average, 13-16 Satisfactory and 17-20 Excellent.

B. Correlation amonst features

Our initial exploratory data analysis is the correlation matrix. We check for independent variables with high correlation and remove them. This is done as the variance in the target variable can be explained by any one of the correlated feature.

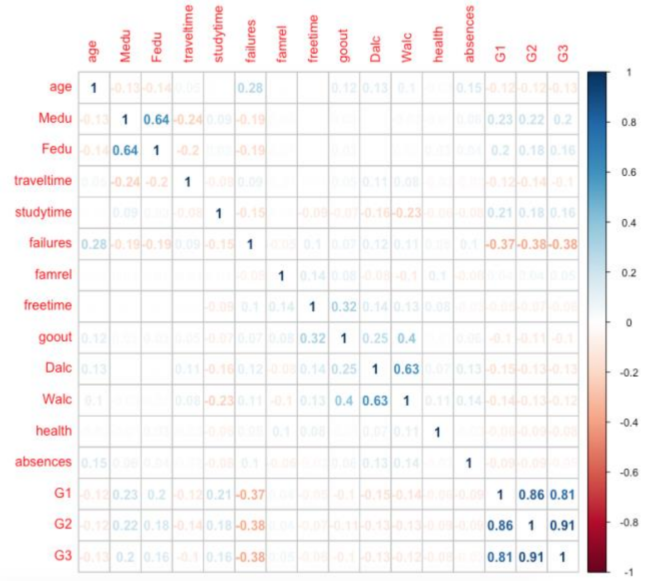


Fig. 1 Correlation Plot for Independent Variables

We notice a high correlation between 3 sets of variables – Parents education (Medu & Fedu), alcohol consumption (Dalc & Walc) and grades in the first & second period (G1 & G2).

We choose one variable in each case at random and remove the following features from our dataset due to its high correlation – Mother's education (Medu) and alcohol consumption on weekends (Walc). We need to remove one of G1 and G2 but since they are the grades in the previous periods of the course, it might be crucial to understand the more important feature of the two.

C. Rank of Important Features

Our next analyses is to rank the features of our dataset in terms of importance. This enables in better understanding the factors that are affecting our target variable – the final grade outcome of the student. This ranking will help us later on in selecting only the relevant features for our model.

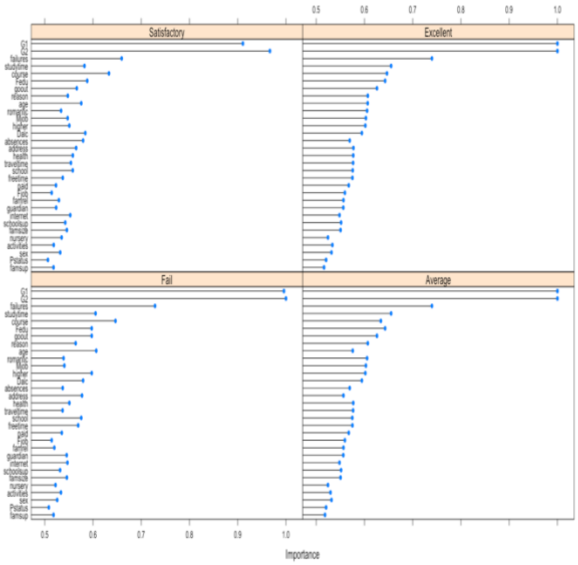


Fig. 2 Rank of Important Features

From the above figure we are able to see that some features contribute more to the prediction of outcome class. Now that we know which are the important features it would be a good practice to remove the unrequired variables. It can be seen that G2 grade is more significant than the G1 grade. From our findings in the previous section, we know that the 2 variables are highly correlated. Hence we can drop G1 grade. But for comparative purposes, we will create a machine learning model with and without G1. This will help identify the best case for explaining our target variable.

However it would not be wise to only remove the features on the basis of above. Hence we further our selection by performing another feature shrinkage algorithm.

D. Recursive Feature Elimination

Since all features do not contribute to the target variable, we are now going to select the relevant features to build our model. We run a recursive feature algorithm on our data, which builds *n* models for the *n* independent variables by adding one feature in each iteration. For each size of the model, it computes all the possibilities and returns the best performing model (Brownlee, 2018).

It is observable from Fig. 3 that all models, whether it has 2 features or all 33, give an accuracy ranging from 83% to 85.5%. It is also observable that a model containing only 8 relevant features will give the highest accuracy. Of these 8, we will only select the top 5 important features as discussed in the above subsection. These top 5 features are G2, failures, course, absences and higher, which are described in section 1.2.

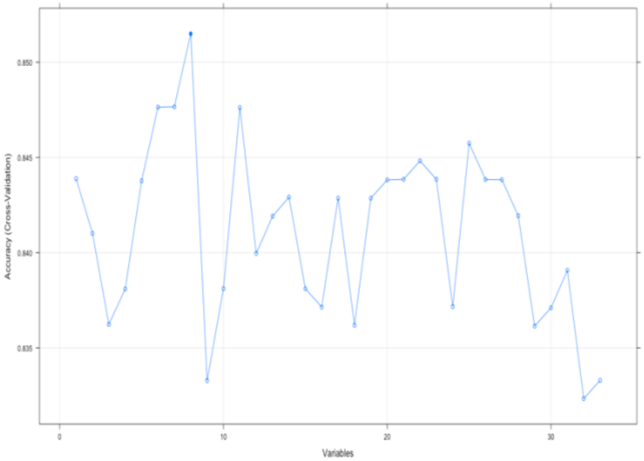


Fig. 3 Recursive Feature Elimination Plot

IV. Research and Investigation in Techniques

Initial research on this data was performed by Cortez et al. (2008). They applied 4 different data mining models and were able to conclude that student achievement can be predicted more accurately when one of the previous grades are known. They also suggested the use of feature elimination as their research model consisted of all the features available.

Previous research in this topic has shown that decision tree tends to outperform other machine learning classification models while analyzing student performance.

Since our dataset is relatively small, it acts as a limitation in applying regression techniques to our data (Iqbal et al., 2017). They were able to achieve a very high model performance using Restricted Boltzmann Machines to predict the final grade of the student, using real-world collected large set of data.

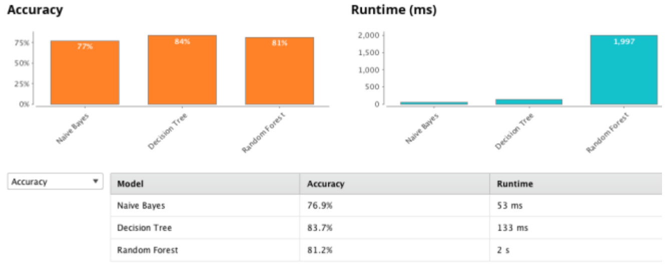
Lakkaraju et al. (2015), in his research found random forests to perform better than decision trees. However his dataset comprised of 200,000 unique students' information. A random forest is able to work on larger datasets with greater number of features. However we aim to develop a model by selecting only the relevant ones.

A decision tree is effective when there are not large number of features and lots of numeric data (Prasad et al., 2006). Hence it is essential to reduce dimensionality before proceeding. In particular the C5.0 algorithm is a highly effective classifier and is known to be very robust to the dataset. It may to tend to over or under fit when the number of classes of the explanatory variable is large. It is also more efficient and less complex than other models and can be used on a relatively small training data (Lantz, 2013).

Hence we will explore decision trees to analyze the problem at hand. We will apply the C5.0 algorithm as it is yet to be implemented on such a dataset. Feature selection method shall also be applied in order to reduce the dimensionality as having redundant features can affect the performance of the model (Preetha et al., 2013). This makes it necessary to implement a feature selection and shrinkage method such that there is no contribution by irrelevant features to our model. This will crucial for our model as previous researchers have not focused on the area of feature selection and elimination.

V. Building The Model

For our classification problem, we have implemented the C5.0 decision tree. Before running this model, we compared the general performance of various models built with all features.



Model	Accuracy	Runtime
Naive Bayes	76.9%	53 ms
Decision Tree	83.7%	133 ms
Random Forest	81.2%	2 s

Fig. 4 Overview of Model Performances

From Fig. 4, we see that decision tree outperforms other classification models when built using all the features in our dataset.

Now that we have narrowed down our model selection, using investigative and comparative means, we run our C5.0 decision tree thrice – each with different set of features.

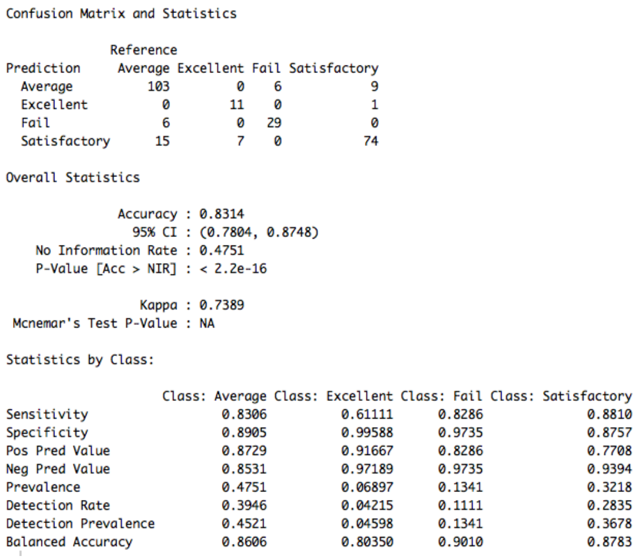
The first model contains all 31 features, whereas for model 2 we drop the variable G1 as it highly correlated to variable G2. In the third C5.0 model, we use only the top 5 features as discussed in section 3.4.

VI. Quantitative Results

A. Model 1 – All features

This C5.0 predictive model consists of all the features. The only features dropped from this model are the correlated variables.

The model accuracy is ***83.14%***.



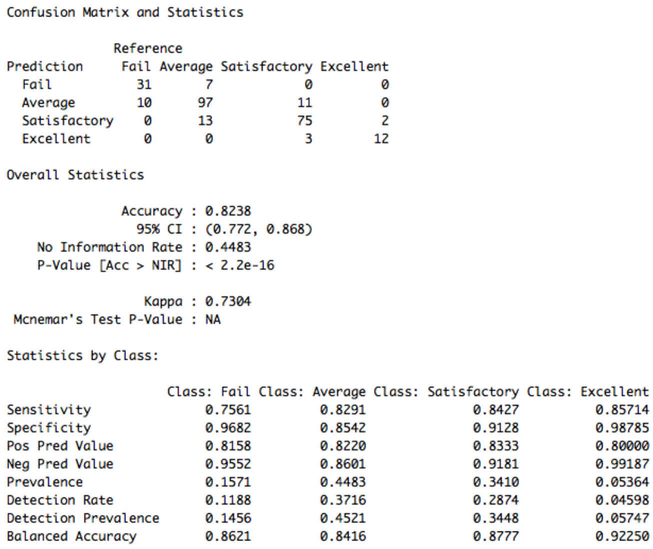
```
Confusion Matrix and Statistics

              Reference
Prediction     Average Excellent Fail Satisfactory
  Average          103         0    6            9
  Excellent          0        11    0            1
  Fail               6         0   29            0
  Satisfactory      15         7    0           74

Overall Statistics

               Accuracy : 0.8314
                 95% CI : (0.7804, 0.8748)
    No Information Rate : 0.4751
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7389
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Average Class: Excellent Class: Fail Class: Satisfactory
Sensitivity                  0.8306          0.61111      0.8286              0.8810
Specificity                  0.8905          0.99588      0.9735              0.8757
Pos Pred Value               0.8729          0.91667      0.8286              0.7708
Neg Pred Value               0.8531          0.97189      0.9735              0.9394
Prevalence                   0.4751          0.06897      0.1341              0.3218
Detection Rate               0.3946          0.04215      0.1111              0.2835
Detection Prevalence         0.4521          0.04598      0.1341              0.3678
Balanced Accuracy            0.8606          0.80350      0.9010              0.8783
```

Fig. 5 Confusion Matrix – Model 1

B. Model 2 – Dropping G1

Since G1 & G2 are highly correlated, we have chosen to drop G1. This model contains all other 30 features used to predict the outcome of the target variable.

The model accuracy reduces to only ***82.38%.***



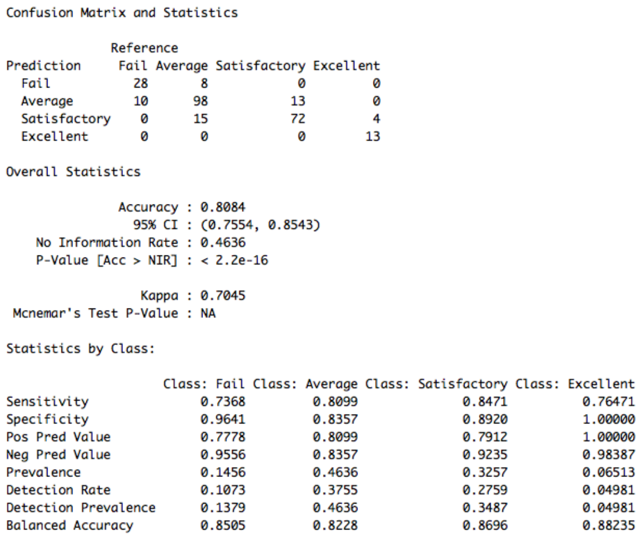
```
Confusion Matrix and Statistics

              Reference
Prediction     Fail Average Satisfactory Excellent
  Fail           31       7            0         0
  Average        10      97           11         0
  Satisfactory    0      13           75         2
  Excellent       0       0            3        12

Overall Statistics

               Accuracy : 0.8238
                 95% CI : (0.772, 0.868)
    No Information Rate : 0.4483
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7304
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Fail Class: Average Class: Satisfactory Class: Excellent
Sensitivity               0.7561         0.8291              0.8427          0.85714
Specificity               0.9682         0.8542              0.9128          0.98785
Pos Pred Value            0.8158         0.8220              0.8333          0.80000
Neg Pred Value            0.9552         0.8601              0.9181          0.99187
Prevalence                0.1571         0.4483              0.3410          0.05364
Detection Rate            0.1188         0.3716              0.2874          0.04598
Detection Prevalence      0.1456         0.4521              0.3448          0.05747
Balanced Accuracy         0.8621         0.8416              0.8777          0.92250
```

Fig. 6 Confusion Matrix - Model 2

C. Model 3 – Top 5 Features only

For our final C5.0 decision tree, we have only eliminated certain features and selected the top five features on the ranking of importance to our model. These features are the grade in the second period, number of previous failures in this course, the course itself, number of school absences and the wish to pursue higher studies.

The model accuracy reduced marginally again to ***80.84%.***



```
Confusion Matrix and Statistics

               Reference
Prediction     Fail Average Satisfactory Excellent
  Fail           28       8            0         0
  Average        10      98           13         0
  Satisfactory    0      15           72         4
  Excellent       0       0            0        13

Overall Statistics

               Accuracy : 0.8084
                 95% CI : (0.7554, 0.8543)
    No Information Rate : 0.4636
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7045
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Fail Class: Average Class: Satisfactory Class: Excellent
Sensitivity               0.7368         0.8099              0.8471          0.76471
Specificity               0.9641         0.8357              0.8920          1.00000
Pos Pred Value            0.7778         0.8099              0.7912          1.00000
Neg Pred Value            0.9556         0.8357              0.9235          0.98387
Prevalence                0.1456         0.4636              0.3257          0.06513
Detection Rate            0.1073         0.3755              0.2759          0.04981
Detection Prevalence      0.1379         0.4636              0.3487          0.04981
Balanced Accuracy         0.8505         0.8228              0.8696          0.88235
```

Fig. 7 Confusion Matrix – Model 3

As observable from Fig. 8 below, reduction in features only affects model performance marginally. Hence, in terms of computation and complexity of decision tree, our model with the top 5 features will satisfy the business needs.

Model	Feature Count	Model Accuracy	Description
1	31	83.14%	Model with all features
2	30	82.38%	Feature G1 dropped from the first model
3	5	80.84%	Only top 5 features retained in the model

Fig. 8 Model Performances

VII. Qualitative Results

We ran 3 models with different counts of features each time using the classification technique of C5.0 decision tree. Our model with even 5 features performed well with a 80.84% accuracy. Hence even this model is able to correctly classify our variable.

Therefore the variance in the target variable can be explained by the 5 features we selected using feature elimination.

A student's final grade is affected by 2 aspects – *classroom and academic factor* which include his previous grade (G2), number of times the student has failed the module, course enrolled in, number of times absent in class; and secondly the student's *personal ambition* to attain higher education.

Majority of the variance in our target variable is explain by the top 5 variables. A decision tree formed using just these 5 variables is a good predictor whether a student will fail the particular course or not. Through these variables, the school teachers, management board and other stakeholders know what areas to evaluate and assess in order to find the students more likely to fail a grade. Analyzing simple factors like absences, previous attempts and failures etc. can help the school to improve their student passing rate. This will put school resources at optimal use.

VIII. Conclusion

Our decision tree model developed using the C5.0 algorithm was effective in correctly classifying 80.84% instances while only considering 5 features of the dataset. This intelligence will help the school to take measures in order to ensure that the at-risk students get attention and are able to cope up.

We suggest an installation of a learning management system in the school to track and maintain their records instead of manually collecting data. The LMS will also act as an early warning system for the school teacher's, in highlighting cases where students may fail their final grade of the course. They can now help these students through extra classes, special assignments or social/psychological sessions. This will improve the quality of education through adequate attention to all students. In turn, this will also increase student retention.

Future scope in this field would involve collecting data from multiple cohorts from multiple schools for a particular district, state and eventually country (Lakkaraju et al., 2015). This would help to ensure an unbiased and large dataset. There is also a need to understand the social and psychological factors that may affect a student's grade and how can they be ingrained in our model.

Other machine learning models should also be used to assess the performance of a student through their grades. There has already been research conducted in Support Vector Machines, Decision Trees, Random Forests and Neural Networks. However the area of deep learning and applying it to analyze student grades are untapped. Perhaps our model obtained from applying C5.0 decision tree using Recursive Feature Elimination can be improved.

References

[1] Archive.ics.uci.edu. (2018). *UCI Machine Learning Repository: Student Performance Data Set*. [online] Available at: https://archive.ics.uci.edu/ml/datasets/student+performance [Accessed 14 Jul. 2018].

[2] Brownlee, J. (2018). *Feature Selection with the Caret R Package*. [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/feature-selection-with-the-caret-r-package/ [Accessed 21 Jul. 2018].

[3] Cortez, P., & Silva, A. (2008). Using data mining to predict secondary school student performance. In the *Proceedings of 5th Annual Future Business Technology Conference*, Porto, Portugal, 5-12.

[4] Iqbal, Z., Qadir, J., Mian, A. and Kamiran, F. (2017). Machine Learning Based Student Grade Prediction: A Case Study.

[5] Lakkaraju, H., Aguiar, E., Shan, C., Miller, D., Bhanpuri, N., Ghani, R. and Addison, K. (2015). A Machine Learning Framework to Identify Students at Risk of Adverse Academic Outcomes. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15*.

[6] Lantz, B. (2013). Machine learning with R. Birmingham: Packt Publishing.

[7] Prasad, A., Iverson, L. and Liaw, A. (2006). Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction. *Ecosystems*, 9(2), pp.181-199.

[8] Preetha Evangeline, D., Sandhiya, C., Anandhakumar, P., Deepti Raj, G. and Rajendran, T. (2013). Feature subset selection for irrelevant data removal using Decision Tree Algorithm. *2013 Fifth International Conference on Advanced Computing (ICoAC)*.