

**BIRZEIT UNIVERSITY**

**Department of Computer Science**

**Machine Learning (comp438)**

**Project#1:**

**Prepared by:**

Laith Ghnemat 1200610

**Instructor:**

Dr. Radi Jarrar

**Section**: 1

**Date**: 22/12/2023

I removed the smallest and largest data from the outliers, and the data with a value of zero, and this is a solution that does not affect the quality. Note the following image:

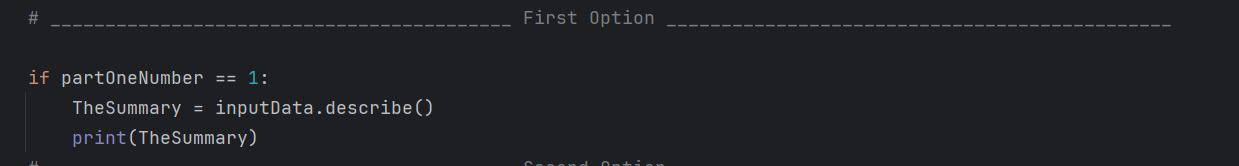


Part1: EDA

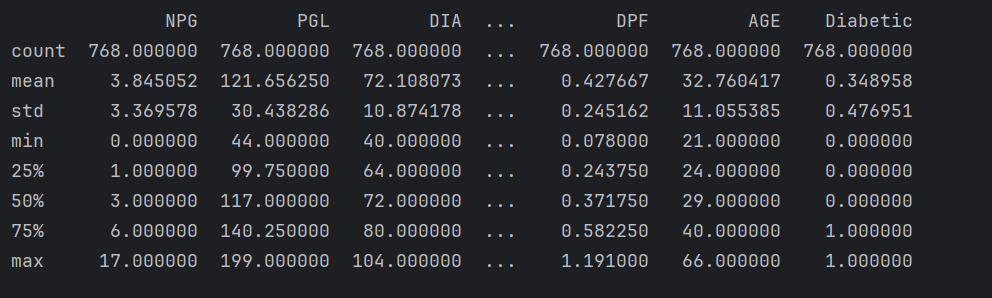
* 1.1

Following the steps in EDA, I printed the summary statistics of the attributes in the dataset.

I calculated the Max, Min, Mean (Average), Standard Deviation and the quarters of the distribution.

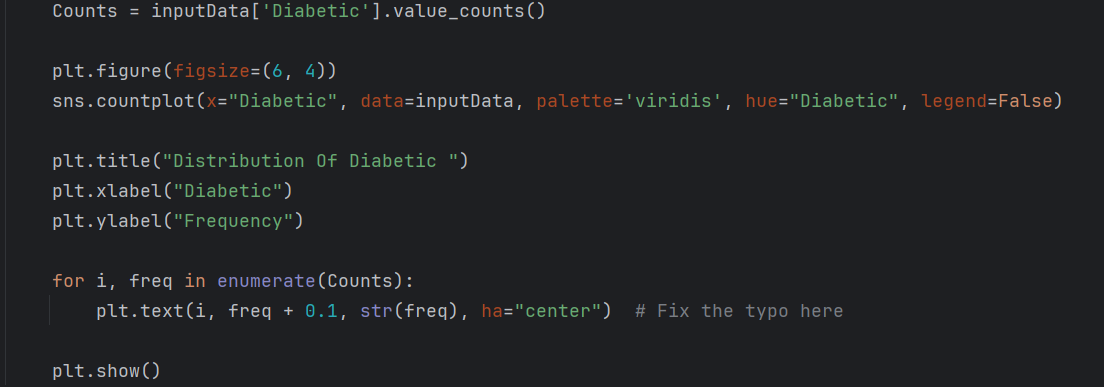


And the output was as the following picture:

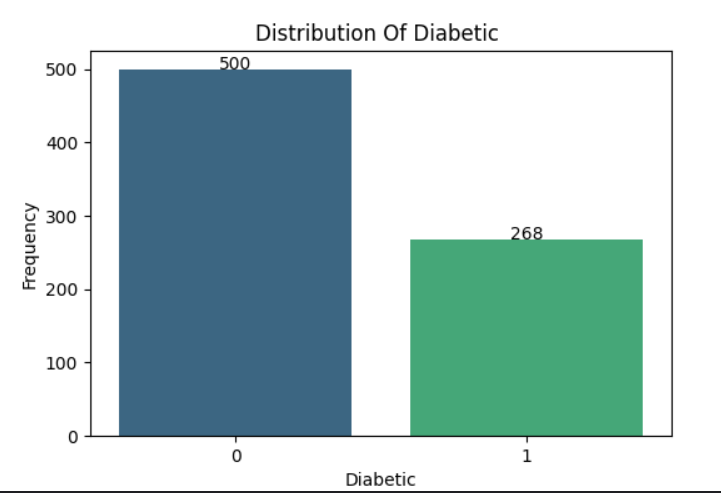


* 1.2

Here I calculated the number of Diabetes people (value of 1) and the number of Non - Diabetes people (value of 0). Then I plotted them on a curve, the x-axis represents the Diabetic status and the y-axis represents the frequency. Note this below code, this code is to show how many of each category:

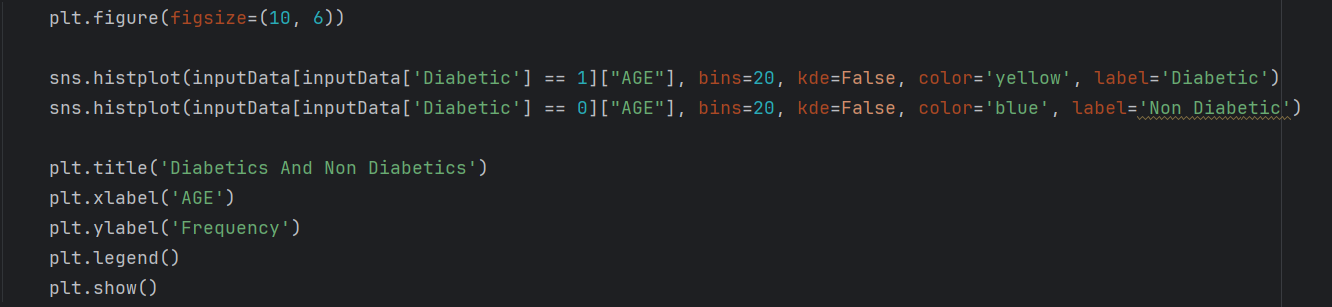


And the output was 500 for Diabetic and 268 for Non – Diabetic, note the following picture:

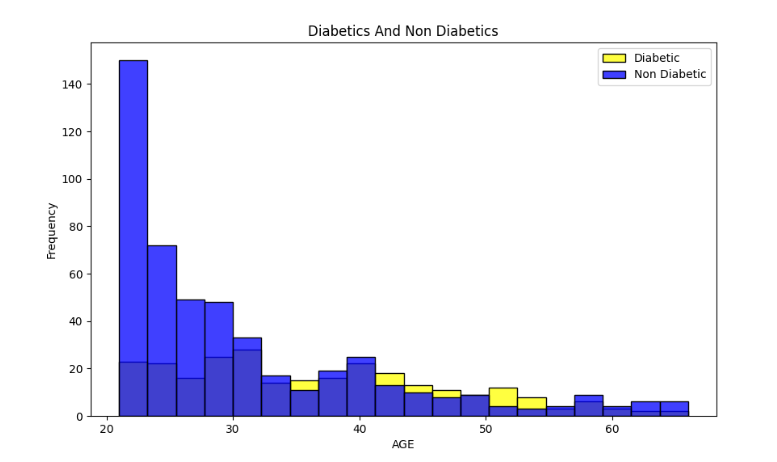


* 1.3

For each group age, I get the draw histogram with age on the x-axis and frequency on the y-axis. In each group we get the mount of diabetics, and the following was the code:

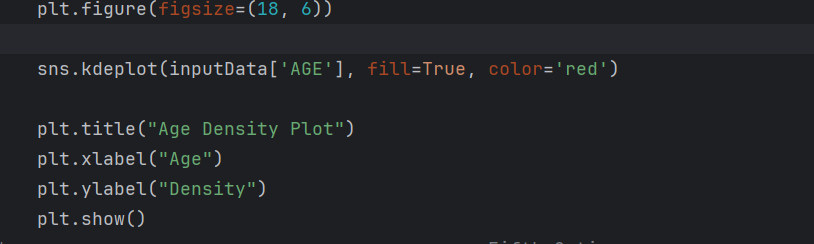


And the output was as following:

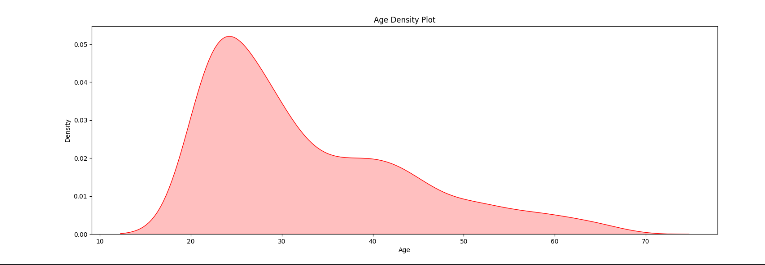


* 1.4

Here is the Density-Plot for the age, note the following code:

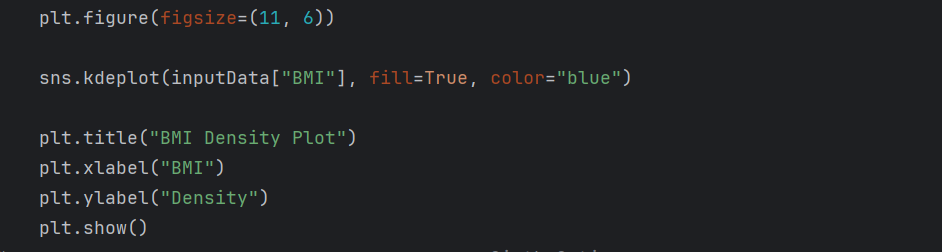


And the output was as following, the x-axis represents the age and the y-axis represents Density:

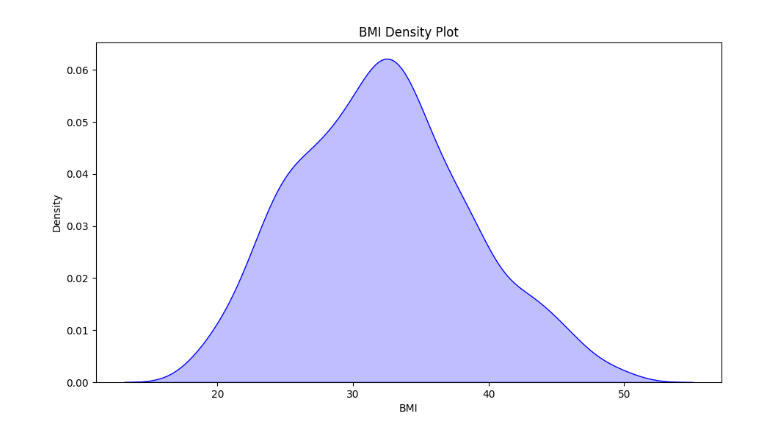


* 1.5

Here is the Density-Plot for the BMI, note the following code:



And the output was as following, the x-axis represents the BMI and the y-axis represents Density:

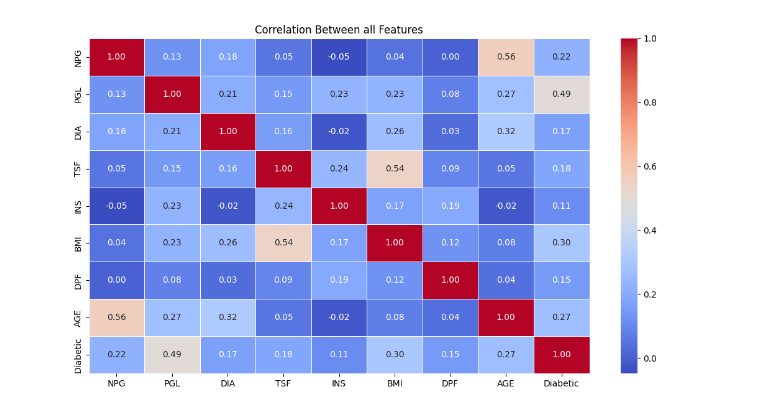


* 1.6

By the correlation, we get the relationships between all features with each other. Note the following code:



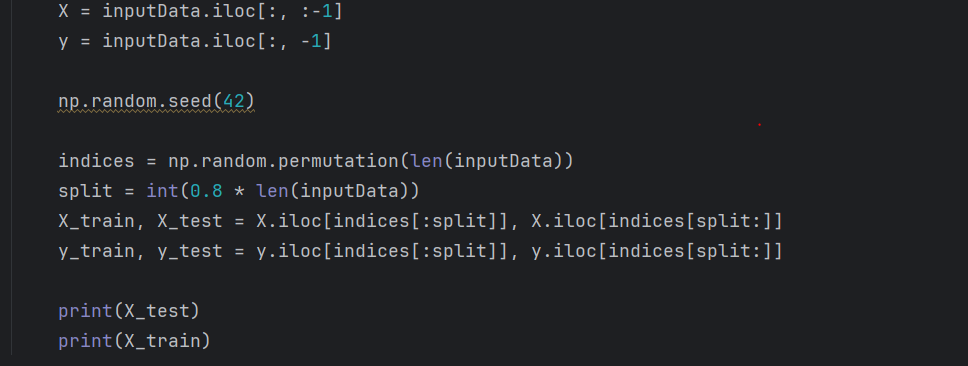
And the output was as the following:



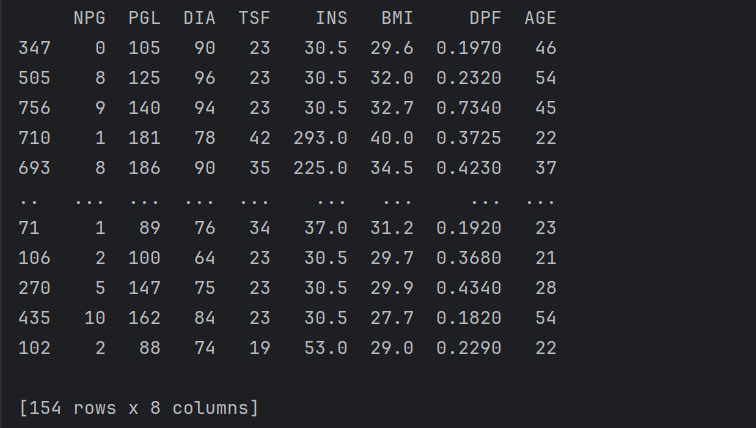
* 1.7

I split the data into training dataset (80% of the data), and testing dataset (20% of the data).

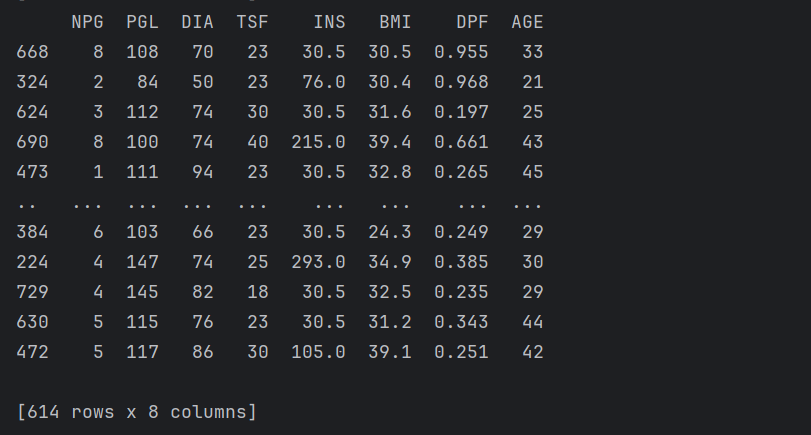
I removed the last column (which had been used for prediction) and from the Schedule bellow, note the following code:



Note the following table for the testing data which has 154 rows:



Also from the following table, we can see that the training data has 614 rows, which about four times of the testing dataset.



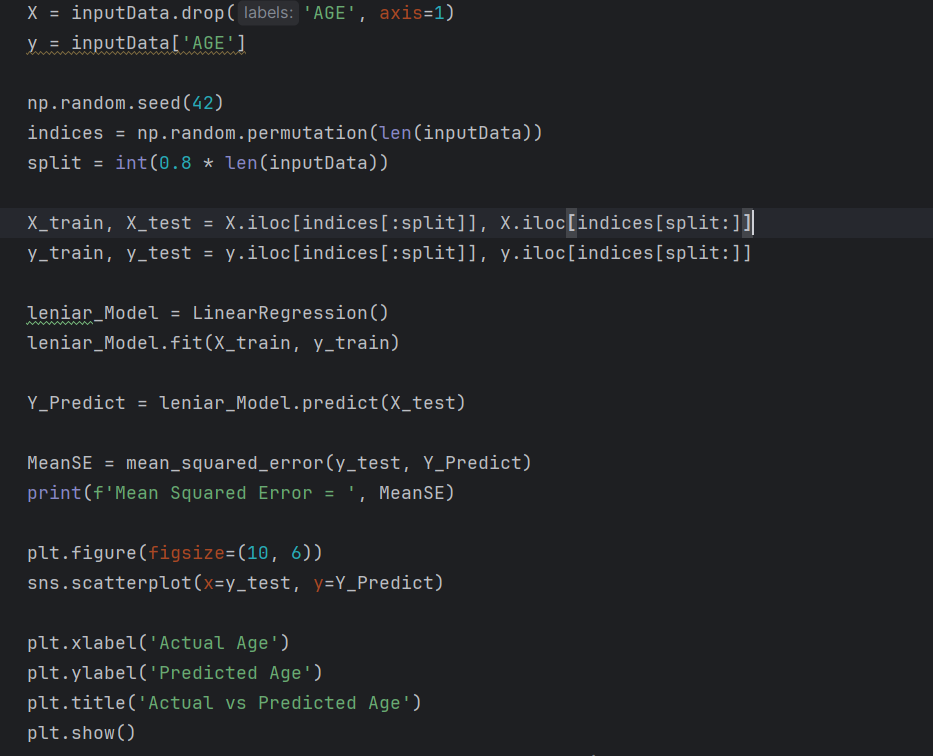
Part2: Linear Regression:

* 2.1

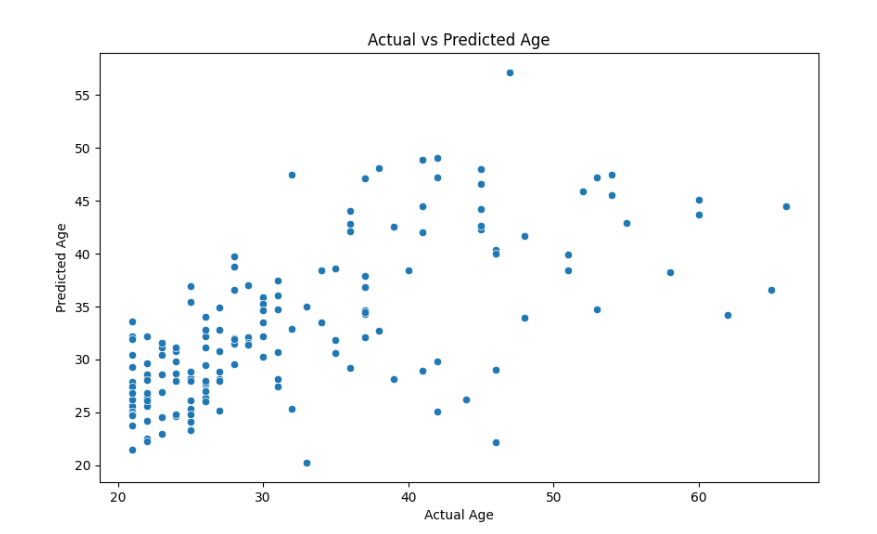
Here I used Linear Regression to learn about the age variable, by using the independent variables.

I used a model for Linear Regression to predict the age variable using other input data (features).

I used Mean-Squared-Error to measure the error, note the following code:

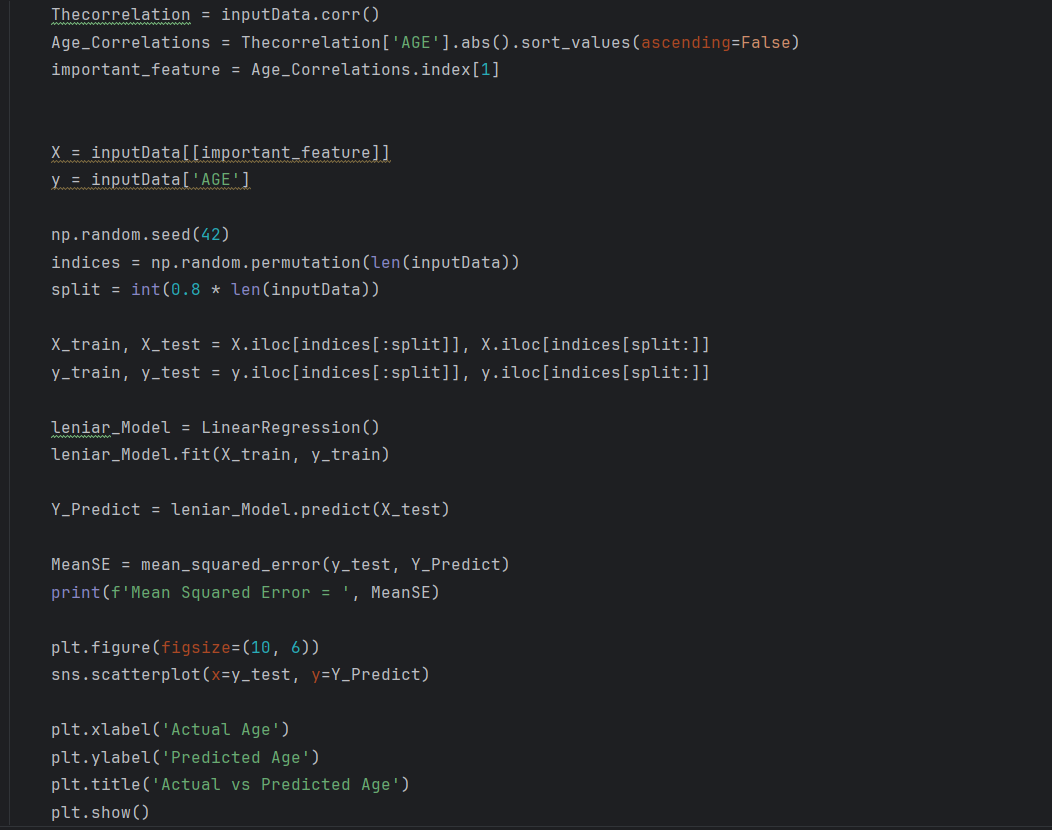


The output was as the following, since the x-axis is the actual age and y-axis is the predicted age, and we can see the Mean-Squared-Error equals almost 66.17:



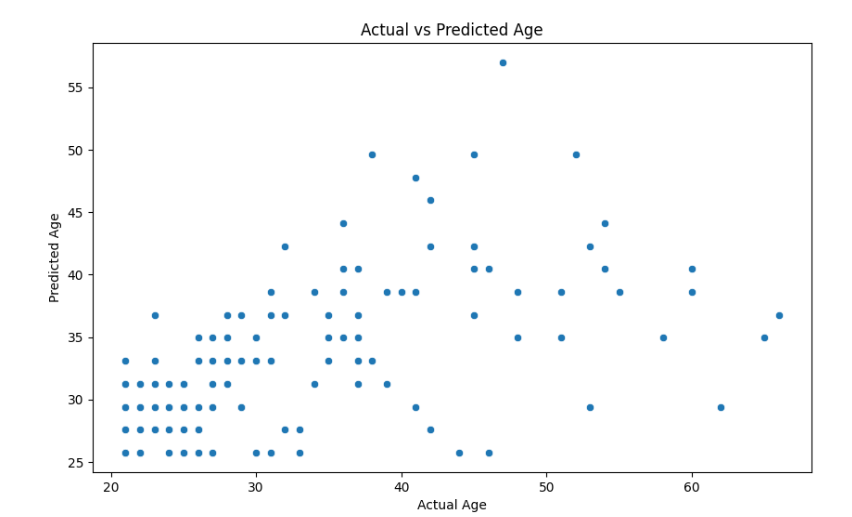
* 2.2

With Linear Regression and the most important feature (from Correlation Matrix). The best correlation has been used is the age. Note the following:



After predicting the age (Target), we predict the value of age.

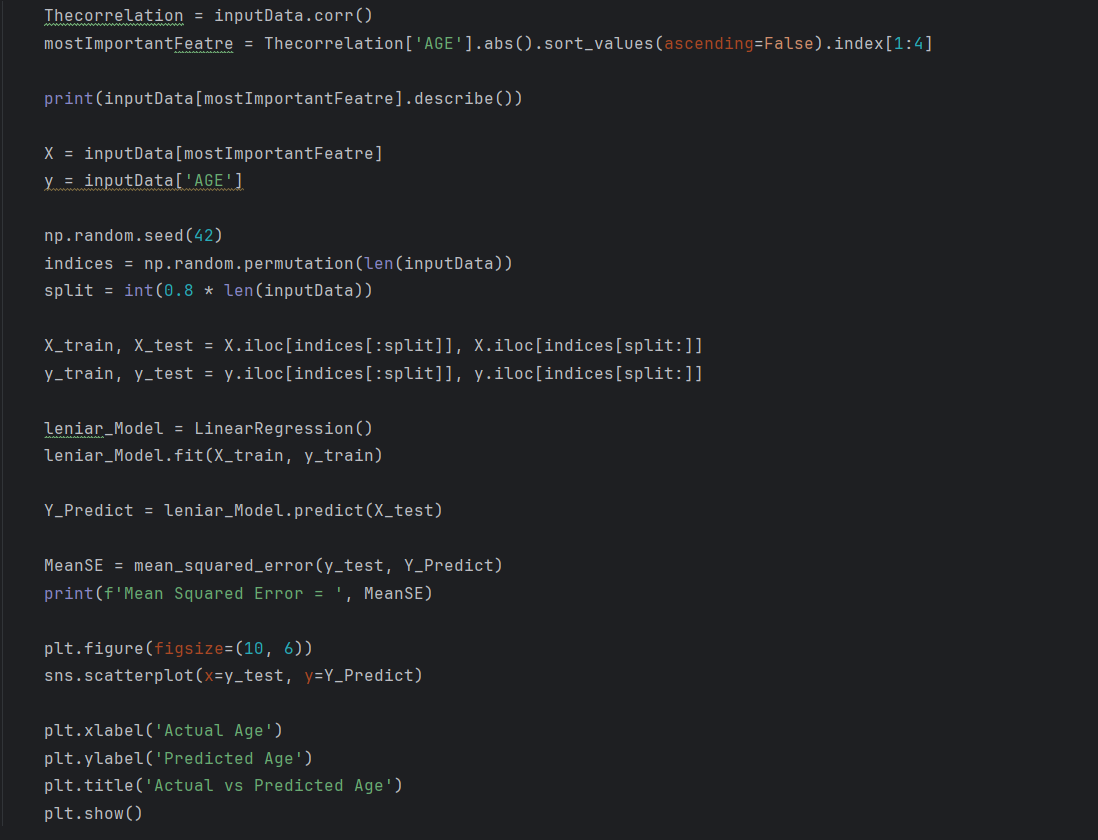
Note the following picture:



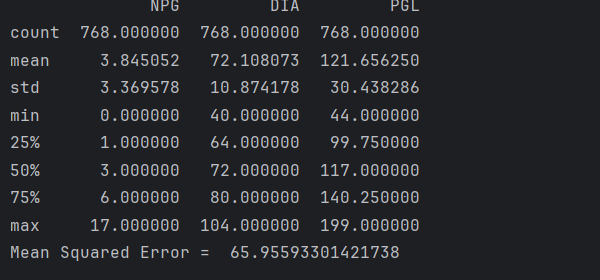
In the following picture I founded the Mean-Squared-Error equlas 77.6, since the x-axis is the actual age and the y-axis is the predicted age.

* 2.3

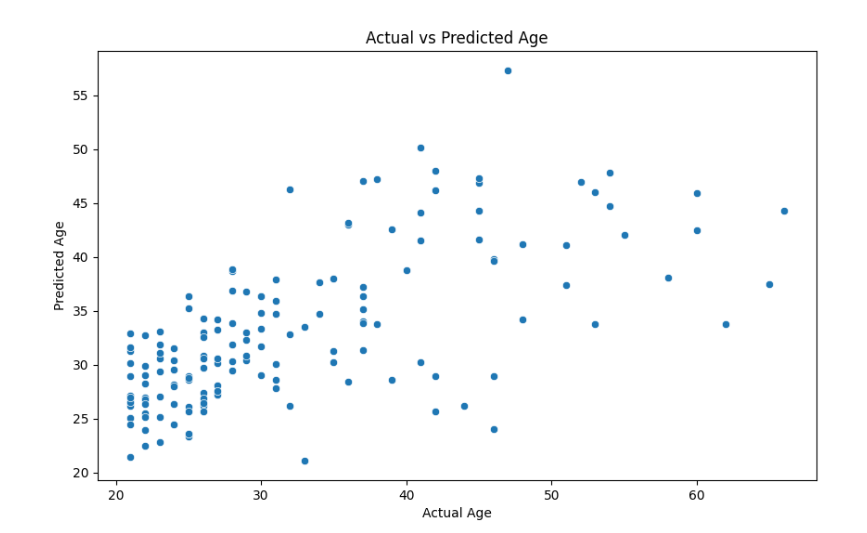
By The Linear Regression I found the three most important features using the Correlation Matrix with age (Target), Note the following code:



The output shows that the best correlations are (NBG, DIA and PGL).



Note the bellow figure which show that the Mean Squared Error equals 65.9, since the x-axis is the Actual age and the y-axis is the predicted age.



* 2.4

In this section I need to compare between the performance of the previous three models.

We know that the relationship between the Mean-Squared-Error and the performance of a model is positive. So the model with lowest Mean-Squared-Error is the best.

Here are the values of the Mean-Squared-Errors for each model:

Mean-Squared-Error for the first model equals 66.17

Mean-Squared-Error for the second model equals 77.6

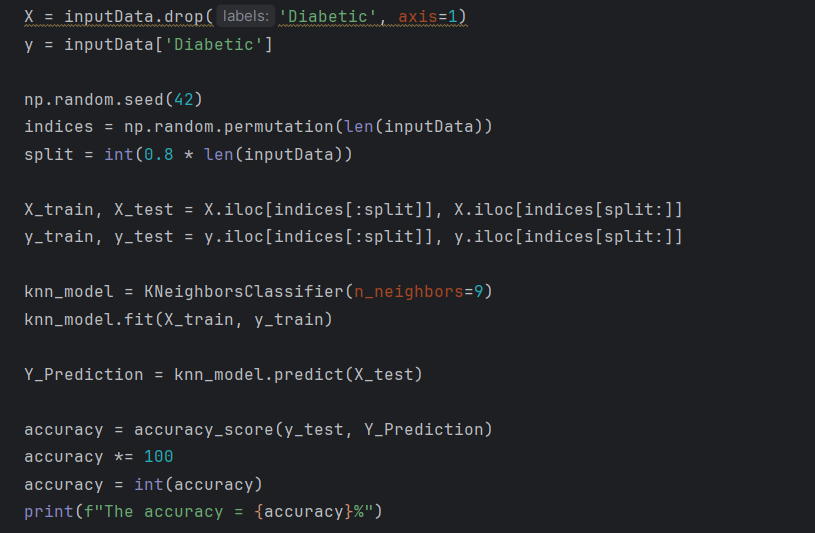
Mean-Squared-Error for the third model equals 65.9

And so that the third model is the best one, then the first and then the second.

Part3: Classification using KNN:

* 3.1

By using the testing set I can predict the classification. So that, through the KNN I can predict if the status is diabetes or not, note the following code:



The accuracy means the percentage of the correctly predicated data which equals 69%.



3.2 For different values of k.

Auc: is the ROC/AUC score, indicating the model's ability to discriminate between classes.

Confusoin Matrix: A table shows the frequencies of true positives (TP), false negatives (FN), ), false positives (FP), and true negatives (TN).

Note the following code:



And we can see the result by taking the values (1, 4, 5, 8) of K:

