Wisconsin Breast Cancer Diagnosis Model

Charles Aldrich

April 2, 2019

Overview

The purpose of this machine learning project is to build a breast cancer diagnostic model based on analysis of the University of Wisconsin's Breast Cancer (Diagnostic) dataset. The dataset is a collection of features computed from digitized images of a fine needle aspirate(FNA) of breast mass. These features describe the cell nuclei captured in the image. The diagnostic machine learning model's goal is to predict the mass as benign or malignant. The dataset was split into test and train datasets with the best model selected based on accuracy.

Attribute information in the dataset as described at https://www.kaggle.com/uciml/breast-cancer-wisconsin-data (https://www.kaggle.com/uciml/breast-cancer-wisconsin-data):

1. ID number 2) Diagnosis (M = malignant, B = benign) 3-32)

Ten real-valued features are computed for each cell nucleus: a) radius (mean of distances from center to points on the perimeter) b) texture (standard deviation of gray-scale values) c) perimeter d) area e) smoothness (local variation in radius lengths) f) compactness (perimeter^2 / area - 1.0) g) concavity (severity of concave portions of the contour) h) concave points (number of concave portions of the contour) i) symmetry j) fractal dimension ("coastline approximation" - 1)

The mean, standard error and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius.

All feature values are recoded with four significant digits.

Missing attribute values: none

Class distribution: 357 benign, 212 malignant

NOTE: The class distribution is not balanced. When comparing model results, 'Balanced Accuracy' calculations are used to compensate for the significantly large majority of benign class instances.

Pre-Processing

During Pre-Processing, NULLs were checked and removed, the id column was removed, and the Caret library zero variance check function was run against the predictors to scan for non-descriptive features. No non-variant features were identified by the zero variance check function:

```
## integer(0)
```

Analysis

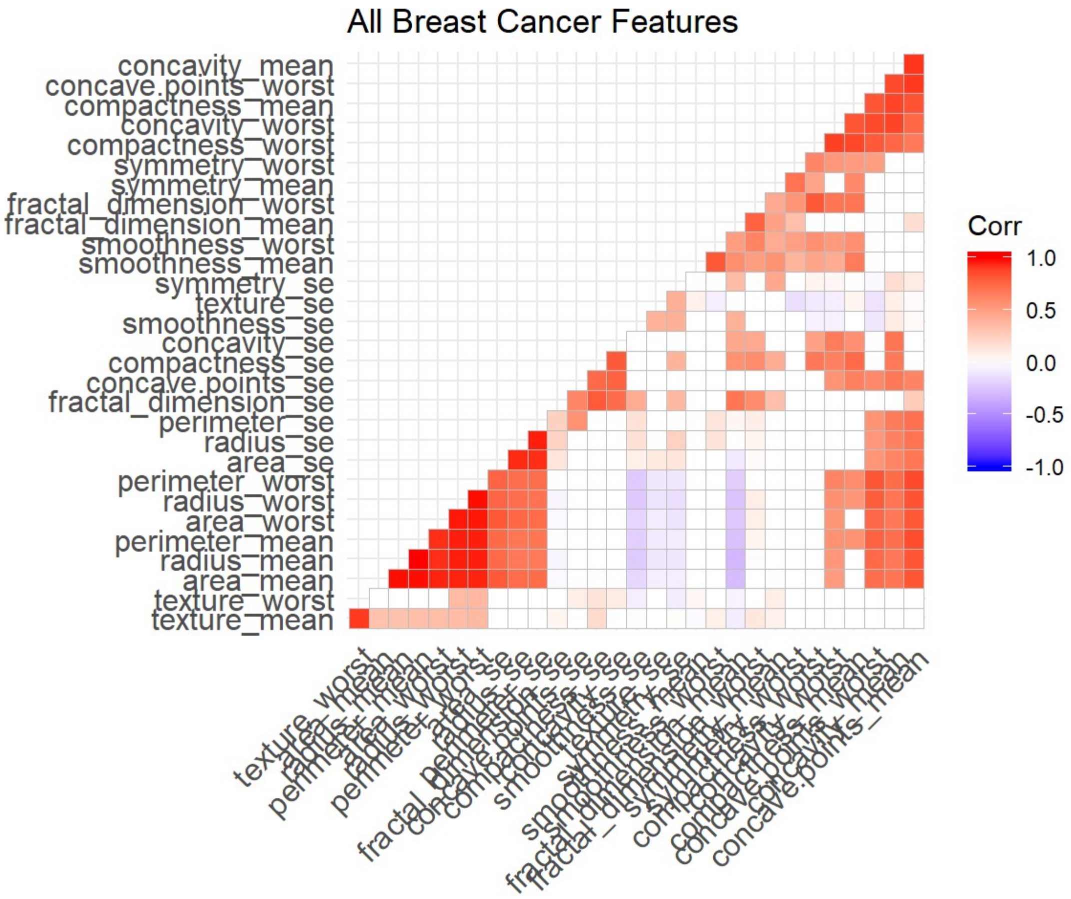
Dataset Structure:

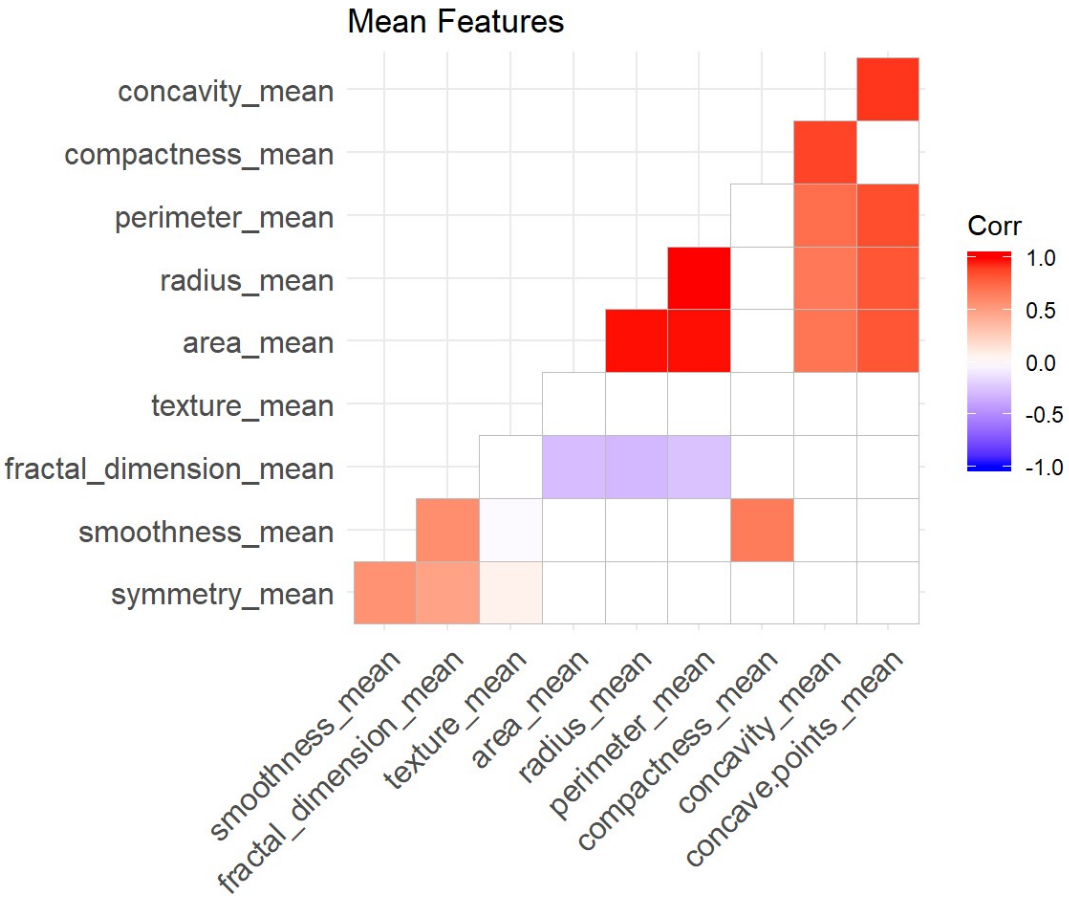
```
## 'data.frame':    569 obs. of  31 variables:
##  $ diagnosis              : Factor w/ 2 levels "B","M": 2 2 2 2 2 2 2 2 2 2 ...
##  $ radius_mean            : num  18 20.6 19.7 11.4 20.3 ...
##  $ texture_mean           : num  10.4 17.8 21.2 20.4 14.3 ...
##  $ perimeter_mean         : num  122.8 132.9 130 77.6 135.1 ...
##  $ area_mean              : num  1001 1326 1203 386 1297 ...
##  $ smoothness_mean        : num  0.1184 0.0847 0.1096 0.1425 0.1003 ...
##  $ compactness_mean       : num  0.2776 0.0786 0.1599 0.2839 0.1328 ...
##  $ concavity_mean         : num  0.3001 0.0869 0.1974 0.2414 0.198 ...
##  $ concave.points_mean    : num  0.1471 0.0702 0.1279 0.1052 0.1043 ...
##  $ symmetry_mean          : num  0.242 0.181 0.207 0.26 0.181 ...
##  $ fractal_dimension_mean : num  0.0787 0.0567 0.06 0.0974 0.0588 ...
##  $ radius_se              : num  1.095 0.543 0.746 0.496 0.757 ...
##  $ texture_se             : num  0.905 0.734 0.787 1.156 0.781 ...
##  $ perimeter_se           : num  8.59 3.4 4.58 3.44 5.44 ...
##  $ area_se                : num  153.4 74.1 94 27.2 94.4 ...
##  $ smoothness_se          : num  0.0064 0.00522 0.00615 0.00911 0.01149 ...
##  $ compactness_se         : num  0.049 0.0131 0.0401 0.0746 0.0246 ...
##  $ concavity_se           : num  0.0537 0.0186 0.0383 0.0566 0.0569 ...
##  $ concave.points_se      : num  0.0159 0.0134 0.0206 0.0187 0.0188 ...
##  $ symmetry_se            : num  0.03 0.0139 0.0225 0.0596 0.0176 ...
##  $ fractal_dimension_se   : num  0.00619 0.00353 0.00457 0.00921 0.00511 ...
##  $ radius_worst           : num  25.4 25 23.6 14.9 22.5 ...
##  $ texture_worst          : num  17.3 23.4 25.5 26.5 16.7 ...
##  $ perimeter_worst        : num  184.6 158.8 152.5 98.9 152.2 ...
##  $ area_worst             : num  2019 1956 1709 568 1575 ...
##  $ smoothness_worst       : num  0.162 0.124 0.144 0.21 0.137 ...
##  $ compactness_worst      : num  0.666 0.187 0.424 0.866 0.205 ...
##  $ concavity_worst        : num  0.712 0.242 0.45 0.687 0.4 ...
##  $ concave.points_worst   : num  0.265 0.186 0.243 0.258 0.163 ...
##  $ symmetry_worst         : num  0.46 0.275 0.361 0.664 0.236 ...
##  $ fractal_dimension_worst: num  0.1189 0.089 0.0876 0.173 0.0768 ...
```

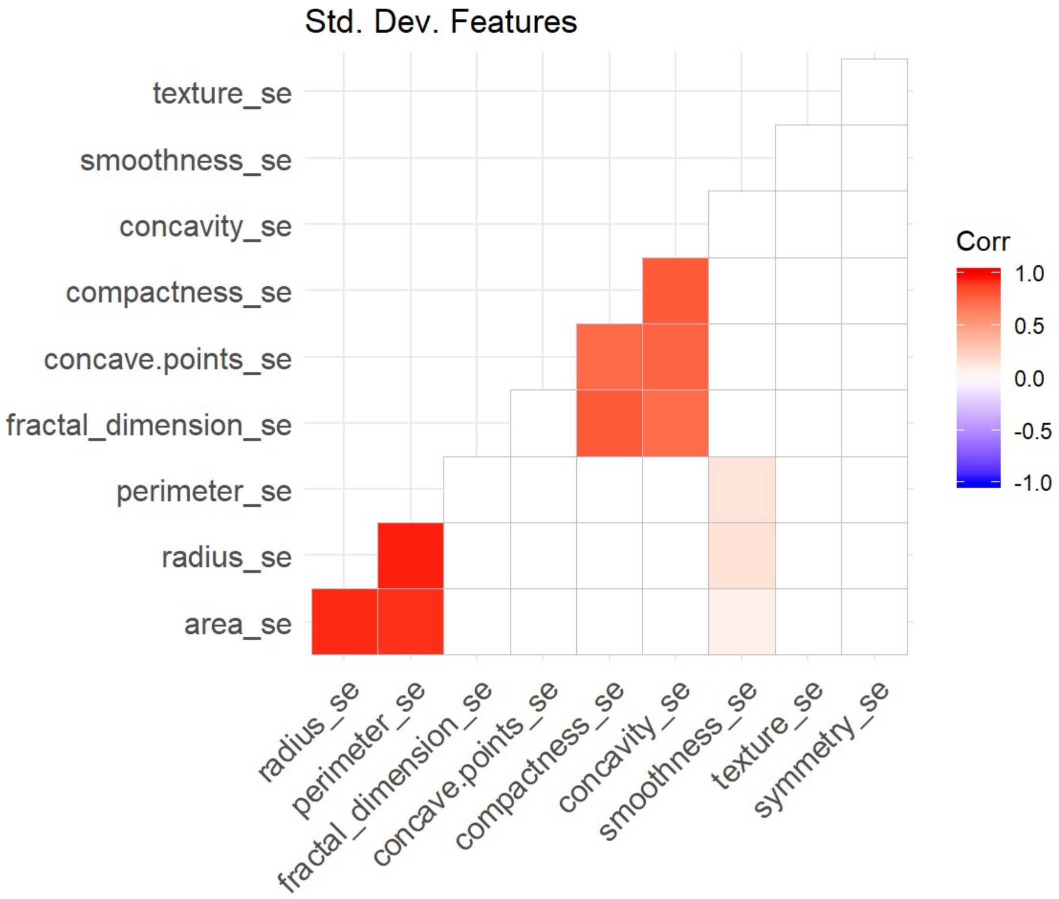
Dataset Summary Statistics:

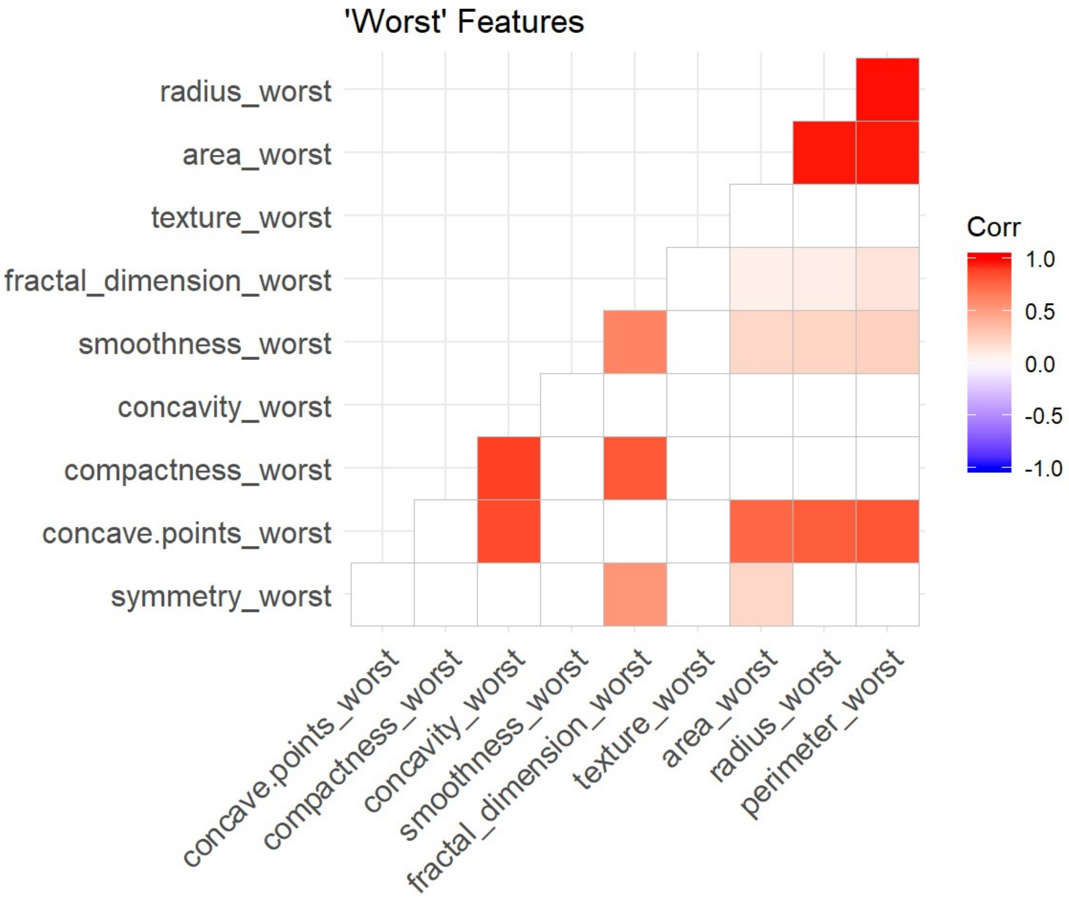
```
##  diagnosis  radius_mean      texture_mean   perimeter_mean
##  B:357     Min.   : 6.981   Min.   : 9.71   Min.   : 43.79
##  M:212     1st Qu.:11.700   1st Qu.:16.17   1st Qu.: 75.17
##            Median :13.370   Median :18.84   Median : 86.24
##            Mean   :14.127   Mean   :19.29   Mean   : 91.97
##            3rd Qu.:15.780   3rd Qu.:21.80   3rd Qu.:104.10
##            Max.   :28.110   Max.   :39.28   Max.   :188.50
##    area_mean      smoothness_mean   compactness_mean  concavity_mean
##  Min.   : 143.5   Min.   :0.05263   Min.   :0.01938   Min.   :0.00000
##  1st Qu.: 420.3   1st Qu.:0.08637   1st Qu.:0.06492   1st Qu.:0.02956
##  Median : 551.1   Median :0.09587   Median :0.09263   Median :0.06154
##  Mean   : 654.9   Mean   :0.09636   Mean   :0.10434   Mean   :0.08880
##  3rd Qu.: 782.7   3rd Qu.:0.10530   3rd Qu.:0.13040   3rd Qu.:0.13070
##  Max.   :2501.0   Max.   :0.16340   Max.   :0.34540   Max.   :0.42680
##  concave.points_mean symmetry_mean    fractal_dimension_mean
##  Min.   :0.00000     Min.   :0.1060   Min.   :0.04996
##  1st Qu.:0.02031     1st Qu.:0.1619   1st Qu.:0.05770
##  Median :0.03350     Median :0.1792   Median :0.06154
##  Mean   :0.04892     Mean   :0.1812   Mean   :0.06280
##  3rd Qu.:0.07400     3rd Qu.:0.1957   3rd Qu.:0.06612
##  Max.   :0.20120     Max.   :0.3040   Max.   :0.09744
##    radius_se        texture_se      perimeter_se       area_se
##  Min.   :0.1115   Min.   :0.3602   Min.   : 0.757   Min.   :  6.802
##  1st Qu.:0.2324   1st Qu.:0.8339   1st Qu.: 1.606   1st Qu.: 17.850
##  Median :0.3242   Median :1.1080   Median : 2.287   Median : 24.530
##  Mean   :0.4052   Mean   :1.2169   Mean   : 2.866   Mean   : 40.337
##  3rd Qu.:0.4789   3rd Qu.:1.4740   3rd Qu.: 3.357   3rd Qu.: 45.190
##  Max.   :2.8730   Max.   :4.8850   Max.   :21.980   Max.   :542.200
##  smoothness_se      compactness_se      concavity_se
##  Min.   :0.001713   Min.   :0.002252   Min.   :0.00000
##  1st Qu.:0.005169   1st Qu.:0.013080   1st Qu.:0.01509
##  Median :0.006380   Median :0.020450   Median :0.02589
##  Mean   :0.007041   Mean   :0.025478   Mean   :0.03189
##  3rd Qu.:0.008146   3rd Qu.:0.032450   3rd Qu.:0.04205
##  Max.   :0.031130   Max.   :0.135400   Max.   :0.39600
##  concave.points_se   symmetry_se       fractal_dimension_se
##  Min.   :0.000000   Min.   :0.007882   Min.   :0.0008948
##  1st Qu.:0.007638   1st Qu.:0.015160   1st Qu.:0.0022480
##  Median :0.010930   Median :0.018730   Median :0.0031870
##  Mean   :0.011796   Mean   :0.020542   Mean   :0.0037949
##  3rd Qu.:0.014710   3rd Qu.:0.023480   3rd Qu.:0.0045580
##  Max.   :0.052790   Max.   :0.078950   Max.   :0.0298400
##   radius_worst   texture_worst   perimeter_worst    area_worst
##  Min.   : 7.93   Min.   :12.02   Min.   : 50.41   Min.   : 185.2
##  1st Qu.:13.01   1st Qu.:21.08   1st Qu.: 84.11   1st Qu.: 515.3
##  Median :14.97   Median :25.41   Median : 97.66   Median : 686.5
##  Mean   :16.27   Mean   :25.68   Mean   :107.26   Mean   : 880.6
##  3rd Qu.:18.79   3rd Qu.:29.72   3rd Qu.:125.40   3rd Qu.:1084.0
##  Max.   :36.04   Max.   :49.54   Max.   :251.20   Max.   :4254.0  
##  smoothness_worst  compactness_worst concavity_worst  concave.points_worst
##  Min.   :0.07117   Min.   :0.02729   Min.   :0.0000   Min.   :0.00000     
##  1st Qu.:0.11660   1st Qu.:0.14720   1st Qu.:0.1145   1st Qu.:0.06493     
##  Median :0.13130   Median :0.21190   Median :0.2267   Median :0.09993     
##  Mean   :0.13237   Mean   :0.25427   Mean   :0.2722   Mean   :0.11461     
##  3rd Qu.:0.14600   3rd Qu.:0.33910   3rd Qu.:0.3829   3rd Qu.:0.16140     
##  Max.   :0.22260   Max.   :1.05800   Max.   :1.2520   Max.   :0.29100     
##  symmetry_worst   fractal_dimension_worst
##  Min.   :0.1565   Min.   :0.05504        
##  1st Qu.:0.2504   1st Qu.:0.07146        
##  Median :0.2822   Median :0.08004        
##  Mean   :0.2901   Mean   :0.08395        
##  3rd Qu.:0.3179   3rd Qu.:0.09208        
##  Max.   :0.6638   Max.   :0.20750
```

Features were analyzed for correlations. Plots were split by type of feature (mean, sd, or 'worst') to improve visual. Insignifcant features were removed from the triangle matrix. Some correlations are expected due to geometry (i.e. radius and area). Of interest are correlations of concavity and compactness across the four plots.




Mean Features
concavity_mean
compactness_mean
perimeter_mean
radius_mean
area_mean
texture_mean
fractal_dimension_mean
smoothness_mean
symmetry_mean
smoothness_mean
fractal_dimension_mean
texture_mean
area_mean
radius_mean
perimeter_mean
compactness_mean
concavity_mean
concave.points_mean
Corr
1.0
0.5
0.0
-0.5
-1.0


Std. Dev. Features
texture_se
smoothness_se
concavity_se
compactness_se
concave.points_se
fractal_dimension_se
perimeter_se
radius_se
area_se
radius_se
perimeter_se
fractal_dimension_se
concave.points_se
compactness_se
concavity_se
smoothness_se
texture_se
symmetry_se
Corr
1.0
0.5
0.0
-0.5
-1.0



Modeling

The dataset was split into train and test datasets with a 60/40 ratio. Train and Test dataset diagnosis (B/M) column proportions were approximately equal. However the B and M class ratio is not balanced.

```
## 
##        B        M 
## 0.627566 0.372434
```

```
## 
##        B        M 
## 0.627193 0.372807
```

The following models were evaluated in order of increasing Balanced Accuracy. A Confusion Matrix (cm function) was used to rely on the function’s calculation of Balanced Accuracy. Balanced Accuracy was chosen as the deciding metric vs simple Accuracy given Benign (B) class instances greatly outnumber Malignant (M) instances in the dataset.

1. Decision Tree (rpart)
2. Random Forest
3. SVM Linear

4. SVM Radial

Confusion Matrix for Decision Tree (rpart):

```
## Confusion Matrix and Statistics
## 
##           Reference
## Prediction   B   M
##          B 138  10
##          M   5  75
##                                           
##                Accuracy : 0.9342          
##                  95% CI : (0.8938, 0.9627)
##     No Information Rate : 0.6272          
##     P-Value [Acc > NIR] : <2e-16          
##                                           
##                   Kappa : 0.8576          
##  Mcnemar's Test P-Value : 0.3017          
##                                           
##             Sensitivity : 0.9650          
##             Specificity : 0.8824          
##          Pos Pred Value : 0.9324          
##          Neg Pred Value : 0.9375          
##              Prevalence : 0.6272          
##          Detection Rate : 0.6053          
##    Detection Prevalence : 0.6491          
##       Balanced Accuracy : 0.9237          
##                                           
##        'Positive' Class : B               
## 
```

Confusion Matrix for Random Forest shows a significant bump in Balanced Accuracy to over .95:

```
## Confusion Matrix and Statistics
## 
##           Reference
## Prediction   B   M
##          B 141   7
##          M   2  78
## 
##                Accuracy : 0.9605
##                  95% CI : (0.9264, 0.9818)
##     No Information Rate : 0.6272
##     P-Value [Acc > NIR] : <2e-16
## 
##                   Kappa : 0.9146
##  Mcnemar's Test P-Value : 0.1824
## 
##             Sensitivity : 0.9860
##             Specificity : 0.9176
##          Pos Pred Value : 0.9527
##          Neg Pred Value : 0.9750
##              Prevalence : 0.6272
##          Detection Rate : 0.6184
##    Detection Prevalence : 0.6491
##       Balanced Accuracy : 0.9518
## 
##        'Positive' Class : B
## 
```

Confusion Matrix for SVM Linear increases Balanced Accuracy just slightly:

```
## Confusion Matrix and Statistics
## 
##           Reference
## Prediction   B   M
##          B 142   7
##          M   1  78
##                                           
##                Accuracy : 0.9649          
##                  95% CI : (0.932, 0.9847)
##     No Information Rate : 0.6272          
##     P-Value [Acc > NIR] : <2e-16          
##                                           
##                   Kappa : 0.9239          
##  Mcnemar's Test P-Value : 0.0771          
##                                           
##             Sensitivity : 0.9930          
##             Specificity : 0.9176          
##          Pos Pred Value : 0.9530          
##          Neg Pred Value : 0.9873          
##              Prevalence : 0.6272          
##          Detection Rate : 0.6228          
##    Detection Prevalence : 0.6535          
##       Balanced Accuracy : 0.9553          
##                                           
##        'Positive' Class : B               
## 
```

Confusion Matrix for SVM Radial gives another bump to Balanced Accuracy to over .97. This reflects how the dataset predictors better fit a non-linear SVM model prividing for 3 more accurate predictions and accurately predicts *all* Benign tumors!

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   B   M
##          B 143   5
##          M   0  80
##
##                Accuracy : 0.9781
##                  95% CI : (0.9496, 0.9928)
##     No Information Rate : 0.6272
##     P-Value [Acc > NIR] : < 2e-16
##
##                   Kappa : 0.9525
##  Mcnemar's Test P-Value : 0.07364
##
##             Sensitivity : 1.0000
##             Specificity : 0.9412
##          Pos Pred Value : 0.9662
##          Neg Pred Value : 1.0000
##              Prevalence : 0.6272
##          Detection Rate : 0.6272
##    Detection Prevalence : 0.6491
##       Balanced Accuracy : 0.9706
##
##        'Positive' Class : B
##
```

Conclusion

The Balanced Accuracy for a SVM Radial model against the entire Wisconsion Breast Cancer (Diagnosis) dataset is over .98:

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   B   M
##          B 357   8
##          M   0 204
##
##                Accuracy : 0.9859
##                  95% CI : (0.9725, 0.9939)
##     No Information Rate : 0.6274
##     P-Value [Acc > NIR] : < 2e-16
##
##                   Kappa : 0.9697
##  Mcnemar's Test P-Value : 0.01333
##
##             Sensitivity : 1.0000
##             Specificity : 0.9623
##          Pos Pred Value : 0.9781
##          Neg Pred Value : 1.0000
##              Prevalence : 0.6274
##          Detection Rate : 0.6274
##    Detection Prevalence : 0.6415
##       Balanced Accuracy : 0.9811
##
##        'Positive' Class : B
##
```

Like during testing, all benign (B) tumors were accurately predicted using the SVM Radial model. SVM Radial was over 96% accurate in identifying malignant (M) tumors, 204 out of 212 malignancies were predicted.