

# Predicting GPA and Interpreting Results

Brandon Chan, Lucas Duncan, Owen Macgowan

2025-03-07

## Contents

<b>1</b>	<b>Executive Summary and Contents</b>	<b>2</b>
<b>2</b>	<b>Introduction and Preprocessing</b>	<b>2</b>
2.1	The Student Performance Dataset . . . . .	2
2.2	Variable Distributions . . . . .	2
2.3	Encoded Data Frame . . . . .	4
2.4	Handling Multi-Collinearity . . . . .	4
<b>3</b>	<b>Data Exploration with Unsupervised Methods (PCA and Clustering)</b>	<b>5</b>
3.1	PCA: Generating and Visualizing Principle Components . . . . .	5
3.2	K-means Clustering: Ideal Number of Clusters . . . . .	6
3.3	K-Means Clustering on First 2 Principal Components . . . . .	6
3.4	Impact of Variables on the First Principle Component . . . . .	7
3.5	Proportion of Variance Explained (PVE) . . . . .	8
3.6	Partial Least Squares Prediction for GPA . . . . .	8
<b>4</b>	<b>Variable Selection for Supervised Methods</b>	<b>9</b>
4.1	Evaluating Criterion for Best Subset Selection . . . . .	9
4.2	Ideal Number of Variables for Each Criterion . . . . .	10
4.3	Names of Ideal Predictors . . . . .	10
<b>5</b>	<b>Regression for Identifying Important Predictors of Academic Success</b>	<b>11</b>
5.1	Choosing the Most Effective Linear Method . . . . .	11
5.2	Extracting Insight And Evaluating Chosen Linear Method . . . . .	11
5.3	Non-linear Regression: GAM . . . . .	12
5.4	Model Comparison . . . . .	12
<b>6</b>	<b>Linear Classification for Evaluating Predictive Accuracy of GPA Predictors</b>	<b>13</b>
6.1	LDA Model . . . . .	13
6.2	Logistic Model . . . . .	16
<b>7</b>	<b>Classification Analysis and Visualization with Tree-Based Methods and SVM</b>	<b>17</b>
7.1	Decision Tree . . . . .	17
7.2	Random Forest . . . . .	17
7.3	SVM Kernel Analysis . . . . .	18
7.4	Deep Learning for Analyzing Specific GPA Scores . . . . .	18
<b>8</b>	<b>Conclusions and Future Research</b>	<b>19</b>
8.1	Conclusions . . . . .	19
8.2	Future Research . . . . .	19

# 1 Executive Summary and Contents

Index	Category	Method
1	Introduction and Tidying	Importing, Wrangling and Encoding
2	Unsupervised Learning and Analysis	PCA and Clustering
3	Preprocessing	Variable Selection
4	Supervised Regression	Linear and Non-Linear Regression
5	Supervised Linear Classification	LDA and Logistic Regression
6	Supervised Non-Linear Classification	Tree-Based Methods and Support Vector Machines
7	Deep Learning	Neural Network
8	Conclusion and Future Research	Written Description

## 2 Introduction and Preprocessing

By analyzing different factors that contribute to student performance, we aim to provide insight into how institutions can improve educational and environmental factors that foster student success. This is an issue that we can relate to directly. As students ourselves, we understand that success in academics is influenced by several factors that include preparation and ability, but go beyond to include factors outside of school.

### 2.1 The Student Performance Dataset

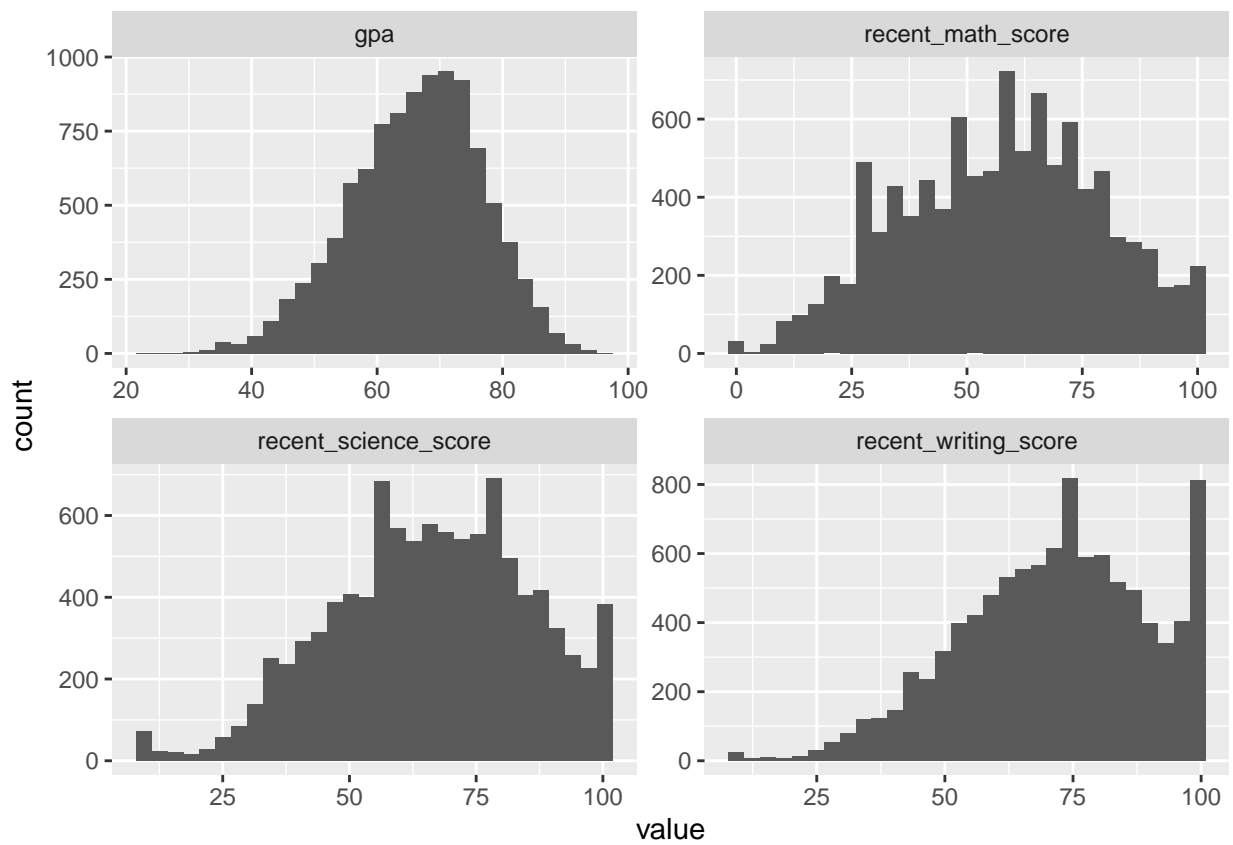
The dataset consists of attributes that may have an impact on performance in school. There are academic variables that include grades in different subjects and other variables like ethnicity, gender, and economic status of the student families. There are options for regression and classification problems with this data. With the given variables, GPA seems like the most likely response variable to represent student performance. After some wrangling, gpa\_letter, gpa, and gpa\_desirable represent different forms of the possible response variable to predict.

We start by cleaning our data; renaming columns for clarity, converting column types, generating attributes, and filtering rows and columns. We filter duplicate rows, and interpolate using column means for missing numerical variables, we fill variables where possible to fix inconsistencies between gpa and gpa\_letter, then drop the remaining rows missing key factor variables.

	is_male	race_ethnicity	parent_education	education_subsidized	education_prep	math_score	science_score	reading_score	writing_score	gpa_letter	gpa	gpa_desirable
1	D	some college	1	1	89	85	26	C	59.5	TRUE		
1	B	high school	1	0	65	67	96	A	82.0	FALSE		

### 2.2 Variable Distributions

We now want to get more familiar with our data, looking first at the numeric columns, we observe all seem approximately normal, which is to be expected given their nature. We note they all seem to be within acceptable range (0,100). We next analyze the boolean columns, none of which have one boolean option which dominated the data. Notably there seems to be an approximately equal number of males and females, which is a good sign that our sample is proportional to an entire group. Finally we analyze the true factors, education and race, and once again no one factor dominates the data, while factors are not exactly normally distributed, nor should they be, there is healthy variation between rows with no one factor representing too few rows.





	VIF.Original	VIF.with.Dropped.Ref.
race_ethnicityE	321.711292	1.526049
race_ethnicityA	238.054413	1.827008
race_ethnicityB	385.188891	2.071960
race_ethnicityC	517.000750	2.004011

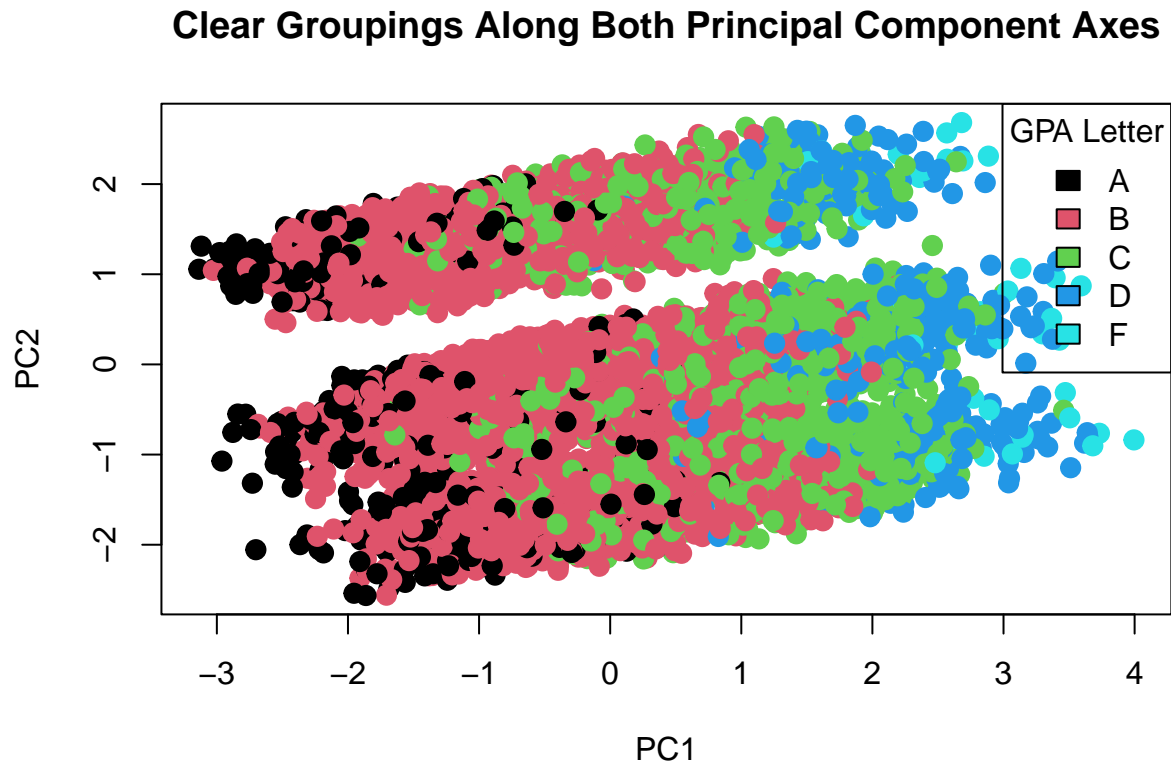
```
{r. include=FALSE} df4<- df4 |> select(-c("race_ethnicityn", "race_ethnicityE"))
```

### 3 Data Exploration with Unsupervised Methods (PCA and Clustering)

Before attempting to make any conclusive analysis, we want to get a better understanding of the data and relationships between variables. We start this process with unsupervised learning to understand variables that cause significant variance and groupings of observations.

#### 3.1 PCA: Generating and Visualizing Principle Components

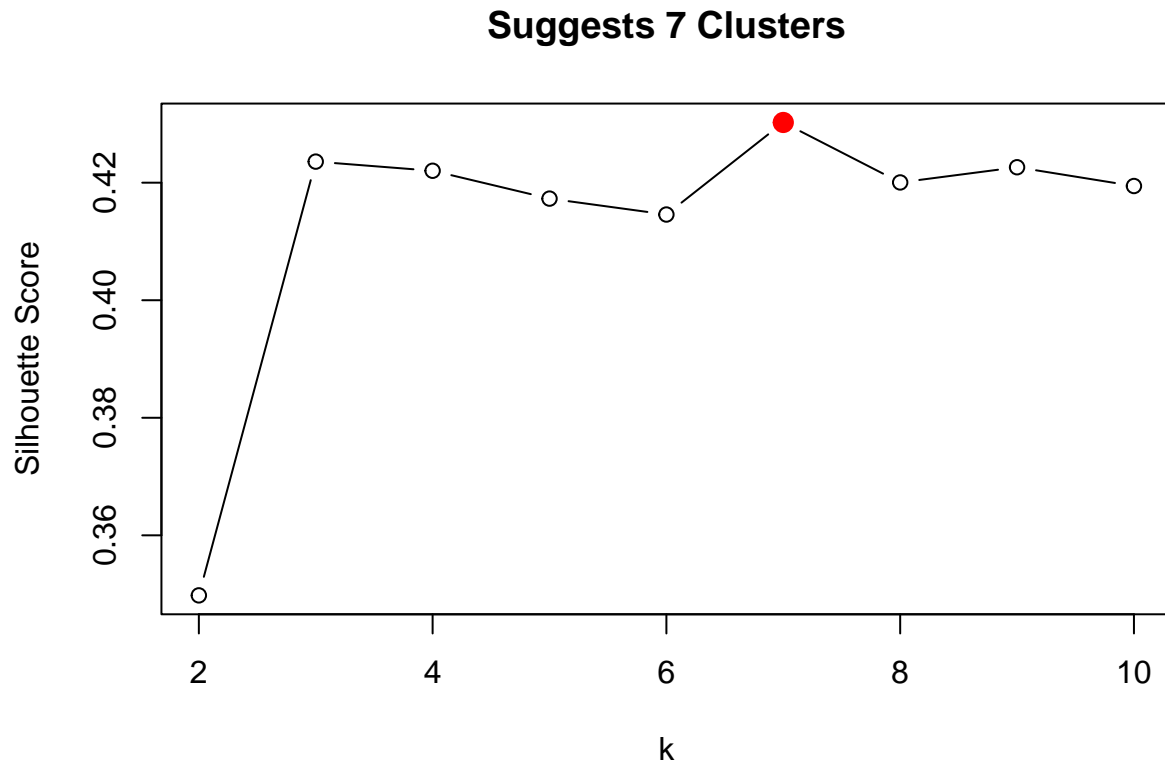
We begin exploration by generating the principal components to create a 2-dimensional visualization. The first 2 principal components represent uncorrelated linear combinations that capture the most variance in the data. We plot them to see if any patterns emerge.



There appear to be 3 distinct groupings in the data, separable by the second principal component. Based on the GPA letter coloring of points, lower values of the first principal component appear to relate to higher GPA letters and vice-versa.

### 3.2 K-means Clustering: Ideal Number of Clusters

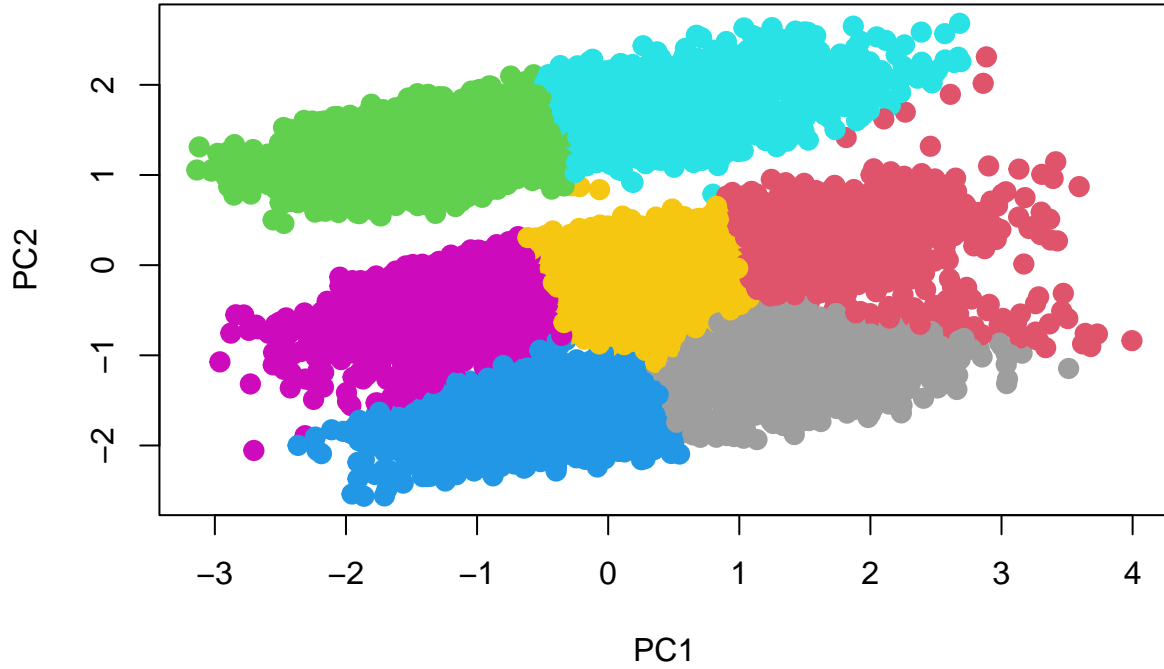
In order to check our theory on the distinct groupings of the first two principal components, we will employ k-means clustering. But first, we need to determine the optimal number of clusters to effectively show potential classes.



A peak in the silhouette score appears where there are 4 different clusters, an ideal number where points are similar within a cluster and different from other clusters.

### 3.3 K-Means Clustering on First 2 Principal Components

We color the points on the same principal component graph, this time by the cluster that they belong to.



Clusters appear to be clearly separated by both principal components, indicating that our hypothesis that the first component significantly splits the data into groups and academic feature such as grades may impact the second principal component appear plausible.

### 3.4 Impact of Variables on the First Principle Component

We wanted to take a closer look at specific features in addition to the overall structure of the data and principal components. So we took the top ten largest loadings (coefficients) of the first principal component. The goal is determine which variables are responsible for variance within the feature set.

Variable Names	PC1 Loading
is_male	0.5916517
recent_math_score	-0.4443770
recent_writing_score	-0.4309222
recent_science_score	-0.3182190
race_ethnicityC	-0.1935130
has_subsidized_lunch	-0.1715074
parent_educationsome high school	-0.1561814
race_ethnicityD	0.1520213
race_ethnicityA	-0.1132013
parent_educationmaster's degree	0.1060432

Having a subsidized lunch and staying prepared appear to be the most significant features impacting variance in the data. Although a subsidized lunch has no obvious direct relationship to academics, it appears that socio-economic factors may be at play in some fashion. Taking an extra preparation course also seems to

have a significant impact in the feature space. Preparation likely has a direct relationship with academic performance features.

### 3.5 Proportion of Variance Explained (PVE)

Only the first principal component was analyzed, and other principal components may tell very different stories about which features may be significant. We show how much each additional principal component contributes to the proportion of variance explained:

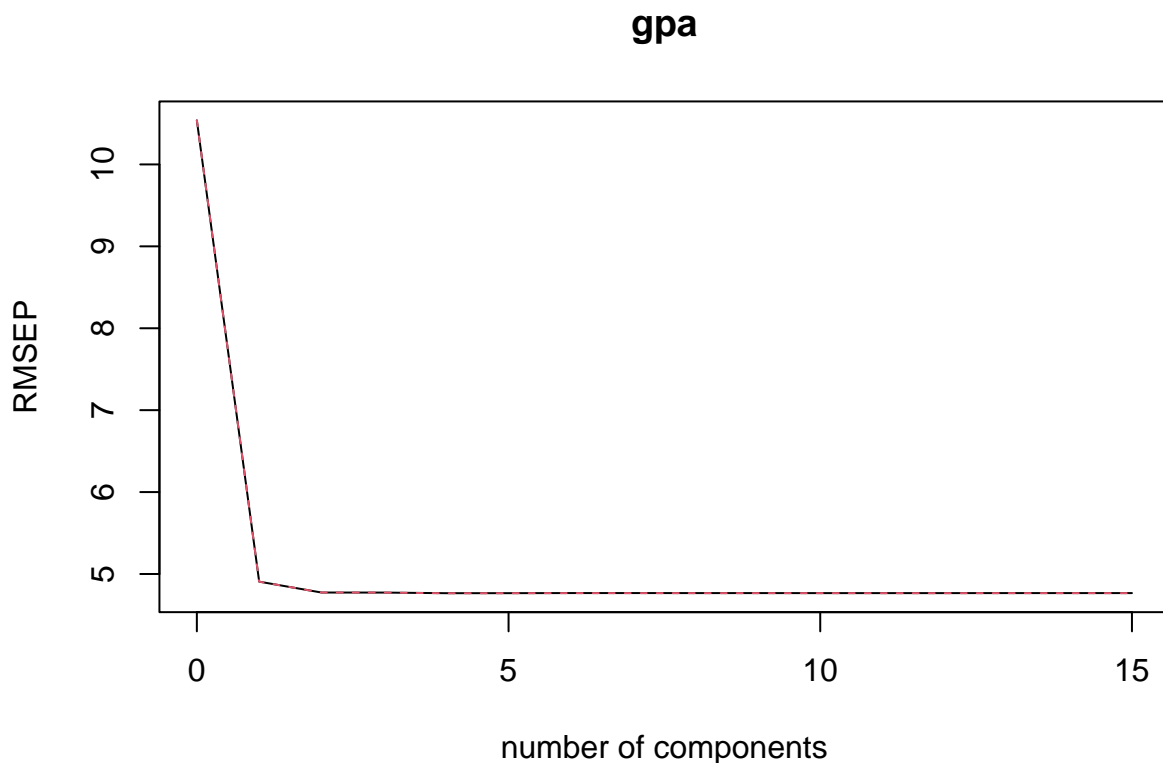
Number of Principal Components	PVE	Cumulative PVE
1	0.0922817	0.0922817
2	0.0814607	0.1737425
3	0.0764538	0.2501962
4	0.0740256	0.3242218
5	0.0733677	0.3975896
6	0.0704879	0.4680774
7	0.0685702	0.5366476
8	0.0673422	0.6039899
9	0.0631607	0.6671506
10	0.0599938	0.7271443

Additional principal components contribute significantly to higher PVE. The first principal component does not contribute to much of the variance explained and the second principal component does not increase it significantly either. We interpret this to mean that there are many notable linear combinations of features rather than a one that significantly explains variance the feature space.

### 3.6 Partial Least Squares Prediction for GPA

Our analysis on the principal components is predicated on the assumption that they having meaning in relation to evaluating academic performance. We evaluate this by using partial least squares to predict GPA.





3 components appears optimal based on the PVE and plot.

RMSE
4.729063

We will compare the error with other methods later in the analysis to determine how effective the principal component are at predicting GPA, our measure for academic performance.

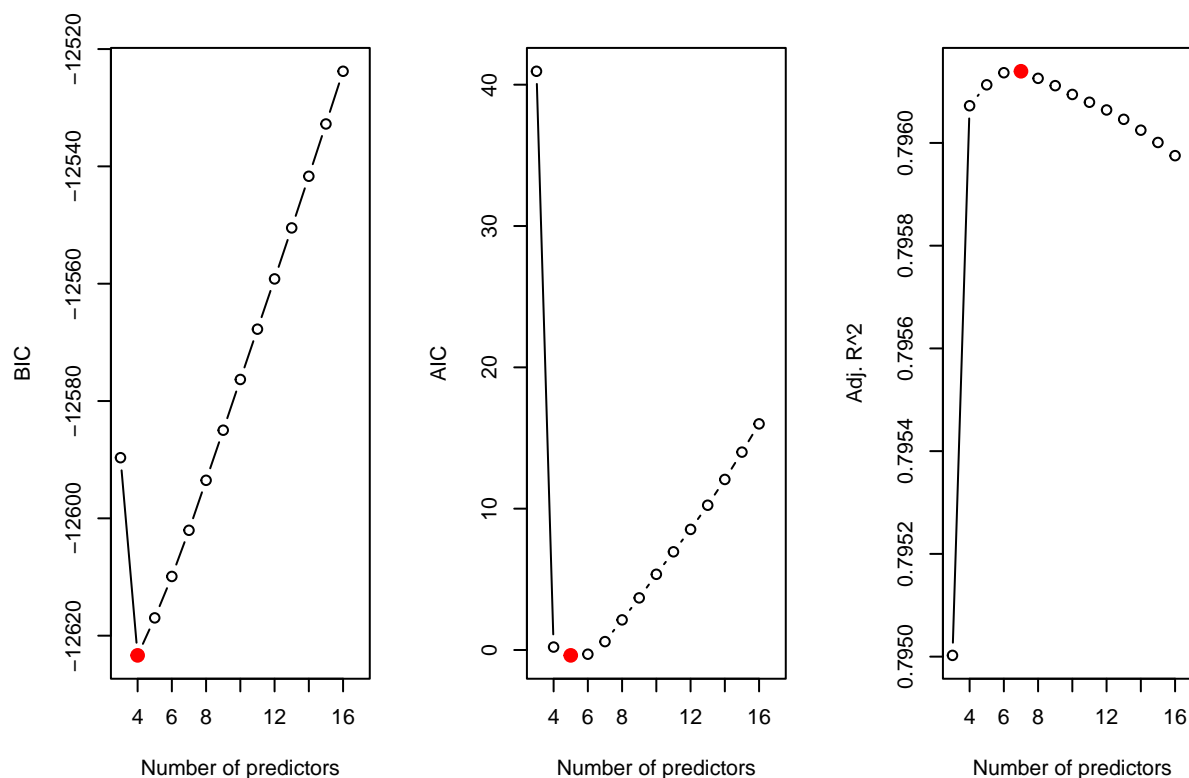
## 4 Variable Selection for Supervised Methods

Before proceeding to predict GPA with supervised linear methods, we want to ensure that predictor variables used are relevant. It could also be helpful to analyze which predictors contribute to academic success. This will help to determine what factors institutions and individuals can improve to aid in academics.

### 4.1 Evaluating Criterion for Best Subset Selection

Since the dataset is not too large (~10000 rows), we feel that using best-subset selection is computationally feasible and is the optimal method for variable selection. 3 types of evaluation criterion are employed to determine the ideal variables; AIC, BIC, and adjusted R-squared.

**## Reordering variables and trying again:**



The best number of predictors appears different but fairly similar given each criterion. BIC heavily penalizes additional variables, so it is expected to favour a lower number of variables. AIC appears to prefer one more variable and few more variables are favoured with  $R^2_{adj}$ .

## 4.2 Ideal Number of Variables for Each Criterion

Best_BIC	Best_Cp	Best_Adjusted_Rsq
4	5	7

Since BIC heavily penalizes extra predictors and adjusted  $R^2_{adj}$  appears to be more lenient, we choose to use 6 predictors, a value in the middle.

## 4.3 Names of Ideal Predictors

The following variables contribute the most according to best subset selection:

---

ChosenPredictors

---

is\_male  
has\_prepared  
recent\_math\_score  
recent\_writing\_score  
recent\_science\_score  
parent\_educationsome high school

---

We update train and test sets to only includes these predictors.

## 5 Regression for Identifying Important Predictors of Academic Success

With the selected variables, we can now test how well they predict academic performance measured in GPA. We start by training supervised regression models to predict the GPA score.

### 5.1 Choosing the Most Effective Linear Method

The base linear regression model, as well as linear models with LASSO and Ridge regularization are created and compared.

Loss Type	Unregularised Linear	LASSO	Ridge
RMSE	4.714030	4.714620	4.738995
MAE	3.874895	3.876661	3.899528

Interestingly, the unregularized model outperforms LASSO and Ridge regression. This may be because we performed regularization after variable selection, so irrelevant predictors were already filtered out. Additional penalties on predictors probably only interfered with making the best predictions that minimize SSE.

### 5.2 Extracting Insight And Evaluating Chosen Linear Method

Before making conclusive analysis, we want to ensure that the model is predicting accurately on test data.

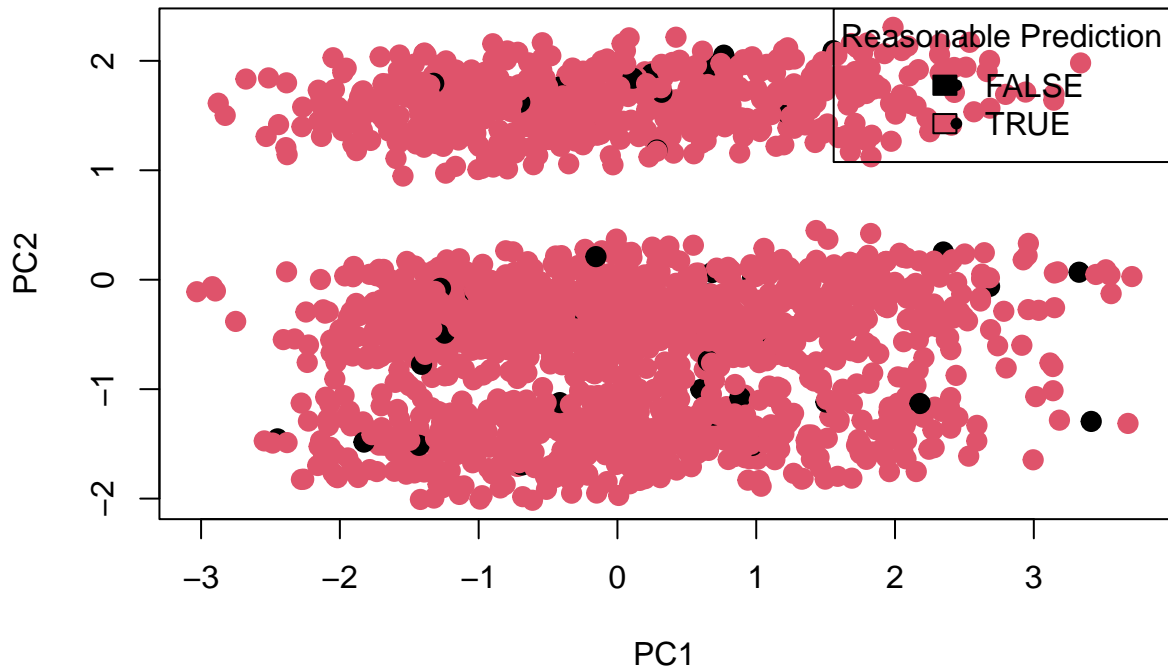
#### 5.2.1 Accuracy Within 10%

Since the linear regression model only predicts continuous values, it wouldn't make sense to calculate accuracy directly, but we still want to ensure that the model is accurate. We resolve this by predicting how frequently predictions are within 10 points of the true GPA.

```
## [1] "Accuracy within 10 GPA points: 0.977409638554217"
```

The vast majority of predictions are within 10 points of the true GPA, so we feel that analysis on the impact of each attribute will be meaningful.

## Nearly All Predictions are Within 10% of True GPA



Since the model is fairly accurate at predicting GPA, we feel that analyzing coefficients directly can provide insight into how each predictor contributes to success.

```
##              (Intercept)                is_male
##              15.4983645                0.7684885
##              has_prepared            recent_math_score
##              0.1515244                0.2453538
##              recent_writing_score    recent_science_score
##              0.2604180                0.2662900
## parent_educationsome_high_school
##              0.2246999
```

Of the encoded variables, gender appears to be the most significant based on the raw coefficient, followed by parent education and preparation. These values suggest that societal factors meaningfully impact academic success, potentially because of different educational opportunities. Pure academic scores appear to contribute similarly to each other. This may be because students successful in one subject tend to be successful overall.

### 5.3 Non-linear Regression: GAM

Although the linear model was successful, but there remains a possibility that prediction is more accurate with non-linear transformations applied to variables. We employ a generalized additive model with splines determined by cross-validation.

### 5.4 Model Comparison

At this point we have the best supervised linear, non-linear, and unsupervised models. By comparing each of them, we hope to obtain the best one and to understand why it performed the best. Once the model is understood, we will attempt to relate the model to real-world academic performance.

Measurement	GAM	LinearModel	PLS
RMSE	4.7125372	4.7140299	4.7290633
R2	0.8056238	0.8055003	0.8042583

The generalized additive model performs the best in terms of RMSE. Non-linear splines applied to predictors appear to improve the model, indicating that GPA is not best determined by a linear combination of features. The PLS predictions formed by the first 3 principal components perform slightly worse, though not by a drastic margin. Reducing dimensionality or the number of variables seems to help the model in all methods.

Based on the results of applying regression to predict GPA, societal and educational factors both contribute to academic success.

## 6 Linear Classification for Evaluating Predictive Accuracy of GPA Predictors

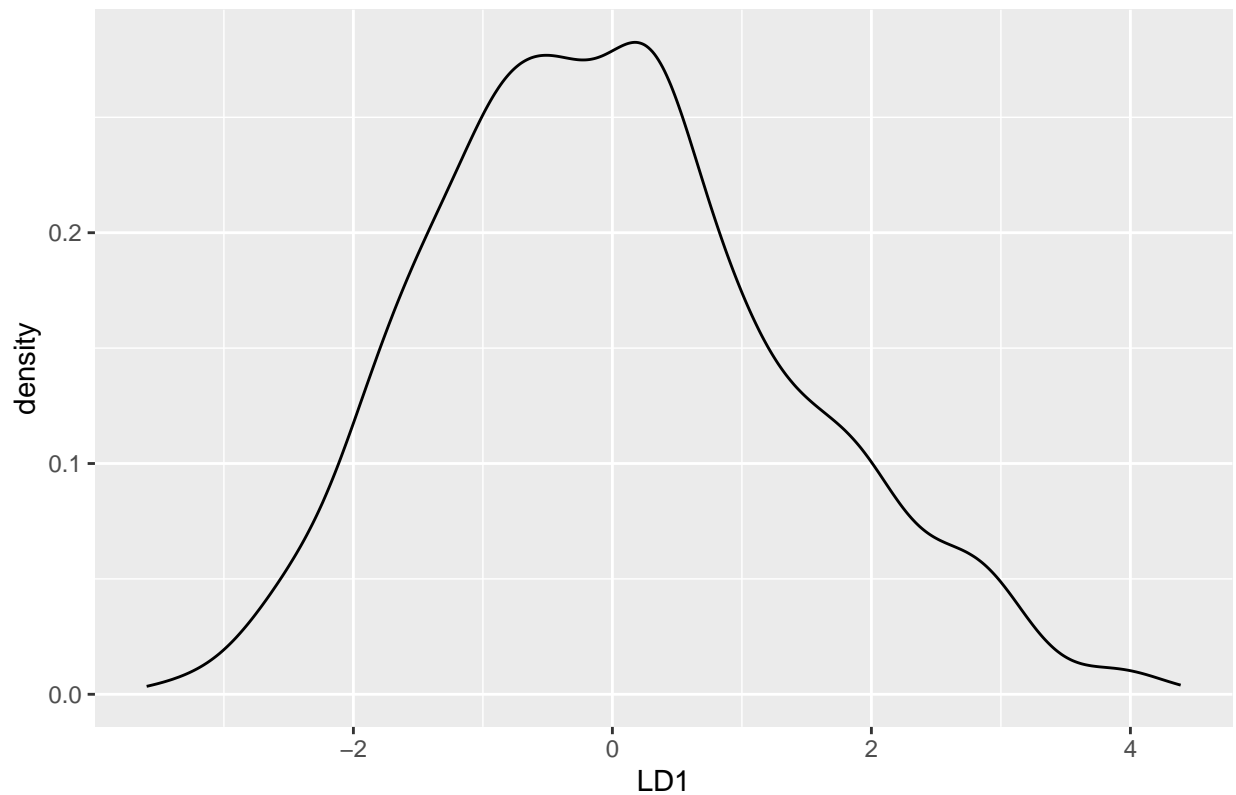
The regression models allowed us show importance within the model of specific variables by analyzing coefficients, but we were unable to compute easily interpretable metrics such as accuracy directly. The purpose of our analysis is to translate our findings to the real world, and we think that a classification perspective will provide more interpretable results. Our goal in this section is to determine whether a linear decision boundary can differentiate between students having a desirable GPA.

### 6.1 LDA Model

The first classification method we attempt is linear discriminant analysis. We train it using the same predictors determined by variable selection.

We plot the distribution of LD1 discriminant scores to look for skew or multimodal properties.

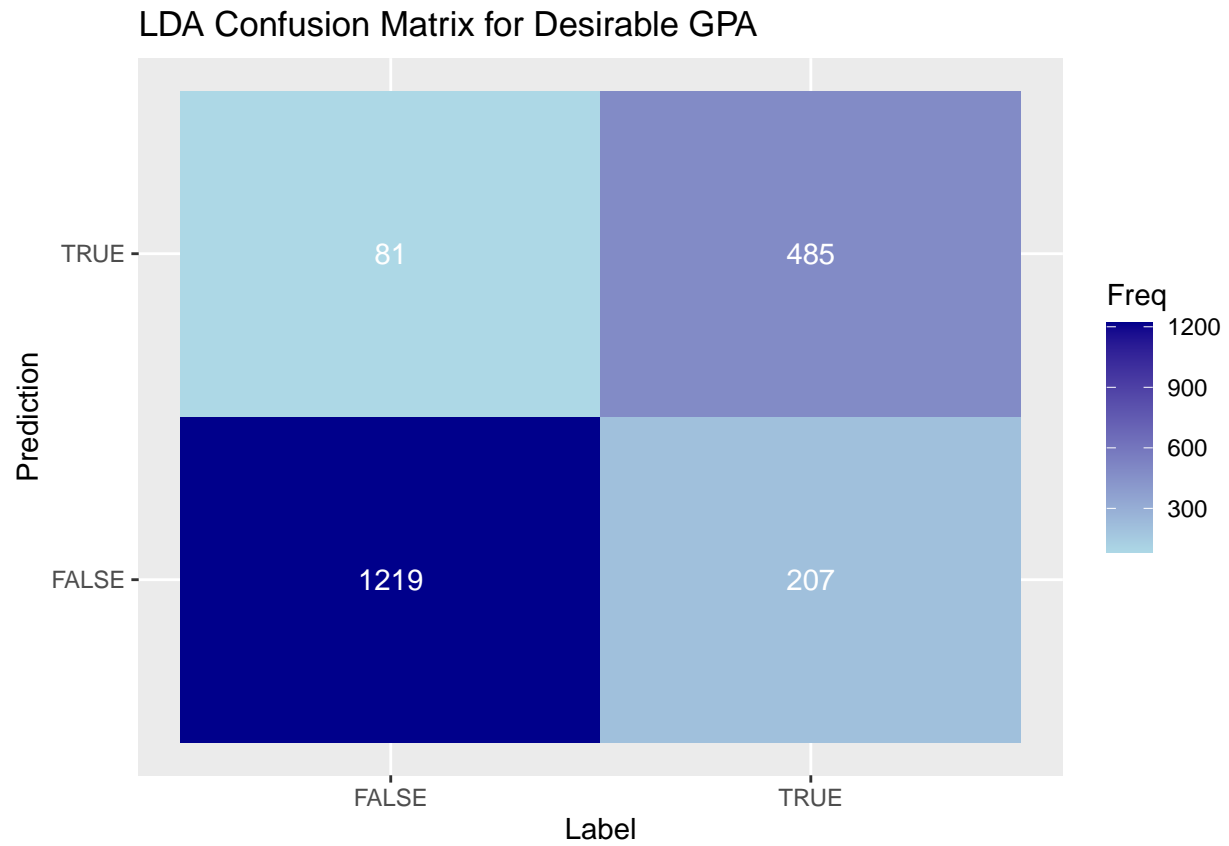
LD1 Discriminant Score Distribution



There appears to be a multimodal peak representing each class although it is very subtle. The distribution also appears to have a slight skew right. We interpret these properties to mean that there is a noticeable distinction between classes and there are more extreme large discriminant scores. This could mean that some students are strongly predicted to have a good GPA given the predictors.

We evaluate the LDA model with a confusion matrix and related metrics.

### 6.1.1 Confusion Matrix



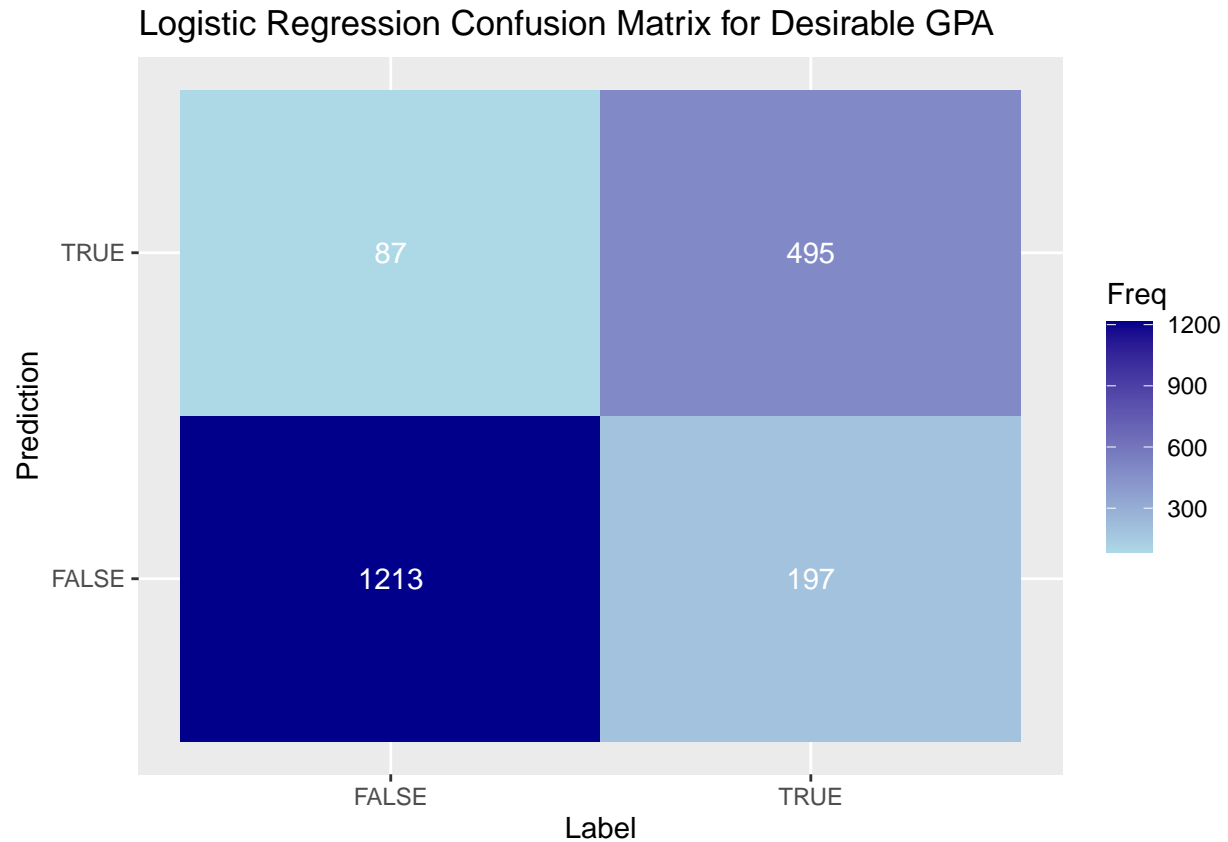
### 6.1.2 Metrics

Metric	Score
Accuracy	0.8554217
Sensitivity	0.9376923
Specificity	0.7008671

The LDA model seem pretty strong based on the confusion matrix and metrics. It seems to predict the true class very well, but is not as good at predicting true negatives. This may be due to a class imbalance.

## 6.2 Logistic Model

### 6.2.1 Confusion Matrix



### 6.2.2 Metrics

Metric	Score
Accuracy	0.8574297
Sensitivity	0.9330769
Specificity	0.7153179

The logistic regression model appears slightly superior to the LDA model at predicting the negative class, but results are fairly similar overall.

### 6.2.3 Plotting Logistic Regression Model Evaluation

Based on the high accuracy of the linear classifiers, a linear decision boundary can effectively differentiate between students with a good GPA and a poor GPA. This shows that the predictors are indicative of academic performance and if students are able to improve some of the attributes, they are likely to be more successful in school.



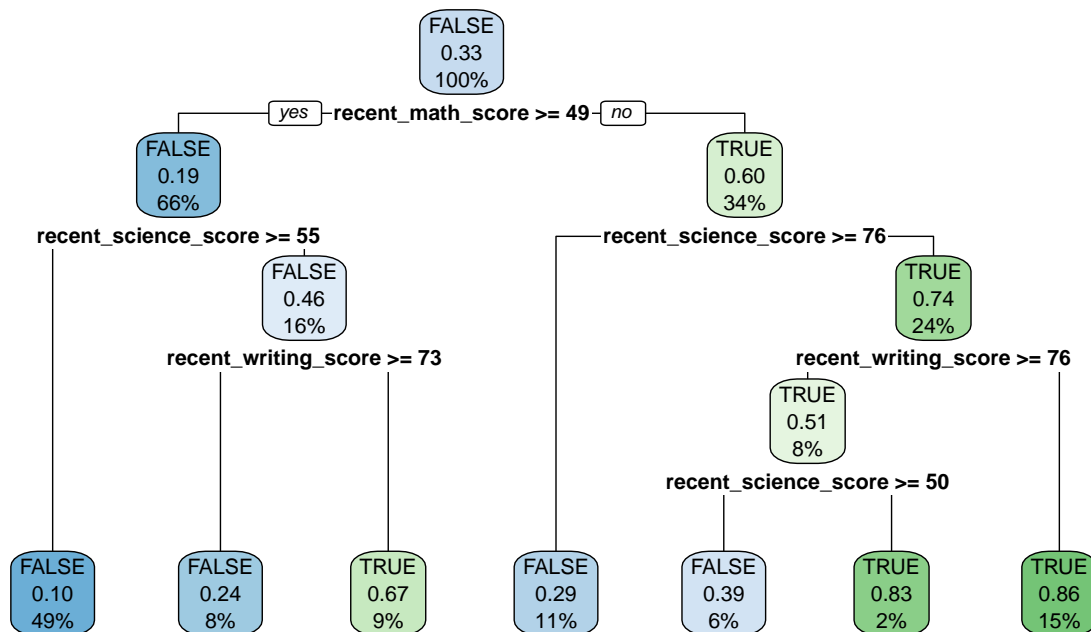
## 7 Classification Analysis and Visualization with Tree-Based Methods and SVM

We now know that the predictors can effectively distinguish between strong and weak academic performers, but other methods may be more accurate and provide more insight into how predictions are being made.

### 7.1 Decision Tree

The decision tree is particularly useful for our case because it is interpretable, and we want to show what predictive variables are causing students to perform well in school.

#### Tree for Classifying Strong Academic Performance

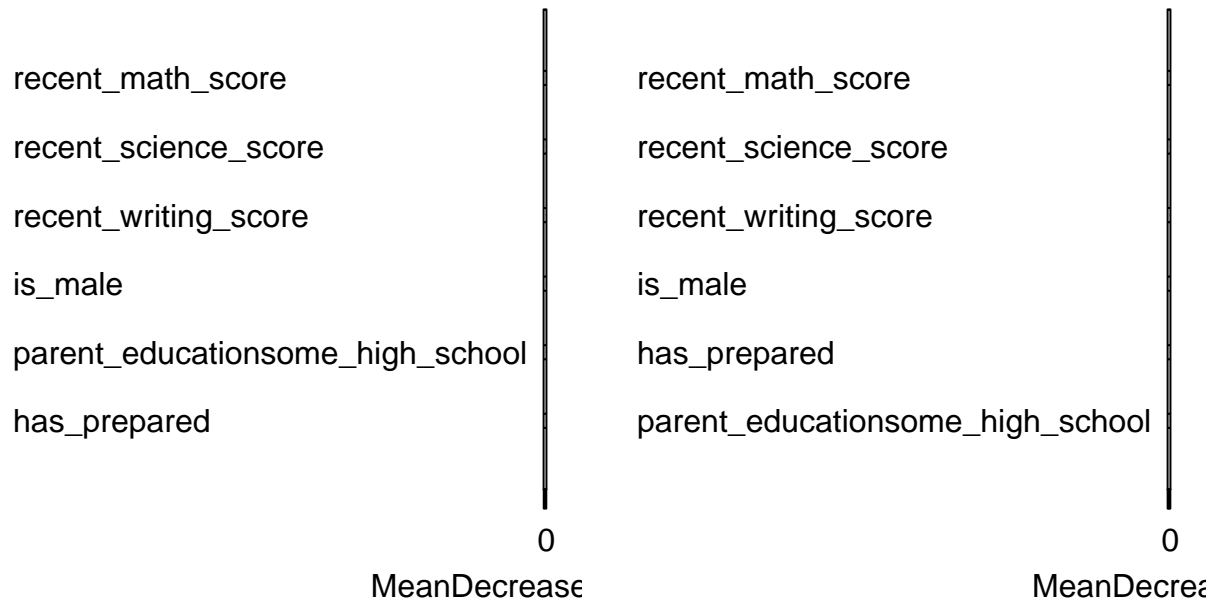


The decision tree with splits based on information gain heavily favours academic scores for determining student performance. Among them, it considers math the most significant split with the highest information gain. Many areas of science rely on a strong math background, which could explain why math score is the first split.

### 7.2 Random Forest

We follow up the decision tree by training a random forest. Since we are attempting to extract inference rather than maximize predictive accuracy, we do not show evaluation metrics on test data. Instead, we plot variable importance to show which predictive variables strongly considered by the random forest model.

## Random Forest Variable Importance



The variable importance plot strengthens the hypothesis that decision trees and random forest heavily favour the academic predictor variables.

### 7.3 SVM Kernel Analysis

We apply the kernel trick to train several support vector machines in order to analyze the ideal shape of the decision boundary.

Kernel.Type	Misclassification
Linear	0.1455823
Radial	0.1445783
Quadratic	0.3283133

The support vector machine with the radial kernel performs the best, followed closely by the linear kernel and distantly by the quadratic kernel. Since both linear and radial kernels perform well, two seemingly contradictory results emerge. The linear decision boundary suggests a simpler relationship between predictor variables to classify academic performers, while the strong performance of the SVM with the radial kernel suggests a more complex one. In the context of academic performance, this may mean that societal and academic predictive factors can be used in a straightforward manner to distinguish strong academic performers, but there are complex relationships between variables that are better separated with a complex decision boundary.

### 7.4 Deep Learning for Analyzing Specific GPA Scores

From analyzing different classification methods and kernels, we determined that complex relationships exist between predictors of academic success. We also previously identified that strong academic performers can be

accurately identified, but specificity tends to be far worse than sensitivity. This leads us to think that certain GPA scores are being misclassified consistently, while others are being classified well. We attempt to verify the claim by creating a deep learning model with softmax activation for multiclass classification output.

```
## [1] "A  accuracy:  11.05 %"
## [1] "B  accuracy:  77.39 %"
## [1] "C  accuracy:  74.45 %"
## [1] "D  accuracy:  72.22 %"
## [1] "F  accuracy:  11.11 %"
```

It appears that the neural networks classifies decent grades (B) very well. The model predicts weak albeit passing grades fairly accurately, but performs much more poorly classifying very good or failing grades. We think this is largely a consequence of the predictors; they are able to identify students with performance at the extremes, either negative or positive. This suggests that students with an A or F GPA distinguish themselves from peers in the same circumstances as themselves. Their individuality sets them on a unique path to achieve either great academic success or failure.

## 8 Conclusions and Future Research

### 8.1 Conclusions

- DOT JOT MEANT TO BE EXPANDED (just my initial ideas -LD)
- previous individual subject marks = best indicator (which is best??)
- impact of parent education (economic + social impact? perspectives on school)
- lunch ~ income range (tends to have less access to resource, or luxury of time)

### 8.2 Future Research

OPTIONAL????????

- DOT JOT MEANT TO BE EXPANDED (just my initial ideas -LD)
- current emphasis on racial and economic factors + previous success
- expansion of analysis possible for geographic, ie between schools with varying urban/rural divide,
- possible opportunity to look at student mental health (IE: psychology, risk behaviours)
- maybe even something about access to technology, a metric about technical capacity ~ assuming technical skills give indication for success in increasingly technical school, idk, that last bit is a stretch