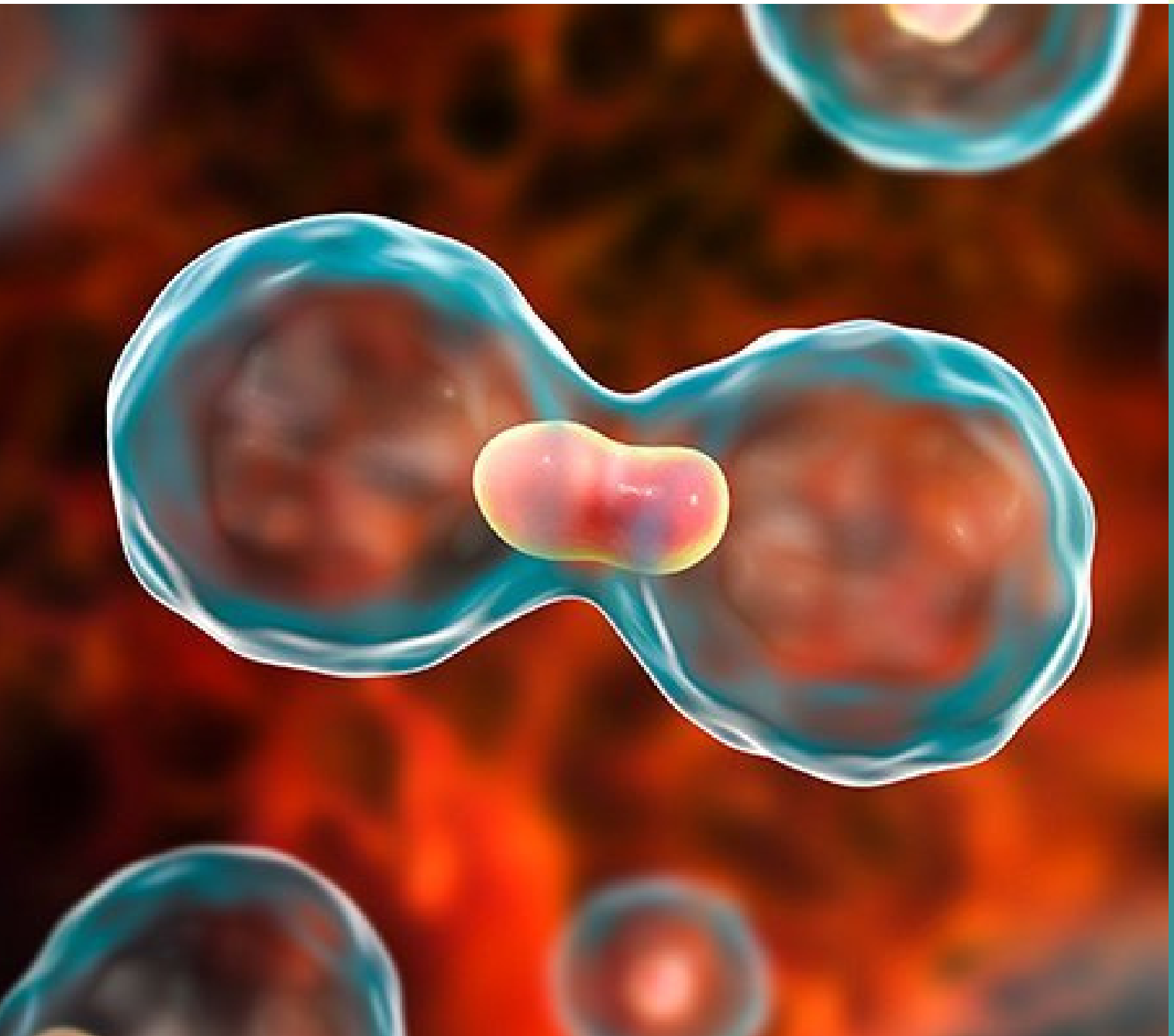
FIRST PROJECT

# BREAST CANCER COIMBRA

**KAMYAB ABEDI**

AIMedic - Summer 2021



# Table of Contents

# Introduction

Breast cancer is the most common malignancy among women worldwide. There is extensive literature on the relationship between body weight and breast cancer risk but some doubts still remain about the role of adipokines per se, the role of insulin and glucose regardless of obesity, as well as the crosstalk between these players. Thus, in this project, we intend to determine the relation between body mass index (BMI), glycaemia, insulinemia, insulin-resistance, blood adipokine levels and tumour.

Data Set Information:
n this dataset there are 10 predictors, all quantitative, and a binary dependent variable, indicating the presence or absence of breast cancer. The predictors are anthropometric data and parameters which can be gathered in routine blood analysis.
Prediction models based on these predictors, if accurate, can potentially be used as a biomarker of breast cancer.

Attribute Information:
This database contains Quantitative Attributes such as Age (years), BMI (kg/m2), Glucose (mg/dL), Insulin (µU/mL), Homeostasis Model Assessment, Leptin (ng/mL), Adiponectin (µg/mL), Resistin (ng/mL), MCP-1(pg/dL), ...
There are 2 class, class 0 for Healthy controls and class 1 is our Patients

Goal:
The goal of this exploratory study was to develop and assess a prediction model which can potentially be used as a biomarker of breast cancer, based on anthropometric data and parameters which can be gathered in routine blood analysis. Nowadays in a medical test, the big indicators of success are specificity and sensitivity. Every medical test strives to reach 100% in both criteria..

# Set up the notebook

Nothing special happens in this part, just load the dataset
Step 1 : Import the libraries
Step 2 : Import the data-set

| | Age | BMI | Glucose | Insulin | HOMA | Leptin | Adiponectin | Resistin | MCP.1 | Classification |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 48 | 23.500000 | 70 | 2.707 | 0.467409 | 8.8071 | 9.702400 | 7.99585 | 417.114 | 1 |
| **1** | 83 | 20.690495 | 92 | 3.115 | 0.706897 | 8.8438 | 5.429285 | 4.06405 | 468.786 | 1 |
| **2** | 82 | 23.124670 | 91 | 4.498 | 1.009651 | 17.9393 | 22.432040 | 9.27715 | 554.697 | 1 |
| **3** | 68 | 21.367521 | 77 | 3.226 | 0.612725 | 9.8827 | 7.169560 | 12.76600 | 928.220 | 1 |
| **4** | 86 | 21.111111 | 92 | 3.549 | 0.805386 | 6.6994 | 4.819240 | 10.57635 | 773.920 | 1 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **107** | 46 | 33.180000 | 92 | 5.750 | 1.304867 | 18.6900 | 9.160000 | 8.89000 | 209.190 | 2 |
| **108** | 68 | 35.560000 | 131 | 8.150 | 2.633537 | 17.8700 | 11.900000 | 4.19000 | 198.400 | 2 |
| **111** | 45 | 26.850000 | 92 | 3.330 | 0.755688 | 54.6800 | 12.100000 | 10.96000 | 268.230 | 2 |
| **112** | 62 | 26.840000 | 100 | 4.530 | 1.117400 | 12.4500 | 21.420000 | 7.32000 | 330.160 | 2 |
| **113** | 65 | 32.050000 | 97 | 5.730 | 1.370998 | 61.4800 | 22.540000 | 10.33000 | 314.050 | 2 |

# Data Preprocessing

Include:

- Data cleansing & editing
- Data exploration
- Data reduction
- Data wrangling

Data Cleaning:

Step 1 : Count and delete all rows with NULL values
Step 2 : Delete duplicate values
Step 3 : Find Negative value
Step 4 : Change column order to better perform splits

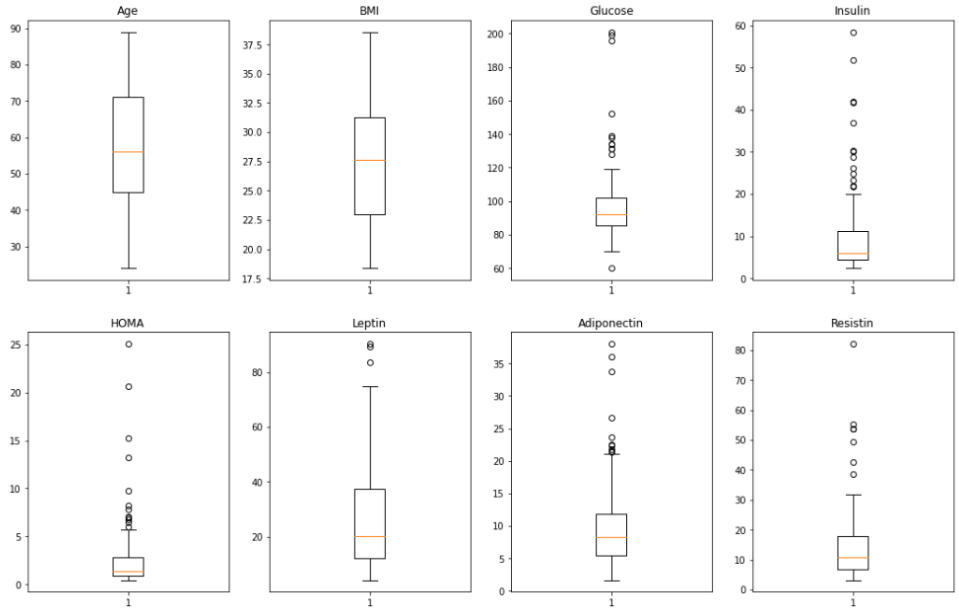
Data Exploration:

Step 1 : Get a concise summary of the dataset
Step 2 : See the Categorical Values

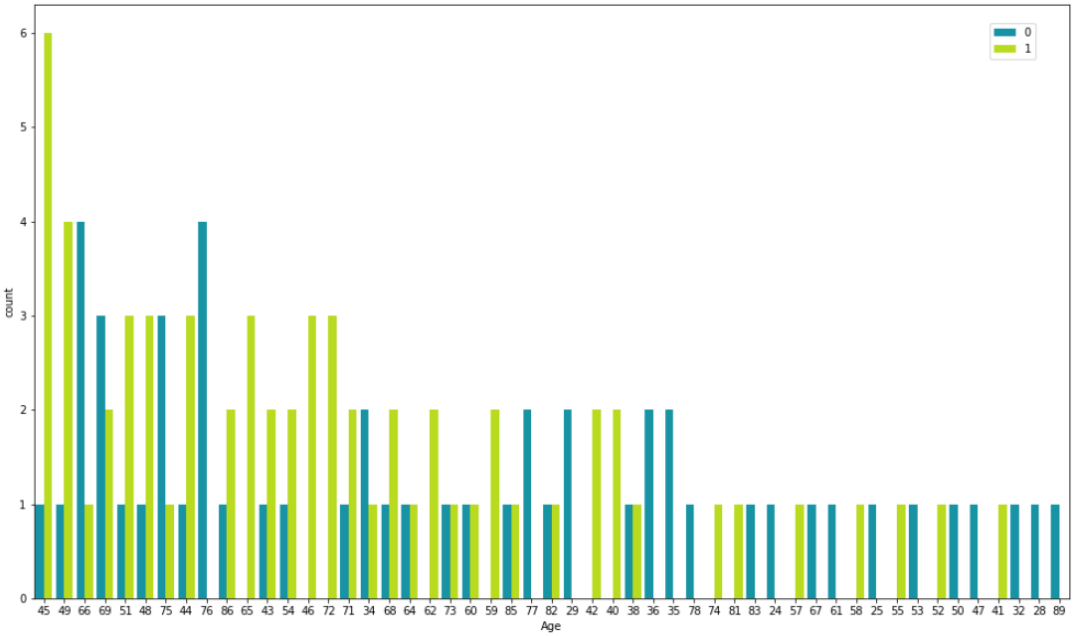
# Visualization

Include:

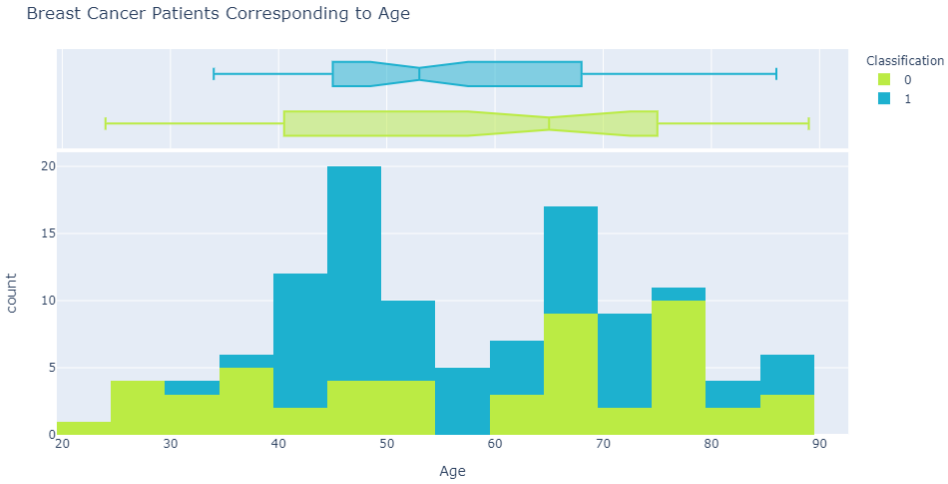
- Create a dictionary of columns representing the features of the dataset and Visualize the data for each feature using box plots.



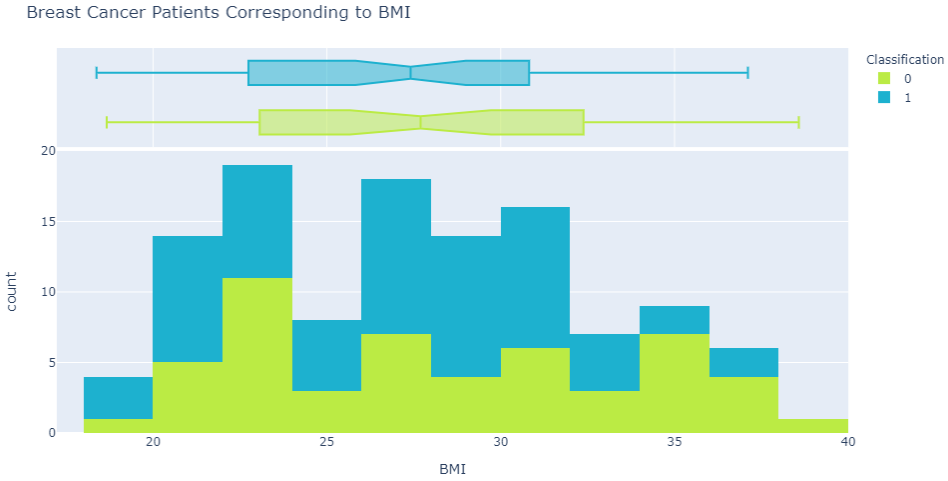

maybe you want to know Which Age related to Patients or Healty Control Health?



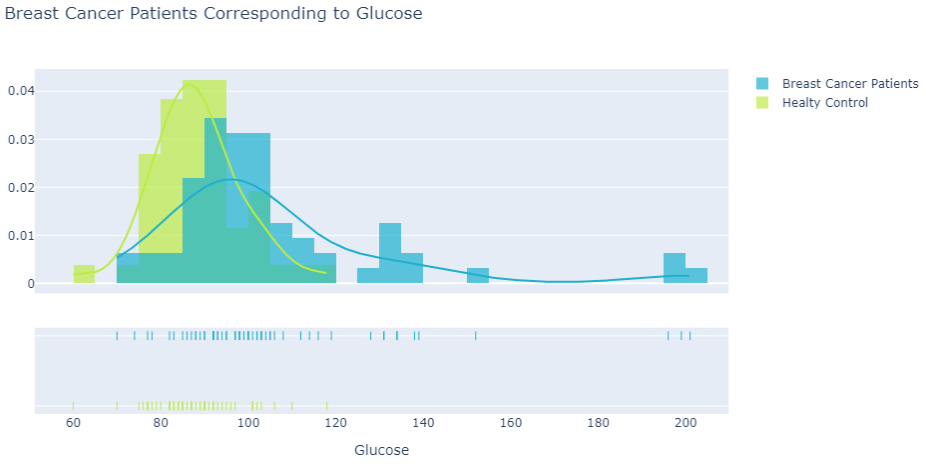
- Breast Cancer Patients Corresponding to Age



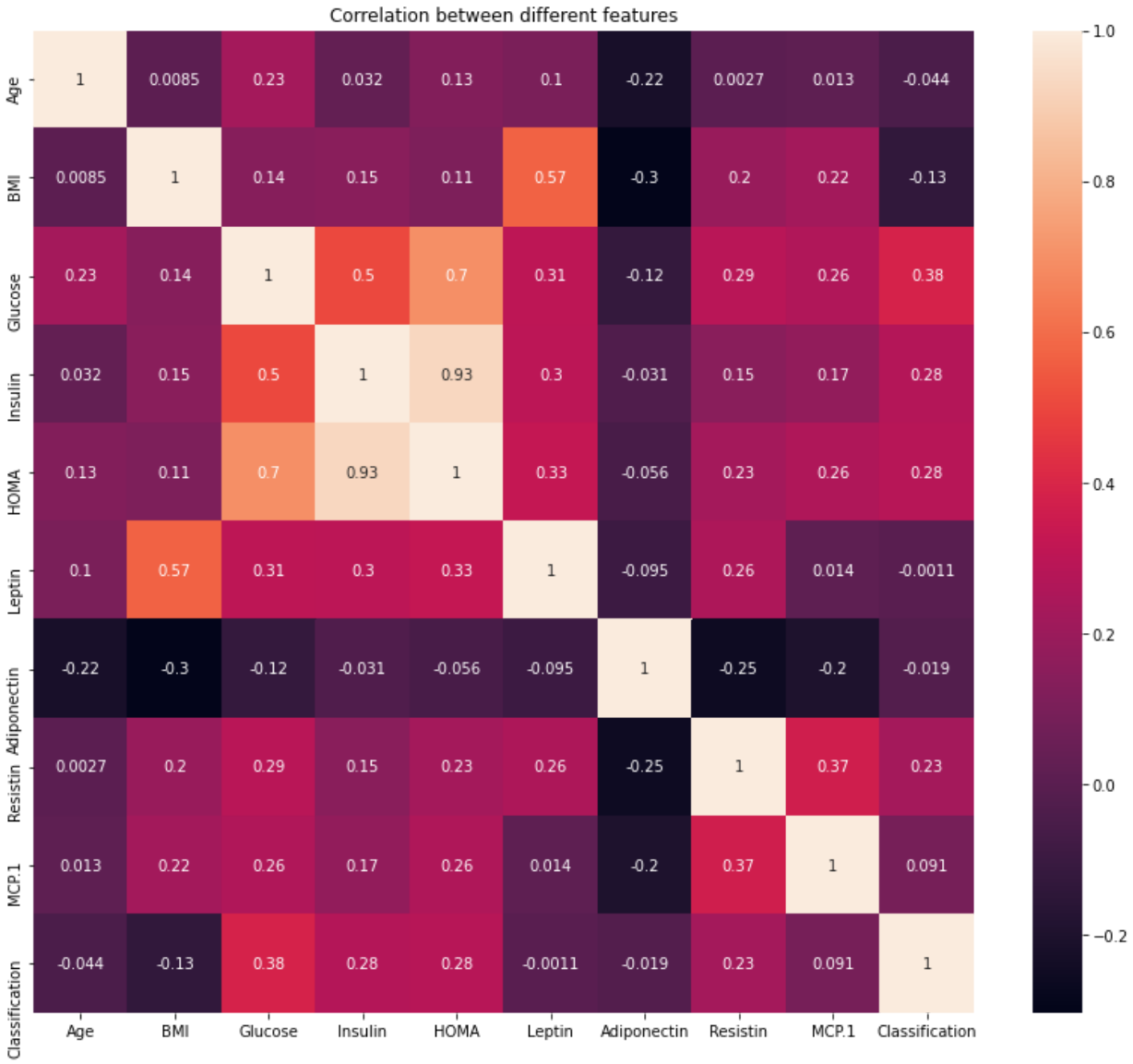

- Breast Cancer Patients Corresponding to BMI




- Breast Cancer Patients Corresponding to Glucose



- Plot heatmap to visualize the correlations.



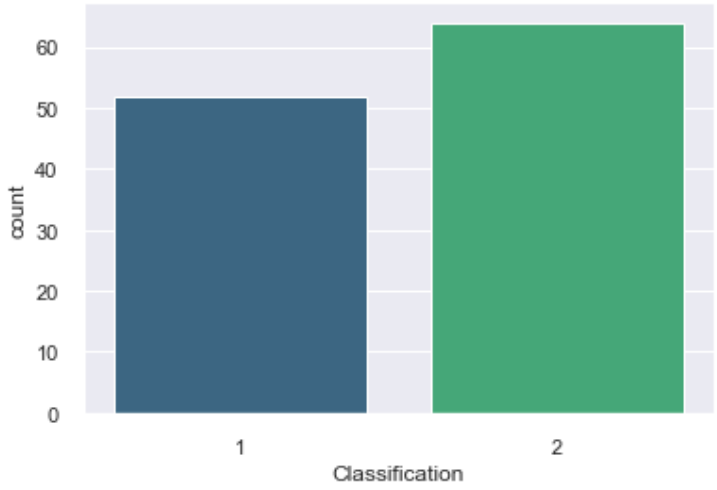

- And a lot of more ...

# Modeling

In this section used 5 Classification algorithms

- Logistic Regression
- K-Nearest Neighbor(KNN)
- Support Vector Machine(with Linear kernel)
- Support Vector Machine(with RBF kernel)
- Decision Tree

Dataset's label is balanced but is not very big




number of class 1 vs class 2?
Class 1 = 52
Class 2 = 64
In the next step I splitted the data-set into Training and Test Set (80% Training set and 20% Testing set)
and in the next step, Feature Scaling!
Feature scaling is the method to limit the range of variables so that they can be compared on common grounds. (most of the Machine Learning models are based on Euclidean Distance. so it's important)

# Next Step

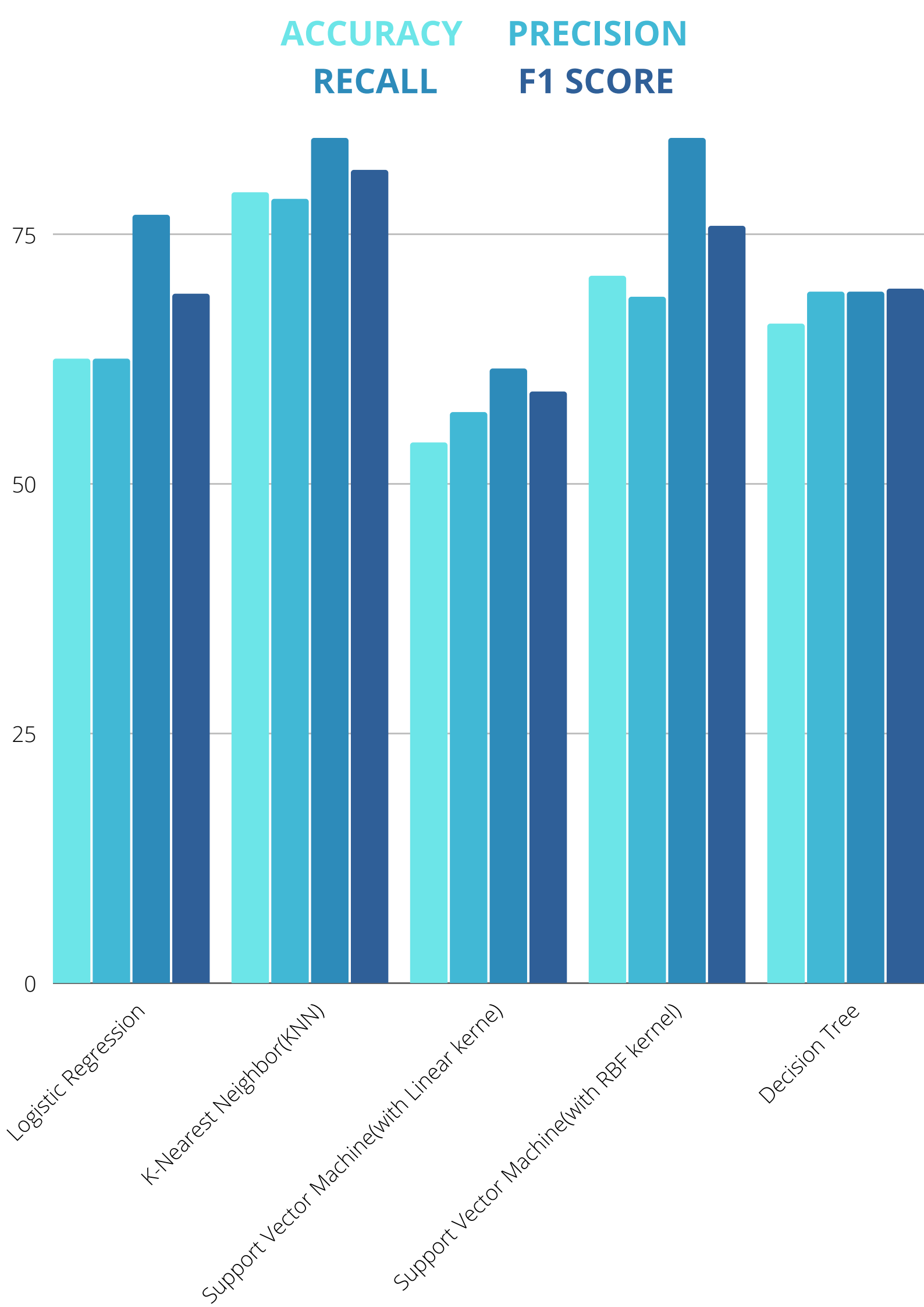
So far I have  built  Some models and trained it on some data...
So in the next step I'm going to evaluating my models
I' used common metrics to evaluate my models.

Accuracy is defined as the percentage of correct predictions for the test data. It can be calculated easily by dividing the number of correct predictions by the number of total predictions.

Precision is defined as the fraction of relevant examples (true positives) among all of the examples which were predicted to belong in a certain class.

Recall is defined as the fraction of examples which were predicted to belong to a class with respect to all of the examples that truly belong in the class.

F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution. Accuracy works best if false positives and false negatives have similar cost. If the cost of false positives and false negatives are very different, it's better to look at both Precision and Recall.


ACCURACY
PRECISION
RECALL
F1 SCORE
75
50
25
0
Logistic Regression
K-Nearest Neighbor(KNN)
Support Vector Machine(with Linear kerne)
Support Vector Machine(with RBF kernel)
Decision Tree

# Improve models

I look into the follow for improvements:
1. Check for outliers?

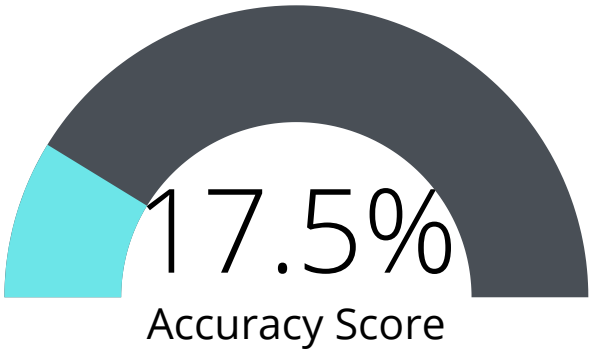
- manually
- Using Z-Score methods
- using IQR score methods

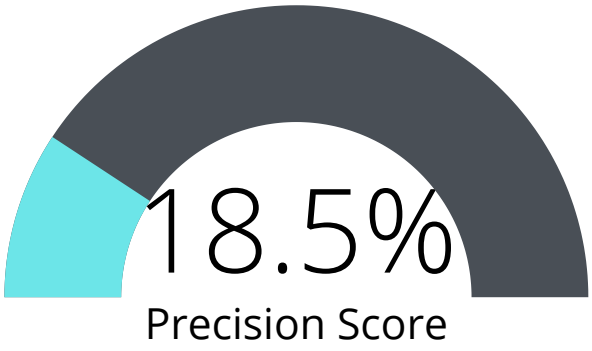
2. Drop unused column for modelling
calculate correlation between all columns and remove highly correlated one...
so I removed HOMA, Leptin, Adiponectin. Insulin and Split the dataset into 80% Training set and 20% Testing set again, Scaling the data and built the models again

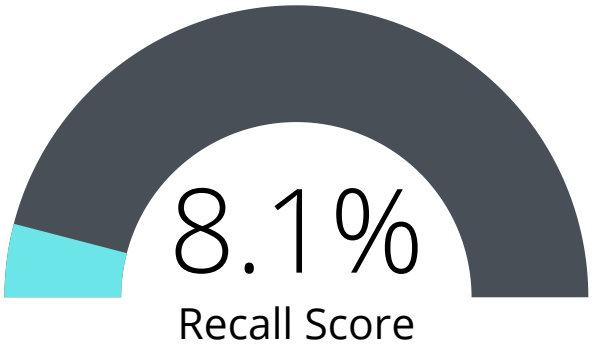
# Percentage Change Increase or Decrease?

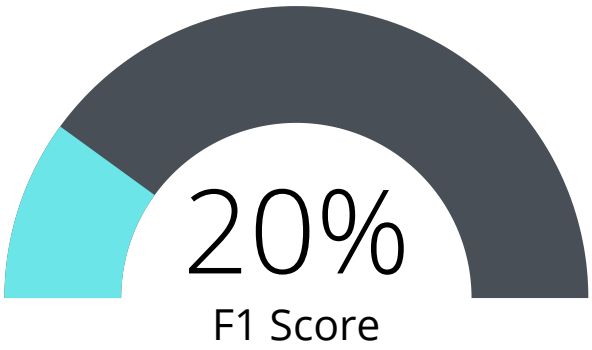
So far I have built  Some models and trained them on some data...
So in the next step, I'm going to evaluating my models
I've used common metrics to evaluate my models.

# Logistic Regression



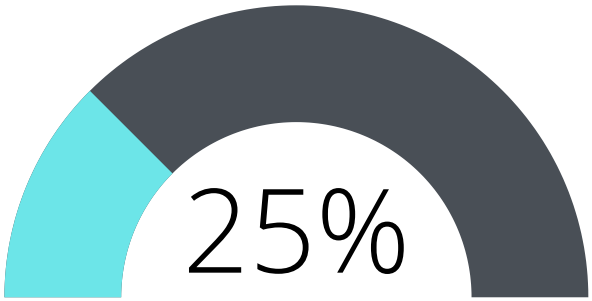



# K-Nearest Neighbor

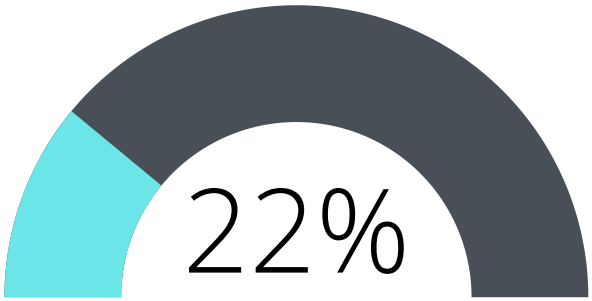
Precision Score

**+ 3%**

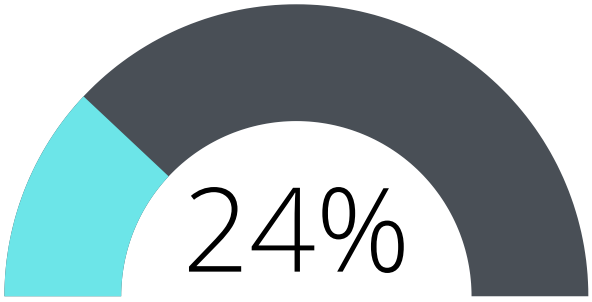
# SVM(Linear)



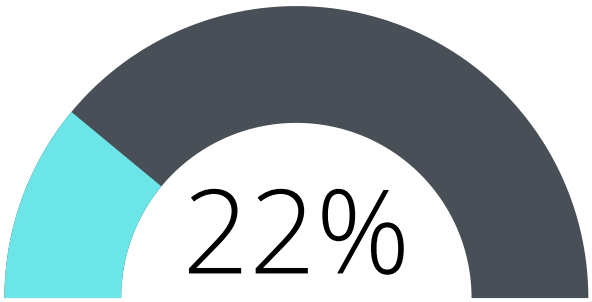

Accuracy Score




Precision Score

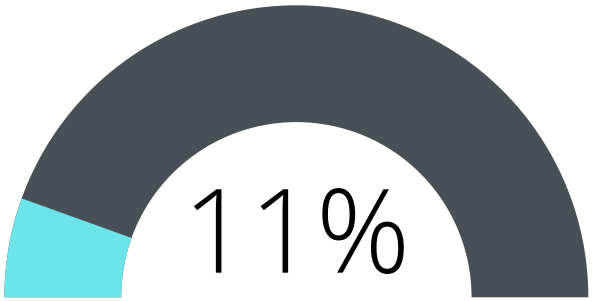



Recall Score

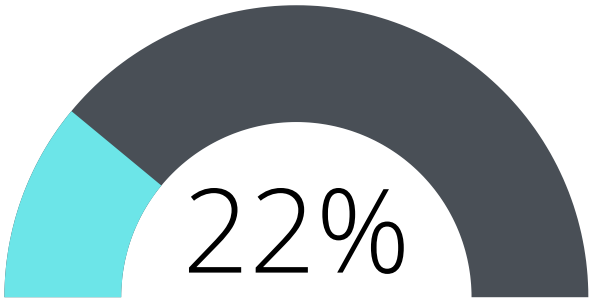



F1 Score

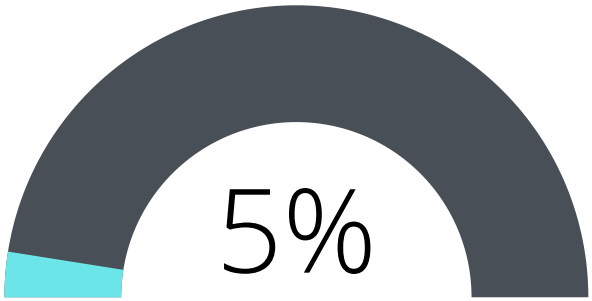
# SVM(RBF)



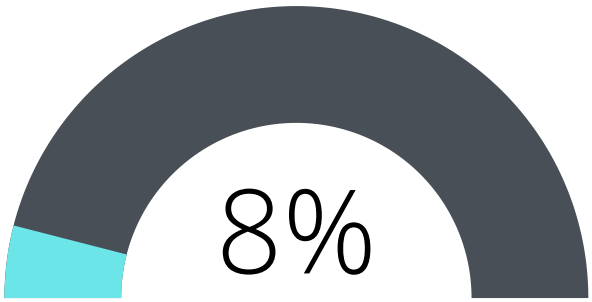

Accuracy Score




Precision Score

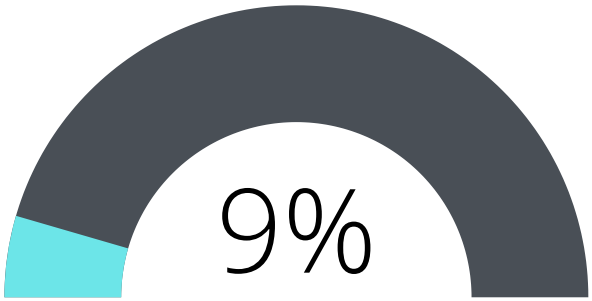



Recall Score

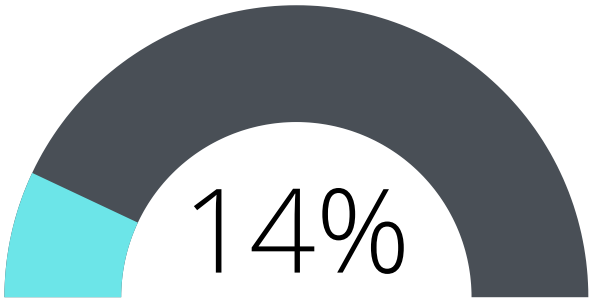



F1 Score

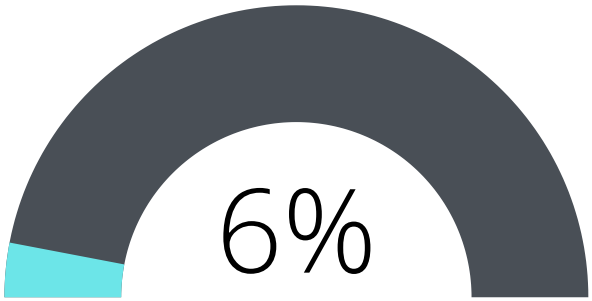
# Decision Tree



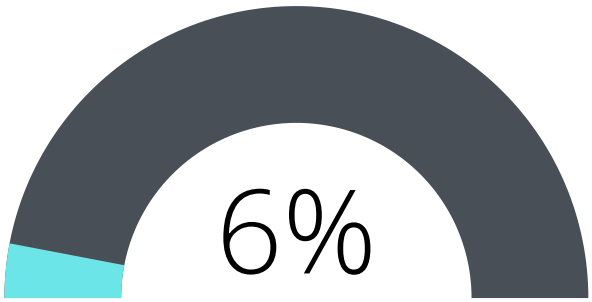

Accuracy Score




Precision Score




Recall Score




F1 Score

# Conclusion

**After removing outliers and drop unused column for modelling our models get better Score at the most criteria assessment**

**Accuracy Score**

| model 1 | model 2 | model 3 | model 4 | model 5 |
|---|---|---|---|---|
| 80 | 78 | 79 | 81 | 75 |

**Precision Score**

| model 1 | model 2 | model 3 | model 4 | model 5 |
|---|---|---|---|---|
| 81 | 82 | 79 | 79 | 83 |

**Recall Score**

| model 1 | model 2 | model 3 | model 4 | model 5 |
|---|---|---|---|---|
| 85 | 79 | 85 | 89 | 75 |

**F1 Score**

| model 1 | model 2 | model 3 | model 4 | model 5 |
|---|---|---|---|---|
| 82 | 79 | 81 | 83 | 75 |