Analyzing Heart Disease Dataset

with Machine Learning in Python

Burcu Acar

Bay Atlantic University

BGDA 555: Business Intelligence

[Micheline Al Harrack](https://learning.bau.edu/user/view.php?id=105&course=152)

April 28, 2024

ANALYZING HEART DISEASE DATASET WITH MACHINE LEARNING IN PYTHON

A dataset named 'Heart Disease Dataset' is downloaded from Kaggle and used to answer the research questions below.

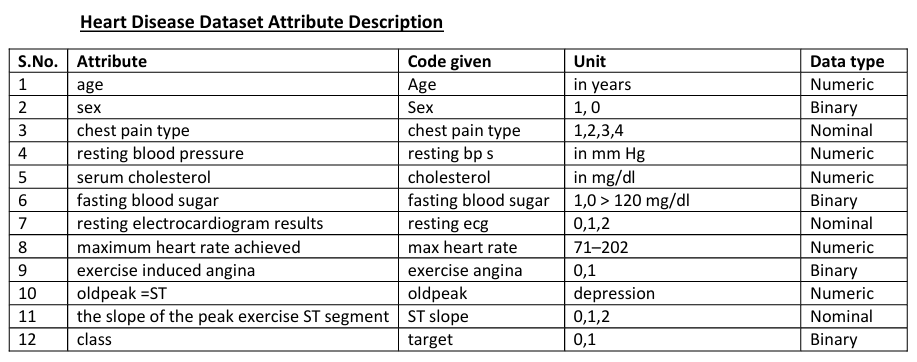
**Research Questions:**

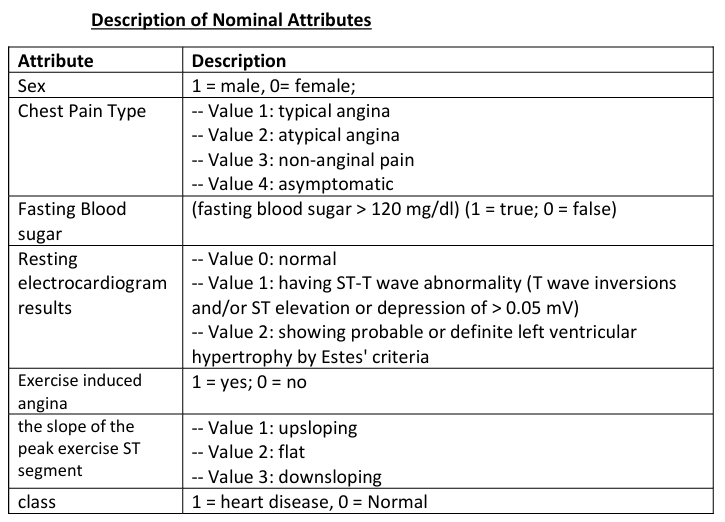
1. What is the accuracy of the model’s predictions?
2. Which risk factors contribute the most to heart disease?
3. Which age groups have high incidence of heart disease?
4. How should the model parameters be adjusted?
5. Examine all the features and find the factors that most influence each other, then comment on them.

Heart Disease Dataset is a comprehensive dataset combined from 5 popular heart disease datasets already available independently which are not combined before. This dataset is the largest heart disease dataset. This dataset created using by five datasets: Cleveland, Hungarian, Switzerland, Long Beach VA, Statlog (Heart) Data Set.

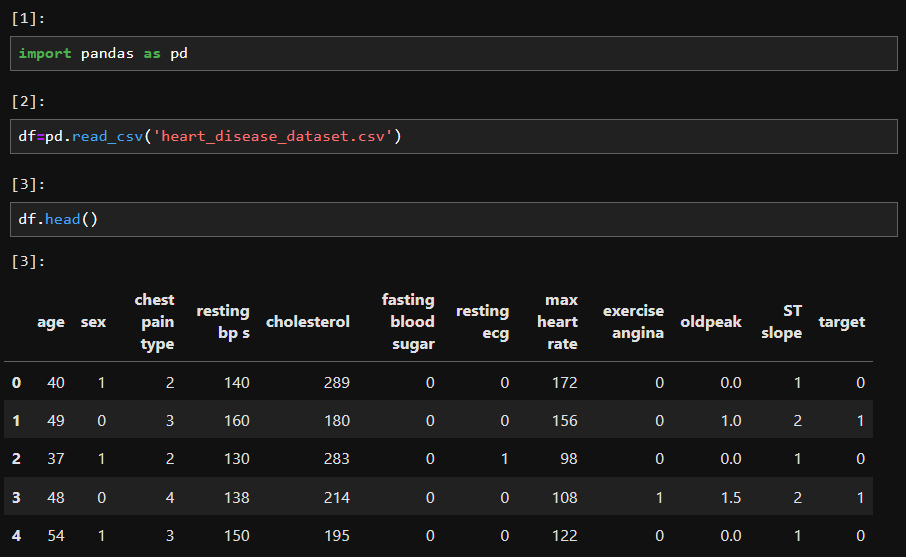
This dataset has 1190 instances with 12 features and combined to help research machine learning and data mining algorithms.

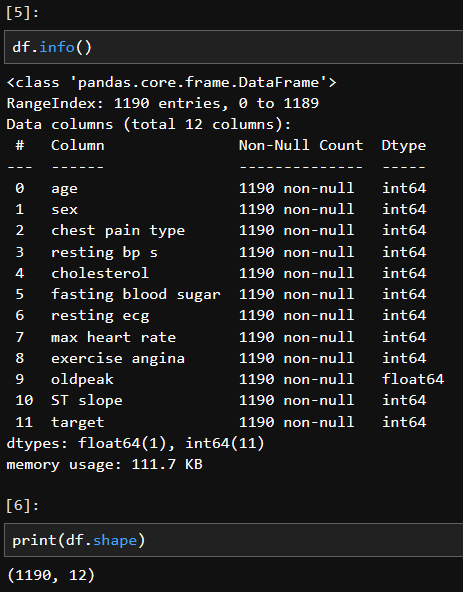
You can find “Heart disease “Dataset Attribute Description” and “Description of Nominal attributes” below.





After I discovered this dataset, I decided to use this dataset and search the most important factors causing heart disease and also wanted to discover the relationship between these features. I wanted to use the dataset for training and testing purpose to see the accuracy of the dataset’s accuracy.





First, I imported panda library and put my dataset to a data frame named as “df” by using read\_csv” function. Then, I wanted to preview the dataset and check columns by using “head” function.

Then, I wanted to be sure if I have null value and to learn what kind of data type it has by using “info” function. And I found out that I have 1190 rows and 12 columns with ”shape” function.

When I examined dataset, I discovered that some columns had Boolean results and some columns had numeric values, but these numeric values assigned to some contents. Therefore, I could not accept these numerical values as increasing or decreasing values. To make these columns more understandable and analyzable, I separated them and turned into different columns. I gave each column a value of 1 or 0, meaning “True” or “False.” For example, there are 4 types of chest pain and I separated and named them as chestpaintype\_1, chestpaintype\_2, chestpaintype\_3 and chestpaintype\_4. I thought that if a row has chestpaintype number 1, all other columns should have a value of 0 and only the chestpaintype column number 1 (chestpaintype\_1) should have a value of 1.

You can follow these values are below:

* Resting ecg (Resting electrocardiogram results):

-- Value 0: normal

-- Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)

-- Value 2: showing probable or definite left ventricular

hypertrophy by Estes' criteria.

* Chest Pain Type:

-- Value 1: typical angina

-- Value 2: atypical angina

-- Value 3: non-anginal pain

-- Value 4: asymptomat

* ST slope (slope of the peak exercise ST segment)

--Value 1: upsloping

-- Value 2: flat

-- Value 3: downsloping

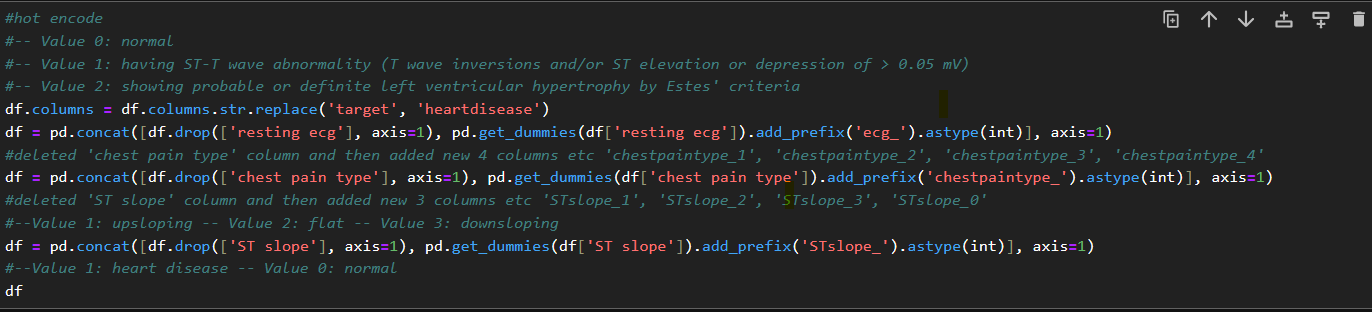
* target

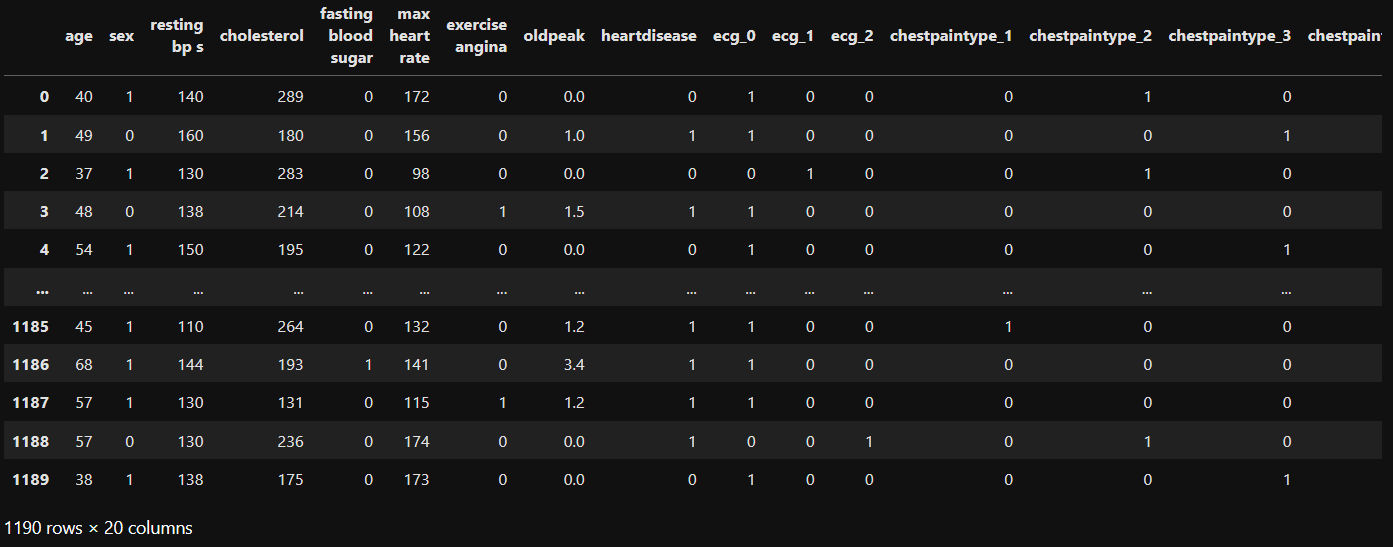
1 = heart disease, 0 = Normal

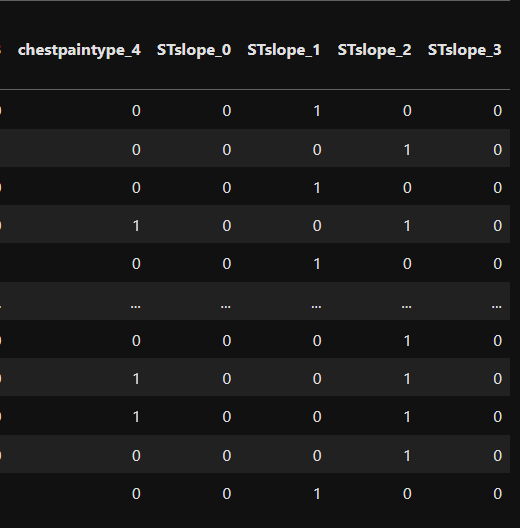
* Fasting Blood sugar

(fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)

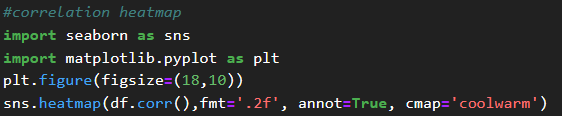
You can find below all this name changings and giving new values to these columns by using some functions.

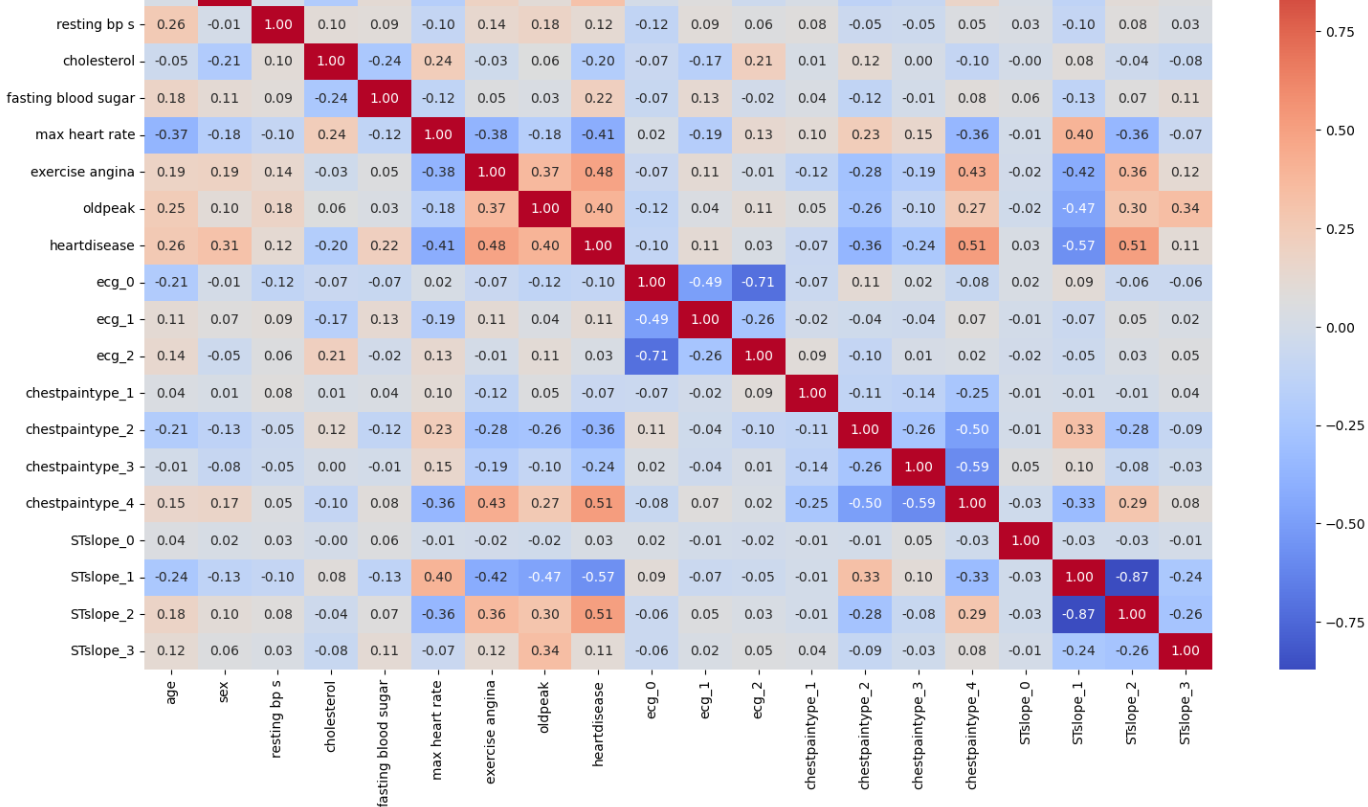


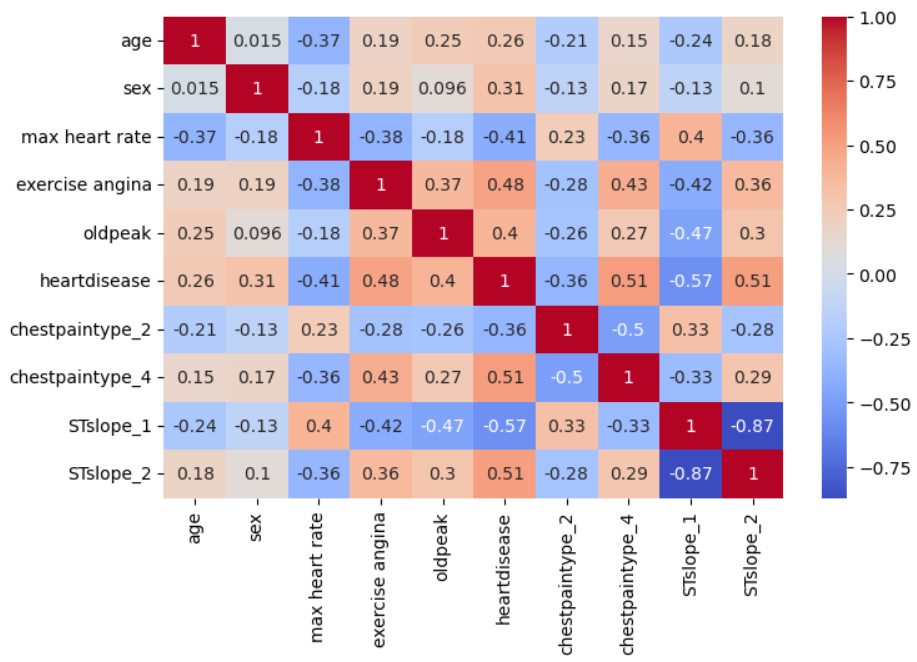
After all these data manipulations I had 20 columns as below. Additionally, I changed the name of the column named "target" to "heart disease".



Then , I created a heatmap by using seaborn library to find out correlation values of all 20 features.





After reviewing the correlation graph, I decided to make it more understandable and after sorting the correlation data from least to most, I removed unimportant values ​​for future use in my research questions.

After all these investigations and data manipulations, I used some special libraries to answer my research questions. To summarize, I started using RandomForestClassifier using the sklearn library to estimate the accuracy of the model I trained and tested.

Then, I found the importance of the features and listed them with the feature name. Then I used “hyperparameter tuning” with the GridSearchCV function and the sklearn library to make better predictions.

Finally, the relationship between age and heart disease was revealed using the #KDE (Kernel Density Estimation) graph and Logistic Regression Model.

You will find details of these techniques under the answers to the research questions.

**RESEARCH QUESTION ANSWERS**

**Question 1: What is the accuracy of the model’s predictions?**

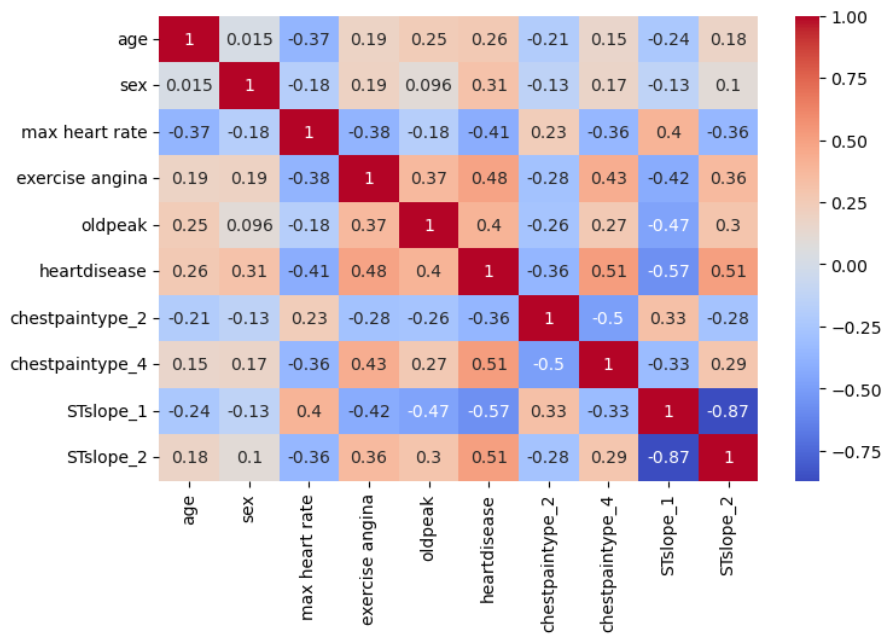
**Answer:** ‘0.9159’

****

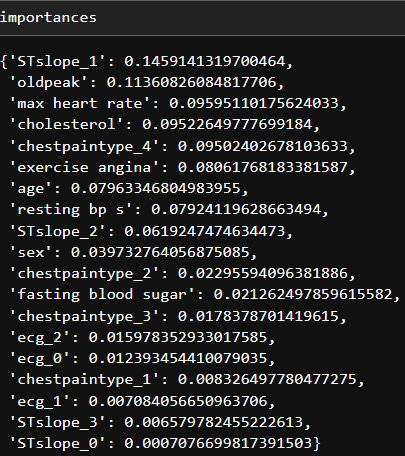
I used “sklearn” library to use “RandomForestClassifier” and “train\_test\_split” functions. After that I trained %70 and tested %30 of the dataset. Then I dropped, target column (changed its name as “heartdisease”) from train\_X and test\_X columns. Then I created an object named as ‘forest’ using ‘RandomForestClassifier’ function then put train\_X and train\_y objects by using ‘fit’ method in forest object. And lastly, I used ‘score’ method to find accuracy of the model’s prediction. The result was: ‘0.9159’

**Question 2 : Which risk factors contribute the most to heart disease?**

1. Most contributing risk factors are below in order of their **“correlation”** values:



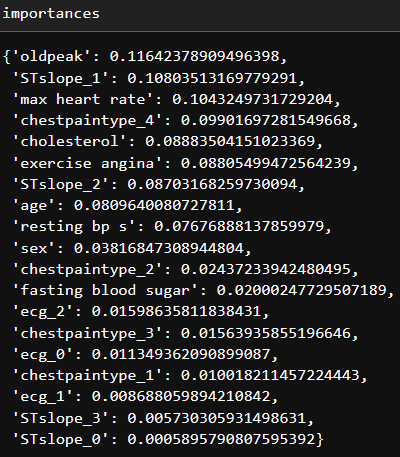
1. STslope\_1: -0.57 (the slope of the peak exercise ST segment- upsloping)
2. STslope\_2: 0.51 (the slope of the peak exercise ST segment- downsloping)
3. Chestpaintype\_4 (asymptomatic): 0.51
4. Exercise angina: 0.48
5. Max heart rate:-0.41
6. Oldpeak :0.4
7. Chestpaintype\_2 (atypical angina): -0.36
8. Most contributing risk factors are below in order of their **“importances”** values:



1. STslope\_1: 0.145
2. Oldpeak:0.113
3. Max heart rate: 0.095
4. Cholesterol: 0.095
5. Exercise angina: 0.081
6. Age: 0.079
7. Resting bp s: 0.079

**After hyperparameter tuning importances values are:**

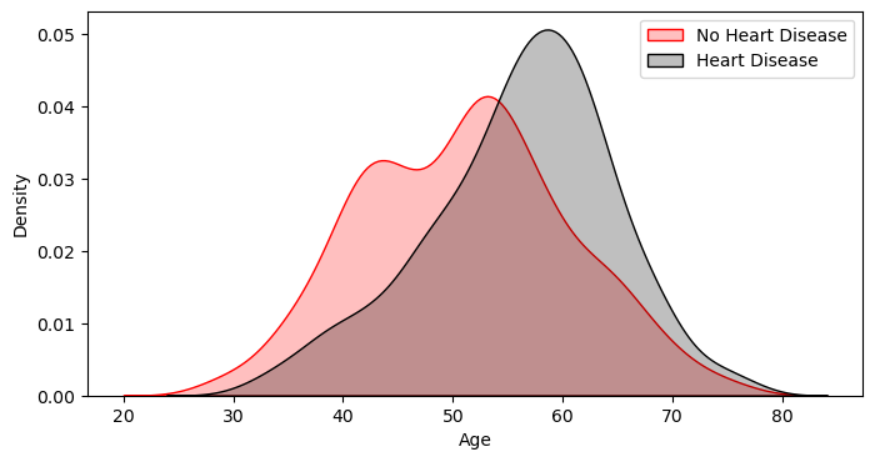
1. Oldpeak:0.116
2. STslope\_1: 0.108
3. Max heart rate: 0.104
4. Chestpaintype\_4:0.099
5. Cholesterol: 0.088
6. STslope\_2: 0.087
7. Age:0.081

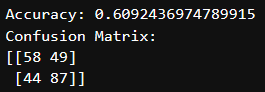


**Question 3 : Which age groups have high incidence of heart disease?**

**Method 1: Kernel Density Estimation Graphic**

I used Kernel Density Estimation Graphic to find age groups which have high incidence of heart disease. As you can see, “heart disease” peaks between ages 55 and 65, but also “no heart disease” peaks between ages 50 and 55 as well. So, we cannot say that “the rate of heart attacks increases in direct proportion to the increase in age.” Because after the age of 65 the “heart disease rate” decreases, and after the age of 50 the “no heart disease” rate also decreases.

**Method 2: Logistic Regression Model**



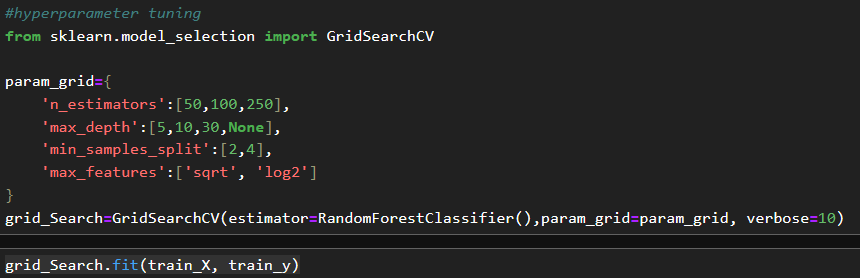
Logistic regression model is predicting whether a person has disease or not based on the “age”. In the confusion matrix 58 indicates the count of “no heart disease” values and the model correctly predicted as 0. 49 represents numbers which are predicted wrong.

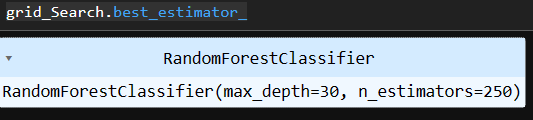
The number 87 represents the count of “heart disease” values which is predicted correctly. 87 represents numbers which are predicted wrong.

As you can see above, after training 80 percent and testing 20 percent of the dataset I had 0.61 accuracy which is quite low and confusion matrix is also not very reliable.

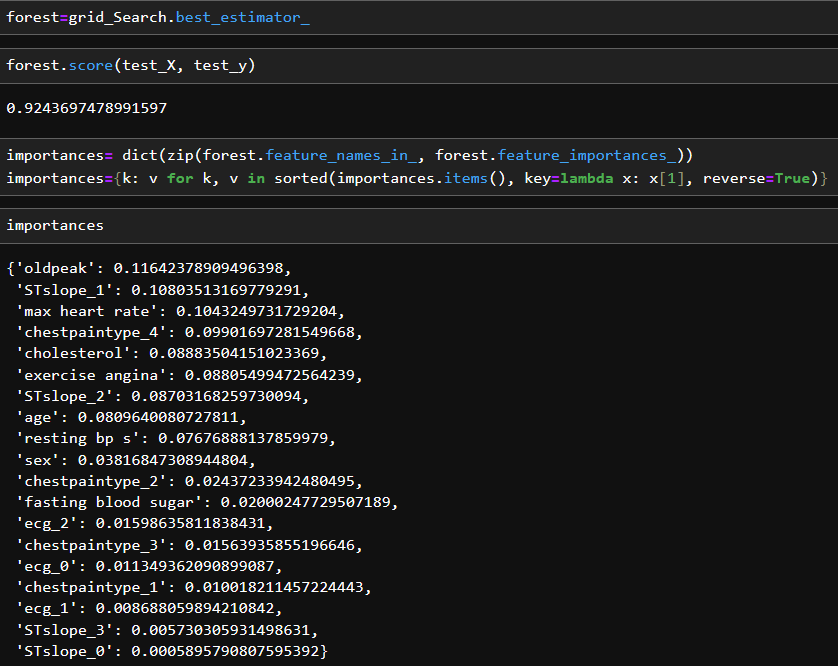
**Question 4: How should the model parameters be adjusted?**

I used hyperparameter optimization with the help of the “sklearn.model\_selection” library and the “GridSearchCV” function. Hyperparameter tuning is used to improve the performance of the machine learning model. These parameters are determined during training and affect performance. Hyperparameter tuning helps find parameters to get the best performance. It does this by trying different hyperparameter values. It helps modelling to achieve better performance.

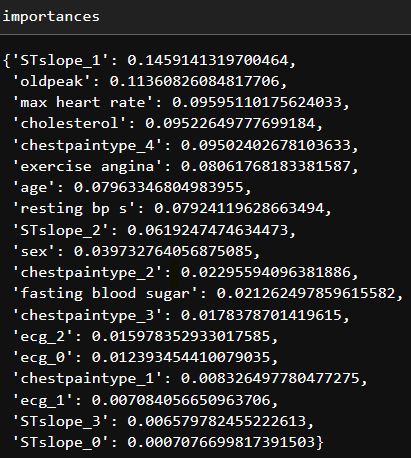




If we observe first importances values we can see differences between them. You can see the second optimized importances values below.



And you can see the non-optimized importances values below:



**Question 5 : Examine all the features and find the factors that most influence each other, then comment on them.**

As I mentioned before, I dropped the insignificant correlation values ​​and finally obtained these values ​​using the “df\_dropped.corr()” code. You can also see the same correlation values ​​in the graph. While correlation values ​​lose their importance as they approach 0, if they approach 1 or -1, it means that the features have a relationship or high correlation between them.

****

I created this image below by seaborn library, heatmap function. When I observed this image, I could say:

1. High negative correlation between STslope\_1 (upsloping) and STslope\_2 (flat) which is

-0.87.

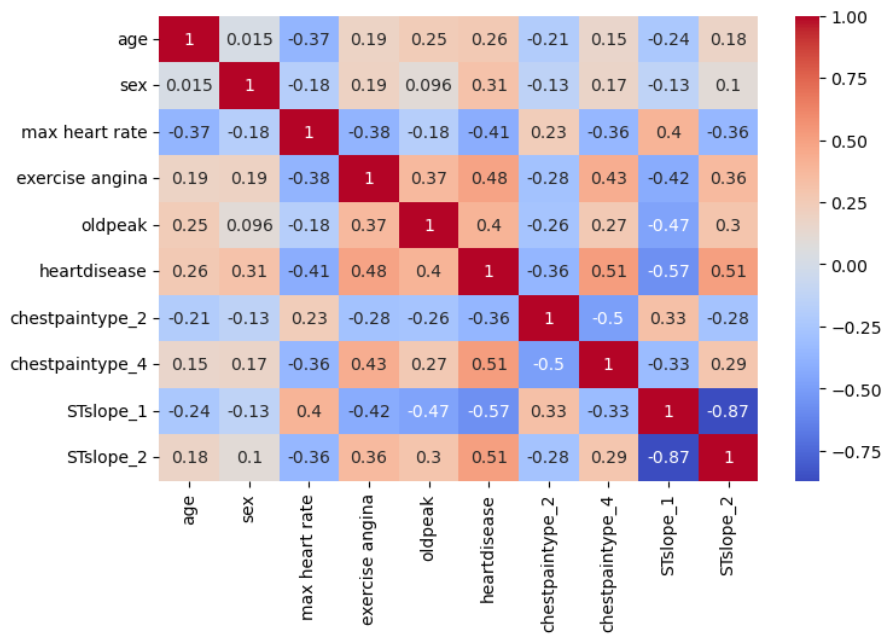
This means that if a person has upsloping ST segment, they probably don’t have flat ST segment. However, since the correlation value is not -1, there is still a possibility that the person may have two different type of ST segments at the same time.

1. High positive correlation between chestpaintype\_4 (asymptomatic) and heartdisease which is 0.51

I can say If someone has asymptomatic chest pain type , they probably have heart disease.

1. High negative correlation between heartdisease and STslope\_1(upsloping) which is -0.57

This means that there is a negative relationship between them. If someone has heart disease , they probably will not have upsloping ST segment.



1. High negative correlation between oldpeak and STslope\_1(upsloping) which is -0.47

There is also a negative relationship between these two features.

1. High positive correlation between exercise angina and heartdisease which is 0.48

This means that if someone has exercise angina, they probably have heart disease.

1. High negative correlation between max heart rate and heart disease which is -0.41

This means that as heart rate increases, heart diseases decrease.

1. Negative correlation between max heart rate and age, exercise angina chestpaintype\_4 , STslope\_2 which is -0.37, -0.38, -0.36, -0.36

This means that heart rate may decrease as age increase. Also, if people have exercise angina or chest pain type 4 or ST slope 2, their heart rate may be low.

# References

NeuralNine. (2023, August 06). *NeuralNine*. Retrieved from youtube: https://www.youtube.com/watch?v=dhoKFqhVJu0&t=847s

SIDDHARTHA, M. (2024, April 08). *Heart Disease Dataset*. Retrieved from kaggle: https://www.kaggle.com/datasets/mexwell/heart-disease-dataset