ASSIGNMENT – 7

The two chosen datasets are from the Breast Cancer Wisconsin UCI repository – "breast-cancer-wisconsin.data" and "wdbc.data".

The primary objective is to predict whether the tumor was benign or malignant based on various features in the two datasets.

BREAST CANCER WISCONSIN DATA:

1. EXPLORATORY DATA ANALYSIS

1. The dataset had 699 observations for 11 variables with 'classes' – benign and malignant as the dependent variable.
2. Of the 9 predictors (excluding the sample code number), 8 were integers and 1 variable was factor.
3. The summary of the data frame revealed that the factor predictor – 'bare_nuclei' had 16 NA's. MICE package was used to impute the missing values after converting it to a numeric variable.
4. All predictors were converted to numeric values.
5. Correlation among the variables was plotted to identify highly correlated variables. Due to high correlations, ML models can fail. So PCA was used later to reduce dimensionality.
6. A feature importance graph was also plotted to identify the variables important for further analysis. It showed 8 variables as important excluding the predictor 'mitosis'.
7. To identify outliers, boxplots were drawn for the 8 variables. There were very few outliers in four variables which wouldn't interfere with the modeling techniques.
8. Histograms were plotted to understand the data distribution and normality. Some of them showed a right skewed distribution, while others were erratic indicating no normality.
9. QQ-plots were also drawn to understand normality in data distribution.
10. Principal Component Analysis (PCA) was performed and it gave nine PC's with the 9th PC explaining all the variance in the data.

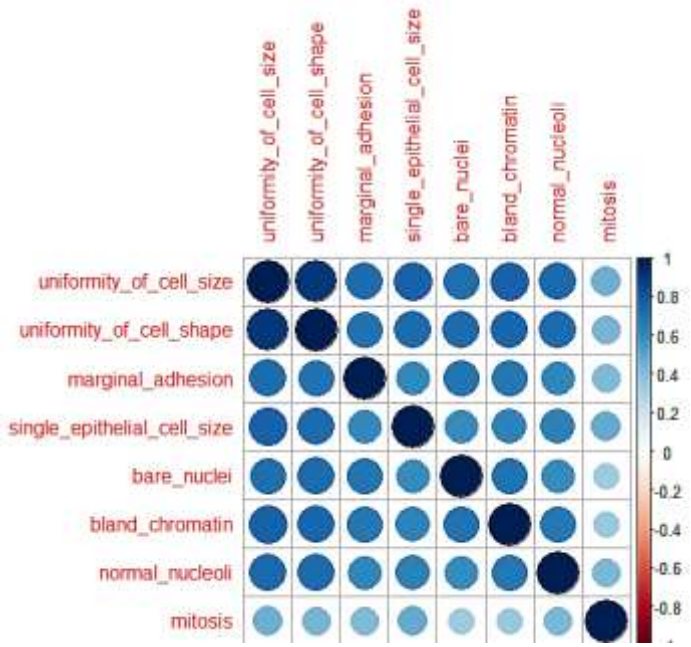
STRUCTURE OF THE DATASET:

```
> str(cancer1.data)
'data.frame':   699 obs. of  11 variables:
 $ sample_code_number         : int  1000025 1002945 1015425 1016277 1017023 1017122 1018099 1018561
 1033078 1033078 ...
 $ clump_thickness            : int  5 5 3 6 4 8 1 2 2 4 ...
 $ uniformity_of_cell_size    : int  1 4 1 8 1 10 1 1 1 2 ...
 $ uniformity_of_cell_shape   : int  1 4 1 8 1 10 1 2 1 1 ...
 $ marginal_adhesion          : int  1 5 1 1 3 8 1 1 1 1 ...
 $ single_epithelial_cell_size: int  2 7 2 3 2 7 2 2 2 2 ...
 $ bare_nuclei                : Factor w/ 11 levels "?","1","10","2",..: 2 3 4 6 2 3 3 2 2 2 ...
 $ bland_chromatin            : int  3 3 3 3 3 9 3 3 1 2 ...
 $ normal_nucleoli            : int  1 2 1 7 1 7 1 1 1 1 ...
 $ mitosis                    : int  1 1 1 1 1 1 1 1 5 1 ...
 $ classes                    : Factor w/ 2 levels "benign","malignant": 1 1 1 1 1 2 1 1 1 1 ...
```

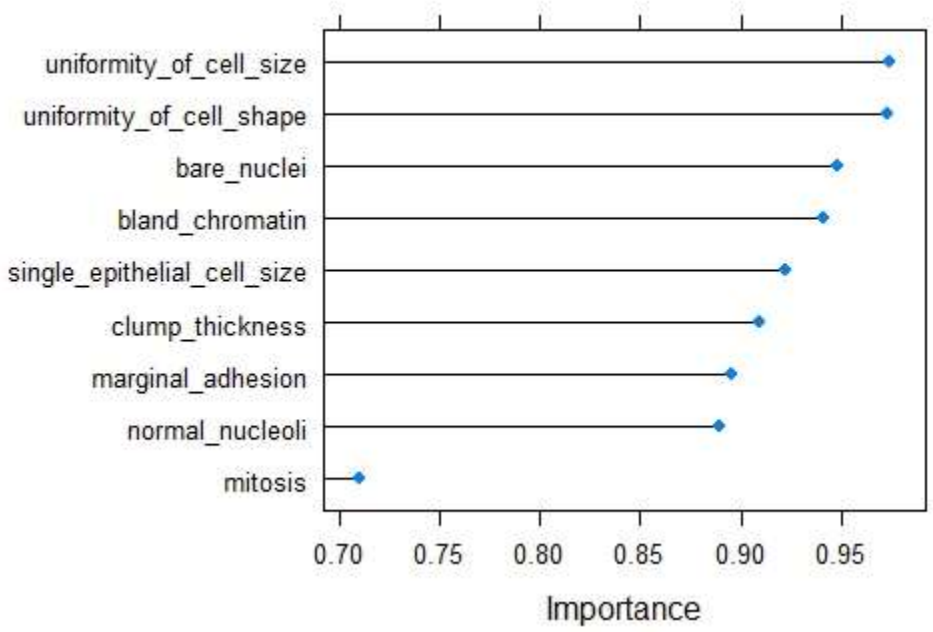
SUMMARY OF THE DATASET:

```
> summary(cancer1.data)
 sample_code_number  clump_thickness  uniformity_of_cell_size uniformity_of_cell_shape
 Min.   :   61634   Min.   : 1.000   Min.   : 1.000          Min.   : 1.000
 1st Qu.:  870688   1st Qu.: 2.000   1st Qu.: 1.000          1st Qu.: 1.000
 Median : 1171710   Median : 4.000   Median : 1.000          Median : 1.000
 Mean   : 1071704   Mean   : 4.418   Mean   : 3.134          Mean   : 3.207
 3rd Qu.: 1238298   3rd Qu.: 6.000   3rd Qu.: 5.000          3rd Qu.: 5.000
 Max.   :13454352   Max.   :10.000   Max.   :10.000          Max.   :10.000

 marginal_adhesion single_epithelial_cell_size  bare_nuclei  bland_chromatin  normal_nucleoli
 Min.   : 1.000    Min.   : 1.000              1      :402   Min.   : 1.000   Min.   : 1.000
 1st Qu.: 1.000    1st Qu.: 2.000              10     :132   1st Qu.: 2.000   1st Qu.: 1.000
 Median : 1.000    Median : 2.000              2      : 30   Median : 3.000   Median : 1.000
 Mean   : 2.807    Mean   : 3.216              5      : 30   Mean   : 3.438   Mean   : 2.867
 3rd Qu.: 4.000    3rd Qu.: 4.000              3      : 28   3rd Qu.: 5.000   3rd Qu.: 4.000
 Max.   :10.000    Max.   :10.000              (Other): 61   Max.   :10.000   Max.   :10.000
                                               NA's   : 16

    mitosis           classes
 Min.   : 1.000   benign   :458
 1st Qu.: 1.000   malignant:241
 Median : 1.000
 Mean   : 1.589
 3rd Qu.: 1.000
 Max.   :10.000
```

CORRELATION MATRIX:



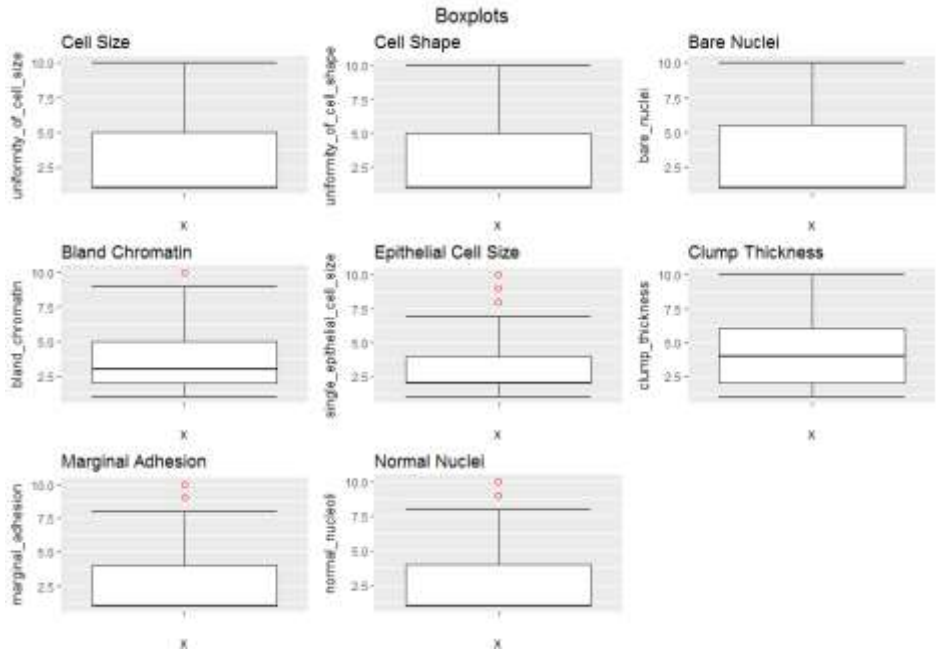
FEATURE IMPORTANCE GRAPH:



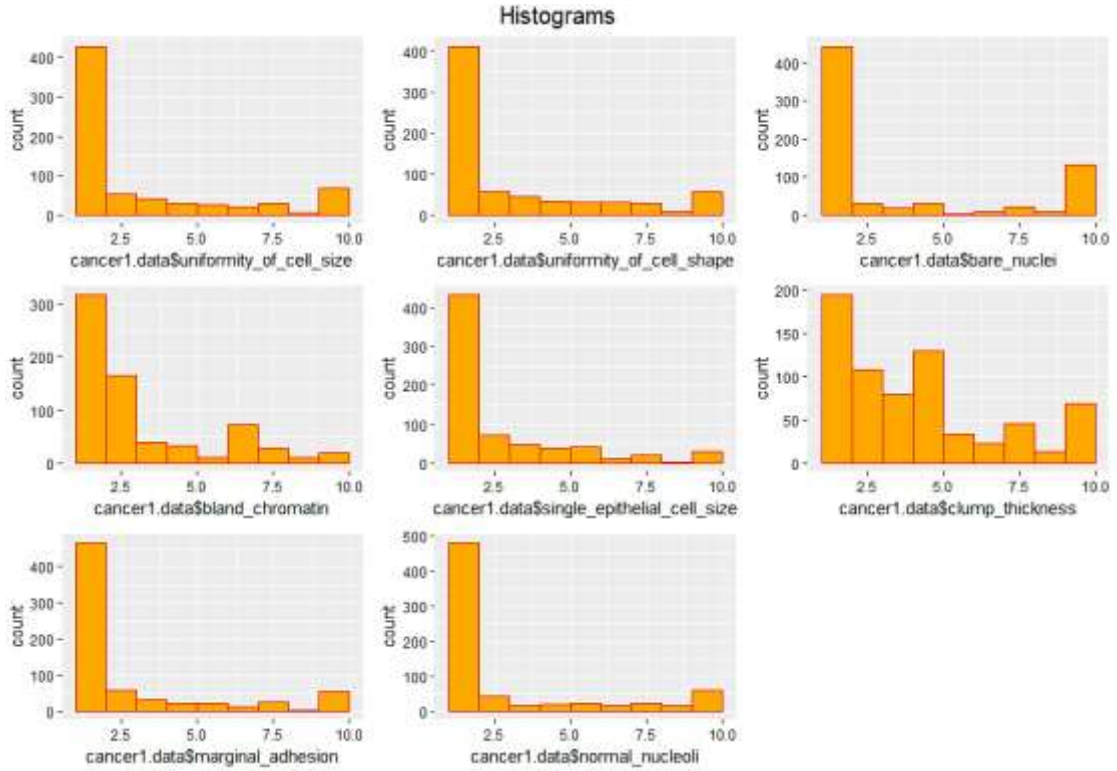
FEATURE IMPORTANCE:

```
> print(importance1)
ROC curve variable importance

                           Importance
uniformity_of_cell_size        0.9740
uniformity_of_cell_shape       0.9735
bare_nuclei                    0.9478
bland_chromatin                0.9409
single_epithelial_cell_size    0.9219
clump_thickness                0.9098
marginal_adhesion              0.8957
normal_nucleoli                0.8897
mitosis                        0.7101
```

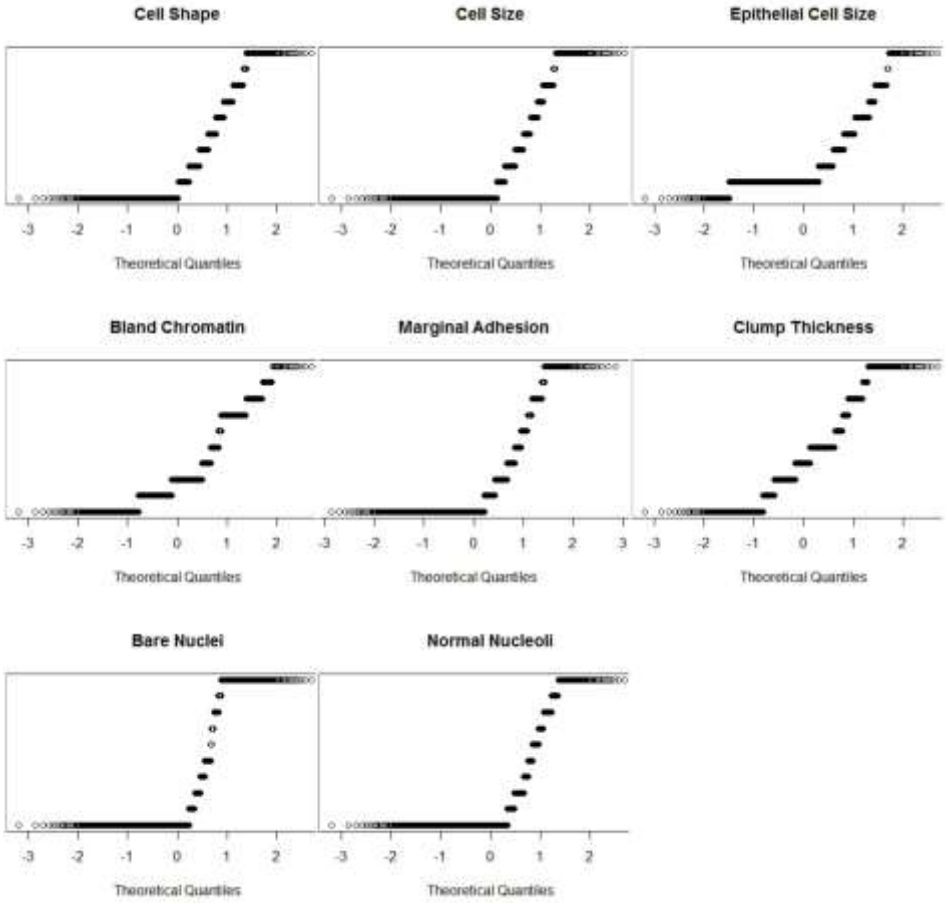
BOXPLOTS:



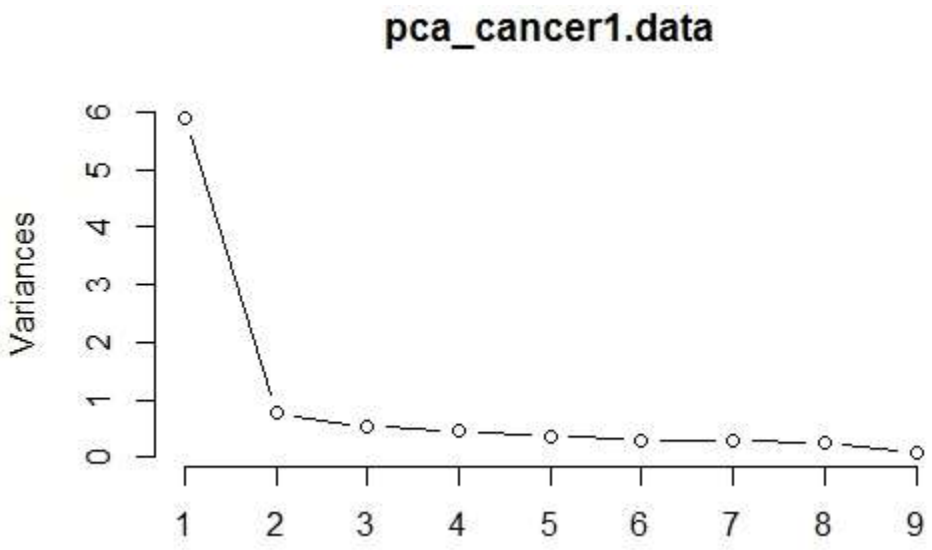
HISTOGRAMS:



QQ-PLOTS:



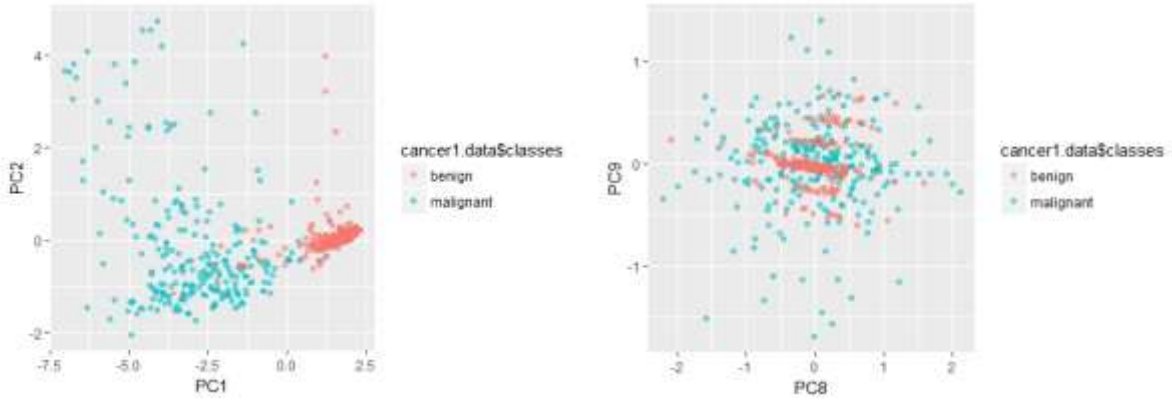
PRINCIPAL COMPONENT ANALYSIS:



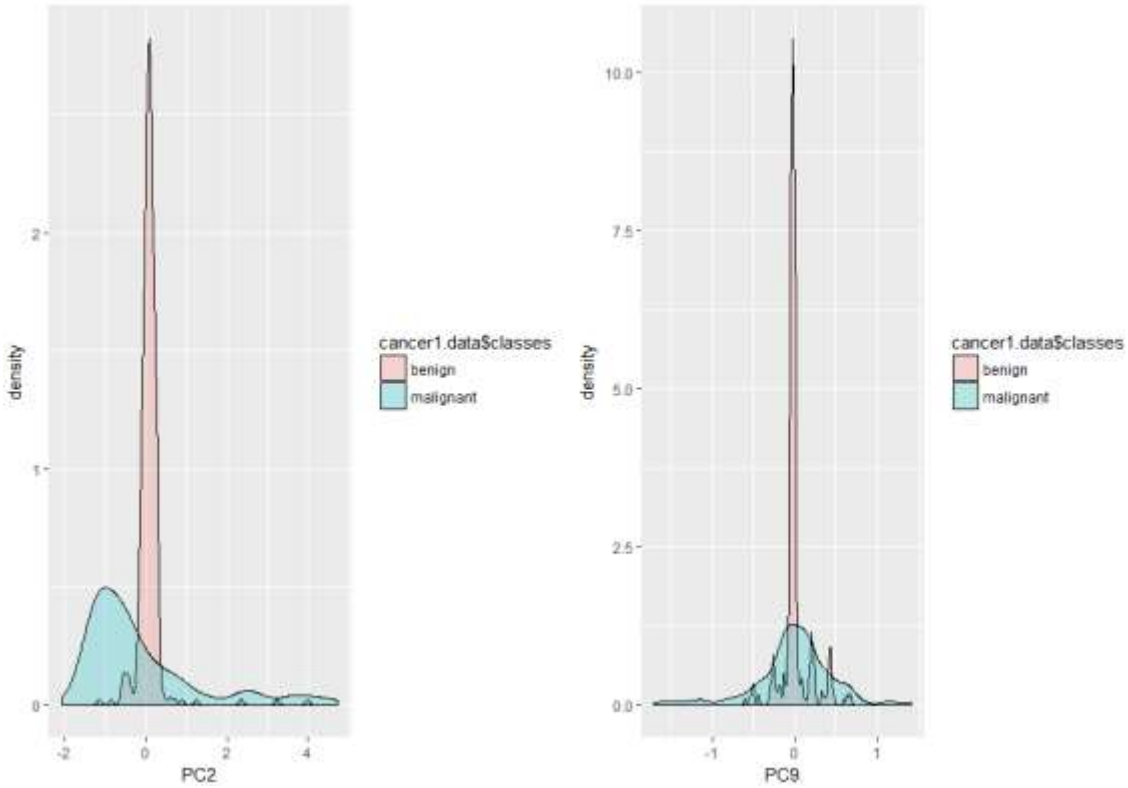
PCA SUMMARY:

```
> summary(pca_cancer1.data)
Importance of components:
                          PC1     PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9
Standard deviation     2.4290 0.88163 0.73406 0.67805 0.61591 0.54754 0.54281 0.51152 0.29790
Proportion of Variance 0.6555 0.08636 0.05987 0.05108 0.04215 0.03331 0.03274 0.02907 0.00986
Cumulative Proportion  0.6555 0.74191 0.80179 0.85287 0.89502 0.92833 0.96107 0.99014 1.00000
```

PCA PLOTS:



While PCA1 vs PCA2 plot shows that the data can be easily separated. The PCA8 vs PCA9 plot shows that the variance is better captured and the data is not so easily separable.



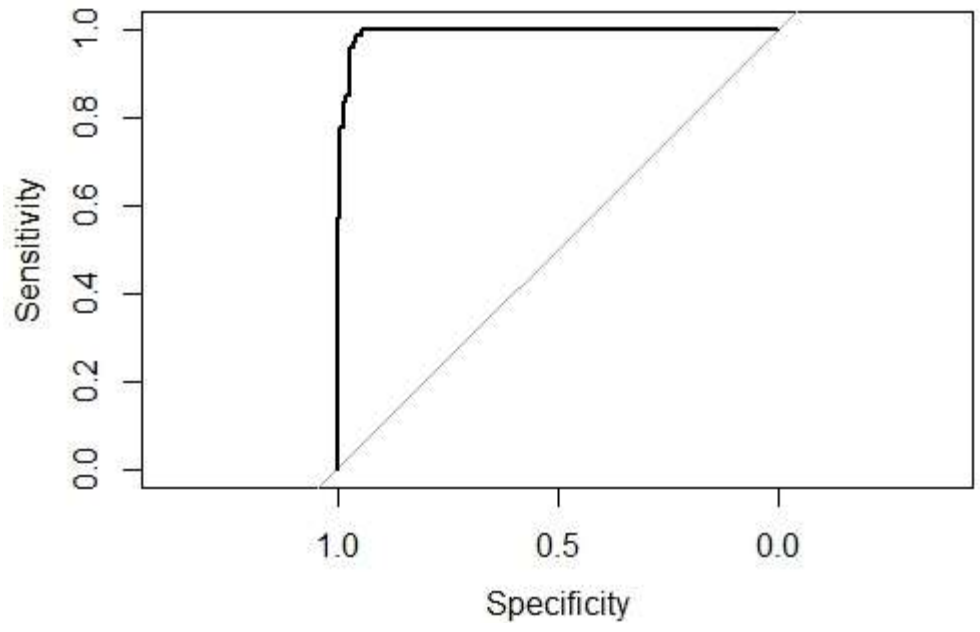
2. MACHINE LEARNING MODELS

1. The resulting data frame from the PCA is used for building models.
2. The data was split into a training set and a test set with 0.7 split data in the former.
3. As the outcome variable is a factor, 'Random Forest', 'Naïve Bayes' and 'K-Nearest Neighbor' algorithms were used to build models.
4. A 'cv (K-fold Cross Validation)' resampling method was used in the 'trainControl' for all the models.
5. The preprocessing options were set to 'center' and 'scale' with a PCA threshold of 0.99, which means that the cutoff for the cumulative percent of variance to be retained by PCA should be 0.99.
6. These models were built on the training set, and predictions were made on the test set.
7. The models used 'Receiver Operating Characteristic' curve as the evaluation metric.
8. Cross validation was performed using the confusion matrix to identify specific ROC characteristics.
9. A table of the best model for each metric was created to understand the models and choose according to specifications.
10. The three models were compared according to their ROC curve metrics and also a correlation matrix was plotted.
11. The ROC curves for each model specified the 'Area Under the Curve' (AUC). The specificity vs sensitivity graphs were also plotted.
12. A boxplot was also plotted for model comparison.

RANDOM FOREST:

```
> cancer1.rf
Random Forest

490 samples
  9 predictor
  2 classes: 'benign', 'malignant'

Pre-processing: centered (9), scaled (9)
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 392, 393, 391, 392, 392
Resampling results across tuning parameters:

  mtry  ROC        Sens       Spec
  2     0.9896016  0.9659135  0.9821747
  5     0.9889429  0.9659135  0.9643494
  9     0.9884792  0.9659135  0.9643494

ROC was used to select the optimal model using  the largest value.
The final value used for the model was mtry = 2.
```

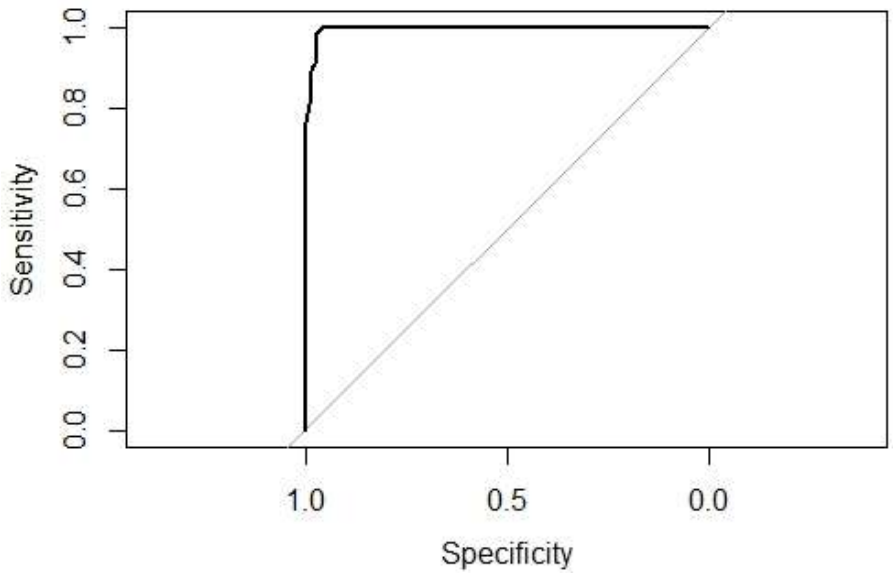
```
Call:
roc.default(response = test1.data$classes, predictor = pred_prob_rf1$malignant)

Data: pred_prob_rf1$malignant in 137 controls (test1.data$classes benign) < 72 cases (test1.data$
classes malignant).
Area under the curve: 0.9922
```



K NEAREST NEIGHBOR:

```
> cancer1.knn
k-Nearest Neighbors

490 samples
  9 predictor
  2 classes: 'benign', 'malignant'

Pre-processing: centered (9), scaled (9)
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 392, 391, 392, 393, 392
Resampling results across tuning parameters:

  k   ROC        Sens       Spec
   5  0.9797917  0.9719231  0.9463458
   7  0.9802938  0.9687981  0.9461676
   9  0.9832733  0.9687981  0.9579323
  11  0.9838346  0.9656731  0.9579323
  13  0.9833145  0.9687981  0.9639929
  15  0.9825638  0.9719231  0.9581105
  17  0.9822422  0.9750481  0.9522282
  19  0.9843393  0.9750481  0.9463458
  21  0.9844041  0.9750481  0.9463458
  23  0.9841743  0.9750481  0.9463458

ROC was used to select the optimal model using  the largest value.
The final value used for the model was k = 21.
```

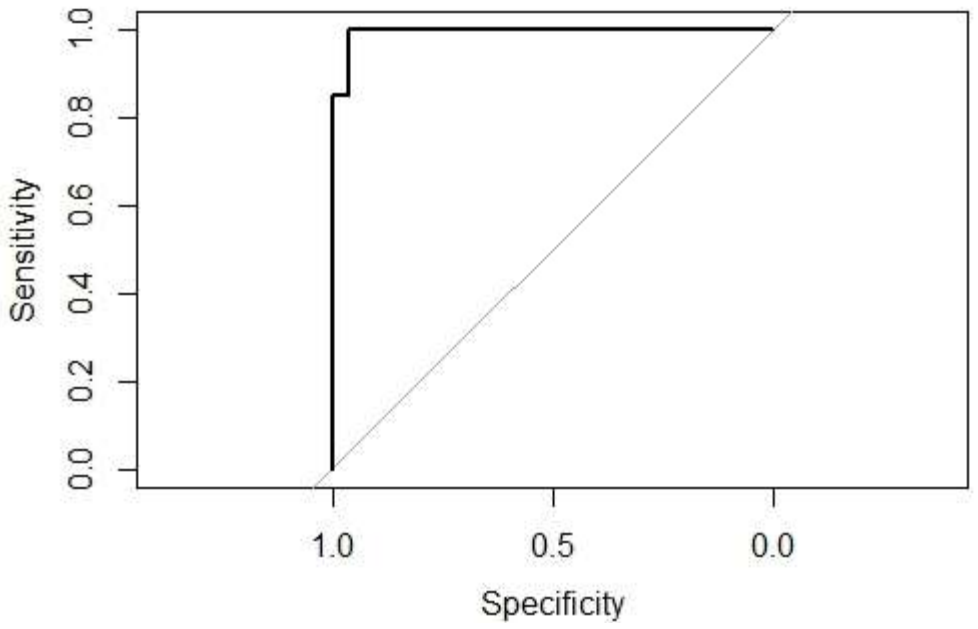
```
> roc1_knn

Call:
roc.default(response = test1.data$classes, predictor = pred_prob_knn1$malignant)

Data: pred_prob_knn1$malignant in 137 controls (test1.data$classes benign) < 72 cases (test1.data
$classes malignant).
Area under the curve: 0.9955
```



NAÏVE BAYES:

```
> cancer1.nb
Naive Bayes

490 samples
  9 predictor
  2 classes: 'benign', 'malignant'

Pre-processing: centered (9), scaled (9)
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 392, 392, 392, 392, 392
Resampling results across tuning parameters:

  usekernel  ROC        Sens       Spec
  FALSE      0.9855914  0.9502404  0.9762923
   TRUE      0.9892519  0.9689904  0.9586453

Tuning parameter 'fL' was held constant at a value of 0
Tuning parameter 'adjust' was
 held constant at a value of 1
ROC was used to select the optimal model using  the largest value.
The final values used for the model were fL = 0, usekernel = TRUE and adjust = 1.
```

```
> roc1_nb

Call:
roc.default(response = test1.data$classes, predictor = pred_prob_nb1$malignant)

Data: pred_prob_nb1$malignant in 137 controls (test1.data$classes benign) < 72 cases (test1.data$
classes malignant).
Area under the curve: 0.9944
```



CONFUSION MATRICES FOR THE THREE MODELS:

RANDOM FOREST:

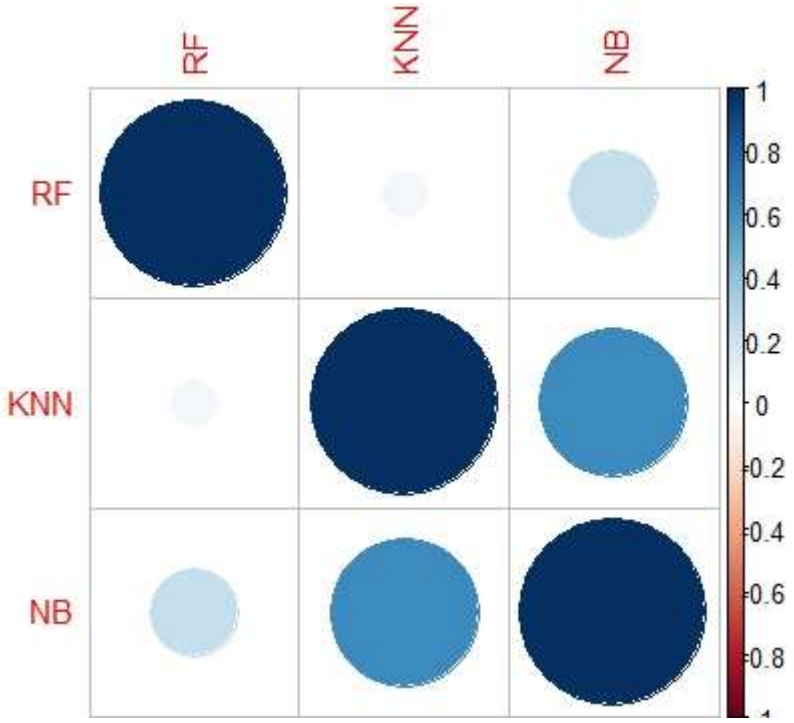
```
> cm1_rf
Confusion Matrix and Statistics

           Reference
Prediction  benign malignant
  benign       132         3
  malignant      5        69

               Accuracy : 0.9617
                 95% CI : (0.926, 0.9833)
    No Information Rate : 0.6555
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9158
 Mcnemar's Test P-Value : 0.7237

            Sensitivity : 0.9583
            Specificity : 0.9635
         Pos Pred Value : 0.9324
         Neg Pred Value : 0.9778
             Prevalence : 0.3445
         Detection Rate : 0.3301
   Detection Prevalence : 0.3541
      Balanced Accuracy : 0.9609

       'Positive' Class : malignant
```

KNN:

```
> cm1_knn
Confusion Matrix and Statistics

           Reference
Prediction  benign malignant
  benign       133         3
  malignant      4        69

               Accuracy : 0.9665
                 95% CI : (0.9322, 0.9864)
    No Information Rate : 0.6555
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9261
 Mcnemar's Test P-Value : 1

            Sensitivity : 0.9583
            Specificity : 0.9708
         Pos Pred Value : 0.9452
         Neg Pred Value : 0.9779
             Prevalence : 0.3445
         Detection Rate : 0.3301
   Detection Prevalence : 0.3493
      Balanced Accuracy : 0.9646

       'Positive' Class : malignant
```

NAÏVE BAYES:

```
> cm1_nb
Confusion Matrix and Statistics

          Reference
Prediction  benign malignant
  benign       132         6
  malignant      5        66

               Accuracy : 0.9474
                 95% CI : (0.9078, 0.9734)
    No Information Rate : 0.6555
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.8831
 Mcnemar's Test P-Value : 1

            Sensitivity : 0.9167
            Specificity : 0.9635
         Pos Pred Value : 0.9296
         Neg Pred Value : 0.9565
             Prevalence : 0.3445
         Detection Rate : 0.3158
   Detection Prevalence : 0.3397
      Balanced Accuracy : 0.9401

       'Positive' Class : malignant
```



- The accuracy is the almost equal and the highest for both ‘Random Forest’ and ‘KNN’ with 0.9617 and 0.9655 respectively.

3. DECISIONS

1. The ROC curve results reveal that
    - KNN gives the highest accuracy.
    - KNN and Random Forest give the least 'False Negative Rate (FNR)' or 'Type-II' error which is an important criteria.
    - It means that people who have a benign tumor are predicted that they have a malignant tumor. This is of grave consequence.
    - The metric for FN is 'Sensitivity', which is 0.958333 for KNN & Random Forest.
    - Type-I error or 'False Positive Rate' is highest in KNN.
    - This means that the people who have a malignant tumor are predicted to have a benign tumor. This is also not desirable, but it is not as costly an error as Type-II.
    - KNN predicts with highest precision the number of malignant cases with 0.9452 given by 'Pos Pred Value' in the results.
    - Recall is higher with KNN. It is given by 'Neg Pred Value' with 0.9779.

Depending on the importance of the situation,

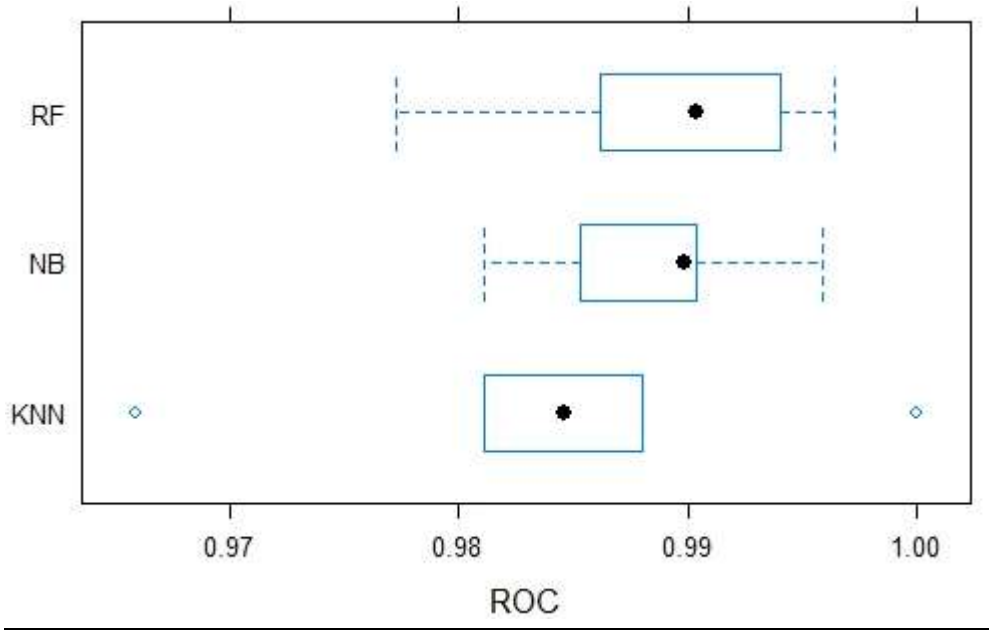
- To minimize wrong classifications for benign, Random Forest mode can be used.
- For a minimal FPR, K-Nearest Neighbor can be used.

ROC CURVE RESULTS COMPARISON FOR THE MODELS:

```
> cm_list_results
                          RF       KNN        NB
Sensitivity        0.9583333 0.9583333 0.9166667
Specificity        0.9635036 0.9708029 0.9635036
Pos Pred Value     0.9324324 0.9452055 0.9295775
Neg Pred Value     0.9777778 0.9779412 0.9565217
Precision          0.9324324 0.9452055 0.9295775
Recall             0.9583333 0.9583333 0.9166667
F1                 0.9452055 0.9517241 0.9230769
Prevalence         0.3444976 0.3444976 0.3444976
Detection Rate     0.3301435 0.3301435 0.3157895
Detection Prevalence 0.3540670 0.3492823 0.3397129
Balanced Accuracy  0.9609185 0.9645681 0.9400852
```

MAXIMUM VALUE FOR EACH ROC COMPONENT ACROSS MODELS:

```
> output_report
                  metric best_model     value
1            Sensitivity         RF 0.9583333
2            Specificity        KNN 0.9708029
3        Pos Pred Value        KNN 0.9452055
4        Neg Pred Value        KNN 0.9779412
5              Precision        KNN 0.9452055
6                 Recall         RF 0.9583333
7                     F1        KNN 0.9517241
8             Prevalence         RF 0.3444976
9         Detection Rate        KNN 0.3301435
10 Detection Prevalence         RF 0.3540670
11    Balanced Accuracy        KNN 0.9645681
```



- The ROC metrics boxplots for all the models in comparison is shown above.

WDBC DATA:

1. EXPLORATORY DATA ANALYSIS

1. The dataset had 569 observations for 32 variables with 'diagnosis' – benign (B) and malignant (M) as the dependent variable.
2. Of the 32 predictors (excluding the sample code number), 30 were numeric values and the dependable variable was a factor.
3. There were no null values in the dataset.
4. Correlation among the variables was plotted to identify highly correlated variables. Due to high correlations, ML models can fail. So PCA was used later to reduce dimensionality.
5. A feature importance graph was also plotted to identify the variables important for further analysis. It showed 20 variables as most important.
6. To identify outliers, boxplots were drawn for some of the variables. There were outliers in most of the variables. These were not removed since the dataset is too small to remove a number of observations.
7. Histograms were plotted to understand the data distribution and normality. Some of them showed a right skewed, left-skewed and a normal distribution, while others were erratic indicating no normality.
8. QQ-plots were also drawn to understand normality in data distribution. It reveals the same characteristics of the distribution as the histograms.
9. Principal Component Analysis (PCA) was performed and it gave 30 PC's with the 10^{th} PC explaining 0.95 variance in the data.

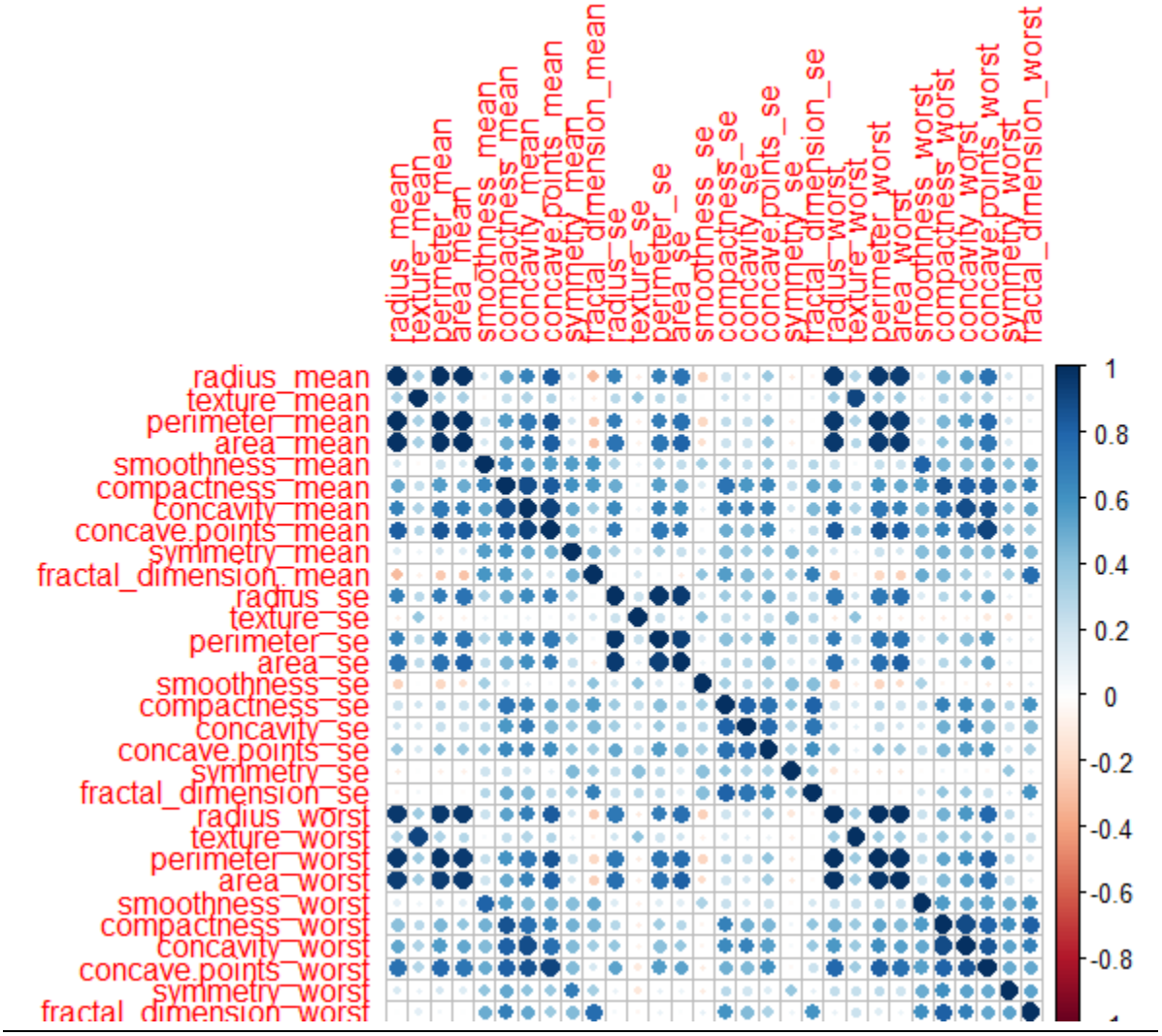
STRUCTURE OF THE DATA:

```
> str(cancer.data)
'data.frame':   569 obs. of  32 variables:
 $ id                     : int  842302 842517 84300903 84348301 84358402 843786 844359 84458202 844
981 84501001 ...
 $ diagnosis              : Factor w/ 2 levels "B","M": 2 2 2 2 2 2 2 2 2 2 ...
 $ radius_mean            : num  18 20.6 19.7 11.4 20.3 ...
 $ texture_mean           : num  10.4 17.8 21.2 20.4 14.3 ...
 $ perimeter_mean         : num  122.8 132.9 130 77.6 135.1 ...
 $ area_mean              : num  1001 1326 1203 386 1297 ...
 $ smoothness_mean        : num  0.1184 0.0847 0.1096 0.1425 0.1003 ...
 $ compactness_mean       : num  0.2776 0.0786 0.1599 0.2839 0.1328 ...
 $ concavity_mean         : num  0.3001 0.0869 0.1974 0.2414 0.198 ...
 $ concave.points_mean    : num  0.1471 0.0702 0.1279 0.1052 0.1043 ...
 $ symmetry_mean          : num  0.242 0.181 0.207 0.26 0.181 ...
 $ fractal_dimension_mean : num  0.0787 0.0567 0.06 0.0974 0.0588 ...
 $ radius_se              : num  1.095 0.543 0.746 0.496 0.757 ...
 $ texture_se             : num  0.905 0.734 0.787 1.156 0.781 ...
 $ perimeter_se           : num  8.59 3.4 4.58 3.44 5.44 ...
 $ area_se                : num  153.4 74.1 94 27.2 94.4 ...
 $ smoothness_se          : num  0.0064 0.00522 0.00615 0.00911 0.01149 ...
 $ compactness_se         : num  0.049 0.0131 0.0401 0.0746 0.0246 ...
 $ concavity_se           : num  0.0537 0.0186 0.0383 0.0566 0.0569 ...
 $ concave.points_se      : num  0.0159 0.0134 0.0206 0.0187 0.0188 ...
 $ symmetry_se            : num  0.03 0.0139 0.0225 0.0596 0.0176 ...
 $ fractal_dimension_se   : num  0.00619 0.00353 0.00457 0.00921 0.00511 ...
 $ radius_worst           : num  25.4 25 23.6 14.9 22.5 ...
 $ texture_worst          : num  17.3 23.4 25.5 26.5 16.7 ...
 $ perimeter_worst        : num  184.6 158.8 152.5 98.9 152.2 ...
 $ area_worst             : num  2019 1956 1709 568 1575 ...
 $ smoothness_worst       : num  0.162 0.124 0.144 0.21 0.137 ...
 $ compactness_worst      : num  0.666 0.187 0.424 0.866 0.205 ...
 $ concavity_worst        : num  0.712 0.242 0.45 0.687 0.4 ...
 $ concave.points_worst   : num  0.265 0.186 0.243 0.258 0.163 ...
 $ symmetry_worst         : num  0.46 0.275 0.361 0.664 0.236 ...
 $ fractal_dimension_worst: num  0.1189 0.089 0.0876 0.173 0.0768 ...
```

SUMMARY:

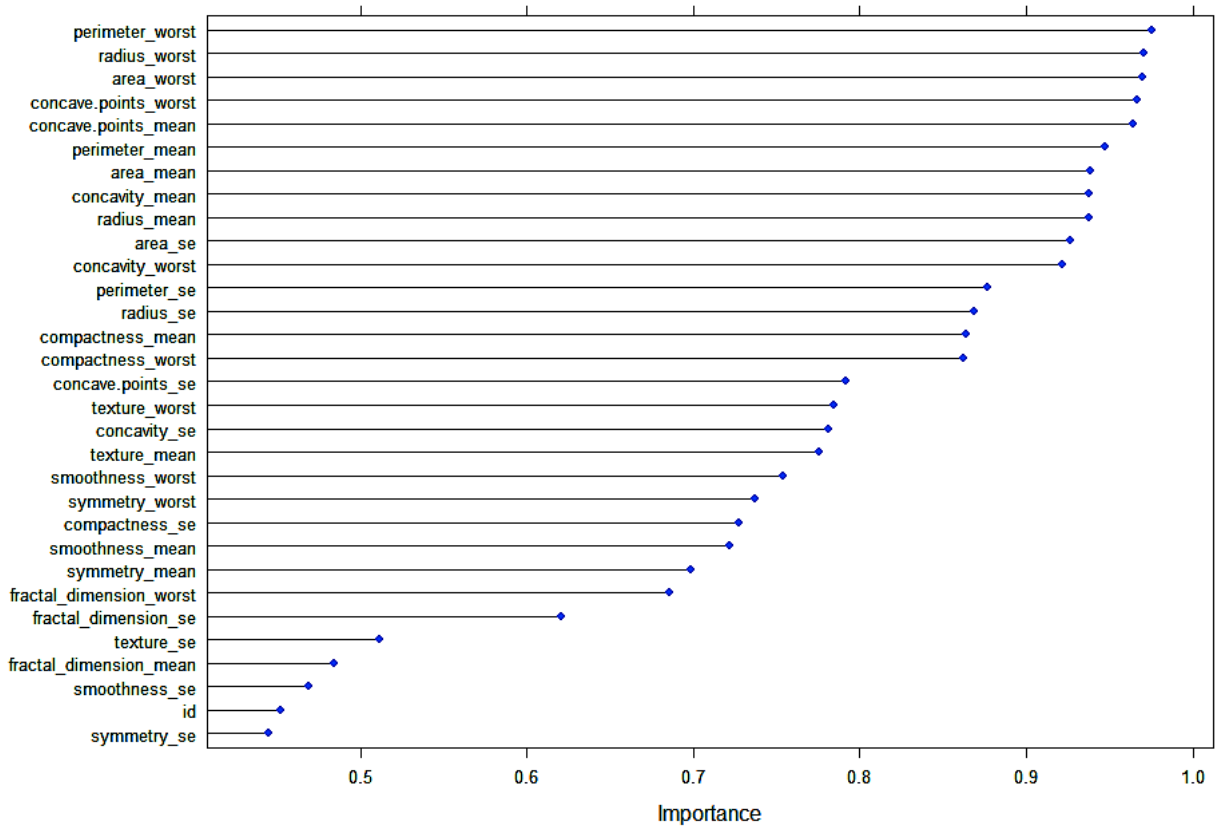
```
> summary(cancer.data)
       id         diagnosis  radius_mean      texture_mean    perimeter_mean
 Min.   :     8670   B:357   Min.   : 6.981   Min.   : 9.71   Min.   : 43.79
 1st Qu.:   869218   M:212   1st Qu.:11.700   1st Qu.:16.17   1st Qu.: 75.17
 Median :   906024           Median :13.370   Median :18.84   Median : 86.24
 Mean   : 30371831           Mean   :14.127   Mean   :19.29   Mean   : 91.97
 3rd Qu.:  8813129           3rd Qu.:15.780   3rd Qu.:21.80   3rd Qu.:104.10
 Max.   :911320502           Max.   :28.110   Max.   :39.28   Max.   :188.50
   area_mean      smoothness_mean   compactness_mean  concavity_mean    concave.points_mean
 Min.   : 143.5   Min.   :0.05263   Min.   :0.01938   Min.   :0.00000   Min.   :0.00000
 1st Qu.: 420.3   1st Qu.:0.08637   1st Qu.:0.06492   1st Qu.:0.02956   1st Qu.:0.02031
 Median : 551.1   Median :0.09587   Median :0.09263   Median :0.06154   Median :0.03350
 Mean   : 654.9   Mean   :0.09636   Mean   :0.10434   Mean   :0.08880   Mean   :0.04892
 3rd Qu.: 782.7   3rd Qu.:0.10530   3rd Qu.:0.13040   3rd Qu.:0.13070   3rd Qu.:0.07400
 Max.   :2501.0   Max.   :0.16340   Max.   :0.34540   Max.   :0.42680   Max.   :0.20120
 symmetry_mean    fractal_dimension_mean   radius_se       texture_se      perimeter_se
 Min.   :0.1060   Min.   :0.04996        Min.   :0.1115   Min.   :0.3602   Min.   : 0.757
 1st Qu.:0.1619   1st Qu.:0.05770        1st Qu.:0.2324   1st Qu.:0.8339   1st Qu.: 1.606
 Median :0.1792   Median :0.06154        Median :0.3242   Median :1.1080   Median : 2.287
 Mean   :0.1812   Mean   :0.06280        Mean   :0.4052   Mean   :1.2169   Mean   : 2.866
 3rd Qu.:0.1957   3rd Qu.:0.06612        3rd Qu.:0.4789   3rd Qu.:1.4740   3rd Qu.: 3.357
 Max.   :0.3040   Max.   :0.09744        Max.   :2.8730   Max.   :4.8850   Max.   :21.980
    area_se        smoothness_se      compactness_se     concavity_se     concave.points_se
 Min.   :  6.802   Min.   :0.001713   Min.   :0.002252   Min.   :0.00000   Min.   :0.000000
 1st Qu.: 17.850   1st Qu.:0.005169   1st Qu.:0.013080   1st Qu.:0.01509   1st Qu.:0.007638
 Median : 24.530   Median :0.006380   Median :0.020450   Median :0.02589   Median :0.010930
 Mean   : 40.337   Mean   :0.007041   Mean   :0.025478   Mean   :0.03189   Mean   :0.011796
 3rd Qu.: 45.190   3rd Qu.:0.008146   3rd Qu.:0.032450   3rd Qu.:0.04205   3rd Qu.:0.014710
 Max.   :542.200   Max.   :0.031130   Max.   :0.135400   Max.   :0.39600   Max.   :0.052790
  symmetry_se      fractal_dimension_se  radius_worst    texture_worst    perimeter_worst
 Min.   :0.007882  Min.   :0.0008948   Min.   : 7.93   Min.   :12.02   Min.   : 50.41
 1st Qu.:0.015160  1st Qu.:0.0022480   1st Qu.:13.01   1st Qu.:21.08   1st Qu.: 84.11
 Median :0.018730  Median :0.0031870   Median :14.97   Median :25.41   Median : 97.66
 Mean   :0.020542  Mean   :0.0037949   Mean   :16.27   Mean   :25.68   Mean   :107.26
 3rd Qu.:0.023480  3rd Qu.:0.0045580   3rd Qu.:18.79   3rd Qu.:29.72   3rd Qu.:125.40
 Max.   :0.078950  Max.   :0.0298400   Max.   :36.04   Max.   :49.54   Max.   :251.20
   area_worst     smoothness_worst  compactness_worst concavity_worst  concave.points_worst
 Min.   : 185.2   Min.   :0.07117   Min.   :0.02729   Min.   :0.0000   Min.   :0.00000
 1st Qu.: 515.3   1st Qu.:0.11660   1st Qu.:0.14720   1st Qu.:0.1145   1st Qu.:0.06493
```

CORRELATION AMONGST THE VARIABLES:



Some variables are highly correlated which may create problems while modeling due to interactions.

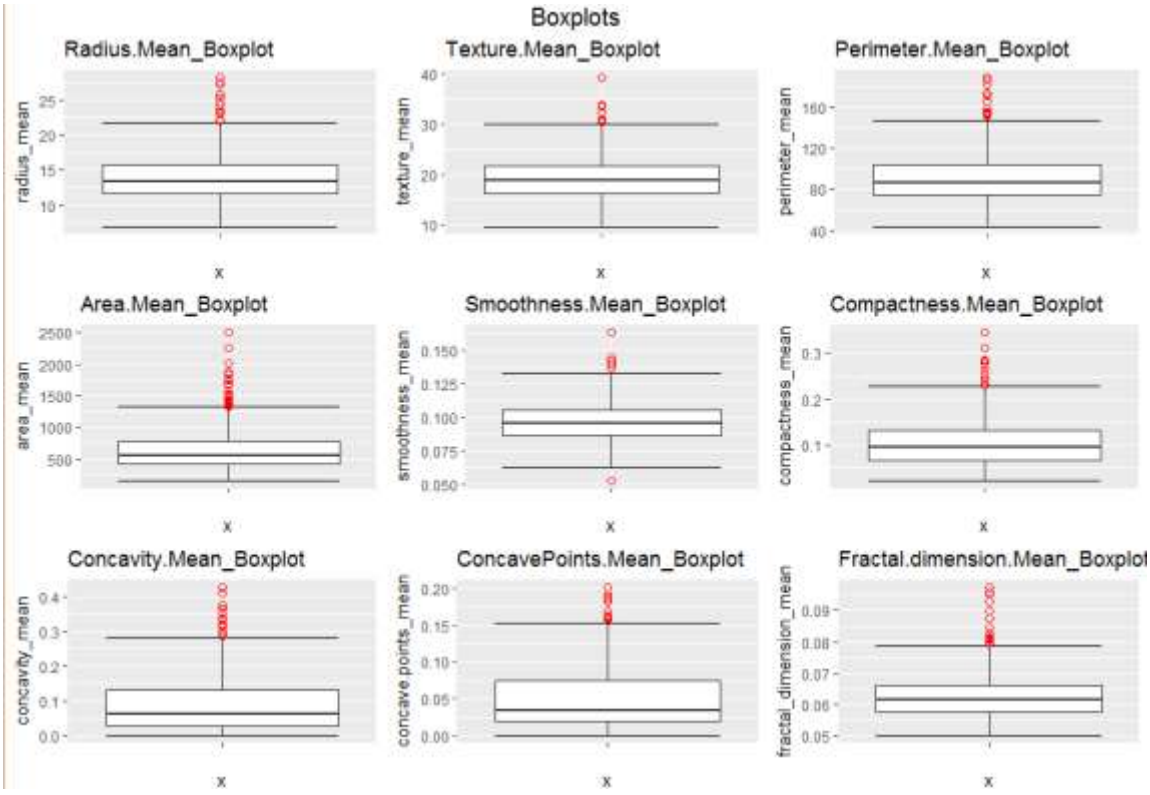
FEATURE IMPORTANCE GRAPH:



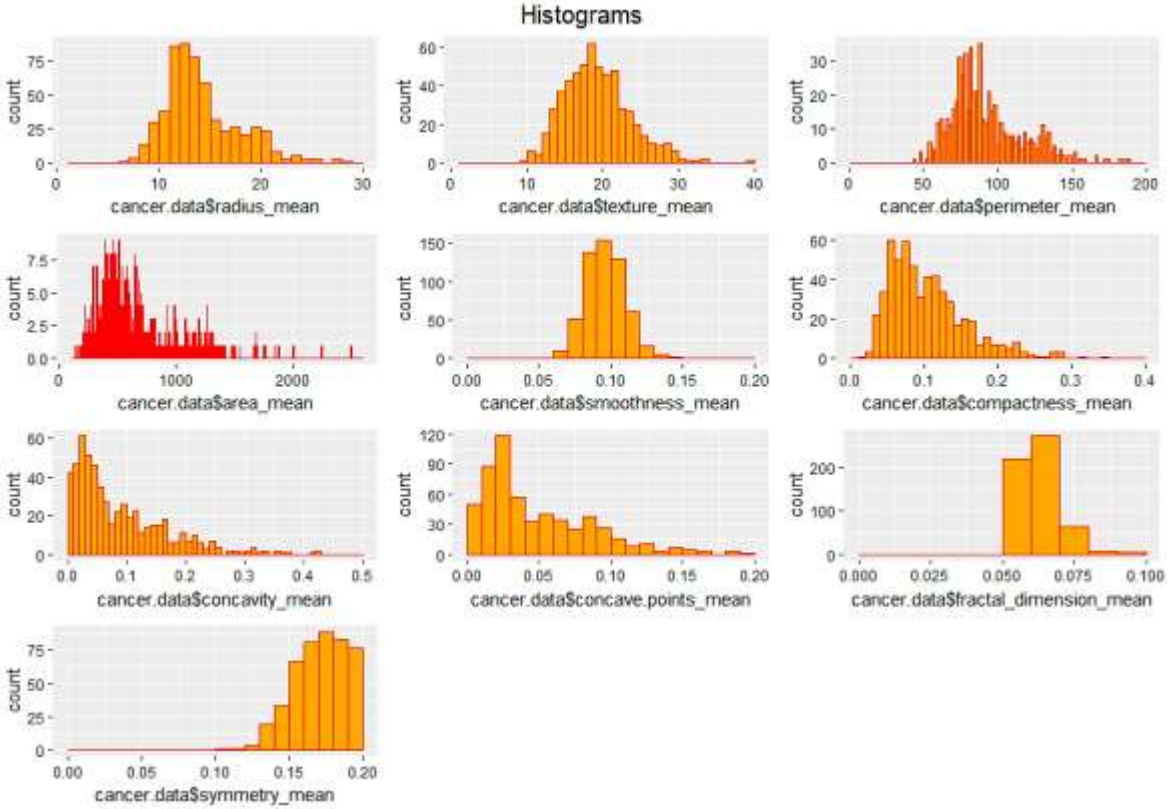
FEATURES:

```
> print(importance)
ROC curve variable importance

  only 20 most important variables shown (out of 31)

                       Importance
perimeter_worst            0.9755
radius_worst               0.9704
area_worst                 0.9698
concave.points_worst       0.9667
concave.points_mean        0.9644
perimeter_mean             0.9469
area_mean                  0.9383
concavity_mean             0.9378
radius_mean                0.9375
area_se                    0.9264
concavity_worst            0.9214
perimeter_se               0.8764
radius_se                  0.8683
compactness_mean           0.8638
compactness_worst          0.8623
concave.points_se          0.7918
texture_worst              0.7846
concavity_se               0.7808
texture_mean               0.7758
smoothness_worst           0.7541
```

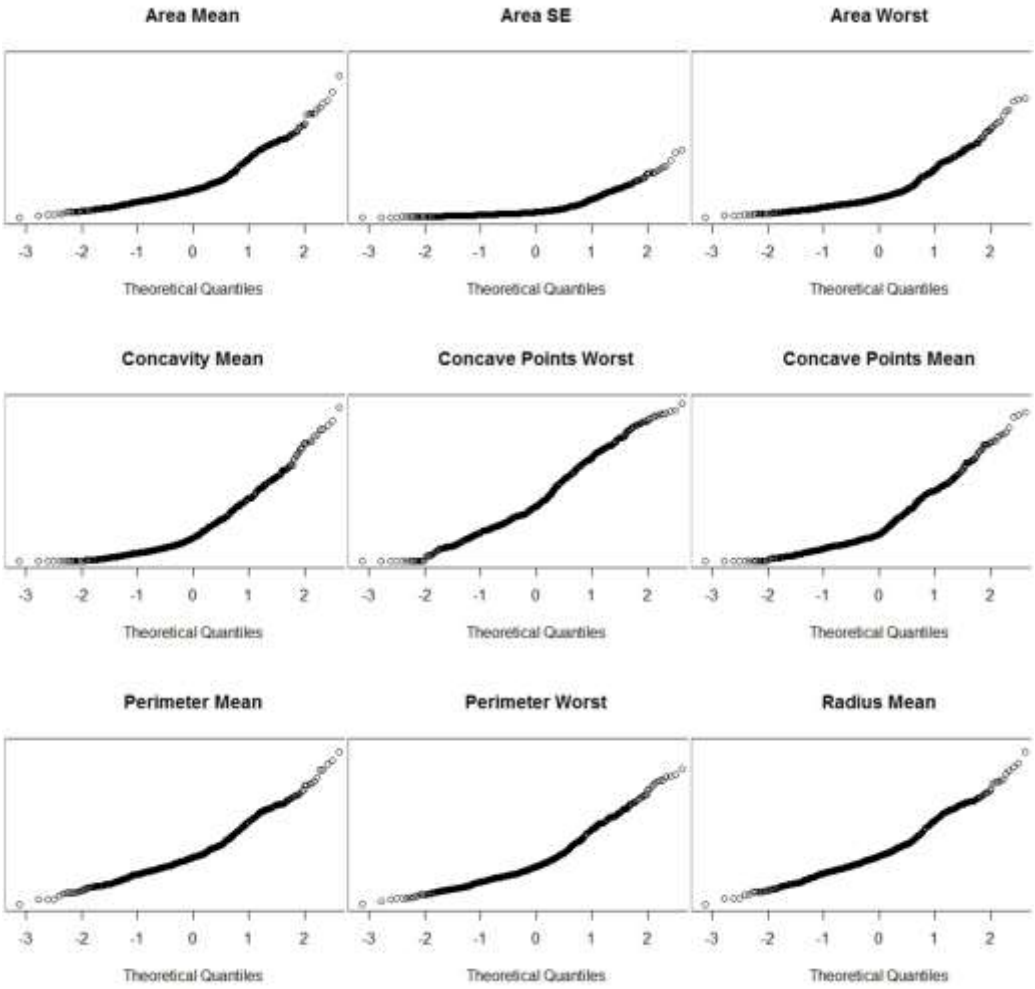
BOXPLOTS:



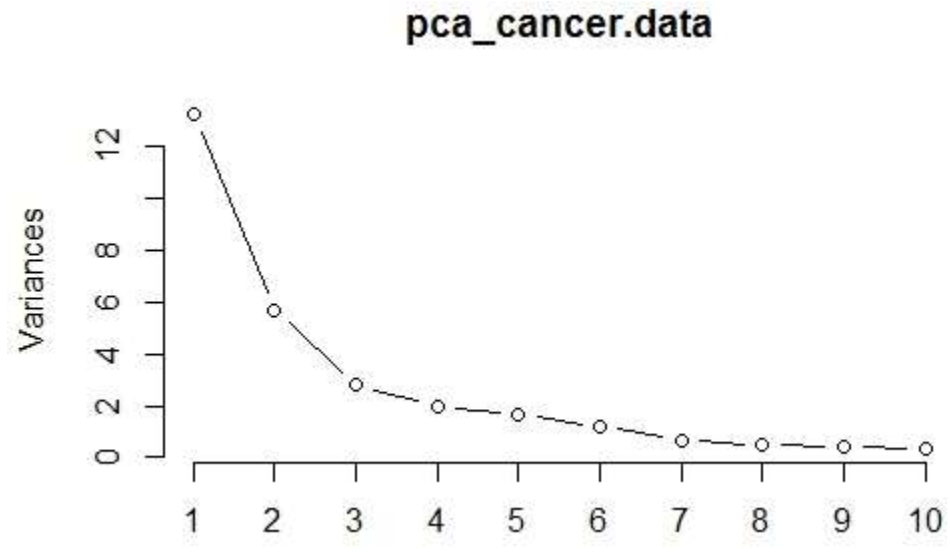
HISTOGRAMS:



QQ-PLOTS:



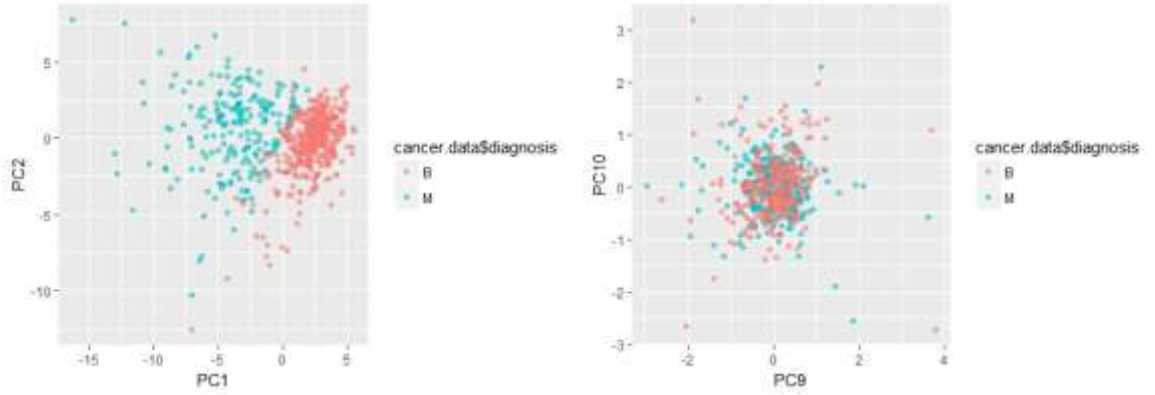
PRINCIPAL COMPONENT ANALYSIS:

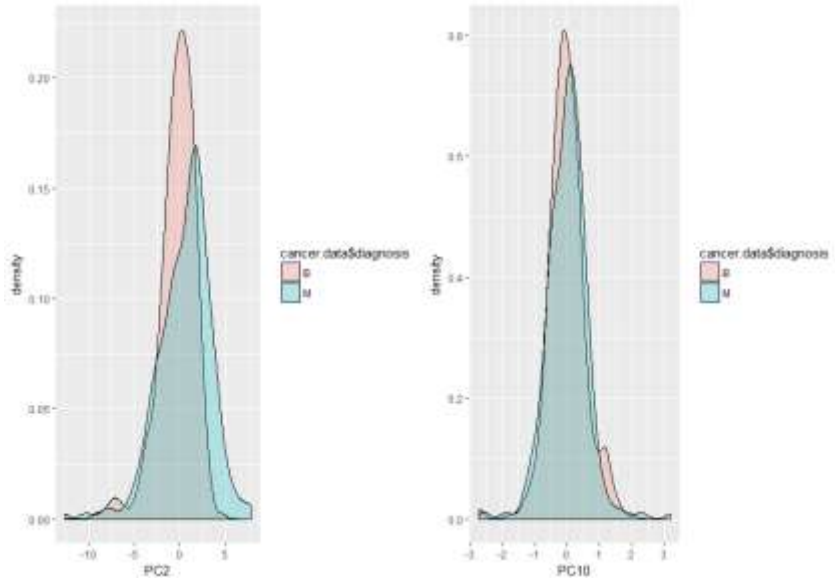


PCA RESULTS:

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

PCA GRAPHS:





While PCA1 vs PCA2 plot shows that the data can be easily separated. The PCA9 vs PCA10 plot shows that the variance is better captured and the data is not so easily separable.

2. MACHINE LEARNING MODELS:

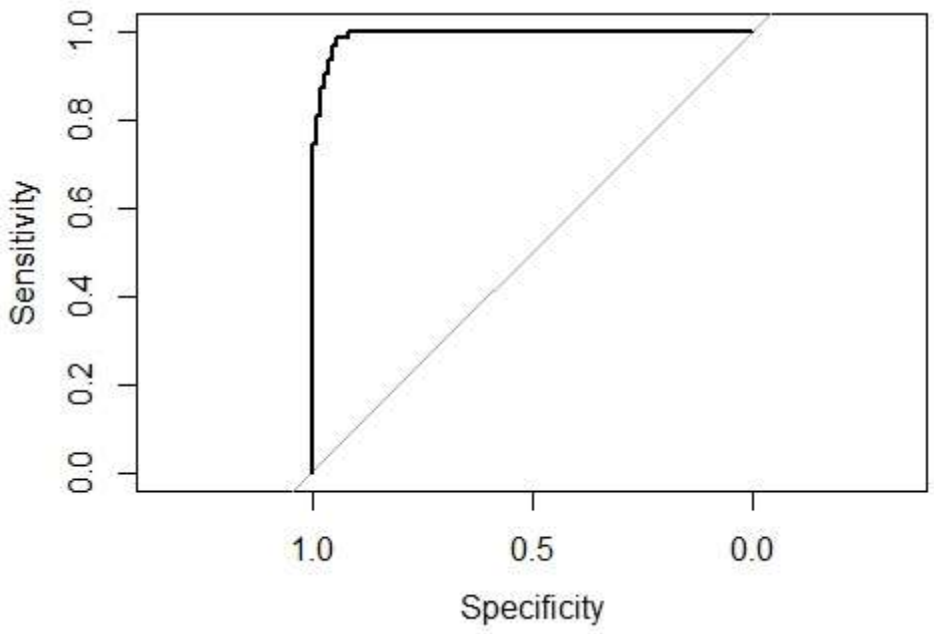
1. The resulting data frame from the PCA is used for building models.
2. The data was split into a training set and a test set with 0.7 split data in the former.
3. As the outcome variable is a factor, 'Random Forest', 'Naïve Bayes' and 'K-Nearest Neighbor' algorithms were used to build models.
4. A 'cv (K-fold Cross Validation)' resampling method was used in the 'trainControl' for all the models.
5. The preprocessing options were set to 'center' and 'scale' with a PCA threshold of 0.99, which means that the cutoff for the cumulative percent of variance to be retained by PCA should be 0.99.
6. These models were built on the training set, and predictions were made on the test set.
7. The models used 'Receiver Operating Characteristic' curve as the evaluation metric.
8. Cross validation was performed using the confusion matrix to identify specific ROC characteristics.
9. A table of the best model for each metric was created to understand the models and choose according to specifications.
10. The three models were compared according to their ROC curve metrics and also a correlation matrix was plotted.
11. The ROC curves for each model specified the 'Area Under the Curve' (AUC). The specificity vs sensitivity graphs were also plotted.
12. A boxplot was also plotted for model comparison.

RANDOM FOREST:

```
> cancer.rf
Random Forest

399 samples
 30 predictor
  2 classes: 'B', 'M'

Pre-processing: centered (30), scaled (30)
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 319, 319, 319, 320, 319
Resampling results across tuning parameters:

  mtry  ROC        Sens   Spec
   2    0.9890667  0.976  0.9200000
  16    0.9848575  0.972  0.9200000
  30    0.9843149  0.964  0.9197701

ROC was used to select the optimal model using  the largest value.
The final value used for the model was mtry = 2.
```

ROC CURVE RESULTS:

```
Call:
roc.default(response = test.data$diagnosis, predictor = pred_prob_rf$M)

Data: pred_prob_rf$M in 107 controls (test.data$diagnosis B) < 63 cases (test.data$diagnosis M)
Area under the curve: 0.9924
```

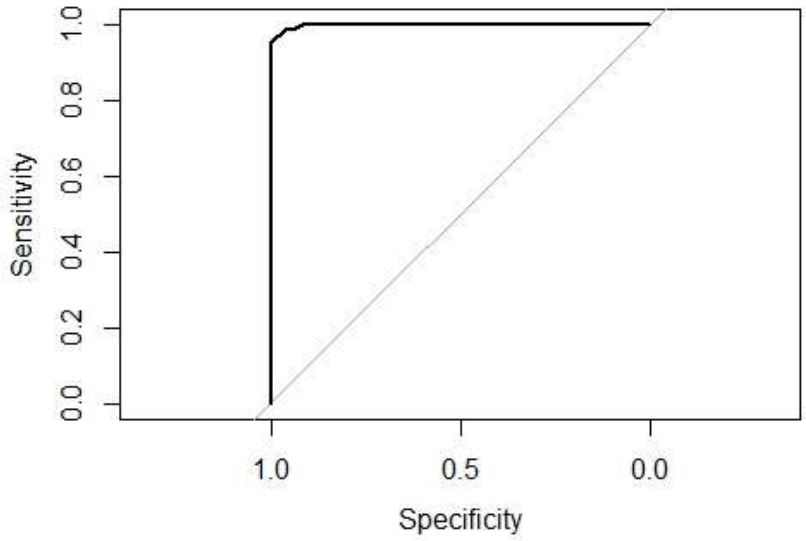


K-NEAREST NEIGHBOR:

```
> cancer.knn
k-Nearest Neighbors

399 samples
 30 predictor
  2 classes: 'B', 'M'

Pre-processing: centered (30), scaled (30)
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 319, 320, 319, 319, 319
Resampling results across tuning parameters:

  k   ROC        Sens   Spec
   5  0.9839126  0.992  0.8924138
   7  0.9862345  0.996  0.8790805
   9  0.9849770  0.992  0.8924138
  11  0.9881126  0.996  0.8926437
  13  0.9897724  0.996  0.8926437
  15  0.9881701  0.996  0.8926437
  17  0.9882391  1.000  0.8859770
  19  0.9878345  0.996  0.8857471
  21  0.9888322  0.996  0.8859770
  23  0.9884989  0.996  0.8790805

ROC was used to select the optimal model using  the largest value.
The final value used for the model was k = 13.
```

ROC CURVE RESULTS:

```
Call:
roc.default(response = test.data$diagnosis, predictor = pred_prob_knn$M)

Data: pred_prob_knn$M in 107 controls (test.data$diagnosis B) < 63 cases (test.data$diagnosis M).
Area under the curve: 0.9983
> plot(roc_knn)
```

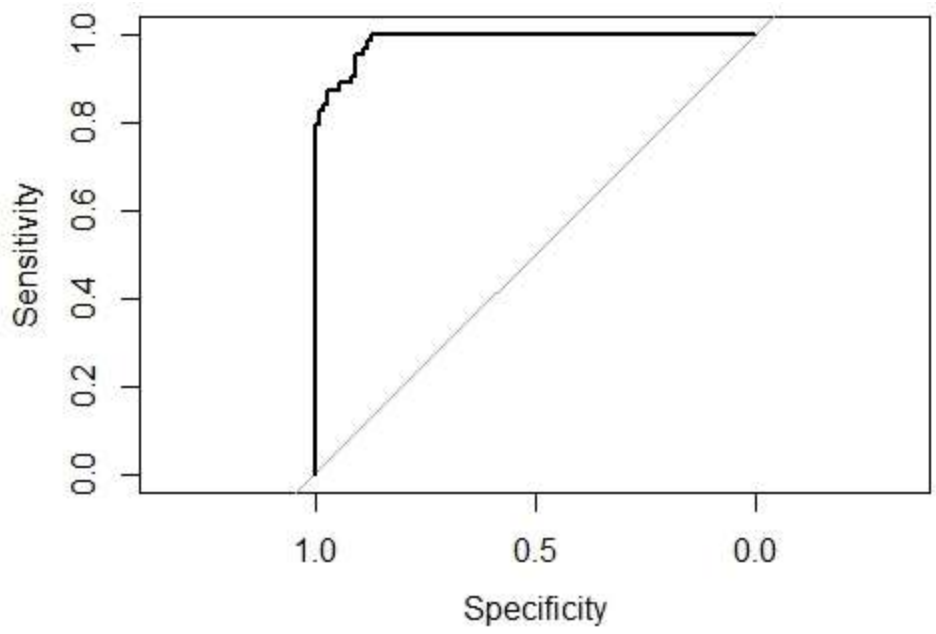


NAÏVE BAYES:

```
> cancer.nb
Naive Bayes

399 samples
 30 predictor
  2 classes: 'B', 'M'

Pre-processing: centered (30), scaled (30)
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 319, 319, 319, 319, 320
Resampling results across tuning parameters:

  usekernel  ROC        Sens   Spec
  FALSE      0.9875540  0.956  0.9062069
   TRUE      0.9825701  0.952  0.9195402

Tuning parameter 'fL' was held constant at a value of 0
Tuning parameter 'adjust' was
 held constant at a value of 1
ROC was used to select the optimal model using  the largest value.
The final values used for the model were fL = 0, usekernel = FALSE and adjust = 1.
```

ROC CURVE RESULTS:

```
> roc_nb

Call:
roc.default(response = test.data$diagnosis, predictor = pred_prob_nb$M)

Data: pred_prob_nb$M in 107 controls (test.data$diagnosis B) < 63 cases (test.data$diagnosis M).
Area under the curve: 0.9861
```



CONFUSION MATRICES FOR THE THREE MODELS:

RANDOM FOREST:

```
> cm_rf
Confusion Matrix and Statistics

          Reference
Prediction   B   M
         B 103   5
         M   4  58

               Accuracy : 0.9471
                 95% CI : (0.9019, 0.9755)
    No Information Rate : 0.6294
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.8861
 Mcnemar's Test P-Value : 1

            Sensitivity : 0.9206
            Specificity : 0.9626
         Pos Pred Value : 0.9355
         Neg Pred Value : 0.9537
             Prevalence : 0.3706
         Detection Rate : 0.3412
   Detection Prevalence : 0.3647
      Balanced Accuracy : 0.9416

       'Positive' Class : M
```

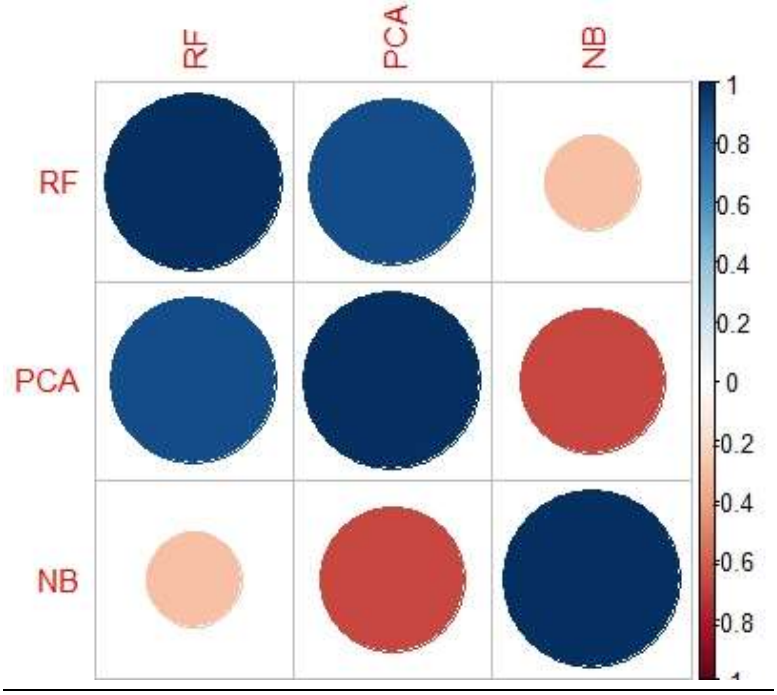
K-NEAREST NEIGHBOR:

```
> cm_knn
Confusion Matrix and Statistics

          Reference
Prediction   B   M
         B 107   6
         M   0  57

               Accuracy : 0.9647
                 95% CI : (0.9248, 0.9869)
    No Information Rate : 0.6294
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.9228
 Mcnemar's Test P-Value : 0.04123

            Sensitivity : 0.9048
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 0.9469
             Prevalence : 0.3706
         Detection Rate : 0.3353
   Detection Prevalence : 0.3353
      Balanced Accuracy : 0.9524

       'Positive' Class : M
```

NAÏVE BAYES:

```
> cm_nb
Confusion Matrix and Statistics

          Reference
Prediction  B  M
         B 99  7
         M  8 56

               Accuracy : 0.9118
                 95% CI : (0.8586, 0.9498)
    No Information Rate : 0.6294
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.8115
 Mcnemar's Test P-Value : 1

            Sensitivity : 0.8889
            Specificity : 0.9252
         Pos Pred Value : 0.8750
         Neg Pred Value : 0.9340
             Prevalence : 0.3706
         Detection Rate : 0.3294
   Detection Prevalence : 0.3765
      Balanced Accuracy : 0.9071

       'Positive' Class : M
```

CORRELATION MATRIX FOR THE MODELS:



The accuracy is the almost equal and the highest for 'KNN' with 0.9647 followed by 'Random forest' with 0.9471.

4. DECISIONS

2. The ROC curve results reveal that
    - KNN gives the highest accuracy.
    - KNN and Random Forest give the least 'False Negative Rate (FNR)' or 'Type-II' error which is an important criteria.
    - It means that people who have a benign tumor are predicted that they have a malignant tumor. This is of grave consequence.
    - The metric for FN is 'Sensitivity', which is 0.921 for KNN & Random Forest.
    - Type-I error or 'False Positive Rate' is highest in KNN.
    - This means that the people who have a malignant tumor are predicted to have a benign tumor. This is also not desirable, but it is not as costly an error as Type-II.
    - KNN predicts with highest precision the number of malignant cases with 0.983 given by 'Pos Pred Value' in the results.
    - Recall is higher with KNN. It is given by 'Neg Pred Value' with 0.9549.

Depending on the importance of the situation,

- To minimize wrong classifications for benign, Random Forest mode can be used.
- For a minimal FPR, K-Nearest Neighbor can be used.

ROC METRICS COMPARISON ACROSS MODELS:

```
> cm_list_results1
                          RF        KNN        NB
Sensitivity        0.9206349 0.9206349 0.8888889
Specificity        0.9626168 0.9906542 0.9252336
Pos Pred Value     0.9354839 0.9830508 0.8750000
Neg Pred Value     0.9537037 0.9549550 0.9339623
Precision          0.9354839 0.9830508 0.8750000
Recall             0.9206349 0.9206349 0.8888889
F1                 0.9280000 0.9508197 0.8818898
Prevalence         0.3705882 0.3705882 0.3705882
Detection Rate     0.3411765 0.3411765 0.3294118
Detection Prevalence 0.3647059 0.3470588 0.3764706
Balanced Accuracy  0.9416259 0.9556446 0.9070613
```

ROC MODEL COMPARISON – MAXIMUM:

```
> output_report1
                  metric best_model     value
1            Sensitivity         RF 0.9206349
2            Specificity        KNN 0.9906542
3        Pos Pred Value        KNN 0.9830508
4        Neg Pred Value        KNN 0.9549550
5             Precision        KNN 0.9830508
6                Recall         RF 0.9206349
7                    F1        KNN 0.9508197
8            Prevalence         NB 0.3705882
9        Detection Rate         RF 0.3411765
10 Detection Prevalence         NB 0.3764706
11    Balanced Accuracy        KNN 0.9556446
```

BOXPLOTS FOR ROC METRICS OF THE THREE MODELS: