OPTIMIZATION FOR DATA SCIENCE

Project - Heart Diagnosis

Submitted by

Kouadio Yao innocent junior

Thanh Tung Trinh

Supervisor

Professor: Loualitene FATAH

EPITA : École Pour l'Informatique et les Techniques Avancées - 19/02/2020

Table of Contents

Introduction

Heart disease is one of the most dangerous human diseases of all times. With the evolution in technology and medicine, we now can have more chances to improve the human's life, especially doctors recently can have a lot of information about the patient from medical devices to support the diagnosis of health condition. Hence, the measurement data from medical devices are the real gold for doctors to have the right decision on each of every person. Since the importance of understanding the data in heart disease and interest in data analysis and prediction model. Our team would like to investigate heart disease through the project Heart Diagnosis with Data Optimization.

Dataset

Dataset One

Our team has chosen a Dataset of Heart Disease from the website Kaggle.com. This Dataset is based on the real case study from the Hungarian Institute of Cardiology, University of Hospital Zurich, University of Hospital Basel and V.A. Medical Center, Long Beach and Cleveland Clinic Foundation USA. The Dataset has 14 columns and 304 rows with 13 types of integer variables (including 4 binary variables) and 1 float variables. Each of the variables describes the factor (index) that is related to heart disease (target). We are going to explain further the variables in the next parts.

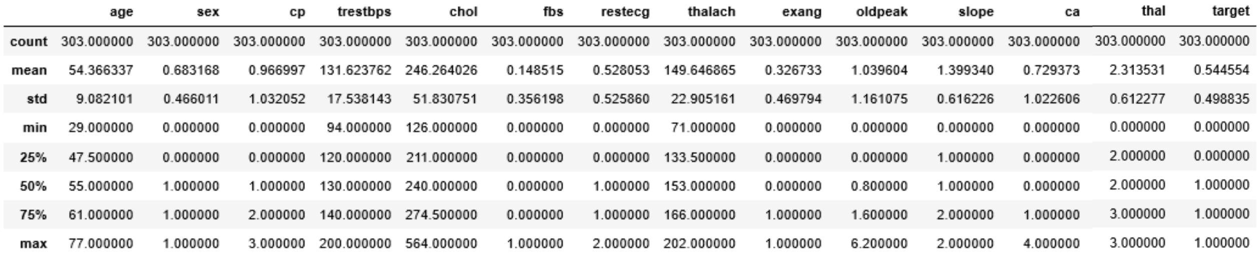
Variables of the first dataset

Features

- Age: All the patient has declared his or her age in the data. The minimum age is 29 and the maximum age is 77
- Sex: 0 means the female gender, 1 means the male gender
- CP: CP measures that different types of chest pain (4 values). There are 4 types of check pain: Value 0 - typical angina, Value 1 - atypical angina, Value 2 - non anginal type, Value 3 - asymptomatic
- Trestbps: This means “resting blood pressure”.
- Chol: This data shows the index of serum cholesterol in mg/dl
- FSB: fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
- Restecg: resting electrocardiographic results (values 0,1,2). There are 3 types of Restecg: Value 0 - norma, Value 1 - having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 005 mV), Value 2 - showing probable or definite left ventricular hypertrophy by Estes criteria
- Thalach: shows the maximum heart rate achieved
- Exang: exercise induced angina (1 = yes, 0 = no)

- Oldpeak: Oldpeak means ST depression induced by exercise relative to rest
- Slope: the slope of the peak exercise ST segment. Value 1 - upsloping, Value 2 - flat, Value 3 - downsloping
- CA: number of major vessels (0-4) colored by fluoroscopy
- Thal: the meaning of the data is “ 3 = normal; 6 = fixed defect; 7 = reversible defect”
- Target: 0 means don’t have heart disease, 1 means that he or she is having heart disease. This is a binary classification problem.

Description of Dataset One:



	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
count	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000
mean	54.366337	0.683168	0.966997	131.623762	246.264026	0.148515	0.528053	149.646865	0.326733	1.039604	1.399340	0.729373	2.313531	0.544554
std	9.082101	0.466011	1.032052	17.538143	51.830751	0.356198	0.525860	22.905161	0.469794	1.161075	0.616226	1.022606	0.612277	0.498835
min	29.000000	0.000000	0.000000	94.000000	126.000000	0.000000	0.000000	71.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	47.500000	0.000000	0.000000	120.000000	211.000000	0.000000	0.000000	133.500000	0.000000	0.000000	1.000000	0.000000	2.000000	0.000000
50%	55.000000	1.000000	1.000000	130.000000	240.000000	0.000000	1.000000	153.000000	0.000000	0.800000	1.000000	0.000000	2.000000	1.000000
75%	61.000000	1.000000	2.000000	140.000000	274.500000	0.000000	1.000000	166.000000	1.000000	1.600000	2.000000	1.000000	3.000000	1.000000
max	77.000000	1.000000	3.000000	200.000000	564.000000	1.000000	2.000000	202.000000	1.000000	6.200000	2.000000	4.000000	3.000000	1.000000

Figure 1 – Dataset 1 Summary table

Dataset Two

The second dataset has been selected after analysis of the first dataset, A question has arisen in us to know why man is a potential subject of heart issue over women , due to that doubt we have download the second dataset from Kaggle which give us some information about human activities that might be a factor of a heart disease . This data set is real data collected from patients after examination.

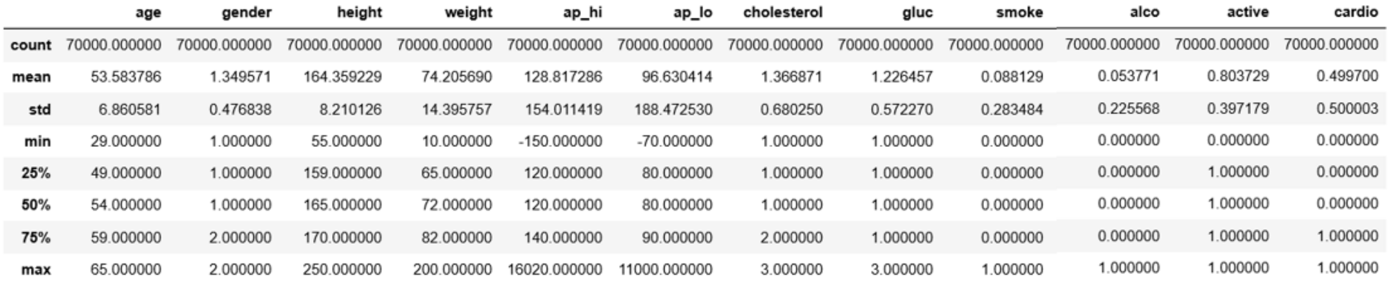
Variable for the second dataset

Features

- Age | Objective Feature | age | int (days)`
- Height | Objective Feature | height | int (cm) |

- Weight | Objective Feature | weight | float (kg) |
- Gender | Objective Feature | gender | categorical code |
- Systolic blood pressure | Examination Feature | ap_hi | int |
- Diastolic blood pressure | Examination Feature | ap_lo | int |
- Cholesterol | Examination Feature | cholesterol | 1: normal, 2: above normal, 3: well above normal |
- Glucose | Examination Feature | gluc | 1: normal, 2: above normal, 3: well above normal |
- Smoke: Smoking | Subjective Feature | binary |
- Alco: Alcohol intake | Subjective Feature | binary |
- Active: Physical activity | Subjective Feature | binary |
- Cardio: Presence or absence of cardiovascular disease | Target Variable | cardio | binary |

Description of Dataset Two



	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
count	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000
mean	53.583786	1.349571	164.359229	74.205690	128.817286	96.630414	1.366871	1.226457	0.088129	0.053771	0.803729	0.499700
std	6.860581	0.476838	8.210126	14.395757	154.011419	188.472530	0.680250	0.572270	0.283484	0.225568	0.397179	0.500003
min	29.000000	1.000000	55.000000	10.000000	-150.000000	-70.000000	1.000000	1.000000	0.000000	0.000000	0.000000	0.000000
25%	49.000000	1.000000	159.000000	65.000000	120.000000	80.000000	1.000000	1.000000	0.000000	0.000000	1.000000	0.000000
50%	54.000000	1.000000	165.000000	72.000000	120.000000	80.000000	1.000000	1.000000	0.000000	0.000000	1.000000	0.000000
75%	59.000000	2.000000	170.000000	82.000000	140.000000	90.000000	2.000000	1.000000	0.000000	0.000000	1.000000	1.000000
max	65.000000	2.000000	250.000000	200.000000	16020.000000	11000.000000	3.000000	3.000000	1.000000	1.000000	1.000000	1.000000

Figure 2 – Dataset Two Summary table

Exploratory Data Analysis

Dataset One

Integer variables

We will discover the insights from the distribution of the variables:

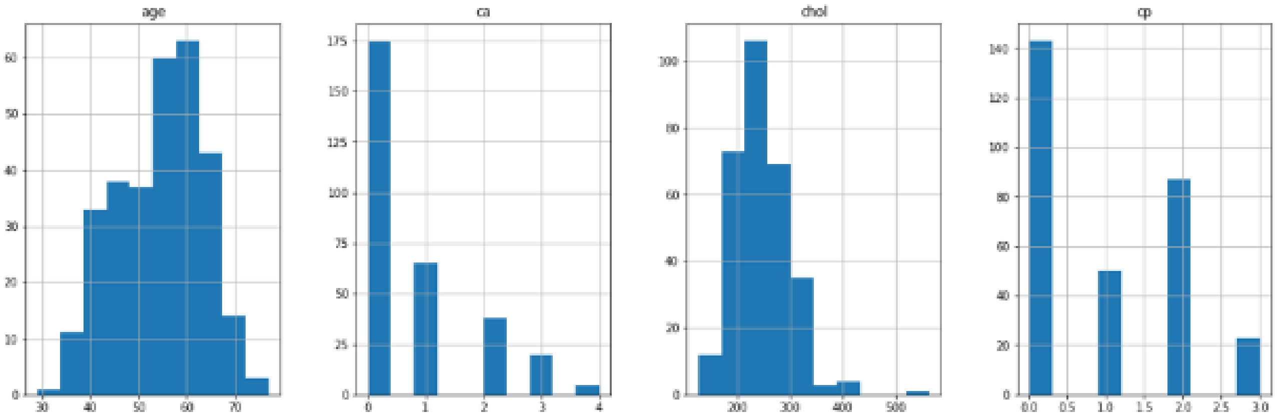


Figure 3 - From left to right: Age, Ca, Chol, Cp

We can observe the frequency of age , chest pain , cholesterol and number of major vessels .

- We observe the age range of people in the data set slice from]30,70[years old ,also from graph we have more subject people in the range of 50 to 60 years old who have been diagnostic from the hospital , we can make a conclusion that people slicing between 50 to 60+ years old check their heart health more frequently compare to others .
- Number of major vessels 0 3 colored by fluoroscopy range from (0 to 4), if the number is greater than zero then is a potential disease otherwise if it is 0 implies no potential

disease, from how dataset we state that we have more people with no potential heart issue .

- A cholesterol level higher than 200 for an adult is considered high ,we can observe some value reaching 300 which is a potential risk patient
- Any chest pain value greater than 0 , is view as a potential risk of heart disease.

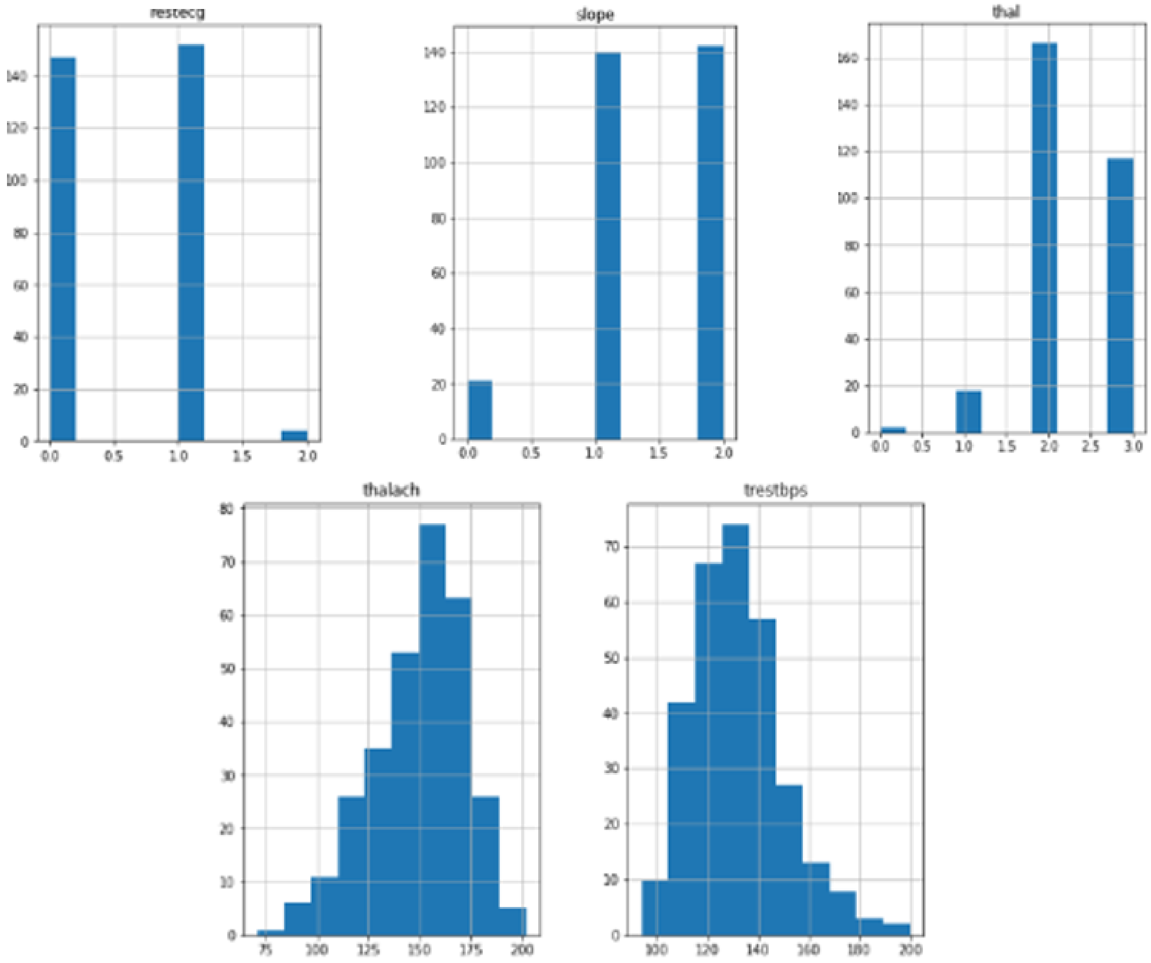


Figure 4: From left to right and from top to bottom, histogram of Restecg, Slope, Thai, Thalach, Trestbps

As in the Figure, we can see the highest value for Thalach skews lies between 125 to 175, meanwhile the Trestbps skews lies between 110 to 150 respectively.. For Thai, the figure also shows the highest type which is 2 and 3, at the same time. In terms of Chol factor, the data is skewed into 150 and 300 which shows the result of the heart beat over a group of patients .

Binary Variables (Integer Variables)

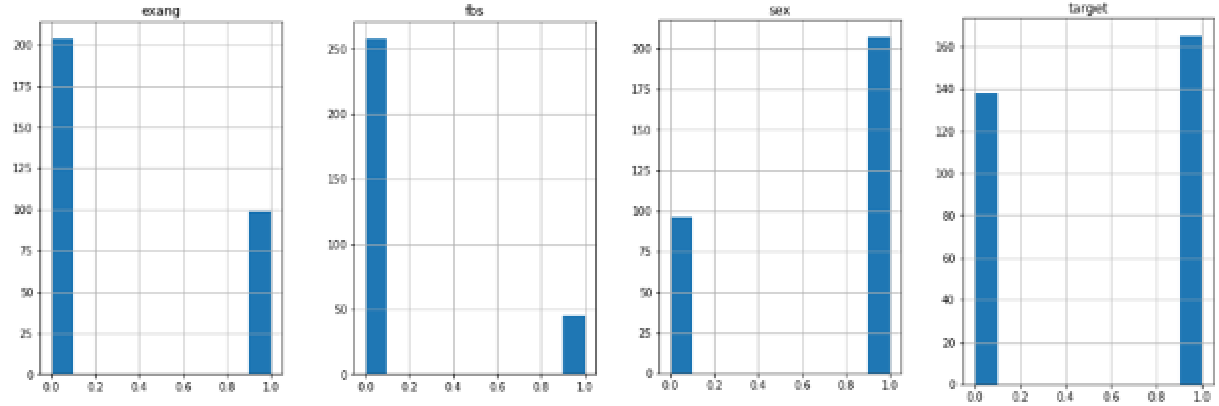


Figure 5: Histogram of Exang, Fbs, Sec and Target

In terms of gender, we can see the number of men is double that of women. On the other side, on Target side we observe the number of heart diseases of people are quite similar versus people with non heart disease. The figure also shows that Majority of Exang and Fbs are 0.

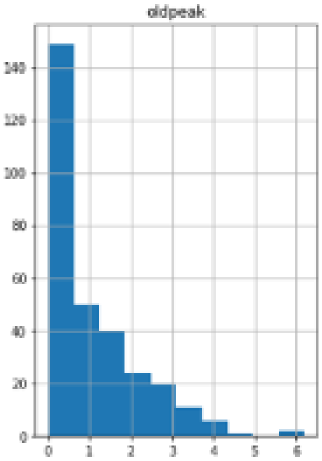


Figure 6: Float Variables - Oldpeak

From the Oldpeak data, we can see that the mean score is around 1.04 points, and the majority of indexes are below 1.

Dataset Two

Since, some of the factors are duplicate with the Dataset One, we chose the most relevant data factors from Dataset Two to analyze.

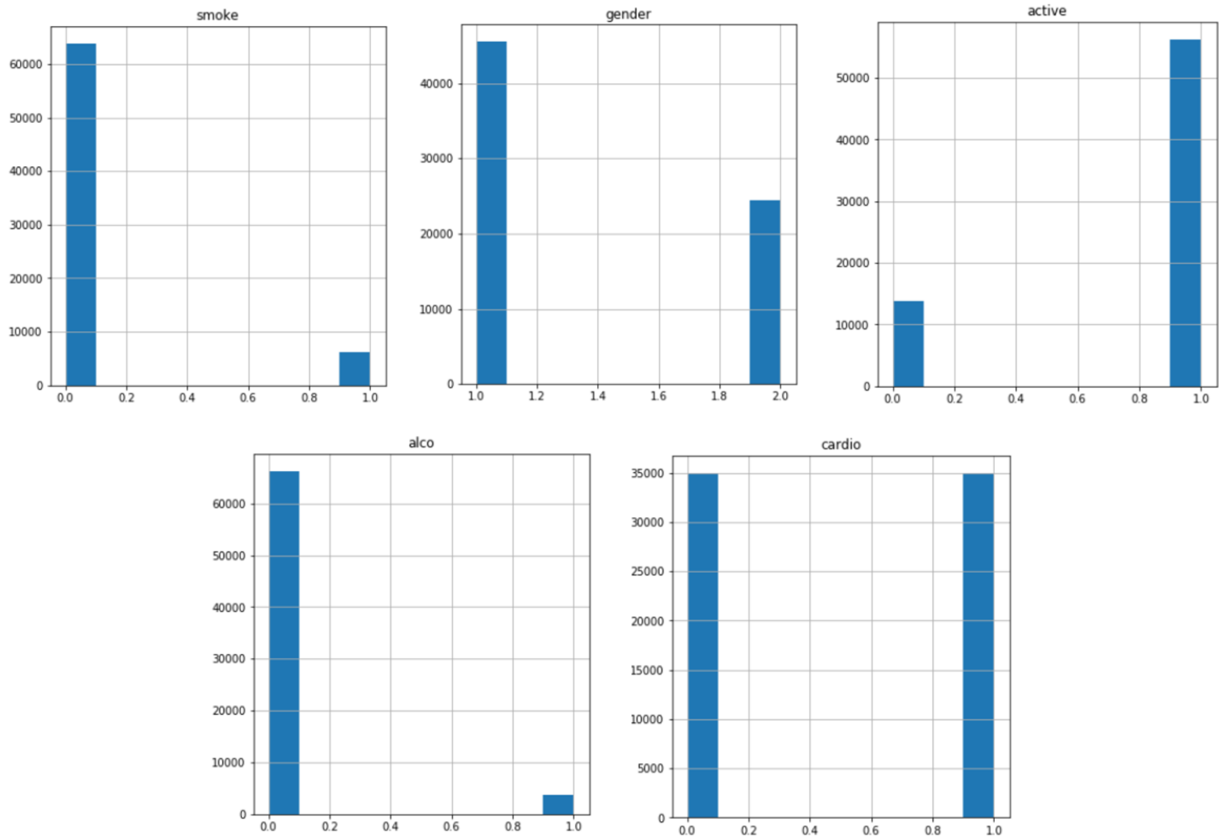


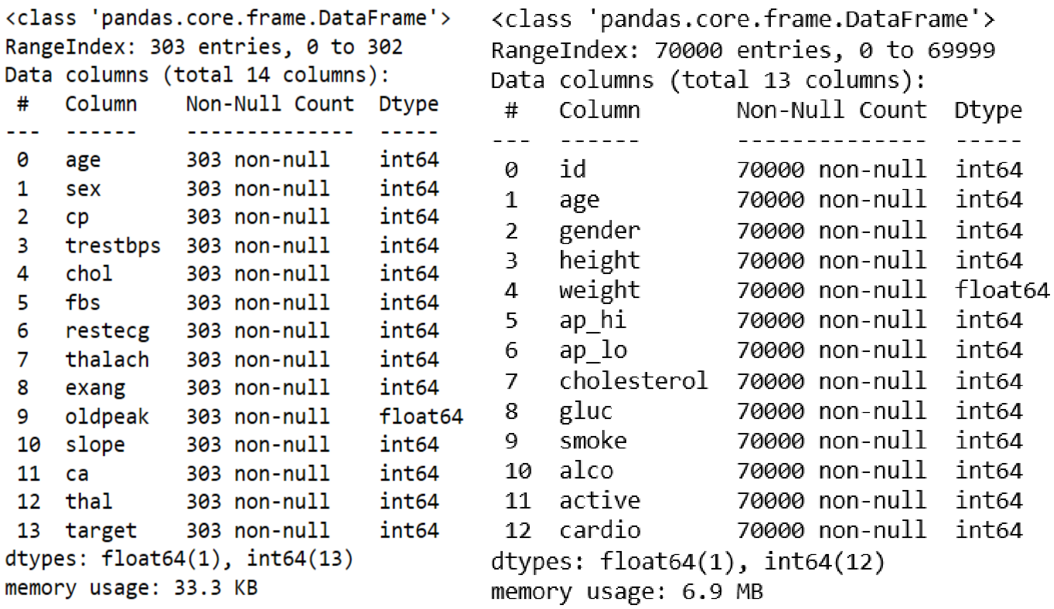
Figure 7 - From left to right, top to bottom - Smoke, Gender, Active, Alco and Cardio

As we have mentioned above, this extra dataset is to find the root cause and activities that lead to heart disease, in this case the classification data is Cardio (0 means normal, 1 means heart disease). There are 4 types of activities: Smoke, Alcohol, Activity (Physical activity). For Smoke and Alcohol, we can see that 90% of patients did not smoke or drink alcohol. In terms of activity, the majority of patients had physical activity.

There is also a small difference vs Dataset One; 1 means woman and 2 means man.

Cleaning Data - Dealing with Missing Data

The dataset has no missing data; hence the data is ready to analyze. For the session of dealing with missing data, but we apply the data standardization.

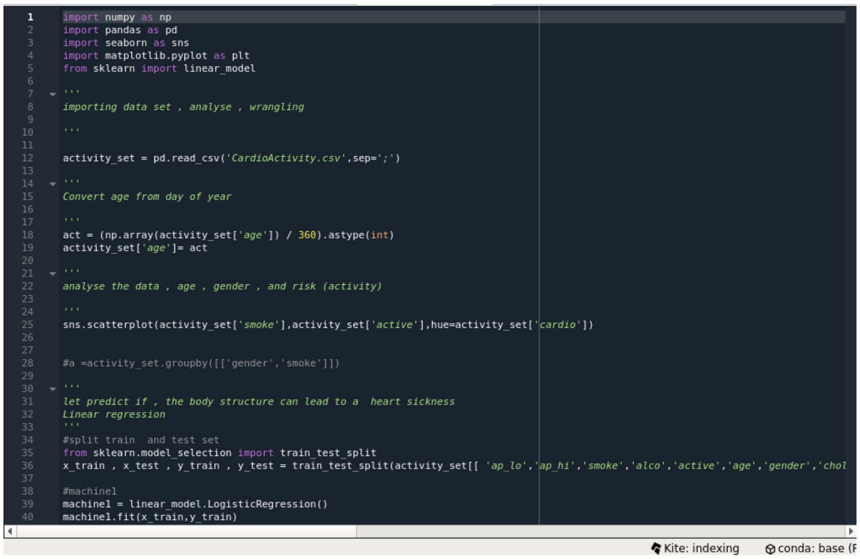


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 303 entries, 0 to 302
Data columns (total 14 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   age       303 non-null    int64  
 1   sex       303 non-null    int64  
 2   cp        303 non-null    int64  
 3   trestbps  303 non-null    int64  
 4   chol      303 non-null    int64  
 5   fbs       303 non-null    int64  
 6   restecg   303 non-null    int64  
 7   thalach   303 non-null    int64  
 8   exang     303 non-null    int64  
 9   oldpeak   303 non-null    float64
 10  slope     303 non-null    int64  
 11  ca        303 non-null    int64  
 12  thal      303 non-null    int64  
 13  target    303 non-null    int64  
dtypes: float64(1), int64(13)
memory usage: 33.3 KB
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 70000 entries, 0 to 69999
Data columns (total 13 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   id           70000 non-null  int64  
 1   age          70000 non-null  int64  
 2   gender       70000 non-null  int64  
 3   height       70000 non-null  int64  
 4   weight       70000 non-null  float64
 5   ap_hi        70000 non-null  int64  
 6   ap_lo        70000 non-null  int64  
 7   cholesterol  70000 non-null  int64  
 8   gluc         70000 non-null  int64  
 9   smoke        70000 non-null  int64  
 10  alco         70000 non-null  int64  
 11  active       70000 non-null  int64  
 12  cardio       70000 non-null  int64  
dtypes: float64(1), int64(12)
memory usage: 6.9 MB
```

Figure 8: Data type and Non-Null Count of Dataset One and Two

From Data Standardization, we change the age from number of days to years to ensure the harmonizing between two Dataset.



```python
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn import linear_model

'''
importing data set , analyse , wrangling

'''

activity_set = pd.read_csv('CardioActivity.csv',sep=';')

'''
Convert age from day of year

'''
act = (np.array(activity_set['age']) / 360).astype(int)
activity_set['age']= act

'''
analyse the data , age , gender , and risk (activity)

'''
sns.scatterplot(activity_set['smoke'],activity_set['active'],hue=activity_set['cardio'])


#a =activity_set.groupby([['gender','smoke']])

'''
let predict if , the body structure can lead to a  heart sickness
Linear regression
'''
#split train  and test set
from sklearn.model_selection import train_test_split
x_train , x_test , y_train , y_test = train_test_split(activity_set[[ 'ap_lo','ap_hi','smoke','alco','active','age','gender','chol

#machine1
machine1 = linear_model.LogisticRegression()
machine1.fit(x_train,y_train)
```

Figure: Data Standardization

Questions raised from the dataset

Deep Analysis the Dataset

To understand the Dataset, we are using the analysis question and investigating the data to answer the questions. Here are the key questions defined:

1. Which gender has the most heart disease?

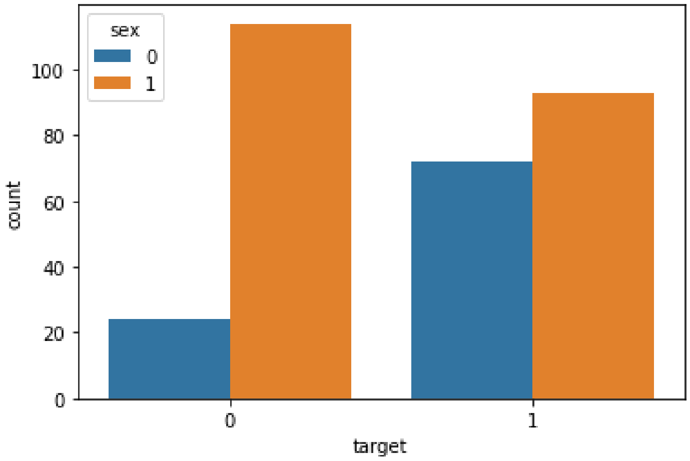


Figure 9 – Data of sex vs Target

(0 is female, non-disease; 1 is male, disease)

We observe that men are potential subjects of heart disease compared to women. To understand the root cause, we will deep dive in the extension part (Dataset Two).

2. What is the age range affecting the most by heart disease?

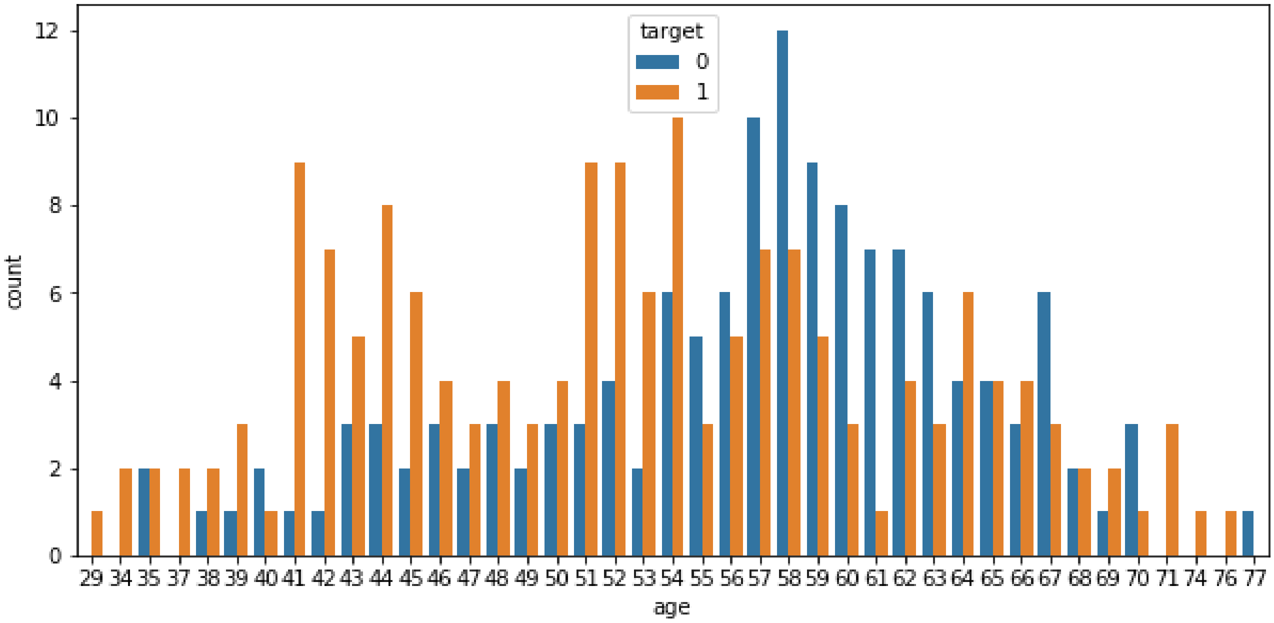


Figure 10 – Data of Age vs Target

(‘Target ’ 0 non-disease; 1 is disease)

The figure showed that all potential heart issues are discovered on people whose age range lies between 41 to 54.

3. Which are the factors that lead to heart disease?

   Relation between old peak and heart disease

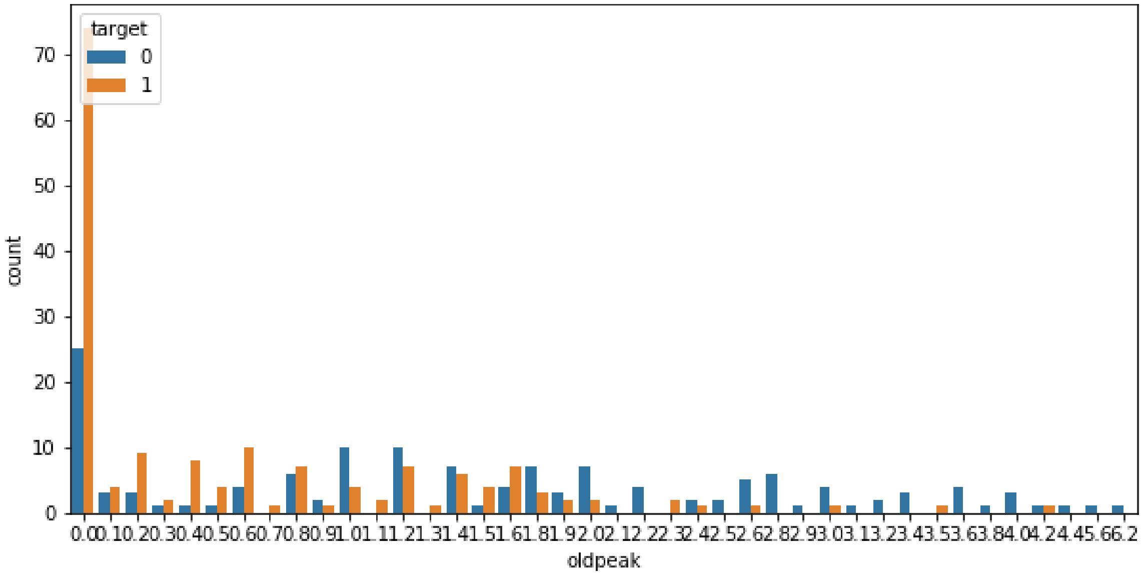


Figure 11 – Data of Oldpeak vs Target

(Target: 0 non-disease; 1 is disease)

The graph shows that the risk of having a heart disease , A subject with an oldpeak less the 0.12 is potential subject to a heart disease , the most subject to a heart issue at 100% is when the oldpeak is equal to zero .

4. Is there a relation between chest pain and heart disease?

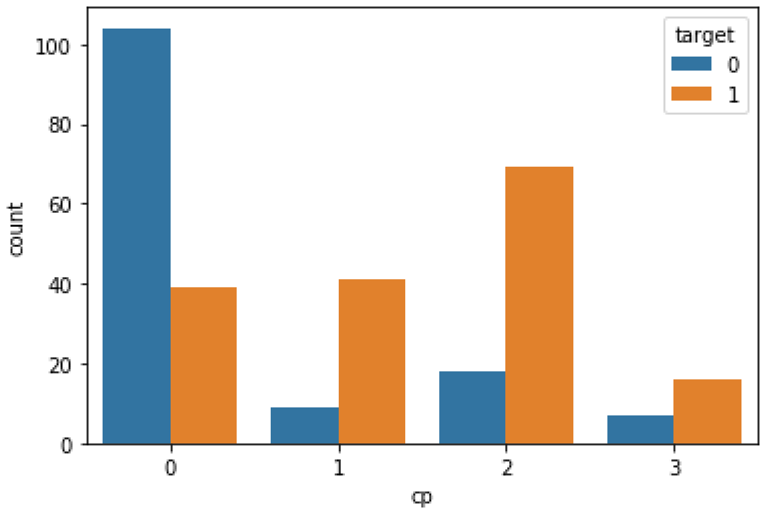


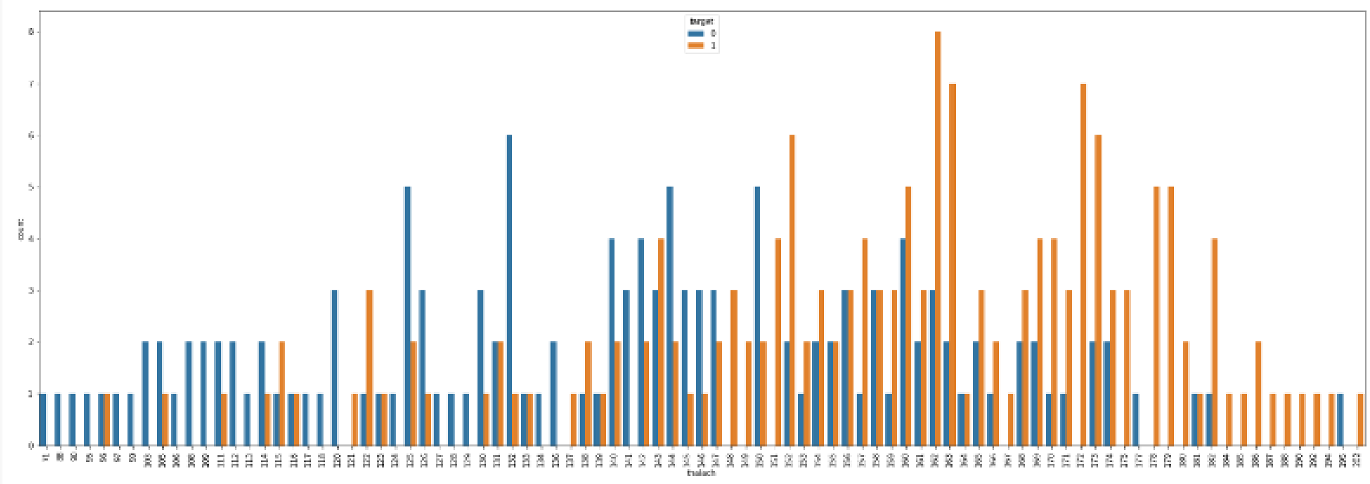
Figure 12 – Data of Cp vs Target

(Target: 0 non-disease; 1 is disease)

We come across a potential relation between chest pain and heart issue, we can state that the fact of having a chest pain level greater than 0 , is a factor of heart dysfonctionnement

6. Relation between chest pain and heart disease?

Figure 13 – Data of Thalach vs Target (Target: 0 non-disease; 1 is disease)



We observe that if the maximum heart rate achieved by a patient lies from 140 to 186 , then there is a potential chance that that patient has a heart issue .

7. Relation between heart issue and resting blood pressure?

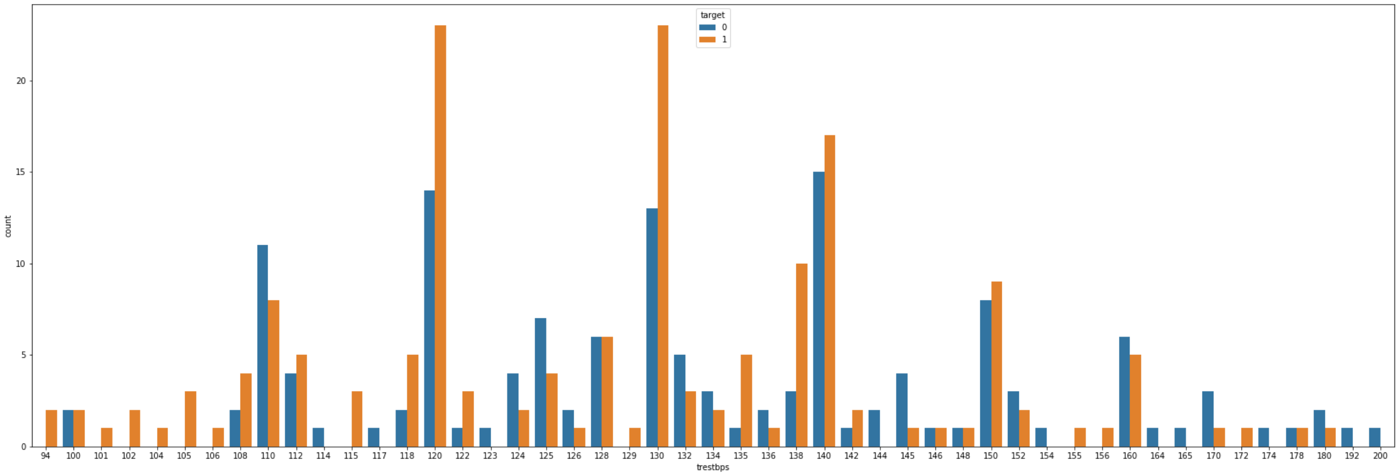


Figure 14 – Data of Trestbps vs Target

(Target: 0 non-disease; 1 is disease)

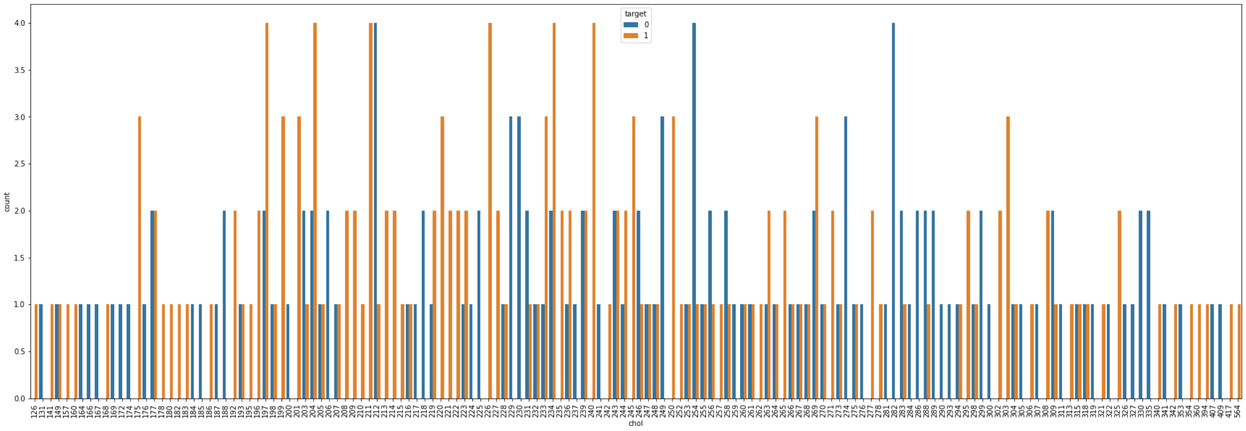


Figure 15 – Data of Chol vs Target

(Target: 0 non-disease; 1 is disease)

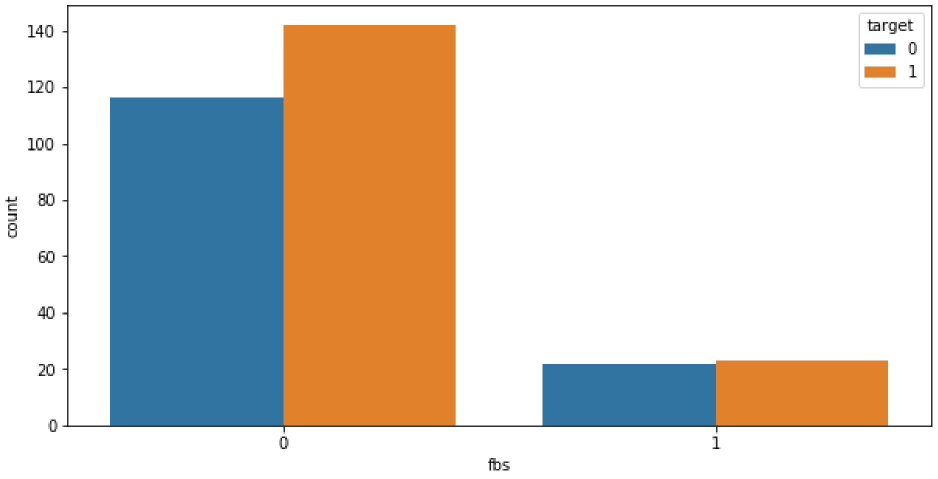


Figure 16 – Data of Fbs vs Target

(Target: 0 non-disease; 1 is disease)

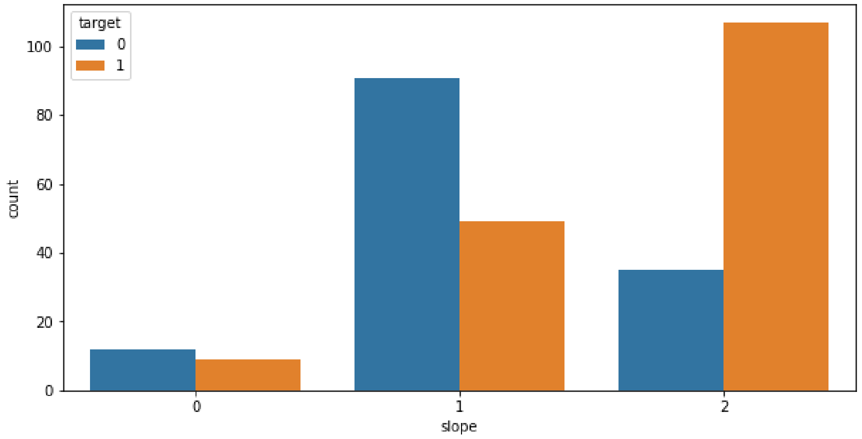


Figure 17 – Data of Slope vs Target

(Target: 0 non-disease; 1 is disease)

From the Figure, we can see that in case the Slope is 1, the chance of heart disease is 35%; in case the Slope is 2, the chance of heart disease is 80%.

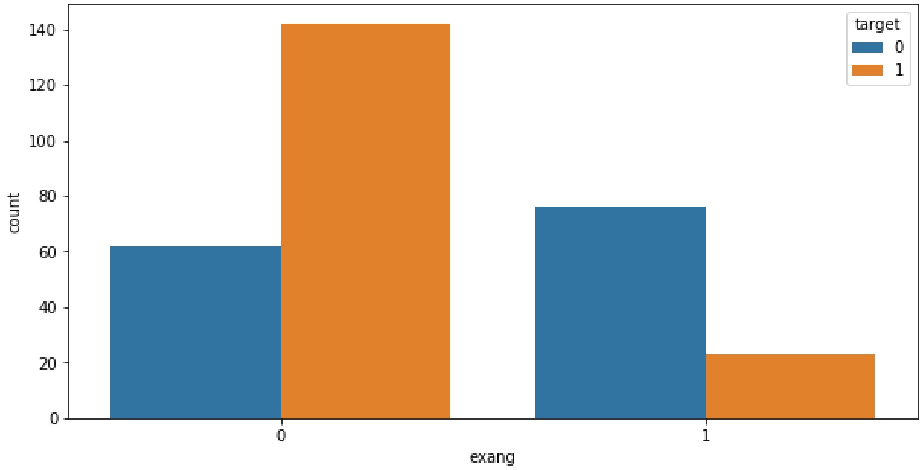


Figure 18 – Data of Exang vs Target

(Target: 0 non-disease; 1 is disease)

From the Figure, we can see that in case the Exang is 0, the chance of heart disease happened more than 80%.

Conclusion

From the multiple figures above, we can see that Oldpeak (at 0.0 to 0.3), CP (type 0, 1, 2), Thalach (from 140 to 186), Trestbps (110 to 150), Slope (1, 2), Exang (0) and Sex (Male) have the tendency to lead heart disease.

Which Factors have the correlation to each other's?

Correlation between variables

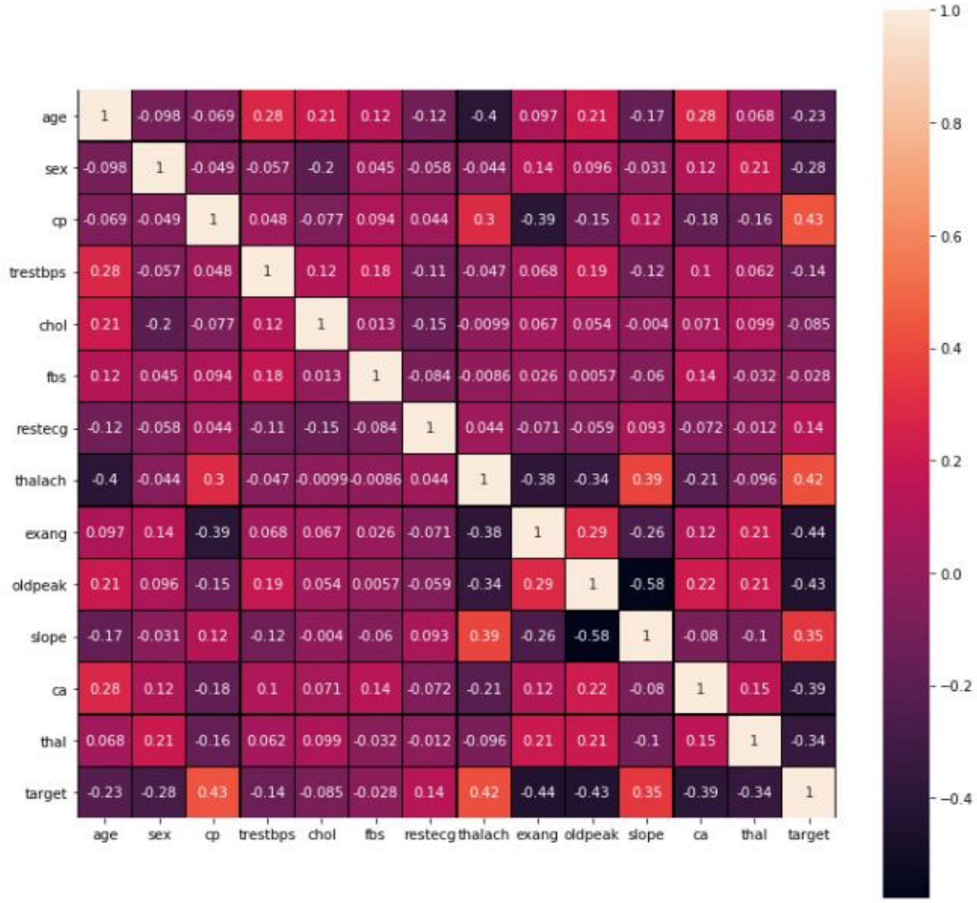


Figure 19 – Correlation Matrix of Dataset

To validate the learning above, we now run the correlation matrix chart to see the relation between the factors. From the figure, we can clearly see these are strong correlation between:

- ✓ Cp and Exang
- ✓ Slope and Oldpeak
- ✓ Target and Ca
- ✓ Target and Thai
- ✓ Target and Exang
- ✓ Target and Oldpeak
- ✓ Thalach and age

Plot of Correlation between variables and Target

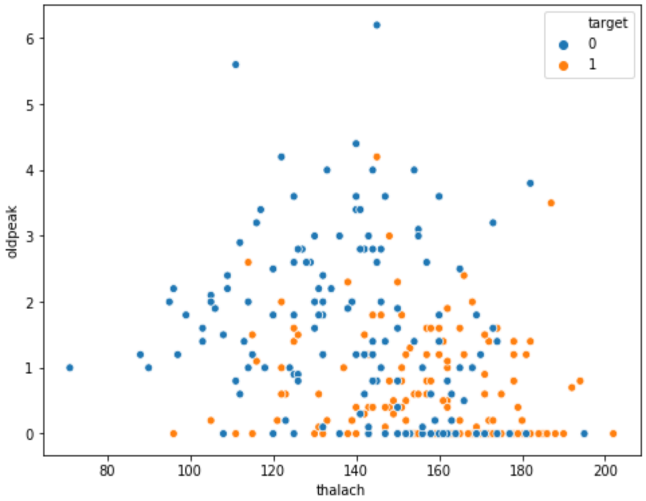


Figure 20 – Scatterplot (Thalach, Oldpeak and Target)

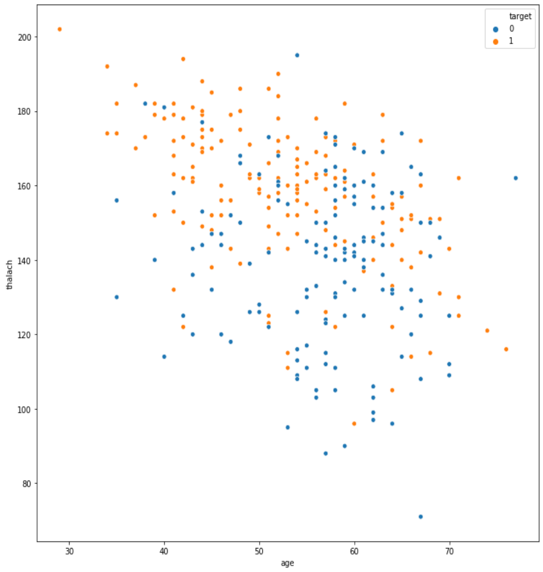


Figure 21 – Scatterplot (Thalach, Age and Target)

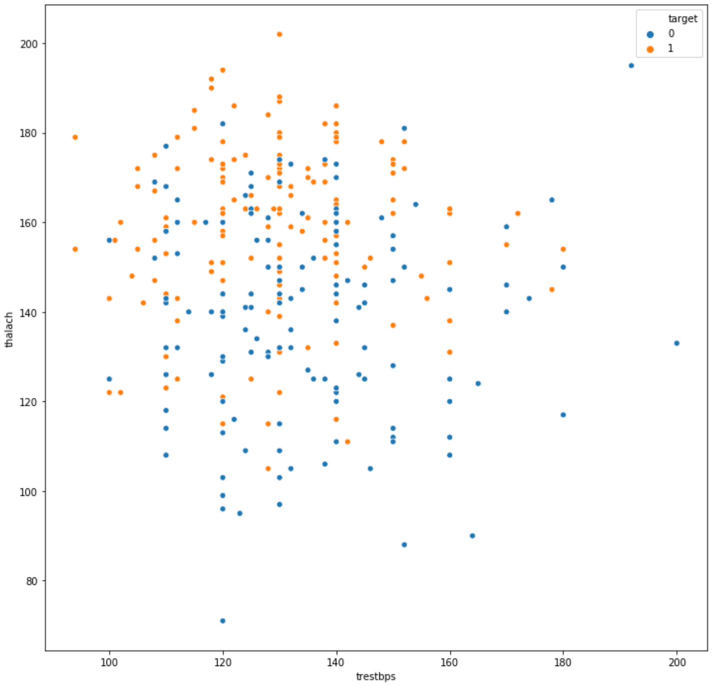


Figure 22 – Scatterplot (Thalach, Trestbps and Target)

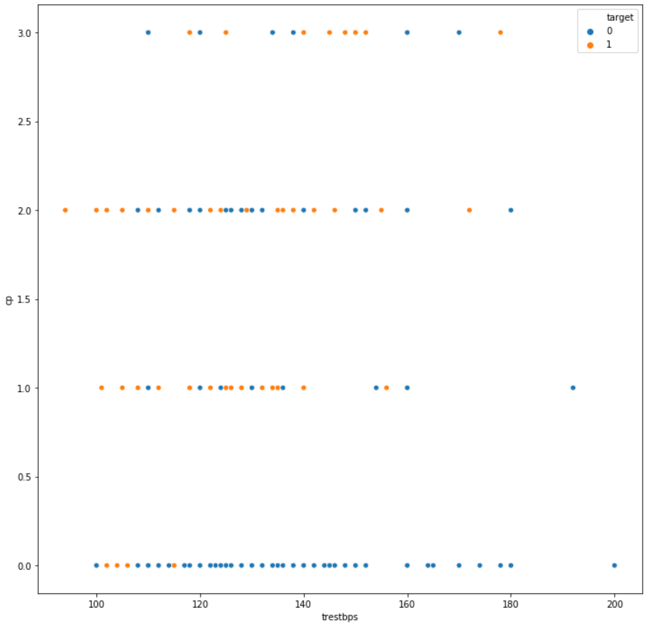


Figure 23 – Scatterplot (Cp, Trestbps and Target)

Prediction Model 1 - The Algorithm to checking Heart Disease

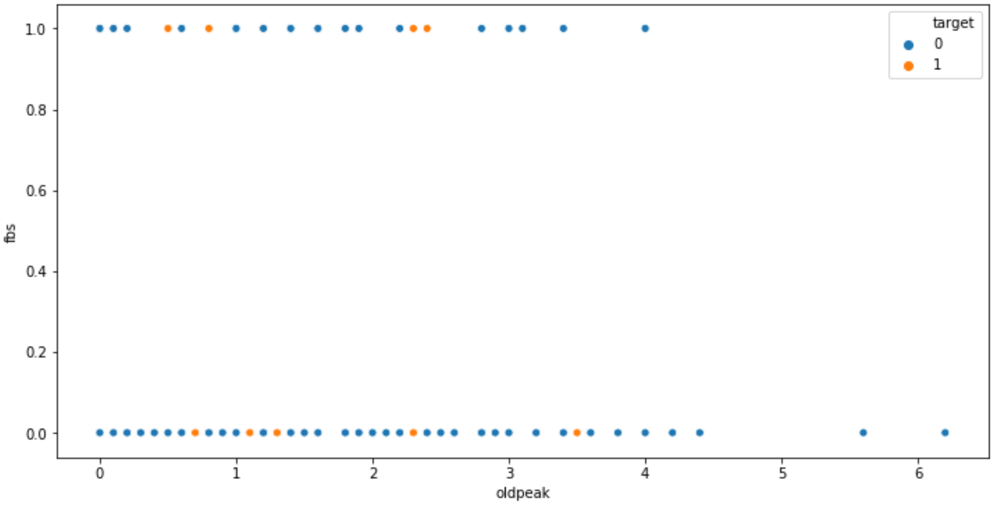


Figure 24

As we can see from the above fig , Our model looks like a Logistic Regression where we have to classify people having heart disease and people not having a heart issue base on 'sex', 'age', 'oldpeak', 'fbs' as input data and predict whether they might be a subject of a heart disease (Dataset one) .

Why Logistic Regression Classification?

A logistic Regression is a type of regression based on the mathematics equation 1/1+e^-x, this regression as a basic shape of S which fits perfectly our model.

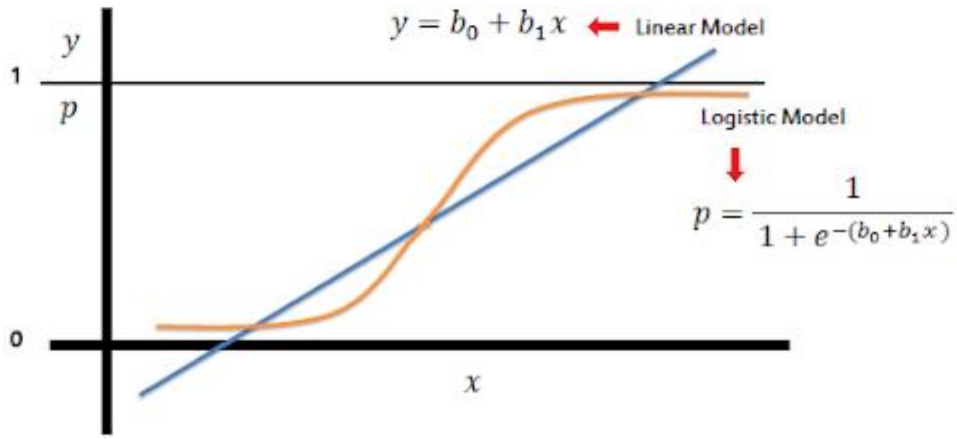


Figure 25 – Logistic Model

Our model can be viewed as multiple regression model , where many features are independent variable and only one dependent variable feature , (y = a + bx + cx^2 +......+).

Different Logistic Regression Tested for Our model

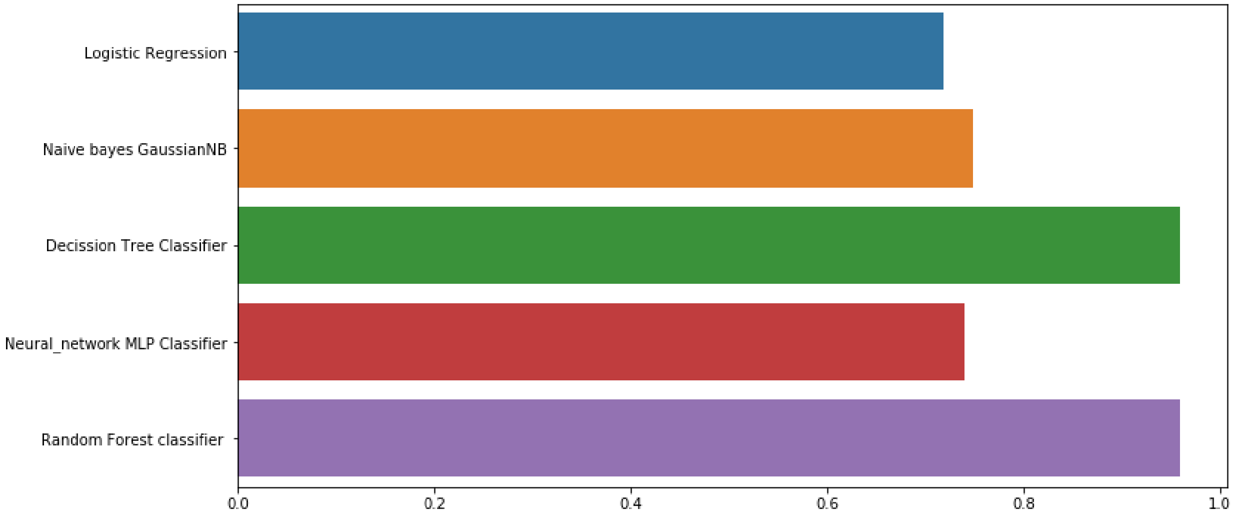


Figure 26 - Plot of different Logistic classifier result

The most accurate algorithm for our model is either Decision Tree classifier or random forest classifier which give us a prediction almost close to 80% percent . from our testing set data .

The extension of the project

As mentioned above, to further understand the root causes that impact the heart disease we have investigated in external data set of cardio_train.csv from the website Kaggle.com to see other related factors to heart disease.

The external dataset analysis

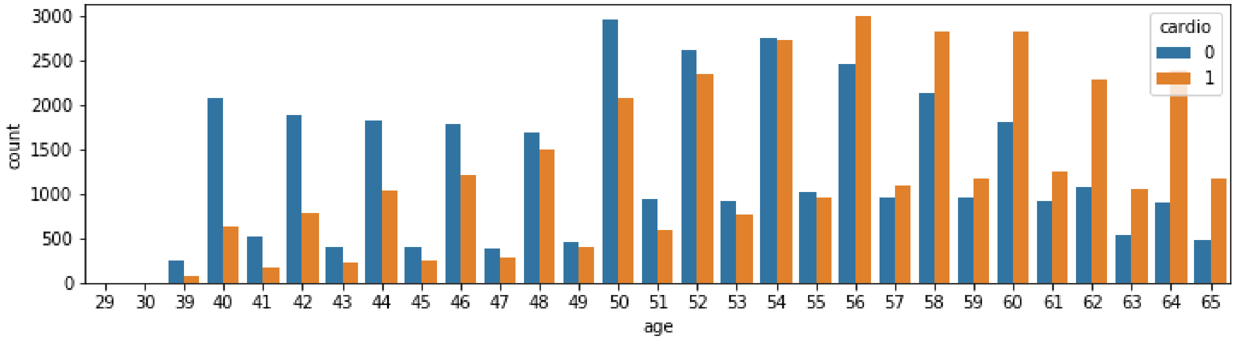


Figure 27 – Age vs Cardio

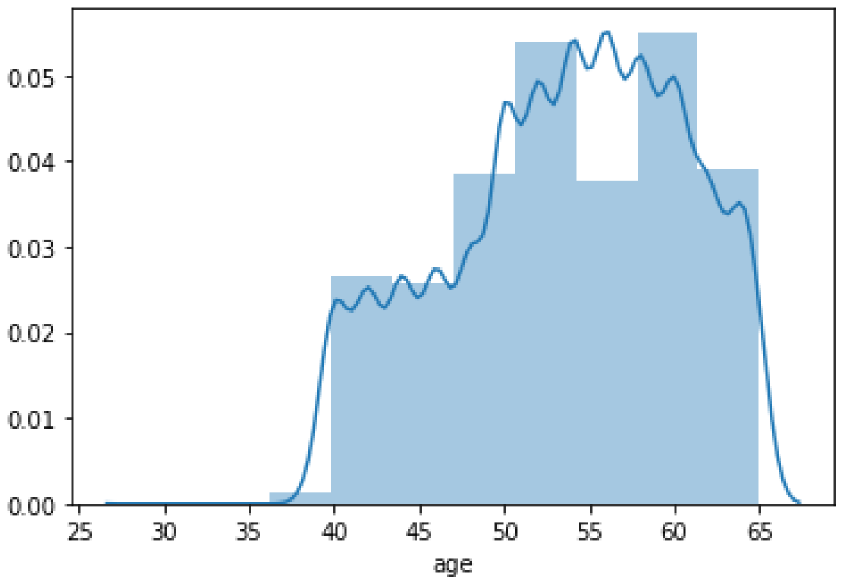


Figure 28 – Distribution of Age

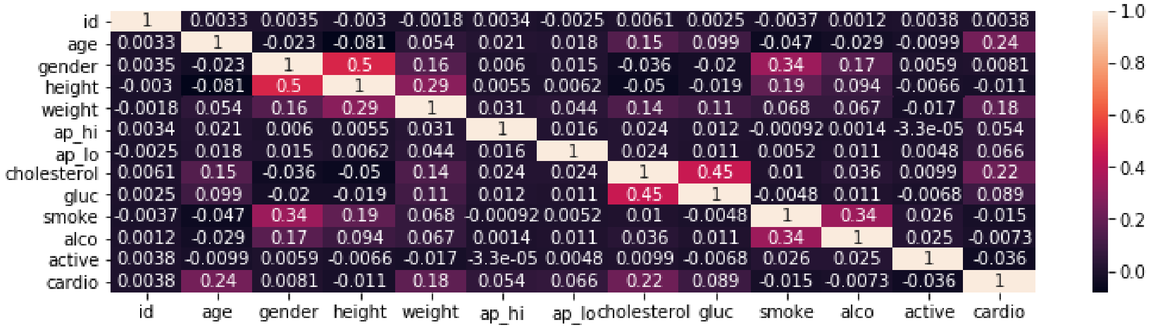


Figure 29 – Correlation Matrix

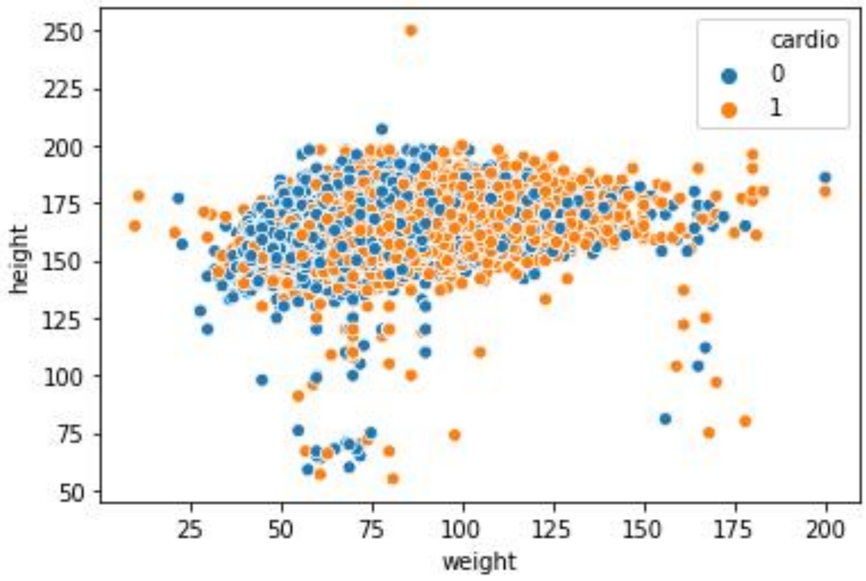


Figure 30 – Correlation Height, Weight and Cardio

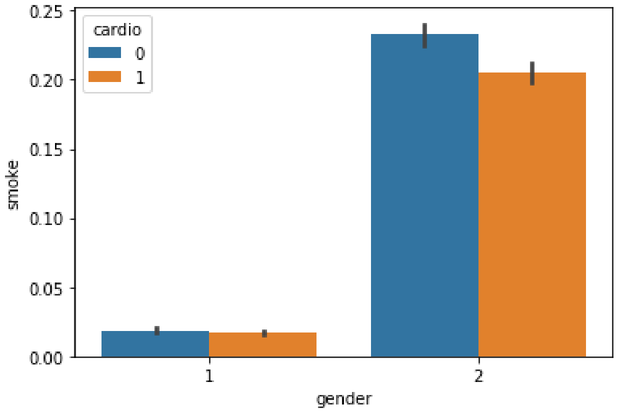
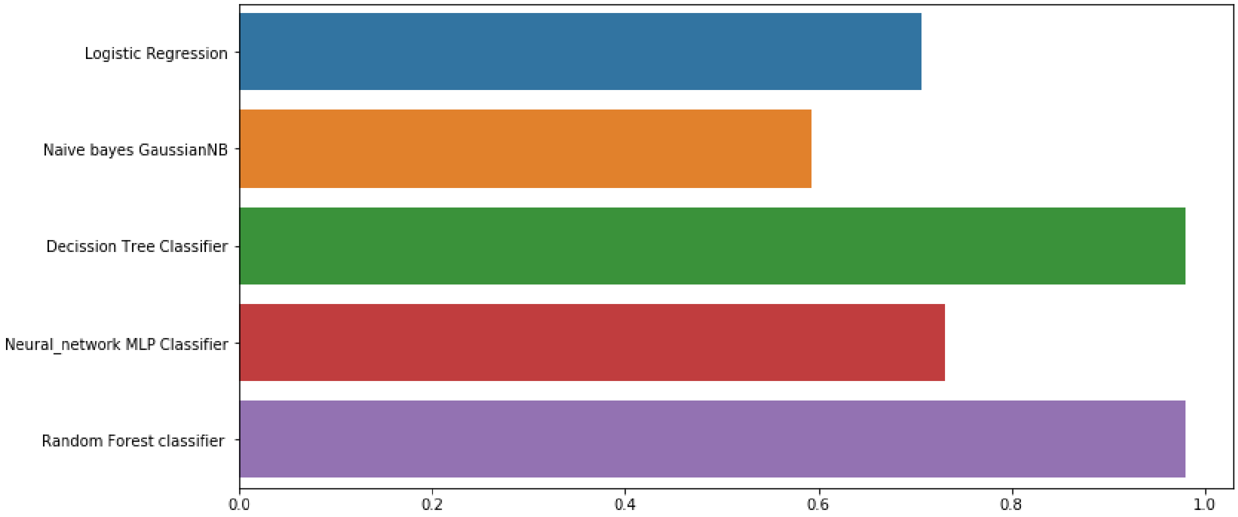


Figure 31 – Correlation Smoke, Gender and Cardio

From the above study (Dataset One), we can see that male had the tendency to have heart disease more than females. In this figure of Dataset Two, smoke shows that men smoke much more women. Hence, smoking can be a factor that relates to the heart disease gender differentiation.

Prediction Model 2 - The Algorithm to checking Heart Disease

How second data sets also classify , people base their activity to predict whether they can be a subject for a heart disease , those factors are based on 'ap_lo', 'ap_hi', 'smoke', 'alco', 'active', 'age', 'gender', 'cholesterol', 'gluc', 'height', 'weight'.



After testing different algorithm Decision Tree and Random Forest fit our model about 70 percent accuracy

Conclusion

Through the project of Heart Diagnosis, we have applied the Logistic Regression Classification with 5 different types of Algorithm including: Logistic Regression, Naïve bayes GaussianNB, Decision Tree Classifier, Neural_network NLP Classifirer, Random Forest classifier. From that we can have the prediction model to help diagnose heart disease with the accuracy 80% for Dataset One and 70% for the Dataset Two. The learning from the project helps us to practice from the Data Analysis, Deep dive understanding, Making the hypothesis and validating the hypothesis by Optimizing data and Applying Logistic Regression Classification.

Acknowledgements

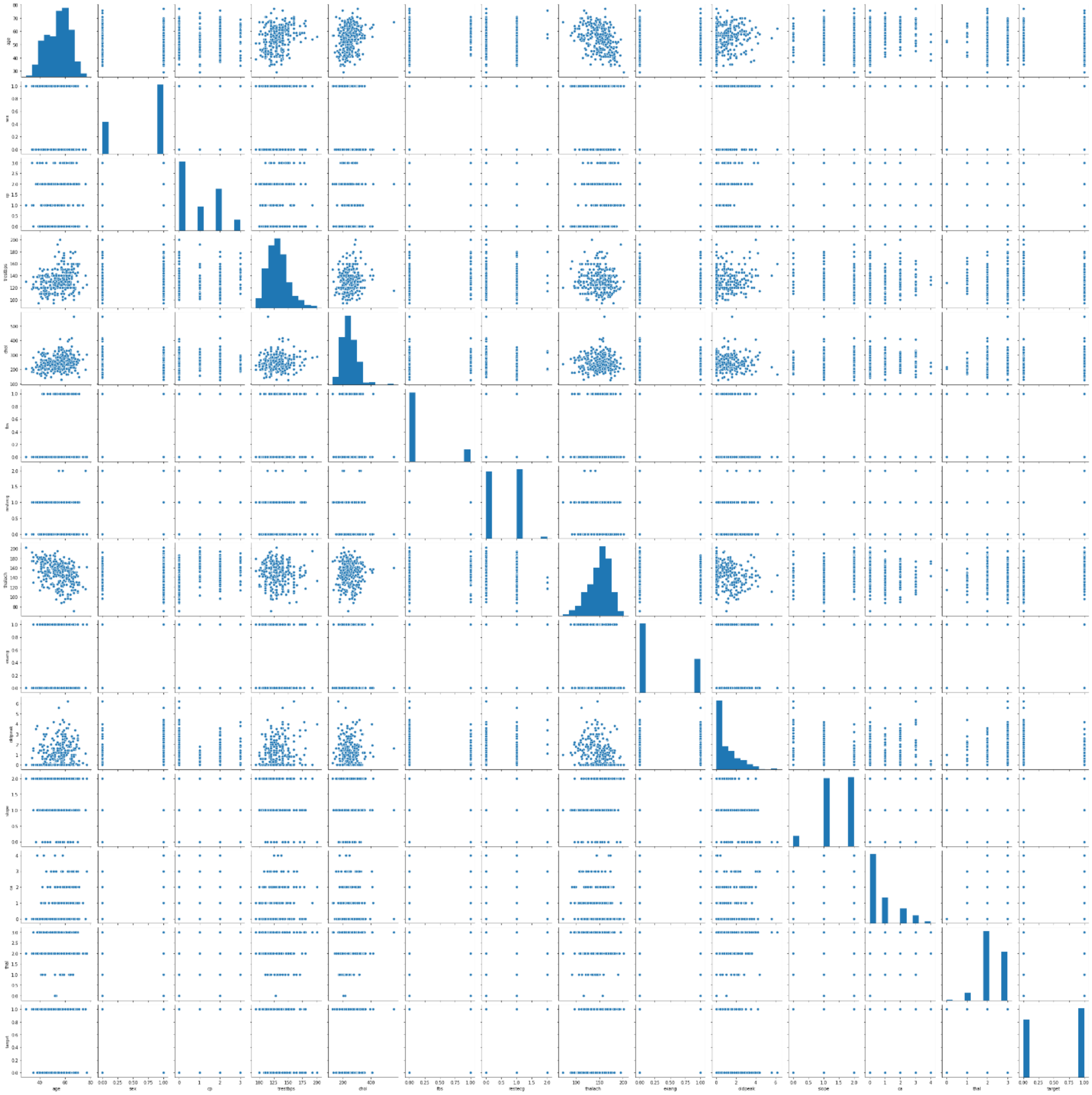
Creators of Dataset:

1. *Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.*
2. *University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.*
3. *University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.*
4. *V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.*

Appendix

Pairplot

(sns.pairplot(mydata)


age
sex
cp
trestbps
chol
fbs
restecg
thalach
exang
oldpeak
slope
ca
thal
target