Data Mining on Canberra Education System Report

Hao Wen

Introduction

Context and Problem Description

In most countries and education systems, final score and GPA is an important measure of student academic achievement. It helps university and potential employer to compare and select candidates with higher final score and GPA. One of the major task of college is to find the factors that affect final score and can be used to predict result, finally help student to achieve the best result.

In Canberra education system, there are two groups of college student and they chose different pathways after college in year 12. The first group is called accredited student. Student from this group is not seeking a tertiary entrance and there is no score but only GPA. Another group is the student who is looking for tertiary entrance after college, they have final score and GPA. For each student, many factors may have contribution on the final score and GPA, including gender, previous school score, NAPLAN test score, first math and English score, STEM score. In addition, NAPLAN test score happens in year 9 and includes Numeracy, Reading, Writing, Spelling and Grammar. This test score is considered to be the most useful predictor on final score. However it is not available for all student due to many reasons, including coming from non-Australian school, illness and parent decline.

Our main goal is using data mining methods to exploring the relationship between final score and other variables and predicting the final score. We already known female has a better performance than male because they tends to make the optimal choice among STEM subjects and non-STEM subjects and maximum GPA. Also, We don't care the effect on the college or final year as we prefer a systemic longer term patterns. So, in this report, we have the following interesting fields we want to explore. Some students have underperformance because their score is below the large amount of average score. The first goal is to select variables which have significant effect on underperformance and classify the underperformance. The second goal is to recommend student the suitable subject, STEM subjects or non-STEM subjects, to maximize their final score based on NAPLAN score, first math score and first English score. The third goal is to explore if there is a significant effect of gender and previous school score on the final score.

These goals are very important and have further impacts on personal, college and society level. Individual student can maximize the final score and get a high GPA by using this result. High GPA will provide opportunity to good university or job interview, which leads to a great development. Also, good score and GPA help student build confidence and reduce pressure. College can use these results to distinguish students and recommend them the most appropriate subject, finally improve the overall performance. Eventually all these benefits form individual and college will contribute the social wellbeing by increase the stability and wealth.

Data Description

This data comes from three different colleges in Year 11 and 12 at ACT, which includes student personal information and academic performance. There are 5641 students and 220 variables.

Among the 220 variables, there are three different types, nominal variable, Numerical ratio scaled variable and Ordinal variable.

Nominal variable includes Gender, College attended, Final year.

Numerical ratio scaled variable includes previous school, Numeracy NAPLAN, Reading NAPLAN, Writing NAPLAN, Spelling NAPLAN, Grammar NAPLAN, First math score, First English score, STEM to total ratio, Final average STEM score, and Final average score.

Ordinal variable includes First math grade, First English grade, Final average STEM GPA, Average GPA, all subject grade and total grades received.

In this report, I mainly discuss the variables excluding the subject variables. Final average score and average GPA are main response (dependent) variables. Other variables are independent variable.

Final average STEM score and Final average STEM GPA are potential response variable. But in most case, I treat them as independent variable because I want to explore the relationship between STEM and final score. The population for this sample is all college student in Year 11 and Year 12 at ACT.

Data quality

There are six basic dimensions for data quality including completeness, consistency, uniqueness, validity, accuracy and timeliness. For this assignment, I mainly talk about completeness and I assume all data are valid and accurate.

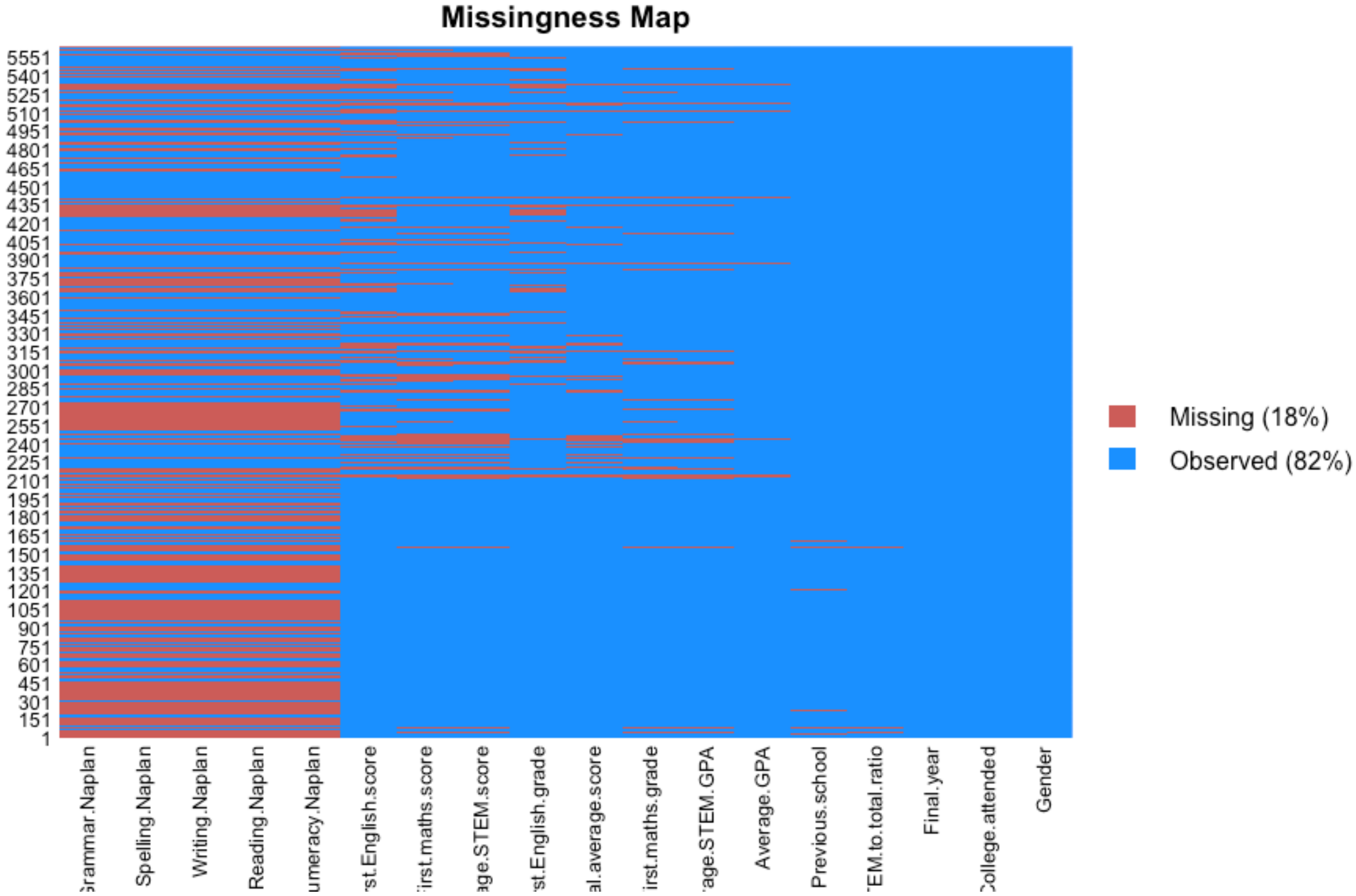
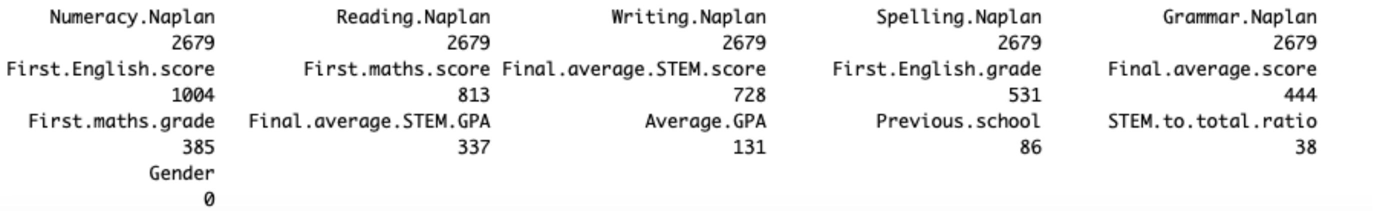


Figure 1



```
     Numeracy.Naplan          Reading.Naplan           Writing.Naplan          Spelling.Naplan          Grammar.Naplan
                2679                    2679                     2679                     2679                    2679
  First.English.score        First.maths.score Final.average.STEM.score      First.English.grade     Final.average.score
                1004                     813                      728                      531                     444
    First.maths.grade   Final.average.STEM.GPA              Average.GPA          Previous.school       STEM.to.total.ratio
                 385                     337                      131                       86                      38
              Gender
                   0
```

(this table shows number of missing value in each variable in decreasing order)

The missingness map(Figure 1) and table shows NAPLAN test(Numeracy, Reading, Writing, Spelling, Grammar) have the highest number of missing values, which is 2679 and account for 47.49% of total observations. Also, for other variable such as First English Score or First Math score, there are around 1000 missing values. Large number of missing value is a serious problem in this dataset and I will deal with it in later section.

Exploratory data analysis

Figure 2 shows the distribution of each independent variable (The basic statistical summary is in Appendix 8). Number of male(2809) and female(2815) student is nearly same. Only small number of student have no gender or gender X. Majority student comes from college 2(4553), total student number from college 1 and 3 is 1088. Most student have final year at 2016(2226). For the most numeric variables, the distribution is symmetric and looks like normal distribution except the STEM to total ratio. STEM to total ratio has a left skewed distribution which means most student have STEM ratio lower than the mean level (mean is 0.3, median is 0.25). Also, for NAPLAN test score, some of them are zero which doesn't make sense. Distribution of previous school score shows there are two cluster with split boundary 50.

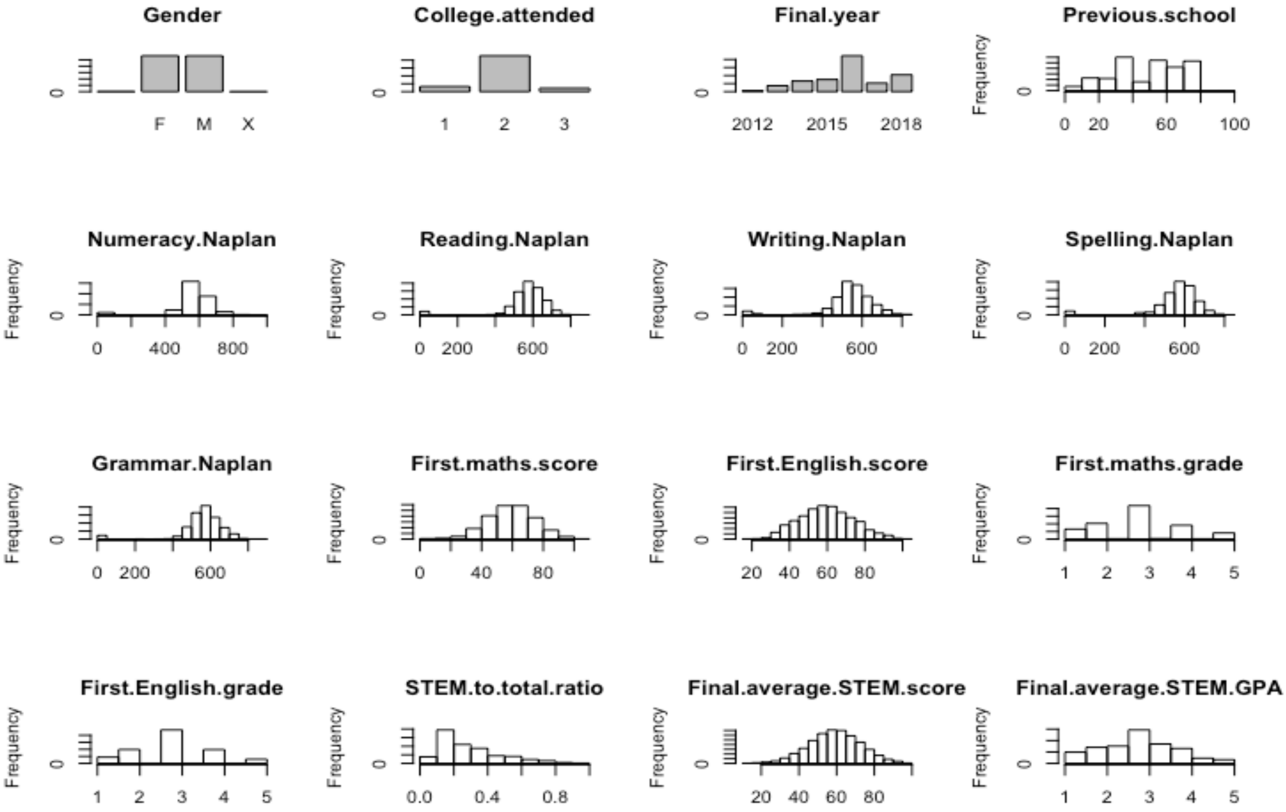


Figure 2

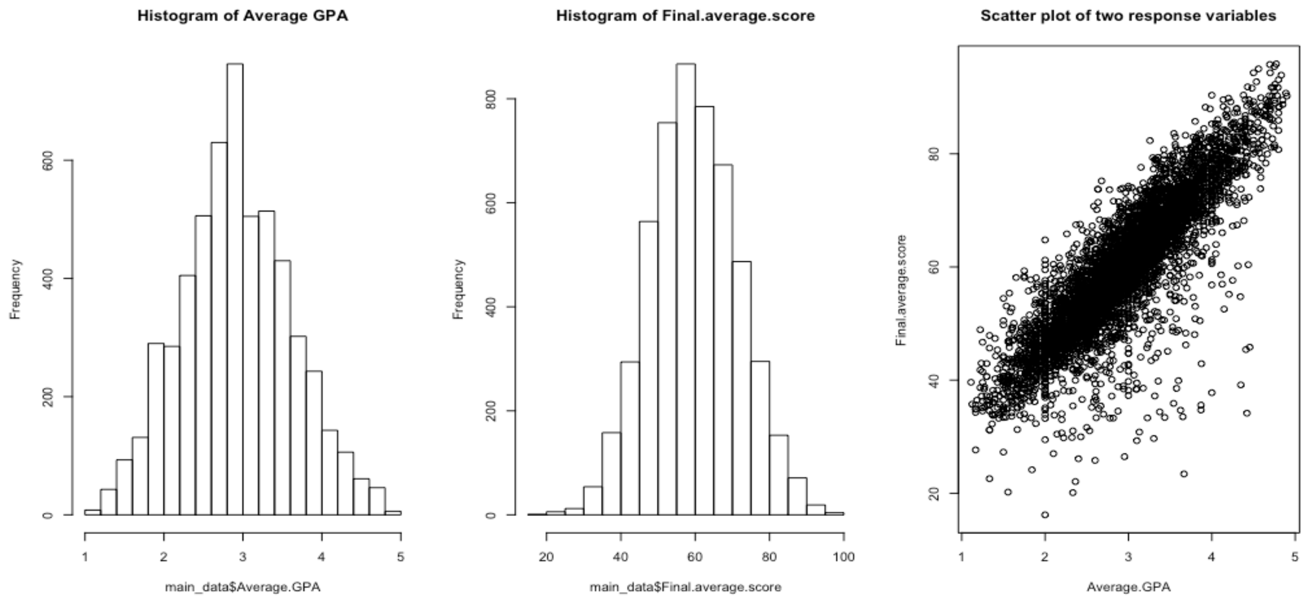


Figure 3

Figure 3 shows the distribution of two response variables are all normal distribution. Most students have Average GPA between 2 to 4 and Final average score between 40 to 80. Also there is a significant positive relationship between Average GPA and Final average score which means higher Average GPA is always associated with higher Final average score.

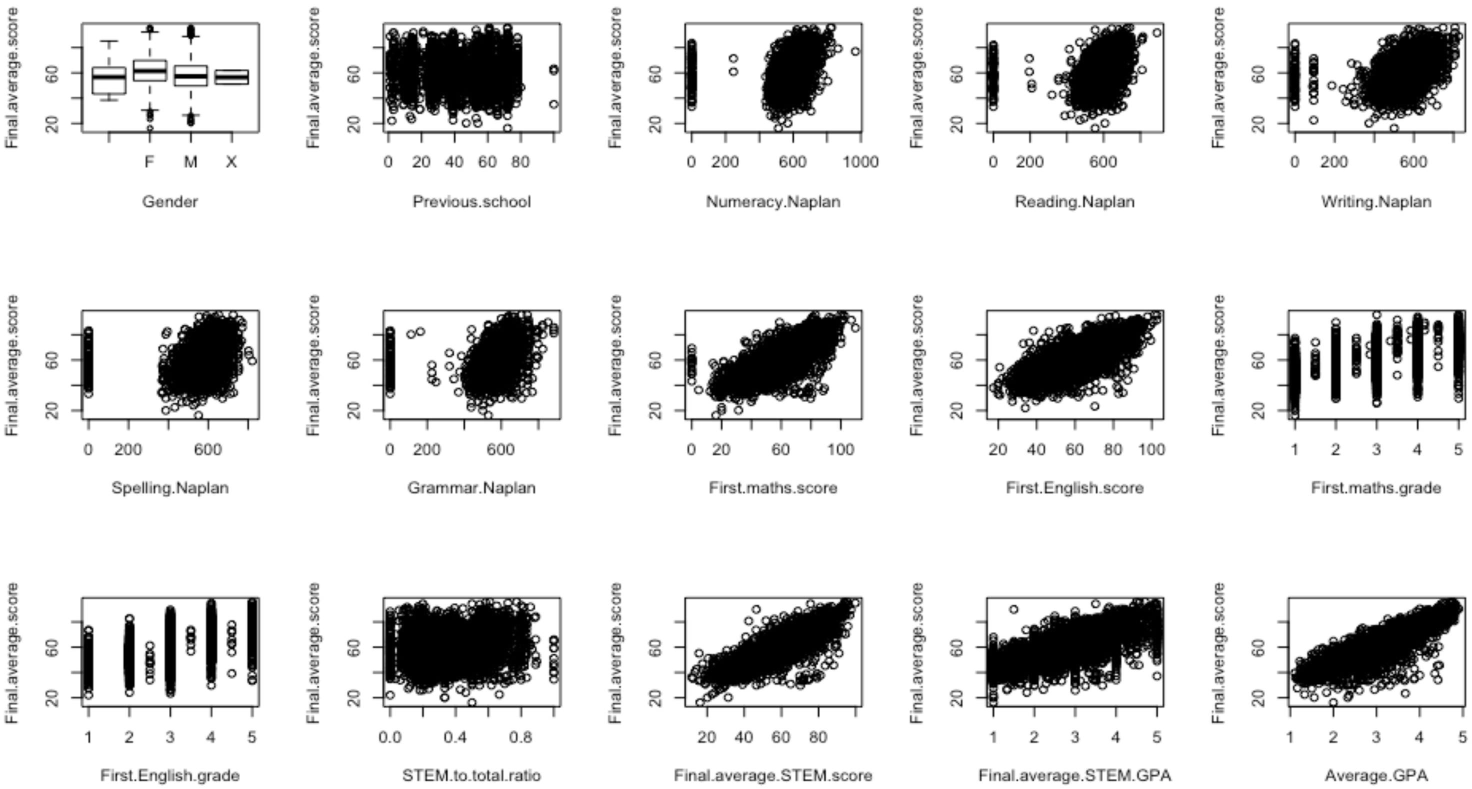


Figure 3

Figure 4 shows the relationship between independent variables and Final average score. First math score, First English score, Final average STEM score and Final average STEM GPA, they all have a positive relationship with Final average score. On the other hand, Previous school, NAPLAN test score, STEM to total ratio, the relationship with Final average score seems not strong and even no relationship. In the later section, I will further explore the relationship and effect of these variables.

Method Description

In this course, we learn many methods for classification and regression(numeric prediction). In this report, I mainly use four methods including Decision tree, Support Vector Machine(SVM) , Linear Regression and Neural Network. The main reasons are Decision tree and SVM are suitable for classification problem and can be easily implemented on Rattle. Also, Linear Regression and Neural Network are suitable for regression problem and can be easily implemented on R.

Data pre-processing

In order to implement the method and achieve a reliable and reasonable result, high quality of input data is necessary. For this data, I remove the students with gender X and unknown gender because there are only 17 students. Also, I treat all zero NAPLAN test score as missing value because it doesn’t make sense according to Q&A from data information document. As discussed before, missing value is a serious problem in this data. Nearly half observation have missing value in NAPLAN score test, so it is not a good idea to impute missing value in NAPLAN score.

Based on the Final average score and NAPLAN test score, I divide student into four groups, students have Final average score and NAPLAN test score(2581 observations), students have Final average score but no NAPLAN test score(2482 observations), students only have Average GPA and NAPLAN test score(154 observations), students have Average GPA but no NAPLAN test score(135 observations). Basic statistical summary for four student group is in Appendix 1. After the grouping, the number of missing value is reduced significantly in each group (NA number is in Basic statistical summary). Now, I need to impute missing value. The basic idea of imputation is very simple. For the variable has normal distribution, I use mean to impute and for variable has skewed distribution, I use median to impute. The result of imputation in Appendix 1 as well. Compare the statistics such as mean and median, the imputation result is good and there is no significant change over the data distribution. Data is all clean, complete and prepared for further analysis.

1.Numerical prediction

First of all, I want to explore what factors influence the final score and GPA in four groups and use these factors to predict final score and GPA. For accredited student, the response variable is Average GPA. For non-accredited (tertiary pathway) student, the response variable is Final average score. Because the sample size is large, I will treat Average GPA as continuous variable for the convenience of building model.

Linear model process

For each data:

1. Divide data into training data(75%) and test data(25%).
2. Using training data to build linear model with all variables.
3. Based on 5% significance level and t-test, select the variables have significant relationship with response variable, and build another linear model only with selected variables. This is the chosen model for this data. Selection detail is in Appendix 2.
4. Use the chosen linear model to predict the test data and calculate mean absolute error(MAE).

Neural network process

For each data:

1. Transfer all value into numeric value and normalize it into 0 to 1.

2. Divide data into training data(75%) and test data(25%). Use the same index as linear model to split data to make sure both methods have same training and testing data.
3. Select network topology and build neural network with the same selected variables from linear model. (In this case, parameter setting is (5,3), which means there are two hidden layers, the first one has 5 units and the second one has 3 units. Actually, after trying different hidden layers and units in each layer, I found the MAE is relative stable and doesn't change a lot. So here the parameter is random number. Also, the activation function is logistic function because it's the widely used function and behave well in most case.)
4. Use the neural network model to predict the same test data, transfer back to original scale, and calculate mean absolute error(MAE).

2.Underperformance Classification

In order to predict the underperformance, I create a categorical variable to indicate if the student has underperformance. I assume student have final score or GPA that is lower than 25% quantile in each group is regarded as underperformance. Build Decision tree and SVM for each group.

Criterion of good model

The criterion for classification method is overall error and ROC-AUC. I prefer the high ROC-AUC, and if the ROC-AUC are same, prefer the low overall error. This is because overall error can be misleading sometimes for unbalanced response variable. For underperformance classification, the response variable is unbalanced in each group (Appendix 9). ROC and AUC can deal with this issue by counting performance over true positive and false positive.

Decision tree process

1. divide data into training data(70%), validation data(15%) and testing data(15%).
2. Set parameter, min_split, min_bucket , max_depth and complexity. Build decision tree on training data with different settings and select appropriate parameter based on performance of validation data. After trying different parameter, I found the default setting of min_split=20, min_bucket=7 and complexity=0.01 usually have good performance (low overall error and high ROC-AUC) on validation data. Also, I set max_depth euqals 30 to make the tree as large as possible and hence improve prediction accuracy.
3. Use decision tree model to predict the test data. Calculate overall error from confusion matrix and plot ROC.

SVM process

For the same data, build SVM on training data and select appropriate kernel according to overall error and ROC-AUC on validation data. Kernel selection process for each group is in Appendix 5.

3.STEM subject suitability Classification

In order to classify student who is suitable for STEM subject, I create a categorical variable to indicate if the student is good at STEM subject. I assume the student have Final average STEM score or GPA that is higher than mean value in each group is regarded as suitable for STEM subject. The logic is there is positive relationship between STEM score and Final score. If a student is good at STEM subject, then he will have a good result on STEM score and hence a good Final score. Similarly, I use decision tree and SVM to make classification. The process is totally same as before.

Method justification and limitation

Basically, the methods are good and meet the purpose of goals regarding to prediction and classification. (1) Linear model can be used to interpret the linear relationship and make prediction. (2) Neural Network can make prediction and deal with all kinds of relationship (not only linear relationship) by different hidden layers and nodes connection, also it has high tolerance to noisy data. (3) Decision tree can make classification and generate rules which can be interpreted and applied. (4)

SVM can make classification and deal with linearly inseparable data by selecting kernel and improve the classification performance. (5) All methods are effective on high-dimensional data.

However, there are some limitations. (1) Some result is unstable and hard to explain the relationship. For example, variable selection based on linear model is not stable and results rely on training data. Given different training data, the significant variable may change each time. Neural Network and SVM is black box. We can only use them to make prediction and don't know the logic or relationship. (2) Not all methods have good performance on all student groups. Bad model makes result unreliable and meaningless. For example, the linear model on the third and fourth group is not good, R^2 is less then 60%. Also, Decision tree and SVM have ROC-AUC less than 0.6 on some groups. (3). For Neural Network and Decision Tree, tuning parameter is a big issue. It is time consuming to try the different combination of parameter and inappropriate parameter will mislead the result. (4) It's hard to detect overfitting problem. For example, we want decision tree as small as possible and have accurate prediction at the same time. But its difficult to find a prefect tree.

Results

Numerical prediction of Linear model and Neural Network

Student Group	Selected variable	MAE on testing data	
		Linear model	Neural network
Have Final average score and NAPLAN	Gender(-), Writing.Naplan(+), First.maths.score(-), First.English.score(+), Final.average.STEM.score(+)	4.00794	3.89
Have Final average score but no NAPLAN	Gender(-), First.maths.score(+), First.English.score(+), Final.average.STEM.score(+)	4.41069	4.25
Have Average GPA and NAPLAN	First.English.grade(+), Final.average.STEM.GPA(+)	0.3280721	0.3215
Have Average GPA but no NAPLAN	First.English.grade(+), Final.average.STEM.GPA(+)	0.3164327	0.3387

The selected variable means the variable has significant relationship with response variable (Symbol + means positive relationship, - means negative relationship). For non-accredited(have final score) student, Writing NAPLAN, First math and English score, final average STEM score has positive relationship with Final average score. Also, negative coefficient of Gender means female has better performance than male. An abnormal pattern is there is negative relationship between First math score, the reason is multicollinearity of Final average stem score with other independent variables. If I remove STEM variable, Gender, Numeracy Naplan, Reading Naplan, Writing Naplan, First maths score, First English score all are significant and have positive relationship with Final score. For accredited student, only First English grade and Final average STEM GPA has positive relationship with Average GPA. If I remove STEM variable, First math and English grade are both significant.

MAE of Neural Network is little lower than linear model except for last group, but basically there is no big difference in terms of prediction accuracy. The plot of predicted test value and true test value also shows no big difference (Appendix 6).

Underperformance Classification on Decision tree and SVM

(Decision tree Result and ROC plot for all groups is provided in appendix 3.)

Student Group	Selected variable (in order)	AUC on testing data	
		Decision Tree	SVM
Have Final average score and NAPLAN	Final.average.STEM.score >First.English.score>STEM.to.total.ratio>Numeracy.Naplan	0.9	0.92
Have Final average score but no NAPLAN	Final.average.STEM.score>First.English.score> Previous school>STEM.to.total.ratio> Gender	0.84	0.91
Have Average GPA and NAPLAN	First.math.grade>First.English.grade	0.61	0.51
Have Average GPA but no NAPLAN	Final.average.STEM.GPA	0.72	0.61

The table shows for tertiary pathway student, Final.average.STEM.score, First.English.score, STEM.to.total.ratio contribute significant information in Final average score. Numeracy.Naplan and Gender have some association with Final average score but it is not strong. If I remove all STEM variable, variables are First math score>First English score>writing Naplan>Reading Naplan. For accredited student, NAPLAN test score has no association with Average GPA. First.math.grade, First.English.grade and Final.average.STEM.GPA are useful predictors for Average GPA. If I remove all STEM variable, variabls are First.math.grade>First.English.grade> Numeracy Naplan (The result is in Appendix 10).

In terms of classification performance, decision tree and SVM are good on first two group and SVM has a higher AUC than Decision Tree. However, both models are not good on the last two groups.

Student Group	Selected variable	AUC on testing data	
		Decision Tree	SVM
Have Final average score and NAPLAN	Numeracy.Naplan	0.68	0.77
Have Final average score but no NAPLAN	First.math.score	0.84	0.9
Have Average GPA and NAPLAN	Numeracy.Naplan,Grammer.Naplan, Reading.Naplan,Spelling.Naplan	0.49	0.57
Have Average GPA but no NAPLAN	First.math.grade	0.9	0.81

Thus, the result for last two groups are not reliable and need to interpret it carefully.

STEM subject suitability Classification on Decision tree and SVM

(Decision tree Result and ROC plot for all groups is provided in appendix 4.)

The table shows for tertiary pathway student, Numeracy.Naplan and First.math.score is the useful predictor to predict suitability of STEM subject. For accredited student, Numeracy.Naplan, Grammer.Naplan, Reading.Naplan, Spelling.Naplan and First.math.grade are useful to predict the suitability of STEM subject.

In terms of classification performance, decision tree and SVM are good on the second and the fourth group. However, both models are not good on the first and the third group because ROC-AUC is low. Thus, we need to interpret the result very carefully.

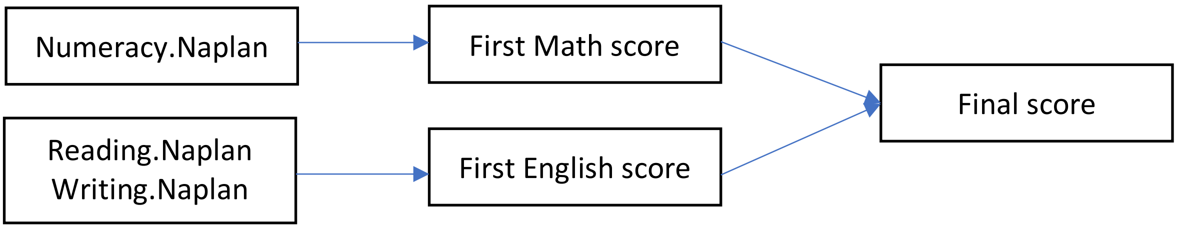
Results summary

Results mentioned above provide some interesting pattern to discuss. However, there is an issue about STEM variables. In reality, we cannot use these variables to predict because they happen at the end of college. So I will ignore the impact of STEM and mainly discuss other variables. In addition, some decision tree rule are complex and tedious, I will not discuss the rule details here. All detail is in Appendix.

For tertiary pathway student, Numeracy, Reading, Writing NAPLAN test, First math and first English score is very useful to predict final score. And they have positive relationship. If there is no NAPLAN score, math score and English score can predict final result. First math score and English score is useful and powerful to classify underperformance. Also, Writing and Reading NAPLAN score is helpful for underperformance classification. Only first math score is useful to classify suitability of STEM subject. NAPLAN score is not reliable for this classification because the model is not good. According to decision tree, if first math score is greater than 59, then STEM subject is suitable for this student and he will achieve a good result on final score.

For accredited student, linear model and decision tree shows First English grade and First Math grade can predict Average GPA and classify underperformance. However, the model is not good and the result has less meaning in real life. For suitability of STEM subject classification, First math grade is useful to predict. According to decision tree, if first math grade is greater than 2.5, then STEM subject is suitable for this student and he will achieve a good result on Final GPA. This result is consistent with plot in Appendix 7. For student is good at STEM subject, most math score is greater than 59 or math grade is greater than 2.5.

Another interesting pattern is for all student group, previous school score and NAPLAN test score have less direct association with their final performance. However, if I use first math score as response variable, Numeracy NAPLAN has significant positive relationship with first math score. Similar for first English score, Reading and Writing NAPLAN have significant positive relationship with first English score. This makes sense and there are many possible reasons. NAPLAN test is in year 9, it have effect on the first math and English score in year 10 or 11, and first math score or English score have impact on final score.



Conclusion and future work

The results can be applied in practice and benefit for individual student and college. Firstly, helping student build confidence. Student don't worry if they have bad result in NAPLAN test or previous school because there is no direct relationship. They have opportunity to change and improve in college and achieve a good final result. Secondly, college can use First English and math score to detect the potential underperformance in tertiary pathway student and find a solution to help them as early as possible. Also, first math score or grade is useful and powerful to help student choose STEM subject or non-STEM subject and maximize final score. If their first math score is greater than

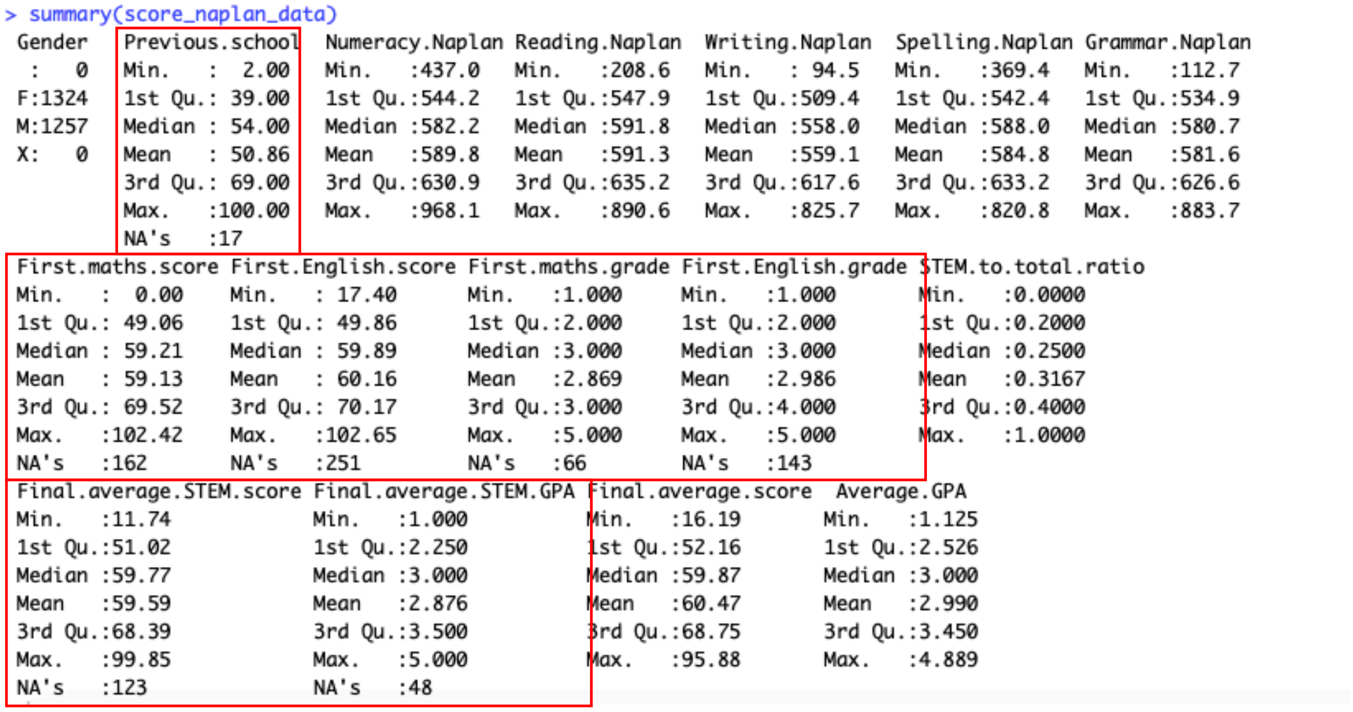
59 or grade is greater than 2.5, then they are suitable for STEM subject. Otherwise they should choose non-STEM subject.

Some challenges are associated with these rules. First of all, some rules are not reliable for certain groups because the model is poor. Especially for accredited student, linear model and decision tree are terrible. Secondly, most student comes from the second college. This may leads the result biased toward student in second college and make it difficult to generalize to other college in ACT. In addition, there are many factors influence final result and hence underperformance. For example, student choose hard course may have low final result but it doesn't means they have problem in intelligence. When deal with underperformance student, college need to be very careful and try to find the real reason that affects underperformance.

There are a lot of works to do in the future to help student and college improve final score. At individual level, overall intelligence, personality, ability to focus may influence the final result. At family level, parent education and income may have potential effect on final result. At college level, teacher quality, class size, teaching facility influence the result potentially as well. We need to collect more data to train our model and make result more accurate and reliable to infer the whole population.

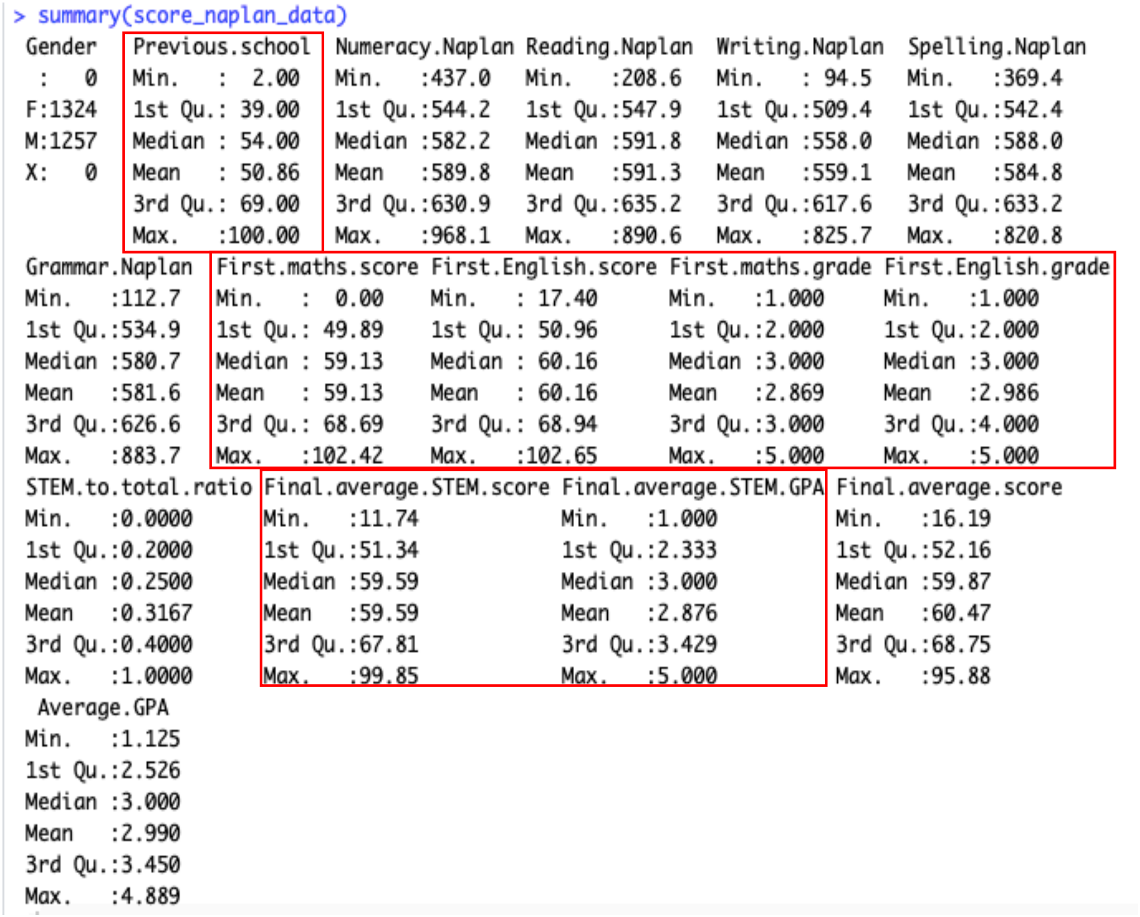
Appendix 1. Basic statistical summary for four groups

1. Score naplan data



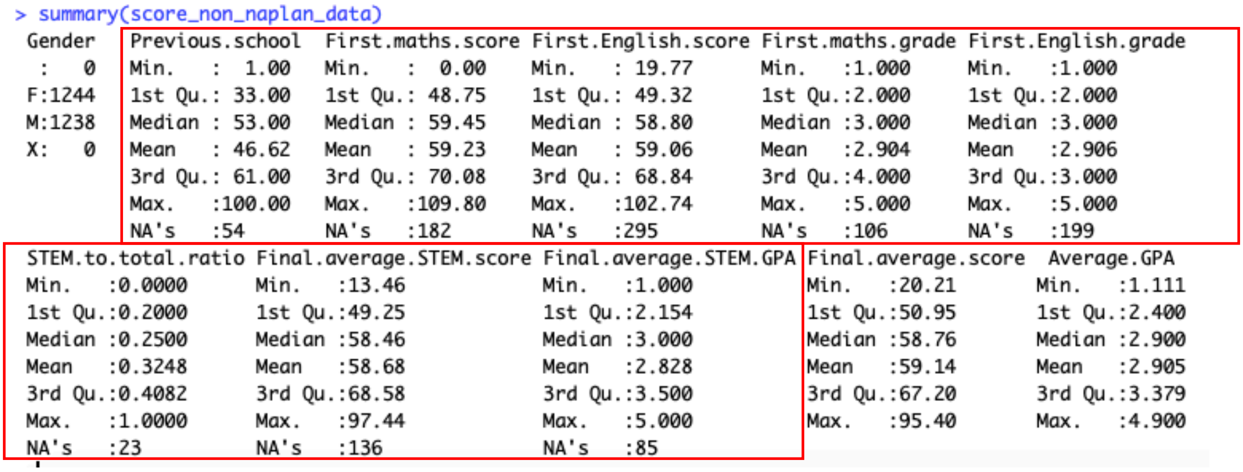
```
> summary(score_naplan_data)
 Gender   Previous.school  Numeracy.Naplan Reading.Naplan  Writing.Naplan  Spelling.Naplan Grammar.Naplan
  :   0   Min.   :  2.00   Min.   :437.0   Min.   :208.6   Min.   : 94.5   Min.   :369.4   Min.   :112.7  
 F:1324   1st Qu.: 39.00   1st Qu.:544.2   1st Qu.:547.9   1st Qu.:509.4   1st Qu.:542.4   1st Qu.:534.9  
 M:1257   Median : 54.00   Median :582.2   Median :591.8   Median :558.0   Median :588.0   Median :580.7  
 X:   0   Mean   : 50.86   Mean   :589.8   Mean   :591.3   Mean   :559.1   Mean   :584.8   Mean   :581.6  
          3rd Qu.: 69.00   3rd Qu.:630.9   3rd Qu.:635.2   3rd Qu.:617.6   3rd Qu.:633.2   3rd Qu.:626.6  
          Max.   :100.00   Max.   :968.1   Max.   :890.6   Max.   :825.7   Max.   :820.8   Max.   :883.7  
          NA's   :17                                                                                      
 First.maths.score First.English.score First.maths.grade First.English.grade STEM.to.total.ratio
 Min.   :  0.00    Min.   : 17.40      Min.   :1.000     Min.   :1.000       Min.   :0.0000     
 1st Qu.: 49.06    1st Qu.: 49.86      1st Qu.:2.000     1st Qu.:2.000       1st Qu.:0.2000     
 Median : 59.21    Median : 59.89      Median :3.000     Median :3.000       Median :0.2500     
 Mean   : 59.13    Mean   : 60.16      Mean   :2.869     Mean   :2.986       Mean   :0.3167     
 3rd Qu.: 69.52    3rd Qu.: 70.17      3rd Qu.:3.000     3rd Qu.:4.000       3rd Qu.:0.4000     
 Max.   :102.42    Max.   :102.65      Max.   :5.000     Max.   :5.000       Max.   :1.0000     
 NA's   :162       NA's   :251         NA's   :66        NA's   :143                            
 Final.average.STEM.score Final.average.STEM.GPA Final.average.score  Average.GPA   
 Min.   :11.74            Min.   :1.000          Min.   :16.19       Min.   :1.125  
 1st Qu.:51.02            1st Qu.:2.250          1st Qu.:52.16       1st Qu.:2.526  
 Median :59.77            Median :3.000          Median :59.87       Median :3.000  
 Mean   :59.59            Mean   :2.876          Mean   :60.47       Mean   :2.990  
 3rd Qu.:68.39            3rd Qu.:3.500          3rd Qu.:68.75       3rd Qu.:3.450  
 Max.   :99.85            Max.   :5.000          Max.   :95.88       Max.   :4.889  
 NA's   :123              NA's   :48                                                
```

After imputation



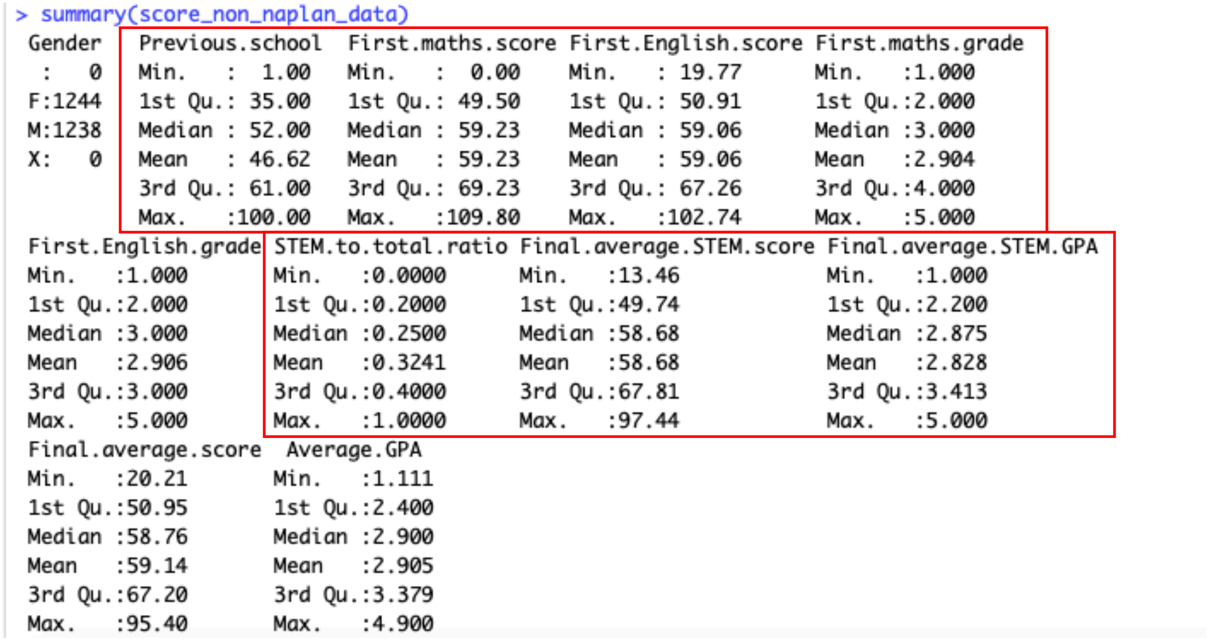
```
> summary(score_naplan_data)
 Gender   Previous.school  Numeracy.Naplan Reading.Naplan  Writing.Naplan  Spelling.Naplan
  :   0   Min.   :  2.00   Min.   :437.0   Min.   :208.6   Min.   : 94.5   Min.   :369.4  
 F:1324   1st Qu.: 39.00   1st Qu.:544.2   1st Qu.:547.9   1st Qu.:509.4   1st Qu.:542.4  
 M:1257   Median : 54.00   Median :582.2   Median :591.8   Median :558.0   Median :588.0  
 X:   0   Mean   : 50.86   Mean   :589.8   Mean   :591.3   Mean   :559.1   Mean   :584.8  
          3rd Qu.: 69.00   3rd Qu.:630.9   3rd Qu.:635.2   3rd Qu.:617.6   3rd Qu.:633.2  
          Max.   :100.00   Max.   :968.1   Max.   :890.6   Max.   :825.7   Max.   :820.8  
 Grammar.Naplan  First.maths.score First.English.score First.maths.grade First.English.grade
 Min.   :112.7   Min.   :  0.00    Min.   : 17.40      Min.   :1.000     Min.   :1.000      
 1st Qu.:534.9   1st Qu.: 49.89    1st Qu.: 50.96      1st Qu.:2.000     1st Qu.:2.000      
 Median :580.7   Median : 59.13    Median : 60.16      Median :3.000     Median :3.000      
 Mean   :581.6   Mean   : 59.13    Mean   : 60.16      Mean   :2.869     Mean   :2.986      
 3rd Qu.:626.6   3rd Qu.: 68.69    3rd Qu.: 68.94      3rd Qu.:3.000     3rd Qu.:4.000      
 Max.   :883.7   Max.   :102.42    Max.   :102.65      Max.   :5.000     Max.   :5.000      
 STEM.to.total.ratio Final.average.STEM.score Final.average.STEM.GPA Final.average.score
 Min.   :0.0000      Min.   :11.74            Min.   :1.000          Min.   :16.19      
 1st Qu.:0.2000      1st Qu.:51.34            1st Qu.:2.333          1st Qu.:52.16      
 Median :0.2500      Median :59.59            Median :3.000          Median :59.87      
 Mean   :0.3167      Mean   :59.59            Mean   :2.876          Mean   :60.47      
 3rd Qu.:0.4000      3rd Qu.:67.81            3rd Qu.:3.429          3rd Qu.:68.75      
 Max.   :1.0000      Max.   :99.85            Max.   :5.000          Max.   :95.88      
  Average.GPA   
 Min.   :1.125  
 1st Qu.:2.526  
 Median :3.000  
 Mean   :2.990  
 3rd Qu.:3.450  
 Max.   :4.889  
```

2.Score_non_naplan_data



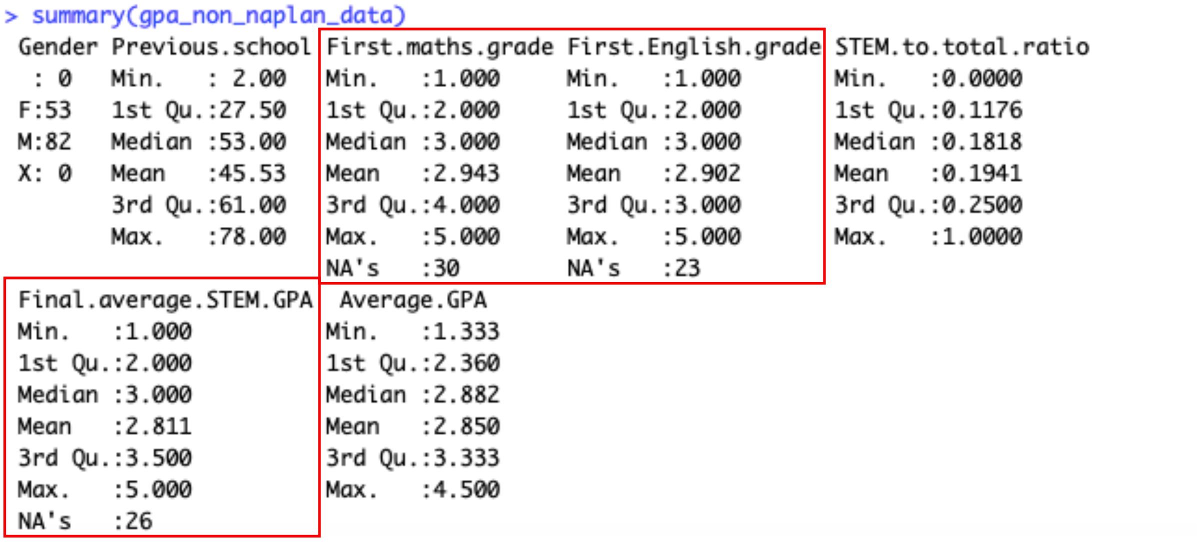
```
> summary(score_non_naplan_data)
 Gender   Previous.school  First.maths.score First.English.score First.maths.grade First.English.grade
  :   0   Min.   :  1.00   Min.   :  0.00    Min.   : 19.77      Min.   :1.000     Min.   :1.000
 F:1244   1st Qu.: 33.00   1st Qu.: 48.75    1st Qu.: 49.32      1st Qu.:2.000     1st Qu.:2.000
 M:1238   Median : 53.00   Median : 59.45    Median : 58.80      Median :3.000     Median :3.000
 X:   0   Mean   : 46.62   Mean   : 59.23    Mean   : 59.06      Mean   :2.904     Mean   :2.906
          3rd Qu.: 61.00   3rd Qu.: 70.08    3rd Qu.: 68.84      3rd Qu.:4.000     3rd Qu.:3.000
          Max.   :100.00   Max.   :109.80    Max.   :102.74      Max.   :5.000     Max.   :5.000
          NA's   :54       NA's   :182       NA's   :295         NA's   :106       NA's   :199
 STEM.to.total.ratio Final.average.STEM.score Final.average.STEM.GPA Final.average.score  Average.GPA
 Min.   :0.0000      Min.   :13.46            Min.   :1.000          Min.   :20.21       Min.   :1.111
 1st Qu.:0.2000      1st Qu.:49.25            1st Qu.:2.154          1st Qu.:50.95       1st Qu.:2.400
 Median :0.2500      Median :58.46            Median :3.000          Median :58.76       Median :2.900
 Mean   :0.3248      Mean   :58.68            Mean   :2.828          Mean   :59.14       Mean   :2.905
 3rd Qu.:0.4082      3rd Qu.:68.58            3rd Qu.:3.500          3rd Qu.:67.20       3rd Qu.:3.379
 Max.   :1.0000      Max.   :97.44            Max.   :5.000          Max.   :95.40       Max.   :4.900
 NA's   :23          NA's   :136              NA's   :85
```

After imputation



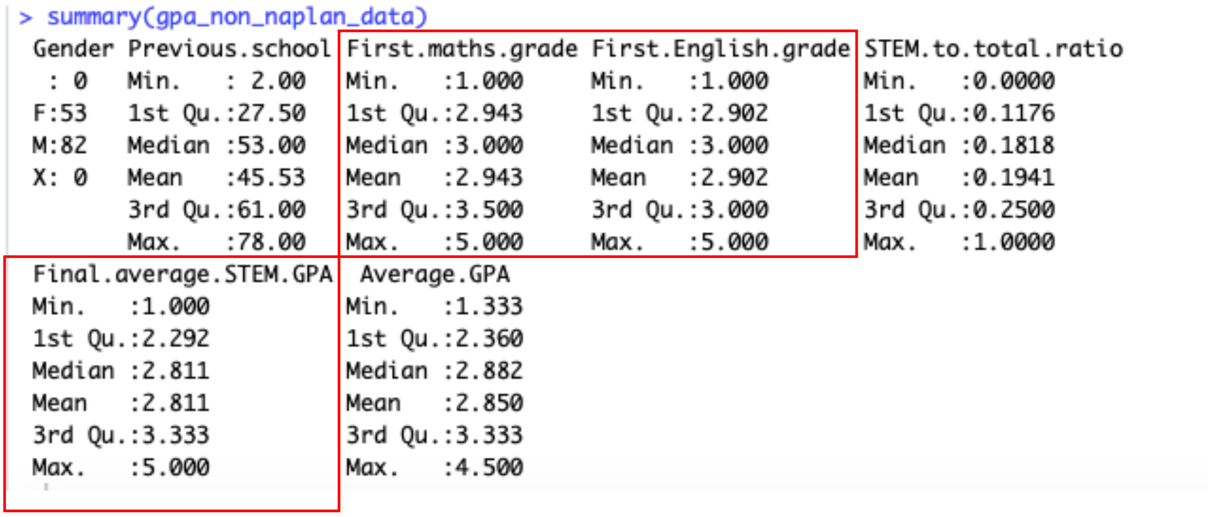
```
> summary(score_non_naplan_data)
 Gender   Previous.school  First.maths.score First.English.score First.maths.grade
  :   0   Min.   :  1.00   Min.   :  0.00    Min.   : 19.77      Min.   :1.000
 F:1244   1st Qu.: 35.00   1st Qu.: 49.50    1st Qu.: 50.91      1st Qu.:2.000
 M:1238   Median : 52.00   Median : 59.23    Median : 59.06      Median :3.000
 X:   0   Mean   : 46.62   Mean   : 59.23    Mean   : 59.06      Mean   :2.904
          3rd Qu.: 61.00   3rd Qu.: 69.23    3rd Qu.: 67.26      3rd Qu.:4.000
          Max.   :100.00   Max.   :109.80    Max.   :102.74      Max.   :5.000
 First.English.grade STEM.to.total.ratio Final.average.STEM.score Final.average.STEM.GPA
 Min.   :1.000       Min.   :0.0000      Min.   :13.46            Min.   :1.000
 1st Qu.:2.000       1st Qu.:0.2000      1st Qu.:49.74            1st Qu.:2.200
 Median :3.000       Median :0.2500      Median :58.68            Median :2.875
 Mean   :2.906       Mean   :0.3241      Mean   :58.68            Mean   :2.828
 3rd Qu.:3.000       3rd Qu.:0.4000      3rd Qu.:67.81            3rd Qu.:3.413
 Max.   :5.000       Max.   :1.0000      Max.   :97.44            Max.   :5.000
 Final.average.score  Average.GPA
 Min.   :20.21       Min.   :1.111
 1st Qu.:50.95       1st Qu.:2.400
 Median :58.76       Median :2.900
 Mean   :59.14       Mean   :2.905
 3rd Qu.:67.20       3rd Qu.:3.379
 Max.   :95.40       Max.   :4.900
```

3.GPA_non_naplan_data



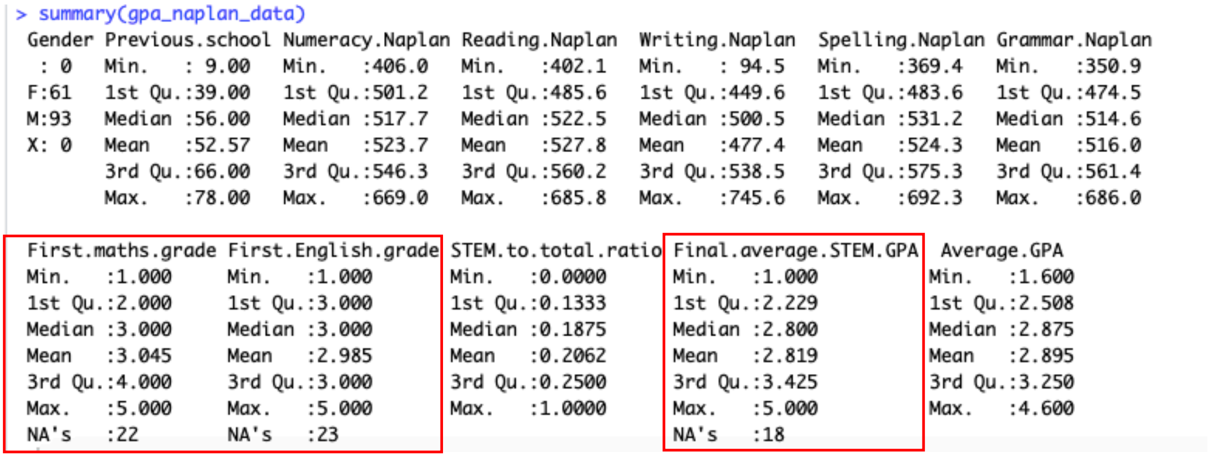
```
> summary(gpa_non_naplan_data)
 Gender Previous.school First.maths.grade First.English.grade STEM.to.total.ratio
  : 0   Min.   : 2.00   Min.   :1.000     Min.   :1.000       Min.   :0.0000
 F:53   1st Qu.:27.50   1st Qu.:2.000     1st Qu.:2.000       1st Qu.:0.1176
 M:82   Median :53.00   Median :3.000     Median :3.000       Median :0.1818
 X: 0   Mean   :45.53   Mean   :2.943     Mean   :2.902       Mean   :0.1941
        3rd Qu.:61.00   3rd Qu.:4.000     3rd Qu.:3.000       3rd Qu.:0.2500
        Max.   :78.00   Max.   :5.000     Max.   :5.000       Max.   :1.0000
                        NA's   :30        NA's   :23
 Final.average.STEM.GPA  Average.GPA
 Min.   :1.000          Min.   :1.333
 1st Qu.:2.000          1st Qu.:2.360
 Median :3.000          Median :2.882
 Mean   :2.811          Mean   :2.850
 3rd Qu.:3.500          3rd Qu.:3.333
 Max.   :5.000          Max.   :4.500
 NA's   :26
```

After imputation



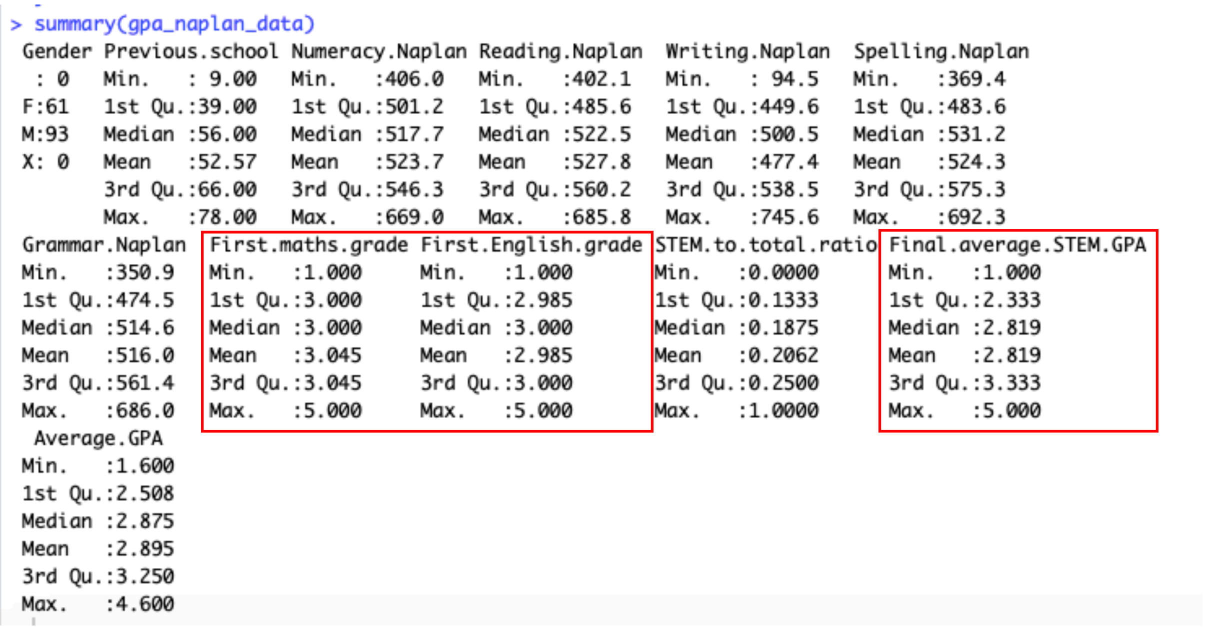
```
> summary(gpa_non_naplan_data)
 Gender Previous.school First.maths.grade First.English.grade STEM.to.total.ratio
  : 0   Min.   : 2.00   Min.   :1.000     Min.   :1.000       Min.   :0.0000
 F:53   1st Qu.:27.50   1st Qu.:2.943     1st Qu.:2.902       1st Qu.:0.1176
 M:82   Median :53.00   Median :3.000     Median :3.000       Median :0.1818
 X: 0   Mean   :45.53   Mean   :2.943     Mean   :2.902       Mean   :0.1941
        3rd Qu.:61.00   3rd Qu.:3.500     3rd Qu.:3.000       3rd Qu.:0.2500
        Max.   :78.00   Max.   :5.000     Max.   :5.000       Max.   :1.0000
 Final.average.STEM.GPA  Average.GPA
 Min.   :1.000          Min.   :1.333
 1st Qu.:2.292          1st Qu.:2.360
 Median :2.811          Median :2.882
 Mean   :2.811          Mean   :2.850
 3rd Qu.:3.333          3rd Qu.:3.333
 Max.   :5.000          Max.   :4.500
```

4.GPA_naplan_data



```
> summary(gpa_naplan_data)
 Gender Previous.school Numeracy.Naplan Reading.Naplan  Writing.Naplan  Spelling.Naplan Grammar.Naplan
  : 0   Min.   : 9.00   Min.   :406.0   Min.   :402.1   Min.   : 94.5   Min.   :369.4   Min.   :350.9
 F:61   1st Qu.:39.00   1st Qu.:501.2   1st Qu.:485.6   1st Qu.:449.6   1st Qu.:483.6   1st Qu.:474.5
 M:93   Median :56.00   Median :517.7   Median :522.5   Median :500.5   Median :531.2   Median :514.6
 X: 0   Mean   :52.57   Mean   :523.7   Mean   :527.8   Mean   :477.4   Mean   :524.3   Mean   :516.0
        3rd Qu.:66.00   3rd Qu.:546.3   3rd Qu.:560.2   3rd Qu.:538.5   3rd Qu.:575.3   3rd Qu.:561.4
        Max.   :78.00   Max.   :669.0   Max.   :685.8   Max.   :745.6   Max.   :692.3   Max.   :686.0

 First.maths.grade First.English.grade STEM.to.total.ratio Final.average.STEM.GPA  Average.GPA
 Min.   :1.000     Min.   :1.000       Min.   :0.0000      Min.   :1.000          Min.   :1.600
 1st Qu.:2.000     1st Qu.:3.000       1st Qu.:0.1333      1st Qu.:2.229          1st Qu.:2.508
 Median :3.000     Median :3.000       Median :0.1875      Median :2.800          Median :2.875
 Mean   :3.045     Mean   :2.985       Mean   :0.2062      Mean   :2.819          Mean   :2.895
 3rd Qu.:4.000     3rd Qu.:3.000       3rd Qu.:0.2500      3rd Qu.:3.425          3rd Qu.:3.250
 Max.   :5.000     Max.   :5.000       Max.   :1.0000      Max.   :5.000          Max.   :4.600
 NA's   :22        NA's   :23                              NA's   :18
```

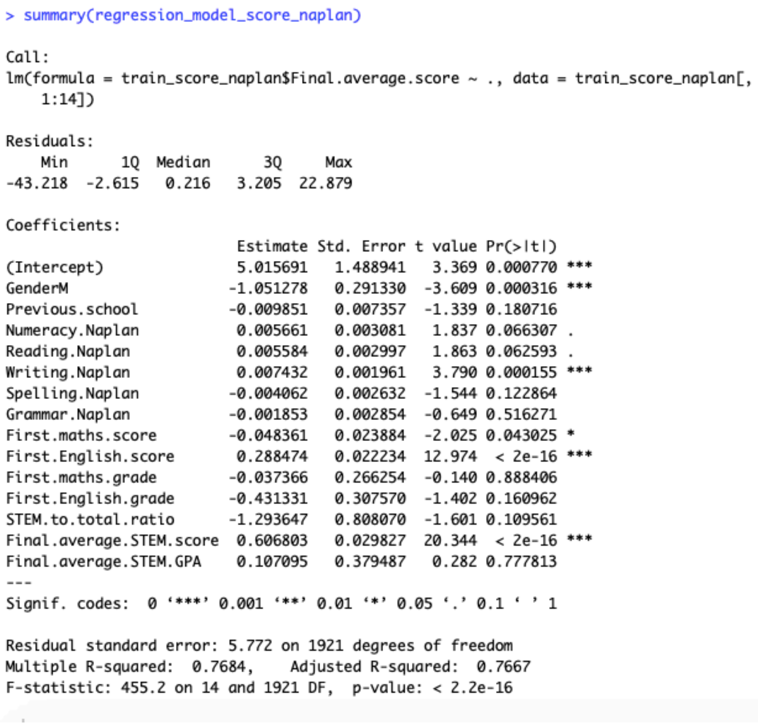
After imputation



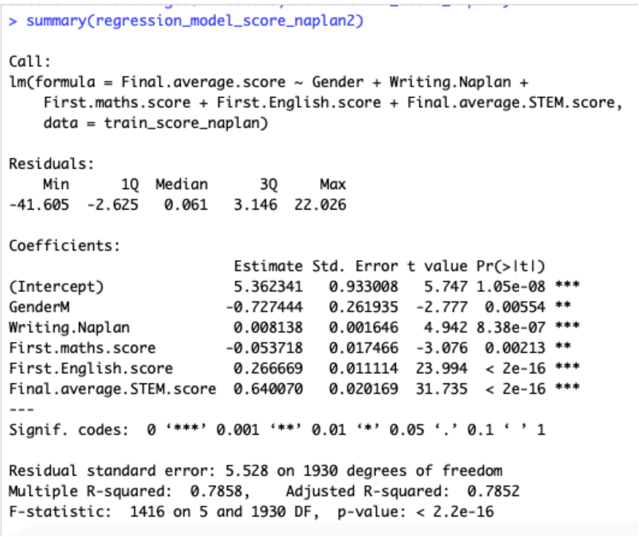
```
> summary(gpa_naplan_data)
 Gender Previous.school Numeracy.Naplan Reading.Naplan  Writing.Naplan  Spelling.Naplan
  : 0   Min.   : 9.00   Min.   :406.0   Min.   :402.1   Min.   : 94.5   Min.   :369.4
 F:61   1st Qu.:39.00   1st Qu.:501.2   1st Qu.:485.6   1st Qu.:449.6   1st Qu.:483.6
 M:93   Median :56.00   Median :517.7   Median :522.5   Median :500.5   Median :531.2
 X: 0   Mean   :52.57   Mean   :523.7   Mean   :527.8   Mean   :477.4   Mean   :524.3
        3rd Qu.:66.00   3rd Qu.:546.3   3rd Qu.:560.2   3rd Qu.:538.5   3rd Qu.:575.3
        Max.   :78.00   Max.   :669.0   Max.   :685.8   Max.   :745.6   Max.   :692.3
 Grammar.Naplan First.maths.grade First.English.grade STEM.to.total.ratio Final.average.STEM.GPA
 Min.   :350.9  Min.   :1.000     Min.   :1.000       Min.   :0.0000      Min.   :1.000
 1st Qu.:474.5  1st Qu.:3.000     1st Qu.:2.985       1st Qu.:0.1333      1st Qu.:2.333
 Median :514.6  Median :3.000     Median :3.000       Median :0.1875      Median :2.819
 Mean   :516.0  Mean   :3.045     Mean   :2.985       Mean   :0.2062      Mean   :2.819
 3rd Qu.:561.4  3rd Qu.:3.045     3rd Qu.:3.000       3rd Qu.:0.2500      3rd Qu.:3.333
 Max.   :686.0  Max.   :5.000     Max.   :5.000       Max.   :1.0000      Max.   :5.000
  Average.GPA
 Min.   :1.600
 1st Qu.:2.508
 Median :2.875
 Mean   :2.895
 3rd Qu.:3.250
 Max.   :4.600
```

Appendix 2

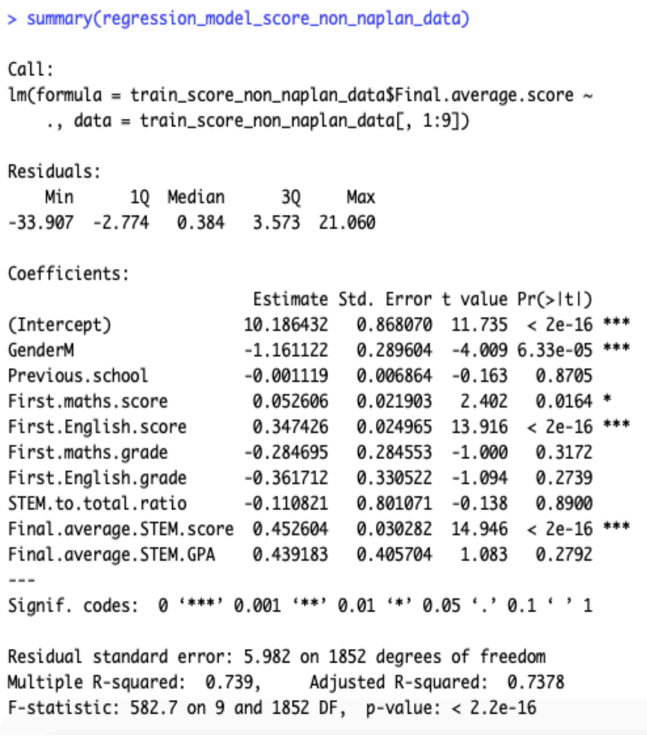
Attribute selection on linear model



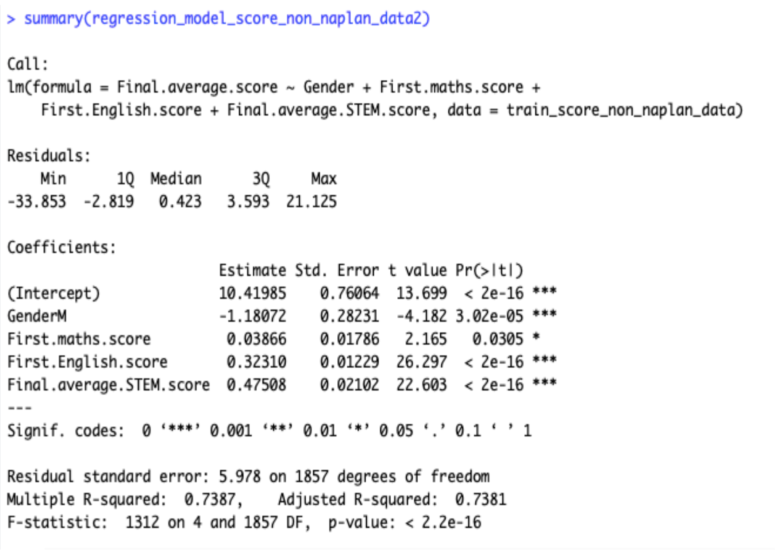
```
> summary(regression_model_score_naplan)

Call:
lm(formula = train_score_naplan$Final.average.score ~ ., data = train_score_naplan[,
    1:14])

Residuals:
    Min      1Q  Median      3Q     Max 
-43.218  -2.615   0.216   3.205  22.879 

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)               5.015691   1.488941   3.369 0.000770 ***
GenderM                  -1.051278   0.291330  -3.609 0.000316 ***
Previous.school          -0.009851   0.007357  -1.339 0.180716    
Numeracy.Naplan           0.005661   0.003081   1.837 0.066307 .  
Reading.Naplan            0.005584   0.002997   1.863 0.062593 .  
Writing.Naplan            0.007432   0.001961   3.790 0.000155 ***
Spelling.Naplan          -0.004062   0.002632  -1.544 0.122864    
Grammar.Naplan           -0.001853   0.002854  -0.649 0.516271    
First.maths.score        -0.048361   0.023884  -2.025 0.043025 *  
First.English.score       0.288474   0.022234  12.974  < 2e-16 ***
First.maths.grade        -0.037366   0.266254  -0.140 0.888406    
First.English.grade      -0.431331   0.307570  -1.402 0.160962    
STEM.to.total.ratio      -1.293647   0.808070  -1.601 0.109561    
Final.average.STEM.score  0.606803   0.029827  20.344  < 2e-16 ***
Final.average.STEM.GPA    0.107095   0.379487   0.282 0.777813    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.772 on 1921 degrees of freedom
Multiple R-squared:  0.7684,	Adjusted R-squared:  0.7667 
F-statistic: 455.2 on 14 and 1921 DF,  p-value: < 2.2e-16
```



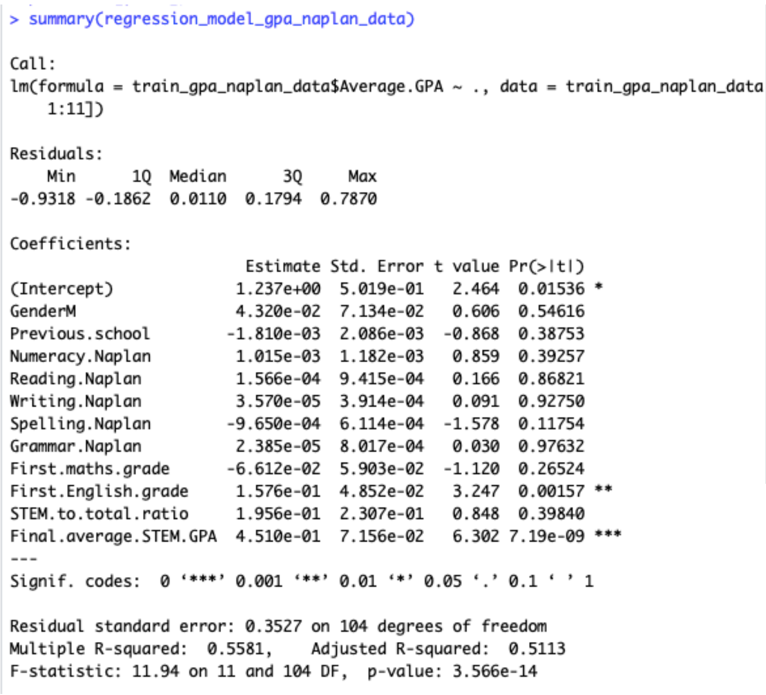
```
> summary(regression_model_score_naplan2)

Call:
lm(formula = Final.average.score ~ Gender + Writing.Naplan + 
    First.maths.score + First.English.score + Final.average.STEM.score, 
    data = train_score_naplan)

Residuals:
    Min      1Q  Median      3Q     Max 
-41.605  -2.625   0.061   3.146  22.026 

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)               5.362341   0.933008   5.747 1.05e-08 ***
GenderM                  -0.727444   0.261935  -2.777  0.00554 ** 
Writing.Naplan            0.008138   0.001646   4.942 8.38e-07 ***
First.maths.score        -0.053718   0.017466  -3.076  0.00213 ** 
First.English.score       0.266669   0.011114  23.994  < 2e-16 ***
Final.average.STEM.score  0.640070   0.020169  31.735  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.528 on 1930 degrees of freedom
Multiple R-squared:  0.7858,	Adjusted R-squared:  0.7852 
F-statistic:  1416 on 5 and 1930 DF,  p-value: < 2.2e-16
```



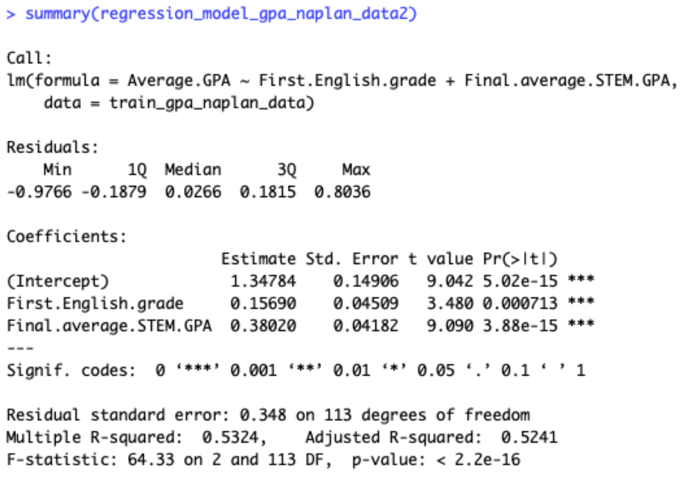
```
> summary(regression_model_score_non_naplan_data)

Call:
lm(formula = train_score_non_naplan_data$Final.average.score ~ 
    ., data = train_score_non_naplan_data[, 1:9])

Residuals:
    Min      1Q  Median      3Q     Max 
-33.907  -2.774   0.384   3.573  21.060 

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)              10.186432   0.868070  11.735  < 2e-16 ***
GenderM                  -1.161122   0.289604  -4.009 6.33e-05 ***
Previous.school          -0.001119   0.006864  -0.163   0.8705    
First.maths.score         0.052606   0.021903   2.402   0.0164 *  
First.English.score       0.347426   0.024965  13.916  < 2e-16 ***
First.maths.grade        -0.284695   0.284553  -1.000   0.3172    
First.English.grade      -0.361712   0.330522  -1.094   0.2739    
STEM.to.total.ratio      -0.110821   0.801071  -0.138   0.8900    
Final.average.STEM.score  0.452604   0.030282  14.946  < 2e-16 ***
Final.average.STEM.GPA    0.439183   0.405704   1.083   0.2792    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.982 on 1852 degrees of freedom
Multiple R-squared:  0.739,	Adjusted R-squared:  0.7378 
F-statistic: 582.7 on 9 and 1852 DF,  p-value: < 2.2e-16
```



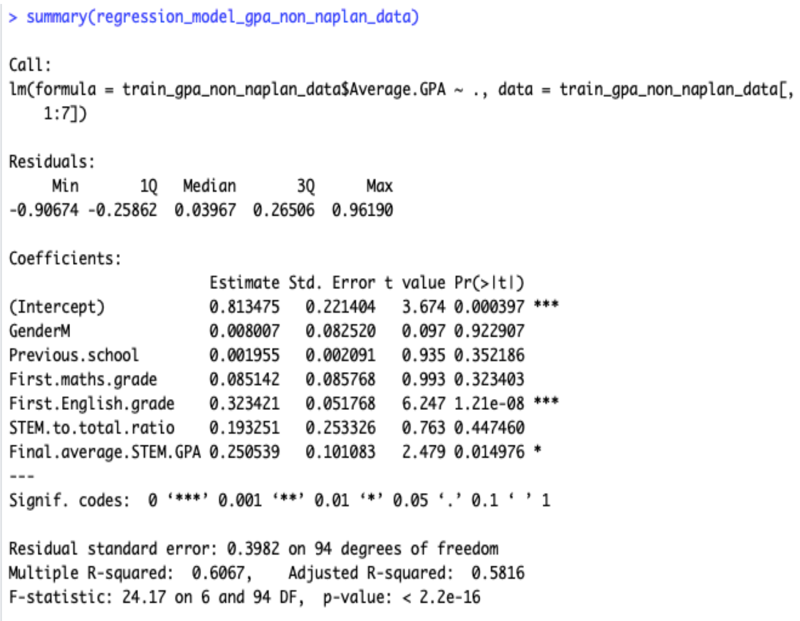
```
> summary(regression_model_score_non_naplan_data2)

Call:
lm(formula = Final.average.score ~ Gender + First.maths.score + 
    First.English.score + Final.average.STEM.score, data = train_score_non_naplan_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-33.853  -2.819   0.423   3.593  21.125 

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)    
(Intercept)              10.41985    0.76064  13.699  < 2e-16 ***
GenderM                  -1.18072    0.28231  -4.182 3.02e-05 ***
First.maths.score         0.03866    0.01786   2.165   0.0305 *  
First.English.score       0.32310    0.01229  26.297  < 2e-16 ***
Final.average.STEM.score  0.47508    0.02102  22.603  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.978 on 1857 degrees of freedom
Multiple R-squared:  0.7387,	Adjusted R-squared:  0.7381 
F-statistic:  1312 on 4 and 1857 DF,  p-value: < 2.2e-16
```



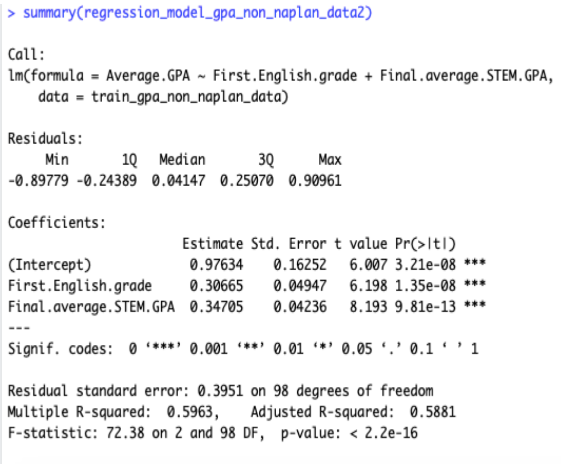
```
> summary(regression_model_gpa_naplan_data)

Call:
lm(formula = train_gpa_naplan_data$Average.GPA ~ ., data = train_gpa_naplan_data
    1:11])

Residuals:
    Min      1Q  Median      3Q     Max 
-0.9318 -0.1862  0.0110  0.1794  0.7870 

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)            1.237e+00  5.019e-01   2.464  0.01536 *  
GenderM                4.320e-02  7.134e-02   0.606  0.54616    
Previous.school       -1.810e-03  2.086e-03  -0.868  0.38753    
Numeracy.Naplan        1.015e-03  1.182e-03   0.859  0.39257    
Reading.Naplan         1.566e-04  9.415e-04   0.166  0.86821    
Writing.Naplan         3.570e-05  3.914e-04   0.091  0.92750    
Spelling.Naplan       -9.650e-04  6.114e-04  -1.578  0.11754    
Grammar.Naplan         2.385e-05  8.017e-04   0.030  0.97632    
First.maths.grade     -6.612e-02  5.903e-02  -1.120  0.26524    
First.English.grade    1.576e-01  4.852e-02   3.247  0.00157 ** 
STEM.to.total.ratio    1.956e-01  2.307e-01   0.848  0.39840    
Final.average.STEM.GPA  4.510e-01  7.156e-02   6.302 7.19e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3527 on 104 degrees of freedom
Multiple R-squared:  0.5581,	Adjusted R-squared:  0.5113 
F-statistic: 11.94 on 11 and 104 DF,  p-value: 3.566e-14
```



```
> summary(regression_model_gpa_naplan_data2)

Call:
lm(formula = Average.GPA ~ First.English.grade + Final.average.STEM.GPA, 
    data = train_gpa_naplan_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.9766 -0.1879  0.0266  0.1815  0.8036 

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)             1.34784    0.14906   9.042 5.02e-15 ***
First.English.grade     0.15690    0.04509   3.480 0.000713 ***
Final.average.STEM.GPA  0.38020    0.04182   9.090 3.88e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.348 on 113 degrees of freedom
Multiple R-squared:  0.5324,	Adjusted R-squared:  0.5241 
F-statistic: 64.33 on 2 and 113 DF,  p-value: < 2.2e-16
```

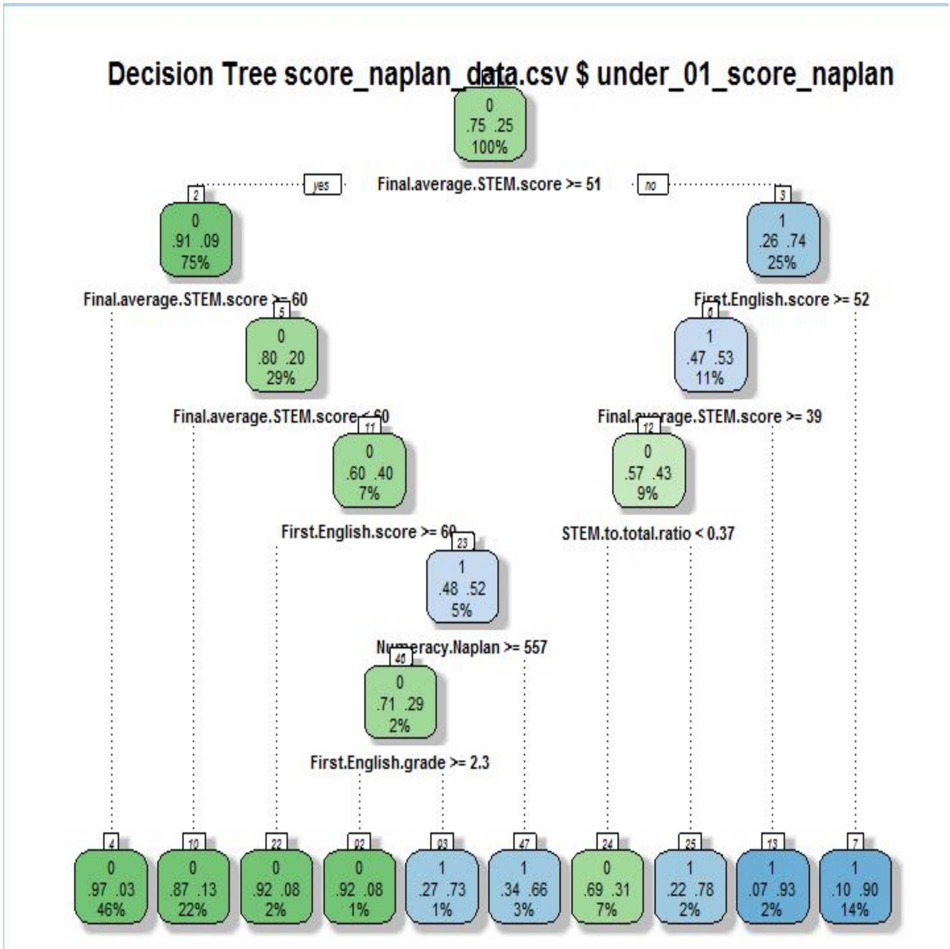


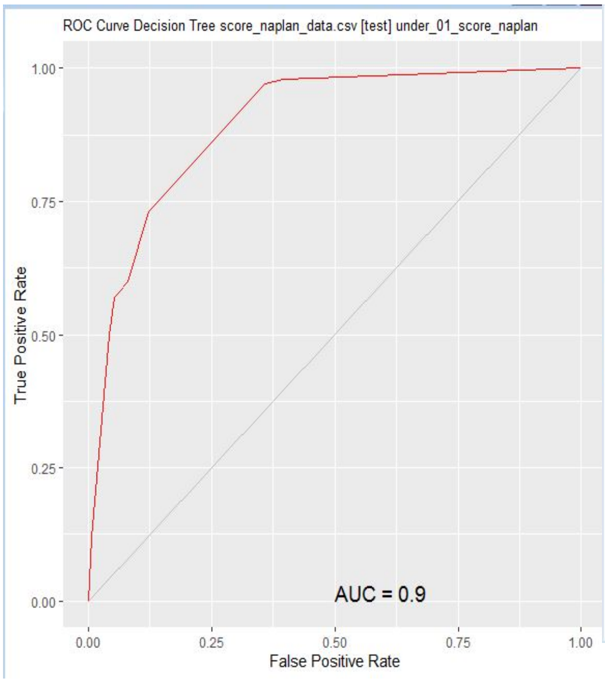
```
> summary(regression_model_gpa_non_naplan_data)

Call:
lm(formula = train_gpa_non_naplan_data$Average.GPA ~ ., data = train_gpa_non_naplan_data[, 
    1:7])

Residuals:
     Min       1Q   Median       3Q      Max 
-0.90674 -0.25862  0.03967  0.26506  0.96190 

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)            0.813475   0.221404   3.674 0.000397 ***
GenderM                0.008007   0.082520   0.097 0.922907    
Previous.school        0.001955   0.002091   0.935 0.352186    
First.maths.grade      0.085142   0.085768   0.993 0.323403    
First.English.grade    0.323421   0.051768   6.247 1.21e-08 ***
STEM.to.total.ratio    0.193251   0.253326   0.763 0.447460    
Final.average.STEM.GPA 0.250539   0.101083   2.479 0.014976 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3982 on 94 degrees of freedom
Multiple R-squared:  0.6067,	Adjusted R-squared:  0.5816 
F-statistic: 24.17 on 6 and 94 DF,  p-value: < 2.2e-16
```

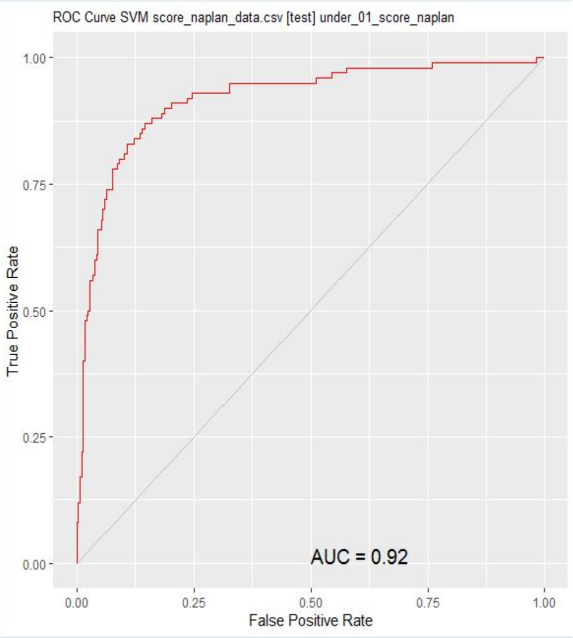


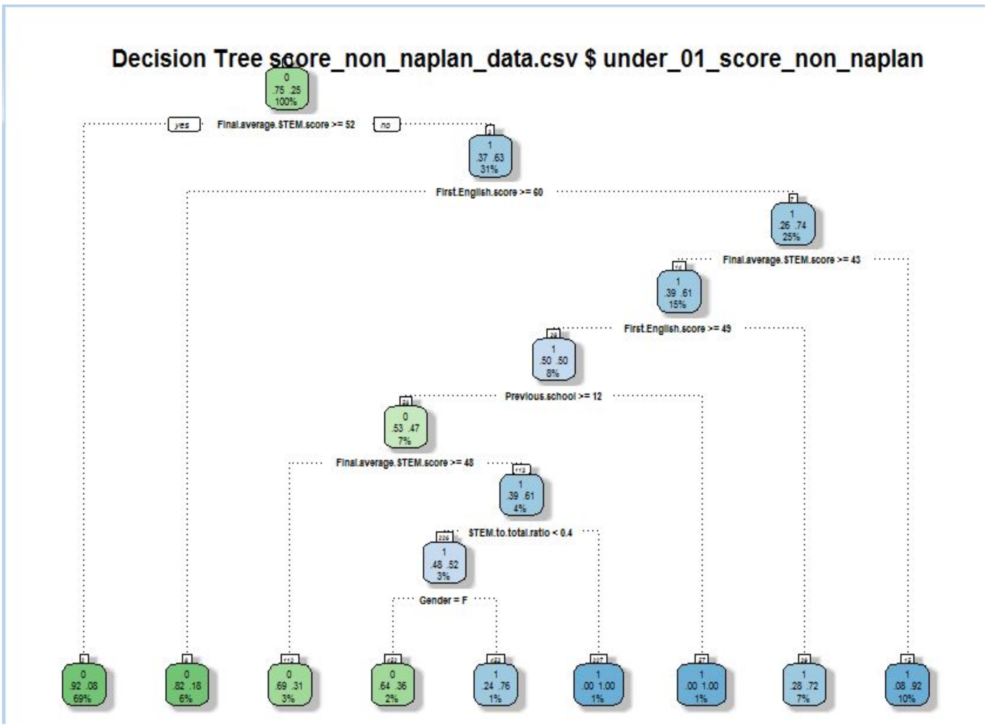
```
> summary(regression_model_gpa_non_naplan_data2)

Call:
lm(formula = Average.GPA ~ First.English.grade + Final.average.STEM.GPA, 
    data = train_gpa_non_naplan_data)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.89779 -0.24389  0.04147  0.25070  0.90961 

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)             0.97634    0.16252   6.007 3.21e-08 ***
First.English.grade     0.30665    0.04947   6.198 1.35e-08 ***
Final.average.STEM.GPA  0.34705    0.04236   8.193 9.81e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3951 on 98 degrees of freedom
Multiple R-squared:  0.5963,	Adjusted R-squared:  0.5881 
F-statistic: 72.38 on 2 and 98 DF,  p-value: < 2.2e-16
```

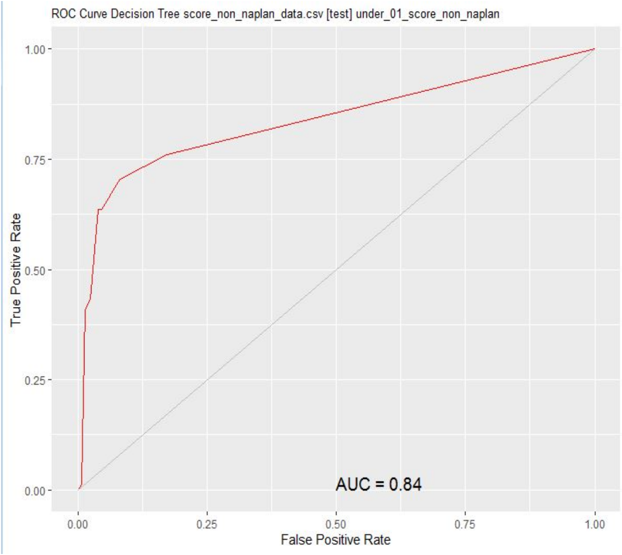
Appendix 3

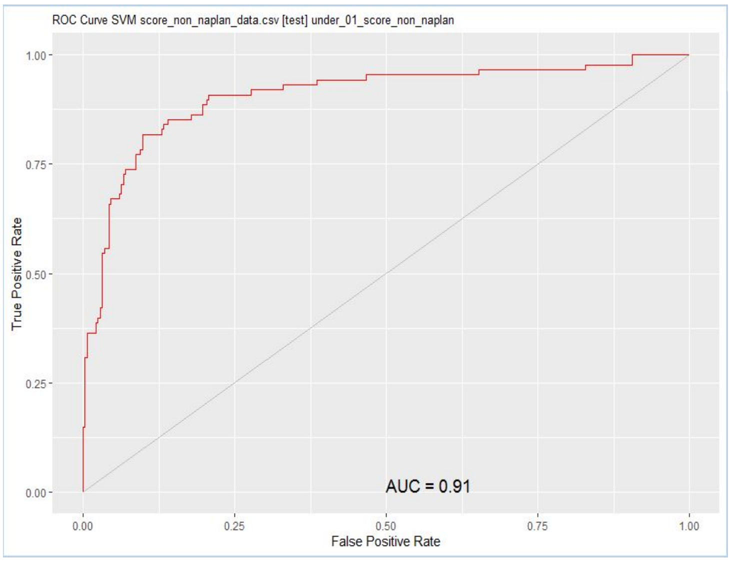

Decision Tree score_naplan_data.csv $ under_01_score_naplan
1
0
.75 .25
100%
Final.average.STEM.score >= 51
yes
no
2
0
.91 .09
75%
3
1
.26 .74
25%
Final.average.STEM.score >= 60
5
0
.80 .20
29%
First.English.score >= 52
6
1
.47 .53
11%
Final.average.STEM.score < 60
11
0
.60 .40
7%
Final.average.STEM.score >= 39
12
0
.57 .43
9%
First.English.score >= 60
23
1
.48 .52
5%
STEM.to.total.ratio < 0.37
Numeracy.Naplan >= 557
46
0
.71 .29
2%
First.English.grade >= 2.3
4
0
.97 .03
46%
10
0
.87 .13
22%
22
0
.92 .08
2%
92
0
.92 .08
1%
93
1
.27 .73
1%
47
1
.34 .66
3%
24
0
.69 .31
7%
25
1
.22 .78
2%
13
1
.07 .93
2%
7
1
.10 .90
14%

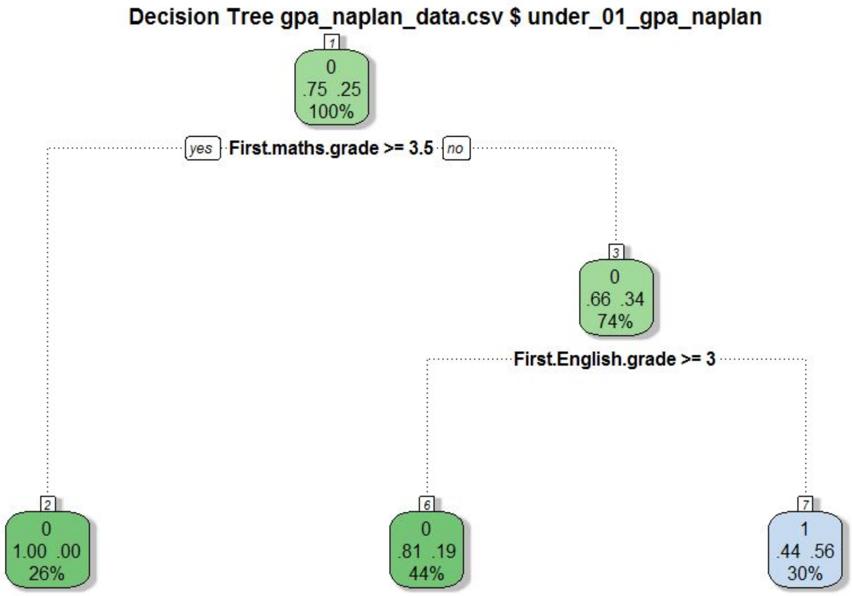

ROC Curve Decision Tree score_naplan_data.csv [test] under_01_score_naplan
1.00
0.75
0.50
0.25
0.00
True Positive Rate
0.00
0.25
0.50
0.75
1.00
False Positive Rate
AUC = 0.9

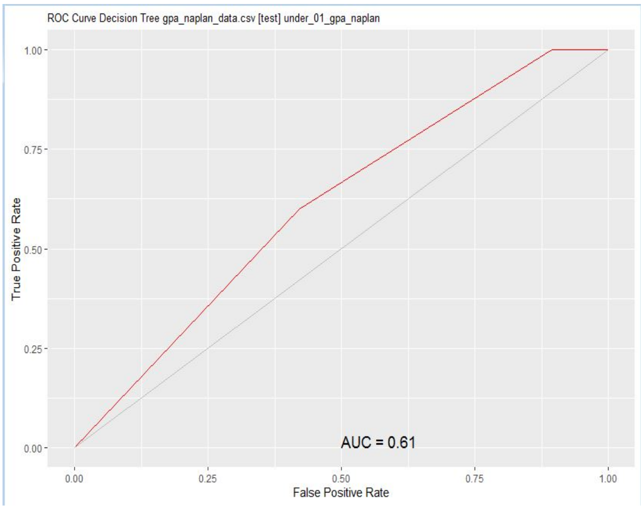

ROC Curve SVM score_naplan_data.csv [test] under_01_score_naplan
1.00
0.75
0.50
0.25
0.00
True Positive Rate
0.00
0.25
0.50
0.75
1.00
False Positive Rate
AUC = 0.92

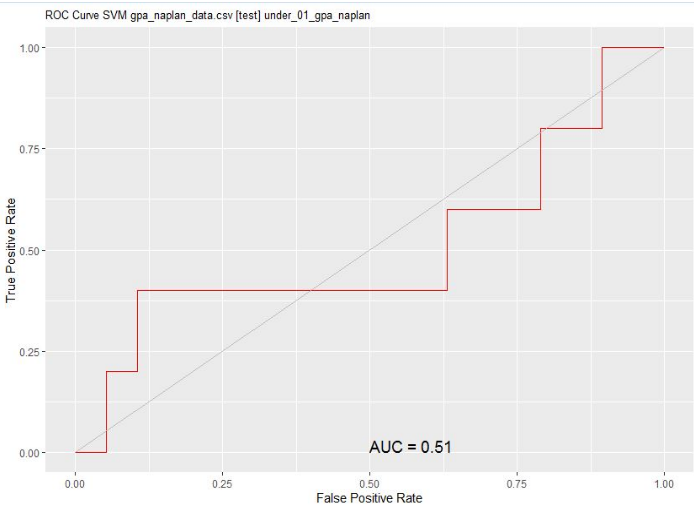

Decision Tree score_non_naplan_data.csv $ under_01_score_non_naplan
0
.75 .25
100%
yes
Final.average.STEM.score >= 52
no
1
.37 .63
31%
First.English.score >= 60
1
.26 .74
25%
Final.average.STEM.score >= 43
1
.39 .61
15%
First.English.score >= 49
1
.50 .50
8%
Previous.school >= 12
0
.53 .47
7%
Final.average.STEM.score >= 48
1
.39 .61
4%
STEM.to.total.ratio < 0.4
1
.48 .52
3%
Gender = F
0
.92 .08
69%
0
.82 .18
6%
0
.69 .31
3%
0
.64 .36
2%
1
.24 .76
1%
1
.00 1.00
1%
1
.00 1.00
1%
1
.28 .72
7%
1
.08 .92
10%

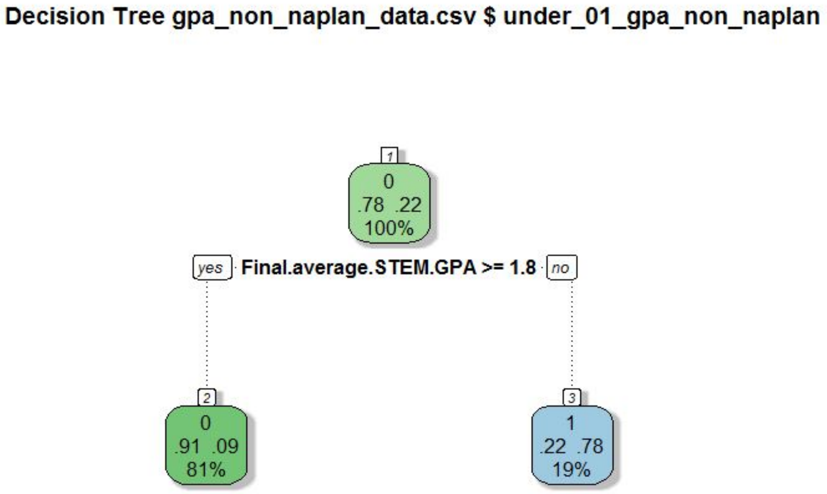

ROC Curve Decision Tree score_non_naplan_data.csv [test] under_01_score_non_naplan
1.00
0.75
0.50
0.25
0.00
True Positive Rate
0.00
0.25
0.50
0.75
1.00
False Positive Rate
AUC = 0.84

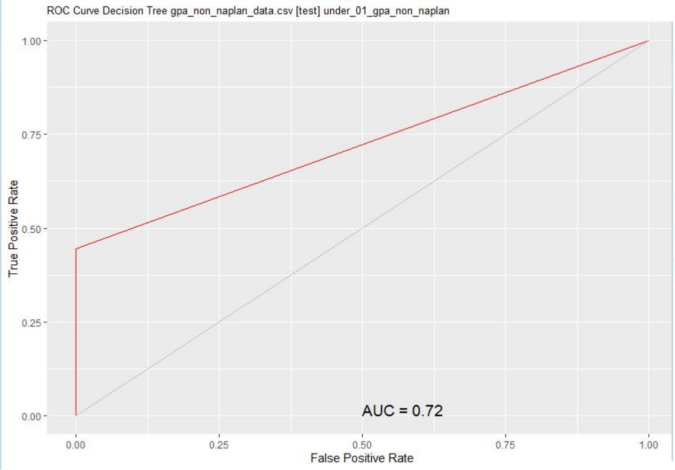

ROC Curve SVM score_non_naplan_data.csv [test] under_01_score_non_naplan
1.00
0.75
0.50
0.25
0.00
True Positive Rate
0.00
0.25
0.50
0.75
1.00
False Positive Rate
AUC = 0.91

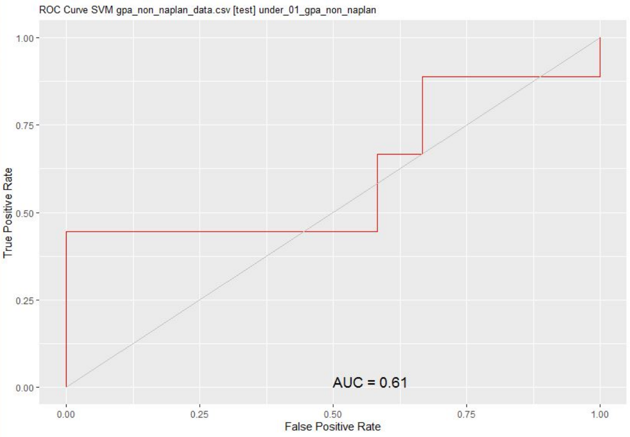

Decision Tree gpa_naplan_data.csv $ under_01_gpa_naplan
1
0
.75 .25
100%
yes
First.maths.grade >= 3.5
no
3
0
.66 .34
74%
First.English.grade >= 3
2
0
1.00 .00
26%
6
0
.81 .19
44%
7
1
.44 .56
30%


ROC Curve Decision Tree gpa_naplan_data.csv [test] under_01_gpa_naplan
True Positive Rate
1.00
0.75
0.50
0.25
0.00
AUC = 0.61
0.00
0.25
0.50
0.75
1.00
False Positive Rate

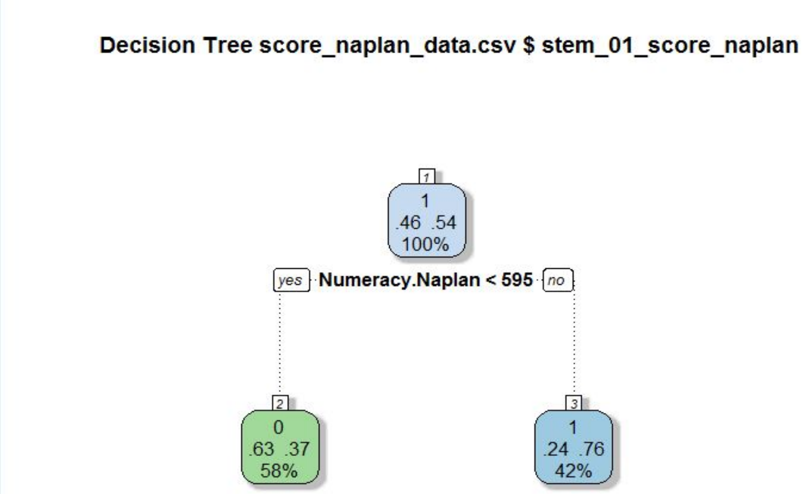

ROC Curve SVM gpa_naplan_data.csv [test] under_01_gpa_naplan
True Positive Rate
1.00
0.75
0.50
0.25
0.00
AUC = 0.51
0.00
0.25
0.50
0.75
1.00
False Positive Rate

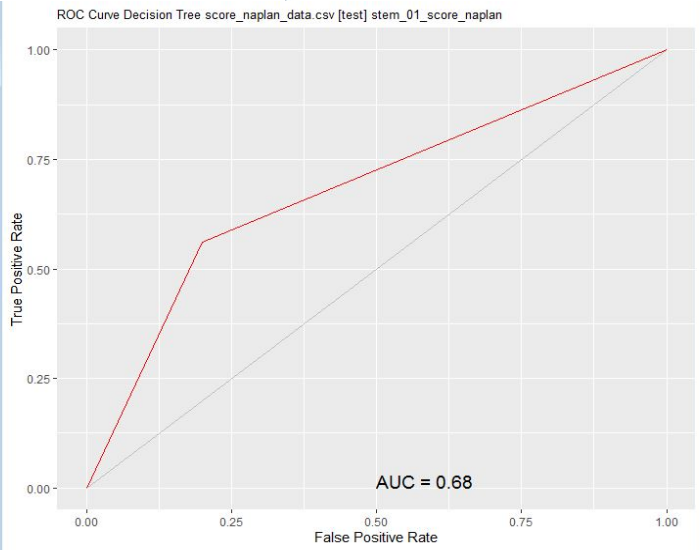

Decision Tree gpa_non_naplan_data.csv $ under_01_gpa_non_naplan
1
0
.78 .22
100%
yes
Final.average.STEM.GPA >= 1.8
no
2
0
.91 .09
81%
3
1
.22 .78
19%

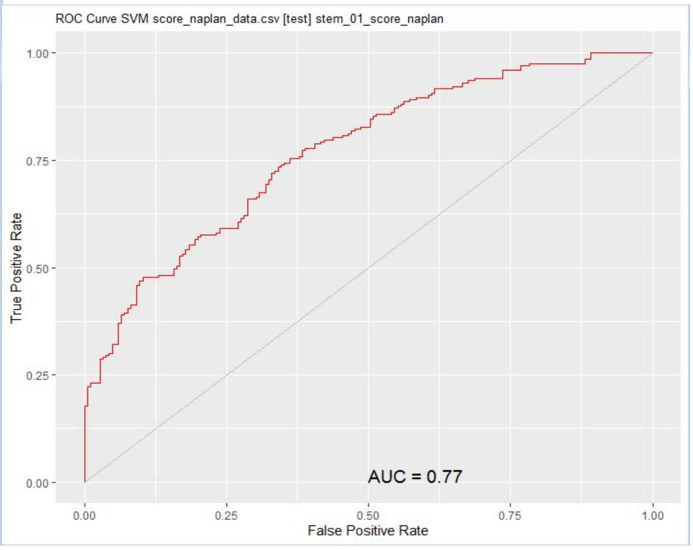

ROC Curve Decision Tree gpa_non_naplan_data.csv [test] under_01_gpa_non_naplan
1.00
0.75
0.50
0.25
0.00
True Positive Rate
0.00
0.25
0.50
0.75
1.00
False Positive Rate
AUC = 0.72

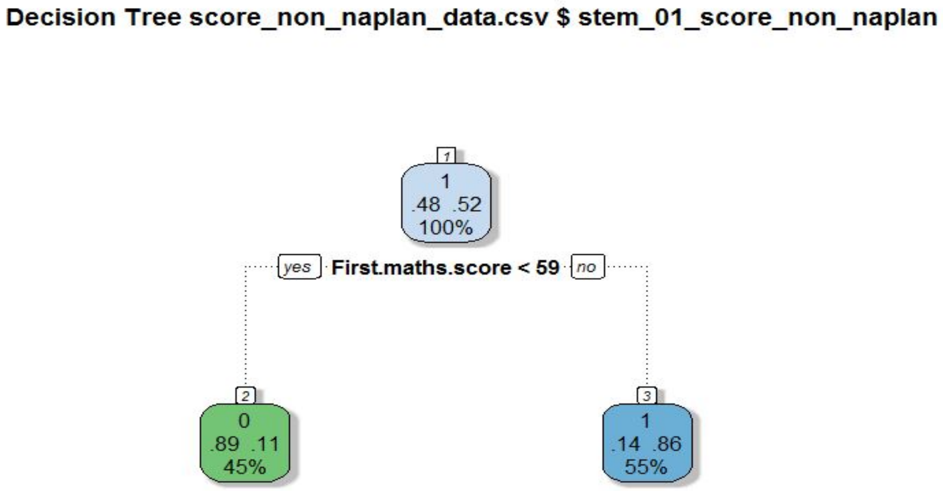

ROC Curve SVM gpa_non_naplan_data.csv [test] under_01_gpa_non_naplan
1.00
0.75
0.50
0.25
0.00
True Positive Rate
0.00
0.25
0.50
0.75
1.00
False Positive Rate
AUC = 0.61

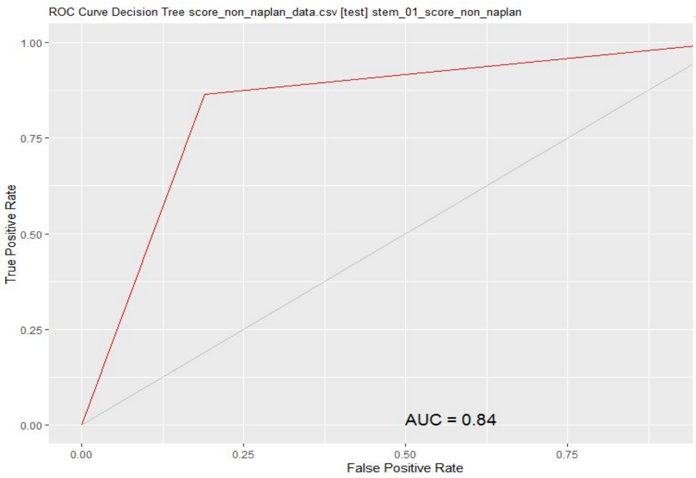
Appendix 4

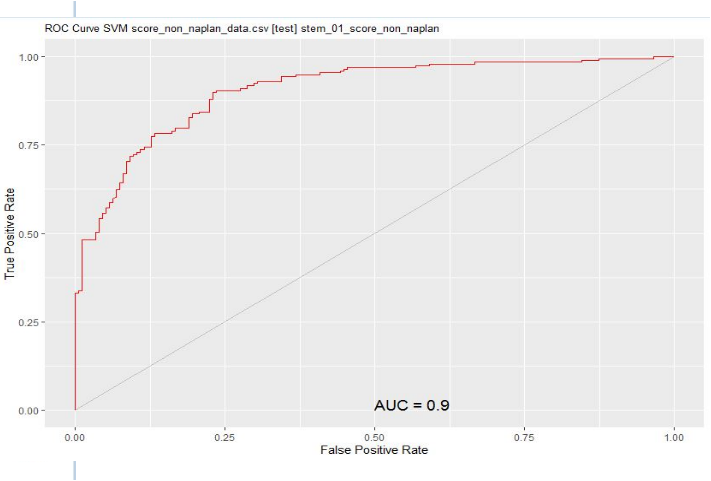

Decision Tree score_naplan_data.csv $ stem_01_score_naplan
1
1
.46 .54
100%
yes
Numeracy.Naplan < 595
no
2
0
.63 .37
58%
3
1
.24 .76
42%

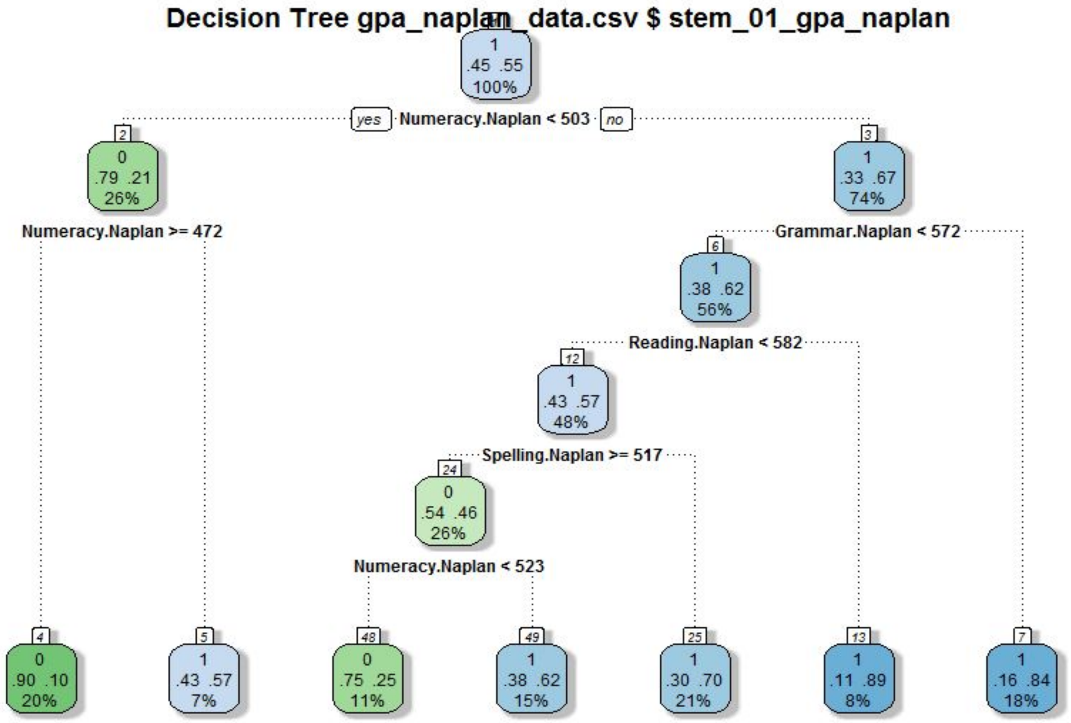

ROC Curve Decision Tree score_naplan_data.csv [test] stem_01_score_naplan
True Positive Rate
1.00
0.75
0.50
0.25
0.00
False Positive Rate
0.00
0.25
0.50
0.75
1.00
AUC = 0.68

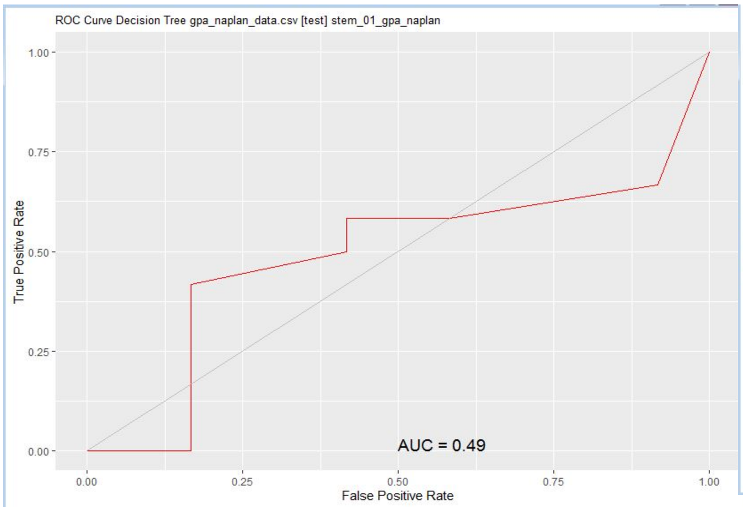

ROC Curve SVM score_naplan_data.csv [test] stem_01_score_naplan
True Positive Rate
1.00
0.75
0.50
0.25
0.00
False Positive Rate
0.00
0.25
0.50
0.75
1.00
AUC = 0.77

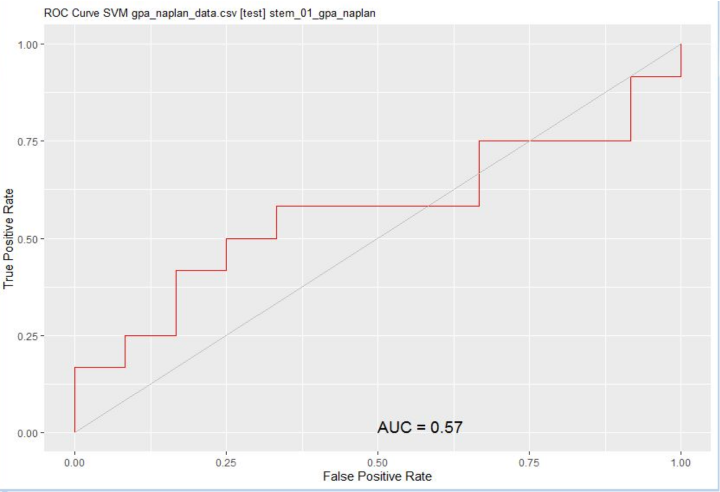

Decision Tree score_non_naplan_data.csv $ stem_01_score_non_naplan
1
1
.48 .52
100%
yes
First.maths.score < 59
no
2
0
.89 .11
45%
3
1
.14 .86
55%

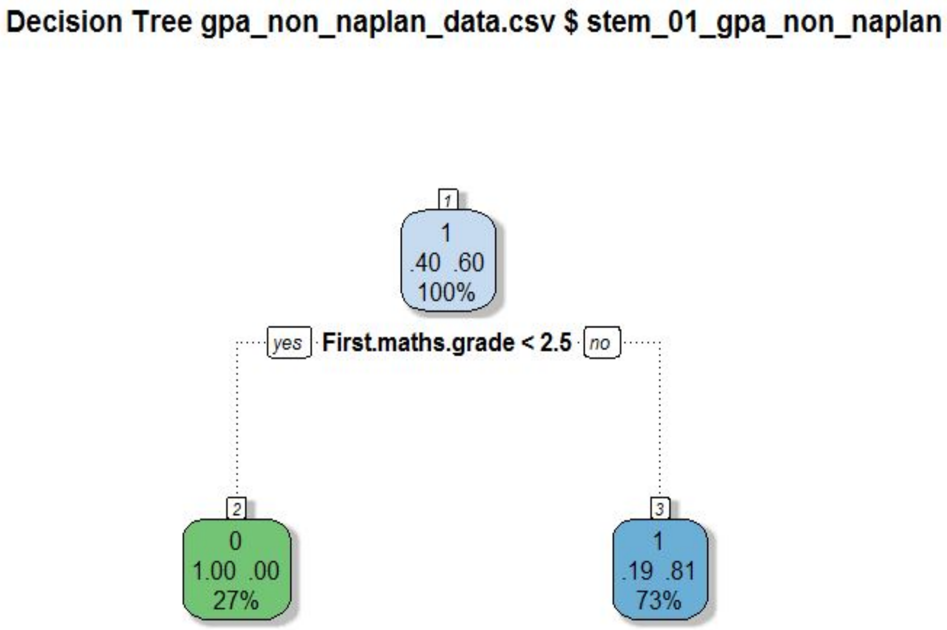

ROC Curve Decision Tree score_non_naplan_data.csv [test] stem_01_score_non_naplan
1.00
0.75
0.50
0.25
0.00
True Positive Rate
AUC = 0.84
0.00
0.25
0.50
0.75
False Positive Rate

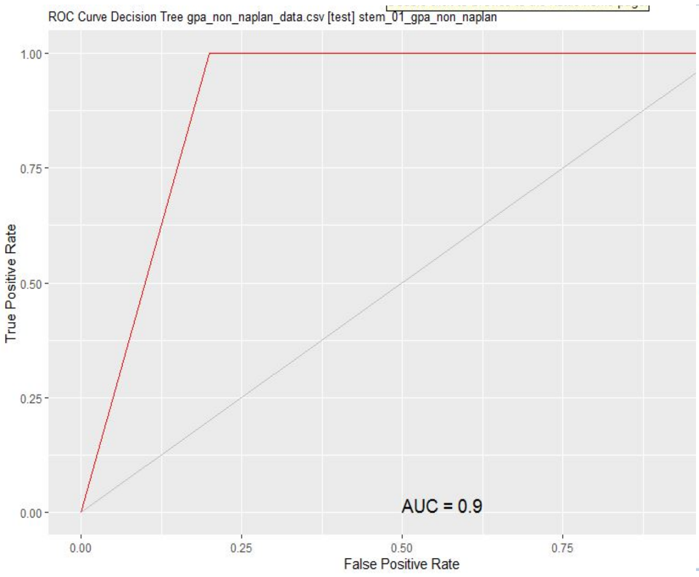

ROC Curve SVM score_non_naplan_data.csv [test] stem_01_score_non_naplan
1.00
0.75
0.50
0.25
0.00
True Positive Rate
AUC = 0.9
0.00
0.25
0.50
0.75
1.00
False Positive Rate

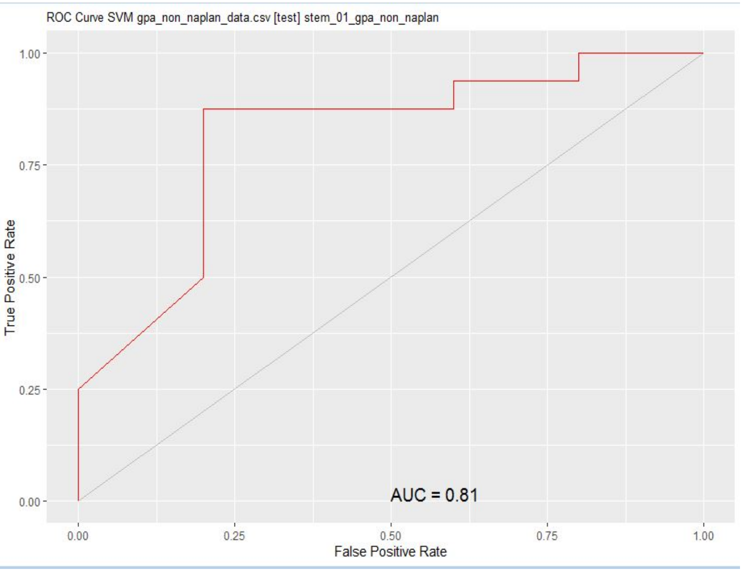

Decision Tree gpa_naplan_data.csv $ stem_01_gpa_naplan
1
1
.45 .55
100%
yes
Numeracy.Naplan < 503
no
2
0
.79 .21
26%
3
1
.33 .67
74%
Numeracy.Naplan >= 472
Grammar.Naplan < 572
6
1
.38 .62
56%
Reading.Naplan < 582
12
1
.43 .57
48%
Spelling.Naplan >= 517
24
0
.54 .46
26%
Numeracy.Naplan < 523
4
0
.90 .10
20%
5
1
.43 .57
7%
48
0
.75 .25
11%
49
1
.38 .62
15%
25
1
.30 .70
21%
13
1
.11 .89
8%
7
1
.16 .84
18%


ROC Curve Decision Tree gpa_naplan_data.csv [test] stem_01_gpa_naplan
1.00
0.75
0.50
0.25
0.00
True Positive Rate
0.00
0.25
0.50
0.75
1.00
False Positive Rate
AUC = 0.49


ROC Curve SVM gpa_naplan_data.csv [test] stem_01_gpa_naplan
1.00
0.75
0.50
0.25
0.00
True Positive Rate
0.00
0.25
0.50
0.75
1.00
False Positive Rate
AUC = 0.57


Decision Tree gpa_non_naplan_data.csv $ stem_01_gpa_non_naplan
1
1
.40 .60
100%
yes
First.maths.grade < 2.5
no
2
0
1.00 .00
27%
3
1
.19 .81
73%


ROC Curve Decision Tree gpa_non_naplan_data.csv [test] stem_01_gpa_non_naplan
1.00
0.75
0.50
0.25
0.00
True Positive Rate
AUC = 0.9
0.00
0.25
0.50
0.75
False Positive Rate


ROC Curve SVM gpa_non_naplan_data.csv [test] stem_01_gpa_non_naplan
1.00
0.75
0.50
0.25
0.00
True Positive Rate
AUC = 0.81
0.00
0.25
0.50
0.75
1.00
False Positive Rate

Appendix 5

Kernel selection of SVM on underperformance

Score_naplan_data

SVM kernel	Overall error	ROC-AUC
Radial Basis(rbfdot)	10.6%	0.95
Polynomial(polydot)	11.6%	0.94
Linear(vanilladot)	11.6%	0.94
Hyperbolin Tangent(Tanhdot)	31.3%	0.74
Laplacian(laplacedot)	11.6%	0.95
Bessel(besseldot)	39%	0.64
ANOVA RBF(anovadot)	73.4%	0.53
Spline(splinedot)	NA	NA

Score_non_naplan_data

SVM kernel	Overall error	ROC-AUC
Radial Basis(rbfdot)	12.4%	0.91
Polynomial(polydot)	12.4%	0.92
Linear(vanilladot)	12.4%	0.92
Hyperbolin Tangent(Tanhdot)	31%	0.74
Laplacian(laplacedot)	12.3%	0.92
Bessel(besseldot)	15.3%	0.86
ANOVA RBF(anovadot)	25.8%	0.23
Spline(splinedot)	27.9%	0.59

Gpa_naplan_data

SVM kernel	Overall error	ROC-AUC
Radial Basis(rbfdot)	30.4%	0.82
Polynomial(polydot)	30.5%	0.79
Linear(vanilladot)	30.5%	0.79
Hyperbolin Tangent(Tanhdot)	43.5%	0.63
Laplacian(laplacedot)	30.4%	0.82
Bessel(besseldot)	34.8%	0.79
ANOVA RBF(anovadot)	26.1%	0.84
Spline(splinedot)	21.7%	0.62

Gpa_non_naplan_data

SVM kernel	Overall error	ROC-AUC
Radial Basis(rbfdot)	5%	0.8
Polynomial(polydot)	10%	0.92
Linear(vanilladot)	10%	0.92
Hyperbolin Tangent(Tanhdot)	35%	0.56
Laplacian(laplacedot)	10%	0.84
Bessel(besseldot)	10%	0.83
ANOVA RBF(anovadot)	5%	0.94
Spline(splinedot)	30%	0.33

Kernel selection of SVM on STEM subject sutability

Score_naplan_data

SVM kernel	Overall error	ROC-AUC
Radial Basis(rbfdot)	30.7%	0.74
Polynomial(polydot)	29.9%	0.77
Linear(vanilladot)	29.9%	0.77
Hyperbolin Tangent(Tanhdot)	50.1%	0.54
Laplacian(laplacedot)	32.6%	0.76
Bessel(besseldot)	31%	0.76
ANOVA RBF(anovadot)	50.1%	0.56
Spline(splinedot)	NA	NA

Score_non_naplan_data

SVM kernel	Overall error	ROC-AUC
Radial Basis(rbfdot)	13.7%	0.92
Polynomial(polydot)	13.7%	0.92
Linear(vanilladot)	15.3%	0.92
Hyperbolin Tangent(Tanhdot)	29%	0.82
Laplacian(laplacedot)	14.2%	0.93
Bessel(besseldot)	15.1%	0.93
ANOVA RBF(anovadot)	47.6%	0.56
Spline(splinedot)	NA	NA

Gpa_naplan_data

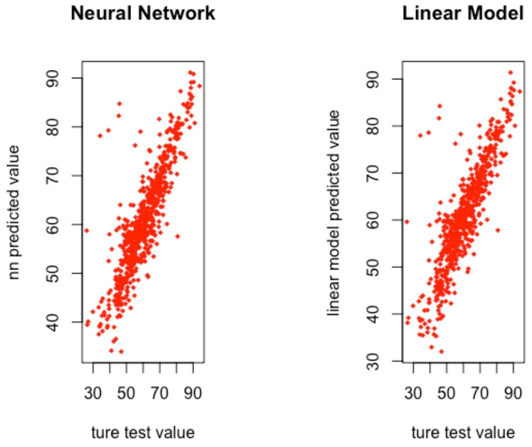
SVM kernel	Overall error	ROC-AUC
Radial Basis(rbfdot)	26.1%	0.87
Polynomial(polydot)	21.7%	0.86
Linear(vanilladot)	21.7%	0.86
Hyperbolin Tangent(Tanhdot)	39.2%	0.49
Laplacian(laplacedot)	26.1%	0.88
Bessel(besseldot)	26.1%	0.85
ANOVA RBF(anovadot)	21.8%	0.79
Spline(splinedot)	52.2%	0.48

Gpa_non_naplan_data

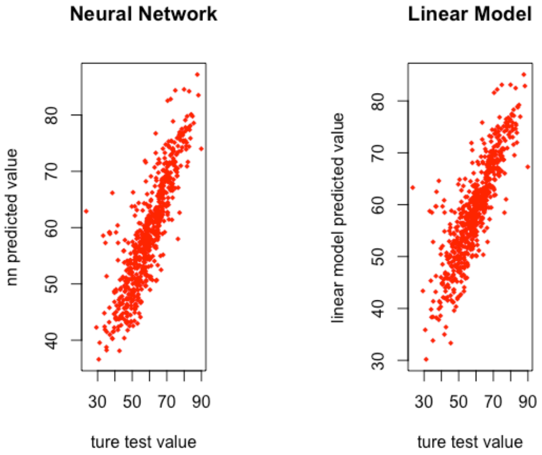
SVM kernel	Overall error	ROC-AUC
Radial Basis(rbfdot)	10%	0.77
Polynomial(polydot)	10%	0.8
Linear(vanilladot)	10%	0.93
Hyperbolin Tangent(Tanhdot)	15%	0.95
Laplacian(laplacedot)	10%	0.92
Bessel(besseldot)	10%	0.81
ANOVA RBF(anovadot)	10%	0.85
Spline(splinedot)	20%	0.51

Appendix 6

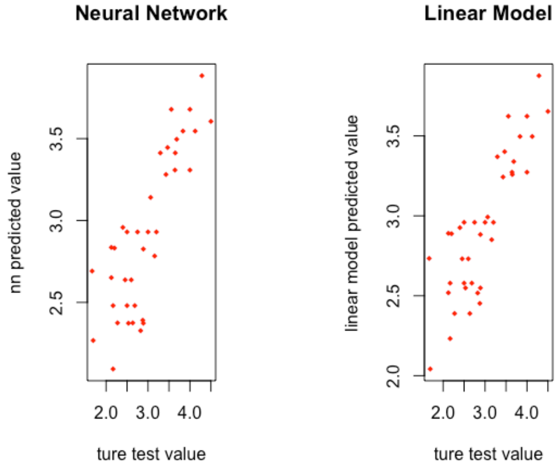
score_naplan_data



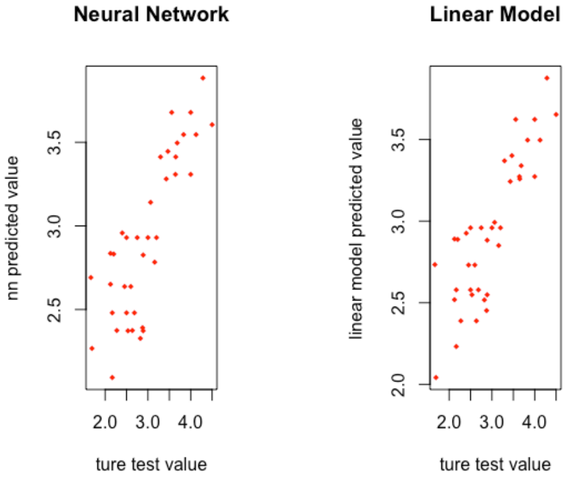
score_non_naplan_data



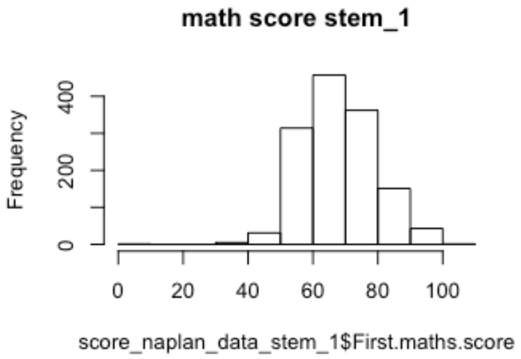
GPA_naplan_data



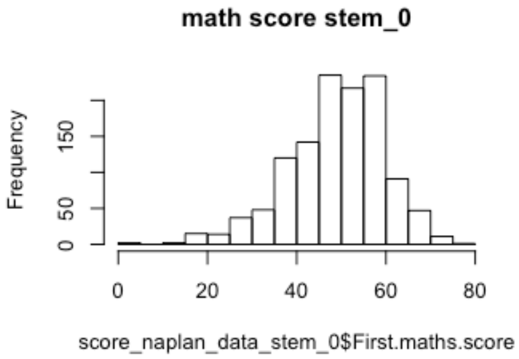
GPA_naplan_data



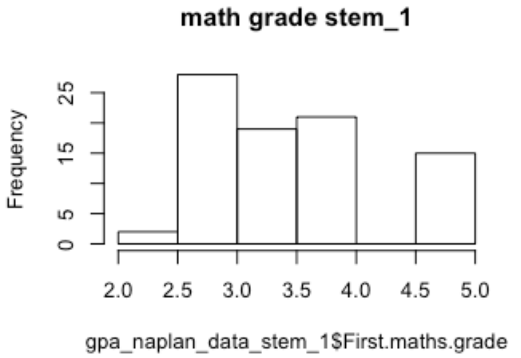
Appendix 7



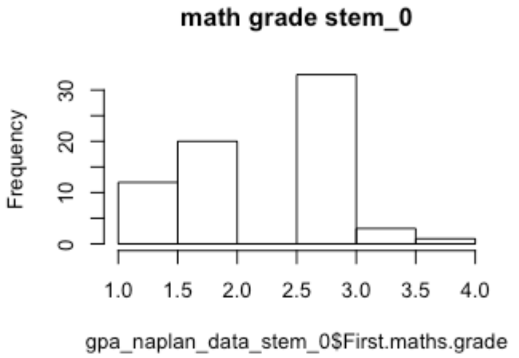
math score stem_1
Frequency
0
200
400
0
20
40
60
80
100
score_naplan_data_stem_1$First.maths.score



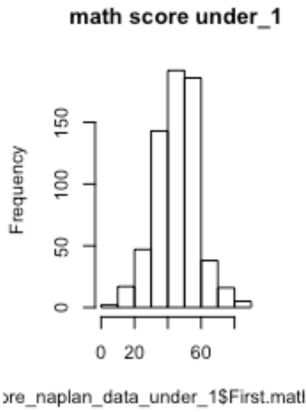
math score stem_0
Frequency
0
50
150
0
20
40
60
80
score_naplan_data_stem_0$First.maths.score



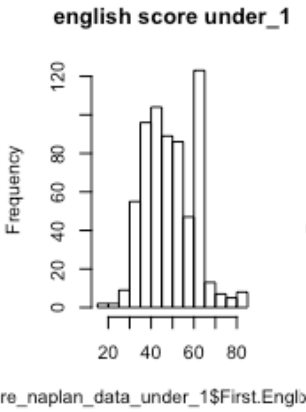
math grade stem_1
Frequency
0
5
15
25
2.0
2.5
3.0
3.5
4.0
4.5
5.0
gpa_naplan_data_stem_1$First.maths.grade



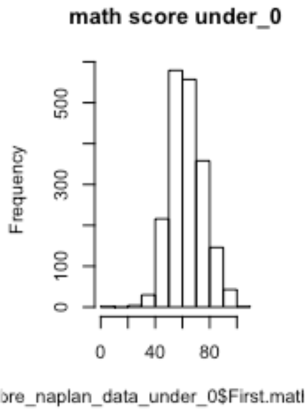
math grade stem_0
Frequency
0
10
20
30
1.0
1.5
2.0
2.5
3.0
3.5
4.0
gpa_naplan_data_stem_0$First.maths.grade



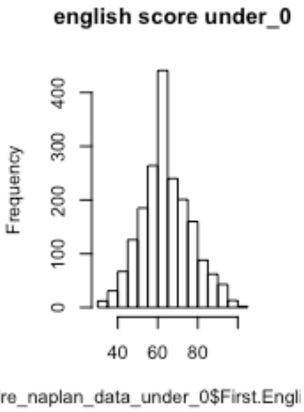
math score under_1
Frequency
0
50
100
150
0
20
60
ɔre_naplan_data_under_1$First.matl



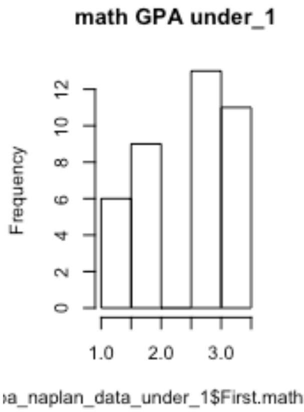
english score under_1
Frequency
0
20
40
60
80
120
20
40
60
80
re_naplan_data_under_1$First.Engl



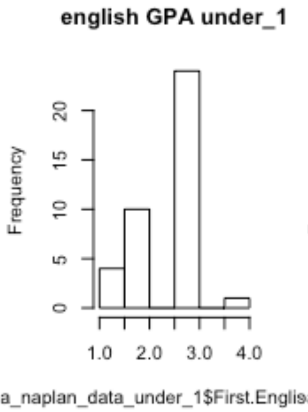
math score under_0
Frequency
0
100
300
500
0
40
80
ɔre_naplan_data_under_0$First.matl



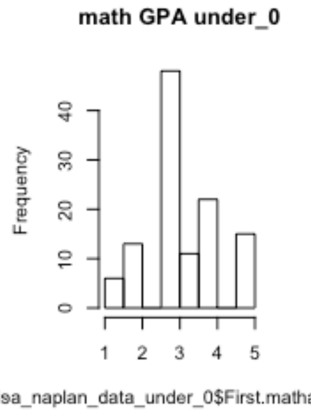
english score under_0
Frequency
0
100
200
300
400
40
60
80
re_naplan_data_under_0$First.Engli



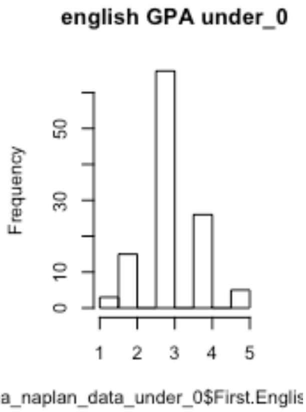
math GPA under_1
Frequency
0
2
4
6
8
10
12
1.0
2.0
3.0
ɔa_naplan_data_under_1$First.math



english GPA under_1
Frequency
0
5
10
15
20
1.0
2.0
3.0
4.0
a_naplan_data_under_1$First.Englis

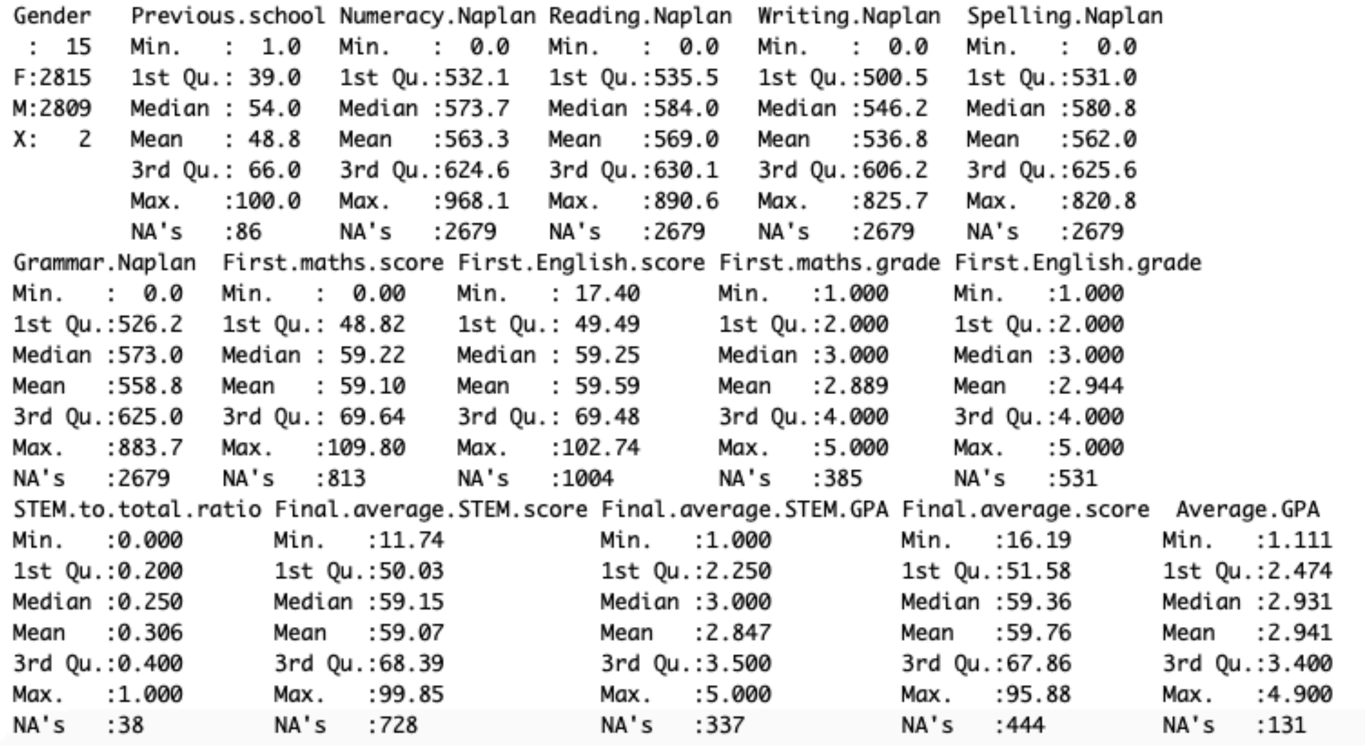


math GPA under_0
Frequency
0
10
20
30
40
1
2
3
4
5
a_naplan_data_under_0$First.math



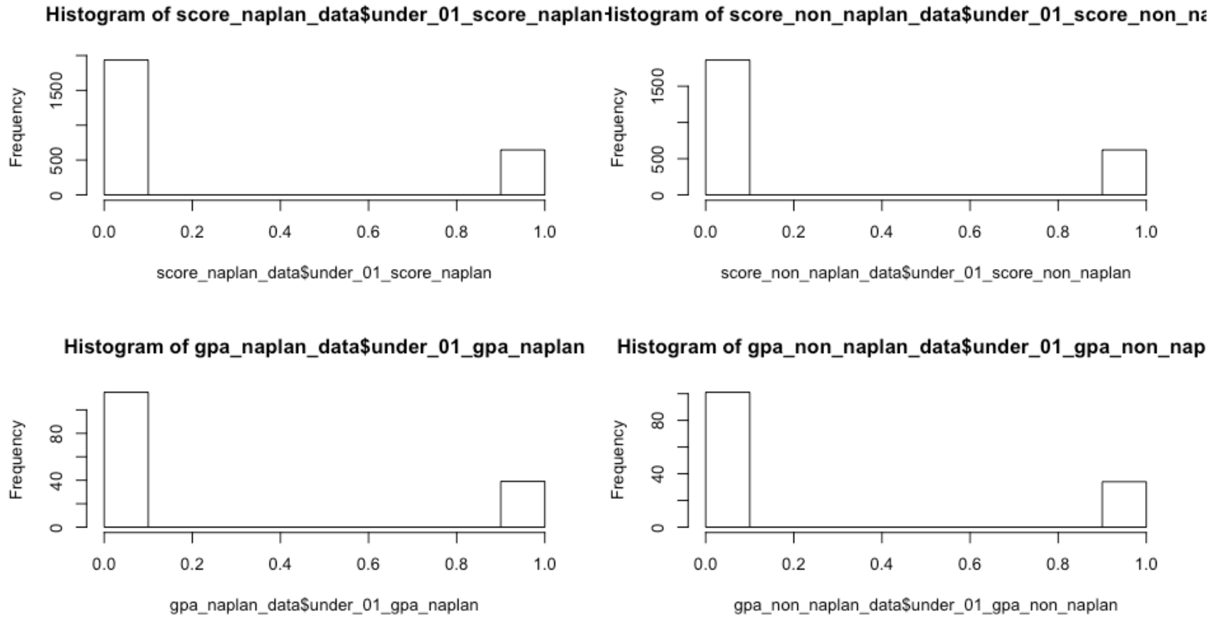
english GPA under_0
Frequency
0
10
30
50
1
2
3
4
5
a_naplan_data_under_0$First.Englis

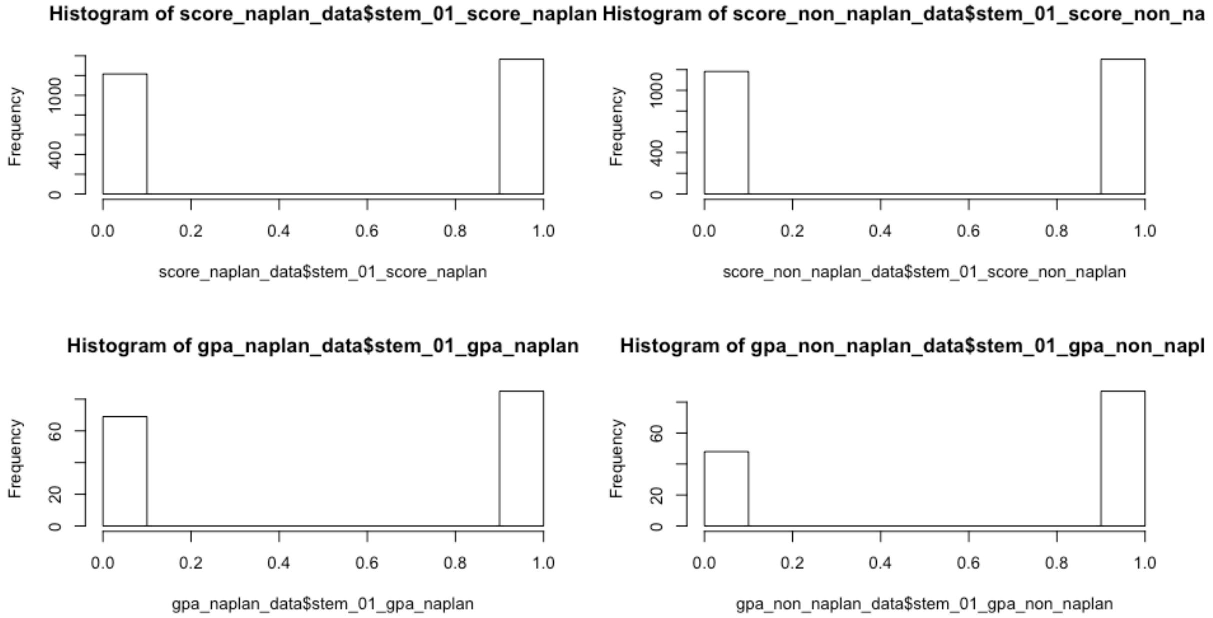
Appendix 8



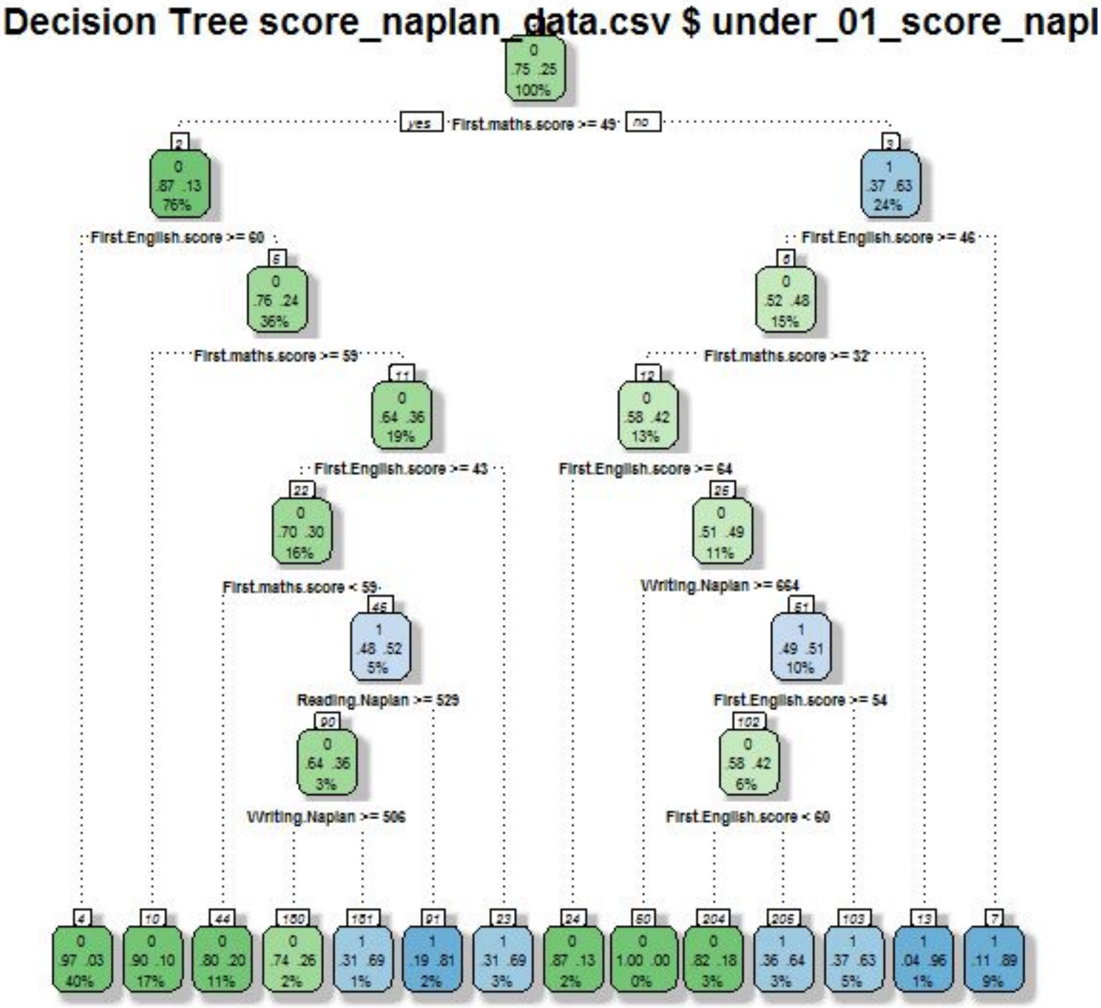
```
Gender   Previous.school Numeracy.Naplan Reading.Naplan  Writing.Naplan  Spelling.Naplan
 :  15   Min.   :  1.0   Min.   :  0.0   Min.   :  0.0   Min.   :  0.0   Min.   :  0.0  
F:2815   1st Qu.: 39.0   1st Qu.:532.1   1st Qu.:535.5   1st Qu.:500.5   1st Qu.:531.0  
M:2809   Median : 54.0   Median :573.7   Median :584.0   Median :546.2   Median :580.8  
X:   2   Mean   : 48.8   Mean   :563.3   Mean   :569.0   Mean   :536.8   Mean   :562.0  
         3rd Qu.: 66.0   3rd Qu.:624.6   3rd Qu.:630.1   3rd Qu.:606.2   3rd Qu.:625.6  
         Max.   :100.0   Max.   :968.1   Max.   :890.6   Max.   :825.7   Max.   :820.8  
         NA's   :86      NA's   :2679    NA's   :2679    NA's   :2679    NA's   :2679   
Grammar.Naplan  First.maths.score First.English.score First.maths.grade First.English.grade
Min.   :  0.0   Min.   :  0.00    Min.   : 17.40      Min.   :1.000     Min.   :1.000  
1st Qu.:526.2   1st Qu.: 48.82    1st Qu.: 49.49      1st Qu.:2.000     1st Qu.:2.000  
Median :573.0   Median : 59.22    Median : 59.25      Median :3.000     Median :3.000  
Mean   :558.8   Mean   : 59.10    Mean   : 59.59      Mean   :2.889     Mean   :2.944  
3rd Qu.:625.0   3rd Qu.: 69.64    3rd Qu.: 69.48      3rd Qu.:4.000     3rd Qu.:4.000  
Max.   :883.7   Max.   :109.80    Max.   :102.74      Max.   :5.000     Max.   :5.000  
NA's   :2679    NA's   :813       NA's   :1004        NA's   :385       NA's   :531    
STEM.to.total.ratio Final.average.STEM.score Final.average.STEM.GPA Final.average.score  Average.GPA
Min.   :0.000       Min.   :11.74            Min.   :1.000          Min.   :16.19       Min.   :1.111  
1st Qu.:0.200       1st Qu.:50.03            1st Qu.:2.250          1st Qu.:51.58       1st Qu.:2.474  
Median :0.250       Median :59.15            Median :3.000          Median :59.36       Median :2.931  
Mean   :0.306       Mean   :59.07            Mean   :2.847          Mean   :59.76       Mean   :2.941  
3rd Qu.:0.400       3rd Qu.:68.39            3rd Qu.:3.500          3rd Qu.:67.86       3rd Qu.:3.400  
Max.   :1.000       Max.   :99.85            Max.   :5.000          Max.   :95.88       Max.   :4.900  
NA's   :38          NA's   :728              NA's   :337            NA's   :444         NA's   :131    
```

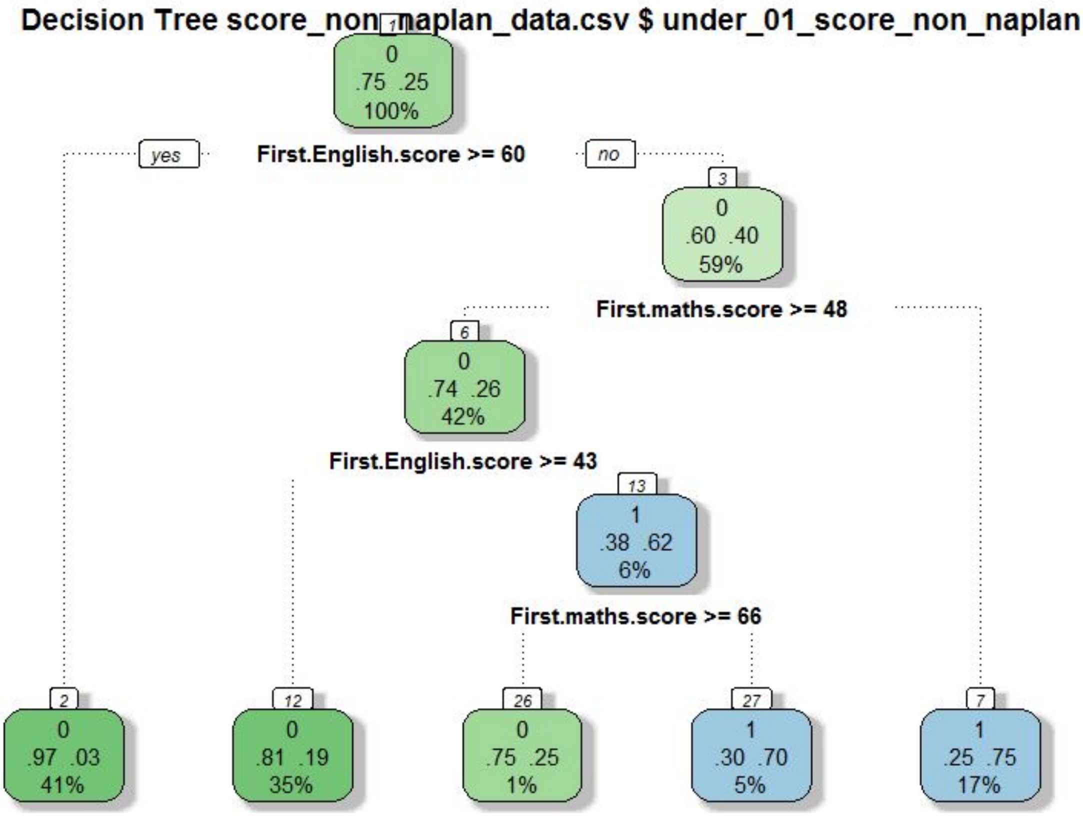
Appendix 9

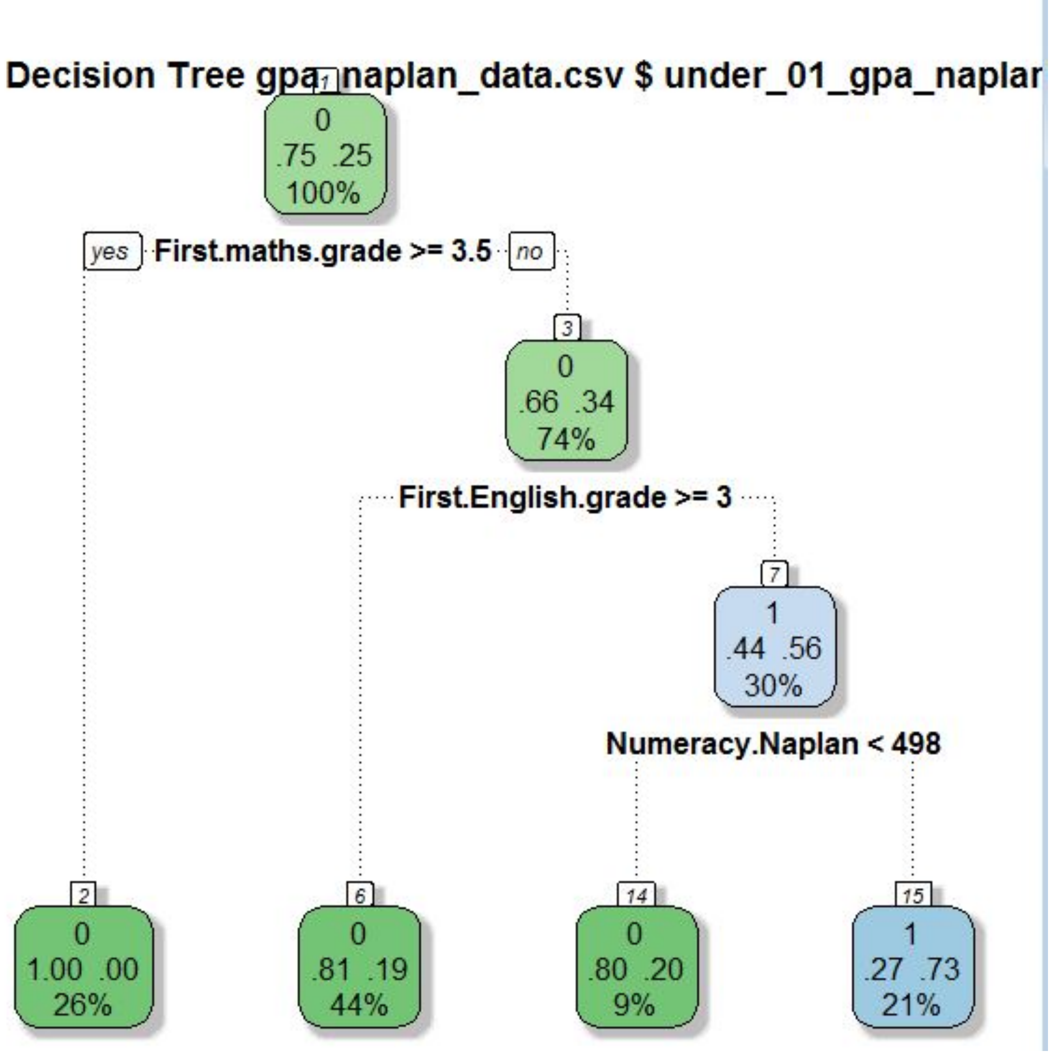

Histogram of score_naplan_data$under_01_score_naplan
Histogram of score_non_naplan_data$under_01_score_non_na
Frequency
0
500
1500
0.0
0.2
0.4
0.6
0.8
1.0
score_naplan_data$under_01_score_naplan
score_non_naplan_data$under_01_score_non_naplan
Histogram of gpa_naplan_data$under_01_gpa_naplan
Histogram of gpa_non_naplan_data$under_01_gpa_non_nap
0
40
80
gpa_naplan_data$under_01_gpa_naplan
gpa_non_naplan_data$under_01_gpa_non_naplan

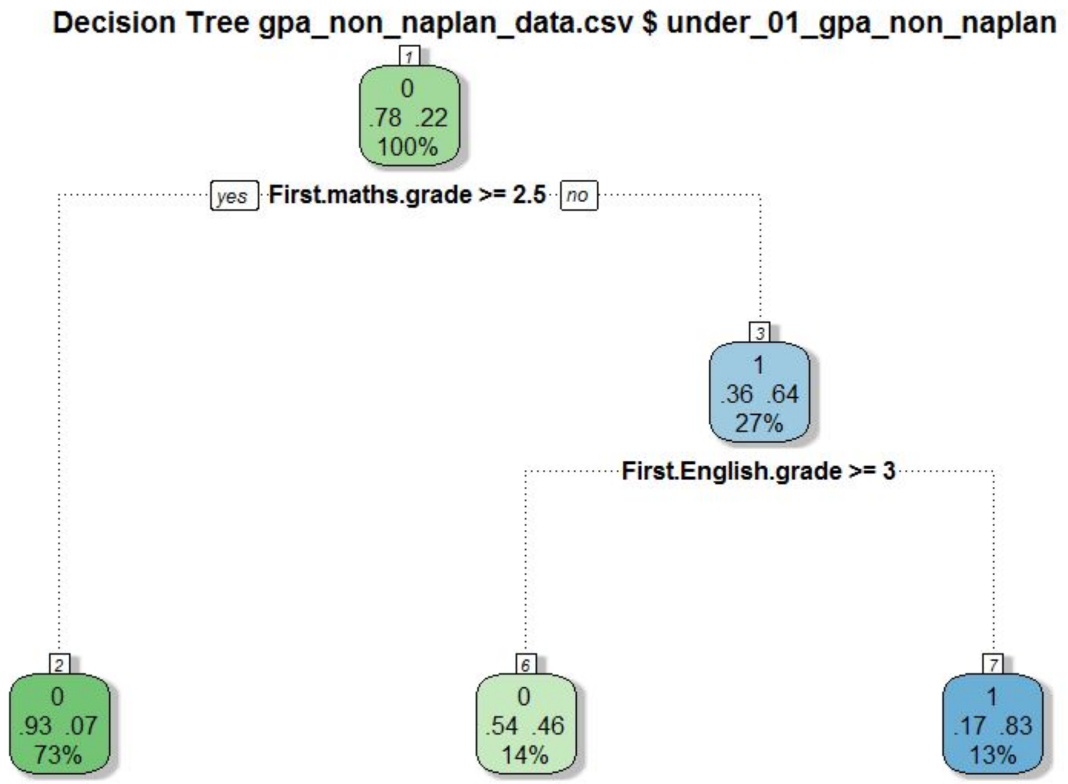

Histogram of score_naplan_data$stem_01_score_naplan
Histogram of score_non_naplan_data$stem_01_score_non_na
Frequency
0
400
1000
0.0
0.2
0.4
0.6
0.8
1.0
score_naplan_data$stem_01_score_naplan
score_non_naplan_data$stem_01_score_non_naplan
Histogram of gpa_naplan_data$stem_01_gpa_naplan
Histogram of gpa_non_naplan_data$stem_01_gpa_non_napl
0
20
60
gpa_naplan_data$stem_01_gpa_naplan
gpa_non_naplan_data$stem_01_gpa_non_naplan

Appendix 10


Decision Tree score_naplan_data.csv $ under_01_score_napl
0
.75 .25
100%
yes
First.maths.score >= 49
no
2
0
.87 .13
76%
3
1
.37 .63
24%
First.English.score >= 60
5
0
.76 .24
36%
First.English.score >= 46
6
0
.52 .48
15%
First.maths.score >= 59
11
0
.64 .36
19%
First.maths.score >= 32
12
0
.58 .42
13%
First.English.score >= 43
22
0
.70 .30
16%
First.English.score >= 64
25
0
.51 .49
11%
First.maths.score < 59
45
1
.48 .52
5%
Writing.Naplan >= 664
51
1
.49 .51
10%
Reading.Naplan >= 529
90
0
.64 .36
3%
First.English.score >= 54
102
0
.58 .42
6%
Writing.Naplan >= 506
First.English.score < 60
4
0
.97 .03
40%
10
0
.90 .10
17%
44
0
.80 .20
11%
180
0
.74 .26
2%
181
1
.31 .69
1%
91
1
.19 .81
2%
23
1
.31 .69
3%
24
0
.87 .13
2%
50
0
1.00 .00
0%
204
0
.82 .18
3%
205
1
.36 .64
3%
103
1
.37 .63
5%
13
1
.04 .96
1%
7
1
.11 .89
9%


Decision Tree score_non_naplan_data.csv $ under_01_score_non_naplan
1
0
.75 .25
100%
yes
First.English.score >= 60
no
3
0
.60 .40
59%
First.maths.score >= 48
6
0
.74 .26
42%
First.English.score >= 43
13
1
.38 .62
6%
First.maths.score >= 66
2
0
.97 .03
41%
12
0
.81 .19
35%
26
0
.75 .25
1%
27
1
.30 .70
5%
7
1
.25 .75
17%


Decision Tree gpa_naplan_data.csv $ under_01_gpa_naplar
1
0
.75 .25
100%
yes
First.maths.grade >= 3.5
no
3
0
.66 .34
74%
First.English.grade >= 3
7
1
.44 .56
30%
Numeracy.Naplan < 498
2
0
1.00 .00
26%
6
0
.81 .19
44%
14
0
.80 .20
9%
15
1
.27 .73
21%


Decision Tree gpa_non_naplan_data.csv $ under_01_gpa_non_naplan
1
0
.78 .22
100%
yes
First.maths.grade >= 2.5
no
3
1
.36 .64
27%
First.English.grade >= 3
2
0
.93 .07
73%
6
0
.54 .46
14%
7
1
.17 .83
13%