# Capstone: Breast Cancer Diagnosis

*Jayashri Gopalakrishnan*

*6/7/2020*

## Introduction

### Objective

Breast cancer is the one of the most common cancers diagnosed in women in the US; second only to skin cancer. According to the American Cancer Society, approximately 1 in 8 women in the US will develop breast cancer in her lifetime. For the purpose of this project, the Breast Cancer Wisconsin (Diagnostic) Data set, created by Dr. William H. Wolberg, W. Nick Street and Olvi L. Mangasarian from the University of Wisconsin, is used for analysis. **The main objective of this project is to predict whether a breast cancer cell is benign or malignant**

Mammograms play a key role in early breast cancer detection and help decrease breast cancer related deaths. If a suspicious breast mass is detected, usually a biopsy test is carried out and masses are analyzed. A fine needle aspiration (FNA) is most often done on swellings or lumps located just under the skin. Features in this dataset are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.

A comparison of six different machine learning algorithms is performed to determine the most effective algorithm for classifying the diagnosis of cancer cells based on accuracy, precision, sensitivity, and specificity.

### Dataset

The dataset's features describe characteristics of the cell nuclei on the image. Link to the dataset: https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/wdbc.data

There are 569 observations of 32 variables. The following are the features specified below:

- Attribute Information:

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

- Ten features were computed for each cell nucleus:

1. radius: mean of distances from center to points on the perimeter
2. texture: standard deviation of grey-scale values
3. perimeter
4. area
5. smoothness: local variation in radius lengths)
6. compactness: perimeter^2 / area – 1.0
7. concavity: severity of concave portions of the contour
8. concave points: number of concave portions of the contour
9. symmetry
10. 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 32 variables (30 + 2 attributes).

# Data Analysis

## Data Cleaning and Exploration

The breastcancer dataset contains 569 observations of 32 variables.

```
dim(breastcancer)
```

```
## [1] 569  32
```

```
head(breastcancer)
```

```
##         id diagnosis radius_mean texture_mean perimeter_mean area_mean
## 1   842302         M       17.99        10.38         122.80    1001.0
## 2   842517         M       20.57        17.77         132.90    1326.0
## 3 84300903         M       19.69        21.25         130.00    1203.0
## 4 84348301         M       11.42        20.38          77.58     386.1
## 5 84358402         M       20.29        14.34         135.10    1297.0
## 6   843786         M       12.45        15.70          82.57     477.1
##   smoothness_mean compactness_mean concavity_mean concave.points_mean
## 1         0.11840          0.27760         0.3001             0.14710
## 2         0.08474          0.07864         0.0869             0.07017
## 3         0.10960          0.15990         0.1974             0.12790
## 4         0.14250          0.28390         0.2414             0.10520
## 5         0.10030          0.13280         0.1980             0.10430
## 6         0.12780          0.17000         0.1578             0.08089
##   symmetry_mean fractal_dimension_mean radius_se texture_se perimeter_se
## 1        0.2419                0.07871    1.0950     0.9053        8.589
## 2        0.1812                0.05667    0.5435     0.7339        3.398
## 3        0.2069                0.05999    0.7456     0.7869        4.585
## 4        0.2597                0.09744    0.4956     1.1560        3.445
## 5        0.1809                0.05883    0.7572     0.7813        5.438
## 6        0.2087                0.07613    0.3345     0.8902        2.217
##   area_se smoothness_se compactness_se concavity_se concave.points_se
## 1  153.40      0.006399        0.04904      0.05373           0.01587
## 2   74.08      0.005225        0.01308      0.01860           0.01340
## 3   94.03      0.006150        0.04006      0.03832           0.02058
## 4   27.23      0.009110        0.07458      0.05661           0.01867
## 5   94.44      0.011490        0.02461      0.05688           0.01885
## 6   27.19      0.007510        0.03345      0.03672           0.01137
##   symmetry_se fractal_dimension_se radius_worst texture_worst perimeter_worst
## 1     0.03003             0.006193        25.38         17.33          184.60
## 2     0.01389             0.003532        24.99         23.41          158.80
## 3     0.02250             0.004571        23.57         25.53          152.50
## 4     0.05963             0.009208        14.91         26.50           98.87
## 5     0.01756             0.005115        22.54         16.67          152.20
## 6     0.02165             0.005082        15.47         23.75          103.40
##   area_worst smoothness_worst compactness_worst concavity_worst
## 1     2019.0           0.1622            0.6656          0.7119
## 2     1956.0           0.1238            0.1866          0.2416
## 3     1709.0           0.1444            0.4245          0.4504
## 4      567.7           0.2098            0.8663          0.6869
```

```
## 5     1575.0           0.1374            0.2050          0.4000
## 6      741.6           0.1791            0.5249          0.5355
##   concave.points_worst symmetry_worst fractal_dimension_worst
## 1               0.2654         0.4601                 0.11890
## 2               0.1860         0.2750                 0.08902
## 3               0.2430         0.3613                 0.08758
## 4               0.2575         0.6638                 0.17300
## 5               0.1625         0.2364                 0.07678
## 6               0.1741         0.3985                 0.12440
```

```
summary(breastcancer)
```

```
##        id            diagnosis  radius_mean      texture_mean
##  Min.   :     8670   B:357     Min.   : 6.981   Min.   : 9.71
##  1st Qu.:   869218   M:212     1st Qu.:11.700   1st Qu.:16.17
##  Median :   906024             Median :13.370   Median :18.84
##  Mean   : 30371831             Mean   :14.127   Mean   :19.29
##  3rd Qu.:  8813129             3rd Qu.:15.780   3rd Qu.:21.80
##  Max.   :911320502             Max.   :28.110   Max.   :39.28
##  perimeter_mean     area_mean      smoothness_mean   compactness_mean
##  Min.   : 43.79   Min.   : 143.5   Min.   :0.05263   Min.   :0.01938
##  1st Qu.: 75.17   1st Qu.: 420.3   1st Qu.:0.08637   1st Qu.:0.06492
##  Median : 86.24   Median : 551.1   Median :0.09587   Median :0.09263
##  Mean   : 91.97   Mean   : 654.9   Mean   :0.09636   Mean   :0.10434
##  3rd Qu.:104.10   3rd Qu.: 782.7   3rd Qu.:0.10530   3rd Qu.:0.13040
##  Max.   :188.50   Max.   :2501.0   Max.   :0.16340   Max.   :0.34540
##  concavity_mean    concave.points_mean symmetry_mean    fractal_dimension_mean
##  Min.   :0.00000   Min.   :0.00000     Min.   :0.1060   Min.   :0.04996
##  1st Qu.:0.02956   1st Qu.:0.02031     1st Qu.:0.1619   1st Qu.:0.05770
##  Median :0.06154   Median :0.03350     Median :0.1792   Median :0.06154
##  Mean   :0.08880   Mean   :0.04892     Mean   :0.1812   Mean   :0.06280
##  3rd Qu.:0.13070   3rd Qu.:0.07400     3rd Qu.:0.1957   3rd Qu.:0.06612
##  Max.   :0.42680   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     concave.points_se
##  Min.   :0.001713   Min.   :0.002252   Min.   :0.00000   Min.   :0.000000
##  1st Qu.:0.005169   1st Qu.:0.013080   1st Qu.:0.01509   1st Qu.:0.007638
##  Median :0.006380   Median :0.020450   Median :0.02589   Median :0.010930
##  Mean   :0.007041   Mean   :0.025478   Mean   :0.03189   Mean   :0.011796
##  3rd Qu.:0.008146   3rd Qu.:0.032450   3rd Qu.:0.04205   3rd Qu.:0.014710
##  Max.   :0.031130   Max.   :0.135400   Max.   :0.39600   Max.   :0.052790
##   symmetry_se       fractal_dimension_se  radius_worst   texture_worst
##  Min.   :0.007882   Min.   :0.0008948    Min.   : 7.93   Min.   :12.02
##  1st Qu.:0.015160   1st Qu.:0.0022480    1st Qu.:13.01   1st Qu.:21.08
##  Median :0.018730   Median :0.0031870    Median :14.97   Median :25.41
##  Mean   :0.020542   Mean   :0.0037949    Mean   :16.27   Mean   :25.68
##  3rd Qu.:0.023480   3rd Qu.:0.0045580    3rd Qu.:18.79   3rd Qu.:29.72
##  Max.   :0.078950   Max.   :0.0298400    Max.   :36.04   Max.   :49.54
```

```
##  perimeter_worst    area_worst     smoothness_worst  compactness_worst
##  Min.   : 50.41   Min.   : 185.2   Min.   :0.07117   Min.   :0.02729
##  1st Qu.: 84.11   1st Qu.: 515.3   1st Qu.:0.11660   1st Qu.:0.14720
##  Median : 97.66   Median : 686.5   Median :0.13130   Median :0.21190
##  Mean   :107.26   Mean   : 880.6   Mean   :0.13237   Mean   :0.25427
##  3rd Qu.:125.40   3rd Qu.:1084.0   3rd Qu.:0.14600   3rd Qu.:0.33910
##  Max.   :251.20   Max.   :4254.0   Max.   :0.22260   Max.   :1.05800
##  concavity_worst  concave.points_worst symmetry_worst   fractal_dimension_worst
##  Min.   :0.0000   Min.   :0.00000      Min.   :0.1565   Min.   :0.05504
##  1st Qu.:0.1145   1st Qu.:0.06493      1st Qu.:0.2504   1st Qu.:0.07146
##  Median :0.2267   Median :0.09993      Median :0.2822   Median :0.08004
##  Mean   :0.2722   Mean   :0.11461      Mean   :0.2901   Mean   :0.08395
##  3rd Qu.:0.3829   3rd Qu.:0.16140      3rd Qu.:0.3179   3rd Qu.:0.09208
##  Max.   :1.2520   Max.   :0.29100      Max.   :0.6638   Max.   :0.20750
```

There are no missing values in this dataset.

```
sum(is.na(breastcancer))
```

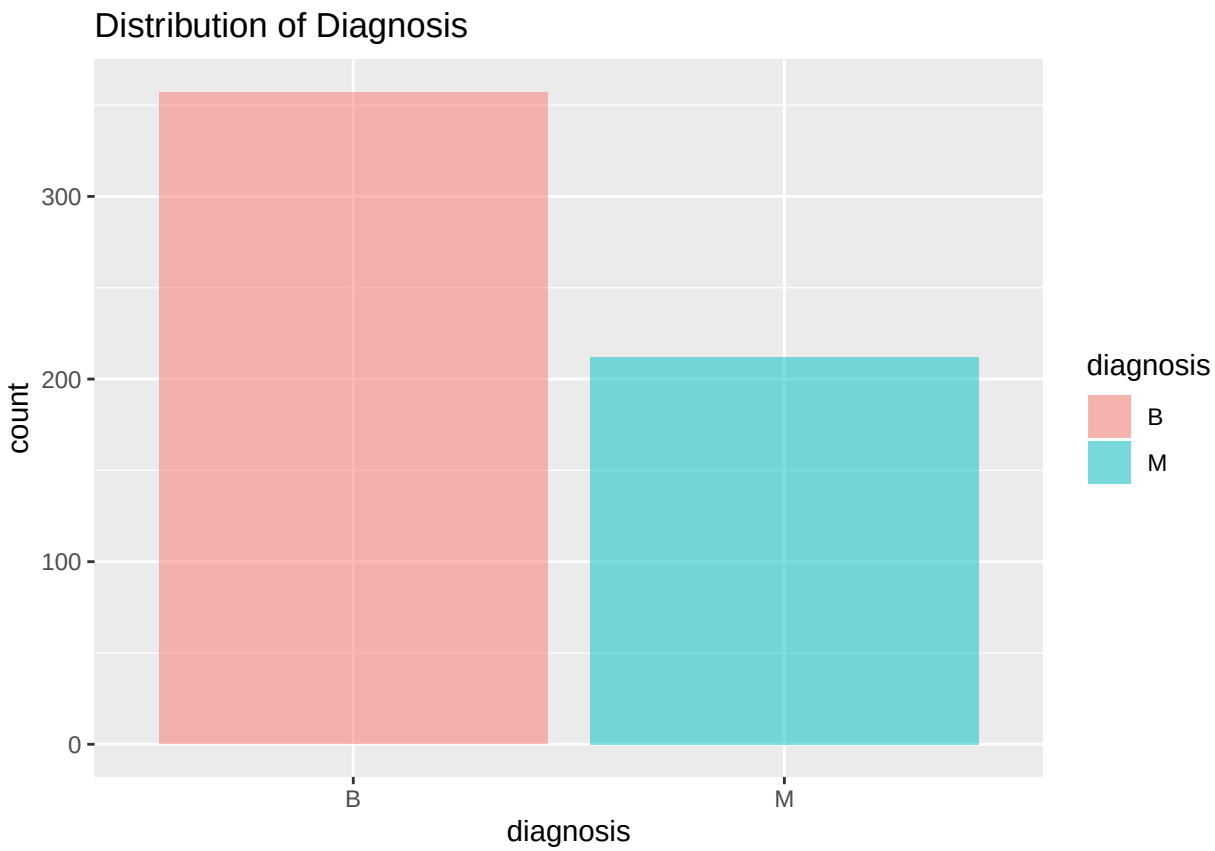
```
## [1] 0
```

### Diagnosis

The diagnosis variable is analyzed and from our previous data exploration it is found that it is a factor with two levels: "B" (benign) and "M" (malignant).

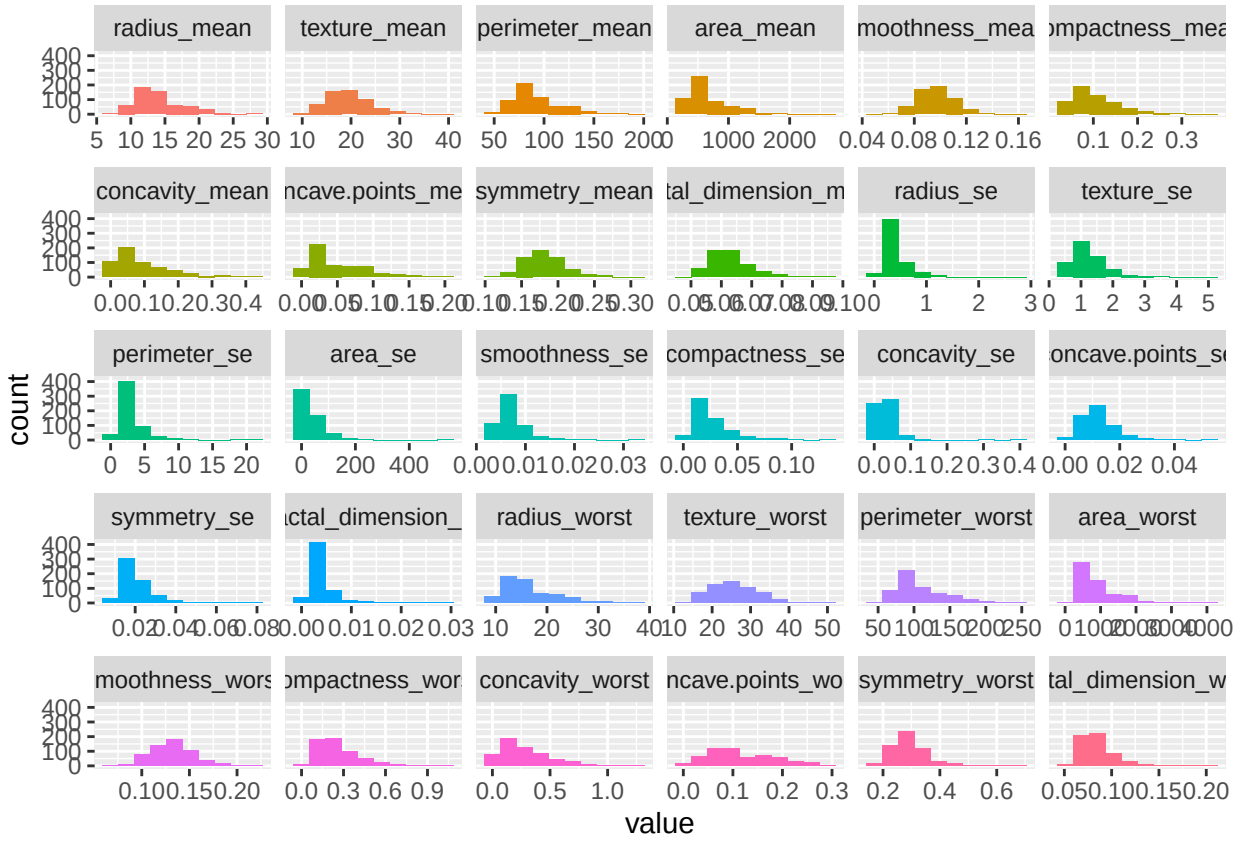
```
##
##         B         M
## 0.6274165 0.3725835
```

The above table indicates that approximately 63% of cancer cells are benign and 37% of cancer cells are malignant. The graph below displays the number of benign and malignant observations.



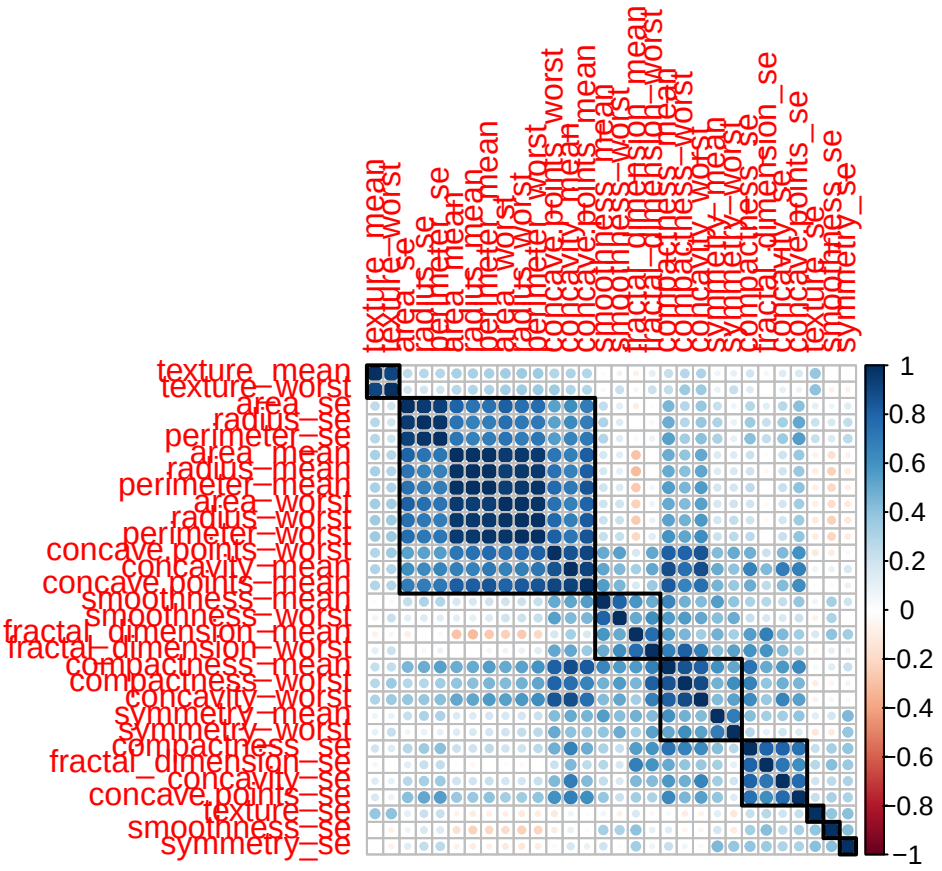

**Visualizing Other Variables**

In the distributions of the 30 features plotted below, most of the variables are normally distributed.




**Exploring Relationships**

Most ML algorithms assume that predictor variables are independent of each other. Multicollinearity can cause issues when we fit the model and interpret results. For this purpose, we check the correlations of the variables.




In the above plot, a lot of predictors are highly correlated. For improving the efficiency and performance of our machine learning algorithms, highly correlated predictors (those with correlations $> 0.90$) are removed from our dataset. Our newly transformed dataset has only *22 variables*.

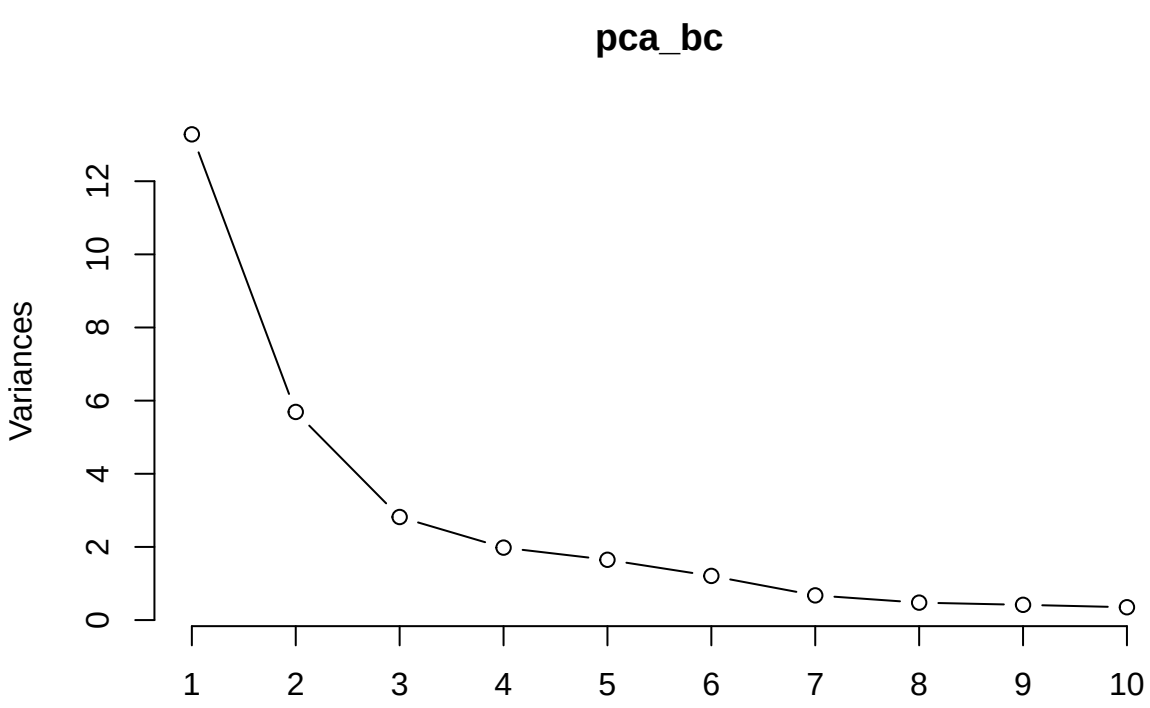
```
# Finding highly correlated variables
highCorrelation <- findCorrelation(correlationMatrix, cutoff=0.9)
# Removing correlated variables
breastcancer2 <- breastcancer %>%select(-highCorrelation)
# Number of columns after removing correlated variables
ncol(breastcancer2)
```

```
## [1] 22
```

## Data Pre-Processing

### Principal Component Analysis (PCA)

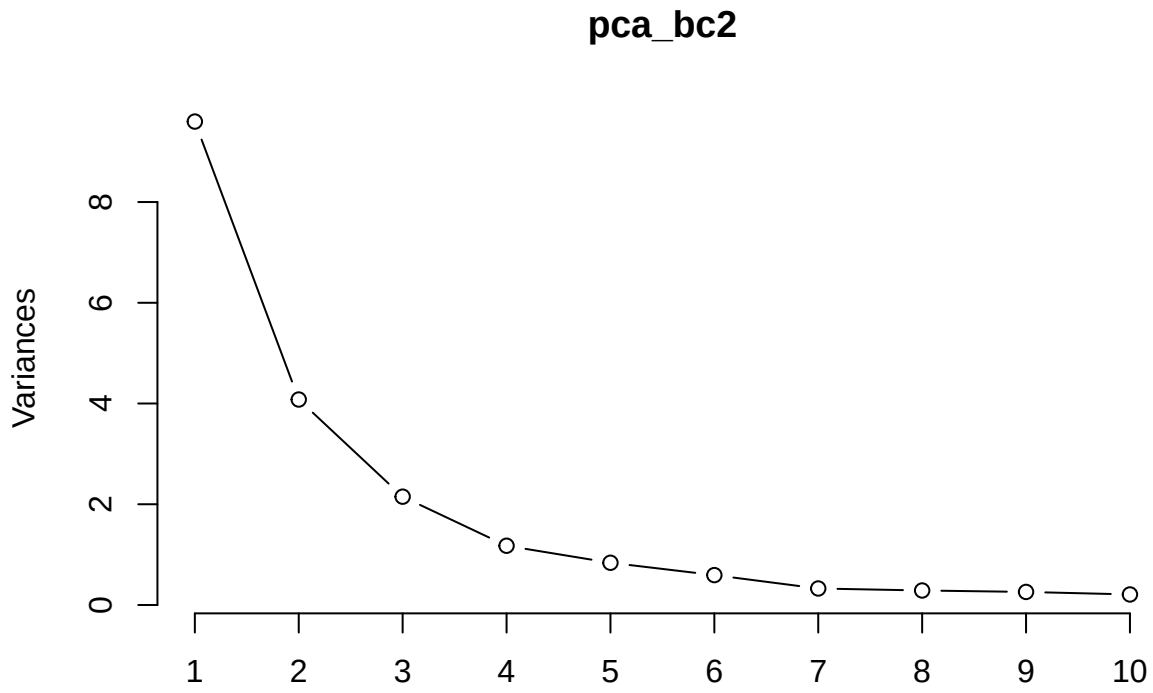
Principal Component Analysis (PCA) is a dimensionality reduction technique to reduce the number of features that are correlated to each other, while retaining as much information as possible. It is performed to avoid redundancy and avoid correlated variables that could be detrimental to clustering analysis.




```
## Importance of components:
##                            PC1    PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     3.6444 2.3857 1.67867 1.40735 1.28403 1.09880 0.82172
## Proportion of Variance 0.4427 0.1897 0.09393 0.06602 0.05496 0.04025 0.02251
## Cumulative Proportion  0.4427 0.6324 0.72636 0.79239 0.84734 0.88759 0.91010
##                            PC8    PC9    PC10   PC11    PC12    PC13    PC14
## Standard deviation     0.69037 0.6457 0.59219 0.5421 0.51104 0.49128 0.39624
## Proportion of Variance 0.01589 0.0139 0.01169 0.0098 0.00871 0.00805 0.00523
## Cumulative Proportion  0.92598 0.9399 0.95157 0.9614 0.97007 0.97812 0.98335
##                           PC15    PC16    PC17    PC18    PC19    PC20   PC21
## Standard deviation     0.30681 0.28260 0.24372 0.22939 0.22244 0.17652 0.1731
## Proportion of Variance 0.00314 0.00266 0.00198 0.00175 0.00165 0.00104 0.0010
## Cumulative Proportion  0.98649 0.98915 0.99113 0.99288 0.99453 0.99557 0.9966
##                           PC22    PC23   PC24    PC25    PC26    PC27    PC28
## Standard deviation     0.16565 0.15602 0.1344 0.12442 0.09043 0.08307 0.03987
## Proportion of Variance 0.00091 0.00081 0.0006 0.00052 0.00027 0.00023 0.00005
## Cumulative Proportion  0.99749 0.99830 0.9989 0.99942 0.99969 0.99992 0.99997
##                           PC29    PC30
## Standard deviation     0.02736 0.01153
## Proportion of Variance 0.00002 0.00000
## Cumulative Proportion  1.00000 1.00000
```

The above PCA performed on the original "breastcancer" dataset indiactes that the first two principle components (PCs) cumulatively explain 63% of the variance, while the first ten PCs explain 95% of the variance.

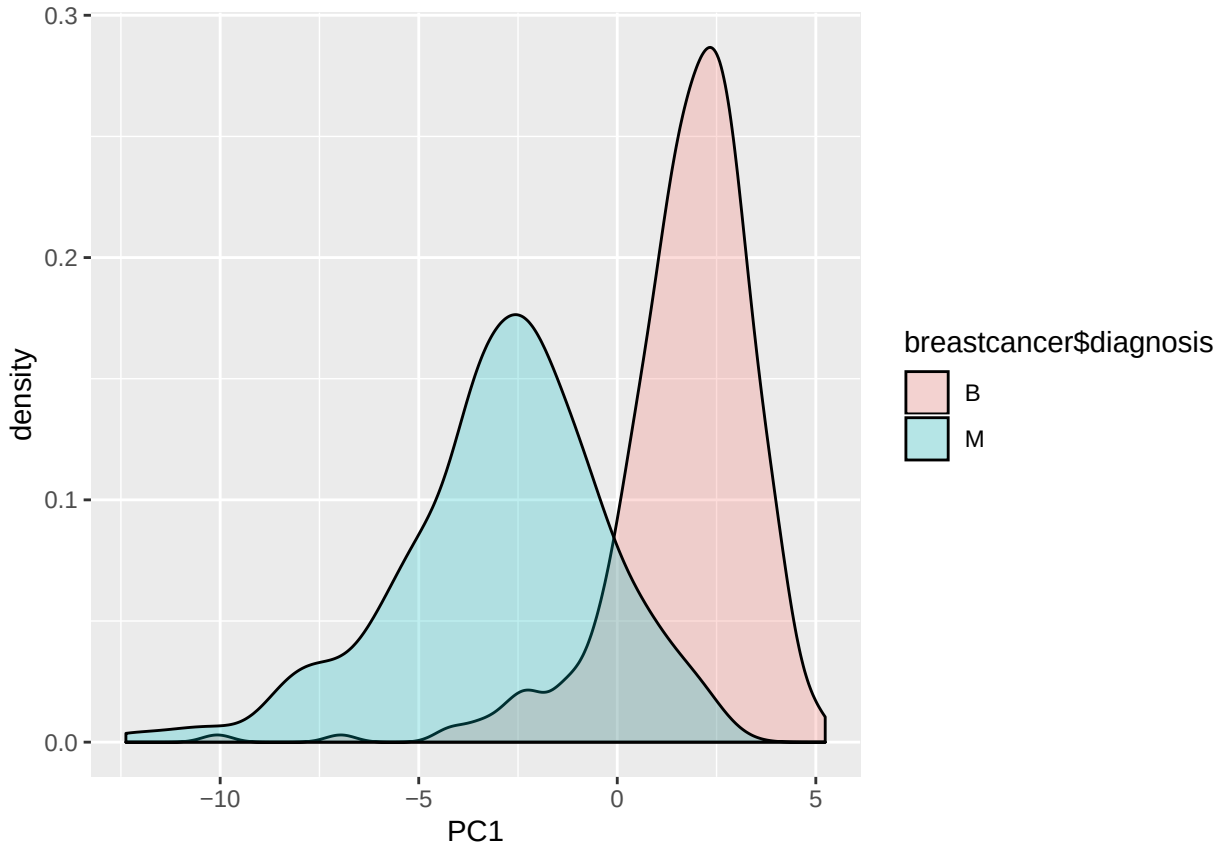
The same analysis is performed on the transformed "breastcancer2" datset which has 22 variables and the summary is presented below:

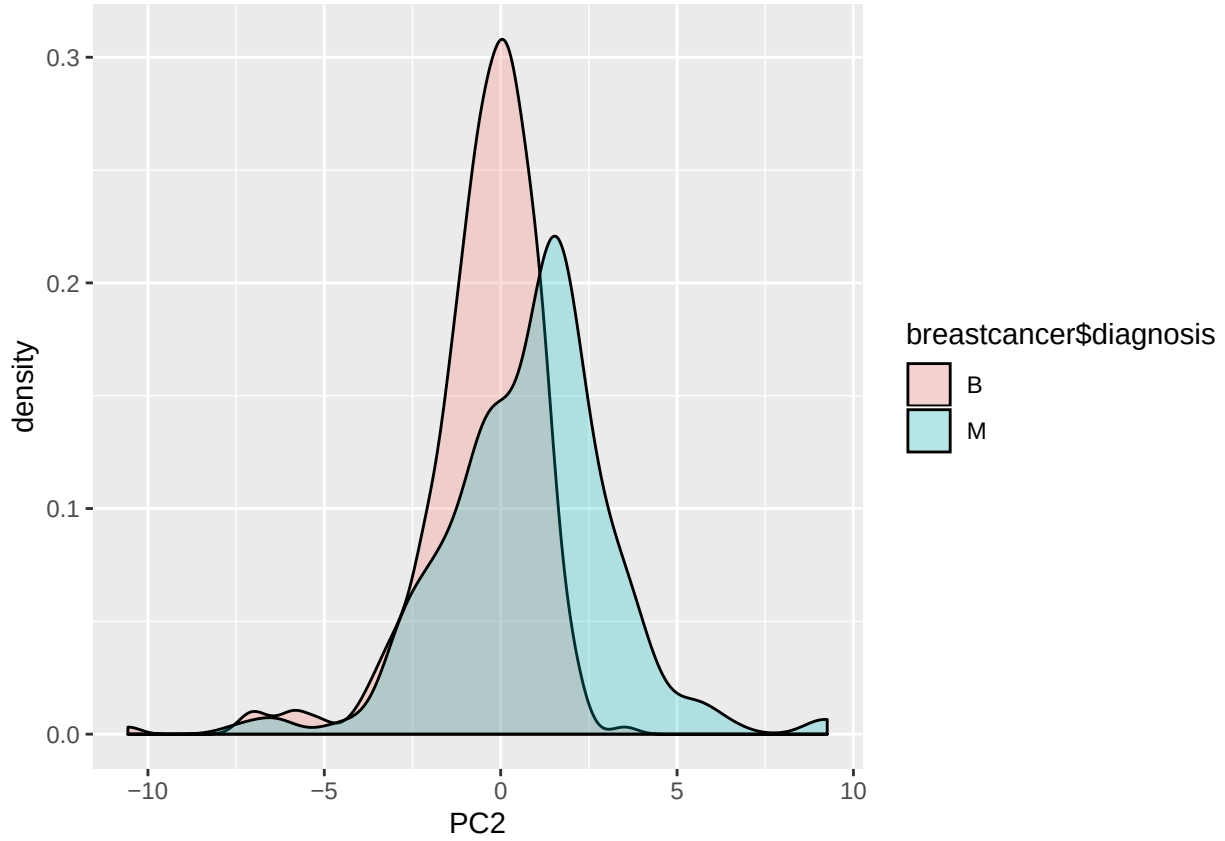



```
## Importance of components:
##                            PC1    PC2    PC3    PC4     PC5     PC6     PC7
## Standard deviation     3.0980 2.0196 1.4663 1.0845 0.91561 0.77019 0.57227
## Proportion of Variance 0.4799 0.2039 0.1075 0.0588 0.04192 0.02966 0.01637
## Cumulative Proportion  0.4799 0.6838 0.7913 0.8501 0.89205 0.92171 0.93808
##                            PC8     PC9    PC10    PC11    PC12    PC13    PC14
## Standard deviation     0.53641 0.50898 0.45726 0.36641 0.31778 0.28802 0.21369
## Proportion of Variance 0.01439 0.01295 0.01045 0.00671 0.00505 0.00415 0.00228
## Cumulative Proportion  0.95247 0.96542 0.97588 0.98259 0.98764 0.99179 0.99407
##                           PC15    PC16    PC17    PC18    PC19    PC20
## Standard deviation     0.1846 0.15579 0.15393 0.14782 0.09636 0.07375
## Proportion of Variance 0.0017 0.00121 0.00118 0.00109 0.00046 0.00027
## Cumulative Proportion  0.9958 0.99699 0.99817 0.99926 0.99973 1.00000
```

In "breastcancer2", 95% of the variance is explained by the first eight PCs.

When visualizing the principle componenets, PC1 and PC2, we find that PCs can be easily separated into two classes as the variance explained by these components is not large.


0.3
0.2
0.1
0.0
density
−10
−5
0
5
PC1
breastcancer$diagnosis
B
M

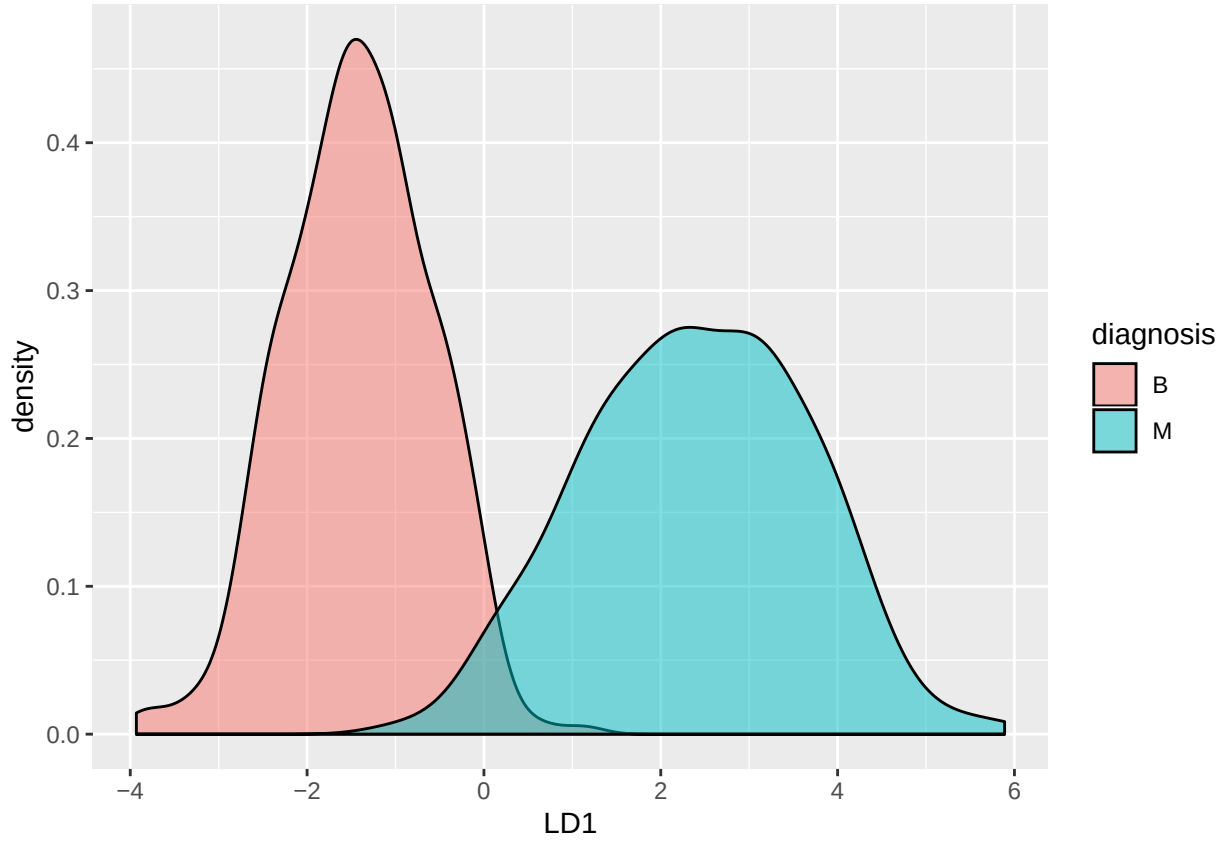



**Linear Discriminant Analysis (LDA)**

Linear Discriminant Analysis is another dimensionality reduction technique and is different from PCA since it takes into consideration different classes. LDA also makes assumptions about normally distributed classes and equal class covariances.

```
## Call:
## lda(diagnosis ~ ., data = breastcancer, center = TRUE, scale = TRUE)
##
## Prior probabilities of groups:
##         B         M
## 0.6274165 0.3725835
##
## Group means:
##         id radius_mean texture_mean perimeter_mean area_mean smoothness_mean
## B 26543825    12.14652     17.91476       78.07541  462.7902      0.09247765
## M 36818050    17.46283     21.60491      115.36538  978.3764      0.10289849
##   compactness_mean concavity_mean concave.points_mean symmetry_mean
## B       0.08008462     0.04605762          0.02571741      0.174186
## M       0.14518778     0.16077472          0.08799000      0.192909
##   fractal_dimension_mean radius_se texture_se perimeter_se  area_se
## B             0.06286739 0.2840824   1.220380     2.000321 21.13515
## M             0.06268009 0.6090825   1.210915     4.323929 72.67241
##   smoothness_se compactness_se concavity_se concave.points_se symmetry_se
## B   0.007195902     0.02143825   0.02599674       0.009857653  0.02058381
```

```
## M   0.006780094     0.03228117   0.04182401       0.015060472  0.02047240
##   fractal_dimension_se radius_worst texture_worst perimeter_worst area_worst
## B          0.003636051     13.37980      23.51507        87.00594   558.8994
## M          0.004062406     21.13481      29.31821       141.37033  1422.2863
##   smoothness_worst compactness_worst concavity_worst concave.points_worst
## B        0.1249595         0.1826725       0.1662377           0.07444434
## M        0.1448452         0.3748241       0.4506056           0.18223731
##   symmetry_worst fractal_dimension_worst
## B      0.2702459              0.07944207
## M      0.3234679              0.09152995
##
## Coefficients of linear discriminants:
##                                    LD1
## id                       -2.512117e-10
## radius_mean              -1.080876e+00
## texture_mean              2.338408e-02
## perimeter_mean            1.172707e-01
## area_mean                 1.595690e-03
## smoothness_mean           5.251575e-01
## compactness_mean         -2.094197e+01
## concavity_mean            6.955923e+00
## concave.points_mean       1.047567e+01
## symmetry_mean             4.938898e-01
## fractal_dimension_mean   -5.937663e-02
## radius_se                 2.101503e+00
## texture_se               -3.979869e-02
## perimeter_se             -1.121814e-01
## area_se                  -4.083504e-03
## smoothness_se             7.987663e+01
## compactness_se            1.387026e-01
## concavity_se             -1.768261e+01
## concave.points_se         5.350520e+01
## symmetry_se               8.143611e+00
## fractal_dimension_se     -3.431356e+01
## radius_worst              9.677207e-01
## texture_worst             3.540591e-02
## perimeter_worst          -1.204507e-02
## area_worst               -5.012127e-03
## smoothness_worst          2.612258e+00
## compactness_worst         3.636892e-01
## concavity_worst           1.880699e+00
## concave.points_worst      2.218189e+00
## symmetry_worst            2.783102e+00
## fractal_dimension_worst   2.117830e+01
```




## Data Modeling

### Data Partition

To build the Machine Learning algorithm, the data is split in 80:20 ratio into training and test sets. The transformed dataset is split into train_data (80%) and test_data (20%) to build models to predict the diagnosis (i.e. benign or malignant). Since the number of observations in the dataset is small, cross validation is used to gives the model the opportunity to train on multiple train-test splits.

Note: Each of the models' metrics below will be described and analyzed in the Results section.

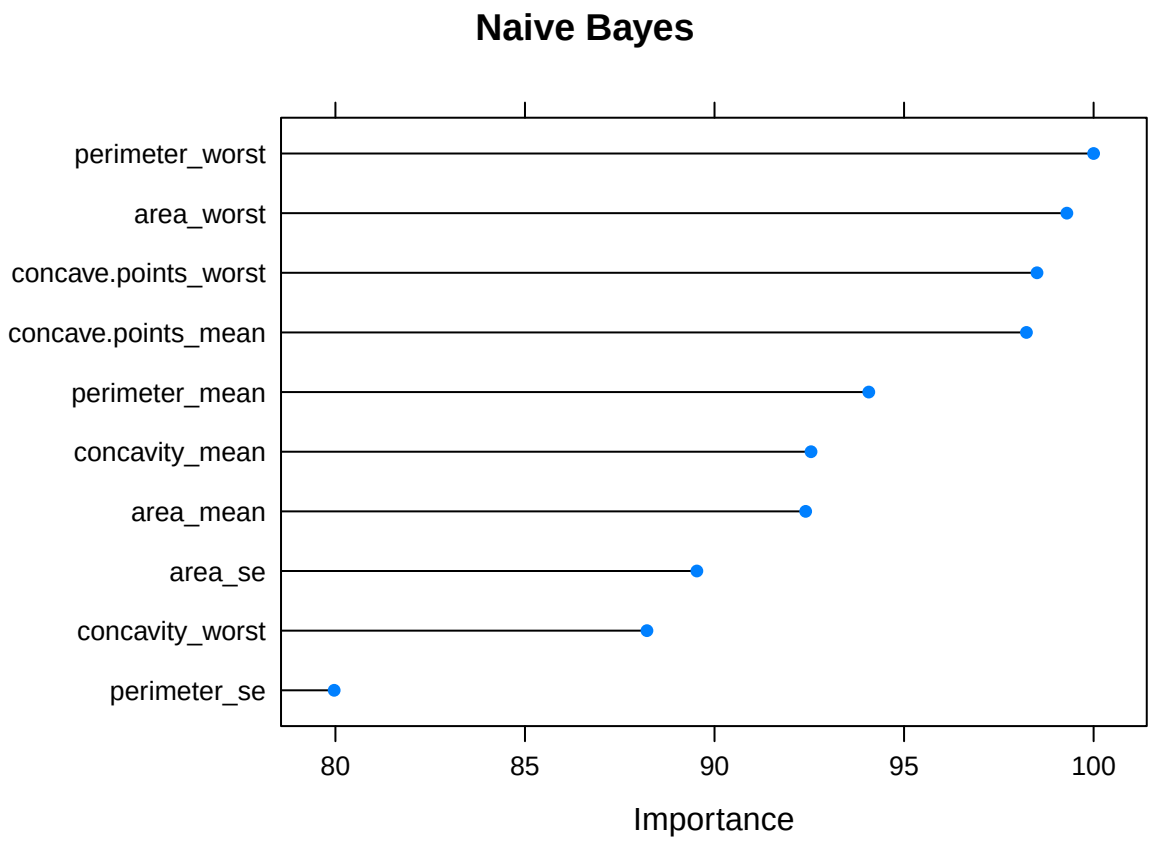
### Naive Bayes Model

The Naive Bayes Model is a classification technique based on the Bayes' theorem and assumes independence among predictors.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  B  M
##          B 66  3
##          M  5 39
##
##                Accuracy : 0.9292
```

```
##                  95% CI : (0.8653, 0.9689)
##     No Information Rate : 0.6283
##     P-Value [Acc > NIR] : 1.372e-13
##
##                   Kappa : 0.8499
##
##  Mcnemar's Test P-Value : 0.7237
##
##             Sensitivity : 0.9286
##             Specificity : 0.9296
##          Pos Pred Value : 0.8864
##          Neg Pred Value : 0.9565
##              Prevalence : 0.3717
##          Detection Rate : 0.3451
##    Detection Prevalence : 0.3894
##       Balanced Accuracy : 0.9291
##
##        'Positive' Class : M
##
```

The most important predictors for the Naive Bayes model are displayed below:



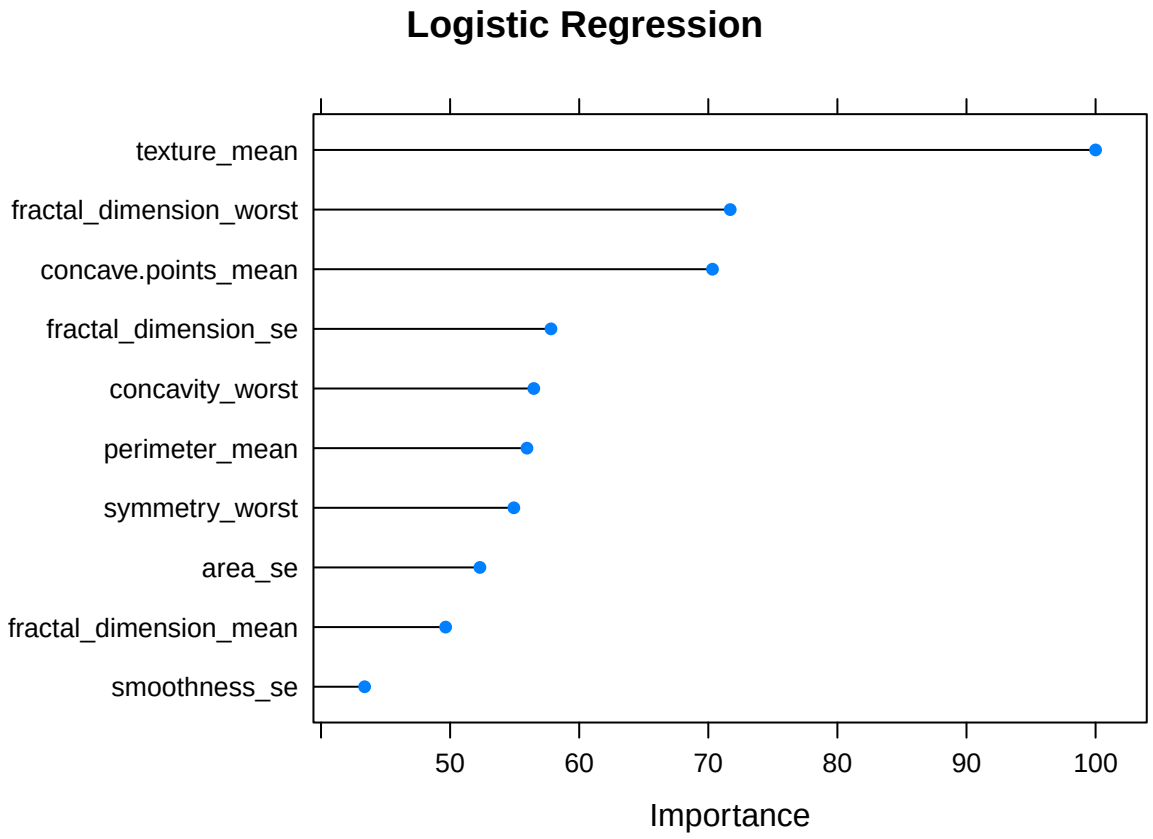

perimeter_worst followed by area_worst, and concave.points_worst are the most important predictors for this model.

**Logistic Regression Model**

Logistic regression is a classification algorithm used to assign observations to a discrete set of classes. Binary logistic model is used to estimate the probability of a binary response based on one or more predictor variables.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  B  M
##          B 67  0
##          M  4 42
##
##                Accuracy : 0.9646
##                  95% CI : (0.9118, 0.9903)
##     No Information Rate : 0.6283
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.9257
##
##  Mcnemar's Test P-Value : 0.1336
##
##             Sensitivity : 1.0000
##             Specificity : 0.9437
##          Pos Pred Value : 0.9130
##          Neg Pred Value : 1.0000
##              Prevalence : 0.3717
##          Detection Rate : 0.3717
##    Detection Prevalence : 0.4071
##       Balanced Accuracy : 0.9718
##
##        'Positive' Class : M
##
```

The most important predictors for the Logistic Regression model are displayed below:




texture_mean is the most important predictor for this model.

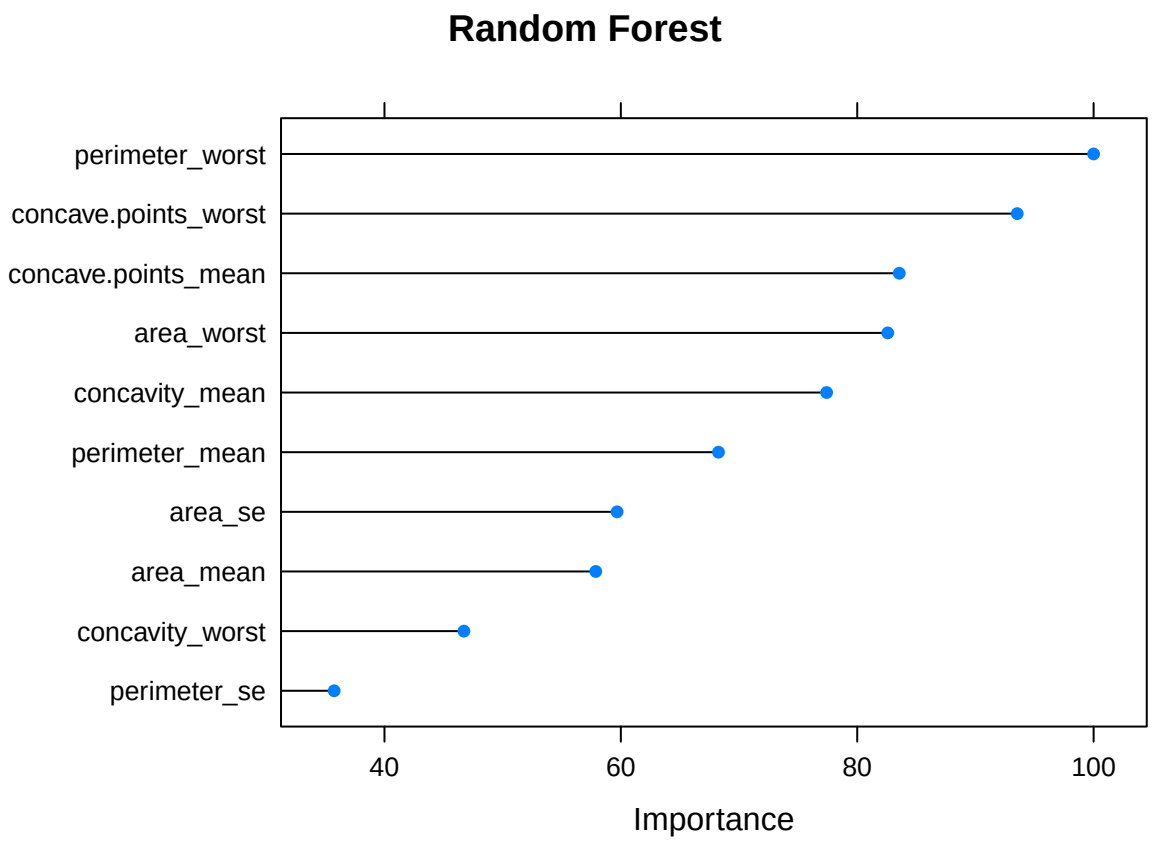
### Random Forest Model

Random Forest, a popular machine learning algorithm, consists of a large number of decision trees that operate as an ensemble. They mostly work well as a large number of relatively uncorrelated models operating together would outperform any of the individual constituent models.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  B  M
##          B 70  1
##          M  1 41
##
##                Accuracy : 0.9823
##                  95% CI : (0.9375, 0.9978)
##     No Information Rate : 0.6283
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.9621
##
##  Mcnemar's Test P-Value : 1
##
##             Sensitivity : 0.9762
```

```
##             Specificity : 0.9859
##          Pos Pred Value : 0.9762
##          Neg Pred Value : 0.9859
##              Prevalence : 0.3717
##          Detection Rate : 0.3628
##    Detection Prevalence : 0.3717
##       Balanced Accuracy : 0.9811
##
##        'Positive' Class : M
##
```

The most important predictors for the Random Forest model are displayed below:




perimeter_worst is the most important predictor for this model.

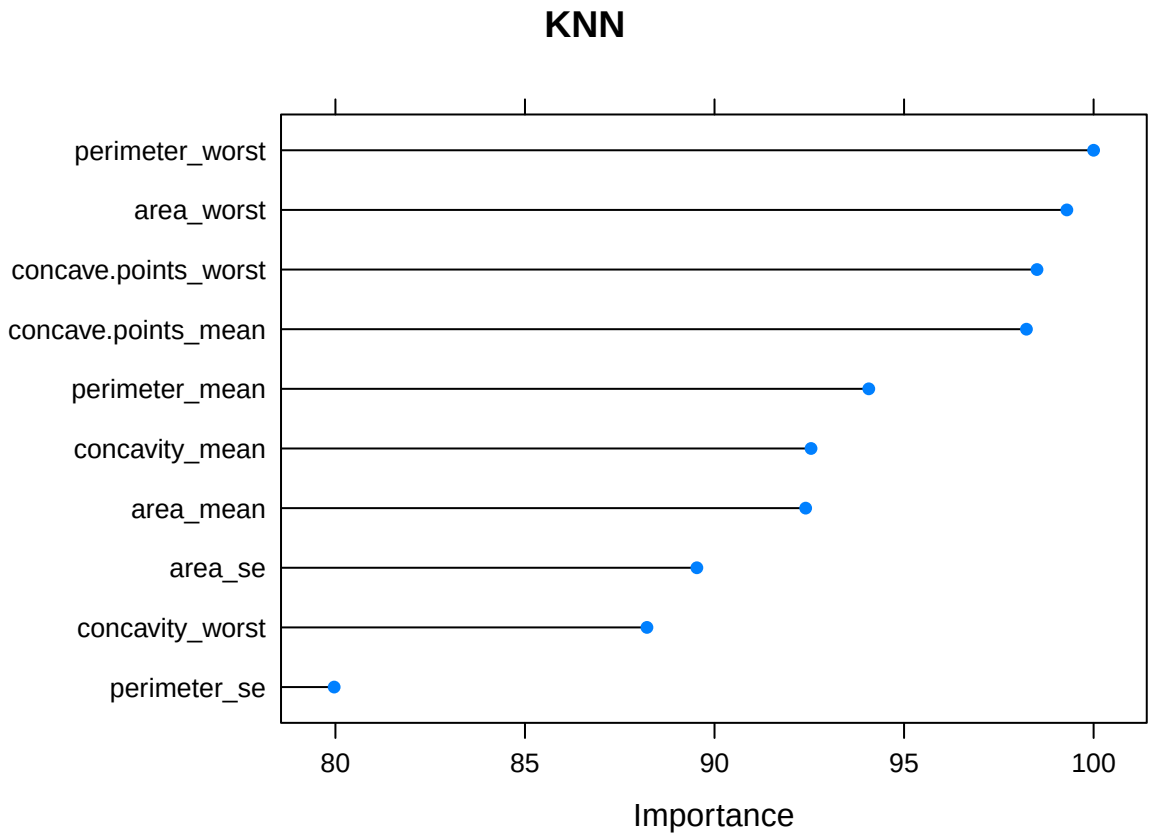
### K-Nearest Neighbor (KNN) Model

KNN algorithms use data and classify new data points based on similarity measures (e.g. distance function). It can be used for both classification and regression.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  B  M
##          B 71  3
##          M  0 39
```

```
##
##                Accuracy : 0.9735
##                  95% CI : (0.9244, 0.9945)
##     No Information Rate : 0.6283
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.9423
##
##  Mcnemar's Test P-Value : 0.2482
##
##             Sensitivity : 0.9286
##             Specificity : 1.0000
##          Pos Pred Value : 1.0000
##          Neg Pred Value : 0.9595
##              Prevalence : 0.3717
##          Detection Rate : 0.3451
##    Detection Prevalence : 0.3451
##       Balanced Accuracy : 0.9643
##
##        'Positive' Class : M
##
```

The most important predictors for the KNN model are displayed below:



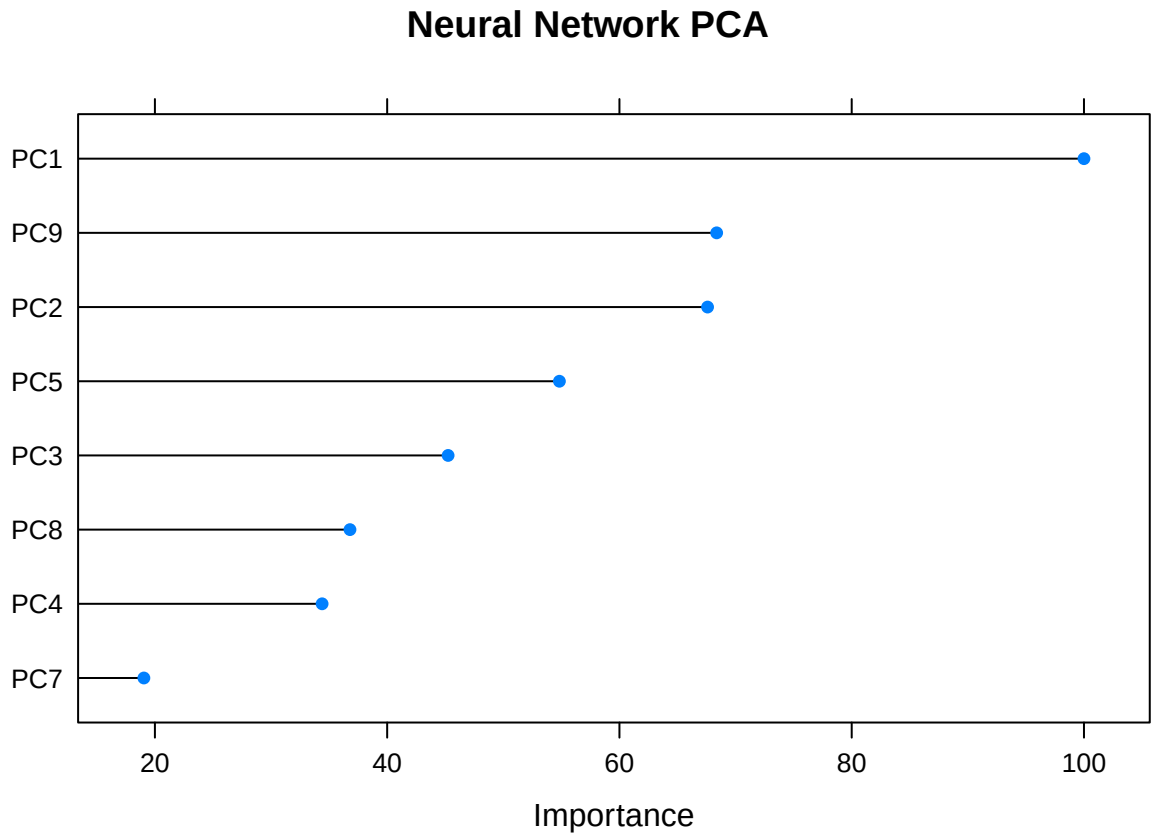

perimeter_worst is the most important predictor for this model.

### Neural Network with PCA Model

Neural networks are a class of machine learning algorithms used to model complex patterns in datasets using multiple hidden layers and non-linear activation functions.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  B  M
##          B 70  0
##          M  1 42
##
##                Accuracy : 0.9912
##                  95% CI : (0.9517, 0.9998)
##     No Information Rate : 0.6283
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.9811
##
##  Mcnemar's Test P-Value : 1
##
##             Sensitivity : 1.0000
##             Specificity : 0.9859
##          Pos Pred Value : 0.9767
##          Neg Pred Value : 1.0000
##              Prevalence : 0.3717
##          Detection Rate : 0.3717
##    Detection Prevalence : 0.3805
##       Balanced Accuracy : 0.9930
##
##        'Positive' Class : M
##
```

The most important predictors for the Neural Network with PCA model are displayed below:




PC1 (principal component) is the most important predictor for this model.

### Neural Network with LDA Model

This Neural Network model is built using the LDA dataframe.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  B  M
##          B 70  1
##          M  1 41
##
##                Accuracy : 0.9823
##                  95% CI : (0.9375, 0.9978)
##     No Information Rate : 0.6283
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.9621
##
##  Mcnemar's Test P-Value : 1
##
##             Sensitivity : 0.9762
##             Specificity : 0.9859
##          Pos Pred Value : 0.9762
```

```
##          Neg Pred Value : 0.9859
##              Prevalence : 0.3717
##          Detection Rate : 0.3628
##    Detection Prevalence : 0.3717
##       Balanced Accuracy : 0.9811
##
##        'Positive' Class : M
##
```

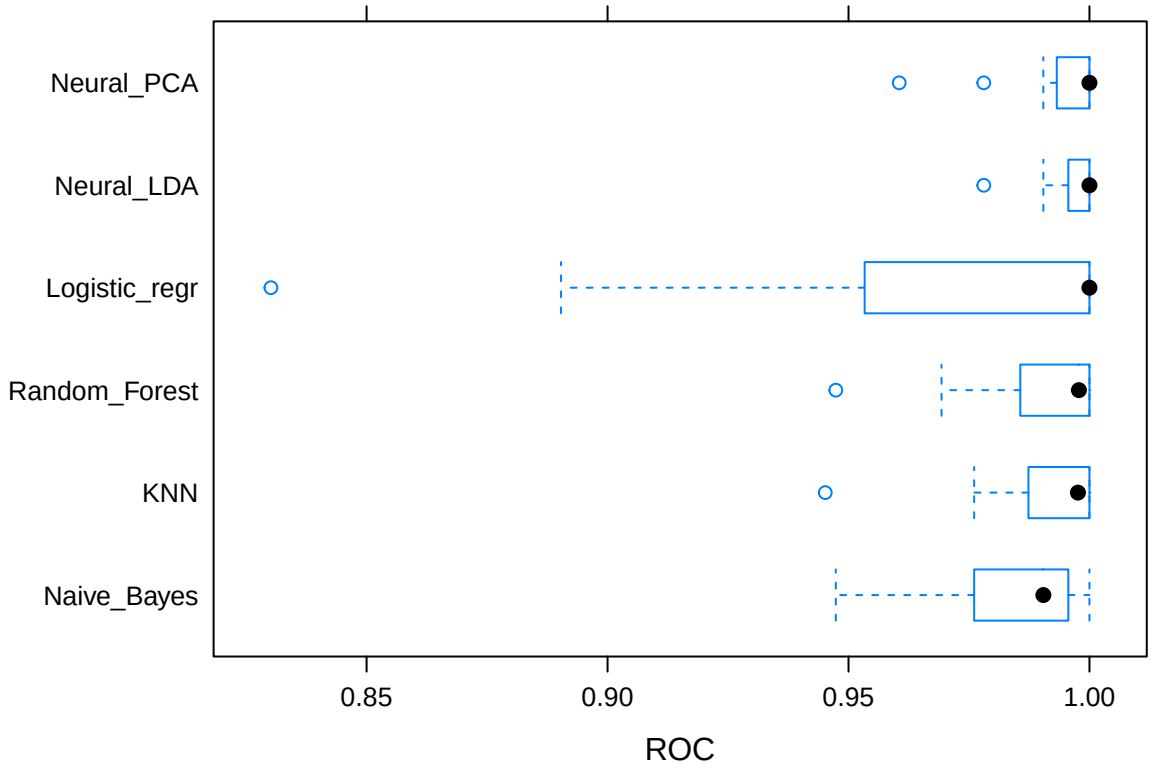
# Results

The previous machine learning algorithms are compared and evaluated.

```
##
## Call:
## summary.resamples(object = models_results)
##
## Models: Naive_Bayes, Logistic_regr, Random_Forest, KNN, Neural_PCA, Neural_LDA
## Number of resamples: 15
##
## ROC
##                    Min.   1st Qu.    Median      Mean  3rd Qu. Max. NA's
## Naive_Bayes   0.9473684 0.9760766 0.9904306 0.9832270 0.995614    1    0
## Logistic_regr 0.8301435 0.9533493 1.0000000 0.9663477 1.000000    1    0
## Random_Forest 0.9473684 0.9856459 0.9978070 0.9902844 1.000000    1    0
## KNN           0.9451754 0.9873405 0.9976077 0.9908347 1.000000    1    0
## Neural_PCA    0.9605263 0.9932217 1.0000000 0.9937267 1.000000    1    0
## Neural_LDA    0.9780702 0.9956140 1.0000000 0.9967092 1.000000    1    0
##
## Sens
##                    Min.   1st Qu.    Median      Mean 3rd Qu. Max. NA's
## Naive_Bayes   0.8421053 0.9473684 0.9473684 0.9512281       1    1    0
## Logistic_regr 0.8421053 0.9236842 1.0000000 0.9582456       1    1    0
## Random_Forest 0.8421053 0.9473684 1.0000000 0.9719298       1    1    0
## KNN           0.8947368 1.0000000 1.0000000 0.9861404       1    1    0
## Neural_PCA    0.9473684 0.9486842 1.0000000 0.9826316       1    1    0
## Neural_LDA    0.9473684 1.0000000 1.0000000 0.9894737       1    1    0
##
## Spec
##                    Min.   1st Qu.    Median      Mean   3rd Qu. Max. NA's
## Naive_Bayes   0.7272727 0.9090909 0.9090909 0.9106061 0.9583333    1    0
## Logistic_regr 0.8181818 0.8712121 1.0000000 0.9419192 1.0000000    1    0
## Random_Forest 0.8181818 0.8712121 0.9166667 0.9303030 1.0000000    1    0
## KNN           0.7272727 0.8712121 0.9166667 0.9136364 1.0000000    1    0
## Neural_PCA    0.8181818 0.9090909 0.9166667 0.9414141 1.0000000    1    0
## Neural_LDA    0.8181818 0.9128788 1.0000000 0.9585859 1.0000000    1    0
```

The plot below indicates the Area Under ROC curve. An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. This curve plots the True Positive Rate and False Positive Rate.

Note that Naive Bayes and Logistic Regression have higher variability while the Neural network with LDA model achieves a great AUC with minimal variability.




## Performance Metrics

Accuracy is the number of correctly predicted data points out of all the data points. It is defined as the number of true positives and true negatives divided by the number of true positives, true negatives, false positives, and false negatives.

Precision is the ratio of the correctly positive (True Positive) labeled by our program to all positive labeled (True Positive and False Positive).

Recall (Sensitivity) is the ratio of the correctly positive labeled by our program to all true positives and false negatives.

F1 Score considers both precision and recall. It is computed as 2 * (Recall * Precision) / (Recall + Precision)

Specificity is the correctly negative labeled (True Negative) by the program to all true negative and false positives.

Below is a summary table of all the algorithms and their performance metrics.

| | Naive_Bayes | Logistic_regr | Random_Forest | KNN | Neural_PCA | Neural_LDA |
|---|---|---|---|---|---|---|
| Sensitivity | 0.9285714 | 1.0000000 | 0.9761905 | 0.9285714 | 1.0000000 | 0.9761905 |
| Specificity | 0.9295775 | 0.9436620 | 0.9859155 | 1.0000000 | 0.9859155 | 0.9859155 |
| Pos Pred Value | 0.8863636 | 0.9130435 | 0.9761905 | 1.0000000 | 0.9767442 | 0.9761905 |
| Neg Pred Value | 0.9565217 | 1.0000000 | 0.9859155 | 0.9594595 | 1.0000000 | 0.9859155 |
| Precision | 0.8863636 | 0.9130435 | 0.9761905 | 1.0000000 | 0.9767442 | 0.9761905 |
| Recall | 0.9285714 | 1.0000000 | 0.9761905 | 0.9285714 | 1.0000000 | 0.9761905 |
| F1 | 0.9069767 | 0.9545455 | 0.9761905 | 0.9629630 | 0.9882353 | 0.9761905 |

| | Naive_Bayes | Logistic_regr | Random_Forest | KNN | Neural_PCA | Neural_LDA |
|---|---|---|---|---|---|---|
| Prevalence | 0.3716814 | 0.3716814 | 0.3716814 | 0.3716814 | 0.3716814 | 0.3716814 |
| Detection Rate | 0.3451327 | 0.3716814 | 0.3628319 | 0.3451327 | 0.3716814 | 0.3628319 |
| Detection Prevalence | 0.3893805 | 0.4070796 | 0.3716814 | 0.3451327 | 0.3805310 | 0.3716814 |
| Balanced Accuracy | 0.9290744 | 0.9718310 | 0.9810530 | 0.9642857 | 0.9929577 | 0.9810530 |

# Conclusion

The table below displays the best model for each performance metric. Through the course of the project, six different machine learning techniques have been performed and evaluated to determine which would be the best to classify a breast cancer cell as benign or malignant. **Neural Networks with PCA has performed the best. Neural Networks with PCA has the highest Accuracy (0.992), Sensitivity (1.00), and F-1 score (0.988) in comparison to the other machine learning models.**

```
##                  metric    best_model     value
## 1           Sensitivity    Neural_PCA 1.0000000
## 2           Specificity           KNN 1.0000000
## 3        Pos Pred Value           KNN 1.0000000
## 4        Neg Pred Value Logistic_regr 1.0000000
## 5             Precision           KNN 1.0000000
## 6                Recall Logistic_regr 1.0000000
## 7                    F1    Neural_PCA 0.9882353
## 8            Prevalence Random_Forest 0.3716814
## 9        Detection Rate Logistic_regr 0.3716814
## 10 Detection Prevalence Logistic_regr 0.4070796
## 11    Balanced Accuracy    Neural_PCA 0.9929577
```

## Future Work

Having a larger dataset would allow for better evaluation of the machine learning algorithms. Despite the time taken to run the code, Neural Network models have performed the best. The results of this project provides a good understanding that the Neural Network with PCA followed by Neural Networks with LDA model yielded the best results and can be used for similar breast cancer classification algorithms.