COLLEGE ADMISSION PROJECT – WRITE UP

OBJECTIVE:

- To analyse the factors that influence the admission of students into colleges using different machine learning techniques
- select the best model with high accuracy rate
- Categorize the data into High, Medium and low based on given criteria

DATA:

Dataset Description:

Attribute	Description
GRE	Graduate Record Exam Scores
GPA	Grade Point Average
Rank	It refers to the prestige of the undergraduate institution. The variable rank takes on the values 1 through 4. Institutions with a rank of 1 have the highest prestige, while those with a rank of 4 have the lowest.
Admit	It is a response variable; admit/don't admit is a binary variable where 1 indicates that student is admitted and 0 indicates that student is not admitted.
SES	SES refers to socioeconomic status: 1 - low, 2 - medium, 3 - high.
Gender_male	Gender_male (0, 1) = 0 -> Female, 1 -> Male
Race	Race – 1, 2, and 3 represent Hispanic, Asian, and African-American

ANALYSIS TASK:

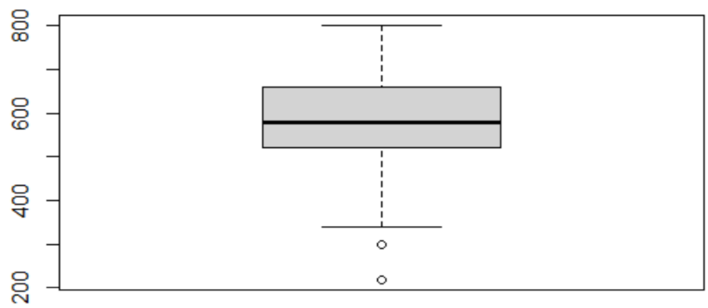
1. Missing Values:

No missing values in the data

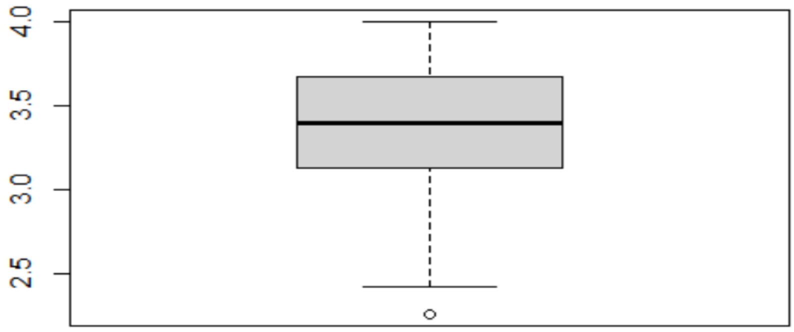
2. Outliers:

Outliers found in GPA, GRE variables using boxplot analysis

- GRE



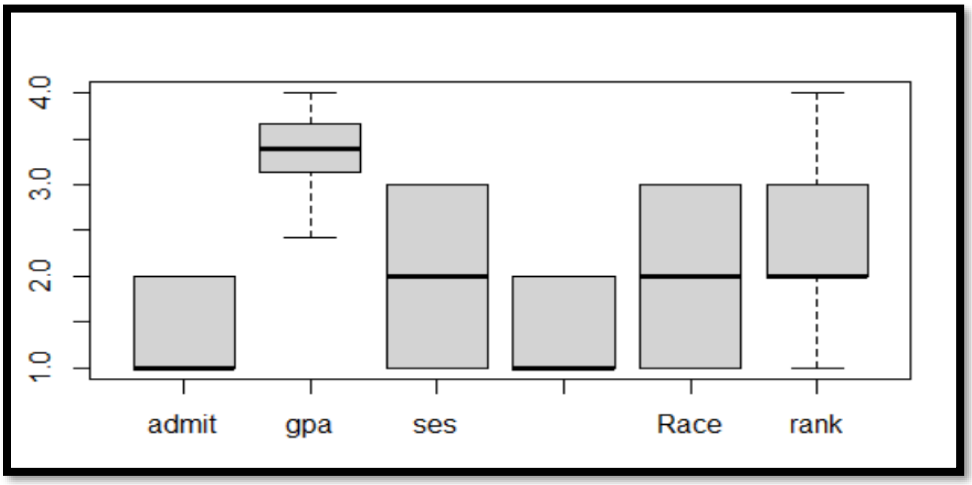
- GPA



Outliers in GPA: 300 300 220   300

Outliers in GRE: 2.26

The consecutive row data was identified and removed



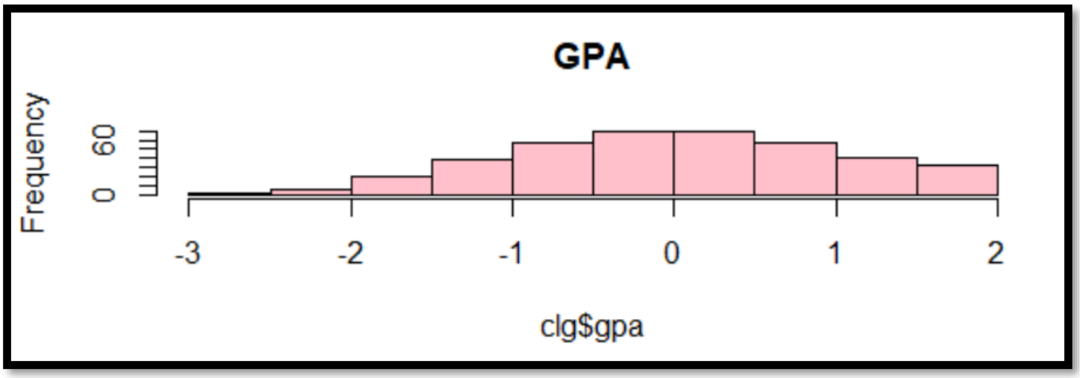
3. Structure transformation of data:

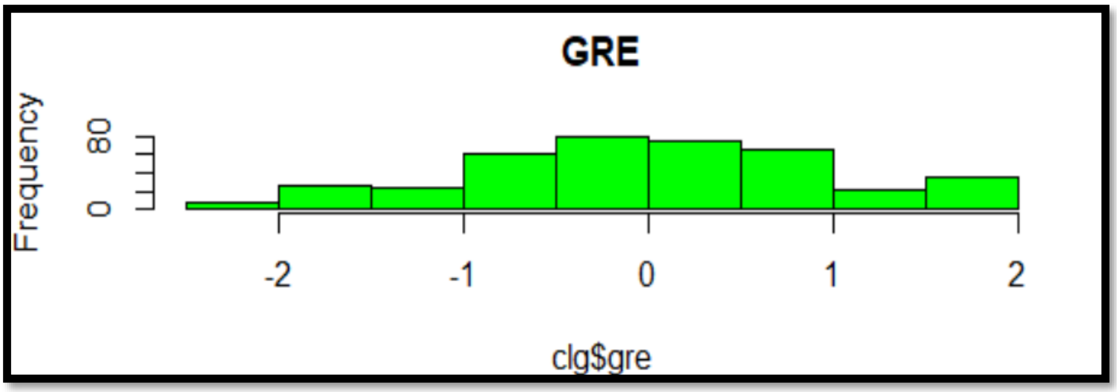
- The structure of the given data set was analysed
- GRE &GPA was set to numeric
- Admit, ses, Gender_male, Race, Rank was set to factor class

```
Console   Terminal ×   Jobs ×
R 4.1.2 · F:/R_programming/
> clg$gre = as.numeric(clg$gre)
> str(clg)
'data.frame':   400 obs. of  7 variables:
 $ admit      : Factor w/ 2 levels "0","1": 1 2 2 2 1 2 2 1 2 1 ...
 $ gre        : num  380 660 800 640 520 760 560 400 540 700 ...
 $ gpa        : num  3.61 3.67 4 3.19 2.93 3 2.98 3.08 3.39 3.92 ...
 $ ses        : Factor w/ 3 levels "1","2","3": 1 2 2 1 3 2 2 2 1 1
 ...
 $ Gender_Male: Factor w/ 2 levels "0","1": 1 1 1 2 2 2 2 1 2 1 ...
 $ Race       : Factor w/ 3 levels "1","2","3": 3 2 2 2 2 1 2 2 1 2
 ...
 $ rank       : Factor w/ 4 levels "1","2","3","4": 3 3 1 4 4 2 1 2 3
 2 ...
```

4. Distribution of dataset

- Since GPA and GRE are numeric
- Histograms were used to analyse the normal distribution of the data
- It was found that the numeric does not follow normal distribution
- Applied scaling to transform the distribution into normality





5. Logistic model

Model 1 – All the independent variables as predictors

Null deviance: 396.77 on 316 degrees of freedom

Residual deviance: 360.51 on 306 degrees of freedom

Model 2 – GPA+rank as predictors

Null deviance: 396.77 on 316 degrees of freedom

Residual deviance: 364.49 on 312 degrees of freedom

Model 3 – GPA+GRE as predictors

Null deviance: 396.77 on 316 degrees of freedom

Residual deviance: 380.41 on 314 degrees of freedom

Accuracy and confusion matrix:

```
> admitPredict = predict(model2,test.data,type ="response")
> test.data$admit1 = ifelse(admitPredict >0.5,1,0)
>
> #confusion matrix
> cf1= table(ActualValue=test.data$admit,
+            PredictedValue= test.data$admit1)
> #accuracy
> accuracy1 = sum(diag(cf1))/sum(cf1)
> accuracy1
[1] 0.7088608
> #accuracy is high with model2 comparatively
> cf1
           PredictedValue
ActualValue  0  1
          0 48  6
          1 17  8
>
```

Model 2 is the best one with high accuracy. GPA and Rank are the most influential factors the student admission

6. SVM model

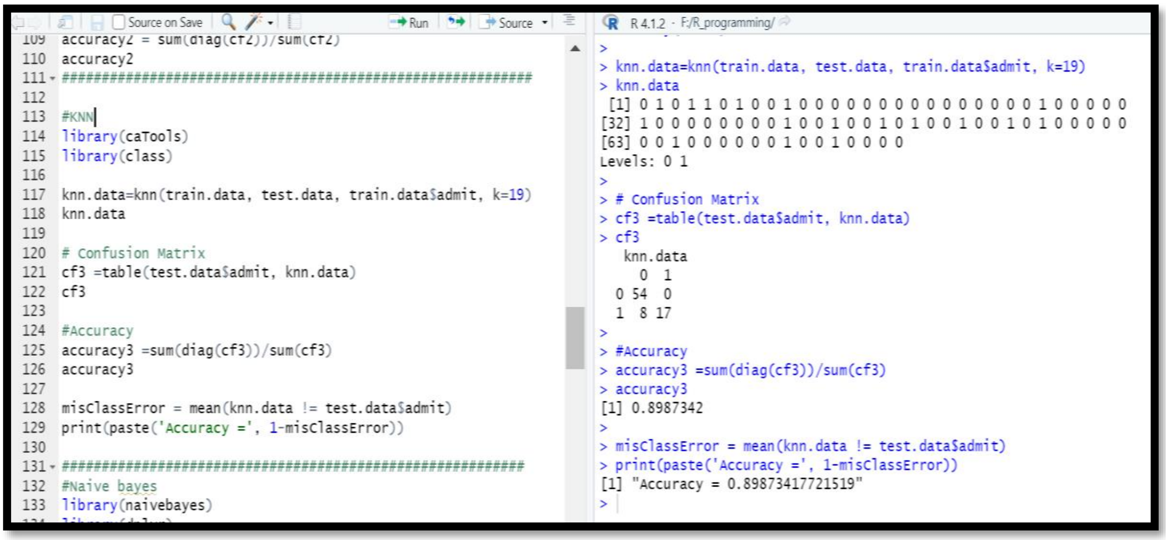
Accuracy and confusion matrix

```
accuracy1 = sum(diag(cf1))/sum(cf1)
accuracy1
#accuracy is high with model2 comparatively

##########################################################
#SVM model
library(e1071)
svm.model = svm(admit~.,
                data = train.data, kernel="linear",scale = T)
summary(svm.model)
head(test.data)
test.data= subset(test.data[-8])
p <-  predict(svm.model,test.data[-1],type="class")
p


#confusion matrix
cf2= table(ActualValue=test.data$admit,
           PredictedValue= p)
cf2

#accuracy
accuracy2 = sum(diag(cf2))/sum(cf2)
accuracy2
##############################################################
```

```
R 4.1.2 · F:/R_programming/
 71  73  76  83  93  94  96  98 102 106 126 129 141 144 149 153
  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
155 158 160 165 168 186 194 213 216 224 229 235 239 240 242 246
  0   1   0   0   0   0   0   0   0   0   0   1   0   0   1   0
258 260 268 269 272 275 278 281 287 293 296 298 299 310 322 327
  0   0   0   0   0   0   1   0   0   0   0   0   0   0   0   0
328 342 347 358 363 365 367 371 372 377 378 386 396 397 398
  0   0   0   0   0   1   0   0   0   0   0   0   0   0   0
Levels: 0 1
>
>
> #confusion matrix
> cf2= table(ActualValue=test.data$admit,
+            PredictedValue= p)
> cf2
           PredictedValue
ActualValue  0  1
          0 52  2
          1 19  6
>
> #accuracy
> accuracy2 = sum(diag(cf2))/sum(cf2)
> accuracy2
[1] 0.7341772
>
```

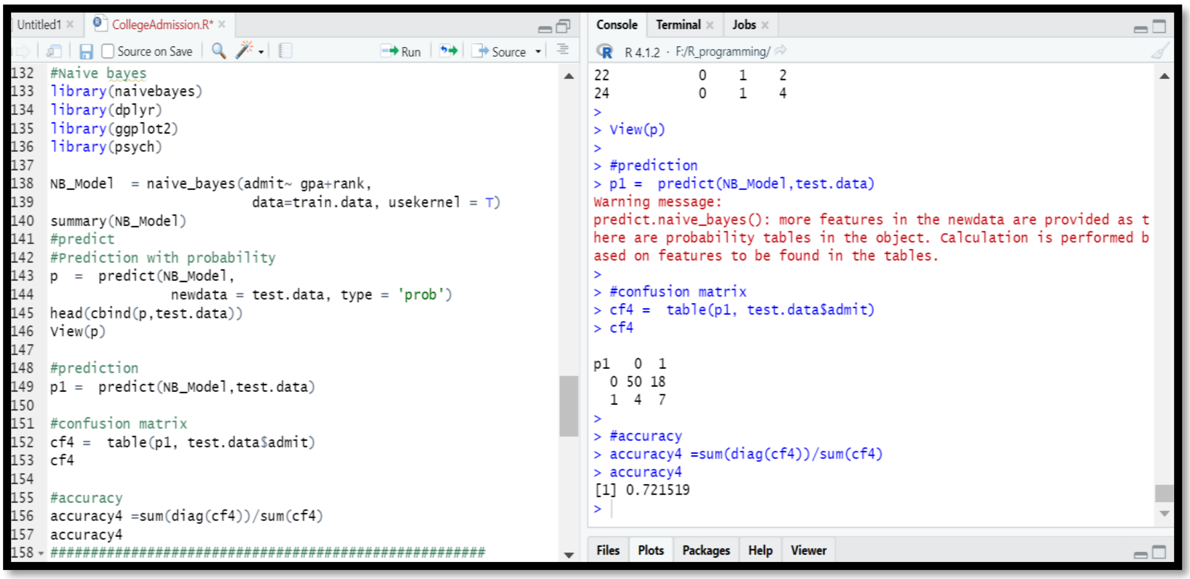
7. KNN model

Accuracy and confusion matrix

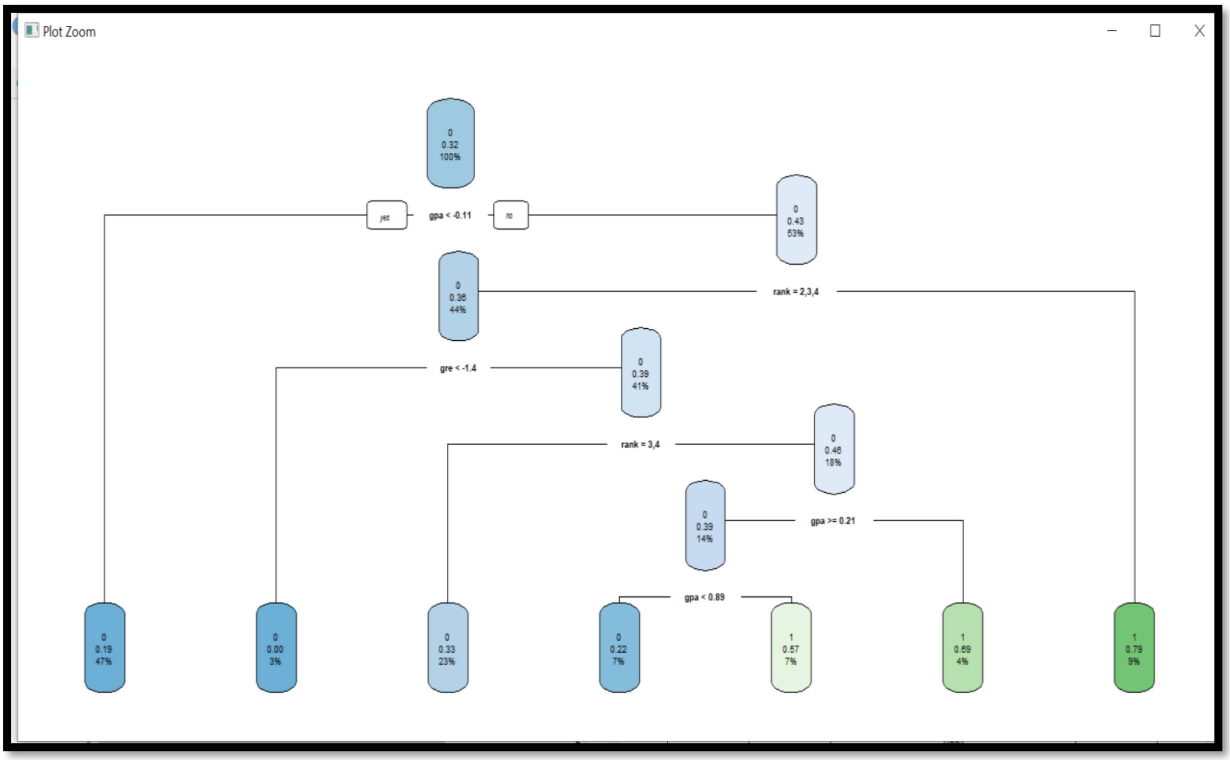


8. Naïve bayes

Accuracy and confusion matrix



9. Decision tree



10. Optimal model

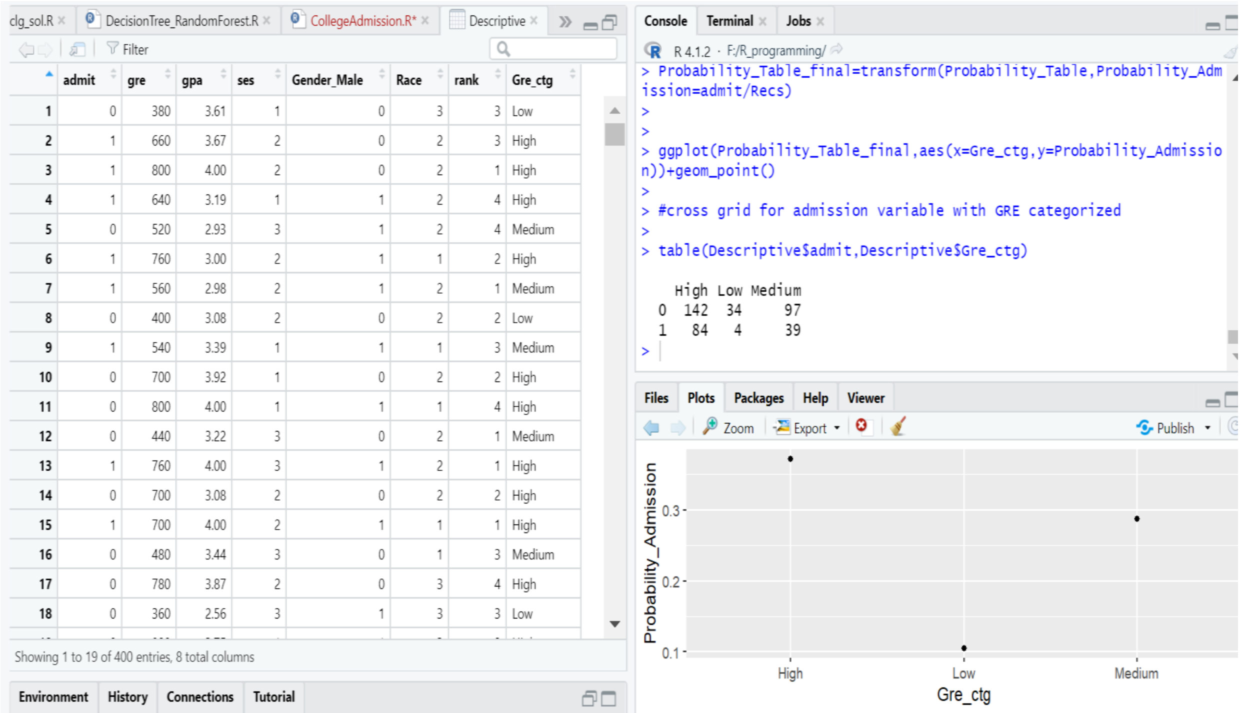
The above analysis shows that KNN is the most accurate model with 89.87%

11. Categorisation of data

Based on the given criteria, the data was categorised into High, Medium and low on the basis of average of grade points

Descriptive:
Categorize the average of grade point into High, Medium, and Low (with admission probability percentages) and plot it on a point chart.
Cross grid for admission variables with GRE Categorization is shown below:

GRE	Categorized
0-440	Low
440-580	Medium
580+	High



CONCLUSION:

This project gave a deeper understanding of concepts and ways to handle the data. By analysing the variables, Grade Point Average and Rank are the most influential rather than other racial factors like gender, race and socio-economic status.