R_DECISION_TREES_RANDOM_FOREST

Ulises Jose Bustamante Mora

2023-10-26

Importing libraries and dataset

```
library(ggplot2)
library(ISLR)
head(College)

##                              Private Apps Accept Enroll Top10perc
Top25perc
## Abilene Christian University     Yes 1660   1232    721        23
52
## Adelphi University               Yes 2186   1924    512        16
29
## Adrian College                   Yes 1428   1097    336        22
50
## Agnes Scott College              Yes  417    349    137        60
89
## Alaska Pacific University        Yes  193    146     55        16
44
## Albertson College                Yes  587    479    158        38
62
##                              F.Undergrad P.Undergrad Outstate
Room.Board Books
## Abilene Christian University        2885         537     7440
3300   450
## Adelphi University                  2683        1227    12280
6450   750
## Adrian College                      1036          99    11250
3750   400
## Agnes Scott College                  510          63    12960
5450   450
## Alaska Pacific University            249         869     7560
4120   800
## Albertson College                    678          41    13500
3335   500
##                              Personal PhD Terminal S.F.Ratio
perc.alumni Expend
## Abilene Christian University     2200  70       78      18.1
12   7041
## Adelphi University               1500  29       30      12.2
16  10527
## Adrian College                   1165  53       66      12.9
```

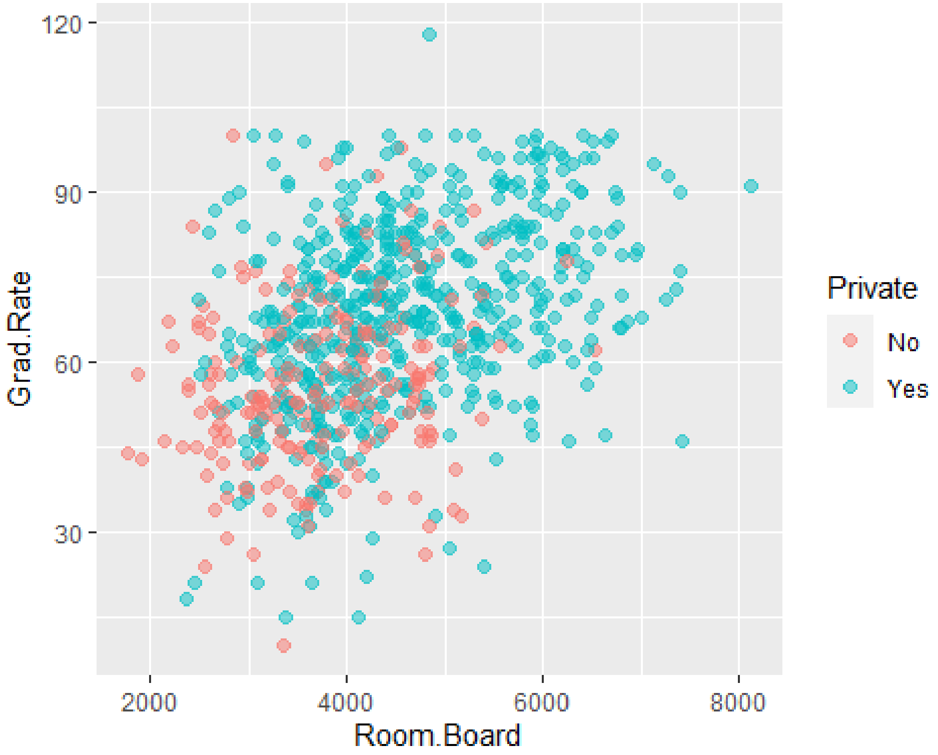
```
30   8735
## Agnes Scott College               875  92        97        7.7
37  19016
## Alaska Pacific University        1500  76        72       11.9
2  10922
## Albertson College                 675  67        73        9.4
11   9727
##                               Grad.Rate
## Abilene Christian University        60
## Adelphi University                  56
## Adrian College                      54
## Agnes Scott College                 59
## Alaska Pacific University           15
## Albertson College                   55

df = College
```

Exploratory Data Analysis

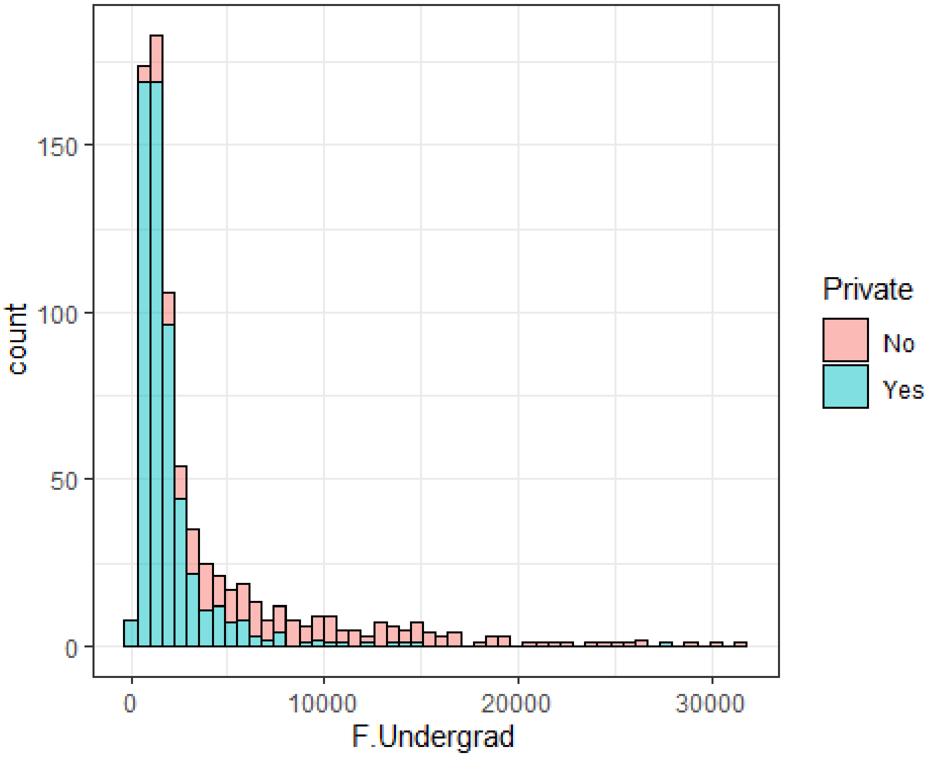
Correlation plot

```
ggplot(df, aes(Room.Board, Grad.Rate)) + geom_point(aes(color = Private),
size = 2, alpha = 0.5)
```



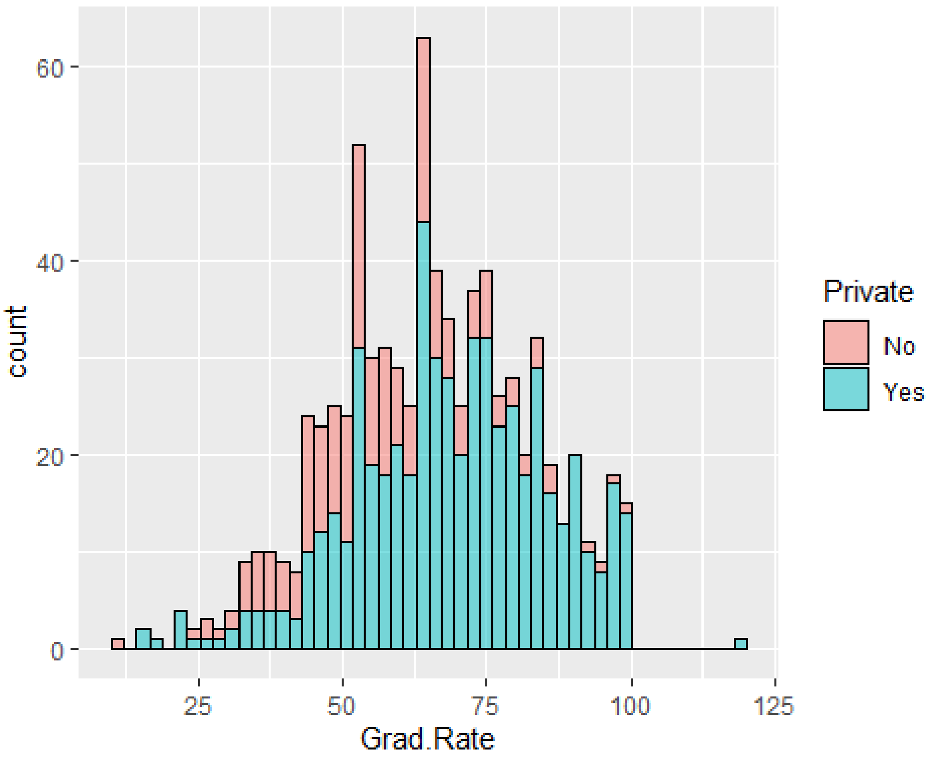
Histogram plot

```
ggplot(df, aes(F.Undergrad)) + geom_histogram(aes(fill = Private),
                                              color = "black", bins = 50,
                                              alpha = 0.5) + theme_bw()
```



We noticed an outlier, so we are going to change that value.

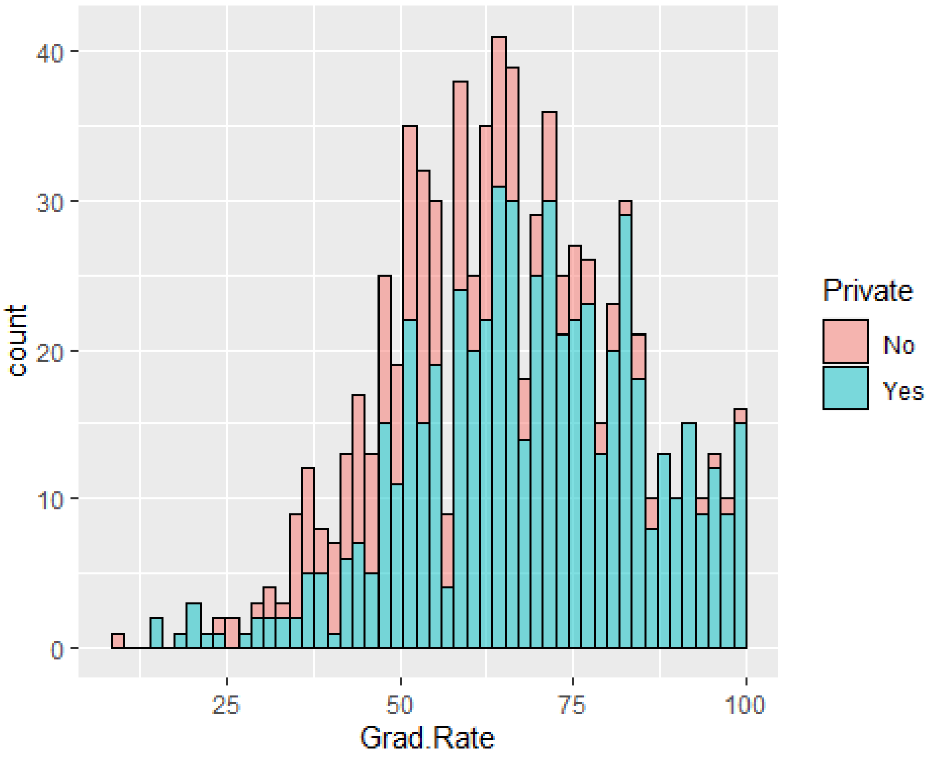
```
ggplot(df, aes(Grad.Rate)) + geom_histogram(aes(fill = Private),
                                            color = "black",
                                            bins = 50,
                                            alpha = 0.5)
```



Confirming that the outlier is changed

```
df["Cazenovia College", "Grad.Rate"] = 100

ggplot(df, aes(Grad.Rate)) + geom_histogram(aes(fill = Private),
                                            color = "black",
                                            bins = 50,
                                            alpha = 0.5)
```



Creating the Desicion Tree model

Splitting the dataset to start with the ML model.

```
library(caTools)
sample = sample.split(df$Private, SplitRatio= 0.7 )
dfTrain = subset(df,sample == T)
dfTest = subset(df,sample == F)
```

Training the Decision Tree model

```
library(rpart)
tree = rpart(Private~., method = "class", data = dfTrain)
summary(tree)

## Call:
## rpart(formula = Private ~ ., data = dfTrain, method = "class")
##   n= 544
##
```

```
##           CP nsplit rel error    xerror       xstd
## 1 0.45945946      0 1.0000000 1.0000000 0.07013217
## 2 0.25000000      1 0.5405405 0.6554054 0.06032238
## 3 0.03378378      2 0.2905405 0.4189189 0.05007929
## 4 0.01801802      3 0.2567568 0.4459459 0.05145466
## 5 0.01000000      7 0.1756757 0.4459459 0.05145466
##
## Variable importance
## F.Undergrad      Enroll      Accept        Apps    Outstate 
P.Undergrad 
##          17          15          13          11          10 
8 
##   S.F.Ratio  Room.Board      Expend   Grad.Rate   Top10perc 
Terminal 
##           7           6           5           4           3 
1 
##         PhD 
##           1 
##
## Node number 1: 544 observations,    complexity param=0.4594595
##   predicted class=Yes  expected loss=0.2720588  P(node) =1
##     class counts:   148   396
##    probabilities: 0.272 0.728 
##   left son=2 (176 obs) right son=3 (368 obs)
##   Primary splits:
##       F.Undergrad < 2995    to the right, improve=92.28086, (0 
missing)
##       Outstate    < 7960    to the left,  improve=90.46910, (0 
missing)
##       Enroll      < 754.5   to the right, improve=73.60392, (0 
missing)
##       P.Undergrad < 896     to the right, improve=63.18460, (0 
missing)
##       S.F.Ratio   < 14.55   to the right, improve=56.98459, (0 
missing)
##   Surrogate splits:
##       Enroll      < 754.5   to the right, agree=0.960, adj=0.875, (0 
split)
##       Accept      < 2040.5  to the right, agree=0.926, adj=0.773, (0 
split)
##       Apps        < 2995    to the right, agree=0.892, adj=0.665, (0 
split)
##       P.Undergrad < 870     to the right, agree=0.831, adj=0.477, (0 
split)
##       S.F.Ratio   < 17.85   to the right, agree=0.752, adj=0.233, (0 
split)
##
## Node number 2: 176 observations,    complexity param=0.25
##   predicted class=No   expected loss=0.3068182  P(node) =0.3235294
##     class counts:   122    54
##     probabilities: 0.693 0.307
##    left son=4 (117 obs) right son=5 (59 obs)
##    Primary splits:
##        Outstate   < 10218.5 to the left,  improve=45.58072, (0 missing)
##        Room.Board < 5460.5  to the left,  improve=42.82834, (0 missing)
##        Expend     < 16636   to the left,  improve=22.91525, (0 missing)
##        S.F.Ratio  < 14.85   to the right, improve=21.41266, (0 missing)
##        Grad.Rate  < 79.5    to the left,  improve=18.86120, (0 missing)
##    Surrogate splits:
##        Room.Board < 5460.5  to the left,  agree=0.881, adj=0.644, (0
split)
##        Expend     < 9436    to the left,  agree=0.858, adj=0.576, (0
split)
##        S.F.Ratio  < 14.55   to the right, agree=0.824, adj=0.475, (0
split)
##        Grad.Rate  < 72.5    to the left,  agree=0.824, adj=0.475, (0
split)
##        Top10perc  < 52.5    to the left,  agree=0.784, adj=0.356, (0
split)
##
## Node number 3: 368 observations,    complexity param=0.01801802
##    predicted class=Yes  expected loss=0.07065217  P(node) =0.6764706
##      class counts:    26   342
##     probabilities: 0.071 0.929
##    left son=6 (67 obs) right son=7 (301 obs)
##    Primary splits:
##        Outstate   < 7960    to the left,  improve=13.546690, (0
missing)
##        Room.Board < 2790    to the left,  improve= 7.299453, (0
missing)
##        Expend     < 6717.5  to the left,  improve= 5.366087, (0
missing)
##        S.F.Ratio  < 14.35   to the right, improve= 5.093944, (0
missing)
##        Grad.Rate  < 52.5    to the left,  improve= 4.687918, (0
missing)
##    Surrogate splits:
##        Room.Board  < 3072.5  to the left,  agree=0.867, adj=0.269, (0
split)
##        Expend      < 6206.5  to the left,  agree=0.853, adj=0.194, (0
split)
##        perc.alumni < 8.5     to the left,  agree=0.845, adj=0.149, (0
split)
##        Grad.Rate   < 45.5    to the left,  agree=0.840, adj=0.119, (0
split)
##        Top25perc   < 13.5    to the left,  agree=0.826, adj=0.045, (0
split)
##
## Node number 4: 117 observations
##    predicted class=No   expected loss=0.05128205  P(node) =0.2150735
##     class counts:   111     6
##    probabilities: 0.949 0.051 
## 
## Node number 5: 59 observations,    complexity param=0.03378378
##   predicted class=Yes  expected loss=0.1864407  P(node) =0.1084559
##     class counts:    11    48
##    probabilities: 0.186 0.814 
##   left son=10 (13 obs) right son=11 (46 obs)
##   Primary splits:
##       Accept      < 5934    to the right, improve=8.533757, (0 
missing)
##       F.Undergrad < 9781.5  to the right, improve=8.533757, (0 
missing)
##       Enroll      < 2477.5  to the right, improve=7.909669, (0 
missing)
##       Apps        < 14155.5 to the right, improve=7.145558, (0 
missing)
##       Room.Board  < 5406    to the left,  improve=5.949587, (0 
missing)
##   Surrogate splits:
##       F.Undergrad < 9781.5  to the right, agree=0.966, adj=0.846, (0 
split)
##       Enroll      < 1985    to the right, agree=0.949, adj=0.769, (0 
split)
##       P.Undergrad < 1653    to the right, agree=0.898, adj=0.538, (0 
split)
##       Apps        < 14155.5 to the right, agree=0.864, adj=0.385, (0 
split)
##       Outstate    < 10674   to the left,  agree=0.814, adj=0.154, (0 
split)
## 
## Node number 6: 67 observations,    complexity param=0.01801802
##   predicted class=Yes  expected loss=0.358209  P(node) =0.1231618
##     class counts:    24    43
##    probabilities: 0.358 0.642 
##   left son=12 (41 obs) right son=13 (26 obs)
##   Primary splits:
##       F.Undergrad < 955     to the right, improve=5.010473, (0 
missing)
##       Terminal    < 90      to the right, improve=4.852580, (0 
missing)
##       Top10perc   < 7.5     to the left,  improve=4.588426, (0 
missing)
##       Outstate    < 4905    to the left,  improve=3.600842, (0 
missing)
##       P.Undergrad < 186.5   to the right, improve=3.570256, (0 
missing)
##   Surrogate splits:
##       Enroll   < 220     to the right, agree=0.866, adj=0.654, (0 
split)
```

```
##       Accept   < 402.5   to the right, agree=0.776, adj=0.423, (0
split)
##       Apps     < 451     to the right, agree=0.746, adj=0.346, (0
split)
##       PhD      < 47      to the right, agree=0.731, adj=0.308, (0
split)
##       Terminal < 61.5    to the right, agree=0.716, adj=0.269, (0
split)
##
## Node number 7: 301 observations
##   predicted class=Yes  expected loss=0.006644518  P(node) =0.5533088
##     class counts:     2   299
##    probabilities: 0.007 0.993
##
## Node number 10: 13 observations
##   predicted class=No   expected loss=0.3076923  P(node) =0.02389706
##     class counts:     9     4
##    probabilities: 0.692 0.308
##
## Node number 11: 46 observations
##   predicted class=Yes  expected loss=0.04347826  P(node) =0.08455882
##     class counts:     2    44
##    probabilities: 0.043 0.957
##
## Node number 12: 41 observations,    complexity param=0.01801802
##   predicted class=No   expected loss=0.4878049  P(node) =0.07536765
##     class counts:    21    20
##    probabilities: 0.512 0.488
##   left son=24 (8 obs) right son=25 (33 obs)
##   Primary splits:
##       Terminal    < 85.5    to the right, improve=4.730229, (0
missing)
##       Top10perc   < 17      to the left,  improve=4.416376, (0
missing)
##       Top25perc   < 53.5    to the left,  improve=3.773519, (0
missing)
##       P.Undergrad < 114.5   to the right, improve=3.710027, (0
missing)
##       Grad.Rate   < 55      to the left,  improve=3.573929, (0
missing)
##   Surrogate splits:
##       PhD         < 78.5    to the right, agree=0.927, adj=0.625, (0
split)
##       Top10perc   < 5.5     to the left,  agree=0.854, adj=0.250, (0
split)
##       Room.Board  < 4665    to the right, agree=0.854, adj=0.250, (0
split)
##       P.Undergrad < 1426.5  to the right, agree=0.829, adj=0.125, (0
split)
##
```

```
## Node number 13: 26 observations
##   predicted class=Yes  expected loss=0.1153846  P(node) =0.04779412
##     class counts:     3    23
##    probabilities: 0.115 0.885
##
## Node number 24: 8 observations
##   predicted class=No   expected loss=0  P(node) =0.01470588
##     class counts:     8     0
##    probabilities: 1.000 0.000
##
## Node number 25: 33 observations,    complexity param=0.01801802
##   predicted class=Yes  expected loss=0.3939394  P(node) =0.06066176
##     class counts:    13    20
##    probabilities: 0.394 0.606
##   left son=50 (14 obs) right son=51 (19 obs)
##   Primary splits:
##       Top10perc   < 17      to the left,  improve=3.013215, (0
missing)
##       Expend      < 6623.5  to the left,  improve=2.647282, (0
missing)
##       Grad.Rate   < 59.5    to the left,  improve=2.647282, (0
missing)
##       Top25perc   < 53.5    to the left,  improve=2.472960, (0
missing)
##       P.Undergrad < 114.5   to the right, improve=1.979798, (0
missing)
##   Surrogate splits:
##       Top25perc  < 42.5    to the left,  agree=0.909, adj=0.786, (0
split)
##       Expend     < 5661.5  to the left,  agree=0.758, adj=0.429, (0
split)
##       Grad.Rate  < 58.5    to the left,  agree=0.758, adj=0.429, (0
split)
##       Room.Board < 2805    to the left,  agree=0.697, adj=0.286, (0
split)
##       Enroll     < 723     to the right, agree=0.667, adj=0.214, (0
split)
##
## Node number 50: 14 observations
##   predicted class=No   expected loss=0.3571429  P(node) =0.02573529
##     class counts:     9     5
##    probabilities: 0.643 0.357
##
## Node number 51: 19 observations
##   predicted class=Yes  expected loss=0.2105263  P(node) =0.03492647
##     class counts:     4    15
##    probabilities: 0.211 0.789
```

Making Predictions

```
treePreds = predict(tree, dfTest)
head(treePreds)

##                              No       Yes
## Agnes Scott College 0.006644518 0.9933555
## Albertson College   0.006644518 0.9933555
## Amherst College     0.006644518 0.9933555
## Aquinas College     0.006644518 0.9933555
## Augustana College   0.006644518 0.9933555
## Baker University    0.006644518 0.9933555
```

Creating the confusion matrix

First we need to make a dummy variable

```
treePreds = as.data.frame(treePreds)

joiner = function(x){
  if (x >= 0.5){
    return("Yes")
  } else{
    return("No")
  }
}

treePreds$Private = sapply(treePreds$Yes, joiner)
head(treePreds)

##                              No       Yes Private
## Agnes Scott College 0.006644518 0.9933555     Yes
## Albertson College   0.006644518 0.9933555     Yes
## Amherst College     0.006644518 0.9933555     Yes
## Aquinas College     0.006644518 0.9933555     Yes
## Augustana College   0.006644518 0.9933555     Yes
## Baker University    0.006644518 0.9933555     Yes
```

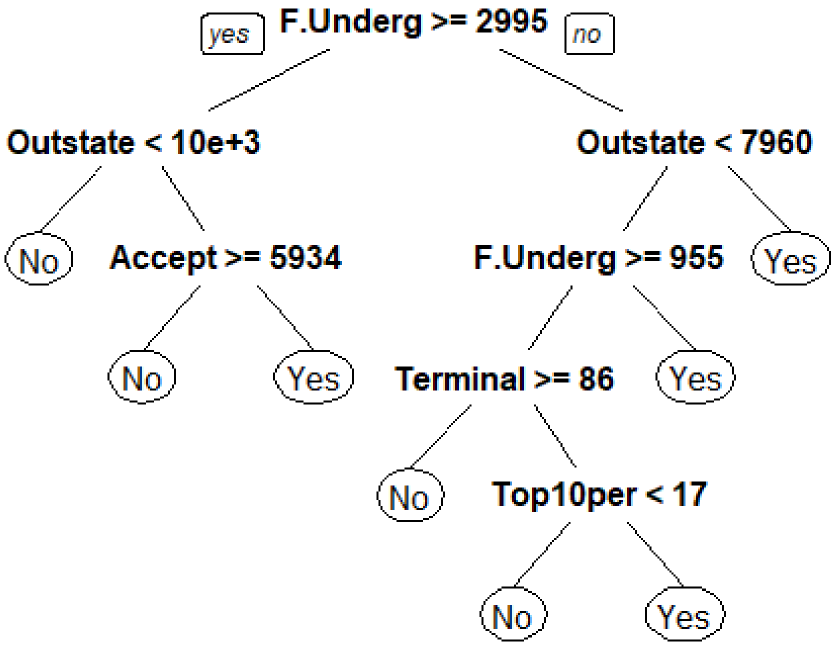
Confusion Matrix

```
table(treePreds$Private, dfTest$Private)

##
##        No Yes
##   No   60   9
##   Yes   4 160
```

Plotting the model

```
#install.packages("rpart.plot")
library(rpart.plot)
prp(tree)
```



Creating a Random Forest model

```
#install.packages("randomForest")
library(randomForest)

## randomForest 4.7-1.1

## Type rfNews() to see new features/changes/bug fixes.

##
## Attaching package: 'randomForest'

## The following object is masked from 'package:ggplot2':
##
##     margin

rfModel = randomForest(Private ~., data = dfTrain, importance = T )

rfModel$confusion

##      No Yes class.error
## No  124  24  0.16216216
## Yes  14 382  0.03535354
```

More info about the configuration of the Random Forest configuration

```
rfModel$importance

##                          No         Yes MeanDecreaseAccuracy
MeanDecreaseGini
## Apps          0.019644949 0.009420944         0.0121952782
8.504058
## Accept        0.022989110 0.012614294         0.0154104886
11.183018
## Enroll        0.043154967 0.030628605         0.0338984148
22.698295
## Top10perc     0.009255763 0.003593597         0.0051316264
4.717788
## Top25perc     0.006804887 0.003006713         0.0040544520
3.771558
## F.Undergrad   0.146791497 0.070376957         0.0909161310
41.555166
## P.Undergrad   0.039214644 0.008148831         0.0164449899
17.119321
## Outstate      0.151277928 0.061442054         0.0854828819
44.750812
## Room.Board    0.020907343 0.018122874         0.0188477754
11.817341
## Books        -0.002248360 0.000480029        -0.0002285399
2.286619
## Personal      0.002545971 0.002065830         0.0022172322
3.824844
## PhD           0.012257379 0.005647215         0.0074292547
4.351807
## Terminal      0.005512467 0.007153085         0.0066905877
3.937940
## S.F.Ratio     0.031537815 0.006891188         0.0134351017
12.338404
## perc.alumni   0.030090264 0.004197012         0.0112169506
6.141727
## Expend        0.017987767 0.014976355         0.0157928524
9.435223
## Grad.Rate     0.011403514 0.007071716         0.0082546110
6.541857
```

Predictions and confusion Matrix

```
rfPreds = predict(rfModel, dfTest)
table(rfPreds, dfTest$Private)

## 
## rfPreds  No Yes
##     No   59   4
##     Yes   5 165
```