Breast Cancer Prediction

By: Brianna Roseberry

Background

- Breast cancer is a disease in which cells in the breast grow out of control
- It is the most common cancer in women in the United States
  - Each year in the United States, about 245,000 cases of breast cancer are diagnosed in women
- Breast cancer can metastasize by spreading outside the breast through blood vessels and lymph vessels

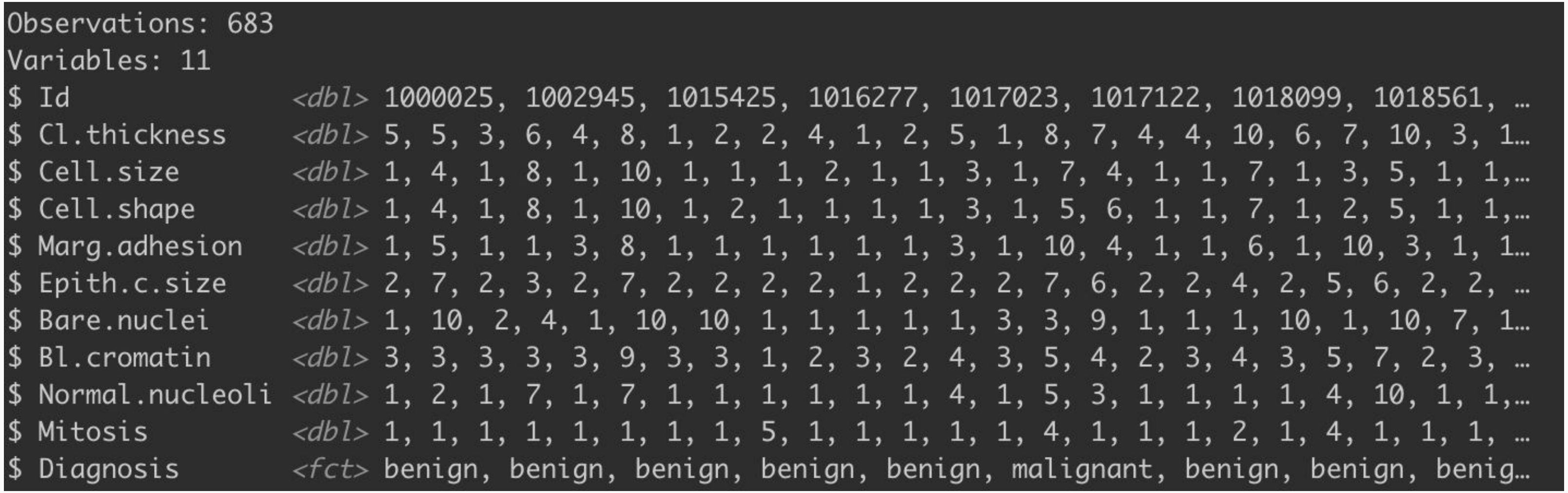
- Deaths from breast cancer have declined over time, but remain the second leading cause of cancer death among women overall
  - About 41,000 women in the U.S. die each year

Data

- UCI Breast Cancer Wisconsin Data Set
  - 683 Observations
  - 11 Variables:
    - ID
    - Diagnosis
    - Clump thickness
    - Cell size
    - Cell shape
    - Marginal Adhesion
    - Single epithelial cell size
    - Bare nuclei
    - Bland chromatin
    - Normal nucleoli
    - Mitosis

Cleaning Data:

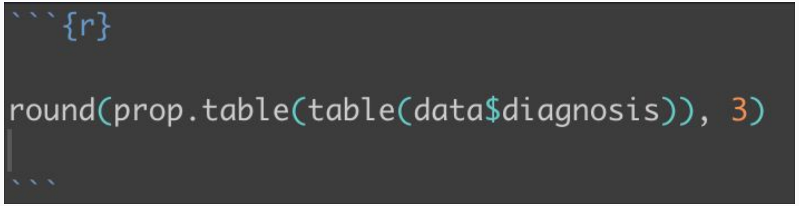
- Checked for NA’s
  - None in data
  - Good to go



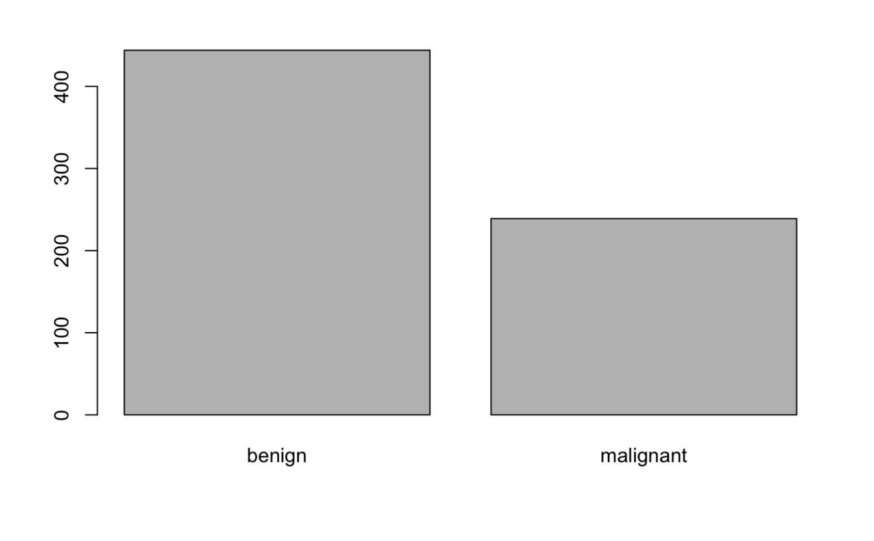
```
Observations: 683
Variables: 11
$ Id              <dbl> 1000025, 1002945, 1015425, 1016277, 1017023, 1017122, 1018099, 1018561, …
$ Cl.thickness    <dbl> 5, 5, 3, 6, 4, 8, 1, 2, 2, 4, 1, 2, 5, 1, 8, 7, 4, 4, 10, 6, 7, 10, 3, 1…
$ Cell.size       <dbl> 1, 4, 1, 8, 1, 10, 1, 1, 1, 2, 1, 1, 3, 1, 7, 4, 1, 1, 7, 1, 3, 5, 1, 1,…
$ Cell.shape      <dbl> 1, 4, 1, 8, 1, 10, 1, 2, 1, 1, 1, 1, 3, 1, 5, 6, 1, 1, 7, 1, 2, 5, 1, 1,…
$ Marg.adhesion   <dbl> 1, 5, 1, 1, 3, 8, 1, 1, 1, 1, 1, 1, 3, 1, 10, 4, 1, 1, 6, 1, 10, 3, 1, 1…
$ Epith.c.size    <dbl> 2, 7, 2, 3, 2, 7, 2, 2, 2, 2, 1, 2, 2, 2, 7, 6, 2, 2, 4, 2, 5, 6, 2, 2, …
$ Bare.nuclei     <dbl> 1, 10, 2, 4, 1, 10, 10, 1, 1, 1, 1, 1, 3, 3, 9, 1, 1, 1, 10, 1, 10, 7, 1…
$ Bl.cromatin     <dbl> 3, 3, 3, 3, 3, 9, 3, 3, 1, 2, 3, 2, 4, 3, 5, 4, 2, 3, 4, 3, 5, 7, 2, 3, …
$ Normal.nucleoli <dbl> 1, 2, 1, 7, 1, 7, 1, 1, 1, 1, 1, 1, 4, 1, 5, 3, 1, 1, 1, 1, 4, 10, 1, 1,…
$ Mitosis         <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 5, 1, 1, 1, 1, 1, 4, 1, 1, 1, 2, 1, 4, 1, 1, 1, …
$ Diagnosis       <fct> benign, benign, benign, benign, benign, malignant, benign, benign, benig…
```

Distribution:

- Looking at distribution of “diagnosis” category
  - Slightly unbalanced
  - M = 0.35
  - B = 0.65

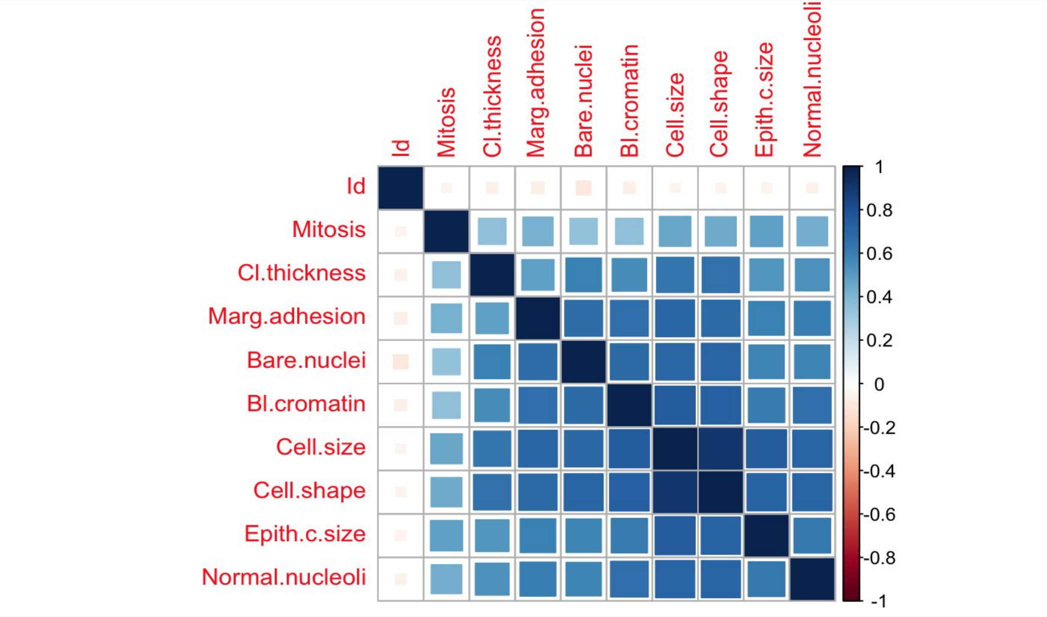


~~~
```{r}

round(prop.table(table(data$diagnosis)), 3)

```
~~~



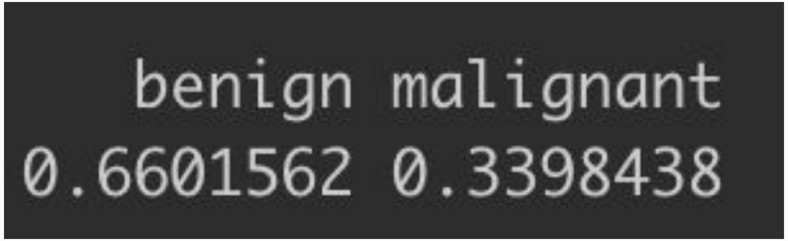
Correlations:

- Looking at correlation in variables
- Used corrplot function
- Many are highly correlated



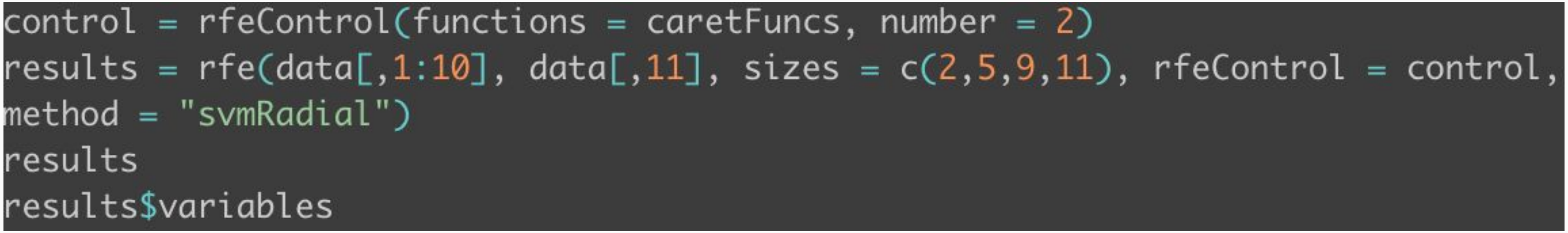
Modeling:

- Split data into test and train dataset
  - Train = 0.75
  - Test = 0.25
- Check to see if test set distribution is accurate

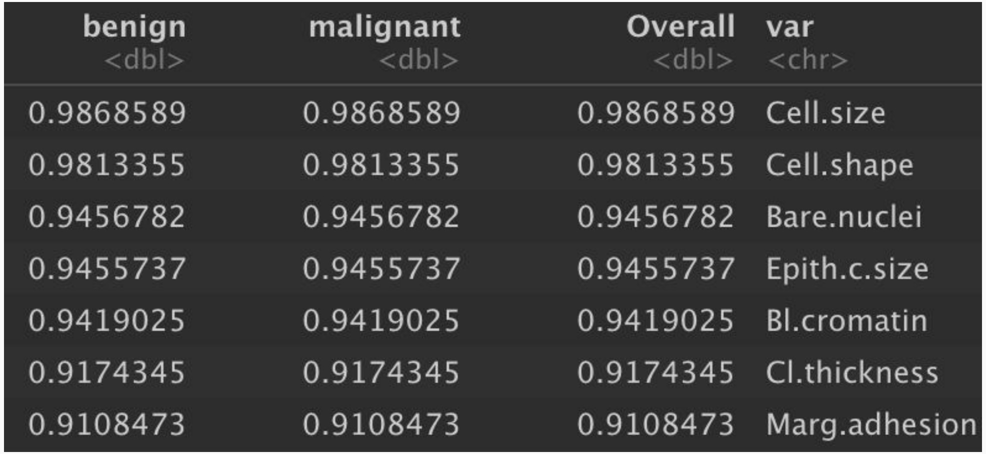


```
   benign malignant
0.6601562 0.3398438
```

Feature Selection: RFE

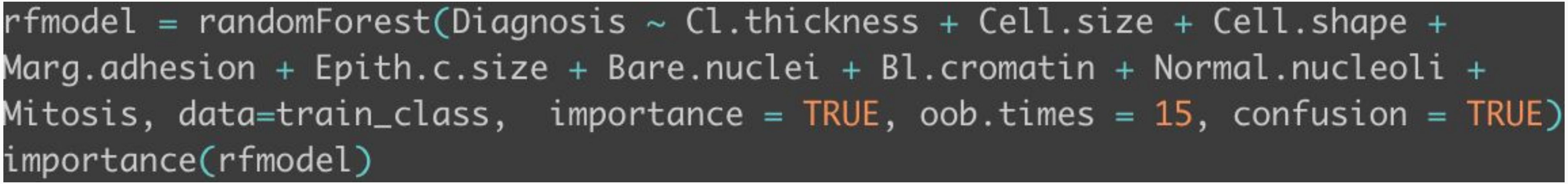


```
control = rfeControl(functions = caretFuncs, number = 2)
results = rfe(data[,1:10], data[,11], sizes = c(2,5,9,11), rfeControl = control,
method = "svmRadial")
results
results$variables
```

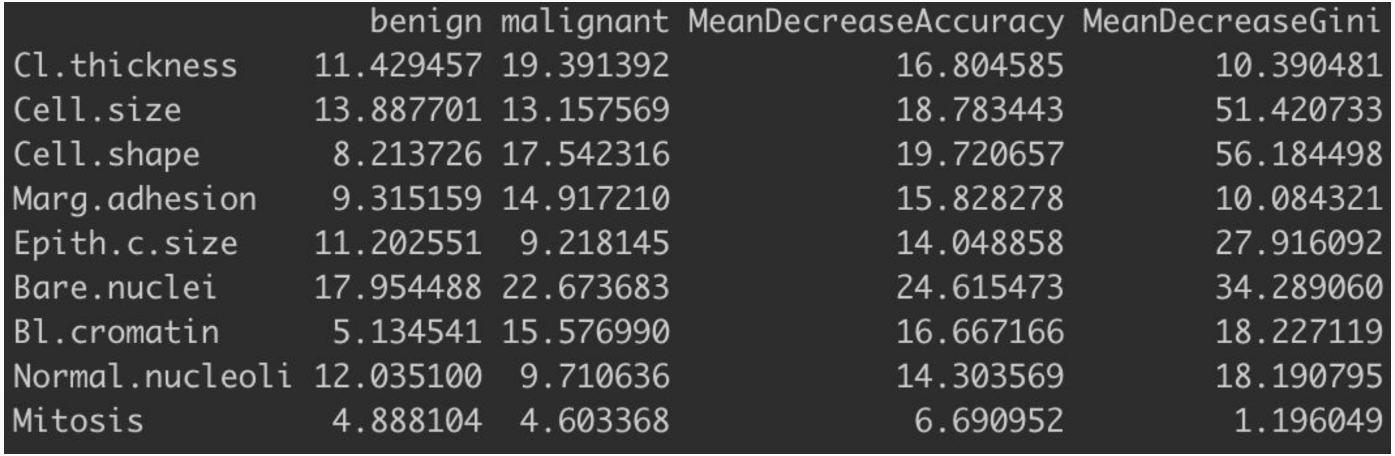


benign <dbl>	malignant <dbl>	Overall <dbl>	var <chr>
0.9868589	0.9868589	0.9868589	Cell.size
0.9813355	0.9813355	0.9813355	Cell.shape
0.9456782	0.9456782	0.9456782	Bare.nuclei
0.9455737	0.9455737	0.9455737	Epith.c.size
0.9419025	0.9419025	0.9419025	Bl.cromatin
0.9174345	0.9174345	0.9174345	Cl.thickness
0.9108473	0.9108473	0.9108473	Marg.adhesion

Feature Selection: Random Forests

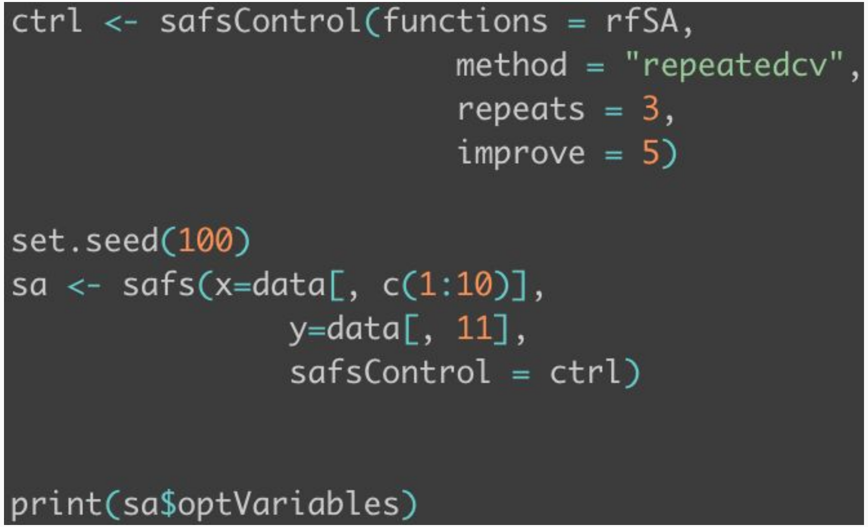


```
rfmodel = randomForest(Diagnosis ~ Cl.thickness + Cell.size + Cell.shape +
Marg.adhesion + Epith.c.size + Bare.nuclei + Bl.cromatin + Normal.nucleoli +
Mitosis, data=train_class,  importance = TRUE, oob.times = 15, confusion = TRUE)
importance(rfmodel)
```

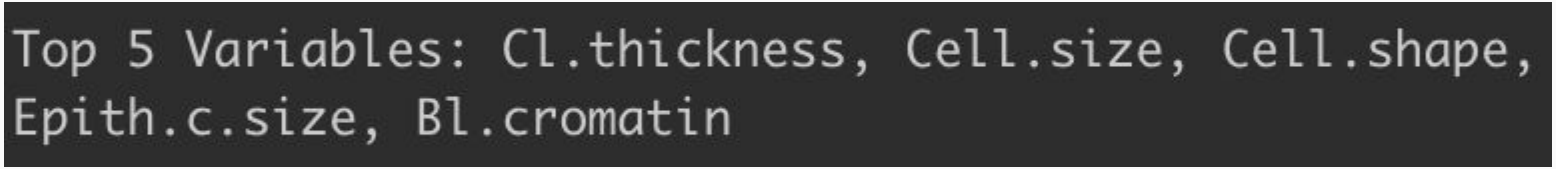


```
                   benign malignant MeanDecreaseAccuracy MeanDecreaseGini
Cl.thickness    11.429457 19.391392            16.804585        10.390481
Cell.size       13.887701 13.157569            18.783443        51.420733
Cell.shape       8.213726 17.542316            19.720657        56.184498
Marg.adhesion    9.315159 14.917210            15.828278        10.084321
Epith.c.size    11.202551  9.218145            14.048858        27.916092
Bare.nuclei     17.954488 22.673683            24.615473        34.289060
Bl.cromatin      5.134541 15.576990            16.667166        18.227119
Normal.nucleoli 12.035100  9.710636            14.303569        18.190795
Mitosis          4.888104  4.603368             6.690952         1.196049
```

Feature Selection: Simulated Annealing



```
ctrl <- safsControl(functions = rfSA,
                        method = "repeatedcv",
                        repeats = 3,
                        improve = 5)

set.seed(100)
sa <- safs(x=data[, c(1:10)],
               y=data[, 11],
               safsControl = ctrl)


print(sa$optVariables)
```

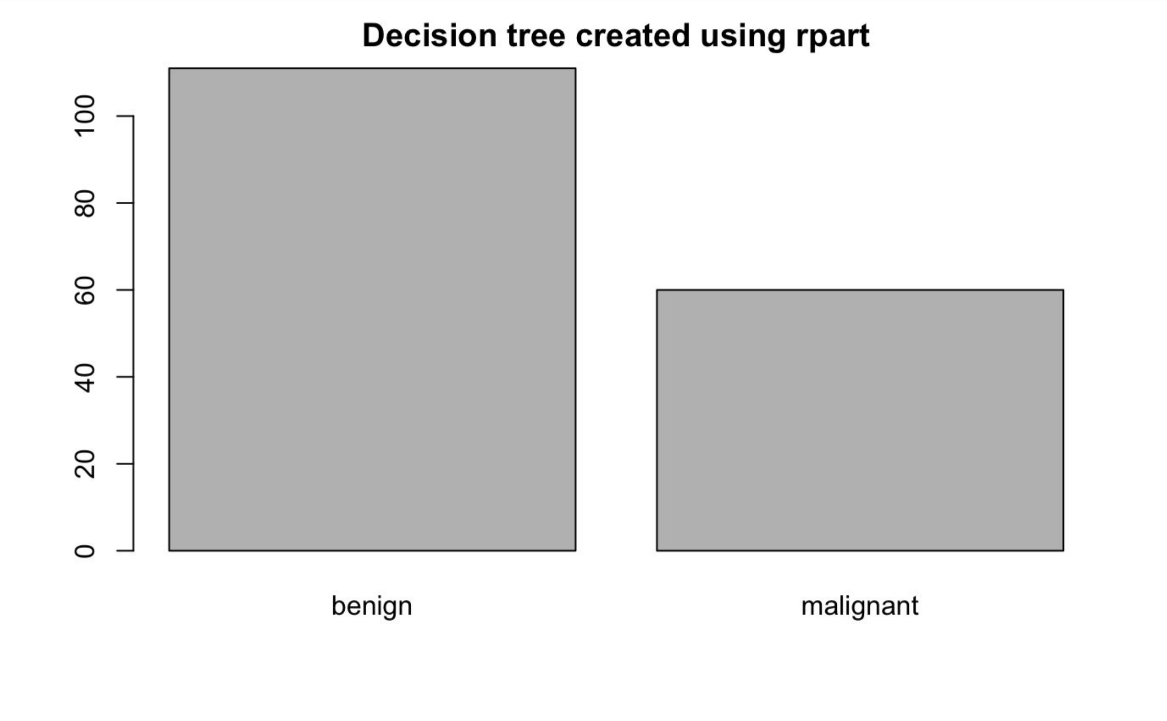


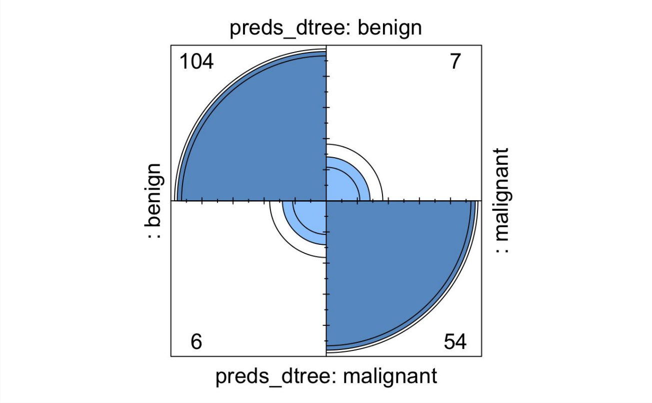
```
Top 5 Variables: Cl.thickness, Cell.size, Cell.shape,
Epith.c.size, Bl.cromatin
```

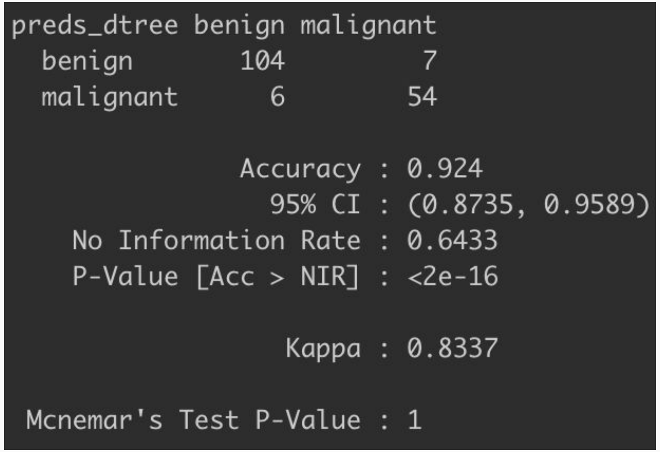
Top 5 Variables:

1. Clump Thickness
2. Cell Size
3. Cell Shape
4. Single Epithelial Cell Size
5. Bland Chromatin

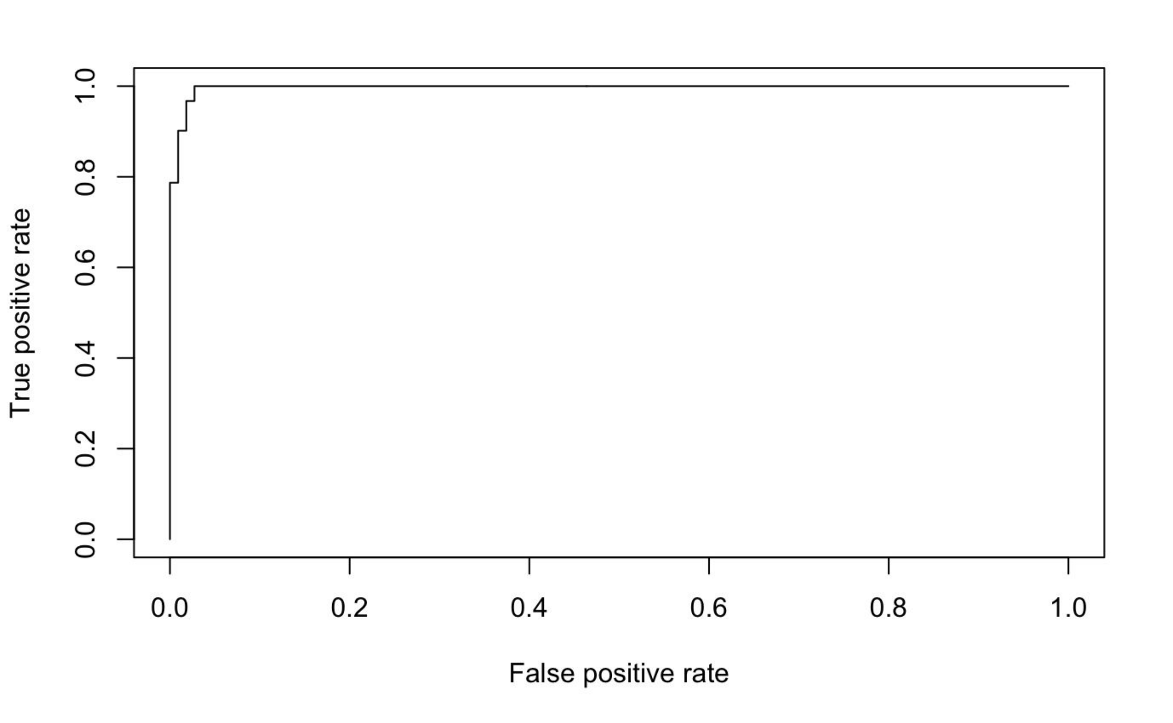
Classification: Decision Tree

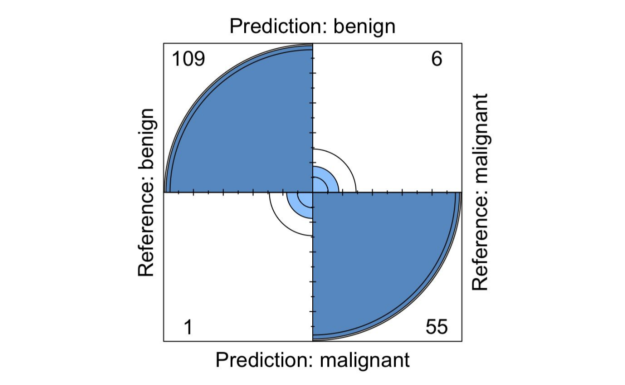

Decision tree created using rpart

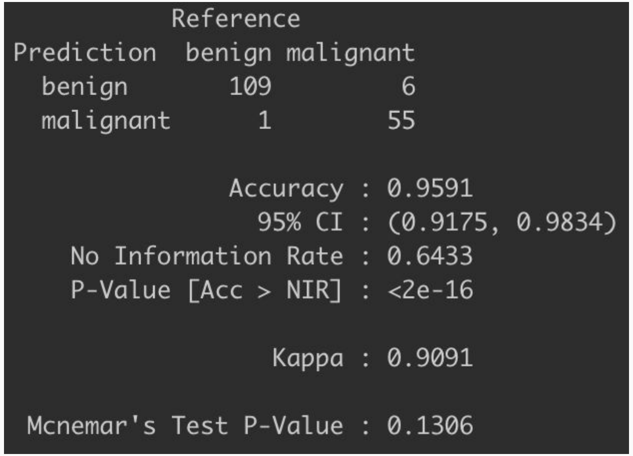



```
preds_dtree benign malignant
  benign       104         7
  malignant      6        54

             Accuracy : 0.924
               95% CI : (0.8735, 0.9589)
   No Information Rate : 0.6433
   P-Value [Acc > NIR] : <2e-16

                Kappa : 0.8337

 Mcnemar's Test P-Value : 1
```

Classification: LDA

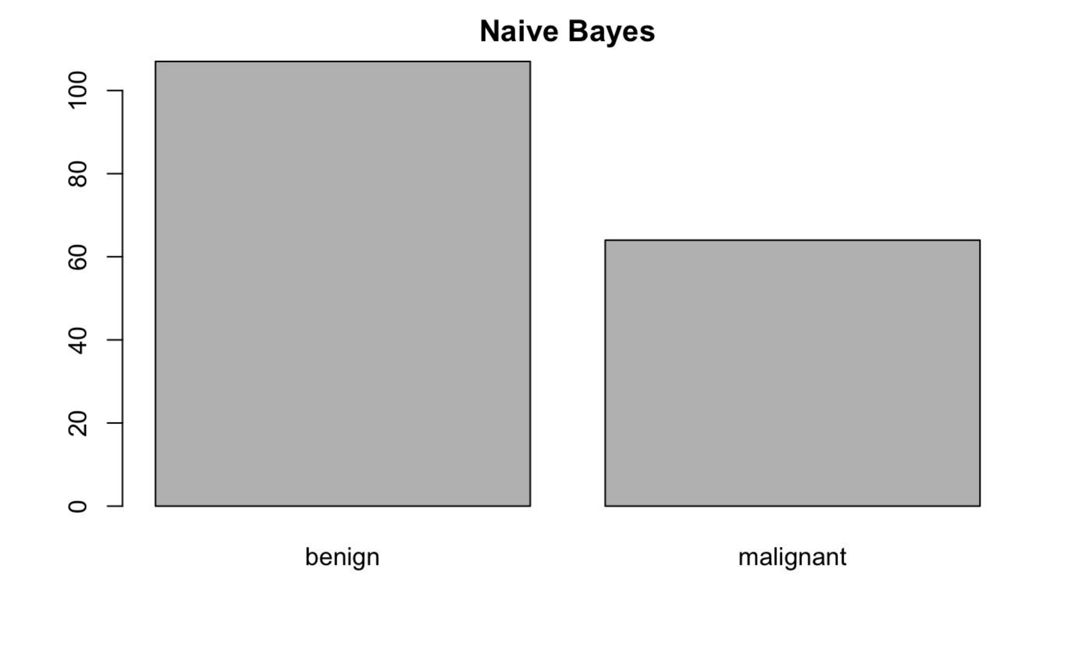


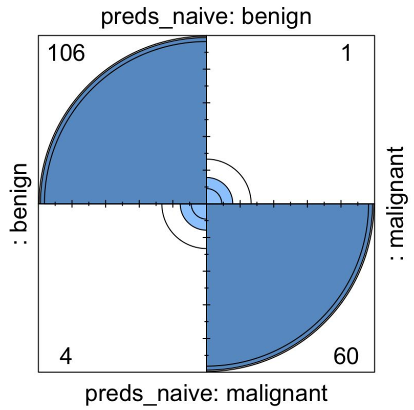


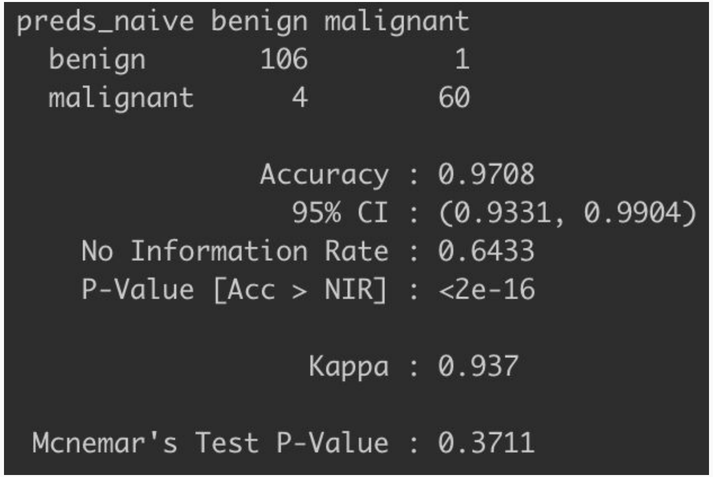


```
          Reference
Prediction  benign malignant
  benign       109         6
  malignant      1        55

               Accuracy : 0.9591
                 95% CI : (0.9175, 0.9834)
    No Information Rate : 0.6433
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9091

 Mcnemar's Test P-Value : 0.1306
```

Classification: Naive Bayes

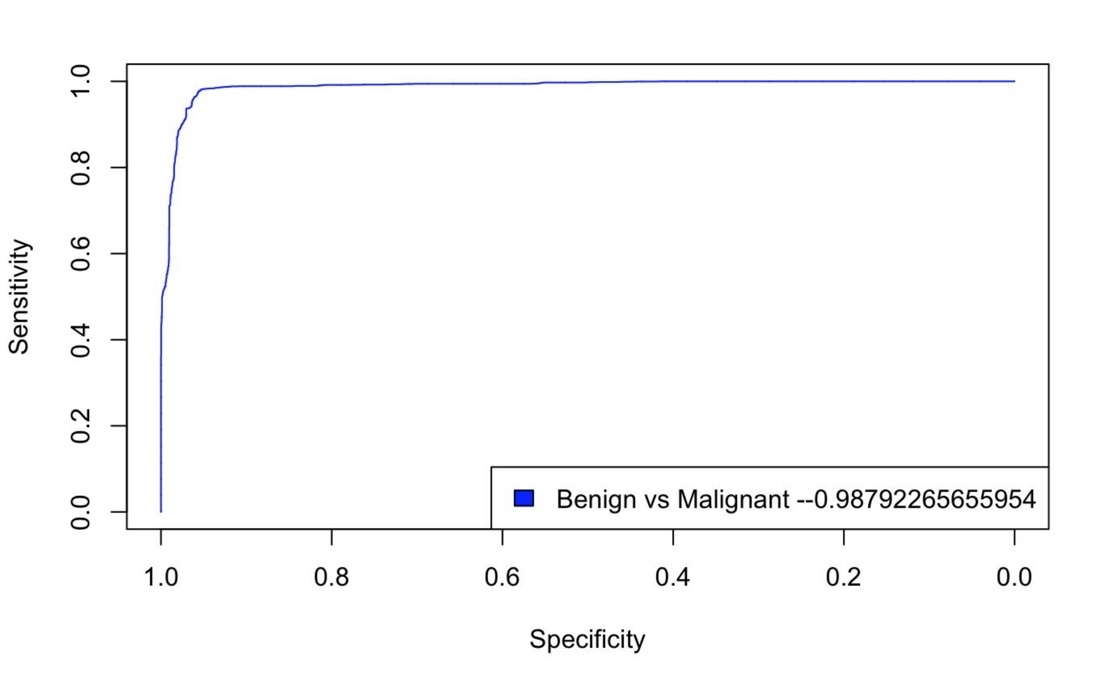


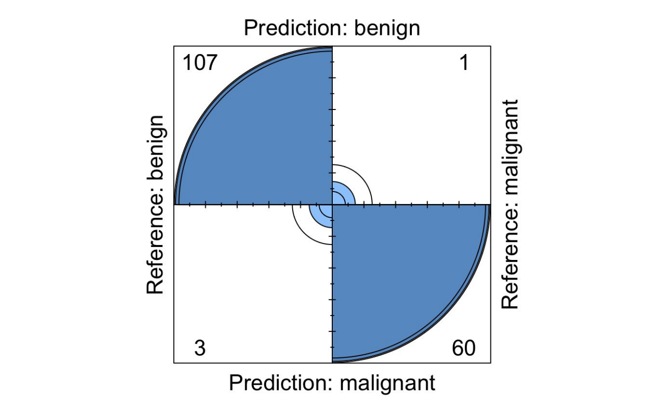


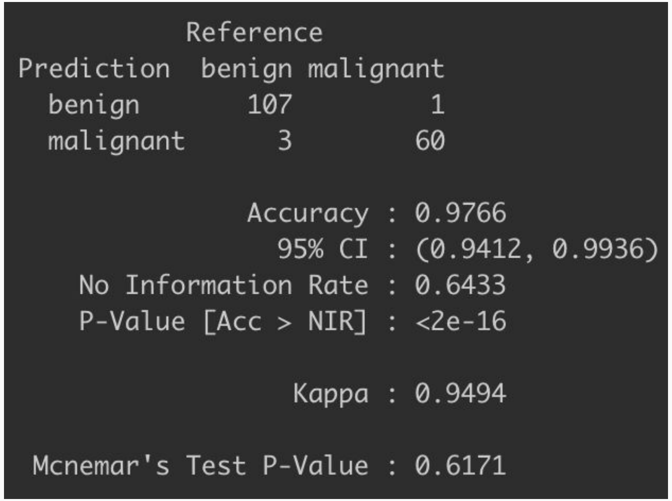


```
preds_naive benign malignant
  benign       106         1
  malignant      4        60

               Accuracy : 0.9708
                 95% CI : (0.9331, 0.9904)
    No Information Rate : 0.6433
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.937

 Mcnemar's Test P-Value : 0.3711
```

Classification: Logistic Regression

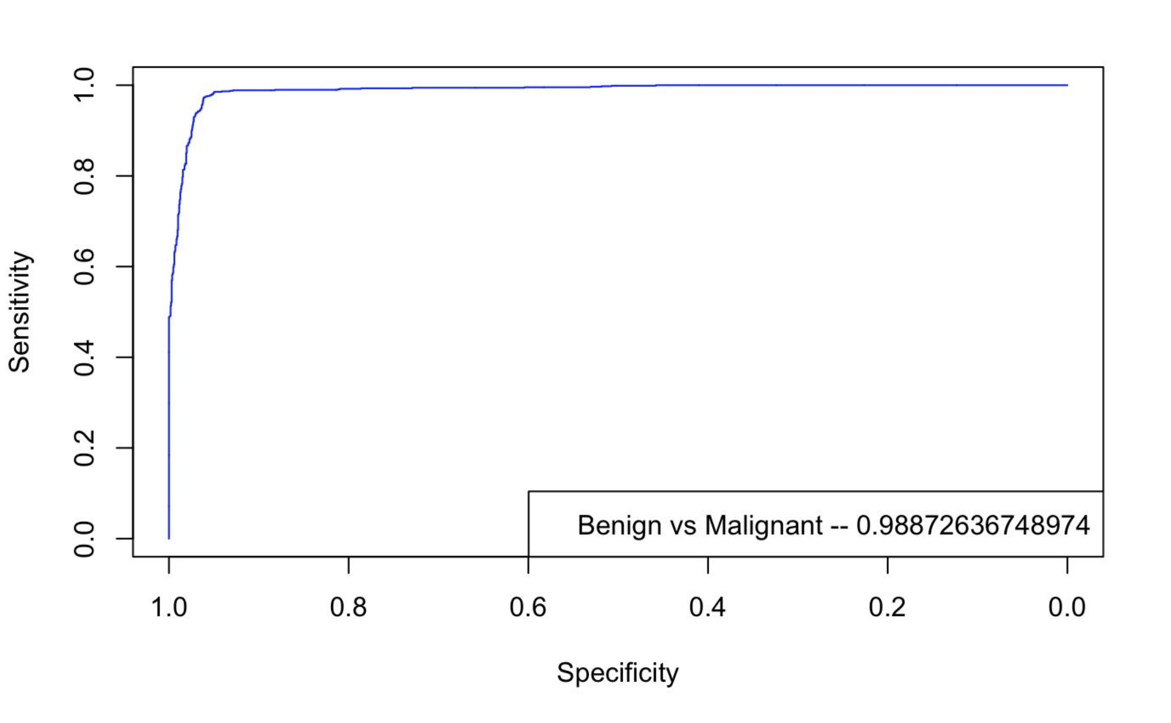


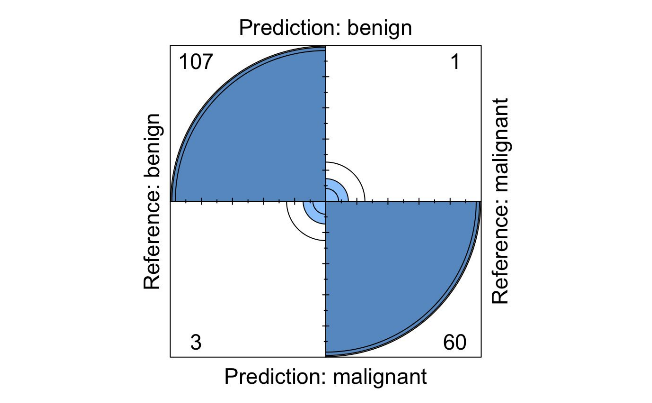


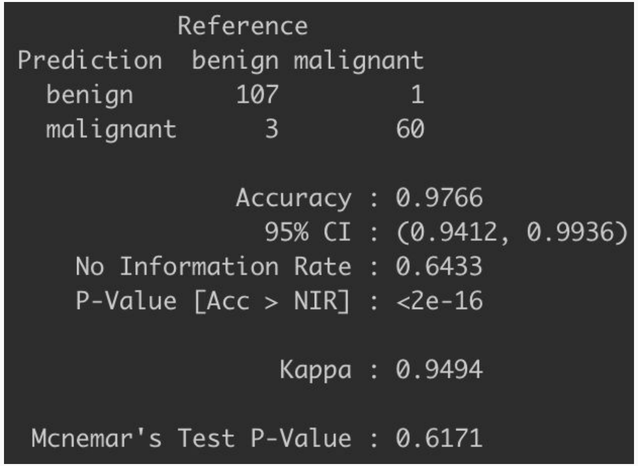


```
            Reference
Prediction  benign malignant
  benign       107         1
  malignant      3        60

               Accuracy : 0.9766
                 95% CI : (0.9412, 0.9936)
    No Information Rate : 0.6433
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9494

 Mcnemar's Test P-Value : 0.6171
```

Classification: SVM







```
            Reference
Prediction  benign malignant
  benign       107         1
  malignant      3        60

               Accuracy : 0.9766
                 95% CI : (0.9412, 0.9936)
    No Information Rate : 0.6433
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9494

 Mcnemar's Test P-Value : 0.6171
```

Thank You!