# Machine learning project

# Overview

In this project I used the breast cancer dataset.

This dataset contains the characteristics of patients diagnosed with cancer. The dataset contains a unique ID for each patient, the type of cancer (diagnosis), the visual characteristics of the cancer and the average values of these characteristics.

# Domain knowledge

**Defining the Questions**:

The key question I am aiming to answer is to predict what the diagnosis of the tumor is: malignant or benign.

**Outcome Definition:**

The desired outcome will be 0 or 1 as 0 indicates benign and 1 indicates malignant.

**Identifying Influences:**

I hypothesize that features that describe the texture and size of the tumor will be beneficial for the prediction of the type of the cancer.

# Project design

**Optimal Design Selection:**To achieve this goal, I will build a classification model.

The model, based on the features in the dataset and potentially other calculated features will be able to classify between benign or malignant tumor.

**Data Handling Strategy:**In order to make sure that the data I am using is optimal I will utilize data cleaning techniques such as IQR outliers analysis and K-nearest neighbors to impute missing data.

**Algorithm Validation:**First, I will validate after each step that the new data frame was created as anticipated by exploring the csv files I save during the data preparation.

Second, To validate the algorithm I will split the data to train, validation and test, to make sure that the algorithm preformed accurately (and not over or under fitting the data).

# Data preparation

**Data Extraction:**

The data is taken from a csv file downloaded from Kaggle. The csv is called: Cancer\_data.csv.

**File Preparation:**  
I changed all the columns that their type was not a number or a string.

In this case, it was only the target variable: ‘diagnosis’ that his type was object.

I changed in to int 64 as M (malignant) is 1 and B (benign) is 0.

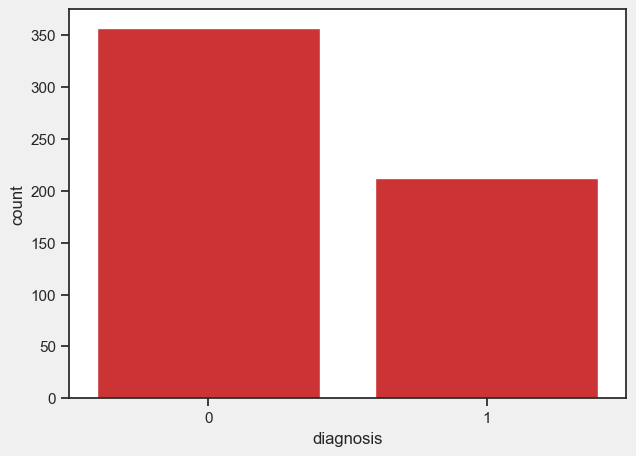
Additionally, I found that I had an extra column without any values so I deleted it.

**Exploratory Data Analysis (EDA):**On this stage I preformed the EDA, I extracted the data statistics and saved it in excel files.

The table below shows the statistics:

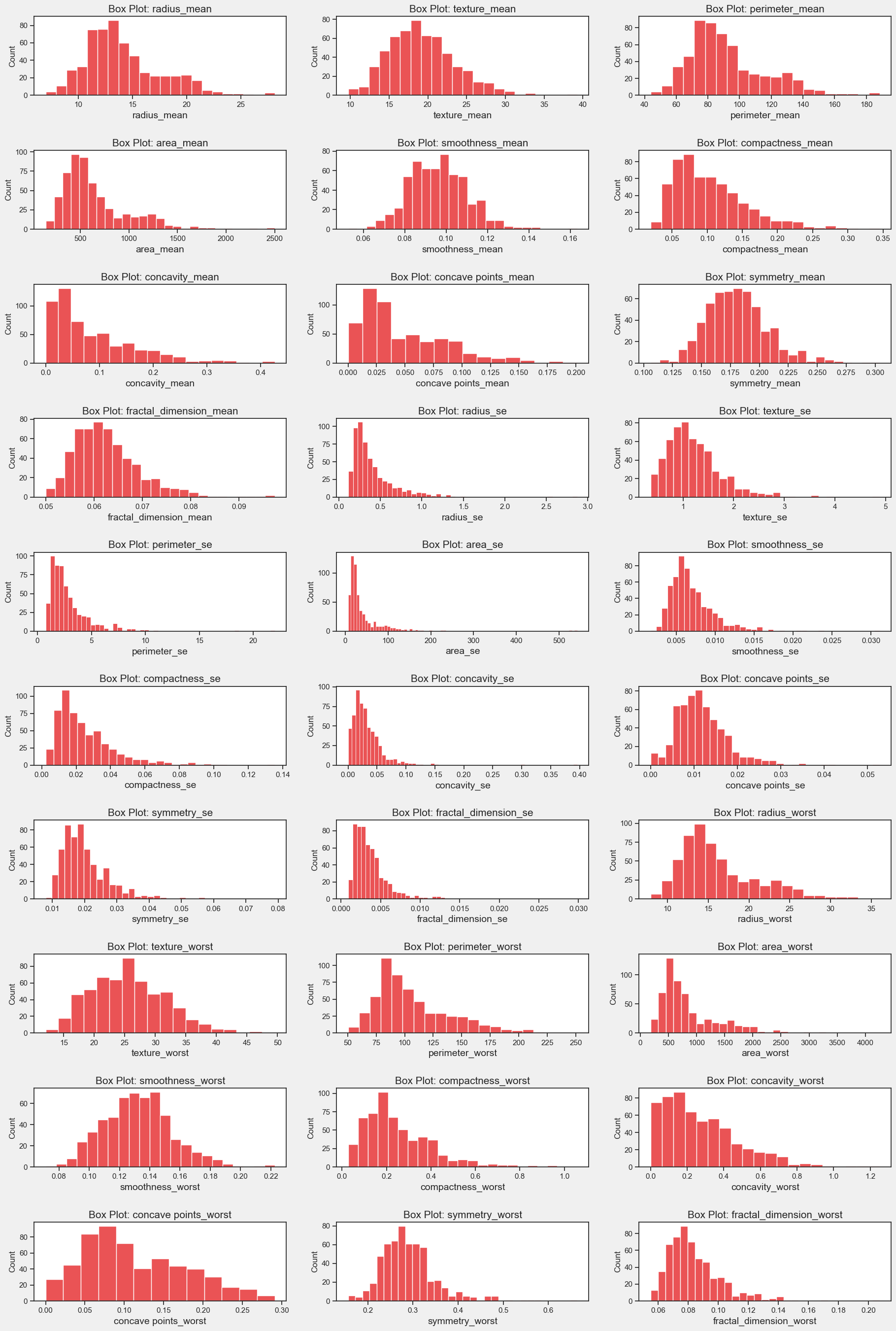
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Feature | count | mean | std | min | 25% | 50% | 75% | max |
| diagnosis | 569 | 0.37 | 0.48 | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 |
| radius\_mean | 569 | 14.13 | 3.52 | 6.98 | 11.70 | 13.37 | 15.78 | 28.11 |
| texture\_mean | 569 | 19.29 | 4.30 | 9.71 | 16.17 | 18.84 | 21.80 | 39.28 |
| perimeter\_mean | 569 | 91.97 | 24.30 | 43.79 | 75.17 | 86.24 | 104.10 | 188.50 |
| area\_mean | 569 | 654.89 | 351.91 | 143.50 | 420.30 | 551.10 | 782.70 | 2501.00 |
| smoothness\_mean | 569 | 0.10 | 0.01 | 0.05 | 0.09 | 0.10 | 0.11 | 0.16 |
| compactness\_mean | 569 | 0.10 | 0.05 | 0.02 | 0.06 | 0.09 | 0.13 | 0.35 |
| concavity\_mean | 569 | 0.09 | 0.08 | 0.00 | 0.03 | 0.06 | 0.13 | 0.43 |
| concave points\_mean | 569 | 0.05 | 0.04 | 0.00 | 0.02 | 0.03 | 0.07 | 0.20 |
| symmetry\_mean | 569 | 0.18 | 0.03 | 0.11 | 0.16 | 0.18 | 0.20 | 0.30 |
| fractal\_dimension\_mean | 569 | 0.06 | 0.01 | 0.05 | 0.06 | 0.06 | 0.07 | 0.10 |
| radius\_se | 569 | 0.41 | 0.28 | 0.11 | 0.23 | 0.32 | 0.48 | 2.87 |
| texture\_se | 569 | 1.22 | 0.55 | 0.36 | 0.83 | 1.11 | 1.47 | 4.89 |
| perimeter\_se | 569 | 2.87 | 2.02 | 0.76 | 1.61 | 2.29 | 3.36 | 21.98 |
| area\_se | 569 | 40.34 | 45.49 | 6.80 | 17.85 | 24.53 | 45.19 | 542.20 |
| smoothness\_se | 569 | 0.01 | 0.00 | 0.00 | 0.01 | 0.01 | 0.01 | 0.03 |
| compactness\_se | 569 | 0.03 | 0.02 | 0.00 | 0.01 | 0.02 | 0.03 | 0.14 |
| concavity\_se | 569 | 0.03 | 0.03 | 0.00 | 0.02 | 0.03 | 0.04 | 0.40 |
| concave points\_se | 569 | 0.01 | 0.01 | 0.00 | 0.01 | 0.01 | 0.01 | 0.05 |
| symmetry\_se | 569 | 0.02 | 0.01 | 0.01 | 0.02 | 0.02 | 0.02 | 0.08 |
| fractal\_dimension\_se | 569 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.03 |
| radius\_worst | 569 | 16.27 | 4.83 | 7.93 | 13.01 | 14.97 | 18.79 | 36.04 |
| texture\_worst | 569 | 25.68 | 6.15 | 12.02 | 21.08 | 25.41 | 29.72 | 49.54 |
| perimeter\_worst | 569 | 107.26 | 33.60 | 50.41 | 84.11 | 97.66 | 125.40 | 251.20 |
| area\_worst | 569 | 880.58 | 569.36 | 185.20 | 515.30 | 686.50 | 1084.00 | 4254.00 |
| smoothness\_worst | 569 | 0.13 | 0.02 | 0.07 | 0.12 | 0.13 | 0.15 | 0.22 |
| compactness\_worst | 569 | 0.25 | 0.16 | 0.03 | 0.15 | 0.21 | 0.34 | 1.06 |
| concavity\_worst | 569 | 0.27 | 0.21 | 0.00 | 0.11 | 0.23 | 0.38 | 1.25 |
| concave points\_worst | 569 | 0.11 | 0.07 | 0.00 | 0.06 | 0.10 | 0.16 | 0.29 |
| symmetry\_worst | 569 | 0.29 | 0.06 | 0.16 | 0.25 | 0.28 | 0.32 | 0.66 |
| fractal\_dimension\_worst | 569 | 0.08 | 0.02 | 0.06 | 0.07 | 0.08 | 0.09 | 0.21 |

To check that the data is not imbalanced i counted the number of malignant tumors and benign tumors. The results are presented in the figure below:



There are 212 malignant tumors in the dataset and 357 benign tumors.

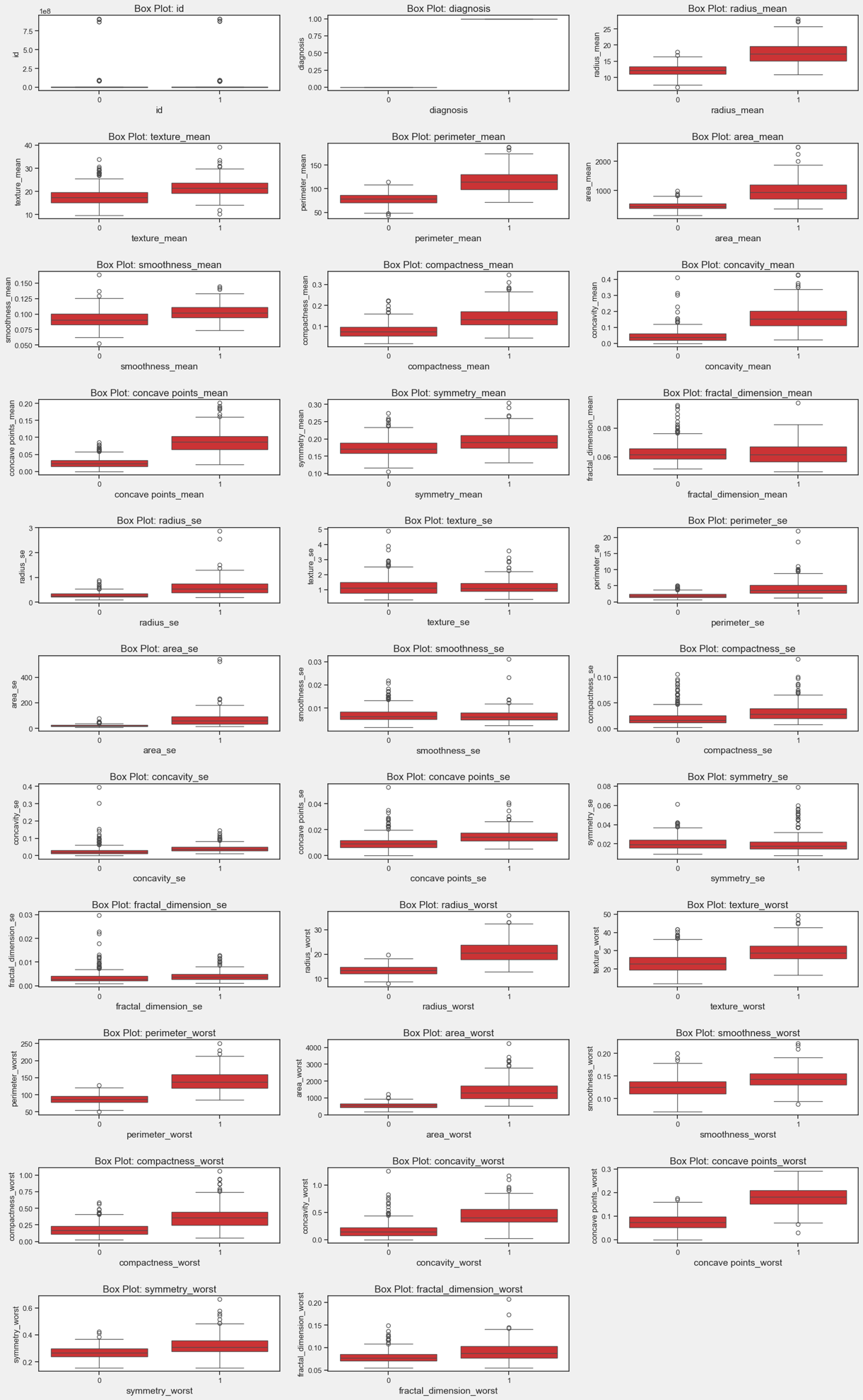
The figures below show the distribution of the features:



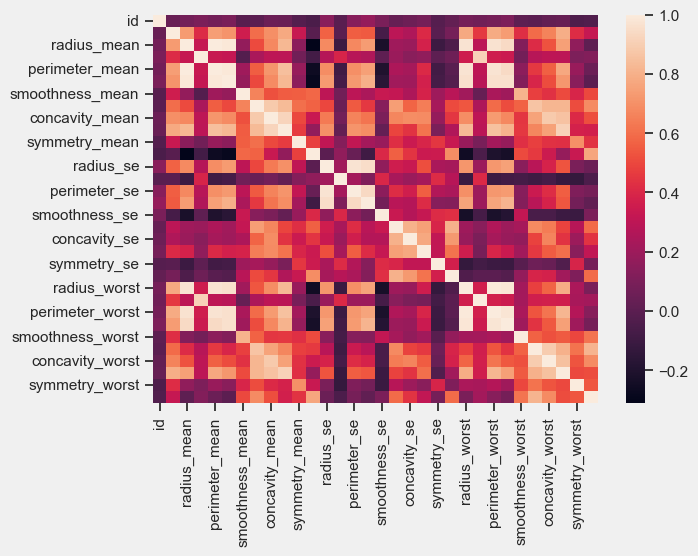
The table below presents the skewness of the dataset:

|  |  |
| --- | --- |
| Feature | skewness |
| area\_se | 5.45 |
| concavity\_se | 5.11 |
| fractal\_dimension\_se | 3.92 |
| perimeter\_se | 3.44 |
| radius\_se | 3.09 |
| smoothness\_se | 2.31 |
| symmetry\_se | 2.20 |
| compactness\_se | 1.90 |
| area\_worst | 1.86 |
| fractal\_dimension\_worst | 1.66 |
| texture\_se | 1.65 |
| area\_mean | 1.65 |
| compactness\_worst | 1.47 |
| concave points\_se | 1.44 |
| symmetry\_worst | 1.43 |
| concavity\_mean | 1.40 |
| fractal\_dimension\_mean | 1.30 |
| compactness\_mean | 1.19 |
| concave points\_mean | 1.17 |
| concavity\_worst | 1.15 |
| perimeter\_worst | 1.13 |
| radius\_worst | 1.10 |
| perimeter\_mean | 0.99 |
| radius\_mean | 0.94 |
| symmetry\_mean | 0.73 |
| texture\_mean | 0.65 |
| texture\_worst | 0.50 |
| concave points\_worst | 0.49 |
| smoothness\_mean | 0.46 |
| smoothness\_worst | 0.42 |

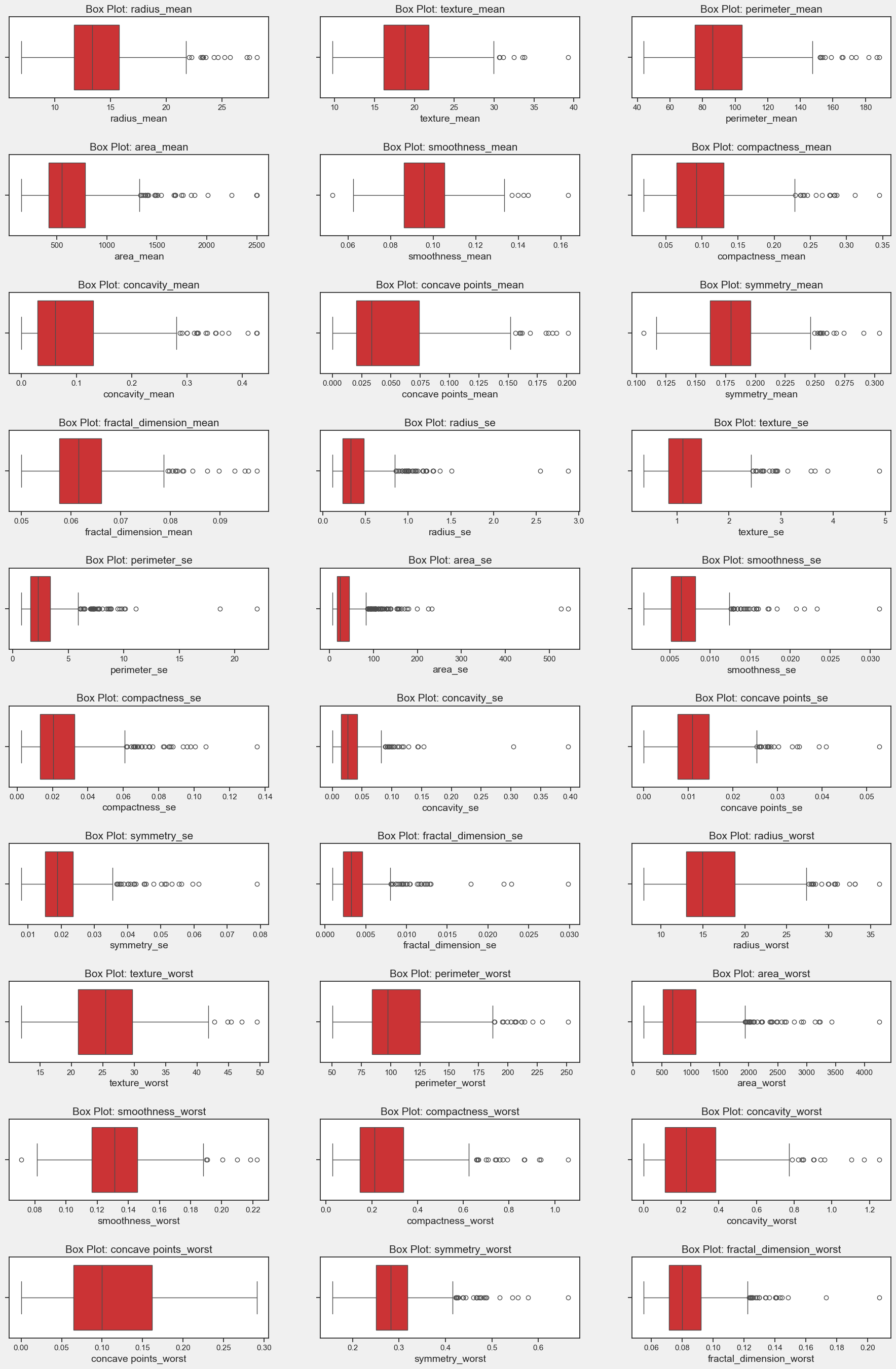
The figures below show the difference between malignant tumor and benign tumor in all features:



The figure below shows the correlations between all features:

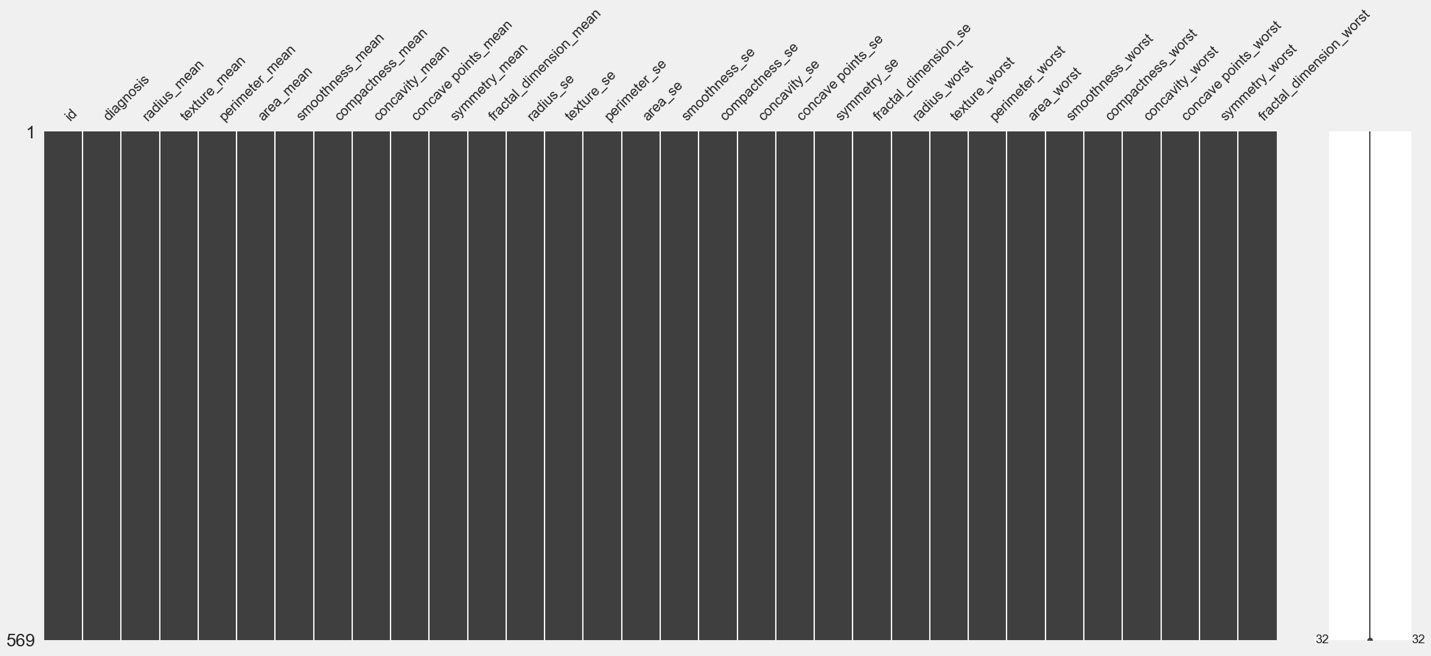


**Data Cleaning:**  
In order to clean the data I checked for outliers using IQR outliers analysis:



I found outliers in some of the features I found outliers but eventually I didn’t drop any of the values since none of them changes the distribution and did not change the correlation.

**Data Imputation:**  
I found no missing data:



**Data Transformation and Engineering:**

# Model selection

I chose to use supervised learning.

I split the dataset to train (70%) and test (30%).

I tested 7 classification models:

* Logistic Regression
* Decision Tree Classifier
* Random Forest
* Support Vector Classifier
* Gradient Boosting
* AdaBoost
* K-Neighbors Regressor

To evaluate their performance I used:

* Accuracy
* Precision
* Recall
* f1-score
* Log-loss
* AUC
* Confusion matrix

The AdaBoost model showed the best performance:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Accuracy | Precision | Recall | F1-score | Log-loss | AUC | Confusion matrix |
| 0.96 | 0.94 | 0.97 | 0.95 | 1.26 | 0.96 | 104,4  2 ,61 |

**Fine tuning:**

I used parameters grid for the fine tuning.

I found that the best parameters for AdaBoost are:

algorithm: 'SAMME'

learning\_rate: 1

n\_estimators: 100

With the final result:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Accuracy | Precision | Recall | F1-score | Log-loss | AUC | Confusion matrix |
| 0.97 | 0.95 | 0.97 | 0.96 | 1.05 | 0.97 | 104,4  2 ,61 |

