Breast Cancer Analysis and Prediction using R Language

2023-08-19

Breast Cancer

Breast cancer is a type of cancer that originates in the cells of the breast tissue. It is the most common cancer among women in the world, although it can also occur in men (albeit much less frequently). It accounts for 25% of all cancer cases and affected over 2.1 Million people in 2015 alone. It starts when cells in the breast begin to grow out of control. These cells usually form tumors that can be seen via X-ray or felt as lumps in the breast area. Breast cancer can develop in various parts of the breast, including the milk ducts, lobules (glands that produce milk), or other tissue.

The key challenge against its detection is how to classify tumors into malignant (cancerous) or benign(non-cancerous). This document provides the analysis of classifying these tumors using R language (with SVMs) and the Breast Cancer Wisconsin (Diagnostic) Dataset.

Data source

https://www.kaggle.com/datasets/nancyalaswad90/breast-cancer-dataset.

Loading required libraries

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
library(caret)
```

```
## Loading required package: lattice
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:purrr':
##
##     lift
```

```
library(e1071)
library(randomForest)
```

```
## randomForest 4.7-1.1
## Type rfNews() to see new features/changes/bug fixes.
##
## Attaching package: 'randomForest'
##
## The following object is masked from 'package:dplyr':
##
##     combine
##
## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
library(corrplot)
library(readxl)
```

Loading the dataset

```
data <- read_excel("data.xlsx")
```

Check the structure and summary of the data

```
head(data)
```

```
## # A tibble: 6 x 32
##         id diagnosis radius_mean texture_mean perimeter_mean area_mean
##      <dbl> <chr>           <dbl>        <dbl>          <dbl>     <dbl>
## 1   842302 M               18.0         10.4          123.      1001
## 2   842517 M               20.6         17.8          133.      1326
## 3 84300903 M               19.7         21.2          130       1203
## 4 84348301 M               11.4         20.4           77.6      386.
## 5 84358402 M               20.3         14.3          135.      1297
## 6   843786 M               12.4         15.7           82.6      477.
## # i 26 more variables: smoothness_mean <dbl>, compactness_mean <dbl>,
## #   concavity_mean <dbl>, concave_points_mean <dbl>, symmetry_mean <dbl>,
## #   fractal_dimension_mean <dbl>, radius_se <dbl>, texture_se <dbl>,
## #   perimeter_se <dbl>, area_se <dbl>, smoothness_se <dbl>,
```

```
## #   compactness_se <dbl>, concavity_se <dbl>, concave_points_se <dbl>,
## #   symmetry_se <dbl>, fractal_dimension_se <dbl>, radius_worst <dbl>,
## #   texture_worst <dbl>, perimeter_worst <dbl>, area_worst <dbl>, ...
```

```
str(data)
```

```
## tibble [569 x 32] (S3: tbl_df/tbl/data.frame)
##  $ id                     : num [1:569] 842302 842517 84300903 84348301 84358402 ...
##  $ diagnosis              : chr [1:569] "M" "M" "M" "M" ...
##  $ radius_mean            : num [1:569] 18 20.6 19.7 11.4 20.3 ...
##  $ texture_mean           : num [1:569] 10.4 17.8 21.2 20.4 14.3 ...
##  $ perimeter_mean         : num [1:569] 122.8 132.9 130 77.6 135.1 ...
##  $ area_mean              : num [1:569] 1001 1326 1203 386 1297 ...
##  $ smoothness_mean        : num [1:569] 0.1184 0.0847 0.1096 0.1425 0.1003 ...
##  $ compactness_mean       : num [1:569] 0.2776 0.0786 0.1599 0.2839 0.1328 ...
##  $ concavity_mean         : num [1:569] 0.3001 0.0869 0.1974 0.2414 0.198 ...
##  $ concave_points_mean    : num [1:569] 0.1471 0.0702 0.1279 0.1052 0.1043 ...
##  $ symmetry_mean          : num [1:569] 0.242 0.181 0.207 0.26 0.181 ...
##  $ fractal_dimension_mean : num [1:569] 0.0787 0.0567 0.06 0.0974 0.0588 ...
##  $ radius_se              : num [1:569] 1.095 0.543 0.746 0.496 0.757 ...
##  $ texture_se             : num [1:569] 0.905 0.734 0.787 1.156 0.781 ...
##  $ perimeter_se           : num [1:569] 8.59 3.4 4.58 3.44 5.44 ...
##  $ area_se                : num [1:569] 153.4 74.1 94 27.2 94.4 ...
##  $ smoothness_se          : num [1:569] 0.0064 0.00522 0.00615 0.00911 0.01149 ...
##  $ compactness_se         : num [1:569] 0.049 0.0131 0.0401 0.0746 0.0246 ...
##  $ concavity_se           : num [1:569] 0.0537 0.0186 0.0383 0.0566 0.0569 ...
##  $ concave_points_se      : num [1:569] 0.0159 0.0134 0.0206 0.0187 0.0188 ...
##  $ symmetry_se            : num [1:569] 0.03 0.0139 0.0225 0.0596 0.0176 ...
##  $ fractal_dimension_se   : num [1:569] 0.00619 0.00353 0.00457 0.00921 0.00511 ...
##  $ radius_worst           : num [1:569] 25.4 25 23.6 14.9 22.5 ...
##  $ texture_worst          : num [1:569] 17.3 23.4 25.5 26.5 16.7 ...
##  $ perimeter_worst        : num [1:569] 184.6 158.8 152.5 98.9 152.2 ...
##  $ area_worst             : num [1:569] 2019 1956 1709 568 1575 ...
##  $ smoothness_worst       : num [1:569] 0.162 0.124 0.144 0.21 0.137 ...
##  $ compactness_worst      : num [1:569] 0.666 0.187 0.424 0.866 0.205 ...
##  $ concavity_worst        : num [1:569] 0.712 0.242 0.45 0.687 0.4 ...
##  $ concave_points_worst   : num [1:569] 0.265 0.186 0.243 0.258 0.163 ...
##  $ symmetry_worst         : num [1:569] 0.46 0.275 0.361 0.664 0.236 ...
##  $ fractal_dimension_worst: num [1:569] 0.1189 0.089 0.0876 0.173 0.0768 ...
```

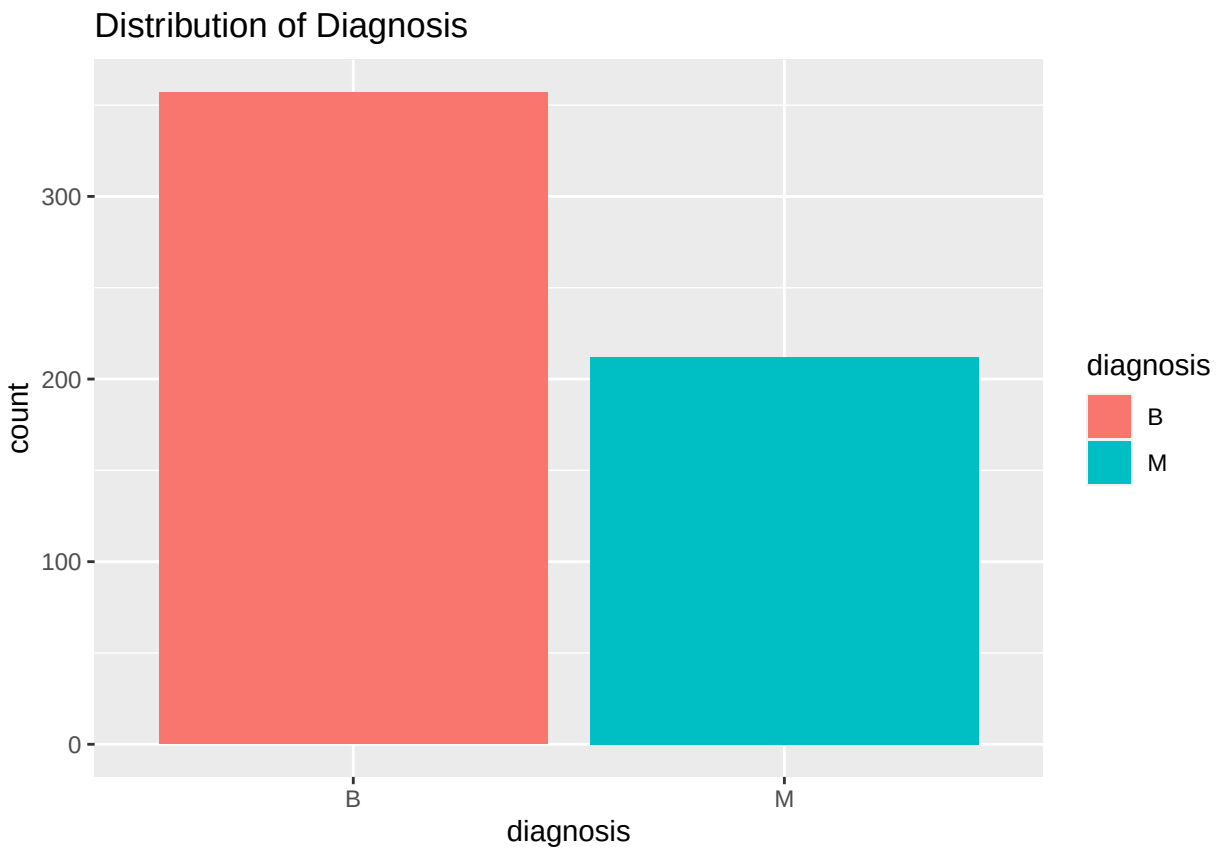
```
summary(data)
```

```
##        id              diagnosis          radius_mean      texture_mean
##  Min.   :     8670   Length:569         Min.   : 6.981   Min.   : 9.71
##  1st Qu.:   869218   Class :character   1st Qu.:11.700   1st Qu.:16.17
##  Median :   906024   Mode  :character   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
```

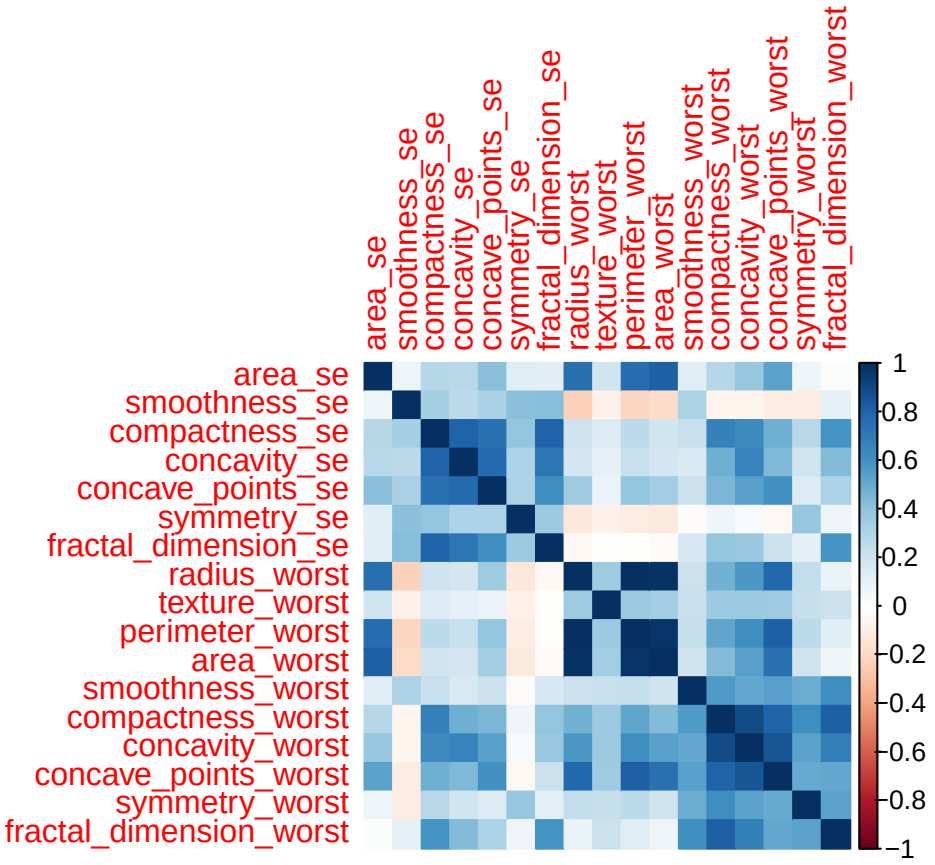
Distribution of diagnosis

```
ggplot(data, aes(x = diagnosis, fill = diagnosis)) +
  geom_bar() +
  labs(title = "Distribution of Diagnosis")
```



Correlation matrix

```
cor_matrix <- cor(data[, 16:32])  # Excluding the 'id' column and 'diagnosis' column
corrplot(cor_matrix, method = "color")
```



Encoding diagnosis as binary variable

```
data$diagnosis <- factor(data$diagnosis, levels = c("B", "M"), labels = c("Benign", "Malignant"))
```

Handling missing values (if needed)

```
data_cleaned <- na.omit(data)
```

Spliting the dataset into train and test sets

```
set.seed(123)
trainIndex <- createDataPartition(data$diagnosis, p = 0.7, list = FALSE, times = 1)
train_data <- data[trainIndex, ]
test_data <- data[-trainIndex, ]
```

Support Vector Machine (SVM) model

```r
svm_model <- svm(diagnosis ~ ., data = train_data)
```

Random Forest model

```r
rf_model <- randomForest(diagnosis ~ ., data = train_data, ntree = 100)
```

Predictions using SVM and Random Forest

```r
svm_pred <- predict(svm_model, newdata = test_data)
rf_pred <- predict(rf_model, newdata = test_data)
```

Evaluate models

```r
confusionMatrix(svm_pred, test_data$diagnosis)
```

```
## Confusion Matrix and Statistics
##
##            Reference
## Prediction  Benign Malignant
##   Benign       103         3
##   Malignant      4        60
##
##                Accuracy : 0.9588
##                  95% CI : (0.917, 0.9833)
##     No Information Rate : 0.6294
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.912
##
##  Mcnemar's Test P-Value : 1
##
##             Sensitivity : 0.9626
##             Specificity : 0.9524
##          Pos Pred Value : 0.9717
##          Neg Pred Value : 0.9375
##              Prevalence : 0.6294
##          Detection Rate : 0.6059
##    Detection Prevalence : 0.6235
##       Balanced Accuracy : 0.9575
##
##        'Positive' Class : Benign
##
```

```r
confusionMatrix(rf_pred, test_data$diagnosis)
```

```
## Confusion Matrix and Statistics
##
##            Reference
## Prediction  Benign Malignant
##   Benign       105         3
##   Malignant      2        60
##
##                Accuracy : 0.9706
##                  95% CI : (0.9327, 0.9904)
##     No Information Rate : 0.6294
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.9367
##
##  Mcnemar's Test P-Value : 1
##
##             Sensitivity : 0.9813
##             Specificity : 0.9524
##          Pos Pred Value : 0.9722
##          Neg Pred Value : 0.9677
##              Prevalence : 0.6294
##          Detection Rate : 0.6176
##    Detection Prevalence : 0.6353
##       Balanced Accuracy : 0.9668
##
##        'Positive' Class : Benign
##
```

Summary of the statistics

For the SVM Model: - Accuracy: 0.9588 - 95% Confidence Interval (CI): (0.917, 0.9833) - No Information Rate: 0.6294 - Kappa: 0.912 - Sensitivity (True Positive Rate): 0.9626 - Specificity (True Negative Rate): 0.9524 - Positive Predictive Value (Precision): 0.9717 - Negative Predictive Value: 0.9375 - Prevalence: 0.6294 - Detection Rate: 0.6059 - Detection Prevalence: 0.6235 - Balanced Accuracy: 0.9575

For the Random Forest Model: - Accuracy: 0.9706 - 95% Confidence Interval (CI): (0.9327, 0.9904) - No Information Rate: 0.6294 - Kappa: 0.9367 - Sensitivity (True Positive Rate): 0.9813 - Specificity (True Negative Rate): 0.9524 - Positive Predictive Value (Precision): 0.9722 - Negative Predictive Value: 0.9677 - Prevalence: 0.6294 - Detection Rate: 0.6176 - Detection Prevalence: 0.6353 - Balanced Accuracy: 0.9668

These statistics provide a comprehensive assessment of how well the models are performing.