Data Science - Capstone - Project Free Breast Cancer

Jose Maria Martin Arribas

Febrary 2025

Contents

1. INTRODUCTION **1**

I. Loading libraries and the dataset . . . 2

II. Creating test and train set (breast_cancer) . . . 3

2. DATA ANALYSYS **4**

I. Data structure . . . 4

II. Summaries data . . . 5

III. Study graphics statistics . . . 6

IV. Conclusion study graphic . . . 11

3. MACHINE LEARNING **11**

I. PCA (principal component analysis) . . . 12

II. Glm (logistic regression) . . . 15

III. Loess (local polynomial regression fitting) . . . 16

IV. K nearest neighbors . . . 16

V. Random forest . . . 17

VI. Ensemble . . . 18

4. RESULTS **19**

5. CONCLUSION **20**

6. REFERENCES **21**

1. INTRODUCTION

The project is based on a dataset provided through: https://archive.ics.uci.edu/dataset/451/breast+cancer+coimbra

We will have to analyze the data using visualization knowledge to make summaries that we observe.

The objective is to use the breast cancer dataset to experiment with multiple predictors (biomarkers in blood analytical) and use cross-validation to create a machine learning with:

- PCA (principal component analysis)
- Glm (logistic regression)
- Loess (local polynomial regression fitting)
- K nearest neighbors
- Random forest.
- Ensemble.

A long to this methods of machine learning we will observe what of all obtain the best accuracy.

I. Loading libraries and the dataset

This loading of the dataset is provided through https://archive.ics.uci.edu/dataset/451/breast+cancer+coimbra

```
############################################################
# 1.Create breast_train and breast_test sets
############################################################

# Note: this process could take a minutes.
if(!require(tidyverse)) install.packages("tidyverse", repos = "http://cran.us.r-project.org")
if(!require(caret)) install.packages("caret", repos = "http://cran.us.r-project.org")
if(!require(kableExtra)) install.packages("kableExtra", repos = "http://cran.us.r-project.org")
if(!require(data.table)) install.packages("data.table", repos = "http://cran.us.r-project.org")
if(!require(vtable)) install.packages("vtable", repos = "http://cran.us.r-project.org")
if(!require(ggthemes)) install.packages("ggthemes", repos = "http://cran.us.r-project.org")
if(!require(corrr)) install.packages("corrr", repos = "http://cran.us.r-project.org")
if(!require(ggcorrplot)) install.packages("ggcorrplot", repos = "http://cran.us.r-project.org")
if(!require(factoextra)) install.packages("factoextra", repos = "http://cran.us.r-project.org")
if(!require(ucimlrepo)) install.packages("recosystem", repos = "http://cran.us.r-project.org")
if(!require(gam)) install.packages("gam", repos = "http://cran.us.r-project.org")
if(!require(splines)) install.packages("splines", repos = "http://cran.us.r-project.org")
if(!require(foreach)) install.packages("foreach", repos = "http://cran.us.r-project.org")

# Libraries required to run the proyect
library(tidyverse)
library(caret)
library(kableExtra)
library(data.table)
library(vtable)
library(ggthemes)
library(corrr)
library(ggcorrplot)
library(factoextra)
library(ucimlrepo)
library(gam)
library(splines)
```

```r
library(foreach)

options(digits = 5)
options(timeout = 120)

# Check which datasets can be imported and choose breast cancer coimbra
# Create a complex HTML table using kableExtra for obtain list of data set of uciml
kable(list_available_datasets(search = "cancer"), caption = "Structure UC Irvine repository") %>%
  kable_styling(bootstrap_options = c("striped", "hover"), html_font = "Palatino") %>%
  column_spec(1:3, bold = TRUE, color = "#27408B") %>%
  footnote(general = "This table lists dataset of cancer (UC Irvine machine learning repository)")
```

Table 1: Structure UC Irvine repository

id	name	url
14	**Breast Cancer**	**https://archive.ics.uci.edu/static/public/14/data.csv**
15	**Breast Cancer Wisconsin (Original)**	**https://archive.ics.uci.edu/static/public/15/data.csv**
16	**Breast Cancer Wisconsin (Prognostic)**	**https://archive.ics.uci.edu/static/public/16/data.csv**
17	**Breast Cancer Wisconsin (Diagnostic)**	**https://archive.ics.uci.edu/static/public/17/data.csv**
62	**Lung Cancer**	**https://archive.ics.uci.edu/static/public/62/data.csv**
383	**Cervical Cancer (Risk Factors)**	**https://archive.ics.uci.edu/static/public/383/data.csv**
451	**Breast Cancer Coimbra**	**https://archive.ics.uci.edu/static/public/451/data.csv**
537	**Cervical Cancer Behavior Risk**	**https://archive.ics.uci.edu/static/public/537/data.csv**
915	**Differentiated Thyroid Cancer Recurrence**	**https://archive.ics.uci.edu/static/public/915/data.csv**

Note:
This table lists dataset of cancer (UC Irvine machine learning repository)

```r
#Save dataset with data and metadata
data_set_uciml = fetch_ucirepo(id=451)

#Save data to study in a data frame breast_file
breast_data <- data.frame(data_set_uciml$data$original)

#Do a backup data frame breast cancer in a csv
write.csv2(breast_data,"breast_cancer.csv")
```

II. Creating test and train set (breast_cancer)

We create two data sets:

1. One breast cancer data train to train of different models with 80% breast data.
2. Second breast cancer data test to probe accuracy different models with 20% breast data.

```r
set.seed(1, sample.kind="Rounding")

test_index <- createDataPartition(y = breast_data$Classification, times = 1, p = 0.2, list = FALSE)
#Breast test set will be 20% of breast cancer data
breast_train <- breast_data[-test_index,]
#Breast data will be 80 % of breast cancer data
breast_test <- breast_data[test_index,]
```

2. DATA ANALYSYS

The dimensions of data frame breast cancer test show we have 92 patients and 10 type of data (predictors-biomarkers).

```
# Check the dimensions of the breast train data
dim(breast_train)
```

```
## [1] 92 10
```

```
# Create a complex HTML table using kableExtra for obtain
# five first registers of data structure breast train
kable(head(breast_train, 5), caption = "Breast cancer") %>%
  kable_styling(bootstrap_options = c("striped", "hover"), html_font = "Tahoma" ) %>%
  column_spec(9:10, bold = TRUE, color = "#27408B") %>%
  footnote(general = "This table lists head of dataset cancer.")
```

Table 2: Breast cancer

Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP.1	Classification
48	23.500	70	2.707	0.46741	8.8071	9.7024	7.9958	**417.11**	**1**
83	20.690	92	3.115	0.70690	8.8438	5.4293	4.0640	**468.79**	**1**
82	23.125	91	4.498	1.00965	17.9393	22.4320	9.2772	**554.70**	**1**
68	21.368	77	3.226	0.61272	9.8827	7.1696	12.7660	**928.22**	**1**
86	21.111	92	3.549	0.80539	6.6994	4.8192	10.5763	**773.92**	**1**

Note:
This table lists head of dataset cancer.

First of all is analysis the structure of data for know, what variables are relevant to realize training to different algorithms of machine learning (glm, loess, kne, random forest, ensamble).

I. Data structure

This dataset originates from a deep learning model trained on the "Coimbra Breast Cancer" dataset, with feature distributions closely resembling the original. The original data includes clinical observations from 64 patients with breast cancer and 52 healthy controls, encompassing 10 quantitative predictors and a binary dependent variable indicating the presence or absence of breast cancer.

Quantitative Attributes:

- **Age (years):** Represents the age of individuals in the dataset.
- **BMI (kg/m²):** Body Mass Index, a measure of body fat based on weight and height.
- **Glucose (mg/dL):** Reflects blood glucose levels, a vital metabolic indicator.
- **Insulin (µU/mL):** Indicates insulin levels, a hormone associated with glucose regulation.
- **HOMA:** Homeostatic Model Assessment, a method assessing insulin resistance and beta-cell function.
- **Leptin (ng/mL):** Represents leptin levels, a hormone involved in appetite and energy balance regulation.

- **Adiponectin (µg/mL):** Reflects adiponectin levels, a protein associated with metabolic regulation.
- **Resistin (ng/mL):** Indicates resistin levels, a protein implicated in insulin resistance.
- **MCP-1 (pg/dL):** Reflects Monocyte Chemoattractant Protein-1 levels, a cytokine involved in inflammation.

Labels classification:

1: Healthy controls.

2: Patients with breast cancer.

These quantitative attributes, including anthropometric data and parameters gathered from routine blood analysis, serve as the foundation for potential biomarkers of breast cancer. The dataset presents an opportunity for developing accurate prediction models, aiding in the identification and understanding of factors associated with breast cancer.

```
#Check the structure of the data
str(breast_train)
```

```
## 'data.frame':    92 obs. of  10 variables:
##  $ Age           : int  48 83 82 68 86 49 76 73 75 34 ...
##  $ BMI           : num  23.5 20.7 23.1 21.4 21.1 ...
##  $ Glucose       : int  70 92 91 77 92 92 118 97 83 78 ...
##  $ Insulin       : num  2.71 3.12 4.5 3.23 3.55 ...
##  $ HOMA          : num  0.467 0.707 1.01 0.613 0.805 ...
##  $ Leptin        : num  8.81 8.84 17.94 9.88 6.7 ...
##  $ Adiponectin   : num  9.7 5.43 22.43 7.17 4.82 ...
##  $ Resistin      : num  8 4.06 9.28 12.77 10.58 ...
##  $ MCP.1         : num  417 469 555 928 774 ...
##  $ Classification: int  1 1 1 1 1 1 1 1 1 1 ...
```

II. Summaries data

We will use summary function to obtain min, 1st quantil, median, mean, 3rd quantil, max to age, bmi, glucose, insulin, homa, leptin, adiponectin, resistin, MCP.1, classification.

```
############################################################
# II. Summaries data
############################################################

#Calculates the following summary statistics for the data frame
summary_breast_train <- summary(breast_train)

#The important statics is rating min, 1st Quantile, median, mean, 3rd Quantil, max
kable(summary_breast_train, caption = "Statics breast cancer") %>%
  kable_styling(bootstrap_options = c("striped", "hover"),fixed_thead = T, html_font = "Tahoma") %>%
  row_spec(3:4, bold = TRUE, color = "#B22222") %>%
  footnote(general = "This table lists(Age, BMI, Glucose, Insulin, HOMA, Leptin,
           Adiponectin, Resistin, MCP.1, Classification, min, 1st Quantile, median, mean, 3rd Quantiles
```

Table 3: Statics breast cancer

Age	BMI	Glucose	Insulin	HOMA	Leptin	Adipone
Min. :24.0	Min. :18.7	Min. : 60.0	Min. : 2.43	Min. : 0.467	Min. : 4.31	Min. : 2
1st Qu.:45.0	1st Qu.:22.9	1st Qu.: 85.8	1st Qu.: 4.03	1st Qu.: 0.822	1st Qu.:12.01	1st Qu.:
Median :53.5	**Median :27.4**	**Median : 92.0**	**Median : 5.80**	**Median : 1.318**	**Median :18.88**	**Median**
Mean :56.4	**Mean :27.5**	**Mean : 96.1**	**Mean : 9.45**	**Mean : 2.472**	**Mean :25.67**	**Mean** :
3rd Qu.:69.0	3rd Qu.:31.2	3rd Qu.:100.0	3rd Qu.:11.01	3rd Qu.: 2.858	3rd Qu.:36.22	3rd Qu.
Max. :86.0	Max. :38.6	Max. :199.0	Max. :51.81	Max. :25.050	Max. :90.28	Max. :3

Note:

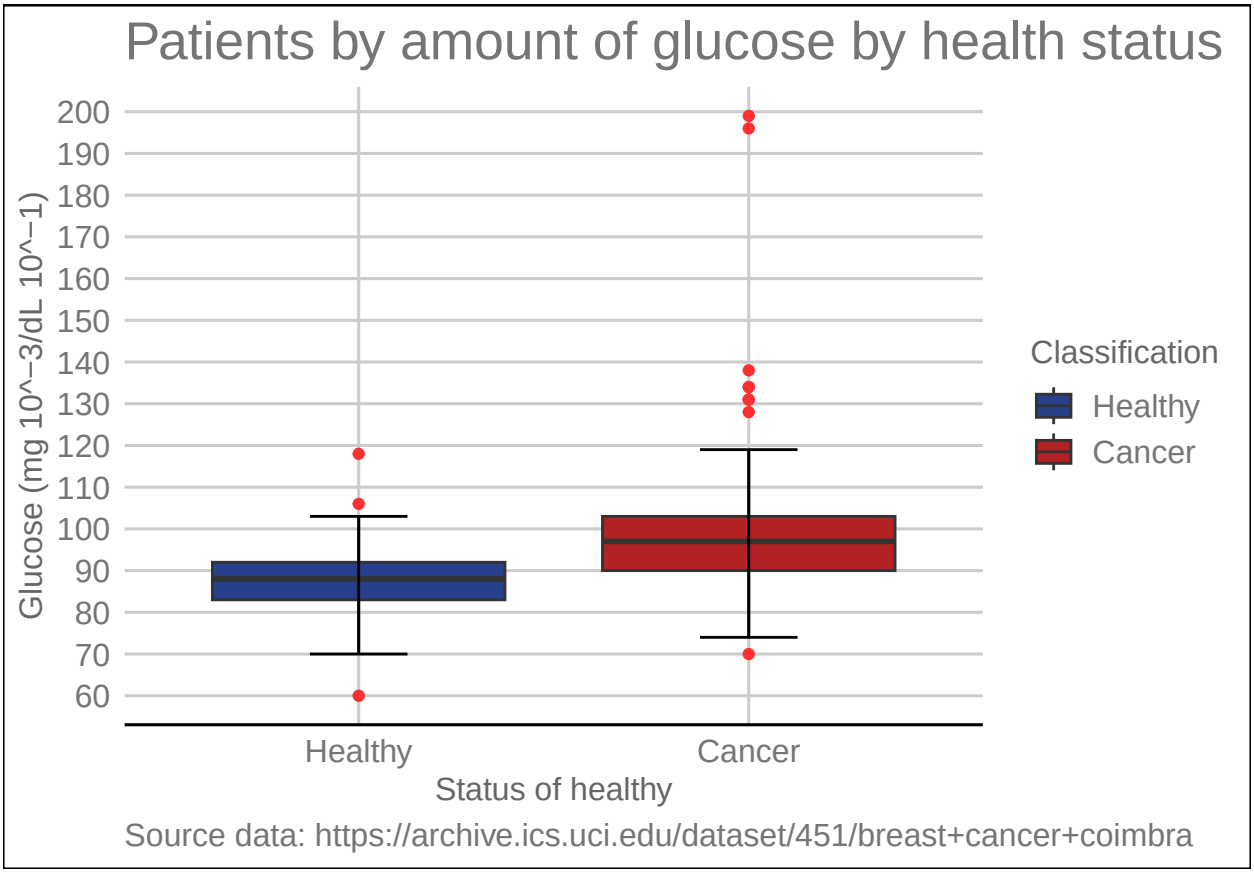
makecell[l]This table lists(Age, BMI, Glucose, Insulin, HOMA, Leptin, Adiponectin, Resistin, MCP.1, Classification, min,

III. Study graphics statistics

The graphs that may be relevant to us are to make a comparison between patients with cancer and those without cancer and to see which predictors are higher in those with cancer.

a. Patients by amount of glucose by health status

```
############################################################
# III. Study graphics statistics
############################################################

############################################################
# a. Patients by amount of glucose by health status
############################################################

#Convert to breast train and breast test field classification in factor value to distinct 1- Healthy 2-
breast_train <- breast_train %>% mutate(Classification = as.factor(Classification))
breast_test <- breast_test %>% mutate(Classification = as.factor(Classification))
levels(breast_train$Classification) <- c("Healthy","Cancer")
levels(breast_test$Classification) <- c("Healthy","Cancer")

ggplot(breast_train, aes(x=Classification, y=Glucose, fill=Classification)) +
  geom_boxplot(outlier.colour = "#FF3030") +
  stat_boxplot(geom = "errorbar",
               width = 0.25) +
  scale_fill_manual(breaks = waiver(),
                    values = c("#27408B", "#B22222")) +
  labs(title="Patients by amount of glucose by health status",
       caption="Source data: https://archive.ics.uci.edu/dataset/451/breast+cancer+coimbra",
       y="Glucose (mg 10^-3/dL 10^-1)",
       x="Status of healthy") +
  scale_y_continuous(breaks = seq(50, 250, by = 10)) +
  theme_gdocs()
```

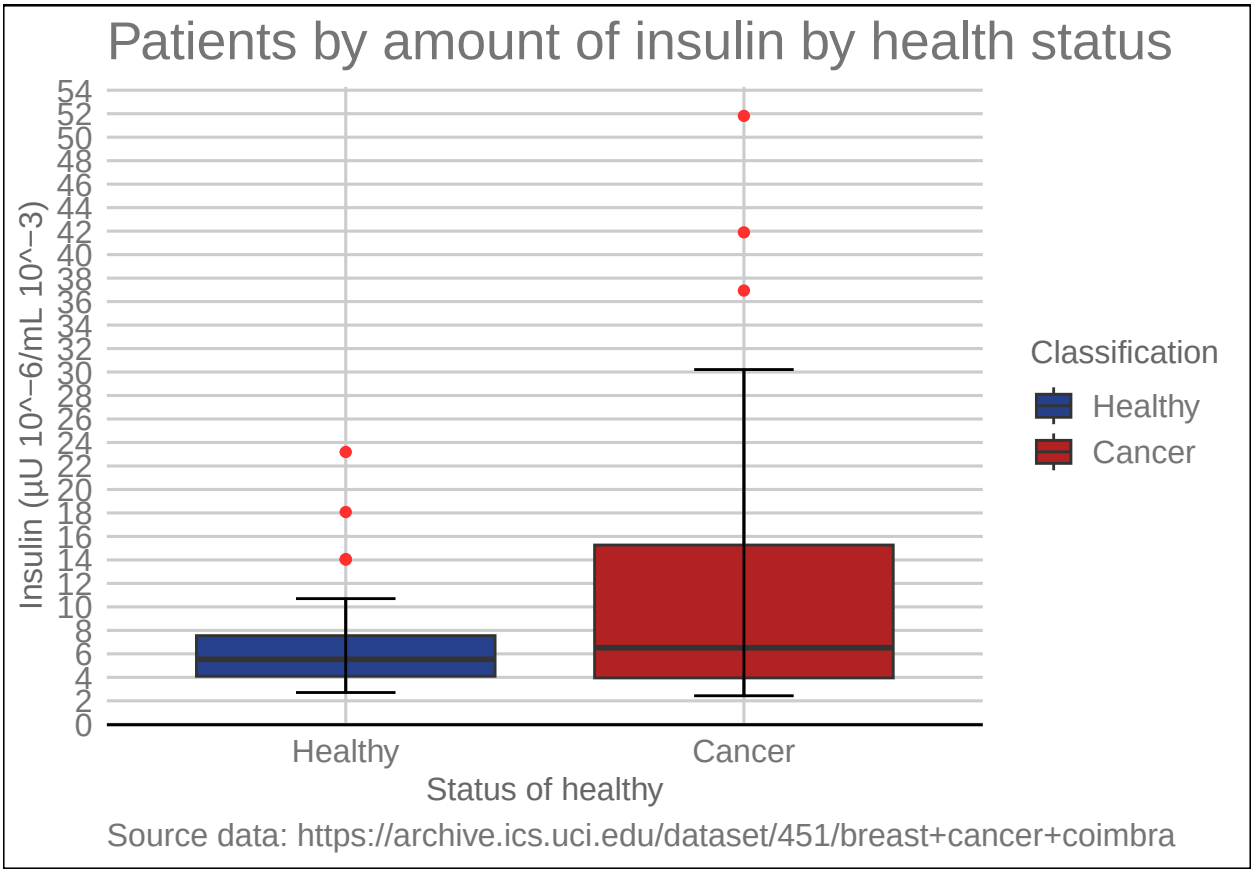


The glucose is a vital metabolic indicator for this reason is a predictor very use full to predict with high accuracy a patient could have a cancer. We have this affirmation for this reason; when a person have malignant cells increase consumption to glucose. These cells need more energy because they reproduce in an uncontrolled manner and increase activity metabolic.

In this box plot, we observe of glucose values are more higher than cancer patients.

b. Patients by amount of insulin by health status

```
#########################################################
# b. Patients by amount of insulin by health status
#########################################################

ggplot(breast_train, aes(x=Classification, y=Insulin, fill=Classification)) +
  geom_boxplot(outlier.colour = "#FF3030") +
  stat_boxplot(geom = "errorbar",
               width = 0.25) +
  scale_fill_manual(breaks = waiver(),
                    values = c("#27408B", "#B22222")) +
  labs(title="Patients by amount of insulin by health status",
       caption="Source data: https://archive.ics.uci.edu/dataset/451/breast+cancer+coimbra",
       y="Insulin (µU 10^-6/mL 10^-3)",
       x="Status of healthy") +
  scale_y_continuous(breaks = seq(0, 60, by = 2)) +
  theme_gdocs()
```



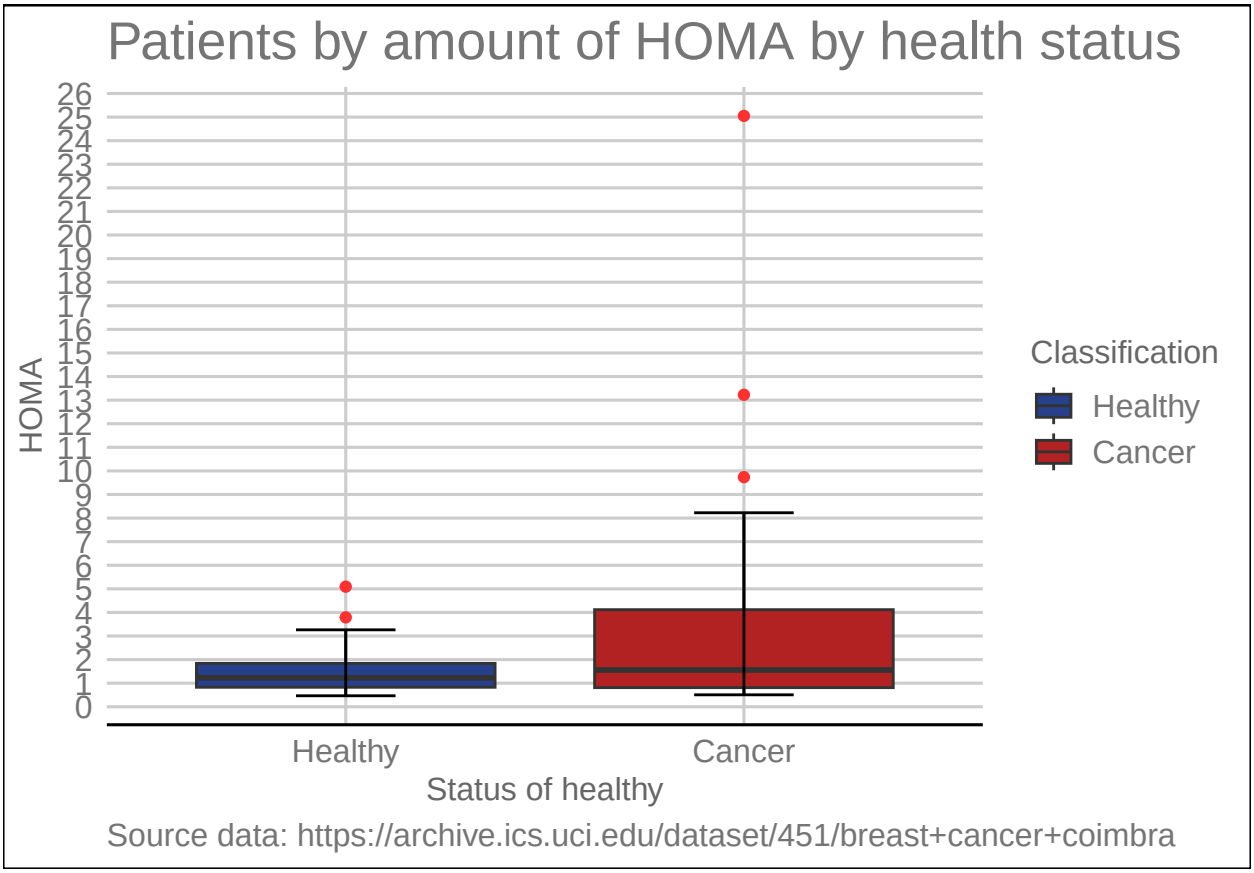
The insulin is a hormone associated with glucose regulation therefore is a predictor correlation to glucose. In this case, produces a reaction chain; malignant cells produce glucose and this produce more insulin. This predictor also a very use full to detect cancer in patients with more accuracy.

For this study, we are missing an important piece of information from patients (field diabetes). If they have diabetes, we need to know the type of diabetes (field type diabetes).

In this box plot, we observe of insulin values are more higher than cancer patients.

c. Patients by amount of HOMA by health status

```
############################################################
# c. Patients by amount of HOMA by health status
############################################################

ggplot(breast_train, aes(x=Classification, y=HOMA, fill=Classification)) +
  geom_boxplot(outlier.colour = "#FF3030") +
  stat_boxplot(geom = "errorbar",
               width = 0.25) +
  scale_fill_manual(breaks = waiver(),
                    values = c("#27408B", "#B22222")) +
  labs(title="Patients by amount of HOMA by health status",
       caption="Source data: https://archive.ics.uci.edu/dataset/451/breast+cancer+coimbra",
       y="HOMA",
       x="Status of healthy") +
  scale_y_continuous(breaks = seq(0, 30, by = 1)) +
  theme_gdocs()
```

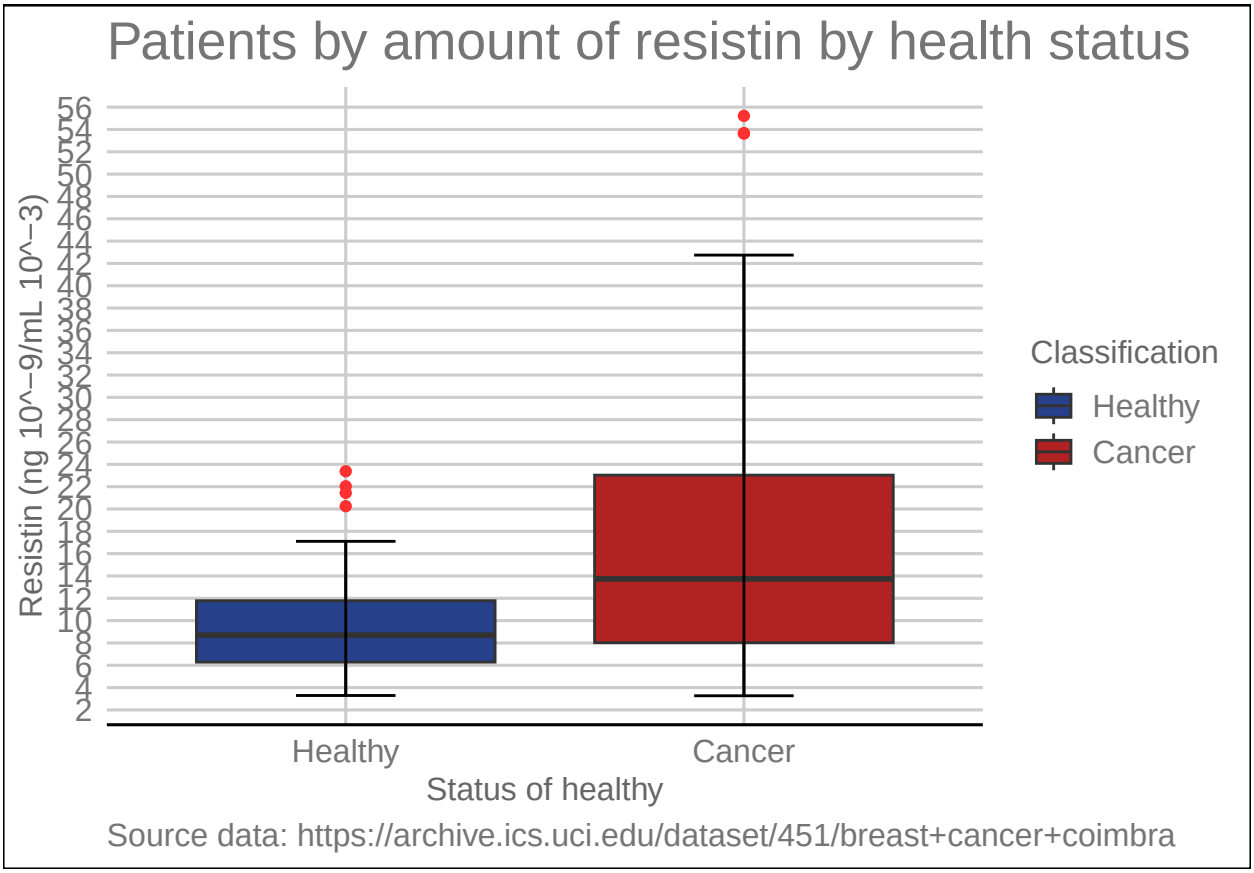


Homeostatic Model Assessment (HOMA), a method assessing insulin resistance and beta-cell function. HOMA is a predictor to correlation to insulin and other reaction change produce to cancer. This predictor also a very use full to detect cancer in patients with more accuracy.

In this box plot, we observe of HOMA values are more higher than cancer patients.

d. Patients by amount of resistin by health status

```
############################################################
# d. Patients by amount of resistin by health status
############################################################

ggplot(breast_train, aes(x=Classification, y=Resistin, fill=Classification)) +
  geom_boxplot(outlier.colour = "#FF3030") +
  stat_boxplot(geom = "errorbar",
               width = 0.25) +
  scale_fill_manual(breaks = waiver(),
                    values = c("#27408B", "#B22222")) +
  labs(title="Patients by amount of resistin by health status",
       caption="Source data: https://archive.ics.uci.edu/dataset/451/breast+cancer+coimbra",
       y="Resistin (ng 10^-9/mL 10^-3)",
       x="Status of healthy") +
  scale_y_continuous(breaks = seq(0, 90, by = 2)) +
  theme_gdocs()
```

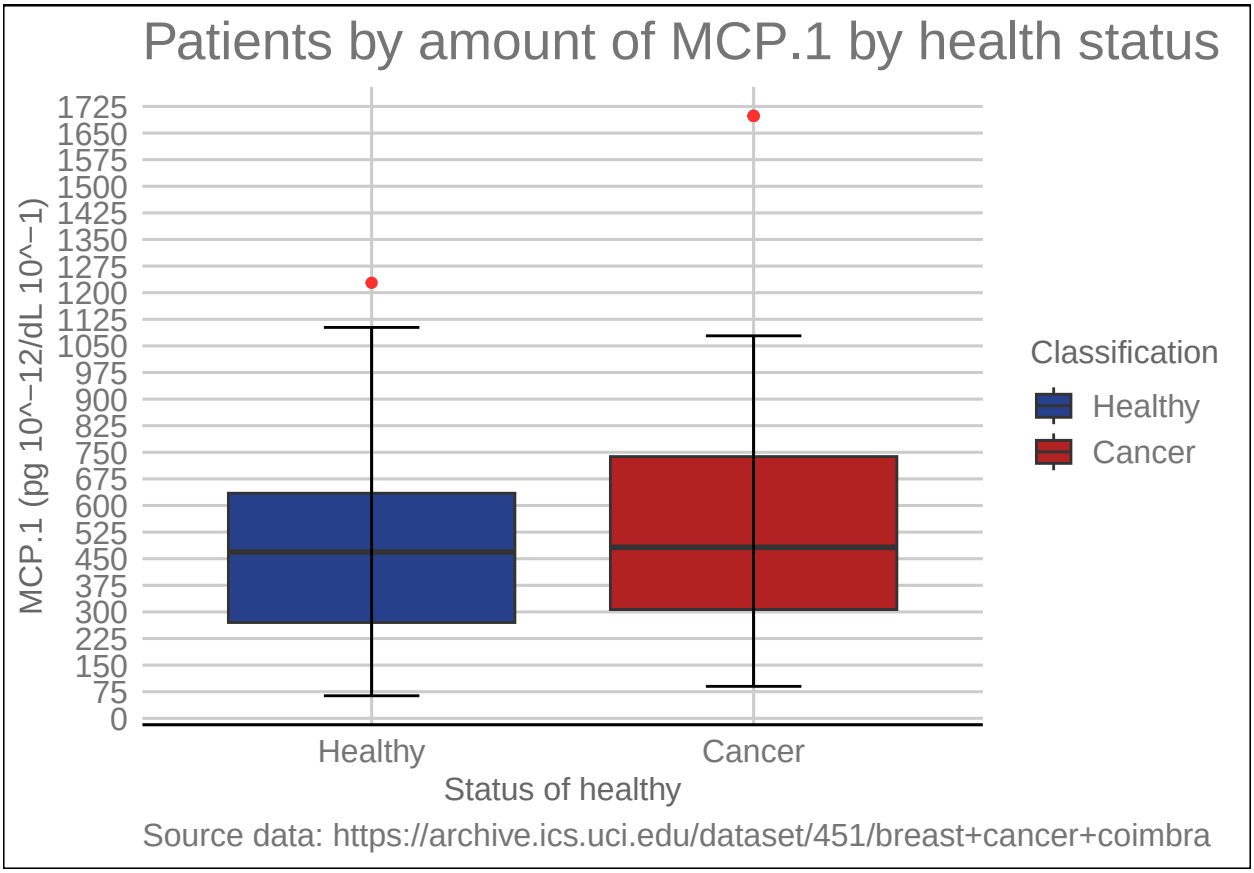


Resistin is a protein implicated in insulin resistance. Resistin is a predictor correlation to insulin and other reaction change produce to malignant cells.

In this box plot, we observe of resistin values are more higher than cancer patients.

e. Patients by amount of MCP.1 by health status

```
############################################################
# e. Patients by amount of MCP.1 by health status
############################################################

ggplot(breast_train, aes(x=Classification, y=MCP.1, fill=Classification)) +
  geom_boxplot(outlier.colour = "#FF3030") +
  stat_boxplot(geom = "errorbar",
               width = 0.25) +
  scale_fill_manual(breaks = waiver(),
                    values = c("#27408B", "#B22222")) +
  labs(title="Patients by amount of MCP.1 by health status",
       caption="Source data: https://archive.ics.uci.edu/dataset/451/breast+cancer+coimbra",
         y="MCP.1 (pg 10^-12/dL 10^-1)",
       x="Status of healthy") +
  scale_y_continuous(breaks = seq(0, 1800, by = 75)) +
  theme_gdocs()
```



Monocyte Chemoattractant Protein-1 (MCP.1) is a cytokine involved in inflammation. This predictor does not provide a conclusive predictor. What I observe is very strange: MCP.1 should be much higher in patients with cancer due to an inflammatory process caused by cancer cells and treatment. However, the boxplot shows that the median is the same in healthy patients or patients with cancer.

MCP.1 should produce an inflammatory reaction when detecting cancer cells. This antibody (monocyte) does not detect cancer cells, which is why it is similar in both healthy patients and those with cancer.

IV. Conclusion study graphic

The predictors most likely to produce high-precision detection cancer because cancer cells produce a reaction chain of correlation to predictors:

$glucose(indicatormetabolic) \rightarrow insulin(hormone) \rightarrow HOMA(betacellfunction) \rightarrow resistin(protein)$

3. MACHINE LEARNING

Machine learning is a scientific field and, more specifically, a subcategory of artificial intelligence. It consists of letting algorithms discover "patterns" recurring patterns in data sets. That data can be numbers, words, images, statistics, etc.

Anything that can be stored digitally can serve as data for machine learning. By detecting patterns in that data, algorithms learn and improve their performance in executing a specific task.

In short, machine learning algorithms autonomously learn to perform a task or make predictions from data and improve their performance over time. Once trained, the algorithm will be able to find the patterns in new data.

I. PCA (principal component analysis)

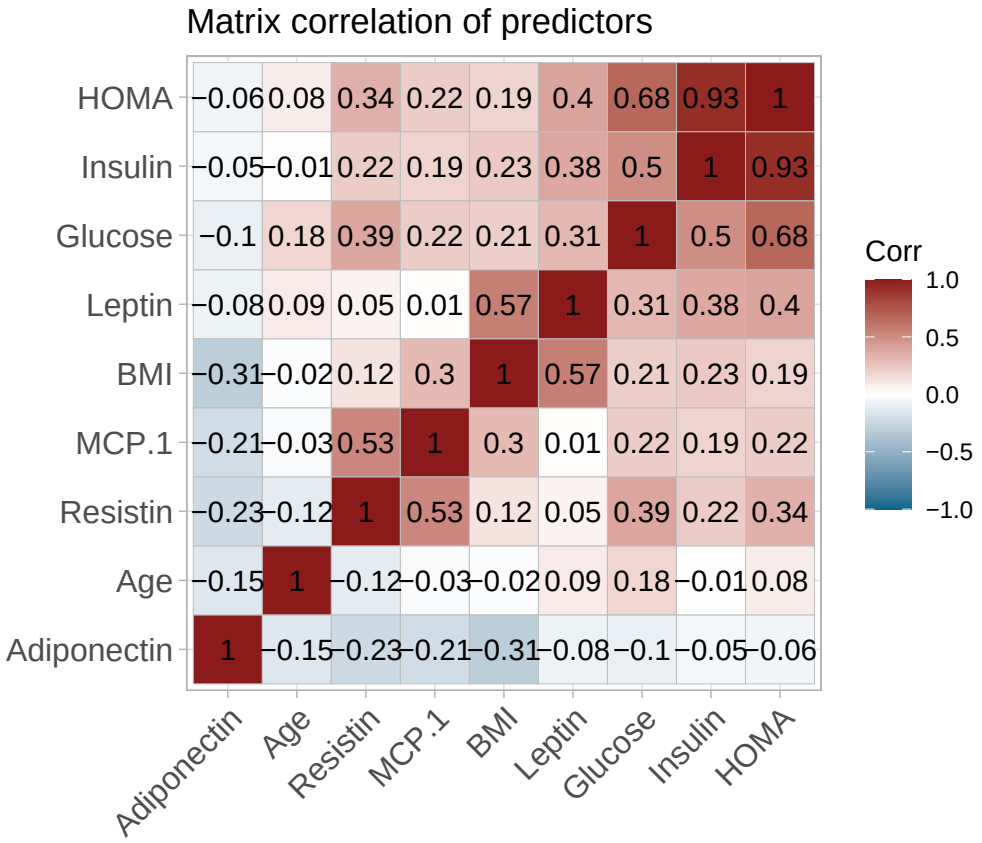
It is a statistical approach that can be used to analyze high-dimensional data and capture the most important information from it. This is done by transforming the original data into a lower-dimensional space and grouping the highly correlated variables.

PCA can be considered as a rotation of the axes of the coordinate system of the original variables to new orthogonal axes, so that these axes coincide with the direction of maximum variance of the data.

```
############################################################
# 3. MACHINE LEARNING
############################################################

############################################################
# I. PCA (principal component analysis)
############################################################

#Convert to breast train and breast test field age and glucose in numeric value to scale
breast_train <- breast_train %>% mutate(Age = as.numeric(Age))
breast_train <- breast_train %>% mutate(Glucose = as.numeric(Glucose))

breast_test <- breast_test %>% mutate(Age = as.numeric(Age))
breast_test <- breast_test %>% mutate(Glucose = as.numeric(Glucose))


PCA <- breast_train %>% select(-Classification)

#check is not null values in PCA data
colSums(is.na(PCA))
```

```
##         Age         BMI     Glucose     Insulin        HOMA      Leptin
##           0           0           0           0           0           0
## Adiponectin    Resistin       MCP.1
##           0           0           0
```

```
#normalization using function scale
data_normalized <- scale(PCA)

#calculate correlation of matrix
corr_matrix <- cor(data_normalized)

#Correlation plot
ggcorrplot(corr_matrix,
           method = "square",
           hc.order = TRUE,
           type = "full",
           lab = TRUE,
           title = "Matrix correlation of predictors",
           ggtheme = ggplot2::theme_light(),
           colors = c("#00688B", "white", "#8B1A1A"))
```



In this matrix correlation corroborates relationship between:

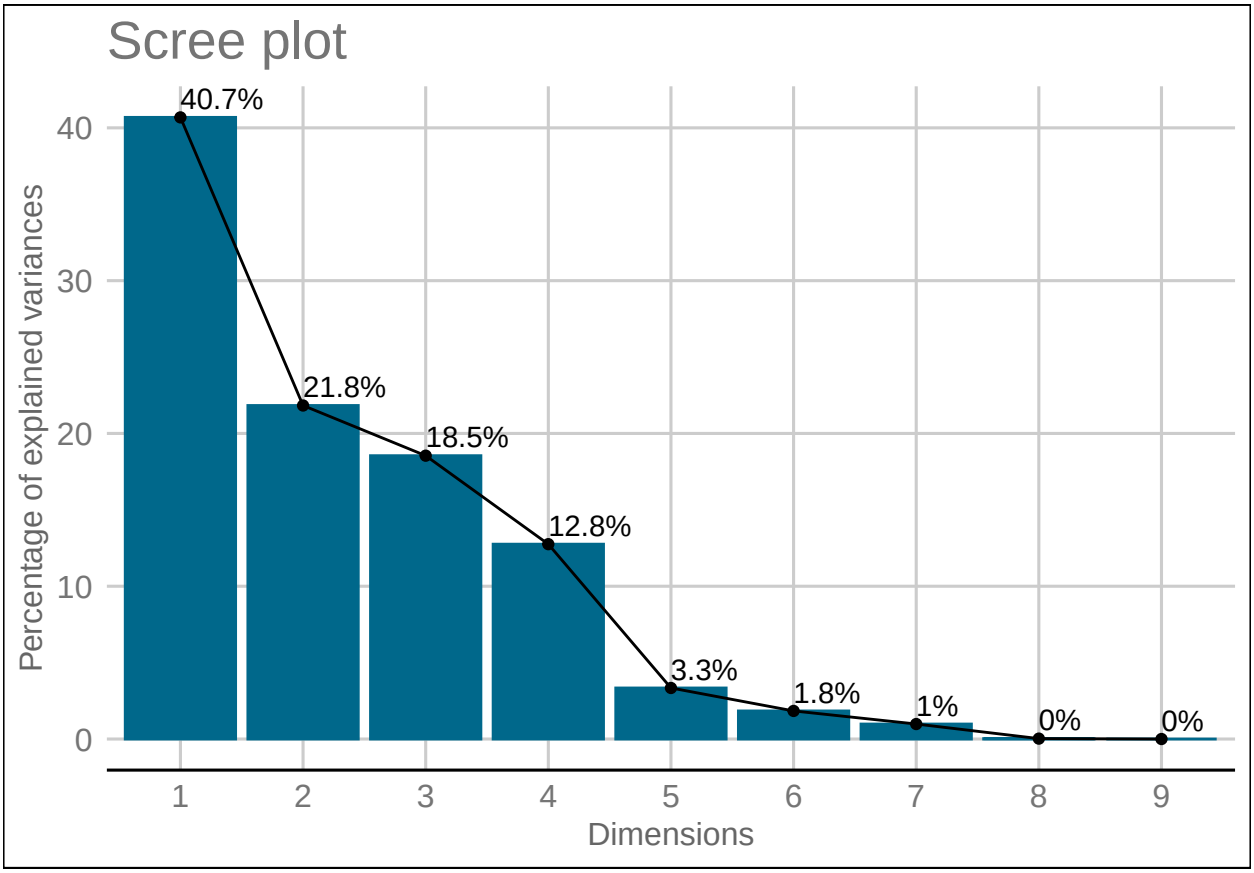
- Homa to insulin (0.93) and glucose (0.63) .
- Insulin to Homa (0.93) and glucose (0.50).
- Glucose to to Homa (0.68) and insulin (0.50).

```
#summary importance of components
data.pca <- princomp(corr_matrix)
summary(data.pca)
```

```
## Importance of components:
##                           Comp.1  Comp.2  Comp.3  Comp.4   Comp.5   Comp.6
## Standard deviation     0.65103 0.47687 0.43952 0.36448 0.186567 0.138379
## Proportion of Variance 0.40686 0.21829 0.18544 0.12753 0.033413 0.018382
## Cumulative Proportion  0.40686 0.62516 0.81060 0.93813 0.971541 0.989923
##                           Comp.7     Comp.8 Comp.9
## Standard deviation     0.1011039 0.01660677      0
## Proportion of Variance 0.0098126 0.00026474      0
## Cumulative Proportion  0.9997353 1.00000000      1
```

```
#Scree plot
fviz_eig(data.pca,
         addlabels = TRUE,
         barfill = "#00688B",
```

```
          barcolor = "#00688B",
          linecolor = "black") +
    theme_gdocs()
```

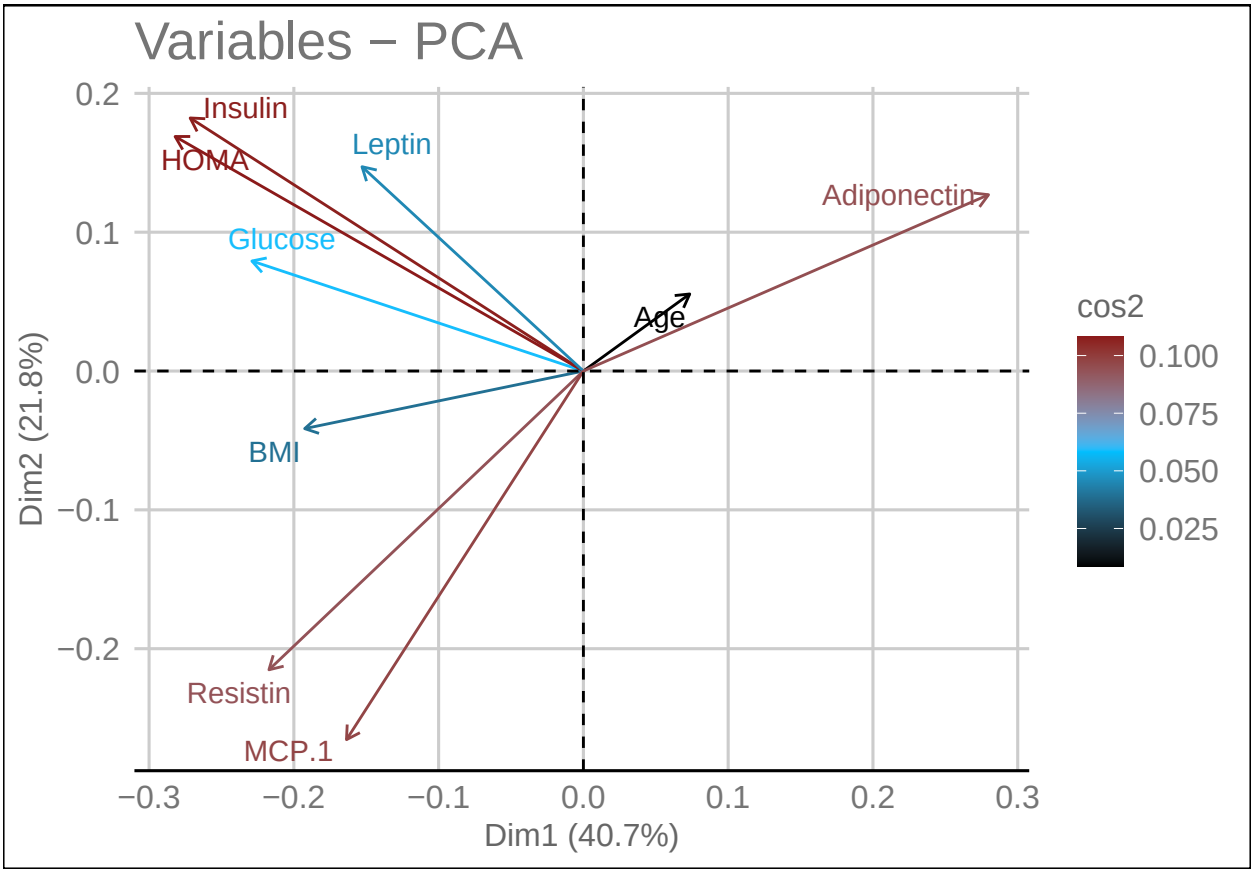


We observe that nine principal components have been generated (Comp.1 to Comp.9), which also correspond to the number of predictors in the breast test. Each component explains a percentage of the total variance in the data set.

In the Cumulative Proportion section, the first principal component explains almost 40.7% of the total variance. This implies that almost half of the data in the set of 9 predictors can be represented by the first principal component alone. The second dimension explains 21.80% of the total variance. The third dimension explains 18.50% of the total variance.

The cumulative proportion Comp.1 to Comp.3 explains almost 81.20% of the total variance.

```
#Graphic pca var
fviz_pca_var(data.pca, col.var = "cos2",
             gradient.cols = c("black", "#00BFFF", "#8B1A1A"),
             repel = TRUE) +
    theme_gdocs()
```



In this biplot we could observe in the right up quadrant are age and adiponect in opposite correlation left down quadrant are BMI, resistin and MCP.1 have high correlation . Finally, left up quadrant we could observe high correlation between homa, insulin.

II. Glm (logistic regression)

GLM (Generalized linear models) are a type of statistical model that is extensively used in the analysis of non-normal data, such as count data or binary data. They enable us to describe the connection between one or more predictor variables and a response variable in a flexible manner.

```
############################################################
# II. Glm (logistic regression)
############################################################

#Create train component x with all predictors except Classification that is a factor use en component y
breast_train_x <- breast_train %>% select(-Classification)

#Normalization train predictors
breast_train_x <- scale(breast_train_x)

#Create train component y to function train to healthy status 1-healthy 2-cancer
breast_train_y <- breast_train$Classification

#Create test component x with all predictors except Classification that is a factor use en component y
breast_test_x <- breast_test %>% select(-Classification)
```

```r
#Normalization test predictors
breast_test_x <- scale(breast_test_x)

#Create test component y to function train to healthy status 1-healthy 2-cancer
breast_test_y <- breast_test$Classification

#Calculate accuracy to detect cancer with 9 predictors to use glm
set.seed(1, sample.kind = "Rounding")

#Use train function to method glm, pass 9 predictors x argument
#and Classification train healthy status 1-healthy 2-cancer y argument
train_glm <- train(breast_train_x, breast_train_y, method = "glm")

#Use predict function to pass train_glm and 9 predictors to test data
glm_preds <- predict(train_glm, breast_test_x)

#To obtain accuracy compare mean to glm predictions results to Classification test 1-healthy 2-cancer
mean(glm_preds == breast_test_y)
```

```
## [1] 0.79167
```

III. Loess (local polynomial regression fitting)

Loess thus build on classical methods, such as linear and nonlinear least squares regression. Loess combines much of the simplicity of linear least squares regression with the flexibility of nonlinear regression. It does this by fitting simple models to localized subsets of the data to build up a function that describes the deterministic part of the variation in the data, point by point.

```r
##########################################################
# III. Loess (local polynomial regression fitting)
##########################################################

#Calculate accuracy to detect cancer with 9 predictors to use loess
set.seed(2, sample.kind = "Rounding")

#Use train function to method gamLoess, pass 9 predictors x argument
#and Classification train  healthy status 1-healthy 2-cancer y argument
train_loess <- train(breast_train_x, breast_train_y, method = "gamLoess")

#Use predict function to pass train_loess and 9 predictors to test data
loess_preds <- predict(train_loess, breast_test_x)

#To obtain accuracy compare mean to loess predictions results to Classification test 1-healthy 2-cancer
mean(loess_preds == breast_test_y)
```

```
## [1] 0.66667
```

IV. K nearest neighbors

In statistics, the k-nearest neighbors algorithm (k-NN) is a non-parametric supervised learning method. Most often, it is used for classification, as a k-NN classifier, the output of which is a class membership.

An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor.

The k-NN algorithm can also be generalized for regression. In k-NN regression, also known as nearest neighbor smoothing, the output is the property value for the object. This value is the average of the values of k nearest neighbors. If k = 1, then the output is simply assigned to the value of that single nearest neighbor, also known as nearest neighbor interpolation.

For both classification and regression, a useful technique can be to assign weights to the contributions of the neighbors, so that nearer neighbors contribute more to the average than distant ones. For example, a common weighting scheme consists of giving each neighbor a weight of 1/d, where d is the distance to the neighbor.

The input consists of the k closest training examples in a data set. The neighbors are taken from a set of objects for which the class (for k-NN classification) or the object property value (for k-NN regression) is known. This can be thought of as the training set for the algorithm, though no explicit training step is required.

A peculiarity (sometimes even a disadvantage) of the k-NN algorithm is its sensitivity to the local structure of the data. In k-NN classification the function is only approximated locally and all computation is deferred until function evaluation. Since this algorithm relies on distance, if the features represent different physical units or come in vastly different scales, then feature-wise normalizing of the training data can greatly improve its accuracy

```
############################################################
# IV. K nearest neihbours
############################################################

#Calculate accuracy to detect cancer with 9 predictors to use knn
set.seed(3, sample.kind = "Rounding")

#Use train function to method knn, pass 9 predictors x argument
#and Classification train healthy status 1-healthy 2-cancer y argument
tuning <- data.frame(k = seq(3, 21, 2))
train_knn <- train(breast_train_x, breast_train_y,
                   method = "knn",
                   tuneGrid = tuning)

#obtain the best parameter of sequence of tuning in this case 3
#train_knn$bestTune

#Use predict function to pass train_knn and 9 predictors to test data
knn_preds <- predict(train_knn, breast_test_x)

#To obtain accuracy compare mean to knn predictions results to Classification test 1-healthy 2-cancer
mean(knn_preds == breast_test_y)
```

```
## [1] 0.83333
```

V. Random forest

Random forests or random decision forests is an ensemble learning method for classification, regression and other tasks that works by creating a multitude of decision trees during training. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the output is the

average of the predictions of the trees. Random forests correct for decision trees' habit of overfitting to their training set.

```
##########################################################
# V. Random forest
##########################################################

#Calculate accuracy to detect cancer with 9 predictors to use random forest

set.seed(4, sample.kind = "Rounding")

#Use train function to method rf, pass 9 predictors x argument
#and Classification train healthy status 1-healthy 2-cancer y argument
tuning <- data.frame(mtry = c(3, 5, 7, 9))
train_rf <- train(breast_train_x, breast_train_y,
                  method = "rf",
                  tuneGrid = tuning,
                  importance = TRUE)

#obtain the best parameter of sequence of tuning in this case 9
#train_rf$bestTune

#Use predict function to pass rf and 9 predictors to test data
rf_preds <- predict(train_rf, breast_test_x)
#To obtain accuracy compare mean to rf predictions results to Classification test 1-healthy 2-cancer
mean(rf_preds == breast_test_y)
```

```
## [1] 0.83333
```

```
#the most important variable in the random forest model is glucose 100
varImp(train_rf)
```

```
## rf variable importance
##
##             Importance
## Glucose          100.0
## Resistin          92.3
## BMI               77.1
## Age               76.4
## Leptin            34.4
## HOMA              25.0
## Adiponectin       12.0
## Insulin           11.2
## MCP.1              0.0
```

The method of machine learning random forest indicate that most important variable is glucose to 100% importance.

VI. Ensemble

In machine learning, one can usually greatly improve the final results by combining the results of different algorithms to obtain a better estimate accuracy.

We compute new class probabilities by taking the average of glm (logistic regression), loess (logical polynominal regression fit), knn(k nearest neighbors) and random forest.

```
##########################################################
# VI. Ensamble
##########################################################

#Calculate accuracy to detect cancer with 9 predictors to use ensamble

#Create a logic matrix to results glm, loess, knn and rf with 24 patients in test data
#results FALSE = healthy and TRUE = cancer
ensemble <- cbind(glm = glm_preds == "Healthy", loess = loess_preds == "Healthy",
                  knn = knn_preds == "Healthy", rf = rf_preds == "Healthy")

#calculate mean to row (glm,loess, knn,rf) to obtain result ensamble
ensemble_preds <- ifelse(rowMeans(ensemble) > 0.5, "Healthy", "Cancer")

#To obtain accuracy compare mean to ensamble predictions to Classification test 1-healthy 2-cancer
mean(ensemble_preds == breast_test_y)
```

```
## [1] 0.79167
```

4. RESULTS

```
#Create a data frame with all results (glm, loess, knn, rf, ensemble)
models <- c("Logistic regression", "Loess", "K nearest neighbors", "Random forest", "Ensemble")
accuracy <- c(mean(glm_preds == breast_test_y),
              mean(loess_preds == breast_test_y),
              mean(knn_preds == breast_test_y),
              mean(rf_preds == breast_test_y),
              mean(ensemble_preds == breast_test_y))
results <- data.frame(Model = models, Accuracy = accuracy)

#show the results
kable(results, caption = "Accuracy to detect breast cancer through 9 biomarkers") %>%
  kable_styling(bootstrap_options = c("striped", "hover"), html_font = "Tahoma" ) %>%
  footnote(general = "Results of differents methods of machine learning use.")
```

Table 4: Accuracy to detect breast cancer through 9 biomarkers

Model	Accuracy
Logistic regression	0.79167
Loess	0.66667
K nearest neighbors	0.83333
Random forest	0.83333
Ensemble	0.79167

Note:
Results of differents methods of machine learning use.

5. CONCLUSION

We had realized to study (EDA-exploration data analysis) about "Coimbra Breast Cancer" dataset.

The original data includes clinical observations from 64 patients with breast cancer and 52 healthy controls, encompassing 10 quantitative predictors and a binary dependent variable indicating the presence or absence of breast cancer.

The data from the Coimbra hospital show 8 biomarkers-predictors of blood analysis (bmi, glucose, insulin, HOMA, leptin, adiponectin, resistin, MCP.1) plus a binary variable (classification - 1. Healthy control patient 2- Cancer patient) apart from age.

Through the graphs, detect the following correlation of predictors.

glucose(metabolic indicator)→insulin(hormone)→HOMA(beta cell function)→resistin(protein)

When a patient has cancer, their blood glucose levels are higher than those of healthy control patients. This is because cancer cells that reproduce in an uncontrolled manner need more glucose than normal for their metabolic activity.

This biomarker produces a chain reaction at a chemical level; as cancer cells increase, glucose increases insulin levels in the blood.

In turn, increasing insulin increases the function of beta cells (HOMA).

Also, increasing insulin increases resistance to it through resistin.

To optimize the accuracy of the analysis methods (machine learning) would be necessary:

1. **On the one hand, more patient records in this case we only have 116 records.**
2. **On the other hand, the more biomarkers the blood test has, more biomarker correlations we can detect.**
3. **Two more fields are also needed to know if the patient has diabetes or not and another that tells us the type I or II diabetes. This allows us to rule out false positives.**

That is why in the PET test (positron emission tomography) TAC (computerized axial tomography) they use the radio tracer F18 -Fluorodeoxyglucose that acts with positrons marking the areas of greater metabolic activity.

This test determines the type of cancer by its location to know which treatment or surgery to apply.

The final conclusion I draw from the 4 biomarkers (glucose, insulin, HOMA, resistin) that could serve as indicators to detect cancer, since they are all derived (correlations) from glucose, would be the one that works for us.

The advantage that this marker would not only be valid for breast cancer, it is valid for any type of cancer. This is because, regardless of the type of cancer cell, they all need more glucose than healthy cells.

Today, cancer screening tests can be done with blood tests using tumor markers, and glucose could be taken into account as another indicator. It must be taken into account that if the patient has diabetes, this marker would not be useful for detecting it.

If we use more biomarkers in blood tests, we will have more predictors to detect future cancer indicators.

The results obtain to accuracy in different machine learning methods are:

- **Logistic regression: 0.79167**

- **Loess: 0.66667**
- **K nearest neighbours: 0.83333**
- **Random forest:0.83333**
- **Ensemble (median combined LM, LOESS, KNN, RF): 0.79167**

6. REFERENCES

[1] Consultation of doubts and knowledge obtain to: https://www.wikipedia.org/