Here I considered the Graduate Admission 2 data set(https://www.kaggle.com/datasets/mohansacharya/graduate-admissions). This data set is created for prediction of Graduate Admissions from an Indian perspective. The data set contains several parameters which are considered important during the application for Masters Programs. The sample size is 400. Take Chance.of.Admit as the quantitative response variable. Among the predictors, Research is a qualitative variable and treat University Rating as a quantitative variable. Take all the data as training data. For all the models ,I used leave-one-out cross-validation (LOOCV) to compute the estimated test MSE.

Initially I fit a regression tree to the data and summarize the results.

```
## Warning: package 'tree' was built under R version 4.2.3
```

```
##
## Regression tree:
## tree(formula = Chance.of.Admit ~ ., data = Admission.data)
## Variables actually used in tree construction:
## [1] "CGPA"      "GRE.Score"
## Number of terminal nodes:  8
## Residual mean deviance:  0.004479 = 1.756 / 392
## Distribution of residuals:
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -0.276300 -0.028820  0.005122  0.000000  0.041180  0.217900
```

The Variables actually used in tree construction are "CGPA" "GRE.Score". There are 8 nodes and residual mean deviance is 0.004479. The distribution of residuals is given below.

```
summary(sumry$residuals)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -0.276250 -0.028824  0.005122  0.000000  0.041176  0.217931
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-0.276250	-0.028824	0.005122	0.000000	0.041176	0.217931

Table 1: The distribution of residuals

Here I display the tree graphically and explicitly describe the regions corresponding to the terminal nodes that provide a partition of the predictor space (i.e., provide expressions for the regions $R_1; ...; R_J$).

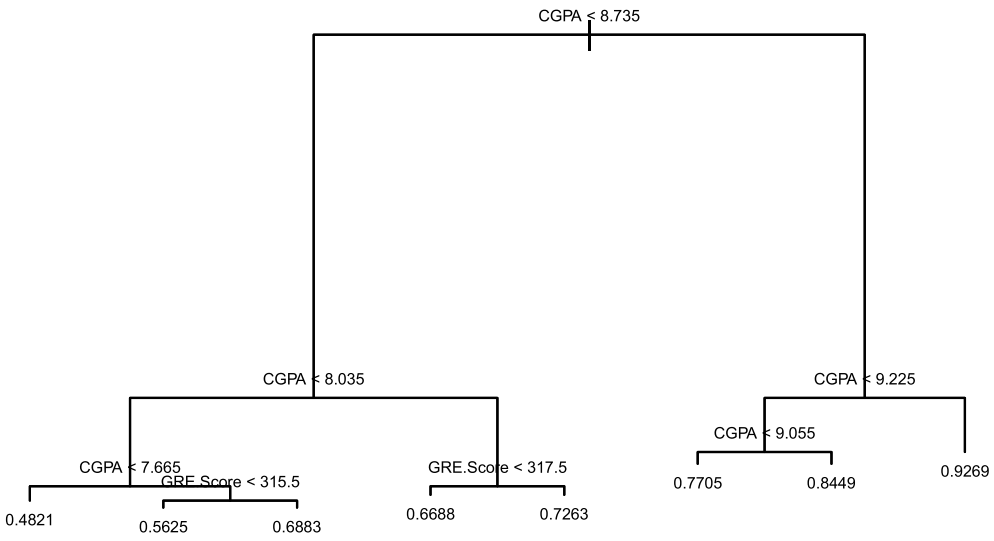

Figure 1: Regression tree for Admission data

Let R_j be the partitions of the predictor space.

$$
\begin{aligned}
R_1 &= \{X \mid CGPA < 7.665\} \\
R_2 &= \{X \mid 7.665 \leq CGPA < 8.035, GRE.Score < 315.5\} \\
R_3 &= \{X \mid 7.665 \leq CGPA < 8.035, 315.5 \leq GRE.Score\} \\
R_4 &= \{X \mid 8.035 \leq CGPA < 8.735, GRE.Score < 317.5\} \\
R_5 &= \{X \mid 8.035 \leq CGPA < 8.735, 317.5 \leq GRE.Score\} \\
R_6 &= \{X \mid 8.735 \leq CGPA < 9.055\} \\
R_7 &= \{X \mid 9.055 \leq CGPA < 9.225\} \\
R_8 &= \{X \mid 9.225 \leq CGPA\}
\end{aligned}
$$

The test MSE using LOOCV

```
## $MSE
## [1] 0.005776329
```

The test MSE using LOOCV is 0.005776329.

Used LOOCV to determine whether pruning is helpful and determined the optimal size for the pruned tree.

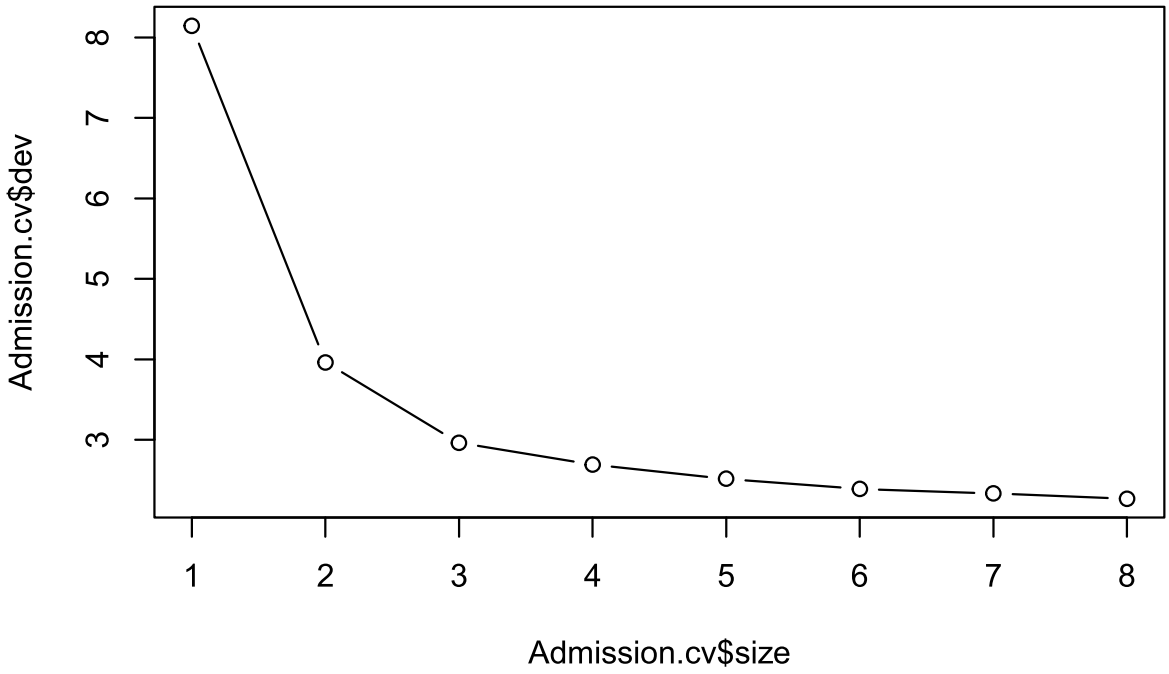

Figure 2: Plot the estimated test error rate

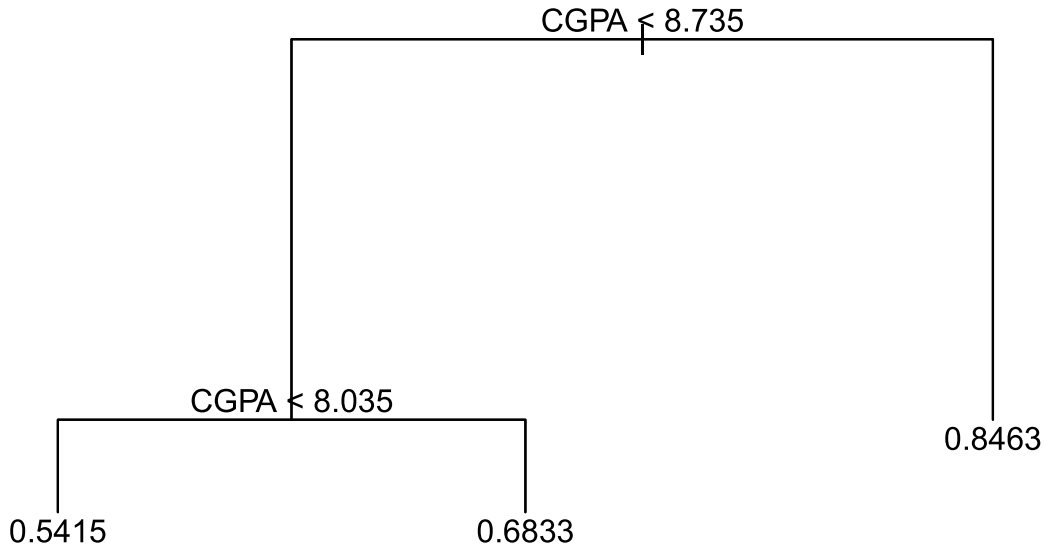

Figure 3: Regression prune Tree for cancer data

```
## $MSE
## [1] 0.007170241
```

The pruned tree has three(3) terminal nodes(Figure 2) and the actual used variable in tree construction is "CGPA"(See Figure 3) and it is seems to be most important predictor. Using LOOCV method the test MSE for pruned tree with three terminal nodes is 0.007170241. Test MSE is greater than the un-pruned tree in part a.

Used a bagging approach to analyze the data with $B = 1000$.

```
##                    %IncMSE IncNodePurity
## GRE.Score          36.26965    0.65288532
## TOEFL.Score        12.36551    0.25280191
## University.Rating  22.38161    0.10401186
## SOP                27.04264    0.22348002
## LOR                15.68851    0.19678760
## CGPA              101.23843    6.42387033
## Research           25.29731    0.08932112
```

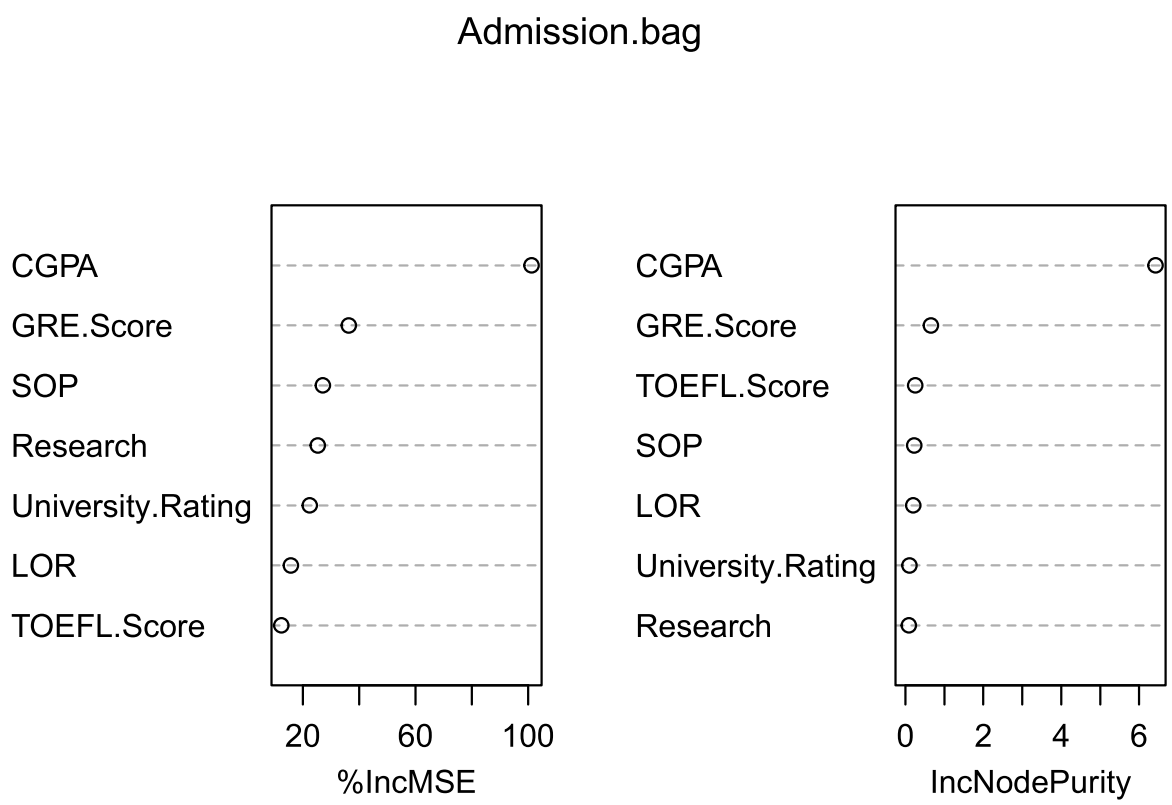

Figure 4: Variable importance measure for each predictor (Bagging)

Using bagging approach with $B = 1000$, the Node purity plot (Figure 4) shows that the variables "CGPA"(IncNodePurity=6.42387033)is the most important predictors.

And the test MSE using LOOCV method is 0.004854975.

Used a random forest approach to analyze the data with $B = 1000$ and $m \approx p/3$.

```
##                    %IncMSE IncNodePurity
## GRE.Score          35.53476     1.8247557
## TOEFL.Score        23.45784     1.2627941
## University.Rating  24.83874     0.7663180
## SOP                20.29365     0.5650116
## LOR                22.83124     0.4759518
```

```
## CGPA              47.87456      2.6027796
## Research          24.47288      0.2501665
```

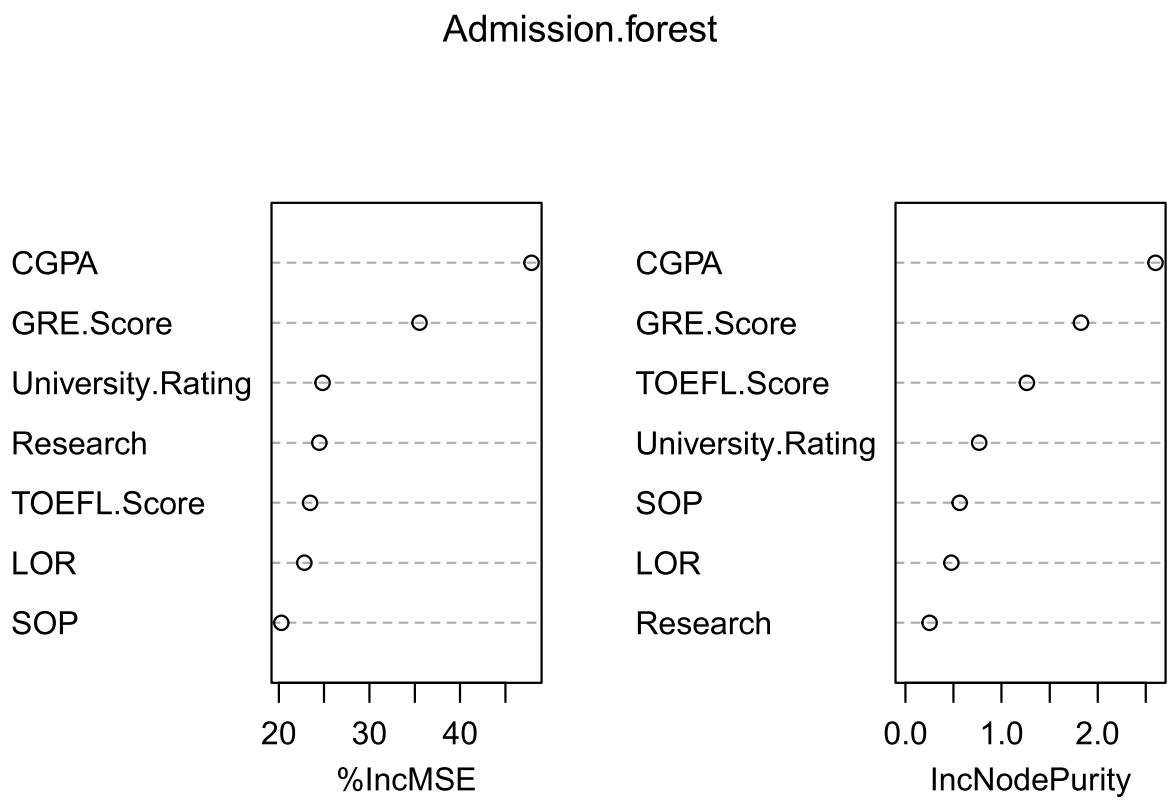

Figure 5: Variable importance measure for each predictor (Random forest)

```
## $MSE
## [1] 0.004408569
```

Using random forest approach with $B = 1000$ the Node purity plot (Figure 5) shows that the variables "CGPA"(IncNodePurity=2.6027796) and "GRE.Score" (IncNodePurity=1.8247557) are most important predictors.

And the test MSE using LOOCV method is 0.004408569.

Used a boosting approach to analyze the data with $B = 1000$, $d = 1$, and $\lambda = 0.01$.

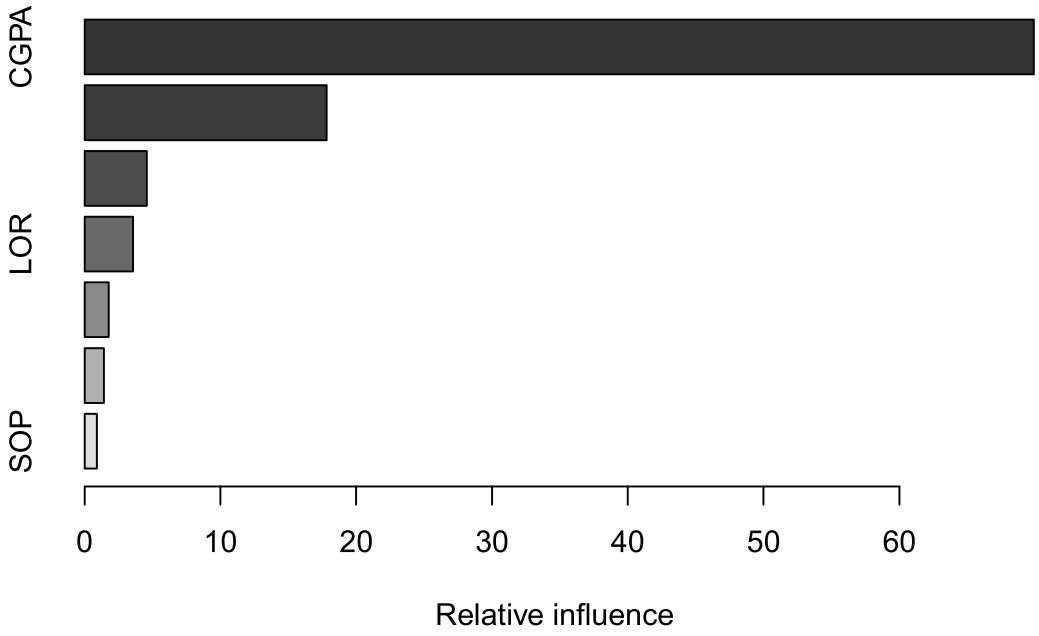

Figure 6: Relative influence Plot

```
##                                   var    rel.inf
## CGPA                             CGPA 69.8978024
## GRE.Score                   GRE.Score 17.8200041
## TOEFL.Score                 TOEFL.Score  4.5783938
## LOR                                 LOR  3.5762225
## University.Rating University.Rating  1.7849804
## Research                     Research  1.4329529
## SOP                               SOP  0.9096439
```

```
## $MSE
## [1] 0.004488337
```

Using bossting approach with $B = 1000$, $d = 1$ and $\lambda = 0.01$, according to the Relative influence plot (Figure 6) it shows that the variables "CGPA" (rel.inf=69.8978024) and "GRE.Score" (rel.inf=17.8200041) are most important predictors. And the test MSE using LOOCV method is 0.004488337.

Comparison of the results from the various methods.

	un-pruned tree	pruned tree	bagging	random-forest	boosting
Test MSE	0.005776329	0.007170241	0.004854975	0.004408569	0.004488337

Table 2: Test MSE for different approches

When consider the four different approaches discussed above, pruned tree approach gives large test MSE(0.007170241) and random-forest approach gives the small test MSE(0.004408569). So random-forest approach should be recommended to analyse Admission data.