BREAST CANCER DETECTION

2023-04-04

Loading the dataset

```
path <- file.choose()
```

```
df <- read.csv(path)
df <- df[,-33]
head(df)
```

```
##         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
```

Handling the NA Values with mean of the feature records and also omitting remaining NA value's records

```
df$radius_mean <- ifelse(is.na(df$radius_mean),
                    ave(df$radius_mean, FUN = function(x)mean(x, na.rm = TRUE)),
                    df$radius_mean)
df$area_mean <- ifelse(is.na(df$area_mean),
                    ave(df$area_mean, FUN = function(x)mean(x, na.rm = TRUE)),
                    df$area_mean)
df$concave.points_worst <- ifelse(is.na(df$concave.points_worst),
                    ave(df$concave.points_worst, FUN = function(x)mean(x, na.rm = TRUE)),
                    df$concave.points_worst)
df$area_worst = ifelse(is.na(df$area_worst),
                    ave(df$area_worst, FUN = function(x)mean(x, na.rm = TRUE)),
                    df$area_worst)
df$concave.points_mean = ifelse(is.na(df$concave.points_mean),
                    ave(df$concave.points_mean, FUN = function(x)mean(x, na.rm = TRUE)),
                    df$concave.points_mean)
df$area_se = ifelse(is.na(df$area_se),
                    ave(df$area_se, FUN = function(x)mean(x, na.rm = TRUE)),
                    df$area_se)
df$concavity_se = ifelse(is.na(df$concavity_se),
                    ave(df$concavity_se, FUN = function(x)mean(x, na.rm = TRUE)),
                    df$concavity_se)
```

```
#Removing the ID Column as this doesn't affect the result
df <- df[,-1]
```

```
#Encoding the Categorical data for Diagnosis where 1 represents M (Malignant) and 2 represents B (Benig
df$diagnosis <- factor(df$diagnosis,levels = c('M','B'),labels = c(1, 2))
head(df)
```

```
##   diagnosis radius_mean texture_mean perimeter_mean area_mean smoothness_mean
## 1         1       17.99        10.38         122.80    1001.0         0.11840
## 2         1       20.57        17.77         132.90    1326.0         0.08474
```

```
## 3         1       19.69        21.25         130.00    1203.0         0.10960
## 4         1       11.42        20.38          77.58     386.1         0.14250
## 5         1       20.29        14.34         135.10    1297.0         0.10030
## 6         1       12.45        15.70          82.57     477.1         0.12780
##   compactness_mean concavity_mean concave.points_mean symmetry_mean
## 1          0.27760         0.3001             0.14710        0.2419
## 2          0.07864         0.0869             0.07017        0.1812
## 3          0.15990         0.1974             0.12790        0.2069
## 4          0.28390         0.2414             0.10520        0.2597
## 5          0.13280         0.1980             0.10430        0.1809
## 6          0.17000         0.1578             0.08089        0.2087
##   fractal_dimension_mean radius_se texture_se perimeter_se area_se
## 1                0.07871    1.0950     0.9053        8.589  153.40
## 2                0.05667    0.5435     0.7339        3.398   74.08
## 3                0.05999    0.7456     0.7869        4.585   94.03
## 4                0.09744    0.4956     1.1560        3.445   27.23
## 5                0.05883    0.7572     0.7813        5.438   94.44
## 6                0.07613    0.3345     0.8902        2.217   27.19
##   smoothness_se compactness_se concavity_se concave.points_se symmetry_se
## 1      0.006399        0.04904      0.05373           0.01587     0.03003
## 2      0.005225        0.01308      0.01860           0.01340     0.01389
## 3      0.006150        0.04006      0.03832           0.02058     0.02250
## 4      0.009110        0.07458      0.05661           0.01867     0.05963
## 5      0.011490        0.02461      0.05688           0.01885     0.01756
## 6      0.007510        0.03345      0.03672           0.01137     0.02165
##   fractal_dimension_se radius_worst texture_worst perimeter_worst area_worst
## 1             0.006193        25.38         17.33          184.60     2019.0
## 2             0.003532        24.99         23.41          158.80     1956.0
## 3             0.004571        23.57         25.53          152.50     1709.0
## 4             0.009208        14.91         26.50           98.87      567.7
## 5             0.005115        22.54         16.67          152.20     1575.0
## 6             0.005082        15.47         23.75          103.40      741.6
##   smoothness_worst compactness_worst concavity_worst concave.points_worst
## 1           0.1622            0.6656          0.7119               0.2654
## 2           0.1238            0.1866          0.2416               0.1860
## 3           0.1444            0.4245          0.4504               0.2430
## 4           0.2098            0.8663          0.6869               0.2575
## 5           0.1374            0.2050          0.4000               0.1625
## 6           0.1791            0.5249          0.5355               0.1741
##   symmetry_worst fractal_dimension_worst
## 1         0.4601                 0.11890
## 2         0.2750                 0.08902
## 3         0.3613                 0.08758
## 4         0.6638                 0.17300
## 5         0.2364                 0.07678
## 6         0.3985                 0.12440
```

DATA SPLITTING

```
library(dplyr)
```

```
##
```

```
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
df$id <- 1:nrow(df)

# Splitting into 80-training and 20-test dataset
trn <- df%>%dplyr::sample_frac(0.80)
tst <- dplyr::anti_join(df,trn,by='id')
```

```
trn <- trn[,-32]
tst <- trn[,-32]
df <- df[,-32]
```

```
head(trn)
```

```
##   diagnosis radius_mean texture_mean perimeter_mean area_mean smoothness_mean
## 1         2       8.671        14.45          54.42     227.2         0.09138
## 2         1      20.180        23.97         143.70    1245.0         0.12860
## 3         1      16.070        19.65         104.10     817.7         0.09168
## 4         2      13.650        13.16          87.88     568.9         0.09646
## 5         1      19.270        26.47         127.90    1162.0         0.09401
## 6         2       9.904        18.06          64.60     302.4         0.09699
##   compactness_mean concavity_mean concave.points_mean symmetry_mean
## 1          0.04276        0.00000             0.00000        0.1722
## 2          0.34540        0.37540             0.16040        0.2906
## 3          0.08424        0.09769             0.06638        0.1798
## 4          0.08711        0.03888             0.02563        0.1360
## 5          0.17190        0.16570             0.07593        0.1853
## 6          0.12940        0.13070             0.03716        0.1669
##   fractal_dimension_mean radius_se texture_se perimeter_se area_se
## 1                0.06724    0.2204     0.7873        1.435   11.36
## 2                0.08142    0.9317     1.8850        8.649  116.40
## 3                0.05391    0.7474     1.0160        5.029   79.25
## 4                0.06344    0.2102     0.4336        1.391   17.40
## 5                0.06261    0.5558     0.6062        3.528   68.17
## 6                0.08116    0.4311     2.2610        3.132   27.48
##   smoothness_se compactness_se concavity_se concave.points_se symmetry_se
## 1      0.009172       0.008007      0.00000          0.000000     0.02711
## 2      0.010380       0.068350      0.10910          0.025930     0.07895
## 3      0.010820       0.022030      0.03500          0.018090     0.01550
## 4      0.004133       0.016950      0.01652          0.006659     0.01371
## 5      0.005015       0.033180      0.03497          0.009643     0.01543
## 6      0.012860       0.088080      0.11970          0.024600     0.03880
##   fractal_dimension_se radius_worst texture_worst perimeter_worst area_worst
## 1             0.003399        9.262         17.04           58.36      259.2
```

```
## 2             0.005987       23.370         31.72          170.30     1623.0
## 3             0.001948       19.770         24.56          128.80     1223.0
## 4             0.002735       15.340         16.35           99.71      706.2
## 5             0.003896       24.150         30.90          161.40     1813.0
## 6             0.017920       11.260         24.39           73.07      390.2
##   smoothness_worst compactness_worst concavity_worst concave.points_worst
## 1           0.1162           0.07057          0.0000              0.00000
## 2           0.1639           0.61640          0.7681              0.25080
## 3           0.1500           0.20450          0.2829              0.15200
## 4           0.1311           0.24740          0.1759              0.08056
## 5           0.1509           0.65900          0.6091              0.17850
## 6           0.1301           0.29500          0.3486              0.09910
##   symmetry_worst fractal_dimension_worst
## 1         0.2592                 0.07848
## 2         0.5440                 0.09964
## 3         0.2650                 0.06387
## 4         0.2380                 0.08718
## 5         0.3672                 0.11230
## 6         0.2614                 0.11620
```

```
head(tst)
```

```
##   diagnosis radius_mean texture_mean perimeter_mean area_mean smoothness_mean
## 1         2       8.671        14.45          54.42     227.2         0.09138
## 2         1      20.180        23.97         143.70    1245.0         0.12860
## 3         1      16.070        19.65         104.10     817.7         0.09168
## 4         2      13.650        13.16          87.88     568.9         0.09646
## 5         1      19.270        26.47         127.90    1162.0         0.09401
## 6         2       9.904        18.06          64.60     302.4         0.09699
##   compactness_mean concavity_mean concave.points_mean symmetry_mean
## 1          0.04276        0.00000             0.00000        0.1722
## 2          0.34540        0.37540             0.16040        0.2906
## 3          0.08424        0.09769             0.06638        0.1798
## 4          0.08711        0.03888             0.02563        0.1360
## 5          0.17190        0.16570             0.07593        0.1853
## 6          0.12940        0.13070             0.03716        0.1669
##   fractal_dimension_mean radius_se texture_se perimeter_se area_se
## 1                0.06724    0.2204     0.7873        1.435   11.36
## 2                0.08142    0.9317     1.8850        8.649  116.40
## 3                0.05391    0.7474     1.0160        5.029   79.25
## 4                0.06344    0.2102     0.4336        1.391   17.40
## 5                0.06261    0.5558     0.6062        3.528   68.17
## 6                0.08116    0.4311     2.2610        3.132   27.48
##   smoothness_se compactness_se concavity_se concave.points_se symmetry_se
## 1      0.009172       0.008007      0.00000          0.000000     0.02711
## 2      0.010380       0.068350      0.10910          0.025930     0.07895
## 3      0.010820       0.022030      0.03500          0.018090     0.01550
## 4      0.004133       0.016950      0.01652          0.006659     0.01371
## 5      0.005015       0.033180      0.03497          0.009643     0.01543
## 6      0.012860       0.088080      0.11970          0.024600     0.03880
##   fractal_dimension_se radius_worst texture_worst perimeter_worst area_worst
## 1             0.003399        9.262         17.04           58.36      259.2
## 2             0.005987       23.370         31.72          170.30     1623.0
## 3             0.001948       19.770         24.56          128.80     1223.0
```

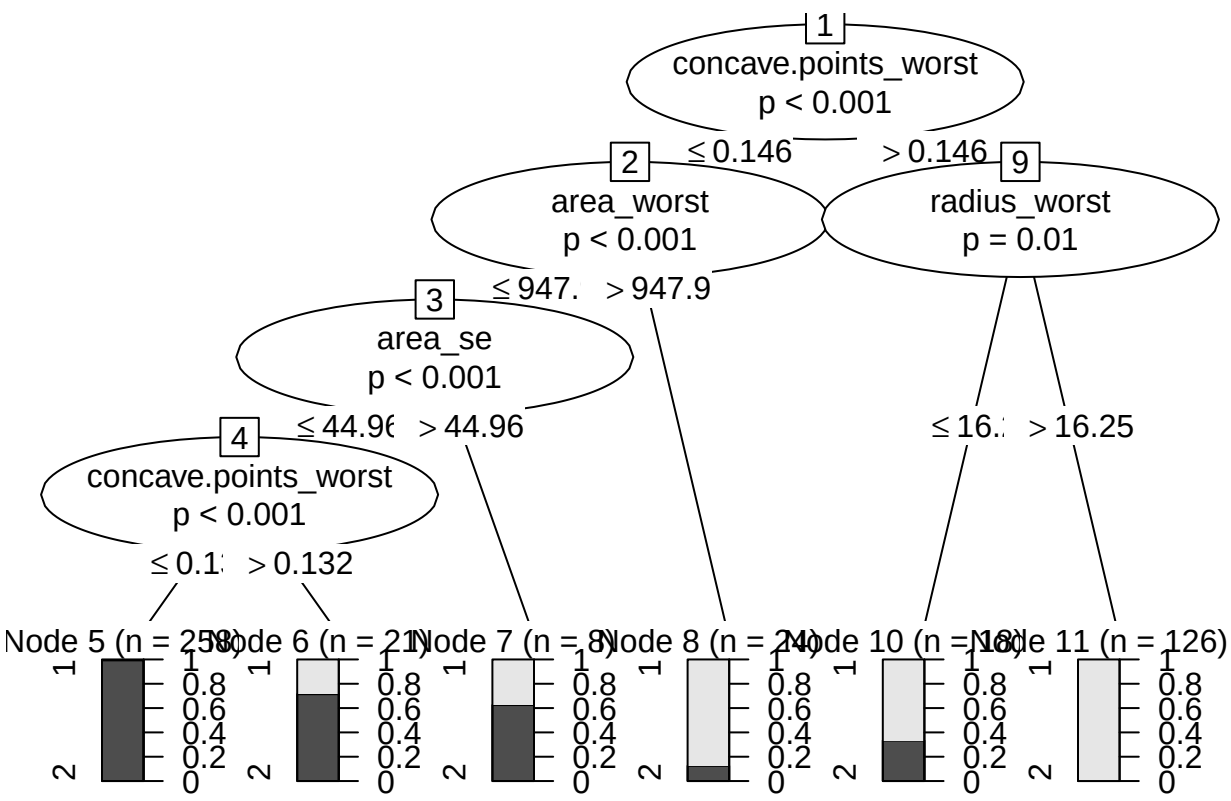
```
## 4             0.002735       15.340         16.35           99.71      706.2
## 5             0.003896       24.150         30.90          161.40     1813.0
## 6             0.017920       11.260         24.39           73.07      390.2
##   smoothness_worst compactness_worst concavity_worst concave.points_worst
## 1           0.1162           0.07057          0.0000               0.00000
## 2           0.1639           0.61640          0.7681               0.25080
## 3           0.1500           0.20450          0.2829               0.15200
## 4           0.1311           0.24740          0.1759               0.08056
## 5           0.1509           0.65900          0.6091               0.17850
## 6           0.1301           0.29500          0.3486               0.09910
##   symmetry_worst fractal_dimension_worst
## 1         0.2592                 0.07848
## 2         0.5440                 0.09964
## 3         0.2650                 0.06387
## 4         0.2380                 0.08718
## 5         0.3672                 0.11230
## 6         0.2614                 0.11620
```

DECISION TREE MODEL

```
#Implementing decision tree classifier
library(party)
```

```
## Warning: package 'party' was built under R version 4.2.3

## Loading required package: grid

## Loading required package: mvtnorm

## Loading required package: modeltools

## Loading required package: stats4

## Loading required package: strucchange

## Warning: package 'strucchange' was built under R version 4.2.3

## Loading required package: zoo

## Warning: package 'zoo' was built under R version 4.2.3

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

## Loading required package: sandwich
```

```
dt <- ctree(diagnosis~., trn)
plot(dt)
```




```
dt
```

```
##
##   Conditional inference tree with 6 terminal nodes
##
## Response:  diagnosis
## Inputs:  radius_mean, texture_mean, perimeter_mean, area_mean, smoothness_mean, compactness_mean, co
## Number of observations:  455
##
## 1) concave.points_worst <= 0.1465; criterion = 1, statistic = 283.472
##   2) area_worst <= 947.9; criterion = 1, statistic = 121.515
##     3) area_se <= 44.96; criterion = 1, statistic = 20.724
##       4) concave.points_worst <= 0.1318; criterion = 1, statistic = 19.709
##         5)*  weights = 258
##       4) concave.points_worst > 0.1318
##         6)*  weights = 21
##     3) area_se > 44.96
##       7)*  weights = 8
##   2) area_worst > 947.9
##     8)*  weights = 24
## 1) concave.points_worst > 0.1465
##   9) radius_worst <= 16.25; criterion = 0.99, statistic = 12.953
```

```
##     10)*  weights = 18
##   9) radius_worst > 16.25
##     11)*  weights = 126
```

```
#Predicted output of test data
dt_pred <- predict(dt,tst)
dt_pred
```

```
##   [1] 2 1 1 2 1 2 2 2 1 2 2 2 2 1 1 2 2 2 2 1 1 2 2 2 2 2 1 1 2 2 2 2 1 1 2 2 2
##  [38] 2 2 1 2 2 2 1 2 1 2 2 2 2 1 1 2 2 1 2 2 2 2 2 2 1 1 2 2 2 2 1 2 2 1 2 2 2
##  [75] 2 1 2 1 2 1 2 2 2 2 1 1 2 2 2 2 1 2 2 1 2 2 2 2 2 2 2 1 1 1 2 2 1 1 1 1 2
## [112] 1 2 1 1 2 1 2 2 1 2 2 1 2 2 2 2 2 1 2 1 1 2 1 2 1 1 2 1 2 2 2 1 2 2 1 2 2
## [149] 1 2 2 1 2 2 2 1 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2 1 2 2 2 1 1 2 2 2 2 2 1 2 2 1
## [186] 2 2 1 1 1 2 2 2 2 2 1 1 2 2 2 2 2 2 1 2 2 1 2 2 2 2 1 2 2 1 1 1 2 2 2 2 2 1
## [223] 2 2 1 2 2 2 2 2 1 2 2 1 1 1 1 1 1 2 2 2 2 1 2 2 2 2 2 2 2 2 1 1 2 2 1 2 1 2 2 1
## [260] 2 1 1 1 1 1 2 2 2 2 2 2 2 1 1 2 2 2 2 2 1 2 2 1 1 2 1 1 1 2 2 2 1 1 1 2 2 2 1 1
## [297] 1 1 2 2 2 1 2 1 2 2 2 1 2 2 1 2 2 1 2 2 2 1 1 2 1 1 2 2 2 2 2 1 2 2 2 2 1 2 1
## [334] 2 2 2 2 2 2 2 1 2 2 2 2 2 1 1 2 2 1 1 2 1 2 2 2 2 2 1 2 1 2 1 2 2 2 2 2 1 2 1 2 2
## [371] 2 2 1 1 1 1 2 2 1 1 2 2 2 1 1 1 1 1 2 2 2 2 2 2 2 2 1 2 1 1 1 2 1 2 1 1 2 2
## [408] 1 1 2 2 1 1 1 2 1 2 2 2 1 1 1 2 1 1 2 2 2 1 2 1 1 1 1 2 2 2 1 2 1 2 2 2 1
## [445] 2 2 1 1 2 2 2 2 2 2 2
## Levels: 1 2
```

```
#Confusion matrix for the decision tree model
library(caret)
```

```
## Loading required package: ggplot2
```

```
## Loading required package: lattice
```

```
confusionMatrix(dt_pred, tst$diagnosis)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   1   2
##          1 159   9
##          2  11 276
##
##                Accuracy : 0.956
##                  95% CI : (0.9329, 0.9729)
##     No Information Rate : 0.6264
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.9059
##
##  Mcnemar's Test P-Value : 0.8231
##
##             Sensitivity : 0.9353
##             Specificity : 0.9684
##          Pos Pred Value : 0.9464
```

```
##          Neg Pred Value : 0.9617
##              Prevalence : 0.3736
##          Detection Rate : 0.3495
##    Detection Prevalence : 0.3692
##       Balanced Accuracy : 0.9519
##
##        'Positive' Class : 1
##
```

Naive Bayes model

```
library(e1071)
nb <- naiveBayes(diagnosis ~ ., data = trn)
nb_pred <- predict(nb, newdata = tst)
confusionMatrix(nb_pred, tst$diagnosis)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   1   2
##          1 150  13
##          2  20 272
##
##                Accuracy : 0.9275
##                  95% CI : (0.8996, 0.9496)
##     No Information Rate : 0.6264
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.8437
##
##  Mcnemar's Test P-Value : 0.2963
##
##             Sensitivity : 0.8824
##             Specificity : 0.9544
##          Pos Pred Value : 0.9202
##          Neg Pred Value : 0.9315
##              Prevalence : 0.3736
##          Detection Rate : 0.3297
##    Detection Prevalence : 0.3582
##       Balanced Accuracy : 0.9184
##
##        'Positive' Class : 1
##
```