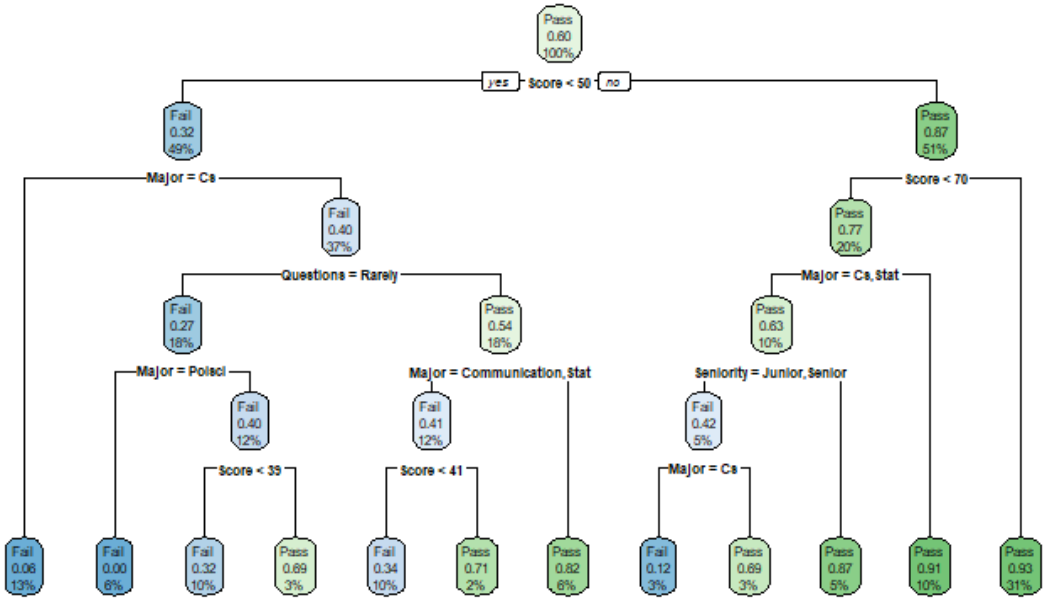
# Prediction Challenge 2 - with rpart

By Shuohao Ping

# Default setting

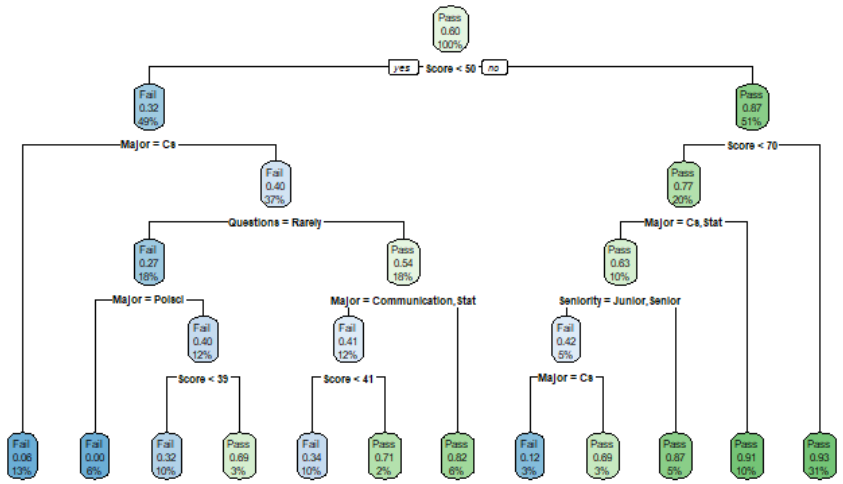
- Decision tree




- Cross validation:
- I use Cross_validate with parameter 100 and 0.8. (Randomly choose 80% of data for training and use remaining 20% data for validation. And repeat this 100 times)
- The average accuracy is 0.8449709

# Control minsplit (minsplit=0)

- Decision tree:




- Cross validation:

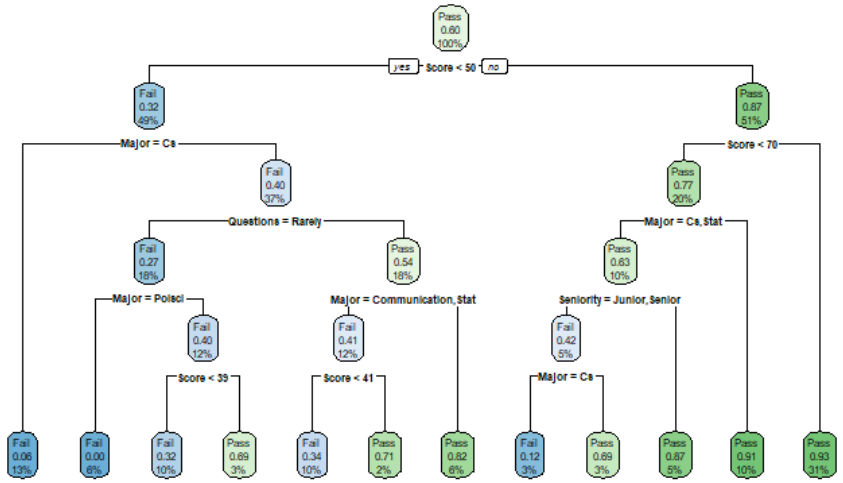
```
[[2]]$average_accuracy_subset
[1] 0.8446646

[[2]]$average_accuracy_all
[1] 0.8446646
```

- When we control minsplit=0, the accuracy is the same as the decision tree in default setting.

# Control minbucket(minbucket=0)

- Decision tree:




- Cross validation:

```
[[2]]$average_accuracy_subset
[1] 0.8446751

[[2]]$average_accuracy_all
[1] 0.8446751
```

- When we control minbucket=0, the accuracy is the same as the decision tree in default setting.

# Control CP

- When CP=0.001
- Cross validation:

```
[[2]]
[[2]]$average_accuracy_subset
[1] 0.8589223

[[2]]$average_accuracy_all
[1] 0.8434654
```

- Interpretation: When CP becomes small, we have better accuracy.
- However, when CP is too small, the accuracy begin to decrease
- When CP=0.0001
- Cross validation:

```
[[2]]$average_accuracy_subset
[1] 0.85243

[[2]]$average_accuracy_all
[1] 0.8442102
```

# Control CP

- When CP=0.0008, we have the highest accuracy.

```
[[2]]$average_accuracy_subset
[1] 0.859514

[[2]]$average_accuracy_all
[1] 0.8441838
```

- Thus, we use command rpart(Grade~Attendance+Score+Seniority+Texting+Questions+Major,data=train,control = rpart.control(cp=0.0008)) to generate decision tree.