Final Project Report

Olivia Fan, Alicia Gong

18 December, 2022

Introduction

Breast cancer is the most common cancer worldwide and the most common cancer diagnosed in the US (Mayo). Each year in the US, about 264,000 cases of breast cancer are diagnosed in women and about 2,400 in men (CDC).

Early diagnosis of the condition is crucial to improve the survival rate and relieve suffering in patients. Mammography is an effective X-ray imaging technology that detects breast cancer early. Classically, benign or malignant breast tumors are diagnosed by radiologists' interpretation of mammograms based on clinical parameters. However, diagnosing cancer is challenging even for the most skilled doctors. Since masses are heterogeneous, clinical parameters supply limited information on mammography mass. The symptoms are often shared with diseases and conditions that are unrelated to cancer, leading doctors to improperly diagnose the disease.

Cancerous lumps are often confused for blocked milk ducts, breast cysts, and other benign conditions. According to an expansive study conducted by Dartmouth College, the University of Vermont, and the Fred Hutchinson Cancer Research Center, and published in the March 2015 issue of the Journal of American Medical Association, approximately 13% of the diagnoses missed Stage 1 breast cancer. Meanwhile, 48% failed to detect atypia hyperplasia, a precursor to breast cancer. A significant number also over-diagnosed atypia hyperplasia.

There is, therefore, an urgent need to find new tools that can identify patients with breast cancer. Our study aims to build supervised machine-learning models to predict the diagnosis of breast cancer and understand the most important variables, to assist doctors and radiologists in accurately interpreting mammography imaging.

We built 3 models in total: Lasso penalized logistic regression, SVM, and Random Forest.

Data

We obtain the Breast Cancer Wisconsin (Diagnosis) Data Set from Kaggle. The dataset contains diagnosis results and features of the cell nuclei computed from a digitized image of a fine needle aspirate (FNA) of a breast mass for 568 patients. The size of the nucleus is expressed by the features radius and area. The shape is expressed by the features smoothness, concavity, compactness, concave points, symmetry, and fractal dimension. The perimeter expresses both the size and shape of the nucleus. A higher value of shape features corresponds to a less regular contour and, therefore, to a higher probability of malignancy. For each of the features the mean value, worst value (mean of the three largest values), and standard error are computed for each image, resulting in 30 features of 568 images.

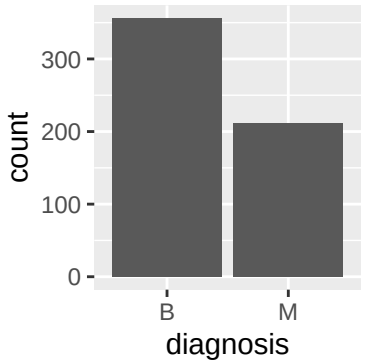
Data Processing

The original dataset contains a blank column '. . . 33,' so we dropped it. We also dropped the 'id' column, and rename several columns that contains blank space in their names.

In order to fit SVM on the data, we encode the `diagnosis` variable into a factor variable with level 1 and -1:

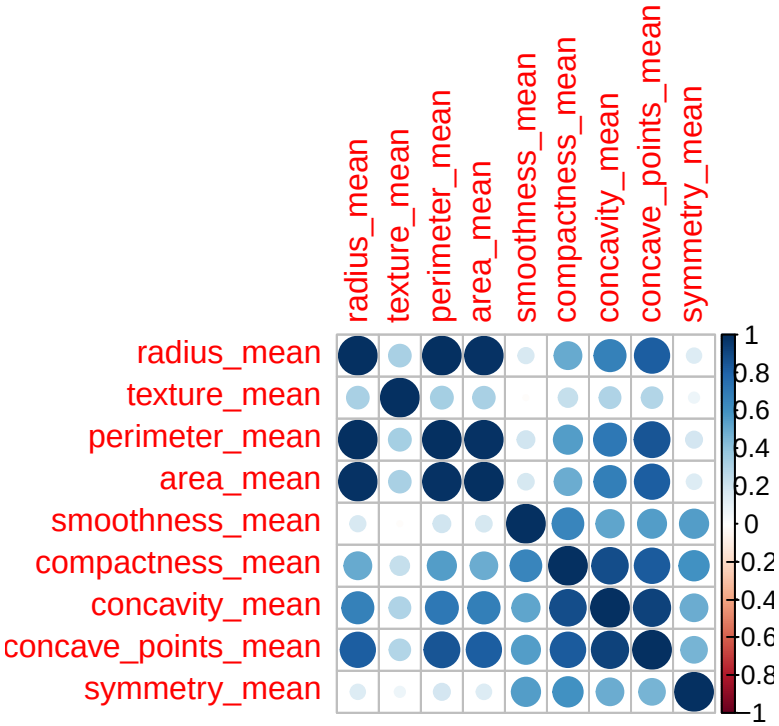
We partition the data into training and testing sets using a 70-30 percentage split(70% of the original data as the training set, and 30% as the testing set):

EDA



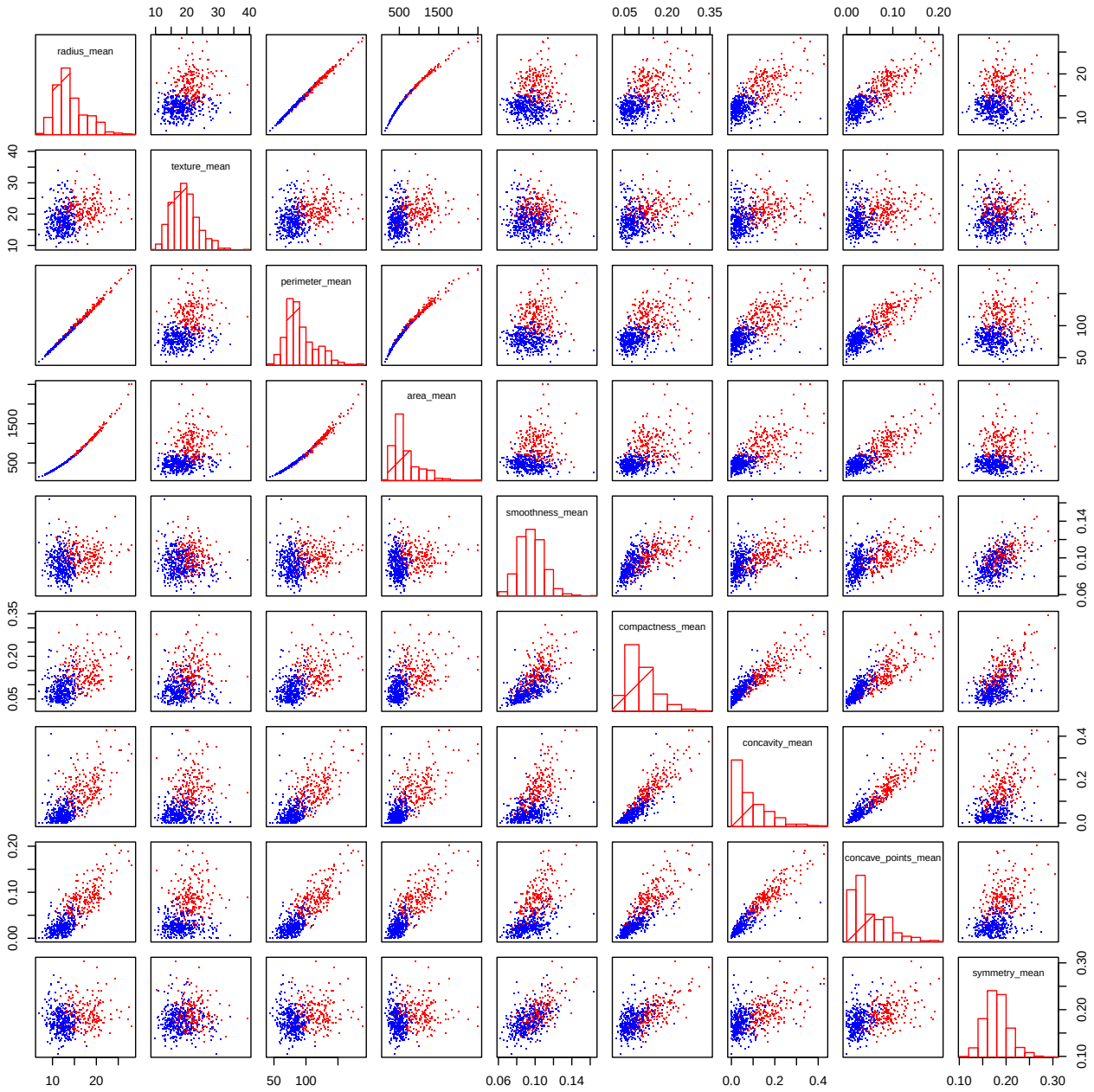
The bar plot shows that there is a larger number of benign than malignant cancer.

We divide the data into 3 categories according to their features.



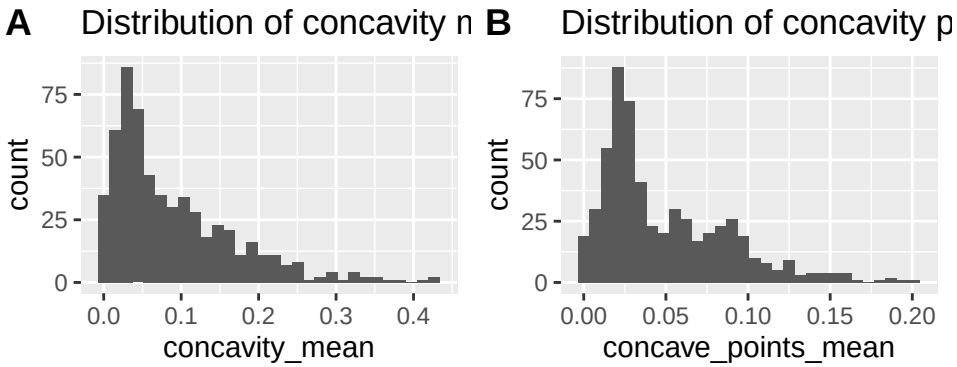
Major observations:

- Radius_mean, perimeter_mean, and area_mean are highly correlated.
- Compactness_mean, concavity_mean and concave_points_mean are highly correlated.

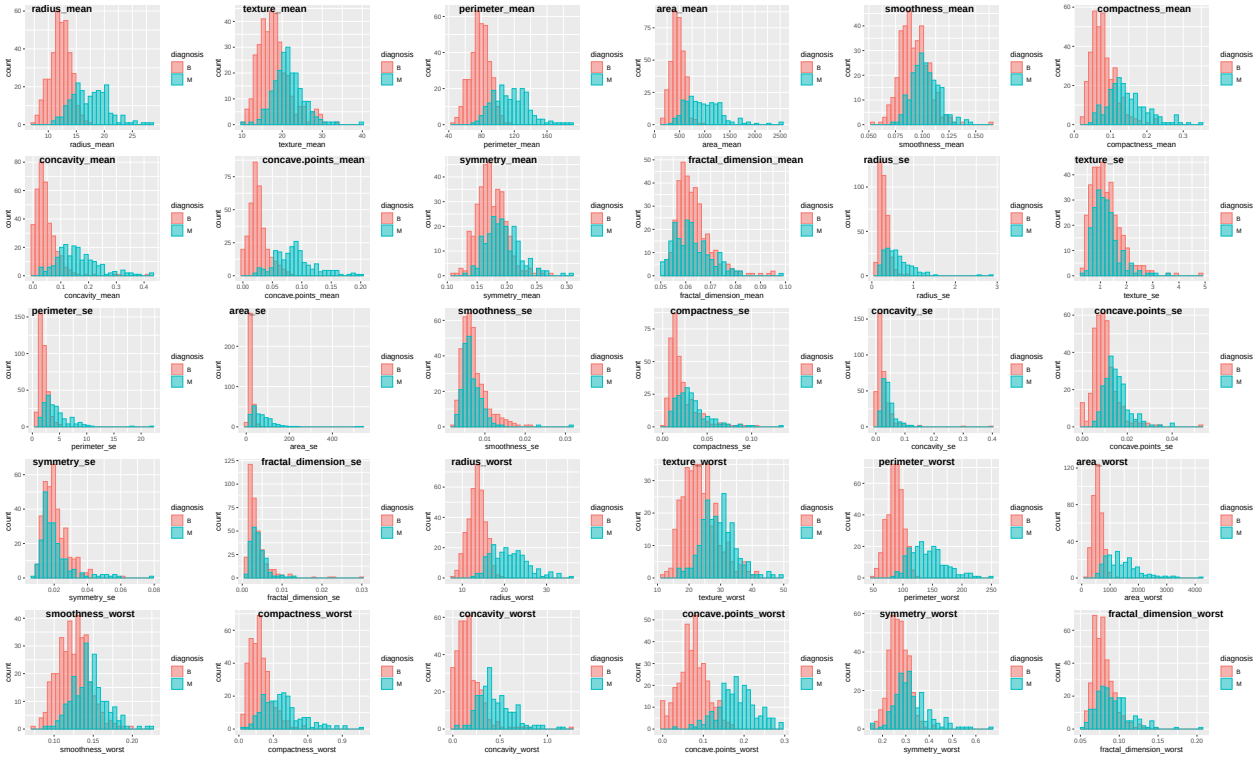


We observe from the pairwise scatterplot matrix above that the two classifications seem to be generally separable, with distinct regions in the visualization that cleanly cluster without much mingling or mixing. Overall across malignant and benign tumors, there seems to be a strong positive linear relationship between `radius_mean` and `parameter_mean`, `radius_mean` and `area_mean`, as well as `area_mean` and `parameter_mean`, which hints again at the collinearity issue which we will later tackle at through variable selection. While the two classifications together constitute a roughly linear relationship between predictors, malignant tumors (red) generally associate with higher values in both predictors accumulating in the right top corner, while benign tumors (blue) generally associate with lower ones in the left bottom corner.

The distributions of the predictors seem to be all unimodal, with no apparant outliers and generally right-skewed, with `concavity_mean` and `concave_points_mean` being particularly right-skewed, hinting at the high correlation between the two predictors. Hence we want to consider including only one of them in our model. We take a closer look at the distributions of these two predictors here:



After deriving the histogram comparing distributions of predictors based on the two classifications, we would like to find features with little overlap between benign and malignant classes which will likely to be significant for diagnosis.



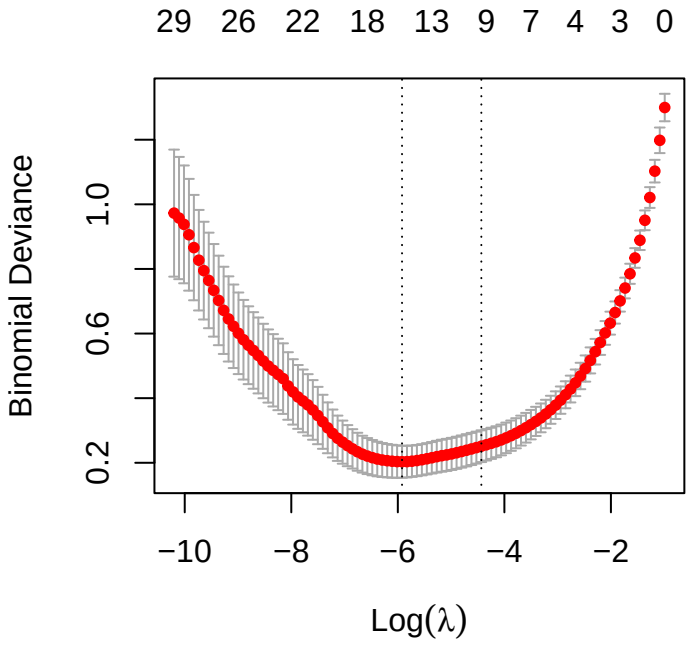
Methodology

Model 1: LASSO-Penalized Logistic Regression

We decided to use a LASSO-penalized logistic regression model to perform variable selection by gauging insights into which predictors are the most contributive, since less significant variables are forced to be exactly zero, and the most significant variables are kept in the final model.

Model Selection

We fitted the LASSO-penalized logistic regression model using the optimal hyperparameter $\lambda = 0.0026830$ via cross validation. To explore the interaction between symmetry and the mean for number of concave portions of the contour, we included the interaction term `symmetry_worst*concave_points_mean`.




```
## [1] 0.002682998
```

	coefficient
(Intercept)	-28.4748
texture_mean	0.0580
concave_points_mean	18.2162
radius_se	4.7933
texture_se	-0.5414
area_se	0.0449
smoothness_se	91.6784
compactness_se	-43.3962
fractal_dimension_se	-82.2476
radius_worst	0.6011
texture_worst	0.2279
smoothness_worst	17.8687
concavity_worst	4.4033
concave_points_worst	20.9383
symmetry_worst	6.5449
concave_points_mean:symmetry_worst	28.5510

Model

WRITE MODEL HERE

Prediction

Using the logistic regression model, besides classification we also want to understand uncertainty - more specifically, predictive probabilities that a tumor is benigh or malignant given the values of the predictors:

texture_mean	concave_points_mean	radius_se	texture_se	area_se	smoothness_se	compactness_se	fractal_dimension_se	radius_worst	texture_worst	smoothness_worst	concavity_worst	concave_points_worst	symmetry_worst	probabilities	predicted_class
10.38	0.147	1.095	0.905	153.40	0.006	0.049	0.006	25.38	17.33	0.162	0.712	0.265	0.460	1.000	M
17.77	0.070	0.543	0.734	74.08	0.005	0.013	0.004	24.99	23.41	0.124	0.242	0.186	0.275	1.000	M

texture_mean	concave_points_mean	radius_se	texture_se	area_se	smoothness_se	compactness_se	fractal_dimension_se	radius_worst	texture_worst	smoothness_worst	concavity_worst	concave_points_worst	symmetry_worst	probabilities	predicted_class
21.82	0.094	0.306	1.002	24.32	0.006	0.035	0.004	15.49	30.73	0.170	0.539	0.206	0.438	0.995	M
22.61	0.080	0.212	1.169	19.21	0.006	0.059	0.008	15.03	32.01	0.165	0.694	0.221	0.360	0.954	M
20.13	0.053	0.473	1.240	45.40	0.006	0.012	0.002	19.07	30.88	0.146	0.291	0.161	0.303	0.995	M
20.68	0.103	0.569	1.073	54.18	0.007	0.025	0.004	20.96	31.48	0.179	0.478	0.207	0.371	1.000	M
22.15	0.095	0.758	1.017	112.40	0.006	0.019	0.002	27.32	30.88	0.151	0.537	0.239	0.277	1.000	M
15.71	0.031	0.185	0.748	14.67	0.004	0.019	0.002	14.50	20.49	0.131	0.189	0.073	0.318	0.002	B
20.25	0.077	0.853	1.849	93.54	0.011	0.027	0.004	21.31	27.26	0.134	0.345	0.149	0.234	1.000	M
18.70	0.052	0.482	1.030	41.00	0.006	0.034	0.006	16.82	28.12	0.164	0.696	0.155	0.476	0.995	M

```
## [1] 0.9883041
```

We achieved a prediction accuracy of 0.9883. To interpret the predictions, we see that a patient with `texture_mean` of 10.38, `concave_points_mean` of 0.147, `radius_se` of 1.095, `texture_se` of 0.905, `area_se` of 153.400, `smoothness_se` of 0.006, `compactness_se` of 0.049, `fractal_dimension_se` of 0.006, `radius_worst` of 25.380, `texture_worst` of 17.33, `smoothness_worst` of 0.162, `concavity_worst` of 0.712, `concave_points_worst` of 0.265, `symmetry_worst` of 0.460 is expected to have a 100% of being diagnosed as malignant tumor. Whereas... is expected to have a 0% of being diagnosed as malignant tumor. Whereas... is expected to have a 87.4% of being diagnosed as malignant tumor.

Model 2: SVM

Linear Kernel SVM

We use the predictors selected by the LASSO penalized logistic regression as predictors for the support vector machine model:

If two predictors have high correlation, we only use one of them:

```
##
## Parameter tuning of 'svm':
##
## - sampling method: 10-fold cross validation
##
## - best parameters:
##  cost
##   0.1
##
## - best performance: 0.02275641
##
## - Detailed performance results:
##    cost      error dispersion
## 1 1e-03 0.09301282 0.03304092
## 2 1e-02 0.05038462 0.02042181
## 3 1e-01 0.02275641 0.02208879
## 4 1e+00 0.03275641 0.02065577
## 5 5e+00 0.04025641 0.02412695
## 6 1e+01 0.04025641 0.02412695
## 7 1e+02 0.04032051 0.02117315
```

```
##
## Call:
## best.tune(METHOD = svm, train.x = diagnosis_binary ~ texture_mean +
##     concave_points_mean + radius_se + texture_se + area_se + smoothness_se +
##     compactness_se + fractal_dimension_se + radius_worst + texture_worst +
##     smoothness_worst + concavity_worst + concave_points_worst + symmetry_worst +
```

```
##     concave_points_mean * symmetry_worst, data = cancer_train, ranges = list(cost = c(0.001,
##     0.01, 0.1, 1, 5, 10, 100)), kernel = "linear")
##
##
## Parameters:
##    SVM-Type:  C-classification
##  SVM-Kernel:  linear
##        cost:  0.1
##
## Number of Support Vectors:  55
##
##  ( 28 27 )
##
##
## Number of Classes:  2
##
## Levels:
##  -1 1
```

	-1	1
-1	98	5
1	0	68

```
## [1] 0.02923977
```

The misclassification rate is 0.02923.

Radial Kernel SVM

```
##
## Parameter tuning of 'svm':
##
## - sampling method: 10-fold cross validation
##
## - best parameters:
##  cost gamma
##     1   0.5
##
## - best performance: 0.04794872
##
## - Detailed performance results:
##     cost gamma      error dispersion
## 1  1e-01   0.5 0.35019231 0.06877558
## 2  1e+00   0.5 0.04794872 0.03456232
## 3  1e+01   0.5 0.05044872 0.03147495
## 4  1e+02   0.5 0.05044872 0.03147495
## 5  1e+03   0.5 0.05044872 0.03147495
## 6  1e-01   1.0 0.35019231 0.06877558
## 7  1e+00   1.0 0.24192308 0.08101092
## 8  1e+01   1.0 0.20929487 0.09689376
## 9  1e+02   1.0 0.20929487 0.09689376
## 10 1e+03   1.0 0.20929487 0.09689376
## 11 1e-01   2.0 0.35019231 0.06877558
## 12 1e+00   2.0 0.34762821 0.06673515
```

```
## 13 1e+01   2.0 0.34012821 0.05872625
## 14 1e+02   2.0 0.34012821 0.05872625
## 15 1e+03   2.0 0.34012821 0.05872625
## 16 1e-01   3.0 0.35019231 0.06877558
## 17 1e+00   3.0 0.35019231 0.06877558
## 18 1e+01   3.0 0.35019231 0.06877558
## 19 1e+02   3.0 0.35019231 0.06877558
## 20 1e+03   3.0 0.35019231 0.06877558
## 21 1e-01   4.0 0.35019231 0.06877558
## 22 1e+00   4.0 0.35019231 0.06877558
## 23 1e+01   4.0 0.35019231 0.06877558
## 24 1e+02   4.0 0.35019231 0.06877558
## 25 1e+03   4.0 0.35019231 0.06877558
```

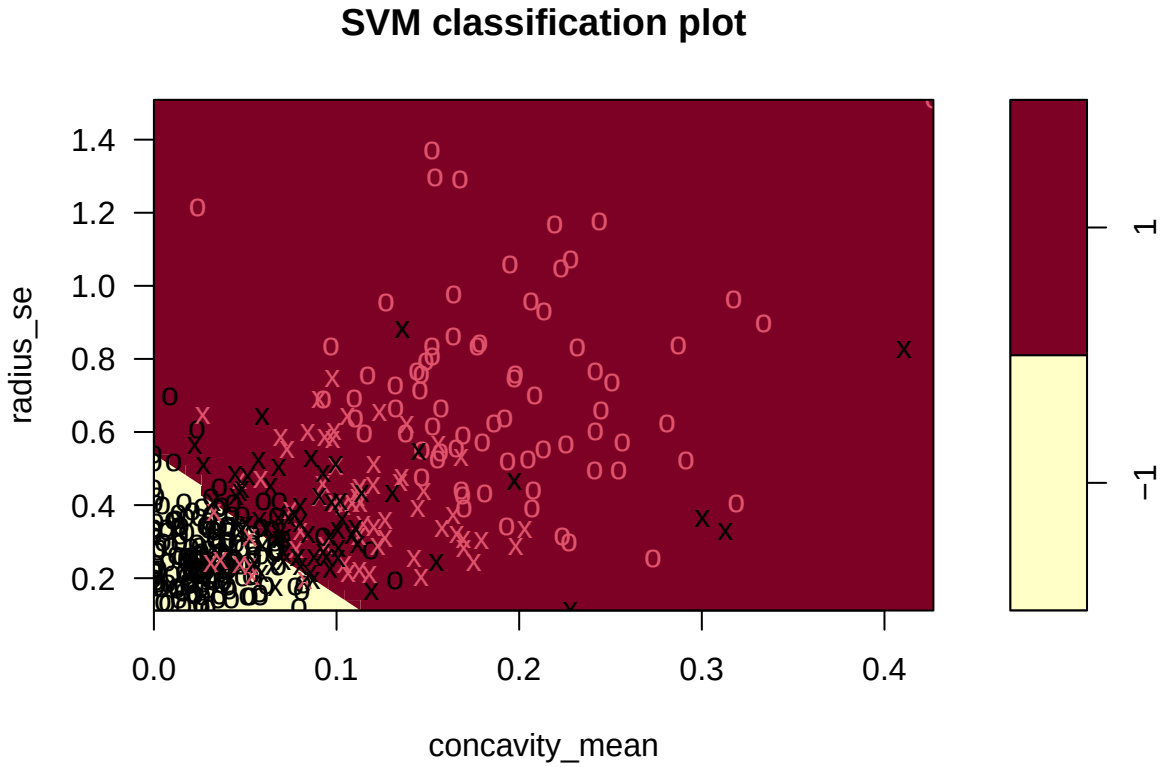
```
##
## Call:
## best.tune(METHOD = svm, train.x = diagnosis_binary ~ texture_mean +
##     concave_points_mean + radius_se + texture_se + area_se + smoothness_se +
##     compactness_se + fractal_dimension_se + radius_worst + texture_worst +
##     smoothness_worst + concavity_worst + concave_points_worst + symmetry_worst +
##     concave_points_mean * symmetry_worst, data = cancer_train, ranges = list(cost = c(0.1,
##     1, 10, 100, 1000), gamma = c(0.5, 1, 2, 3, 4)), kernel = "radial")
##
##
## Parameters:
##    SVM-Type:  C-classification
##  SVM-Kernel:  radial
##        cost:  1
##
## Number of Support Vectors:  290
##
##  ( 130 160 )
##
##
## Number of Classes:  2
##
## Levels:
##  -1 1
```

```
##        truth
## predict -1  1
##      -1 94  2
##      1   4 71
```

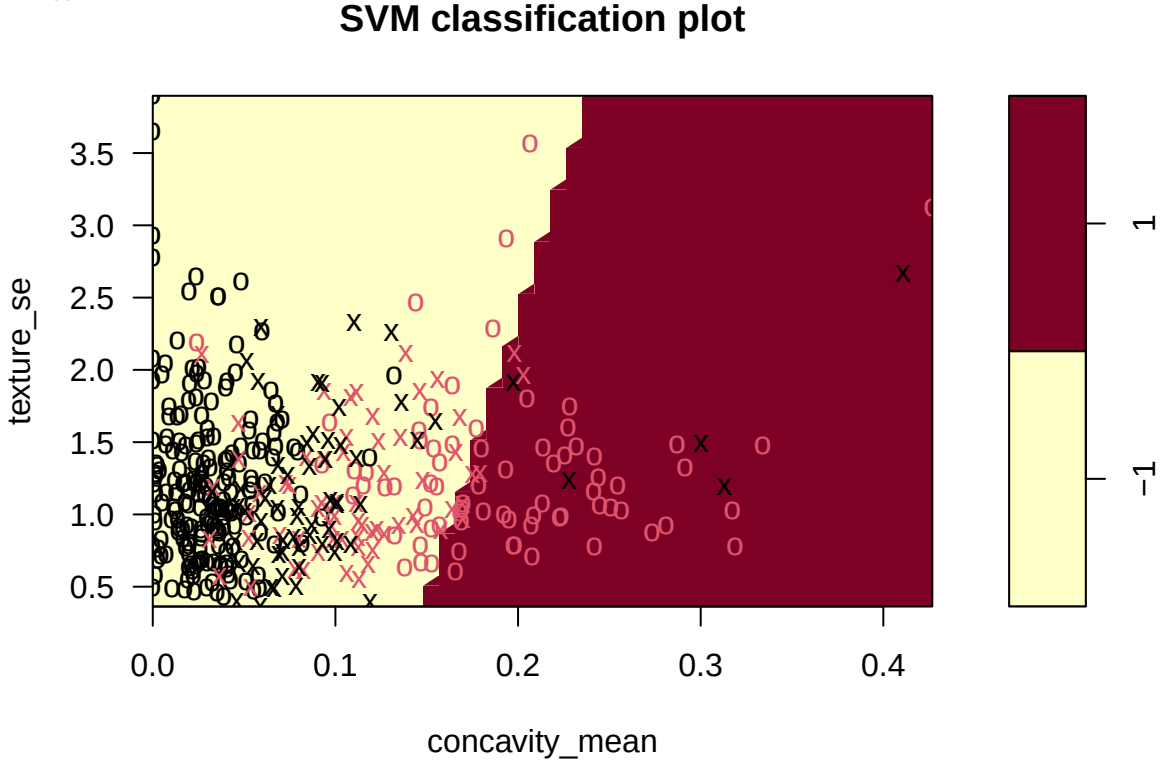
```
## [1] 0.03508772
```

The misclassification rate is 0.03509, which is greater than that of the linear kernel which suggests that the two classes are likely to be linearly separable so that we can find a separating hyperplane using the linear kernel.

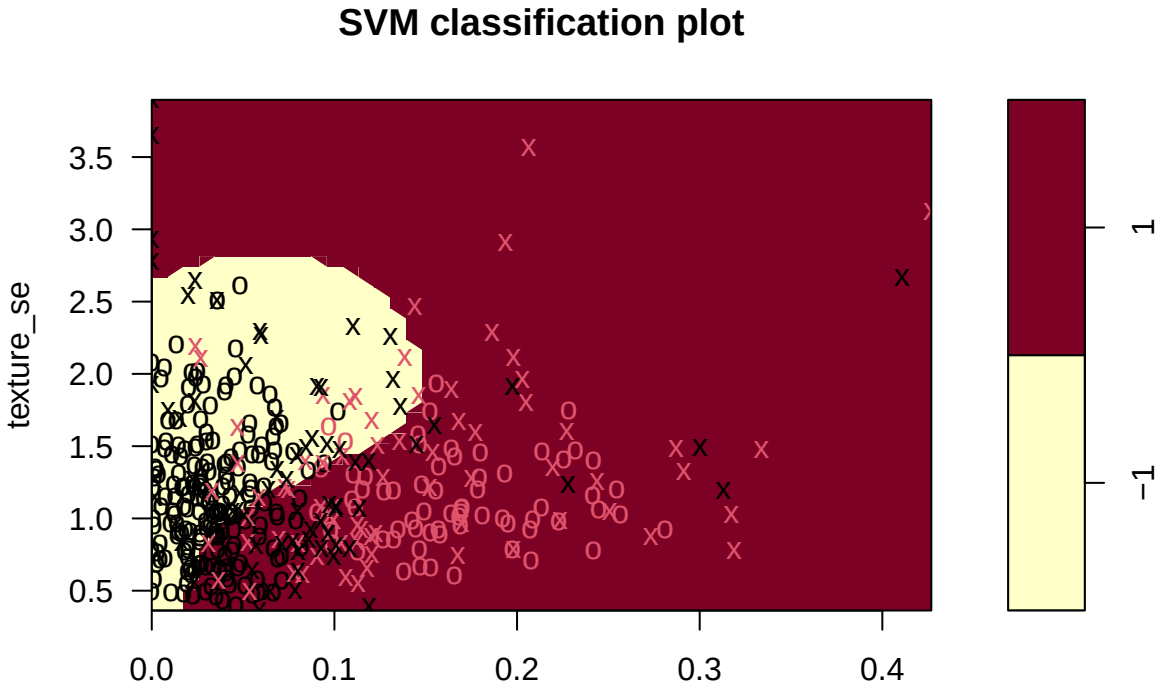
SVM Visualization

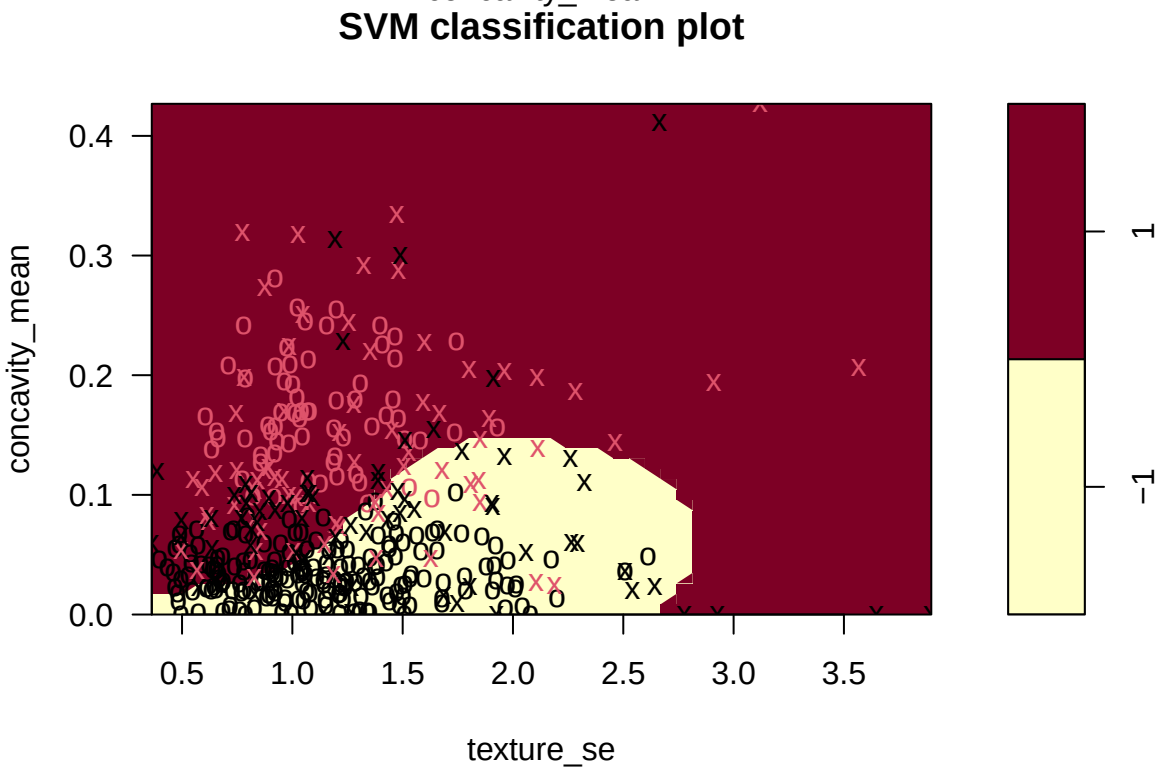


Linear



Radial

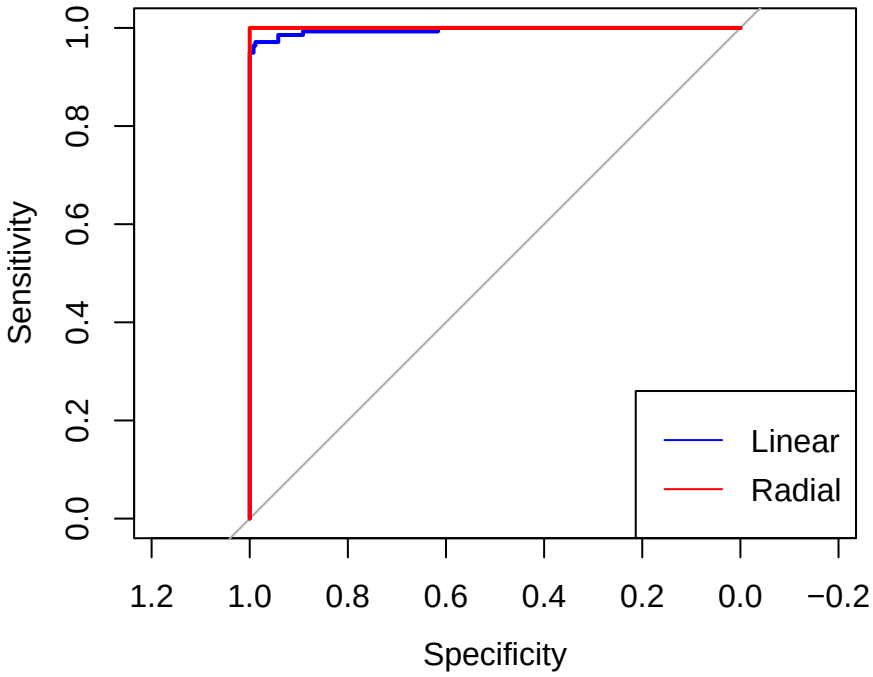




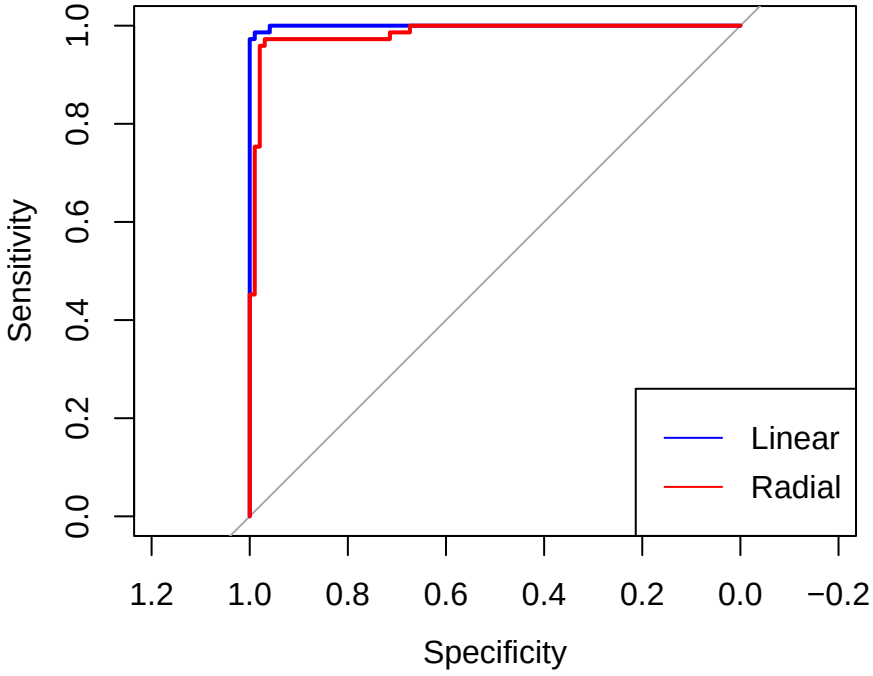
ROC (Linear SVM)

We visualize the ROC curves:

On train data:



On test data:



Even though the radial kernel fits the training data more closely due to its higher complexity, the linear kernel performs better on the test data (since the data is likely to be linearly separable as explained above), we decided to select the model with the linear kernel.

Model 3: Random Forest

For Random Forest, we decided to use group 'Mean' and group 'Worst' separately. This is because, we want to understand the potential difference in prediction given the different level of severity of the patients' conditions. As we have mentioned in the introduction, 'Worst' measures mean of the three largest values. By

building a model with only 'Mean' and 'Worst' variables, we think we can better understand the extreme cases.

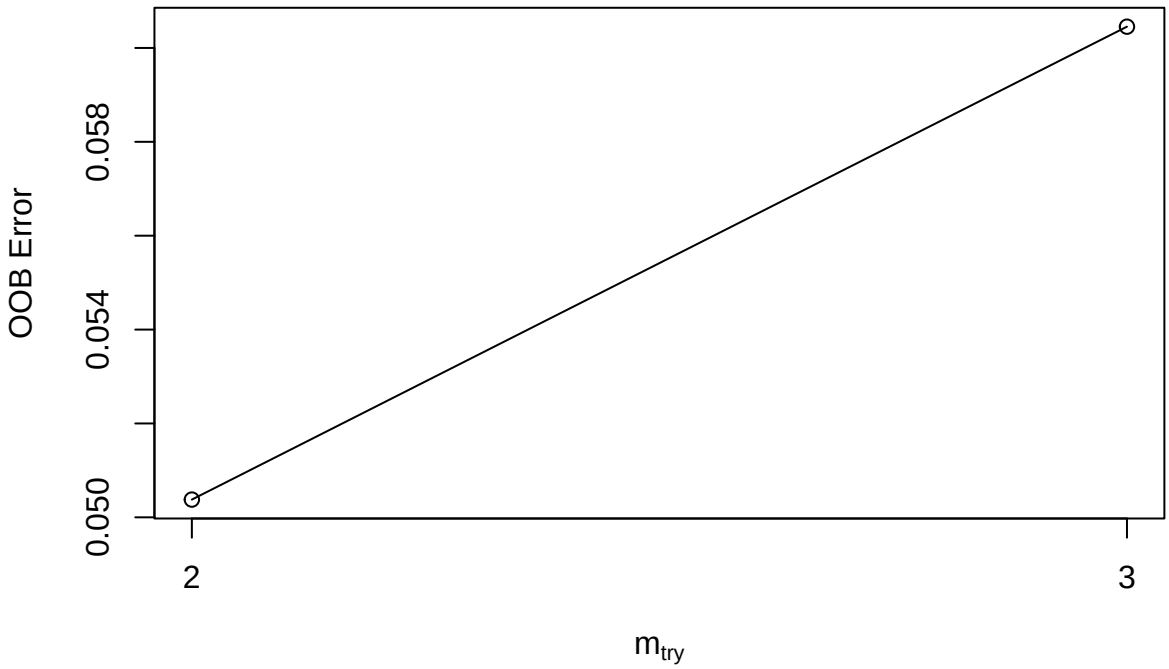
First, We select variables by running a preliminary random forest model with all the variables, to rank their importance.

```
##                          MeanDecreaseGini
## radius_mean                     8.5616407
## texture_mean                    3.0559818
## perimeter_mean                  5.9008019
## area_mean                       9.8692050
## smoothness_mean                 1.1478325
## compactness_mean                1.6773967
## concavity_mean                 10.1532690
## concave_points_mean            18.7904551
## symmetry_mean                   0.6545792
## fractal_dimension_mean          0.7849650
## radius_se                       2.4824456
## texture_se                      0.8773616
## perimeter_se                    2.8646560
## area_se                         6.6899822
## smoothness_se                   1.1900223
## compactness_se                  0.9471109
## concavity_se                    1.1550698
## concave_points_se               1.0486457
## symmetry_se                     0.8049179
## fractal_dimension_se            1.1728247
## radius_worst                   18.0287062
## texture_worst                   3.6374436
## perimeter_worst                20.0049651
## area_worst                     17.7615481
## smoothness_worst                2.9276817
## compactness_worst               3.0046846
## concavity_worst                 6.6002452
## concave_points_worst           25.2111650
## symmetry_worst                  1.8577452
## fractal_dimension_worst         1.4519769
```

After testing different variables based on their important, and taking into accounts the collinearity issue we discussed in the EDA section, we decided to select the following variables as the predictors: concave_points_worst, area_worst, perimeter_worst, radius_worst, concave_points_mean, perimeter_mean, concavity_worst, area_se.

We chose mtry=2, because after tuning mtry, we found that mtry=2 has the lowest OOB error.

```
## mtry = 2  OOB error = 5.04%
## Searching left ...
## Searching right ...
## mtry = 3     OOB error = 6.05%
## -0.2 0.01
```



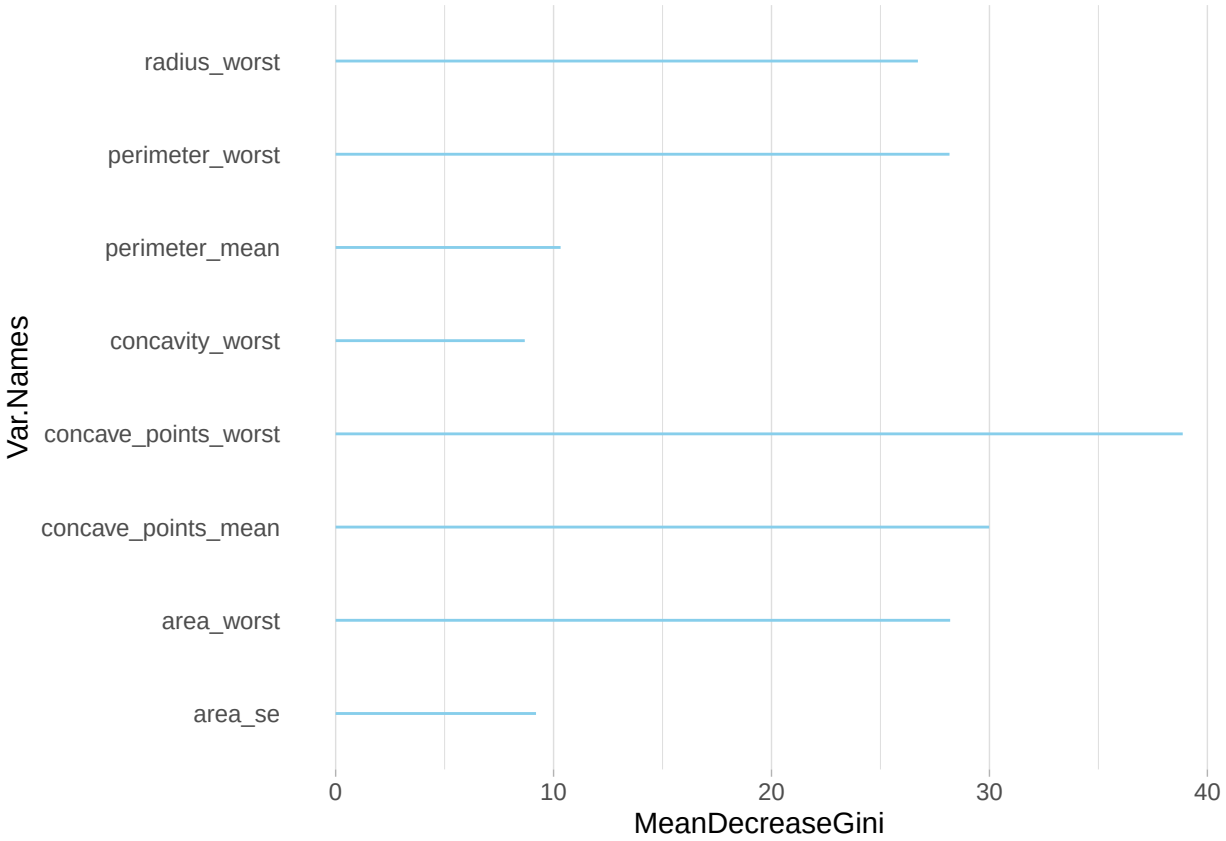
We chose the number of tree to be 500. The number of trees should be chosen carefully, since a high performance of the individual models might lead to overfitting when the number of trees is very high. However, taking 50 trees caused a lower accuracy than taking 500 trees. Therefore, this higher number of trees is chosen.

```
##
## Call:
##  randomForest(formula = diagnosis_binary ~ concave_points_worst +      area_worst + perimeter_worst +
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 2
##
##         OOB estimate of  error rate: 5.04%
## Confusion matrix:
##     -1   1 class.error
## -1 251   7  0.02713178
## 1   13 126  0.09352518
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction -1  1
##         -1 96  5
##         1   2 68
##
##                Accuracy : 0.9591
##                  95% CI : (0.9175, 0.9834)
##     No Information Rate : 0.5731
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.9159
##
##  Mcnemar's Test P-Value : 0.4497
```

```
##
##             Sensitivity : 0.9796
##             Specificity : 0.9315
##          Pos Pred Value : 0.9505
##          Neg Pred Value : 0.9714
##              Prevalence : 0.5731
##          Detection Rate : 0.5614
##    Detection Prevalence : 0.5906
##       Balanced Accuracy : 0.9555
##
##        'Positive' Class : -1
##
```

The RF model yields a 96% accuracy for the testing set.



From the plot we can see that the most important predictors are: concave_points_worst, concave_points_mean, area_worst, perimeter_worst, and radius_worst. Interestingly, perimeter_worst has high gini coefficient, but perimeter_mean ranks second from the last.

Conclusion & Future Work

Concavity is the severity of concave portions of the contour. A high concavity means that the boundary of the cell nucleus has indentations, and thus is rather rough than smooth. Concave points id the number of concave portions of the contour of the cell nucleus.

Citations