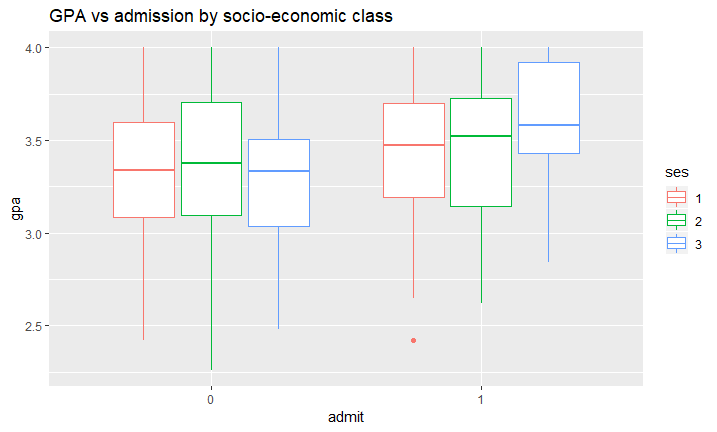
**An education department in the US needs to analyze the factors that influence the admission of a student into a college. Analyze the historical data and determine the key drivers.**

First, when reading in the data, it was noticed that the **admit**, **ses**, **Gender\_Male**, **Race** and **rank** columns needed to be made categorical.

The data was not missing any values and thus no calculations needed to be made to fill missing data and no observations needed to be dropped.

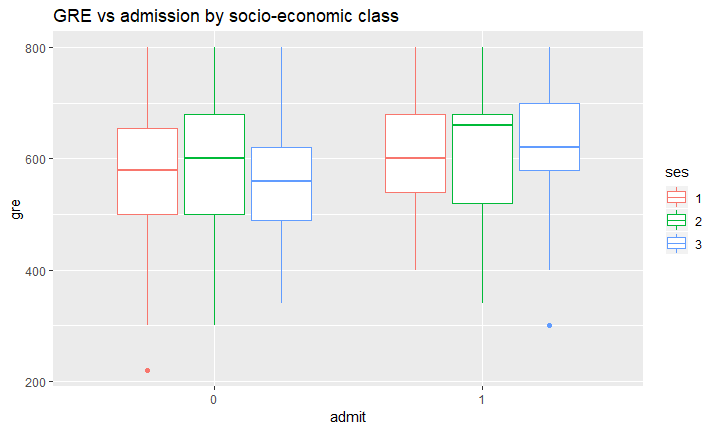
Some basic exploratory data analysis was conducted to identify any factors that could play a large role in college admission.

**GPA and socio-economic class**



GPA certainly looks like it plays a role. It looks like the average student with a GPA of 3.3 doesn’t get admitted, regardless of socio-economic class.

**GRE**



GRE looks like it may play a role as the mean grade point average seems higher amongst those accepted.

**Rank**

University Rank

Admitted 1 2 3 4

0 28 97 93 55

1 33 54 28 12

University rank appears to have an influence – more students applying to universities ranked “1” (higher) are accepted than not and a lot more students applying to universities ranked “4” (lowest) are rejected than not.

**Race**

Race

Admitted 1 2 3

0 90 94 89

1 53 35 39

Race may influence being admitted. From the dataset, a fair bit more Hispanic (1) students are being accepted to college compared to Asian (2) and African-American (3) students.

**Gender**

Male

Admitted 0 1

0 141 132

1 69 58

Gender does not appear to have a role in being admitted to a university.

Given the plots and tables above, it is expected that GPA, GRE, rank and race would be the most significant variables.

**Train and test data**

An 80-20 split was used to split the data into a training set and testing set.

The seed was set to 100 in R so that the results may be reproduced.

**Logistic regression model**

The best suited logistic regression model used the variables **GPA, rank and race**. The model had an accuracy of 0.7188 and a sensitivity of 0.2376.

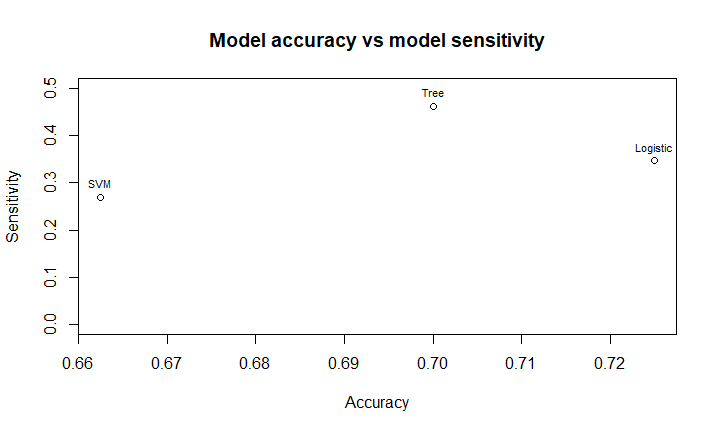
**Decision Tree classifier**

The decision tree model used the variables **GPA, GRE and rank**. It had an accuracy of 0.7719 and a sensitivity of 0.4455.

**Support Vector Machine**

The SVM model was set up with a **linear kernel**. The best fitting model with tuning had an accuracy of 0.7031 and a sensitivity of 0.25743.

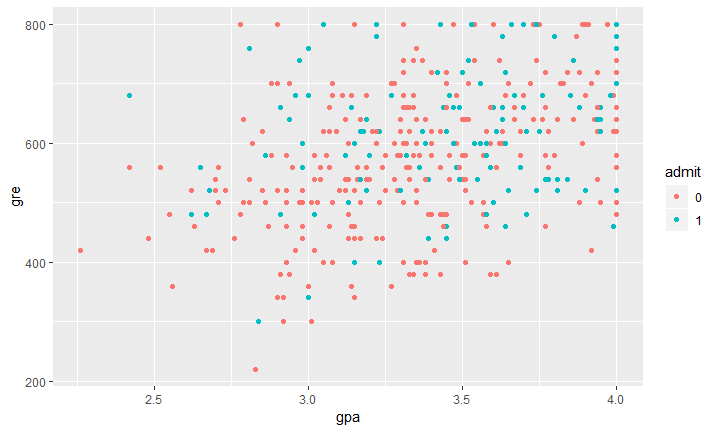
The plot summarized the best models from all three of the above methods:



It was decided that the **Decision Tree model** would be the champion model. This model had a much higher sensitivity rate than the other models and had a fairly good accuracy (> 0.7). This model seemed to strike a good balance between predicting true positives as positive and an overall accuracy.

**Other Machine Learning techniques that could be applied to binomial classification**

KNN (K-Nearest Neighbors) – KNN takes similar characteristics in variables and groups them as clusters. The clusters determine the class of the response. KNN would not be a good tool for modelling categorical data – as the observations will not be easily differentiable. If only the only continuous variables were used, the below is an example of what the clusters would look like. KNN **would not** be useful here.



Naïve Bayes Classifier – This uses Bayes theorem, we can find the probability of **A** happening, given that **B** has occurred. **B** is the evidence and **A** is the hypothesis. The assumption made here is that the predictors/features are independent. That is presence of one particular feature does not affect the other. In the dataset we have many categorical variables and gre and gpa can be categorized into bins as well. This would make naïve bayes a **useful** classifier for this problem.

Random Forest - Random Forest is a supervised learning algorithm. The “forest” it builds, is an ensemble of Decision Trees, most of the time trained with the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall result. Given that the Decision Tree was selected as the champion model and that a Random Forest is an extension of the Decision Tree algorithm, this method could result in a **useful** predictor.