Ian Mulchrone

Fundamentals of Data Science

Homework 3

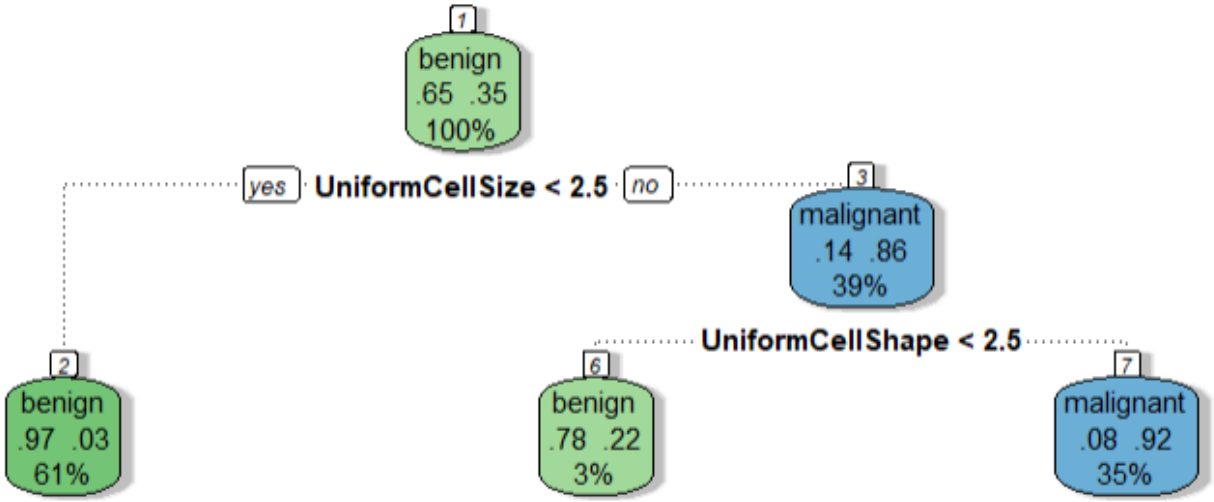
Problem 1

a.

```
CART

683 samples
  9 predictor
  2 classes: 'benign', 'malignant'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 615, 614, 614, 614, 615, 615, ...
Resampling results across tuning parameters:

  cp          Accuracy   Kappa
  0.02510460  0.9488453  0.8886664
  0.05439331  0.9386151  0.8675892
  0.79079498  0.8259120  0.5494330

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was cp = 0.0251046.
```

Accuracy is 94.88%.

b.



c. If UniformCellSize < 2.5 then Class = “benign”

If UniformCellSize >= 2.5 and UniformCellShape < 2.5 then Class = “benign”

If UniformCellSize >= 2.5 and UniformCellShape >= 2.5 then Class = “malignant”

Problem 2

a.

```
CART

2170 samples
  11 predictor
   5 classes: '1', '2', '3', '4', '5'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 1953, 1953, 1952, 1952, 1952, 1954, ...
Resampling results:

  Accuracy   Kappa
  0.8359511  0.7550537
```

Accuracy is 83.6%

b.

Train

```
Confusion Matrix and Statistics

          Reference
Prediction   1   2   3   4   5
         1 867   0   0   0   0
         2   0 348   0   0   0
         3   0   0   0 233   0
         4   0   0   0 238   0
         5   0   0   0  52   0

Overall Statistics

               Accuracy : 0.836
                 95% CI : (0.8178, 0.8531)
    No Information Rate : 0.4988
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7552

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4 Class: 5
Sensitivity            1.0000   1.0000       NA   0.4551       NA
Specificity            1.0000   1.0000   0.8659   1.0000  0.97008
Pos Pred Value         1.0000   1.0000       NA   1.0000       NA
Neg Pred Value         1.0000   1.0000       NA   0.8100       NA
Prevalence             0.4988   0.2002   0.0000   0.3009  0.00000
Detection Rate         0.4988   0.2002   0.0000   0.1369  0.00000
Detection Prevalence   0.4988   0.2002   0.1341   0.1369  0.02992
Balanced Accuracy      1.0000   1.0000       NA   0.7275       NA
```

Test

```
Confusion Matrix and Statistics

          Reference
Prediction   1   2   3   4   5
         1 216   0   0   0   0
         2   0  86   0   0   0
         3   0   0   0  58   0
         4   0   0   0  59   0
         5   0   0   0  13   0

Overall Statistics

               Accuracy : 0.8356
                 95% CI : (0.7973, 0.8694)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7544

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4 Class: 5
Sensitivity               1.0   1.0000       NA   0.4538       NA
Specificity               1.0   1.0000   0.8657   1.0000  0.96991
Pos Pred Value            1.0   1.0000       NA   1.0000       NA
Neg Pred Value            1.0   1.0000       NA   0.8097       NA
Prevalence                0.5   0.1991   0.0000   0.3009  0.00000
Detection Rate            0.5   0.1991   0.0000   0.1366  0.00000
Detection Prevalence      0.5   0.1991   0.1343   0.1366  0.03009
Balanced Accuracy         1.0   1.0000       NA   0.7269       NA
```

The train and test set confusion matrices show that the model performs very similarly on both data sets. They both have problems classifying category 3 and category 5, with neither of them being predicted in both the training and test sets. Because they are so similar in accuracy and categorization, this suggests that there is no concern for overfitting in the model.

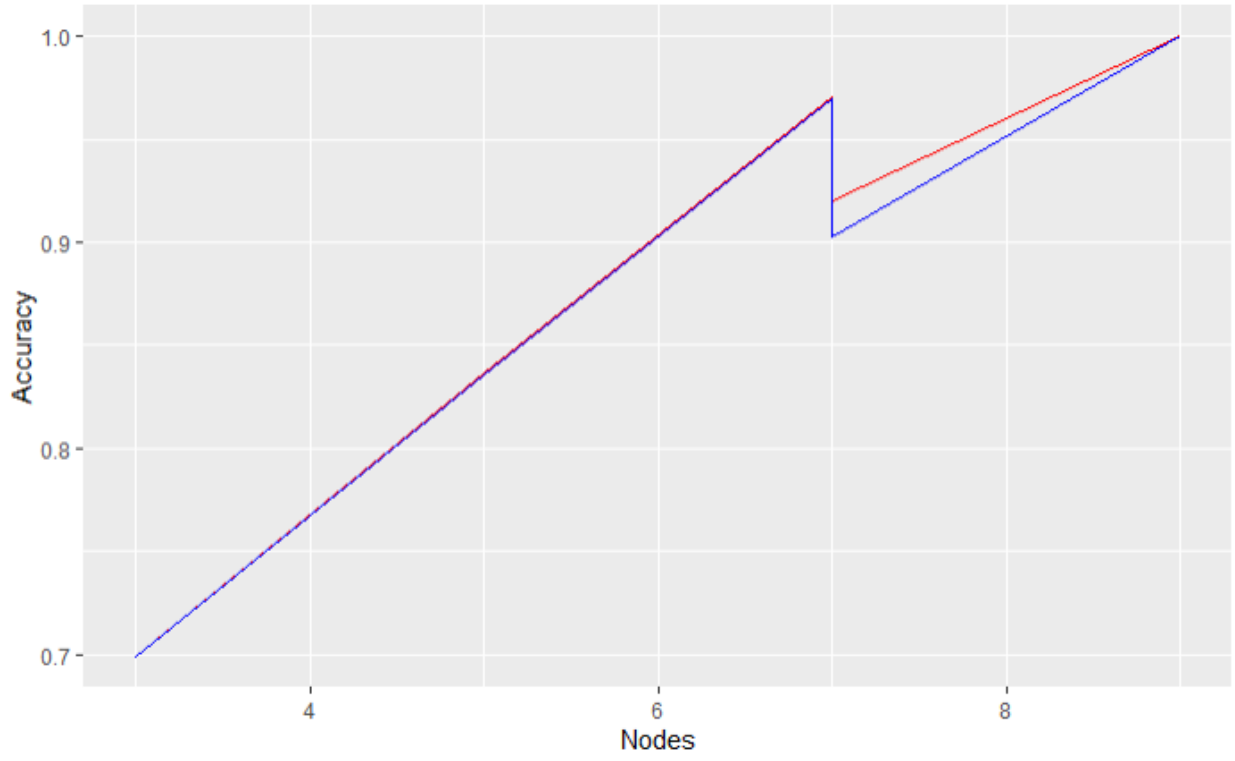
Problem 3

a.

```
index = createDataPartition(y=storms_clean$category, p=0.8, list=FALSE)
train_set = storms_clean[index,]
test_set = storms_clean[-index,]
```

b.

	Nodes <int>	TrainAccuracy <dbl>	TestAccuracy <dbl>	MaxDepth <dbl>	Minsplit <dbl>	Minbucket <dbl>
Accuracy	3	0.6990794	0.6990741	1	3	3
1	5	0.8360184	0.8356481	2	5	5
11	7	0.9700806	0.9699074	3	10	10
12	7	0.9700806	0.9699074	3	15	15
13	9	1.0000000	1.0000000	4	15	15
14	9	1.0000000	1.0000000	4	25	25
15	9	1.0000000	1.0000000	5	50	50
16	7	0.9700806	0.9699074	5	75	75
17	7	0.9700806	0.9699074	6	100	100
18	7	0.9200230	0.9027778	6	250	250
19	3	0.6990794	0.6990741	10	500	500



c. Final model: MaxDepth = 4, MinSplit = 15, Minbucket = 15

Train

```
Confusion Matrix and Statistics

          Reference
Prediction   1   2   3   4   5
         1 867   0   0   0   0
         2   0 348   0   0   0
         3   0   0 233   0   0
         4   0   0   0 238   0
         5   0   0   0   0  52

Overall Statistics

               Accuracy : 1
                 95% CI : (0.9979, 1)
    No Information Rate : 0.4988
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 1

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4 Class: 5
Sensitivity            1.0000   1.0000   1.0000   1.0000  1.00000
Specificity            1.0000   1.0000   1.0000   1.0000  1.00000
Pos Pred Value         1.0000   1.0000   1.0000   1.0000  1.00000
Neg Pred Value         1.0000   1.0000   1.0000   1.0000  1.00000
Prevalence             0.4988   0.2002   0.1341   0.1369  0.02992
Detection Rate         0.4988   0.2002   0.1341   0.1369  0.02992
Detection Prevalence   0.4988   0.2002   0.1341   0.1369  0.02992
Balanced Accuracy      1.0000   1.0000   1.0000   1.0000  1.00000
```

Test

```
Confusion Matrix and Statistics

          Reference
Prediction   1   2   3   4   5
         1 216   0   0   0   0
         2   0  86   0   0   0
         3   0   0  58   0   0
         4   0   0   0  59   0
         5   0   0   0   0  13

Overall Statistics

               Accuracy : 1
                 95% CI : (0.9915, 1)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 1

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4 Class: 5
Sensitivity               1.0   1.0000   1.0000   1.0000  1.00000
Specificity               1.0   1.0000   1.0000   1.0000  1.00000
Pos Pred Value            1.0   1.0000   1.0000   1.0000  1.00000
Neg Pred Value            1.0   1.0000   1.0000   1.0000  1.00000
Prevalence                0.5   0.1991   0.1343   0.1366  0.03009
Detection Rate            0.5   0.1991   0.1343   0.1366  0.03009
Detection Prevalence      0.5   0.1991   0.1343   0.1366  0.03009
Balanced Accuracy         1.0   1.0000   1.0000   1.0000  1.00000
```

Accuracy is 100%.

Problem 4

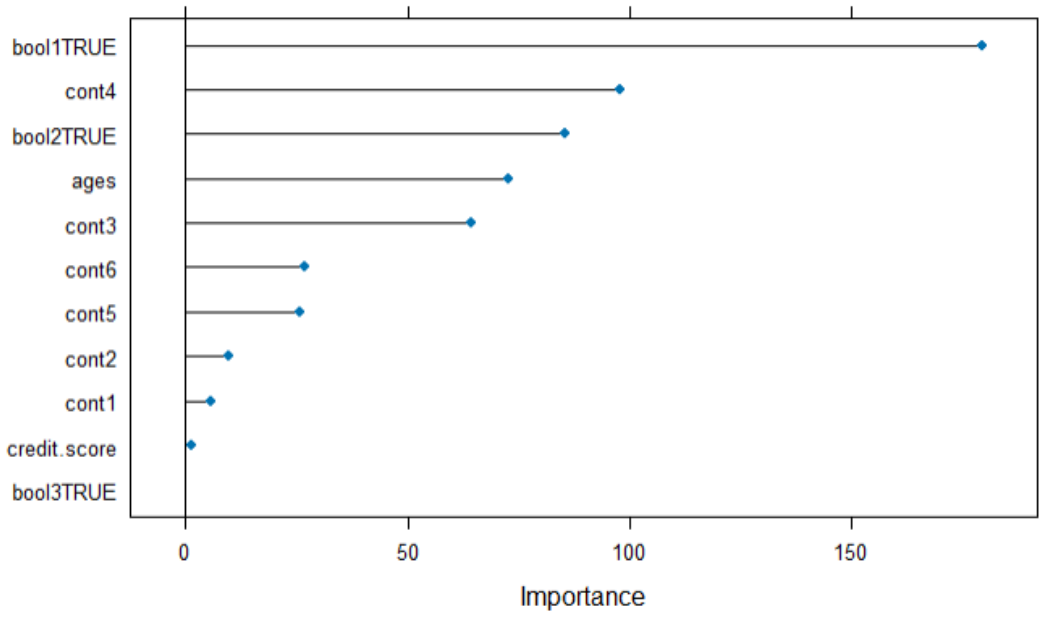
a.

```
CART

666 samples
 11 predictor
  2 classes: '-', '+'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 600, 600, 599, 599, 599, 599, ...
Resampling results:

  Accuracy   Kappa
  0.8844188  0.7668954
```

b.

	Overall <dbl>
bool1TRUE	179.282437
cont4	97.700068
bool2TRUE	85.622001
ages	72.799821
cont3	64.343416
cont6	26.828499
cont5	25.877602
cont2	9.619683
cont1	5.645976
credit.score	1.504253
bool3TRUE	0.000000

c.

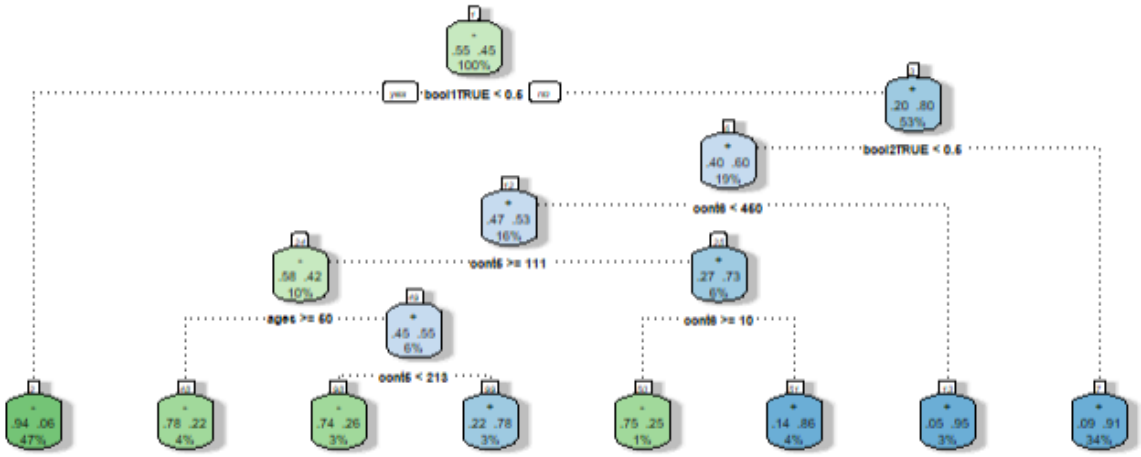


d.

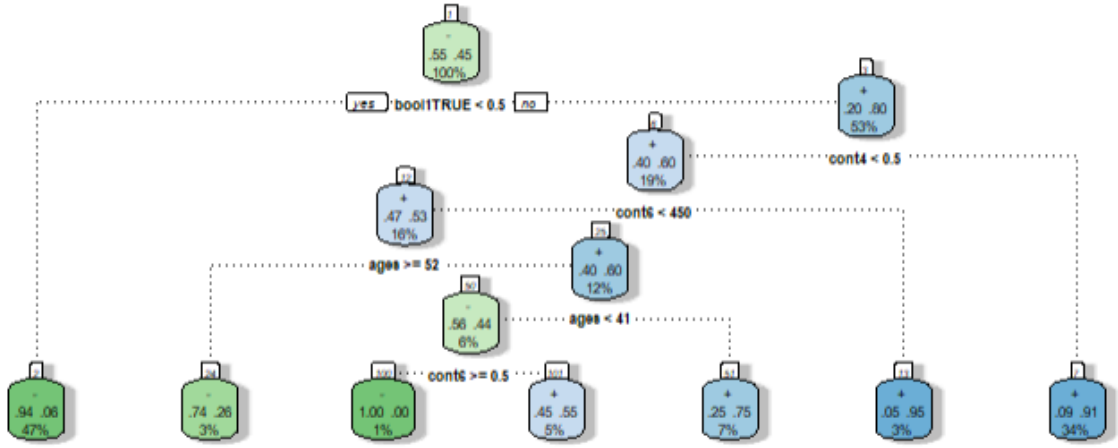
```
CART

666 samples
  6 predictor
  2 classes: '-', '+'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 600, 599, 599, 599, 599, 600, ...
Resampling results:

  Accuracy   Kappa
  0.8663275  0.7310903
```

The new accuracy is 86.63%, down from 88.44% in the full model.

e.



Tree 1



Tree 2

The tree with fewer variables is smaller than the first. With all the variables, the tree has 15 nodes and it is reduced to 13 when using only the top 6 most important variables. The depth, however, remains unchanged at 6 for both trees.