PROSTATE CANCER PROJECT

Graziana Capurso Mat. 2097099

Daniela Di Labbio Mat. 2091677

Agata Garbin Mat. 2072693

Project Objective

Aim : In this project we present a binary classification problem based on the "Prostate Cancer Dataset", freely available on Kaggle

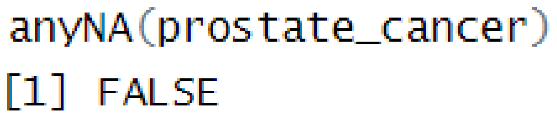
Dataset: The dataset consists of 10 features and 100 instances corresponding to the patients that have been analyzed.

Features

- ID
- Diagnosis result
- Radius
- Texture
- Perimeter
- Area
- Smoothness
- Compactness
- Symmetry
- Fractal Dimension

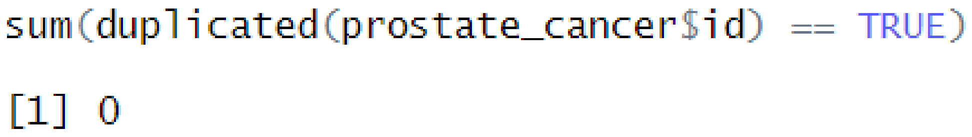
Data Preprocessing

- We check if there are any missing values.



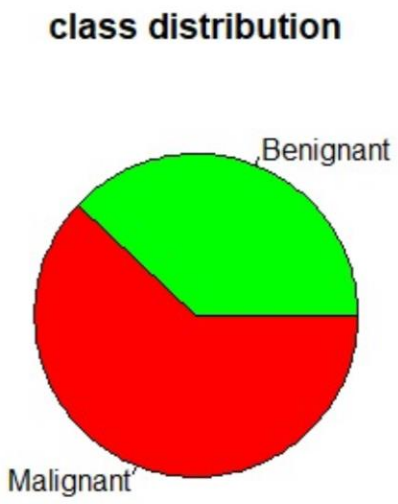
```
anyNA(prostate_cancer)
[1] FALSE
```

- We then check if any of the row is duplicated.



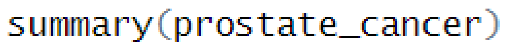
```
sum(duplicated(prostate_cancer$id) == TRUE)

[1] 0
```

- Since each row is unique, we get rid of the first column, referring to the ID of the patients, to have a dataset with only relevant information.
- The classes Benignant and Malignant are respectively 38 % of benignant and 62% of malignant.

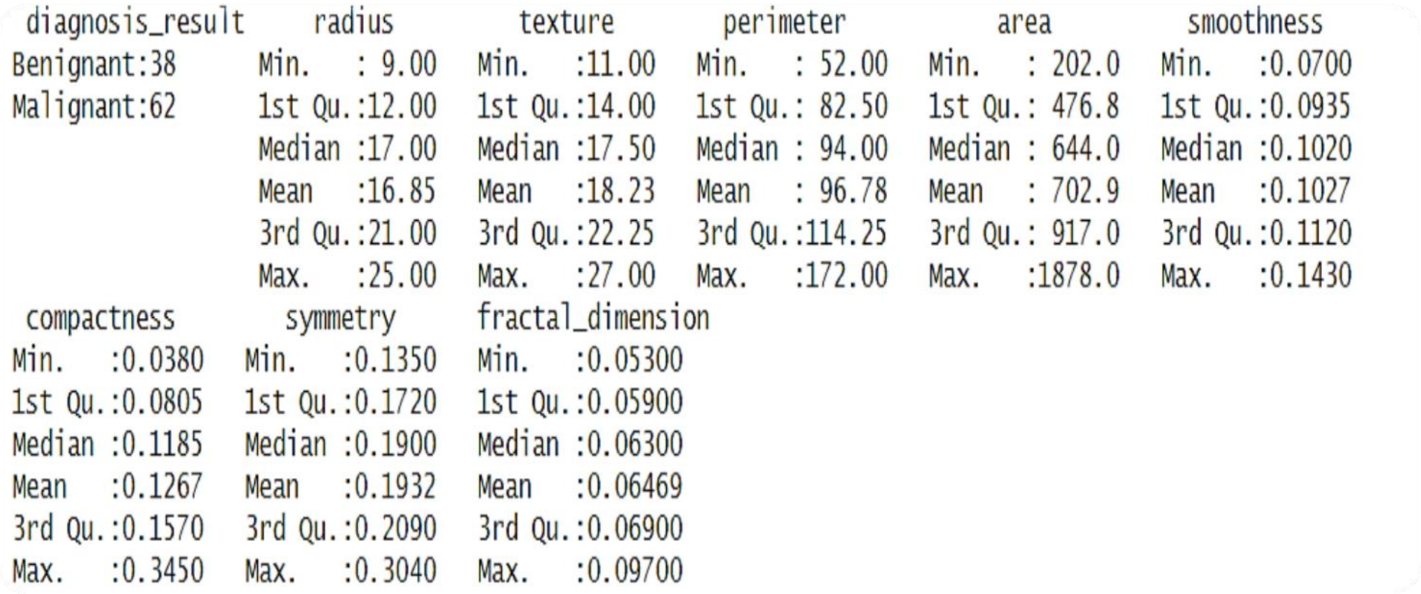


Data Analysis

We produce a summary of our dataset to see how the different values of every features were distributed, and then we calculate the variance and the standard deviation of each variables using the functions var() and sd().



```
summary(prostate_cancer)
```

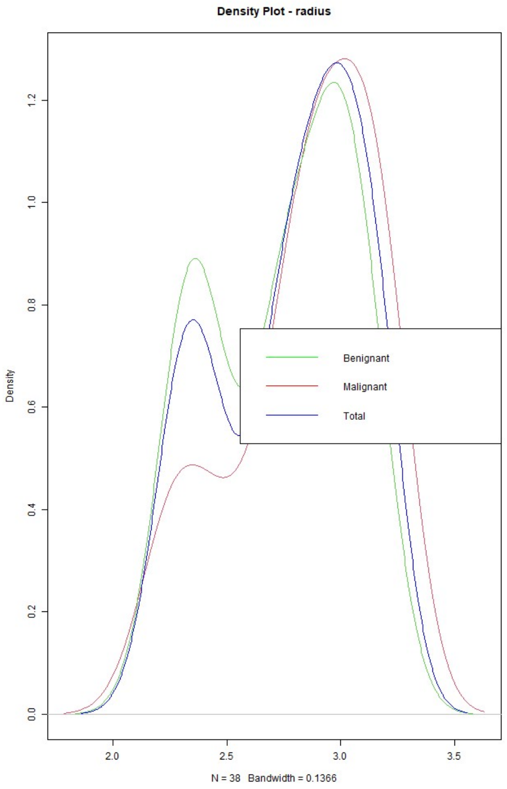


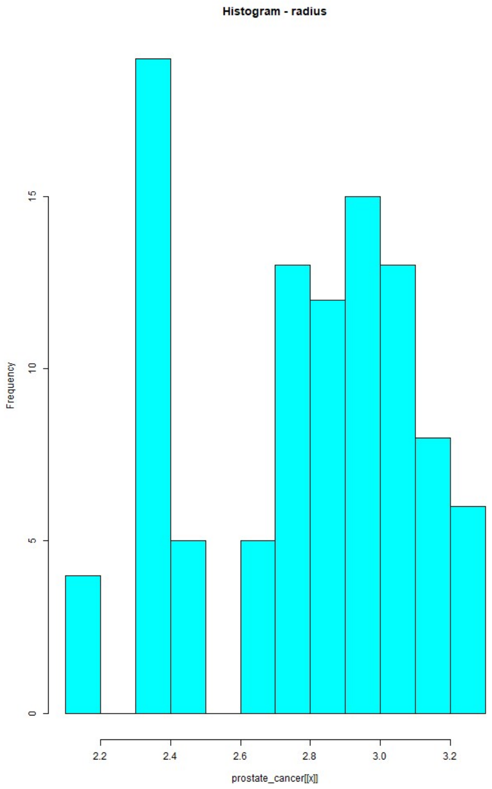
```
 diagnosis_result     radius         texture         perimeter           area          smoothness    
Benignant:38      Min.   : 9.00   Min.   :11.00   Min.   : 52.00   Min.   : 202.0   Min.   :0.0700  
Malignant:62      1st Qu.:12.00   1st Qu.:14.00   1st Qu.: 82.50   1st Qu.: 476.8   1st Qu.:0.0935  
                  Median :17.00   Median :17.50   Median : 94.00   Median : 644.0   Median :0.1020  
                  Mean   :16.85   Mean   :18.23   Mean   : 96.78   Mean   : 702.9   Mean   :0.1027  
                  3rd Qu.:21.00   3rd Qu.:22.25   3rd Qu.:114.25   3rd Qu.: 917.0   3rd Qu.:0.1120  
                  Max.   :25.00   Max.   :27.00   Max.   :172.00   Max.   :1878.0   Max.   :0.1430  
 compactness        symmetry      fractal_dimension
Min.   :0.0380   Min.   :0.1350   Min.   :0.05300  
1st Qu.:0.0805   1st Qu.:0.1720   1st Qu.:0.05900  
Median :0.1185   Median :0.1900   Median :0.06300  
Mean   :0.1267   Mean   :0.1932   Mean   :0.06469  
3rd Qu.:0.1570   3rd Qu.:0.2090   3rd Qu.:0.06900  
Max.   :0.3450   Max.   :0.3040   Max.   :0.09700  
```

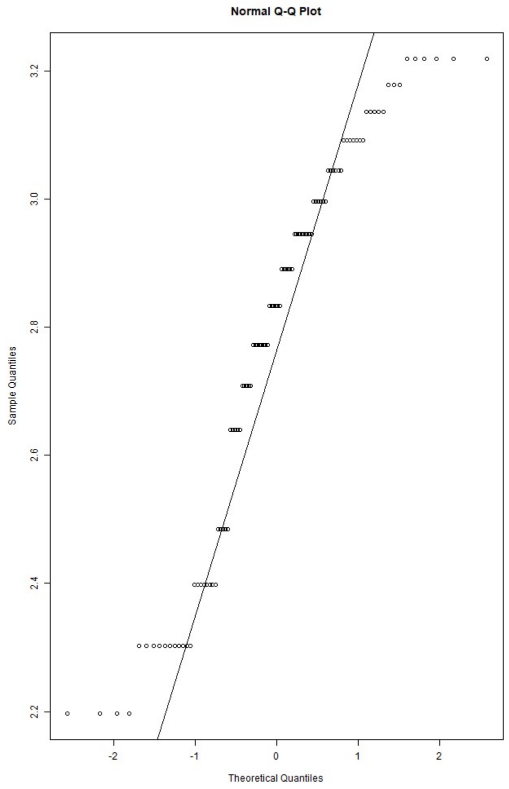
Data Analysis

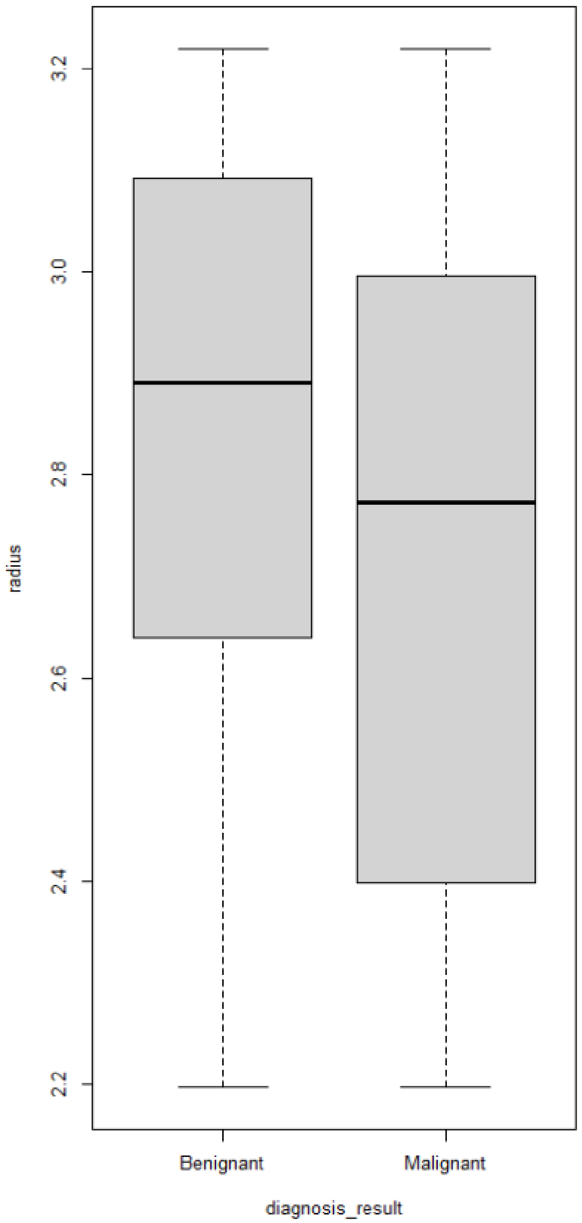
We observe now the plot produced by our features

RADIUS





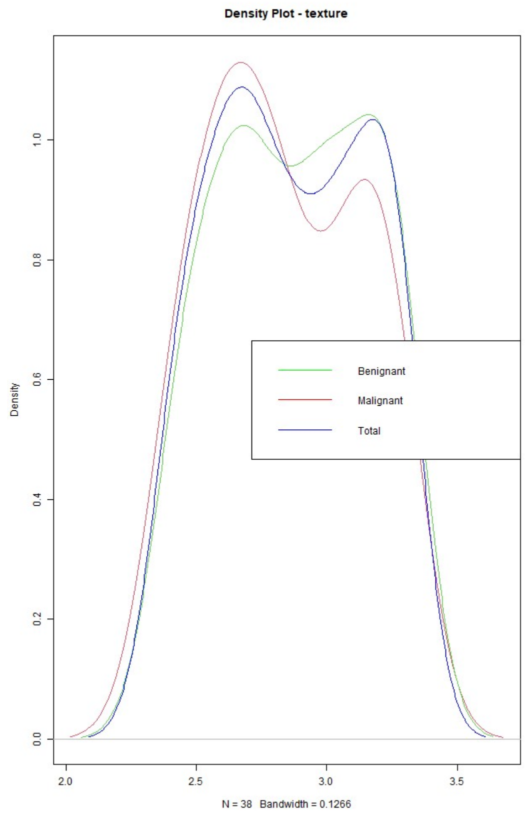


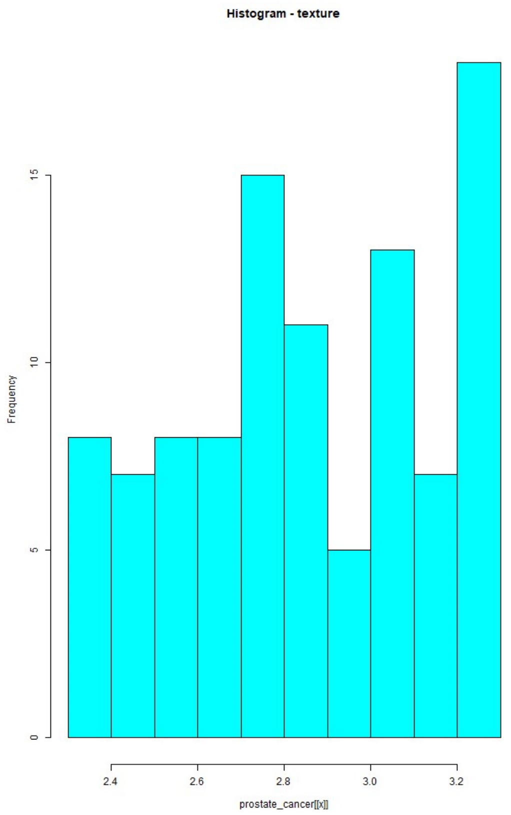


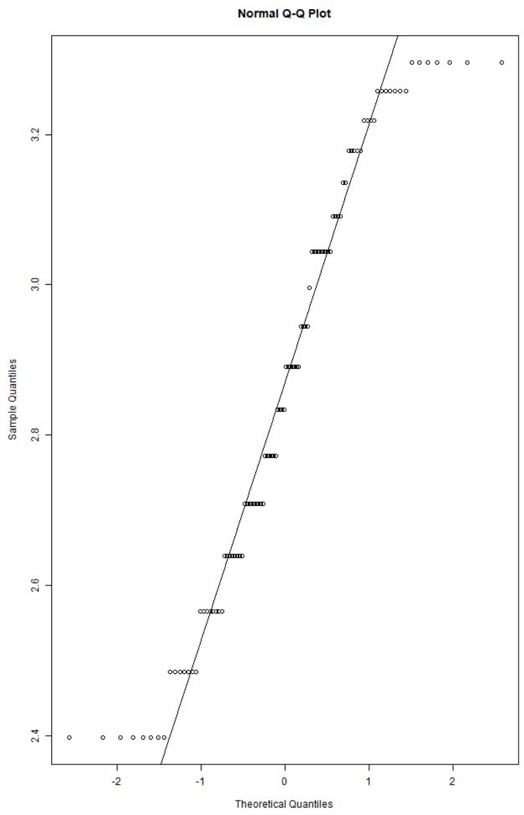
Data Analysis

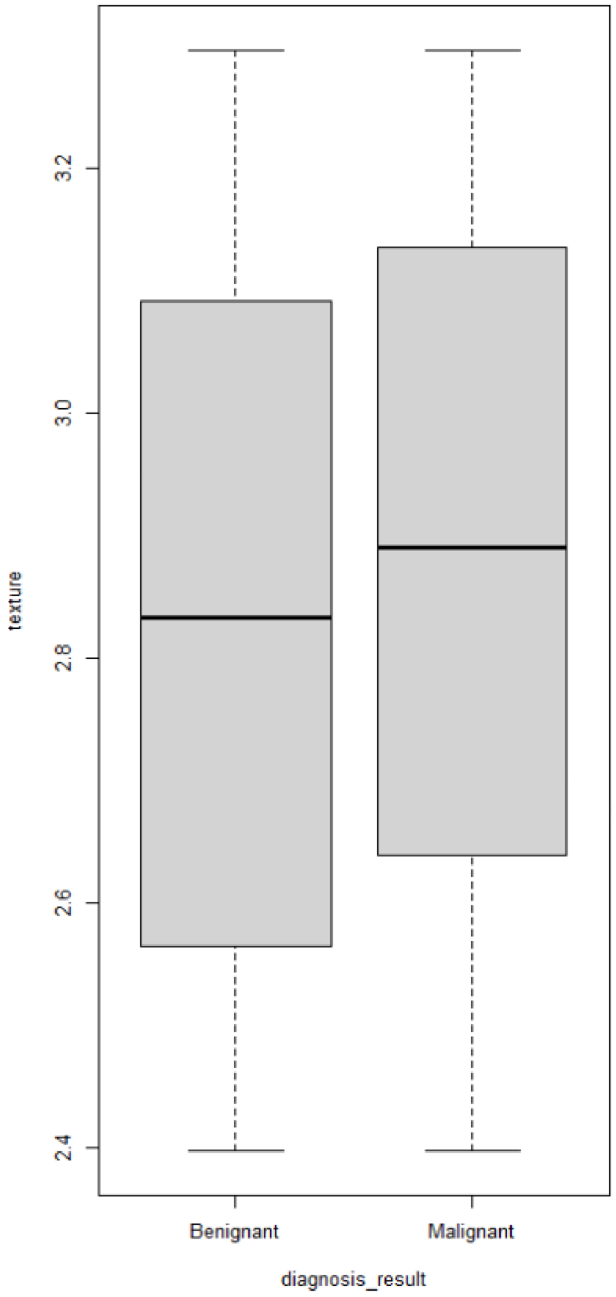
We observe now the plot produced by our features

TEXTURE





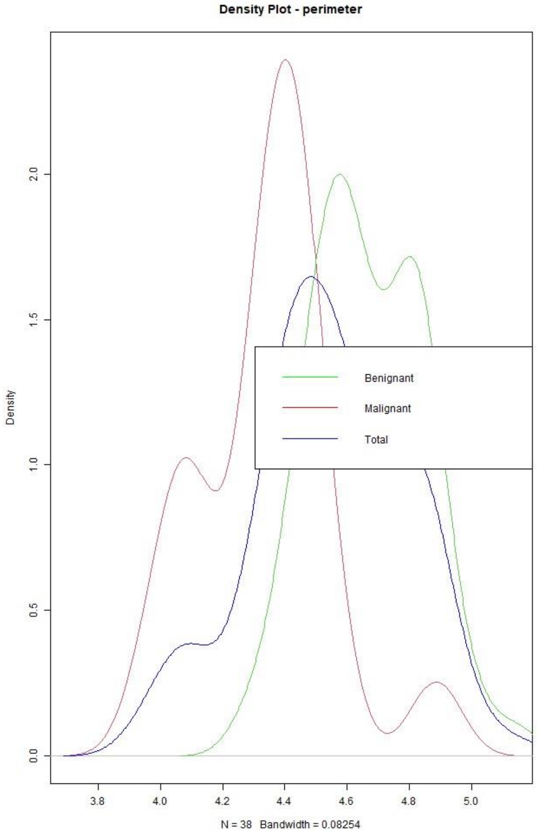


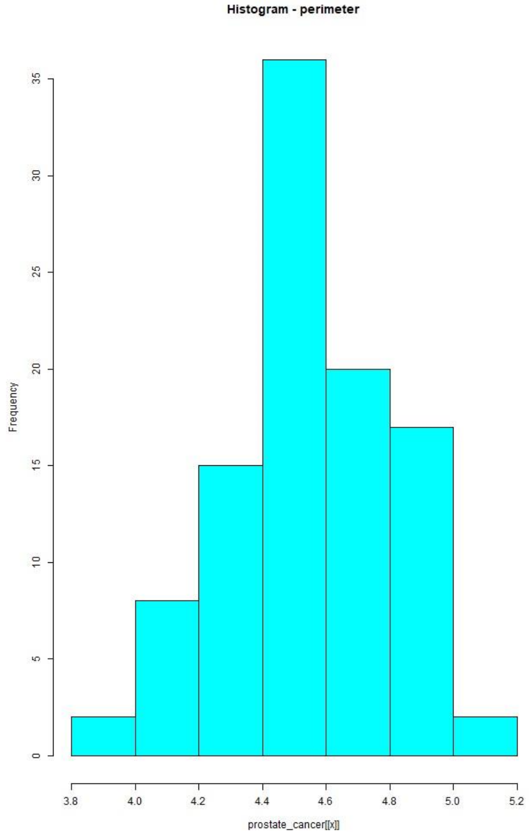


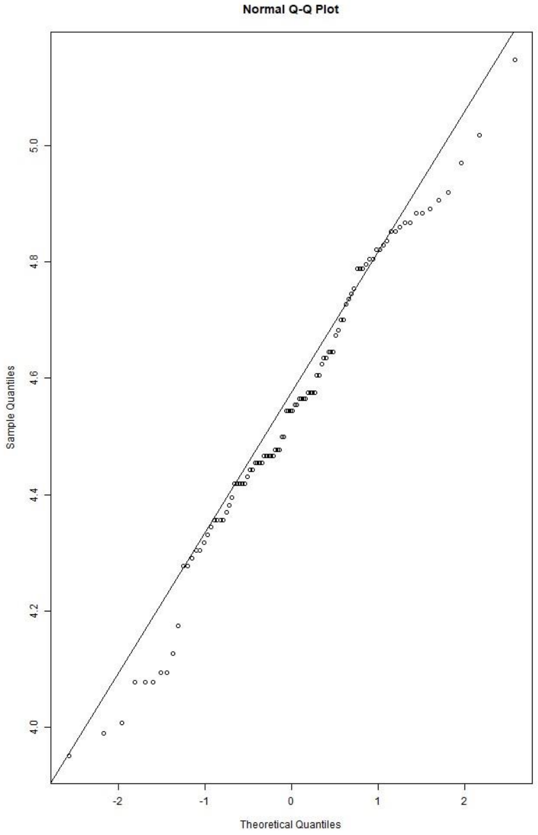
Data Analysis

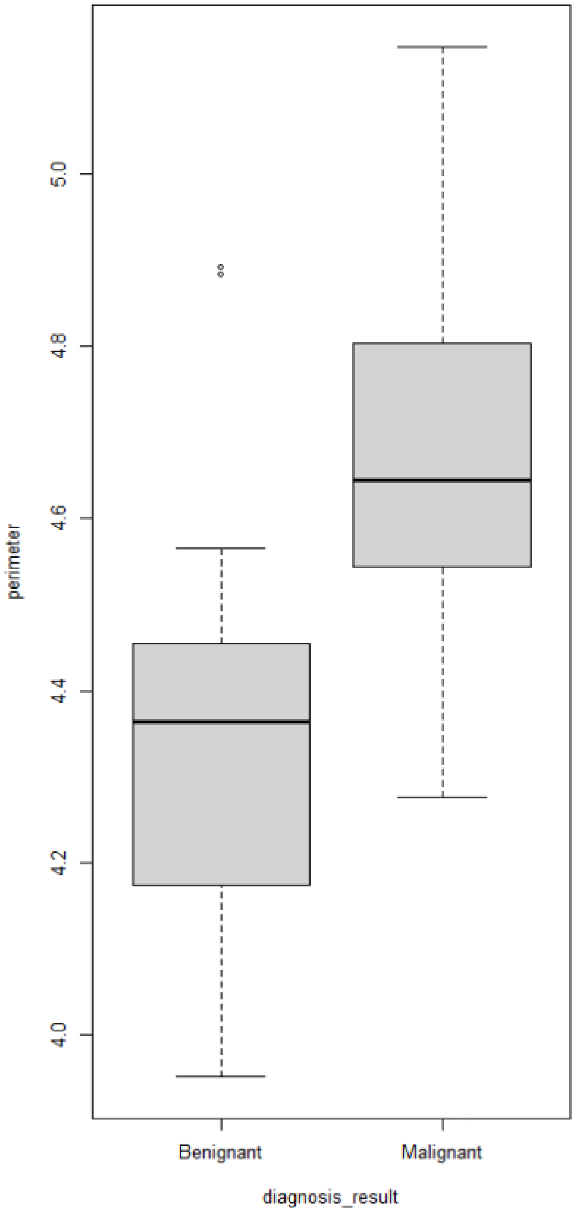
We observe now the plot produced by our features

PERIMETER





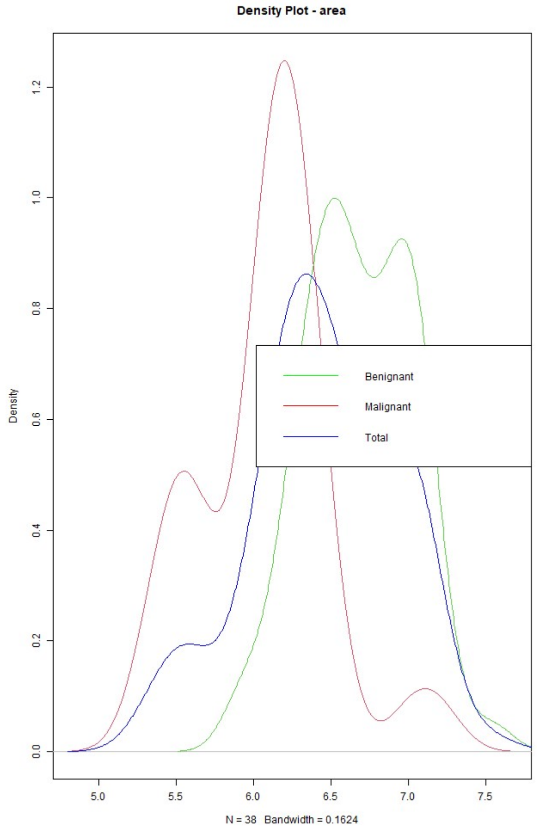


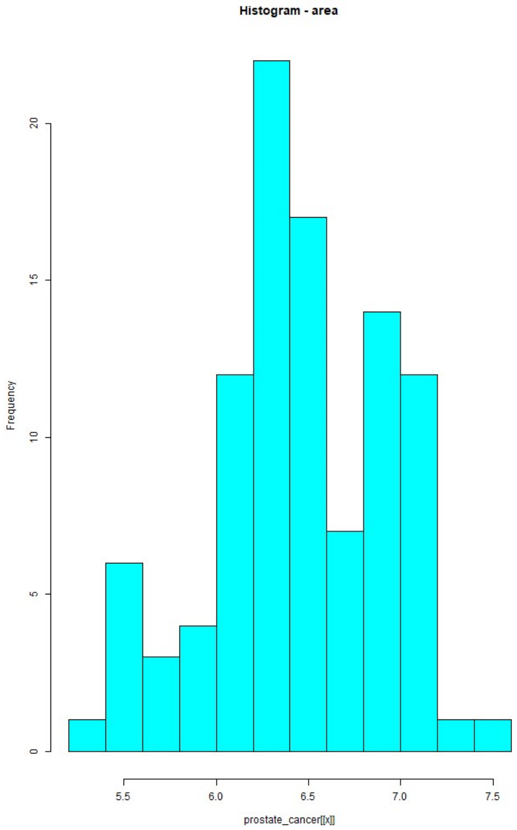


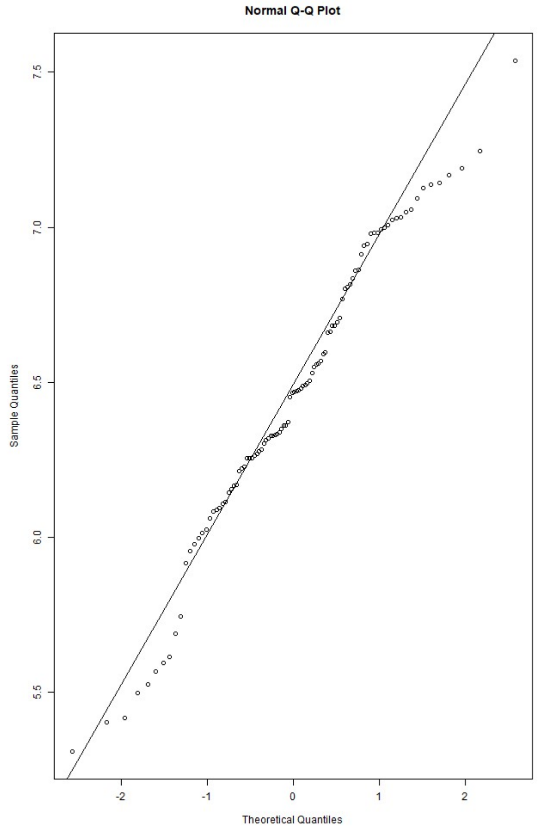
Data Analysis

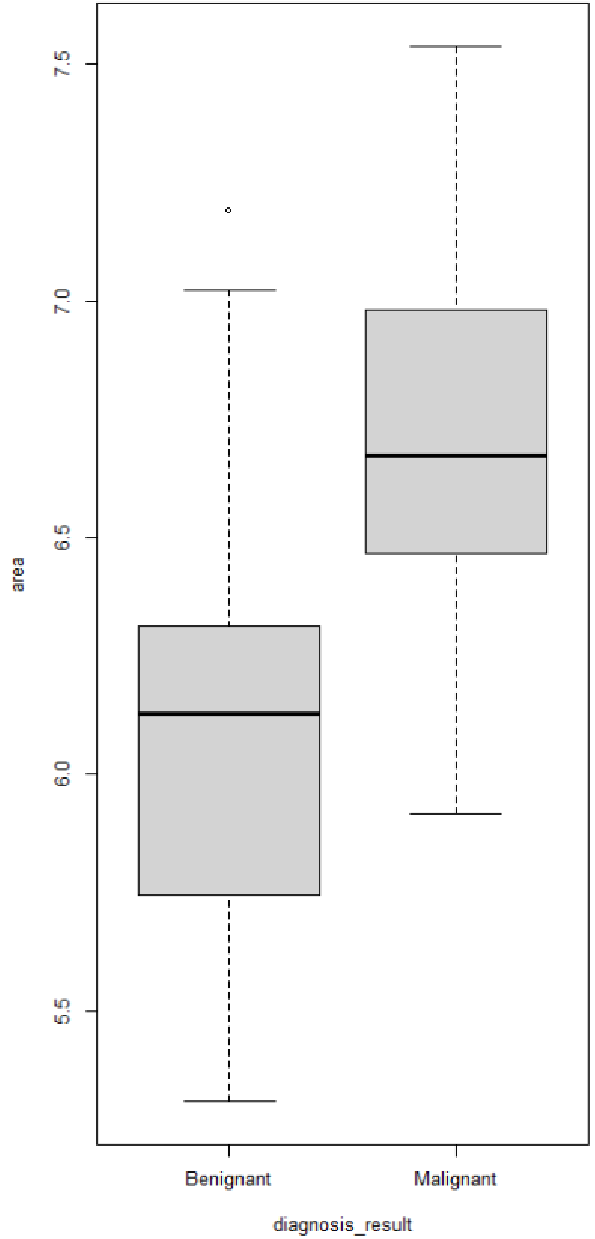
We observe now the plot produced by our features

AREA





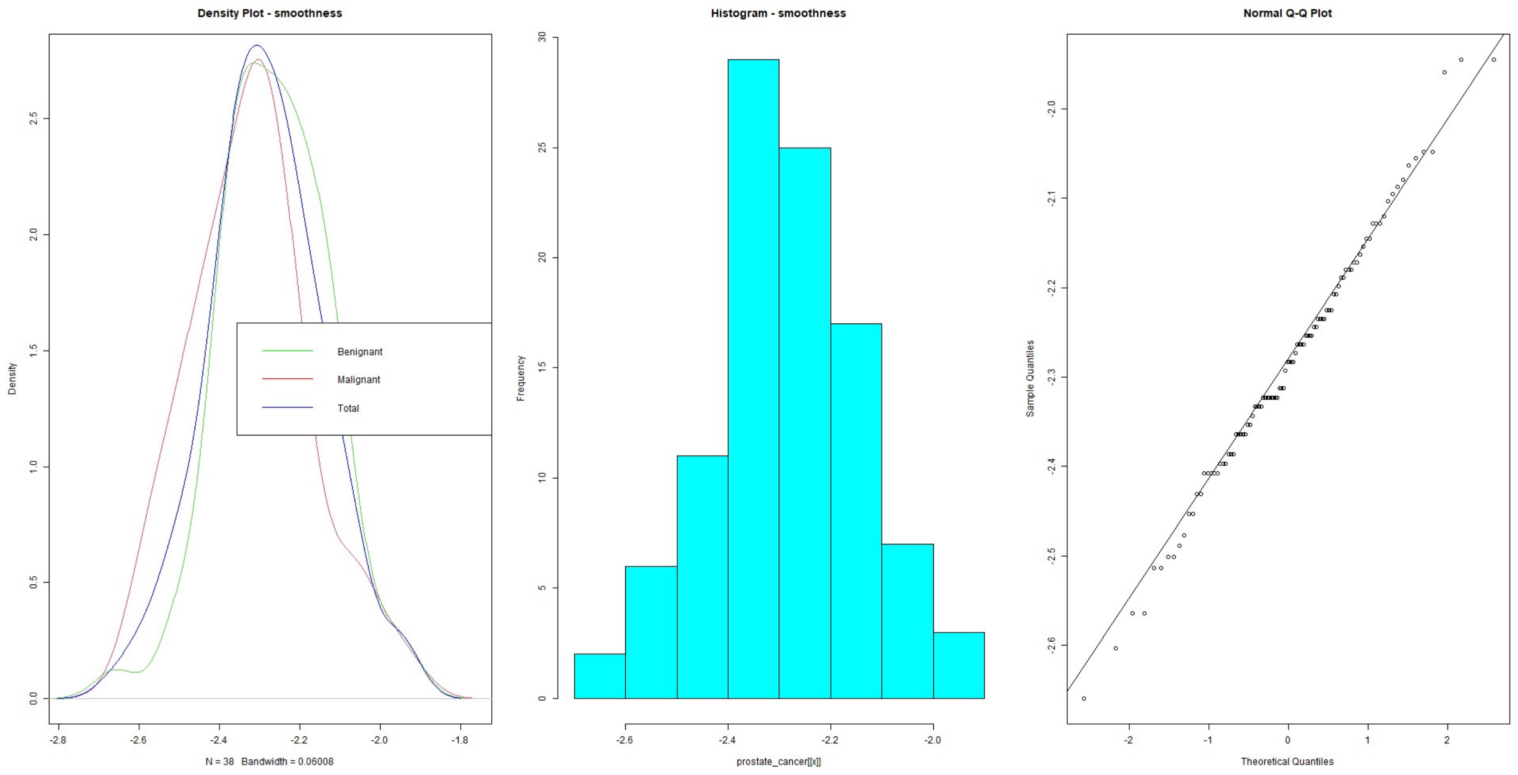


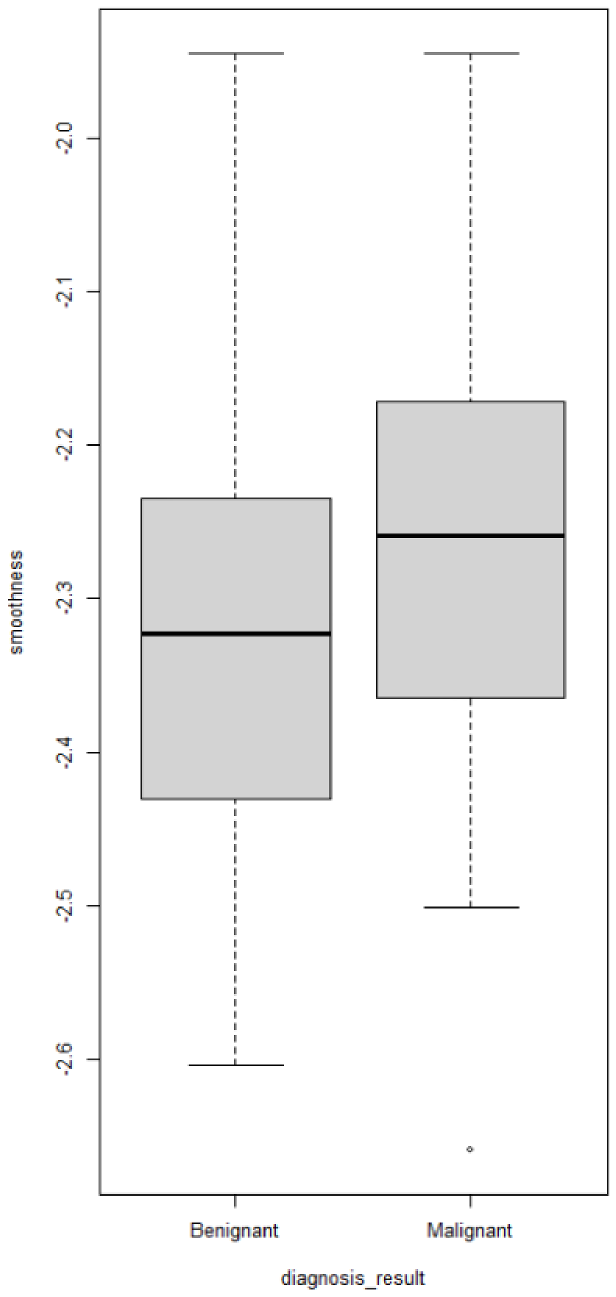


Data Analysis

We observe now the plot produced by our features

SMOOTHNESS

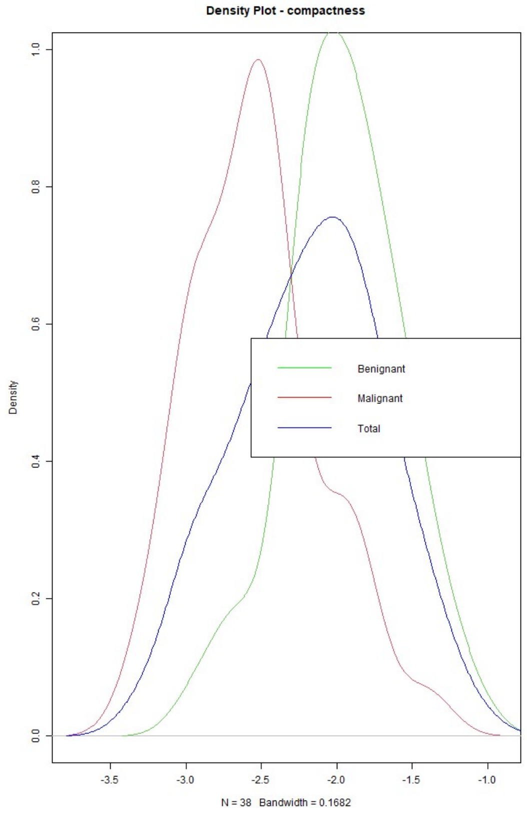


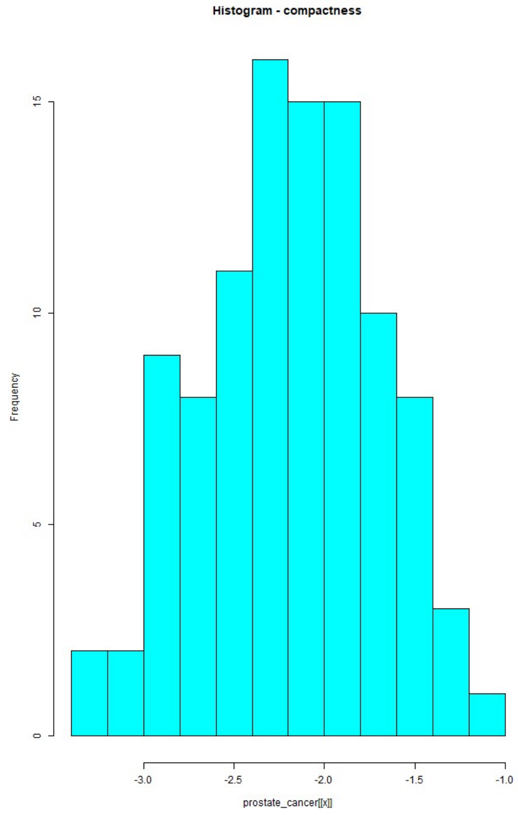


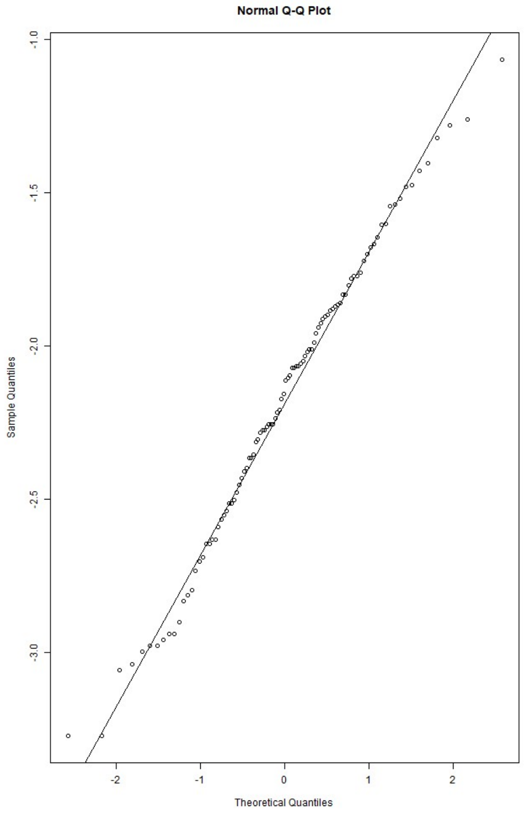
Data Analysis

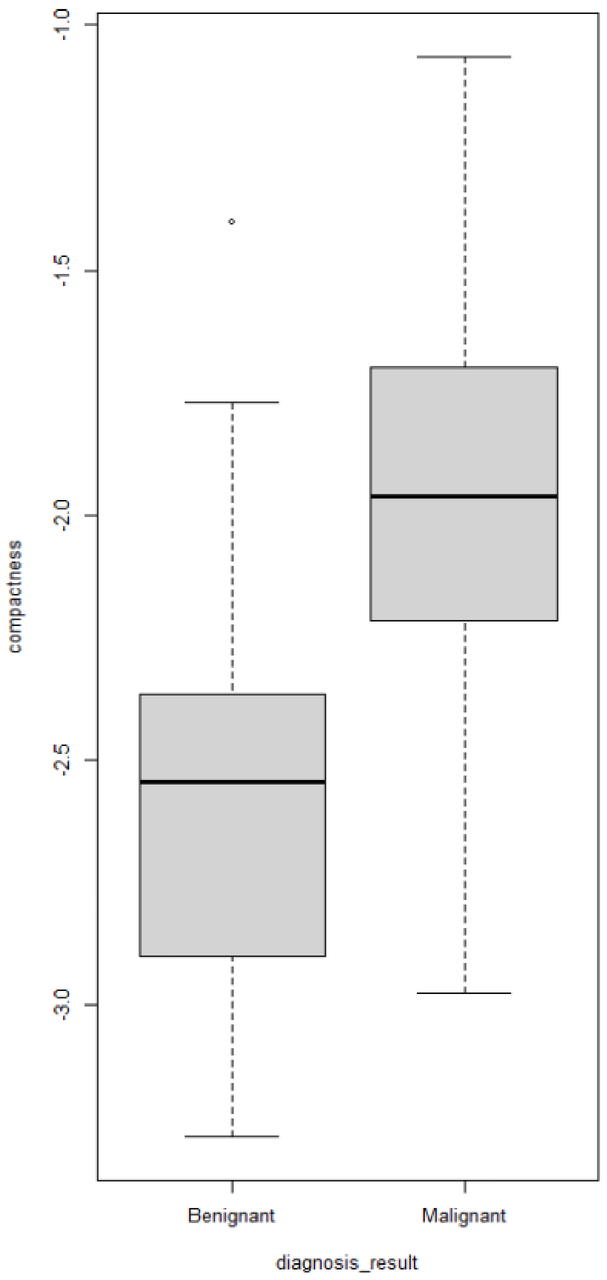
We observe now the plot produced by our features

COMPACTNESS


Density Plot - compactness


Histogram - compactness

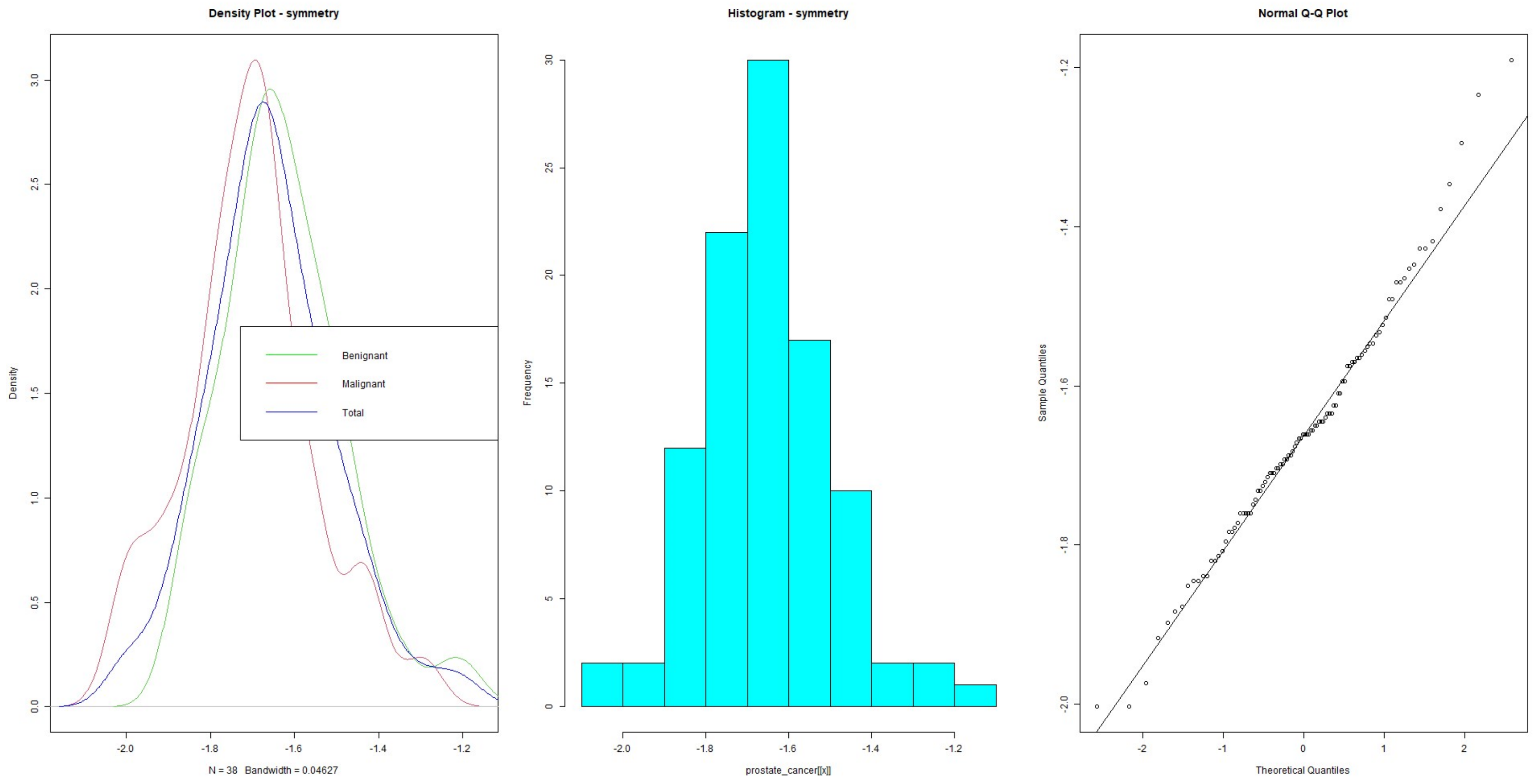

Normal Q-Q Plot

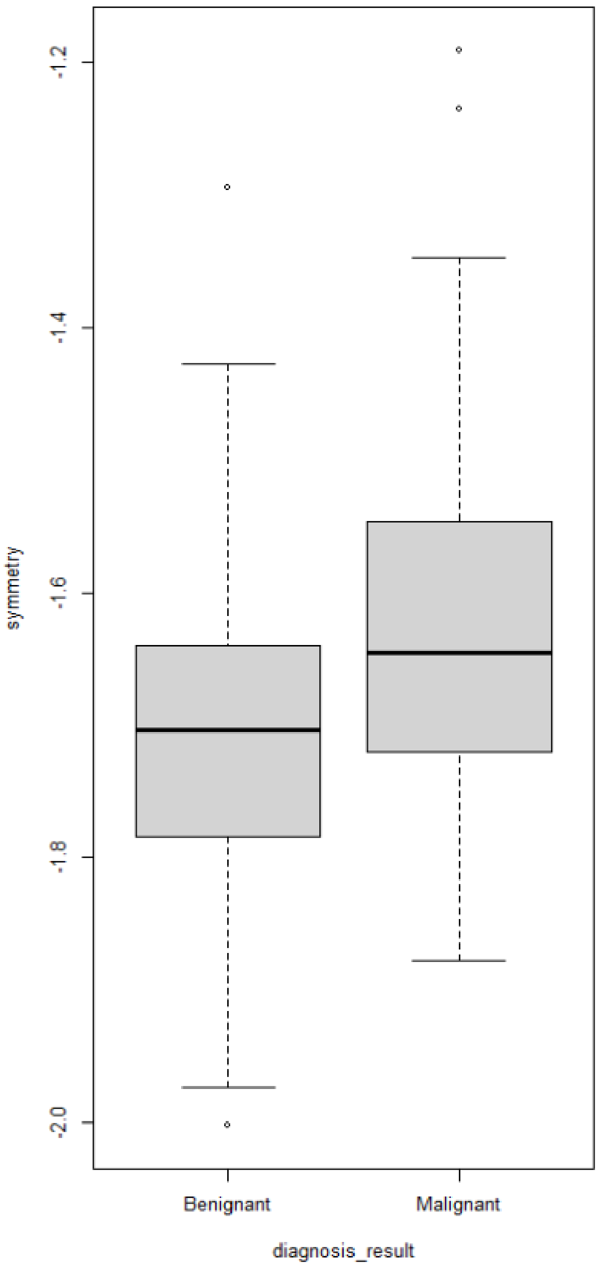


Data Analysis

We observe now the plot produced by our features

SYMMETRY

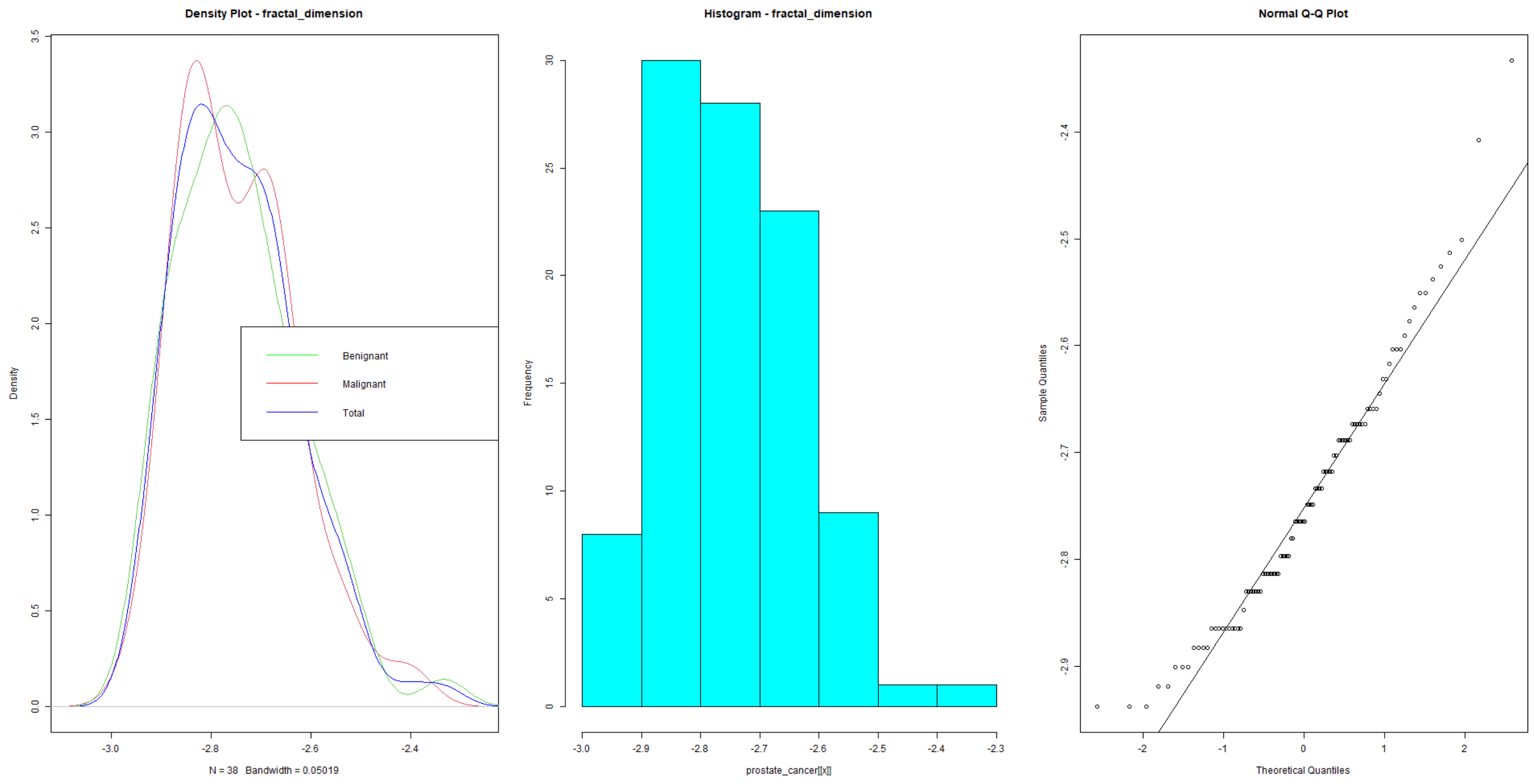


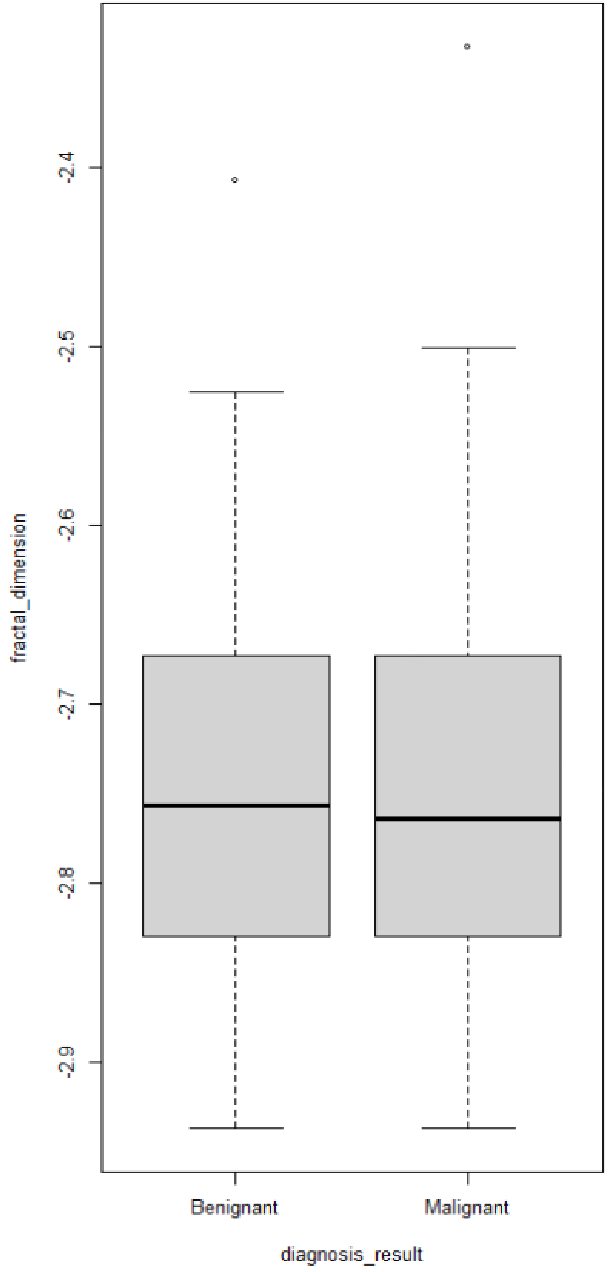


Data Analysis

We observe now the plot produced by our features

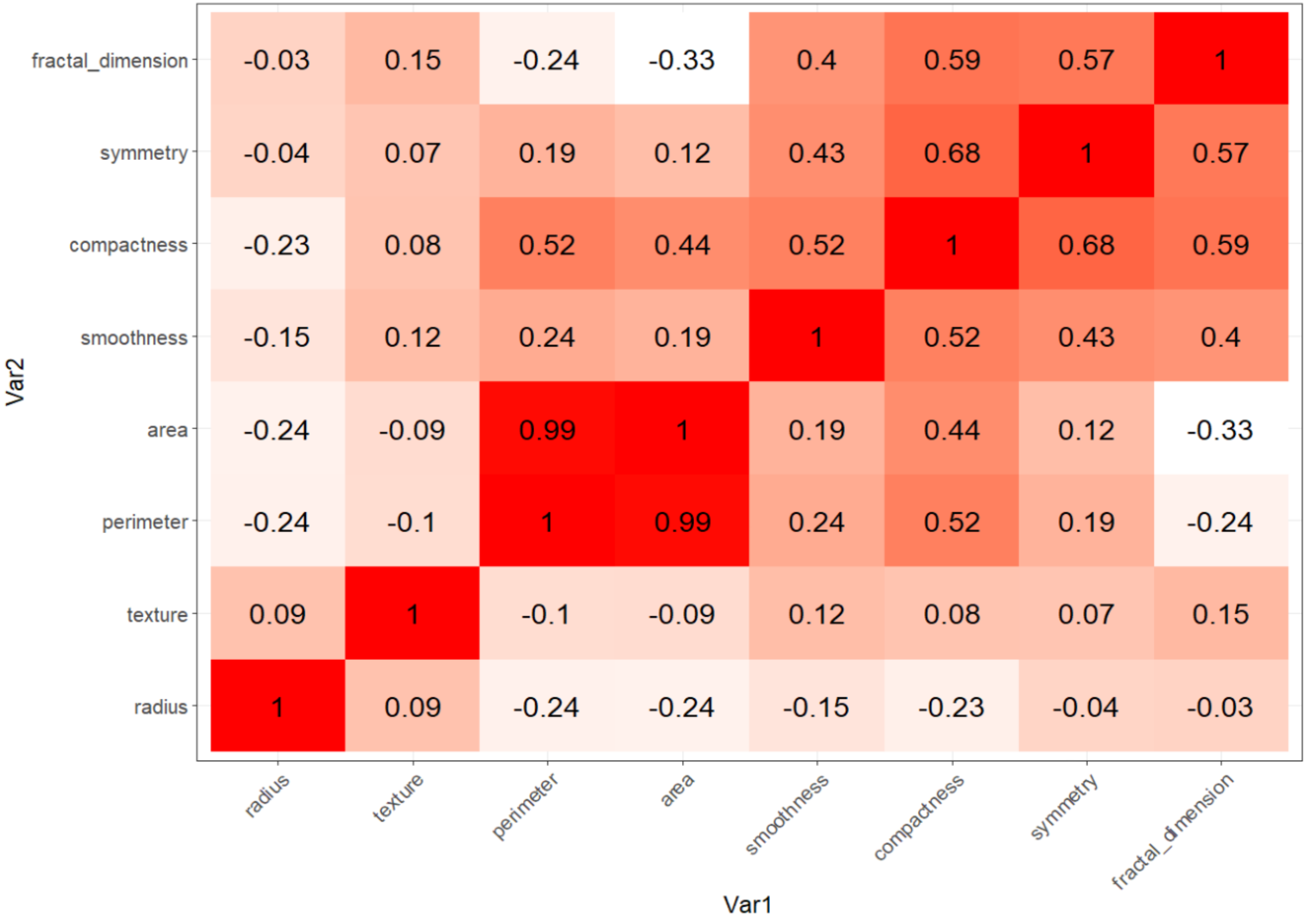
FRACTAL DIMENSION





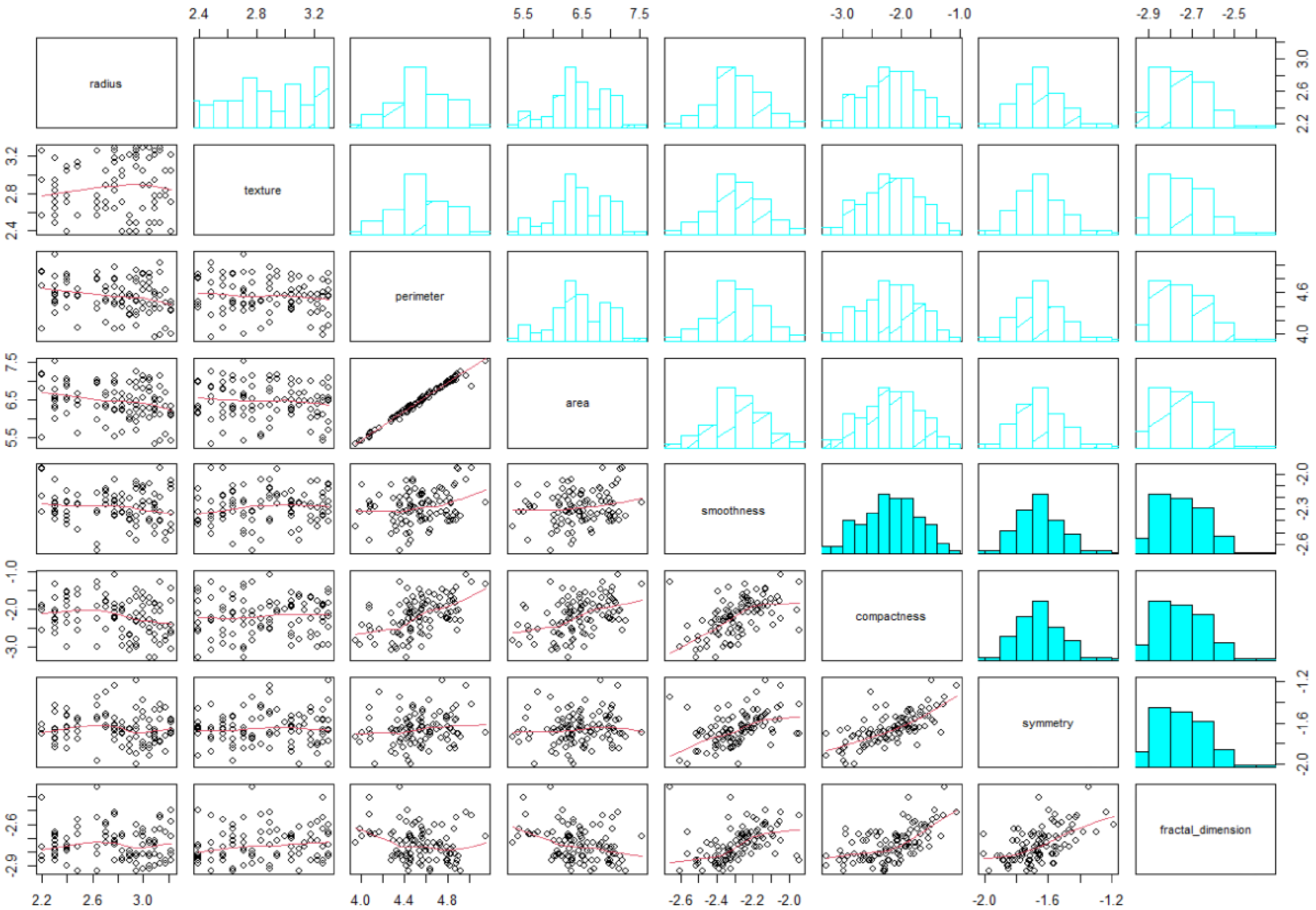
Data Analysis

We now check the **correlation** between all the numerical variables



Data Analysis

To visualize the pairwise comparison of correlations between our variables, we generate a Pair Plot.



Classification Models

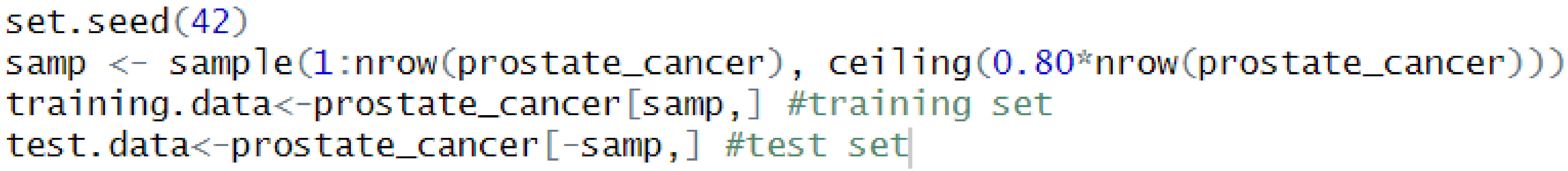
- Logistic Regression
- Ridge Regression
- Lasso Regression
- Naive Bayes
- Linear Discriminant Analysis (LDA)
- Quadratic Discriminant Analysis (QDA)
- K-NN

Train and Test Set

We decided to split the dataset into:

Train: 80%

Test: 20%



```
set.seed(42)
samp <- sample(1:nrow(prostate_cancer), ceiling(0.80*nrow(prostate_cancer)))
training.data<-prostate_cancer[samp,] #training set
test.data<-prostate_cancer[-samp,] #test set
```

AIM

We estimate the probability of developing cancer given the values of its attributes.

STEPWISE

Possibility to apply the Stepwise Selection approach for including and excluding iteratively the covariates inside the model.

VIF

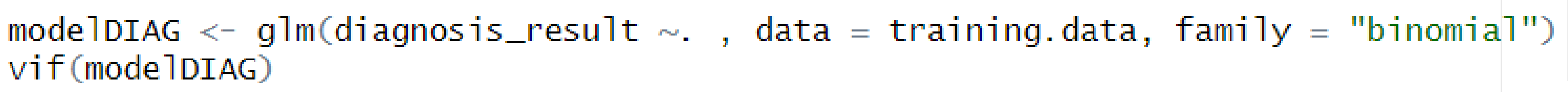
We consider the Variance Inflation Factor for the remotion of the collinearity.

GLM:

General Setting

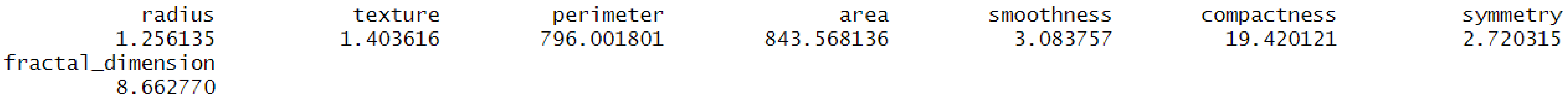
Logistic Regression

First, we look at the VIF of our model



```
modelDIAG <- glm(diagnosis_result ~. , data = training.data, family = "binomial")
vif(modelDIAG)
```

VIF values:



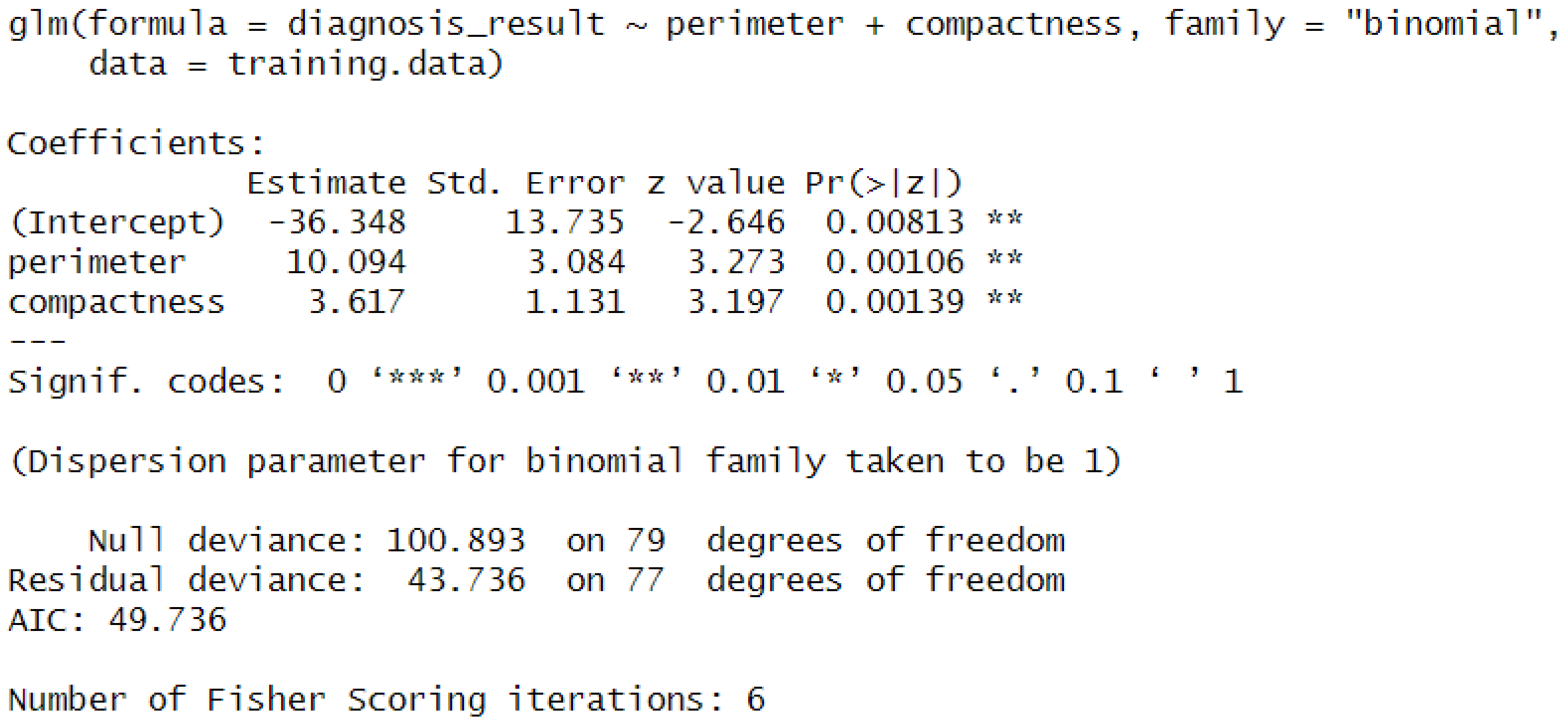
```
           radius          texture        perimeter             area       smoothness      compactness         symmetry 
         1.256135         1.403616       796.001801       843.568136         3.083757        19.420121         2.720315 
fractal_dimension 
         8.662770 
```

We decide to remove the 'area' variable and, subsequently, after reapplying the VIF to the model, the 'fractal_dimension' variable, which showed high values.

Logistic Regression

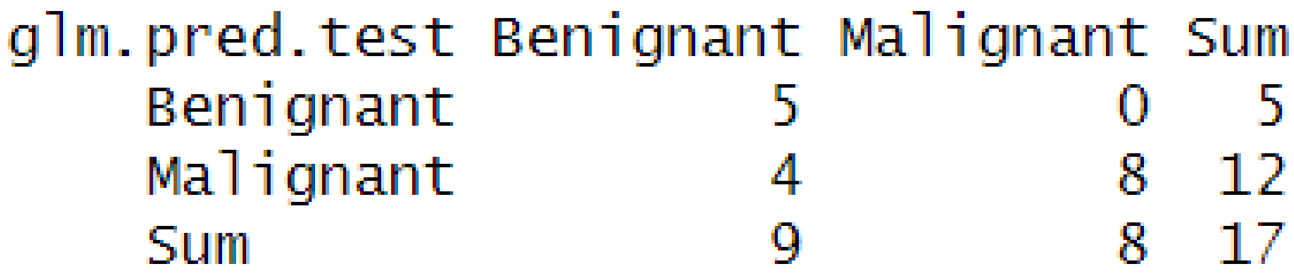
Starting with the model obtained after the application of VIF method, we use the Backward Selection.

At the end, the optimal model is:



```
glm(formula = diagnosis_result ~ perimeter + compactness, family = "binomial",
    data = training.data)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -36.348     13.735  -2.646  0.00813 **
perimeter     10.094      3.084   3.273  0.00106 **
compactness    3.617      1.131   3.197  0.00139 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 100.893  on 79  degrees of freedom
Residual deviance:  43.736  on 77  degrees of freedom
AIC: 49.736

Number of Fisher Scoring iterations: 6
```

Confusion matrix of Logistic Regression



```
glm.pred.test Benignant Malignant Sum
    Benignant         5         0   5
    Malignant         4         8  12
    Sum               9         8  17
```

Error	0.05882353
Specificity	1
Sensitivity	0.5555556
False Positive Rate	0

Regularized Regression

Generalized models with some **penalizations** according to **λ**;

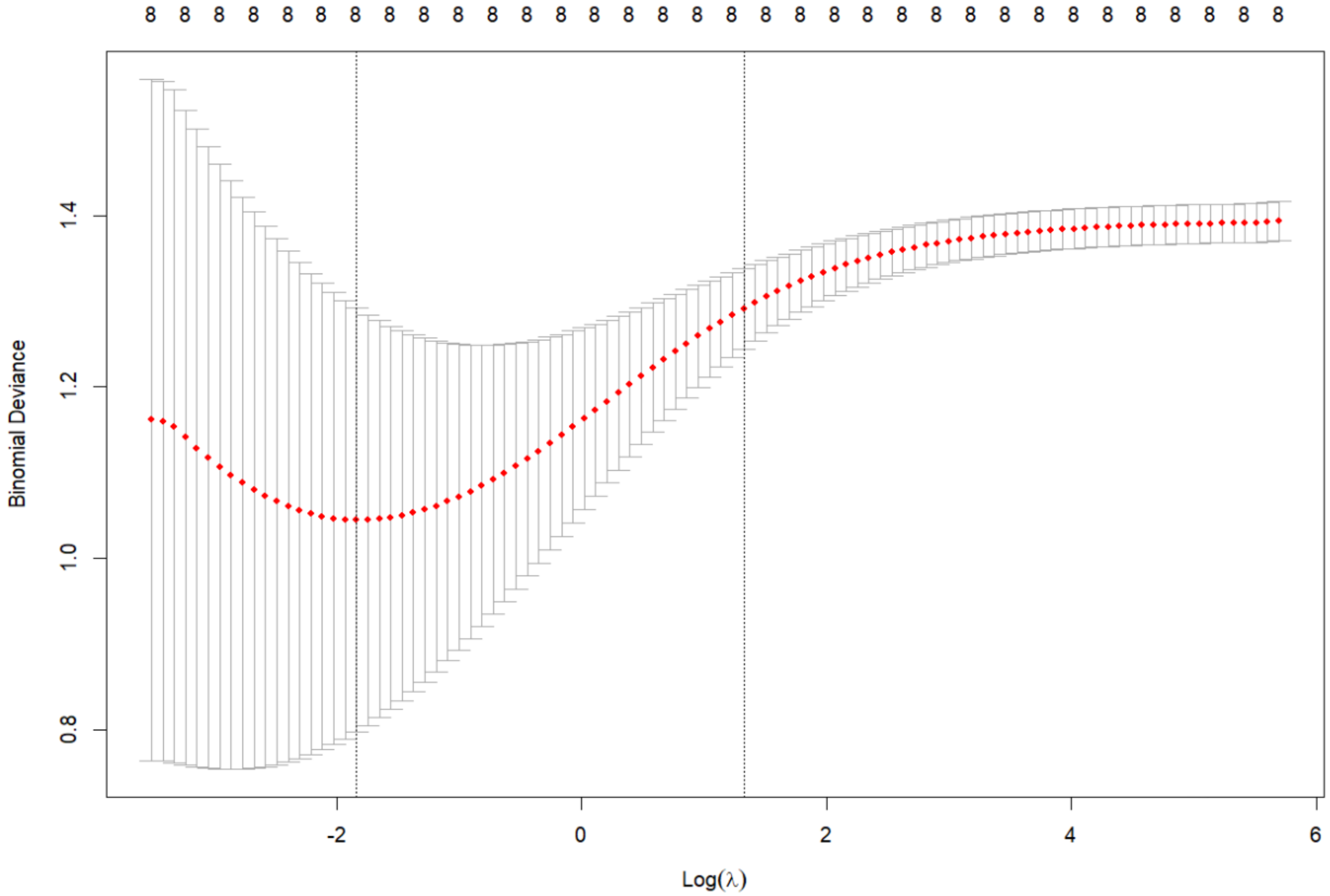
Estimators with very large variants and small bias can lead to **multicollinearity** and produce poor estimates;

Automatic selection of variables through shrinkage on the coefficients of the predictors in such a way that they assume **values very close to zero (Ridge)** or **even zero (Lasso);**

We decide to use **balanced dataset**;

Ridge

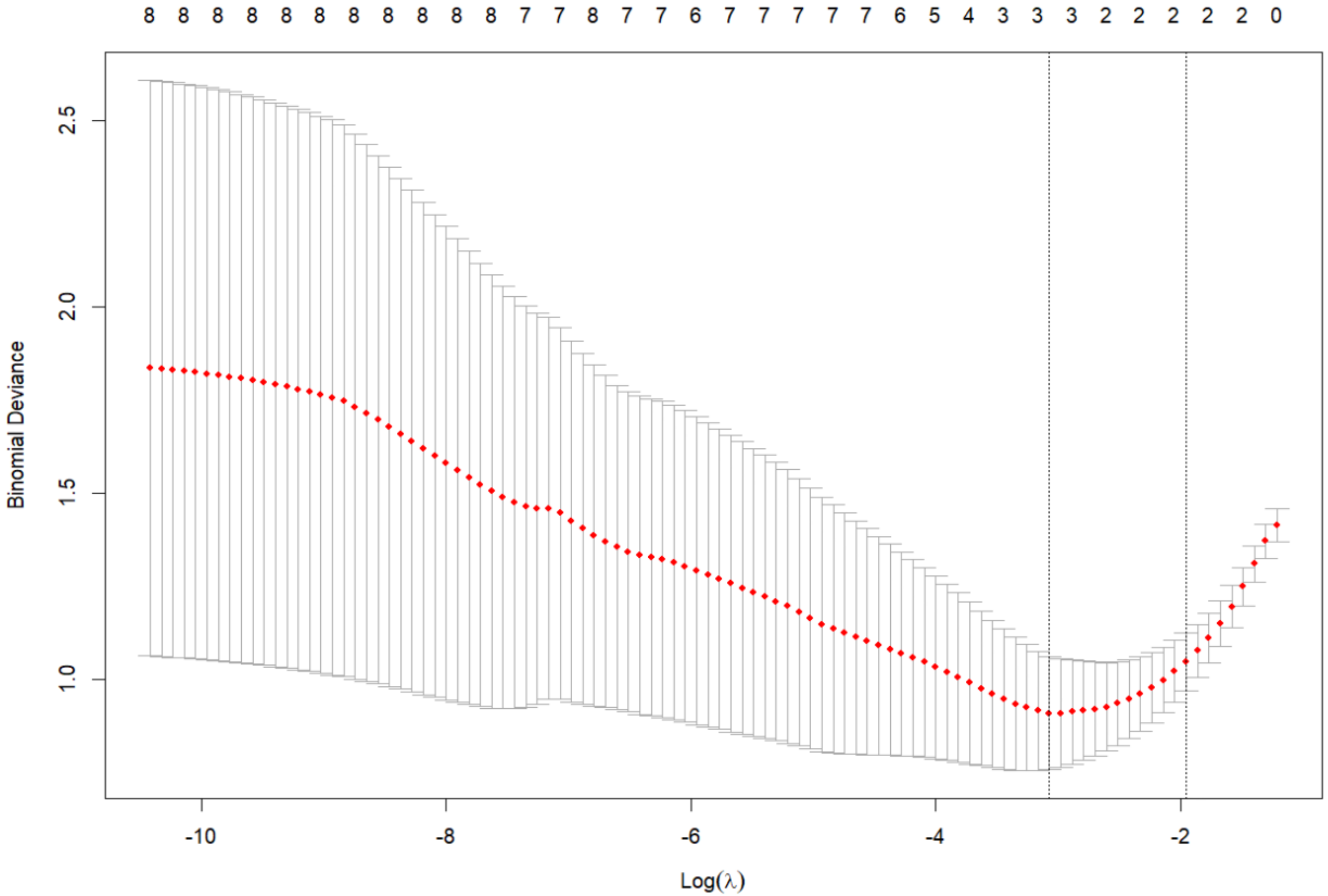
Best λ: 0.15917



Error	0.1
Specificity	0.9142857
Sensitivity	0.8666667
False Positive Rate	0.08571429

Lasso

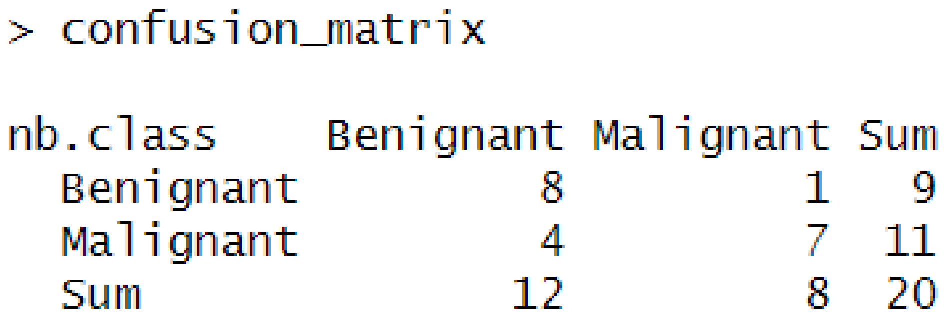
Best λ: 0.04639891



Error	0.14
Specificity	0.9090909
Sensitivity	0.7647059
False Positive Rate	0.09090909

Naive Bayes

Classification model



```
> confusion_matrix

nb.class     Benignant Malignant Sum
  Benignant          8         1   9
  Malignant          4         7  11
  Sum               12         8  20
```

Error	0.25
Specificity	0.875
Sensitivity	0.6666667
False Positive Rate	0.125

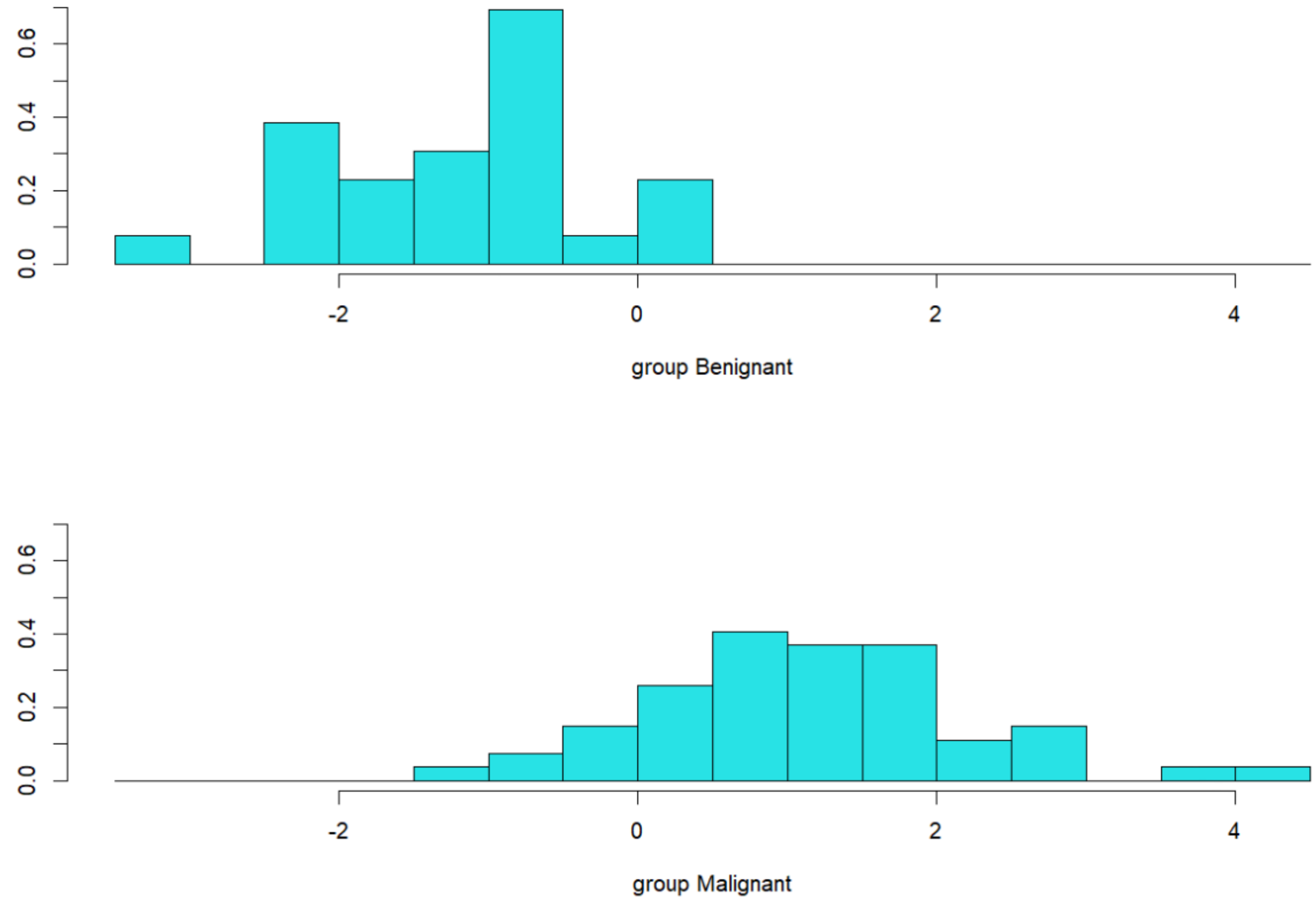
Discriminant Analysis

Other models for classification problems are **Linear Discriminant Analysis** and **Quadratic Discriminant Analysis** which are based on Bayes theorem.

Problems:	the normality of the variables conditioned to the 2 classes;
	No features selection
	Sensitivity to outliers

LDA

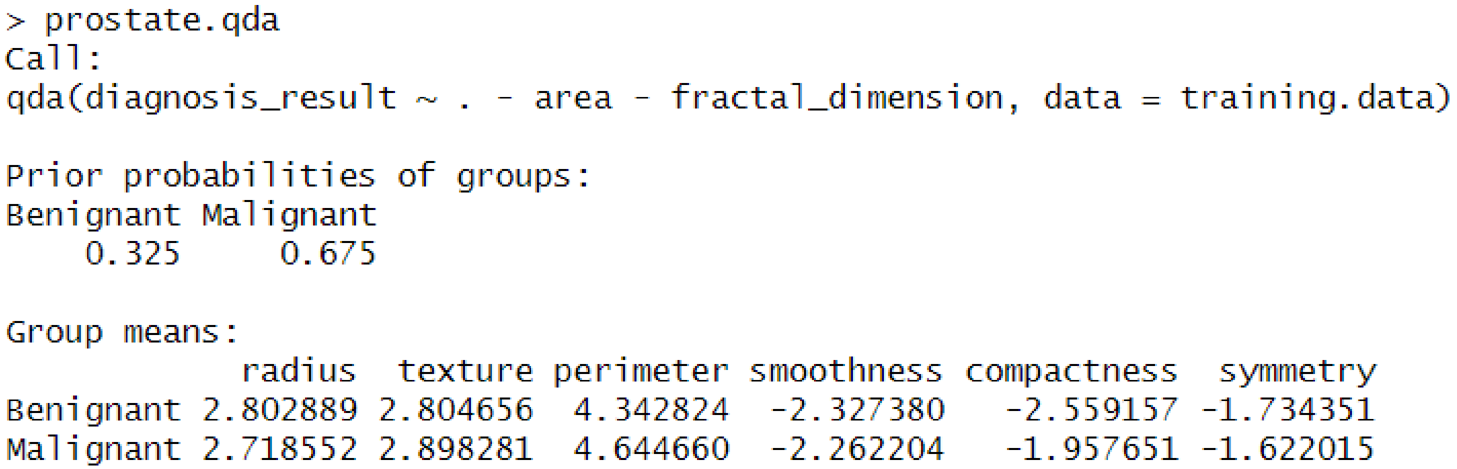
Linear discriminant analysis



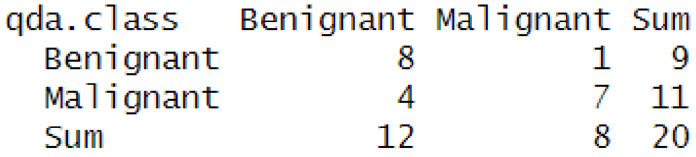
Error	0.15
Specificity	1
Sensitivity	0.75
False Positive Rate	0

QDA

Quadratic Discriminant Analysis



```
> prostate.qda
Call:
qda(diagnosis_result ~ . - area - fractal_dimension, data = training.data)

Prior probabilities of groups:
Benignant Malignant 
    0.325     0.675 

Group means:
            radius  texture perimeter smoothness compactness  symmetry
Benignant 2.802889 2.804656  4.342824  -2.327380   -2.559157 -1.734351
Malignant 2.718552 2.898281  4.644660  -2.262204   -1.957651 -1.622015
```



```
qda.class   Benignant Malignant Sum
  Benignant         8         1   9
  Malignant         4         7  11
  Sum              12         8  20
```

Error	0.25
Specificity	0.875
Sensitivity	0.6666667
False Positive Rate	0.125

K-Nearest Neighbors

K-NN is a completely non parametric approach;

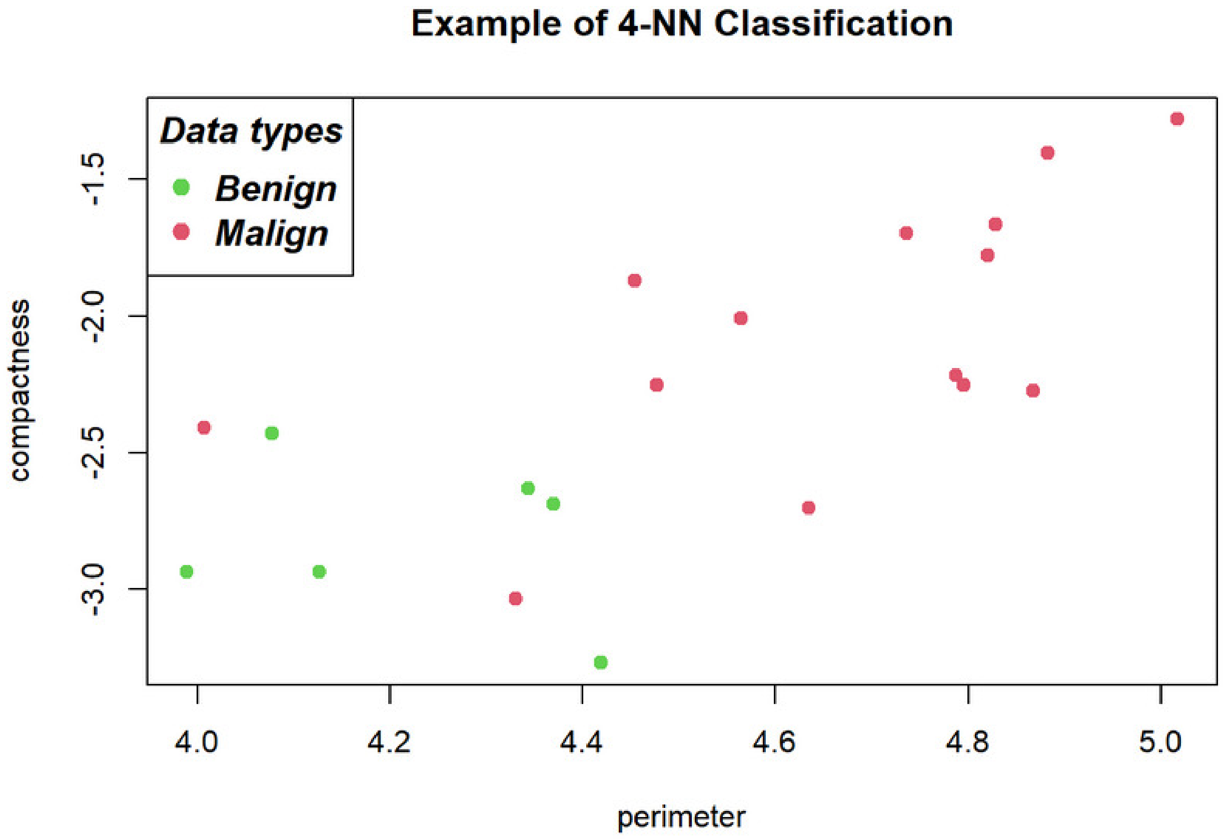
To make a prediction for an observation x using K-NN:

- The K training observations that are closest to x are identified;
- the x is assigned to the class to which the plurality of these observations belong.

KNN

K-Nearest Neighbors

Best **k** = 4

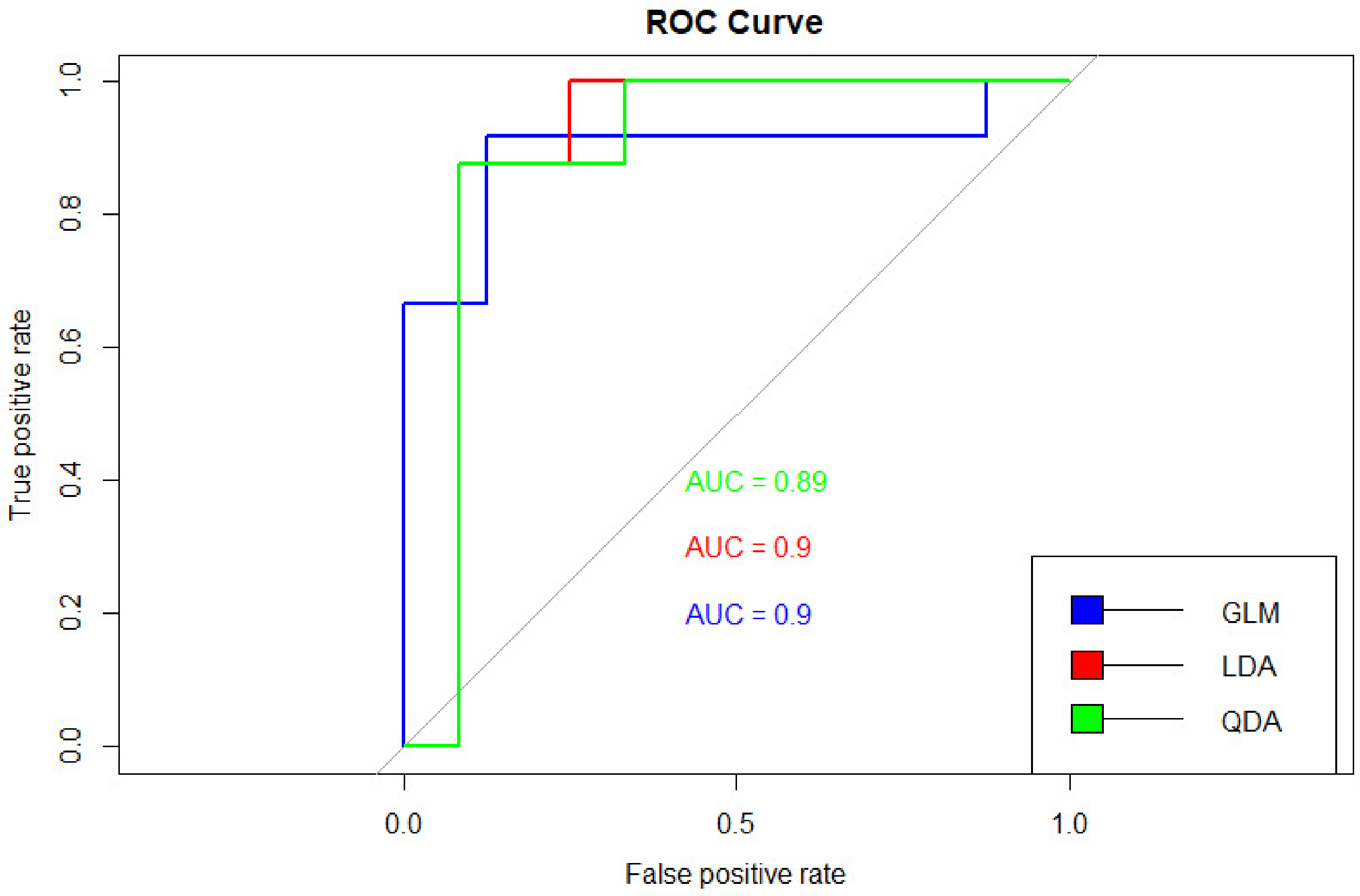


Error	0.3
Specificity	0.5714286
Sensitivity	1
False Positive Rate	0.4285714

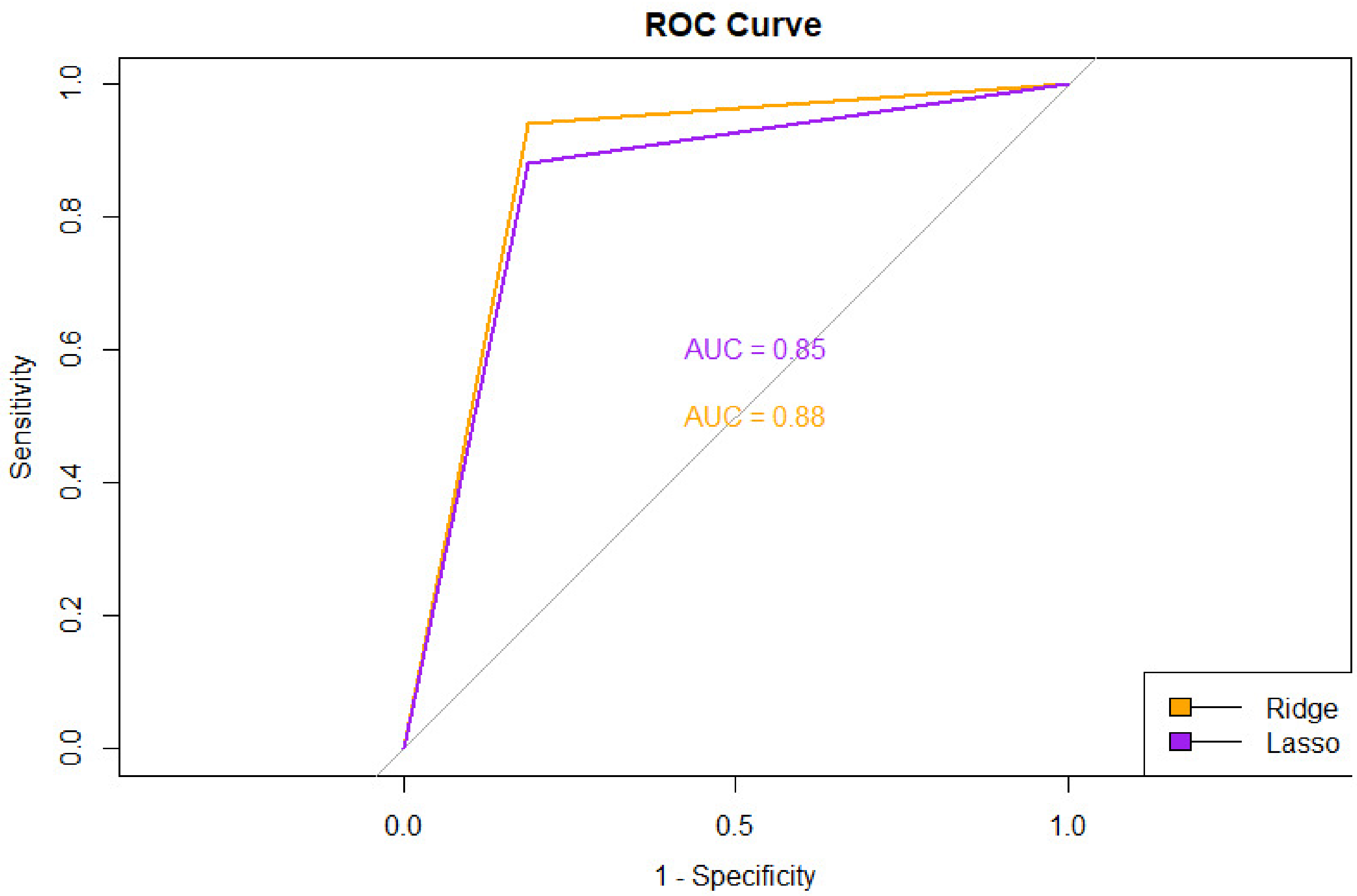
Comparison of models

Model	Error	Specificity	Sensitivity	False Positive Rate
Logistic Regression	0.058	1	0.555	0
Ridge	0.1	0.914	0.866	0.085
Lasso	0.14	0.909	0.764	0.090
Naive Bayes	0.25	0.875	0.666	0.125
LDA	0.15	1	0.75	0
QDA	0.25	0.875	0.666	0.125
K-NN	0.3	0.571	1	0.428

ROC CURVE


ROC Curve
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.5
1.0
AUC = 0.89
AUC = 0.9
AUC = 0.9
🟦 GLM
🟥 LDA
🟩 QDA

ROC CURVE RIDGE/LASSO



THANK YOU

Graziana Capurso Mat. 2097099

Daniela Di Labbio Mat. 2091677

Agata Garbin Mat. 2072693