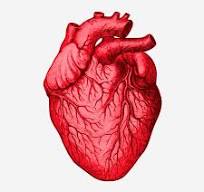
**A study of How to Detect Heart Disease**

****

**Name - Akesh Mark**

**Date - 09.03.2024**

# **Abstract**

Heart disease remains a significant public health concern globally, necessitating accurate

detection methods for timely intervention. In this study, a predictive model was developed using

an Artificial Neural Network (ANN) to identify individuals at risk of heart disease based on

various clinical and demographic features. The model, implemented using the Keras library,

achieved impressive accuracy, reaching 100% on both training and validation datasets. The

architecture comprised three dense layers with dropout regularization to mitigate overfitting.

Model training utilized the Adam optimizer with a binary cross-entropy loss function,

demonstrating robust convergence and performance. Through comprehensive evaluation using

metrics such as accuracy and confusion matrix analysis, the model exhibited exceptional

predictive capability in distinguishing between individuals with and without heart disease.

Additionally, feature scaling techniques using MinMaxScaler were applied to normalize input

data, contributing to improved model performance. The study findings underscore the potential

of deep learning approaches, particularly ANN models, in enhancing heart disease risk prediction

and guiding preventive strategies. The developed model and associated preprocessing scalers

have been saved for future deployment, facilitating widespread utilization in clinical settings.

Overall, this research contributes to advancing predictive modeling techniques for heart disease

detection, thereby aiding in early diagnosis and targeted interventions to mitigate cardiovascular

risk factorts.

Contents

[**Abstract** 2](#_Toc160897854)

[**Chapter 1: Introduction** 5](#_Toc160897855)

[**1.1 Background** 5](#_Toc160897856)

[**1.2 Research problem** 6](#_Toc160897857)

[**1.3 Research questions** 7](#_Toc160897858)

[**1.4 Objectives** 7](#_Toc160897859)

[**1.5 Scope of the research** 7](#_Toc160897860)

[**1.6 Justification of the Research** 8](#_Toc160897861)

[**1.7 Limitations** 8](#_Toc160897862)

[**Chapter 2: Literature Review** 9](#_Toc160897863)

[**2.1 Introduction to the research theme** 9](#_Toc160897864)

[**2.2 Theoretical explanation about the keywords in the topic** 9](#_Toc160897865)

[**2.3 Findings by other researchers** 10](#_Toc160897866)

[**2.4 Research gap** 11](#_Toc160897867)

[**2.5 Variables** 12](#_Toc160897868)

[**Chapter 3: Methodology** 13](#_Toc160897869)

[**3.1 Introduction** 13](#_Toc160897870)

[**3.2 Population Sample and sampling technique** 13](#_Toc160897871)

[**3.3 Types of Data to be Collected and data sources** 14](#_Toc160897872)

[**3.4 Data Preprocessing and data wrangling** 14](#_Toc160897873)

[**3.5 Methods techniques and Tools** 14](#_Toc160897874)

[**3.5.1 Artificial Neural Network in Deep Learning** 14](#_Toc160897875)

[**3.5.2 Testing for statistical significance** 15](#_Toc160897876)

[**3.5.3 Detecting outliers** 15](#_Toc160897877)

[**3.5.4 Visualization of Data and Interpretations** 15](#_Toc160897878)

[**Chapter 4: Data Analysis** 15](#_Toc160897879)

[**4.1 Importing the Dataset** 15](#_Toc160897880)

[**4.2 Data preprocessing** 18](#_Toc160897881)

[**4.3 Exploratary Data Analysis (EDA)** 25](#_Toc160897882)

[**4.3.1 Univariate Analysis** 26](#_Toc160897883)

[**4.3.2 Bivariate Analysis** 34](#_Toc160897884)

[**4.3.3 Statistical Analysis** 40](#_Toc160897885)

[**Chapter 5: Training and Testing the Model** 41](#_Toc160897886)

[**Chapter 6: Creating the web application** 50](#_Toc160897887)

[**Chapter 7: Conclusion, Discussion and Recommendation** 52](#_Toc160897888)

[**7.1 Conclusion** 52](#_Toc160897889)

[**7.2 Discussion** 53](#_Toc160897890)

[**7.3 Recommendation** 53](#_Toc160897891)

[**Appendices** 54](#_Toc160897892)

[**List of references** 69](#_Toc160897893)

List of Figures

[Figure 1 importing the dataset 17](#_Toc160898359)

[Figure 2 head of the dataset. 17](#_Toc160898360)

[Figure 3 tail of the dataset 18](#_Toc160898361)

[Figure 4 information of the dataset 18](#_Toc160898362)

[Figure 5 finding missing values of the dataset. 19](#_Toc160898363)

[Figure 6 finding outliers of the datset 20](#_Toc160898364)

[Figure 7 after removing outliers of the dataset 21](#_Toc160898365)

[Figure 8 renamed the outliers removed columns 22](#_Toc160898366)

[Figure 9 concatenated dataset 22](#_Toc160898367)

[Figure 10 after droping the columns with outliers 23](#_Toc160898368)

[Figure 11 finding missing values of the new dataset 23](#_Toc160898369)

[Figure 12 after removing null values in the dataset 24](#_Toc160898370)

[Figure 13 after removing null values from the dataset 24](#_Toc160898371)

[Figure 14 after renaming and reordering the dataset 25](#_Toc160898372)

[Figure 15 after converting the data types as before dataset 25](#_Toc160898373)

[Figure 16 Importing the preprocessed dataset 26](#_Toc160898374)

[Figure 17 summary of the dataset 26](#_Toc160898375)

[Figure 18 histograms of the numerical variables 27](#_Toc160898376)

[Figure 19 barchart of sex variable 28](#_Toc160898377)

[Figure 20 barchart of cp variable 29](#_Toc160898378)

[Figure 21 barchart of fbs variable 30](#_Toc160898379)

[Figure 22 barchart of restecg variable 31](#_Toc160898380)

[Figure 23 barchart of slope variable 32](#_Toc160898381)

[Figure 24 barchart of ca variable 33](#_Toc160898382)

[Figure 25 barchart of thal variable 34](#_Toc160898383)

[Figure 26 barchart of target variable 35](#_Toc160898384)

[Figure 27 heatmap of numerical variables 36](#_Toc160898385)

[Figure 28 barchart of sex vs age 37](#_Toc160898386)

[Figure 29 barchart of target vs age 38](#_Toc160898387)

[Figure 30 barchart of fbs vs target 39](#_Toc160898388)

[Figure 31 barchart of cp vs target 40](#_Toc160898389)

[Figure 32 barchart of restecg vs target 41](#_Toc160898390)

[Figure 33 importing the dataset to training 42](#_Toc160898391)

[Figure 34 converted the dataset into numpy array 43](#_Toc160898392)

[Figure 35 dividing the dataset into data and target 44](#_Toc160898393)

[Figure 36 normalizing the data and target 45](#_Toc160898394)

[Figure 37 dividing the dataset into training and testing 45](#_Toc160898395)

[Figure 38 creating the model 46](#_Toc160898396)

[Figure 39 training the model 47](#_Toc160898397)

[Figure 40 loss functions of the model 47](#_Toc160898398)

[Figure 41 accuracy functions of the model 48](#_Toc160898399)

[Figure 42 predicting using test data 49](#_Toc160898400)

[Figure 43 accuracy of the model 49](#_Toc160898401)

[Figure 44 testing the model 49](#_Toc160898402)

[Figure 45 saving the model and scaled data and scaled target 50](#_Toc160898403)

[Figure 46 testing the model with random data 50](#_Toc160898404)

[Figure 47 picture of the web application 52](#_Toc160898405)

[Figure 48 picture of results of the web app 53](#_Toc160898406)

# Chapter 1: Introduction

## 1.1 Background

In today’s world, a lot of people want a luxurious life with money so they are working

like robots to achieve their dreams. While people work as robots they forget to care about

their health and themselves. Because of that people’s lifestyles had changed. There are so

many risk factors such as chest pain, cholesterol, smoking, and unhealthy

lifestyle that can affect heart disease. Some of the risks can be controlled by the person.

High blood pressure and high sugar levels can affect a person even at a young age

because of this lifestyle. These things can lead to cardiovascular diseases. Cardiovascular

diseases are also known as heart disease.

It is the world leading cost of death and costs 18.6 million lives annually.

Cardiovascular disease patients have increased more than in the past.

Heart disease has become a major problem in low and middle-income countries. Heart

disease is the leading cause of death for white people, black people and Hispanics. There

are various types of heart disease such as Arrhythmia, Atherosclerosis, Cardiomyopathy,

and Heart infections. Chest pain, slow pulse, coldness in the arms and legs, shortness of

breath, and lightheadedness are some of the symptoms of heart disease. Over 17.9 million

people died from heart disease in 2019 and most of them were due to heart attack and

stroke.

About 80% of deaths from cardiovascular diseases are between the age of 30 and 70 and

it can be prevented. It can be prevented by living a healthy lifestyle and avoiding tobacco.

## **1.2 Research problem**

These days a lot of people are living an unhealthy lifestyles. Most people are addicted to

smoking, and alcohol drinking and some people are addicted to drugs too. So these things

are seriously injurious to health. This has become a major problem in heart disease. Heart

disease patients are increasing today because of this. Physical activity is another problem

that causes heart disease. There are also some other risk factors that we can’t control but

they also affect heart disease. Some of them are age, sex, and family history.

Heart disease can be controlled by living a good healthy lifestyle and with medicines and

in some cases surgeries. In this study, I will show how risk factors affect heart disease.

Some people don’t know whether they have heart disease or not. So it is very important

to know the symptoms of heart disease so the doctors can give the right treatments

quickly.

In this research, I try to predict heart disease using Deep Learning by creating a model and

developing an application so that can easily predict the risk level of the people. So this can be

very helpful for the people to find whether they have a risk of heart disease. I think this research

will help to save lot of people’s lives by knowing their risk levels and living according to them.

My aim is to reduce the number of heart disease patients in Sri Lanka and save a lot of lives so

that can be good for the country’s development too.

## **1.3 Research questions**

- What are the most significant risk factors associated with the early diagnosis of

heart disease, and how can these risk factors be identified for effective prevention

and treatment strategies?

- What is the nature of the relationship between lifestyle factors, genetic

predisposition, and environmental factors and how can this understanding inform

the development of effective preventative measures for heart disease?

## **1.4 Objectives**

Main objective – Detecting a person having heart disease or not

Other objectives

- Identifying risk factors of heart disease for early diagnosis.

- Understanding the relationship between various factors and heart disease to gain insights into

preventative measures.

## **1.5 Scope of the research**

This research is carried out to mainly analyze whether a person has heart disease or not.

The dataset was collected via the Internet and it has been used by other researchers too.

The dataset provides valuable information that may be associated with heart disease.

Therefore Researchers can explore the relationship between these factors and heart

disease incidence to identify the most significant risk factors. The research has used

Deep Learning techniques like Artificial Neural Network model to predict heart disease.

Furthermore, this research has been used to identify the relationship between various

factors using statistical tests and how these factors can affect heart disease. The outcome of the

predictions can be used to help doctors with their decisions. Sometimes doctors make wrong

decisions by their experience and knowledge so this will help them to make correct decisions at

the right time because this model will good accurate results. Also, this research will give an idea

for a person as to whether they have the risk of diagnosing heart disease.

## **1.6 Justification of the Research**

Heart disease is one of the leading causes of death worldwide and it is responsible for a

significant burden of morbidity and mortality. As previously mentioned the purpose of

this research is to detect heart disease as early as possible. This type of research has been

done in Europe countries, the USA, and some other countries like Australia. This type of

research also have been done in Sri Lanka but I think this research is special because of the

results the model predicts. So this research will be very important in Sri Lanka because heart

disease has become a major problem in Sri Lanka too. In Sri Lanka, a lot of people are dying

from heart disease because of their wrong lifestyle and alcohol and cigarettes. Sri Lanka is not a

very good country with a healthcare system. That’s why sometimes patients are transferred to

different countries. So this research will help Sri Lankans to protect their lives early by looking

at the predictions of this research and by the web application. This research will help people to

protect themselves and this will help doctors to make correct decisions.

## **1.7 Limitations**

Like most research projects this research has some limitations too. This dataset is from

one year and it consists of four countries: Cleveland, Hungary, Switzerland, and Long

Beach V. It’s better to have a dataset with countries around the world. Also it is better to have

dataset in Sri Lanka because I’m doing this research in Sri Lanka. Conclusions would be drawn

from this dataset, which contains real-world data. The model is trained by using ANN in Deep

Learning and Machine Learning models have not been used to train the model so that is a limitation too

because this can be done in Machine Learning too.

# **Chapter 2: Literature Review**

## **2.1 Introduction to the research theme**

The theme of this research is to detect Cardiovascular diseases also known as heart

diseases. This research seeks to determine how to reduce heart disease by knowing the

patients. Cardiovascular diseases are the ones that cost the most human deaths in

today’s world. Over three-quarters of CVD deaths take place in low and middle-income

countries. As a result, this chapter reviews the previous studies and tries to give the

maximum to protect human lives from this heart disease.

## **2.2 Theoretical explanation about the keywords in the topic**

Since this is a familiar topic to people this does not have many complex words.

There is an explanation of some of the important keywords below. Some other keywords

might be added throughout this research.

|  |  |
| --- | --- |
| **Keyword** | **Theoretical explanation** |
| Cardiovascular disease | This is a general term for conditions affecting the heart and blood vessels. |
| Diagnosis | The identification of the nature of an illness or other problem by examination of the symptoms |
| Thalach | The person’s maximum heart rate achieved |
| Patients | A person registered to receive medical treatment for heart disease |
| Treatments | Medical care is given to the person for illness |
| Behaviours | A person’s behaviour that can cause heart disease |

## **2.3 Findings by other researchers**

Numerous studies have been done that focus on the diagnosis of heart disease. They have

applied different data mining techniques and machine learning algorithms for the

diagnosis of heart disease and they achieved different probabilities.

(Voon Khai Tick, Ng Yung Meeng, Nur Farahiyah Mohammad, Nor Hazlyna Harun, Hiam

Alquran and Mohamad Farhan Mohamad Mohsin 1997) has done a research using ANN in Deep

Learning and they have normalized the data into floating points and they have divided the dataset

into training 70% and testing 30%. And they have tested the training with different learning rates

0.25,0.5,0.75,1.0 and they have found that 0.25 is the best learning rate for them. They have

trained this by 25 neurons and 1000 epochs. This model got 80.26% accuracy.

(Raniya R. Sarra, Ahmed Musa Dinar, Mazin Abed Mohammed 2023) has done a research using

ANN. The dataset have been devided into 242 training and 61 testing. Then the dataset have

been again dived the training set into 8:2 ration for training and validation. This model has got

93.44% accuracy. In this model they have used 2 layers and there were 30 units in the input

layer, which was also the first hidden layer.

(Polaraju, Durga Prasad, Tech Scholar 2017) have done a prediction of heart disease

using a multiple regression model. This work is done by using training data set consisting

of 3000 instances with 13 different attributes. The data have been divided into two parts

70% of the data are training data and 30% data have been used for testing.

(Hlaudi Daniel Masthe,Mosima Anna Masthe 2014) have researched heart disease. The

researchers used pattern recognition and data mining methods in the domain of

cardiovascular diagnoses. The purpose of predictions in data mining is to help discover

trends in patient data to improve their health. The researchers used data mining

algorithms decision trees, naïve Bayes, neural networks, association classification and

genetic algorithms for predicting and analyzing heart disease from the dataset.

(Purushottam, Saxena, and Sharma 2016) have done a heart prediction using data mining.

This helps medical practitioner to make effective decision making based on certain

parameters. Training and testing provide 86.3% accuracy in the testing phase and 87.3%

in the training phase.

(Yangguang He, Xinlong Li, Ruixian Song) have done a heart disease project to predict

whether a person has heart disease. Different methods have been used in this research

such as logistic regression, SVM, naïve Bayes, random forest and artificial neural work.

(Beyene and Kamat 2018) researched heart disease. This research targets one of the most

common problems in medical centres. It is about all experts do not have equal

knowledge to treat their patients so they give their own decision that may give poor

results. Machine learning algorithms like decision trees, naïve Bayes, k-nearest

neighbour, support vector machine and artificial neural networks have been used in this

research.

**2.4 Research gap**  
In this research I have used Artificial Neural Network in Deep Learning to predict heart disease

using important features. That has been very successful because the model has 100% accuracy

for the validation data and 99% accuracy for the training data. So I think this model will give

incredible results and this will give good help for the doctors. Also I have develop a heart risk

level predictor application so I think that is also a plus point in this research. Also I have used

statistical tests like t-test,chi-squared test,anova test to determine the relationship between the

variables. Most of the research don’t have used this statistical tests to identify relationships and

that is an advantage in this research.

## **2.5 Variables**

|  |  |  |
| --- | --- | --- |
| **Variable** | **Description** | **Values** |
| Age | Patient’s age in years | Continuous value |
| Sex | Sex of patient | 1=Male 0=Female |
| Cp | Chest pain | 0:typical angina 1:atypical angina 2:non-angina pain 3:asymptomatic |

|  |  |  |
| --- | --- | --- |
| Trestbps | Resting blood pressure | Continuous value in mm/Hg |
| Chol | Serum cholesterol in mg/dl | Continuous value in mg/dl |
| Fbs | Fasting blood sugar | 1>=120mg/dl 0<=120mg/dl |
| Restcg | Resting electrocardiographic results | 0=normal 1=having\_ST\_T wave abnormal 2=left ventricular hypertrophy |
| Thalach | Maximum heart rate achieved | Continuous value |
| Exang | Exercise induced angina | 1:yes 0:no |
| Oldpeak | ST depression induced by exercise relative to rest | Continuous value |
| Slope | The slope of the peak exercise ST segment | 0:upsloping 1:flat 2:down sloping |
| Ca | Number of major vessels colored by fluoroscopy | 0-4 value |
| Thal | Defect type | 0=normal 1=fixed defect 2=reversable defect  3=another effect |
| Target | Diagnosis of heart diseases | 1:yes 0:no |

# **Chapter 3: Methodology**

## **3.1 Introduction**

In this study, we're looking closely at information about heart disease that we found on Kaggle, a

website where researchers share data. Our main goal is to detect people that are having heart

disease and understand what things might be related to heart disease and how we can find it

early to help stop it from getting worse. We're using a computer program called Python to sort

through the data and figure out what's important. Then, we'll make graphs and pictures to help us

see the information better. Our hope is that by doing this, we can learn more about heart disease

and maybe find new ways to prevent it from happening in the first place, which would be really

good for people's health.

## **3.2 Population Sample and sampling technique**

Out of the entire population, this dataset is a sample for this study.

Because the population for this study will be large and it is very hard to deal with

population data. So systematic sampling procedures will be employed. As a result of

sampling, precision must be at a high level.

## **3.3 Types of Data to be Collected and data sources**

This dataset was obtained from Kaggle as mentioned earlier. Since this dataset is not

collected by me this is not a primary dataset. It means that this is already available on the

internet so this is secondary data. This dataset has numerical values and categorical

values so it will be easy to analyze. The dataset is in CSV format.

## **3.4 Data Preprocessing and data wrangling**

Data cleaning is very important in data preprocessing. Data cleaning techniques are used

to clean the dataset, before going through with the analysis. Data cleaning is very crucial

because it gives a high-quality research study. Some data-cleaning techniques would be

used in this research to find missing and null values and remove duplicate data and removing

outliers. It is important to mention that this is a secondary dataset that is collected by someone

else. The file is in a CSV format, making it easy to understand. Normalization of the data has

been done before training the model in ANN.

## **3.5 Methods techniques and Tools**

### **3.5.1 Artificial Neural Network in Deep Learning**

In our research on heart disease, we use a special type of computer program called Artificial

Neural Networks (ANNs). These networks are part of a bigger system called deep learning,

which helps us find hidden patterns in lots of data. By training these networks with information

like age, cholesterol levels, and lifestyle habits, we can predict if someone might have heart

disease. These programs learn from examples and get better at making predictions over time.

With their help, we aim to understand and detect heart disease more accurately, which is the

main goal of our research.

### **3.5.2 Testing for statistical significance**

The first step in conducting a statistical significance is to state the null hypothesis and

alternative hypothesis. The second step is selecting the alpha value. After that, we have to

choose a test like a t-test, ANOVA test, or chi-square test to compute the statistical

significance. In the final step, we can interpret the results by rejecting the null hypothesis

or alternative hypothesis.

### **3.5.3 Detecting outliers**

Inter quartile range is the method to identify an outlier. Calculating outliers is very

important because descriptive statistics methods such as mean, correlation coefficient,

and the standard deviation is sensitive to outliers.

### **3.5.4 Visualization of Data and Interpretations**

The data will be visualized using charts and various methods. So the users of this study

can get a good idea along with the interpretation. Matplotlib, seaborn, and other packages

will be used to visualize this study.

# **Chapter 4: Data Analysis**

## **4.1 Importing the Dataset**

First, The Pandas library and the dataset have been imported into the the jupyter notebook.

Jupyter notebook is a platform that is specially used for data analytics, Machine Learning, Deep

Learning and python and for other data science related things. You can see a picture of the

dataset below in Figure 1.

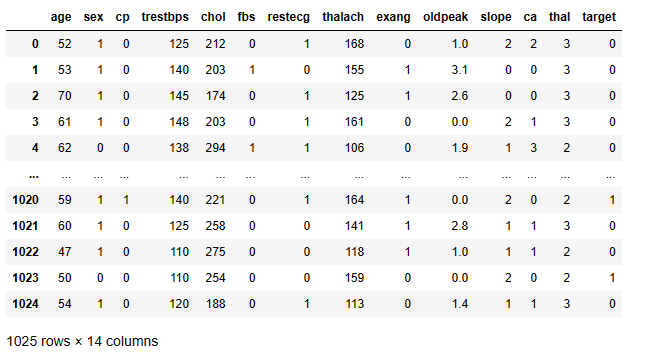


Figure 1 importing the dataset

Then we can see the first 10 rows and the last 10 rows of the dataset from the following Figure 2

and Figure 3

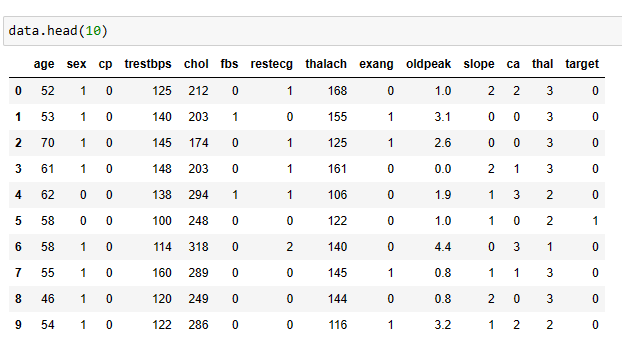


Figure 2 head of the dataset.

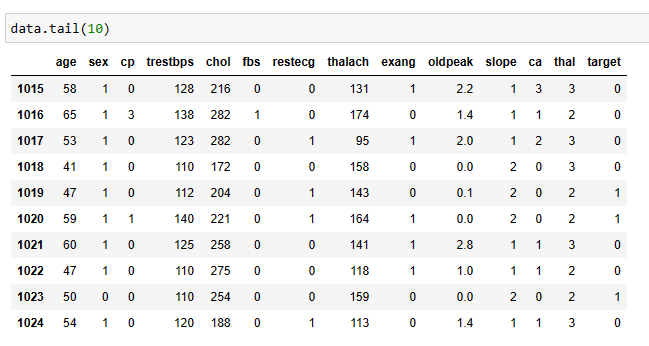


Figure 3 tail of the dataset

The information of the dataset variables can be seen in the following Figure 4.

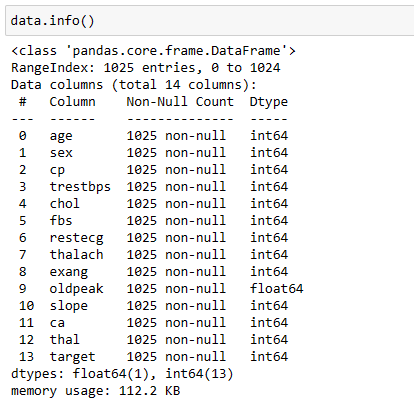


Figure 4 information of the dataset

## **4.2 Data preprocessing**

Data preprocessing is an important step in data analysis because that involves preparing

the data for modeling or analysis. If the dataset is not properly preprocessed then the predictions

will give errors. In this research, preprocessing is used to ensure that the data is consistent,

complete, and accurate and that is in a format suitable for analysis.

First thing is to find if there are any missing values in this dataset.

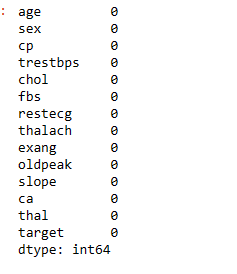


Figure 5 finding missing values of the dataset.

According to the Figure 5 we can see that there are no missing values in this dataset. So there is

no need to drop or fill missing values. Then I have created a boxplot for detecting outliers in the

numerical variables.

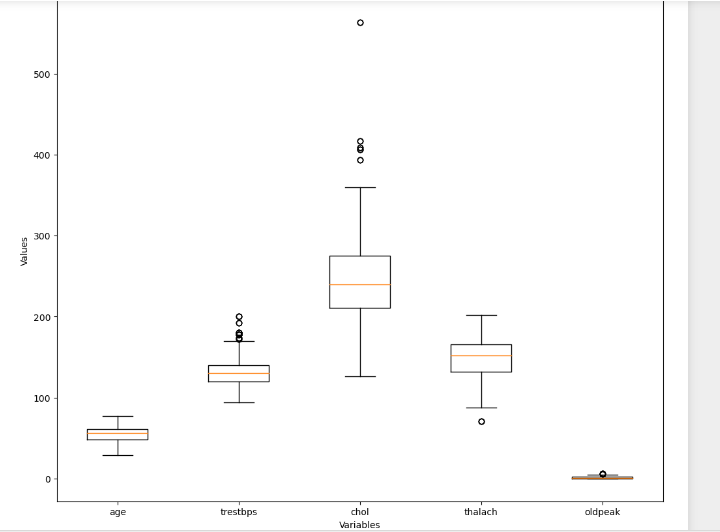


Figure 6 finding outliers of the datset

According to the Figure 6 we can see that there are some outliers in the numerical variables of

the dataset. Then I have removed the outliers of the dataset. That can be seen in the

following Figure 7.

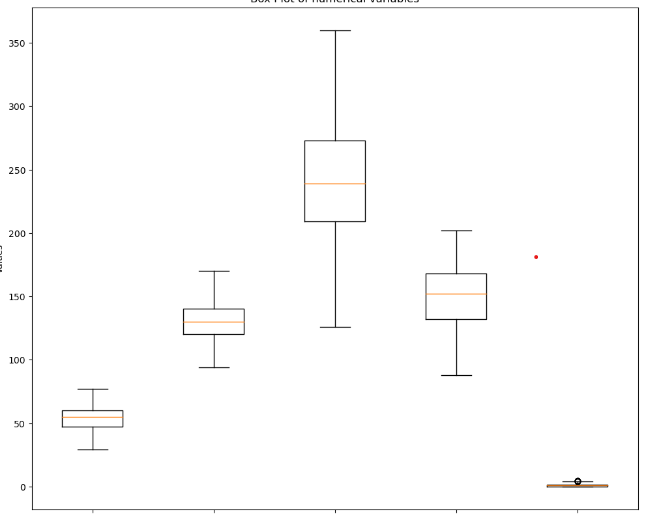


Figure 7 after removing outliers of the dataset

Then I have renamed the outliers removed columns to new names that can be seen in the below

Figure 8.

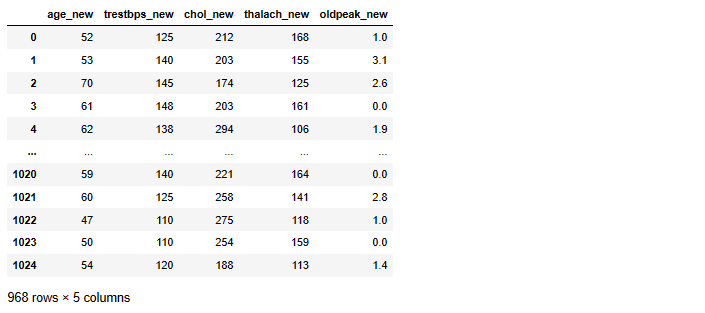


Figure 8 renamed the outliers removed columns

After that I have joined the outliers removed dataset with the original dataset by using the

‘concat’ function in the pandas.

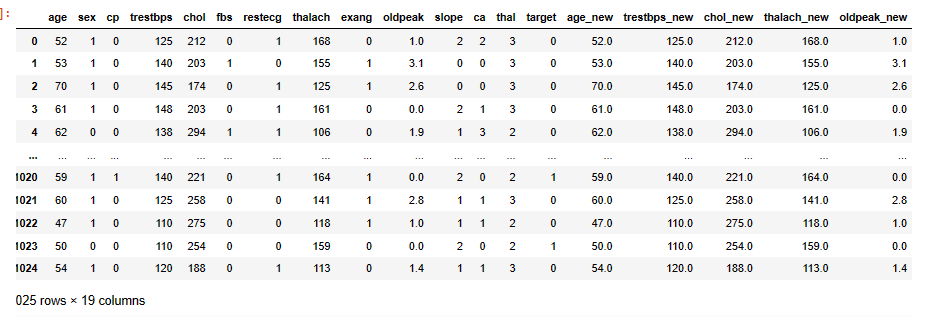


Figure 9 concatenated dataset

Now the dataset have duplicate columns. So we have to remove the numerical variables with

outliers. So I have droped the columns that are having outliers. Then the dataset can be seen as

according to the following Figure 10.

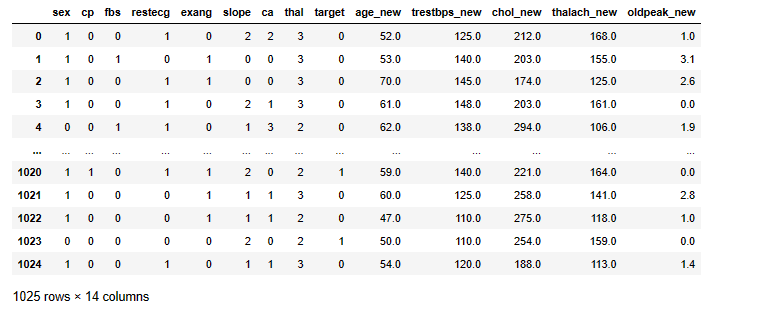


Figure 10 after droping the columns with outliers

After that I have checked whether there are null values in the new dataset because there can be

missing values.

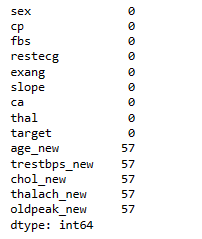


Figure 11 finding missing values of the new dataset

According the above Figure 11 we can see that there are now missing values in the dataset. That

is because I have removed outliers in the numerical variables and the I have concatenated those

variables with the original dataset. Now we have to drop the null values in this dataset.

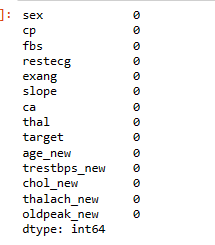


Figure 12 after removing null values in the dataset

According to the above Figure 12 we can see that the null values from the dataset have been

removed. Then I have reordered the columns as to the order of the previous dataset and I have

renamed the numerical variables also according to the previous dataset and changed the data

types of them to previous dataset. That can be seen in the following Figures.

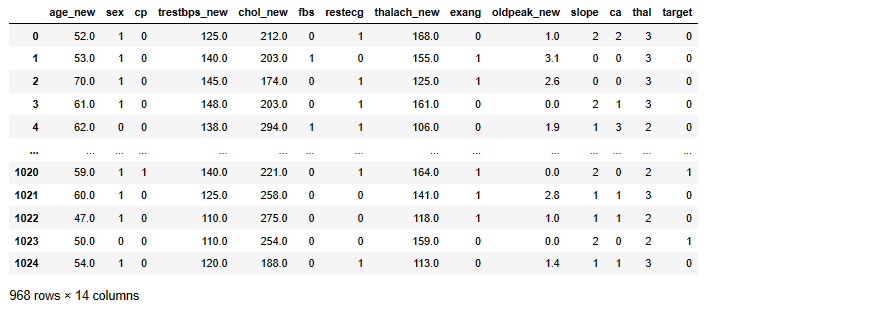


Figure 13 after removing null values from the dataset

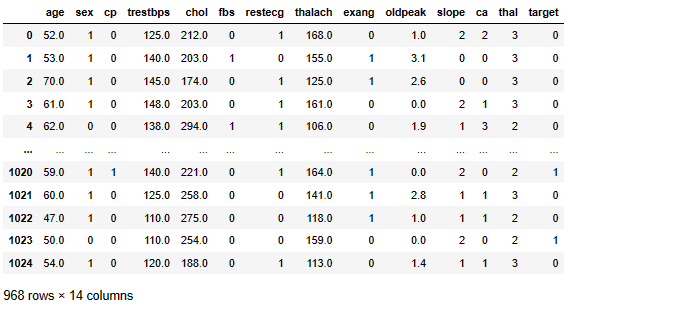


Figure 14 after renaming and reordering the dataset

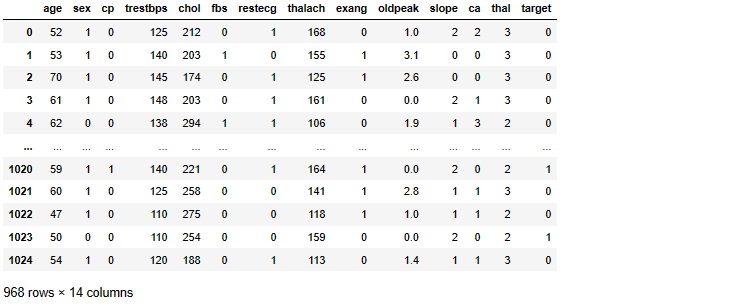


Figure 15 after converting the data types as before dataset

After doing the data preprocessing part I have saved the new dataset in a csv format so that can

be easy for the future use. After that I have imported the preprocessed dataset into the jupyter

notebook.

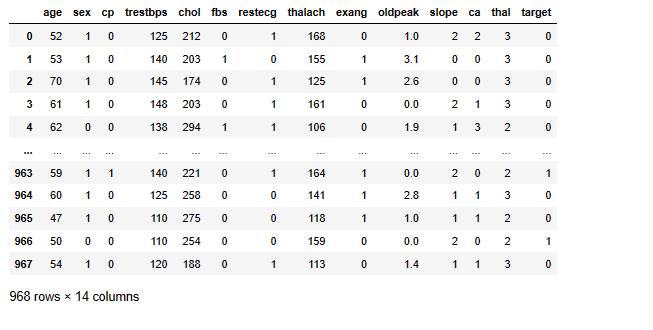


Figure 16 Importing the preprocessed dataset

## **4.3 Exploratary Data Analysis (EDA)**

In this section I have done a summary of the dataset variables, univariate analysis, bivariate

analysis, and statistical analysis. In the following figure shows the summary of the dataset

variables.

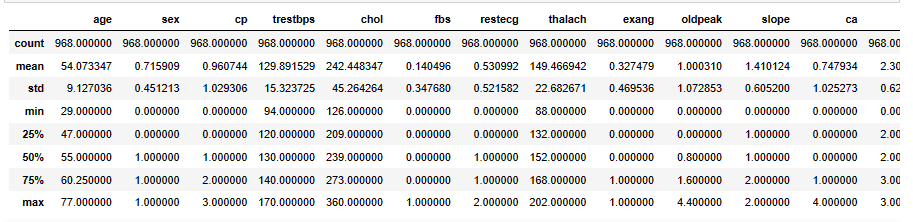


Figure 17 summary of the dataset

According to the above Figure 17 we can see the mean, median, max value and other things

related to the variables.

### **4.3.1 Univariate Analysis**

#### **4.3.1.1 Numerical variables**

In this study univariate analysis have been done for the numerical variables. Histograms have

been used for this analysis.

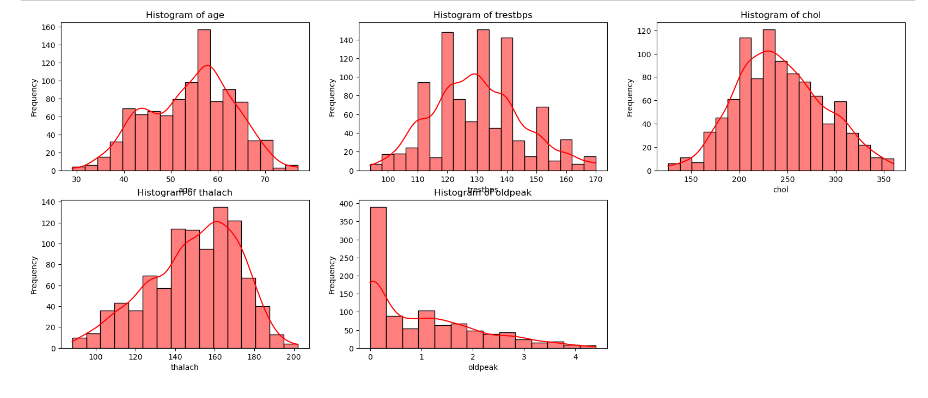


Figure 18 histograms of the numerical variables

According to the above Figure 18 shows the distribution of the numerical variables of the

dataset. The histogram of the age shows that a lot of people are in the age of 55-58 and a small

distribution after the age of 70. According to the trestbps histogram most of the people have a

trestbps level of 130. A small distribution for the trestbps level of below 100. According to the

chol histogram most of the people are in the cholesterol level of 200-250. After the 250 chol

level the distribution is going down. Most of the people having heart rate of 160-170 according

to the thalach histogram chart. A lot of people are in oldpeak=0 level and that is a good sign and

that can be seen in the oldpeak histogram.

#### **4.3.1.2 Categorical variables**

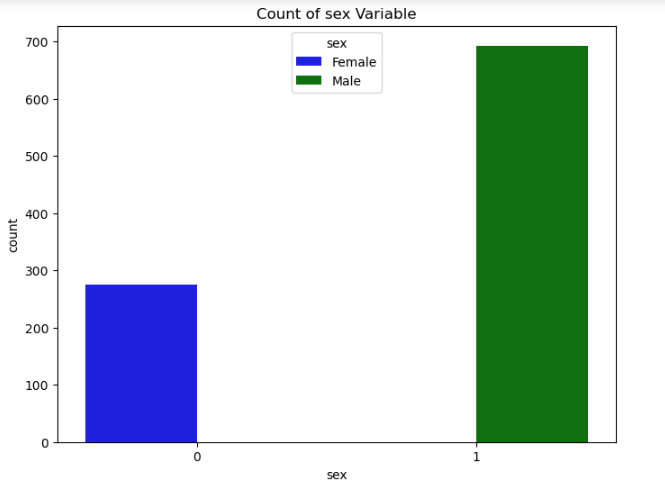


Figure 19 barchart of sex variable

The above Figure 19 shows the distribution of the sex variable. Here 0 represents female and 1

represents 1. So we can see there are lot of males in this dataset according to the above Figure

19.

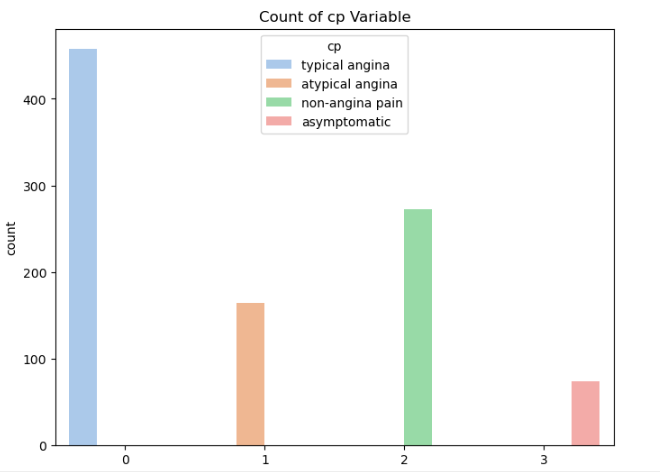


Figure 20 barchart of cp variable

The above Figure 20 shows the distribution of the cp variable. Most of the the people are in the

cp=0 category and less number of people are in the cp=3 category.

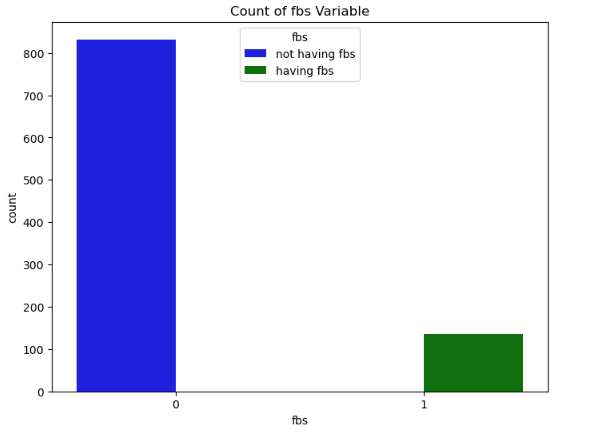


Figure 21 barchart of fbs variable

The above Figure 21 shows that most of the people in this dataset don’t have fasting blood sugar.

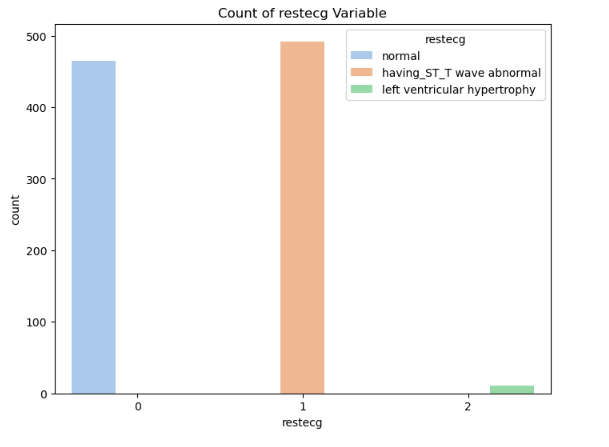


Figure 22 barchart of restecg variable

According to the above Figure 22 shows that there are a small number of people have left

ventricular hypertrophy.

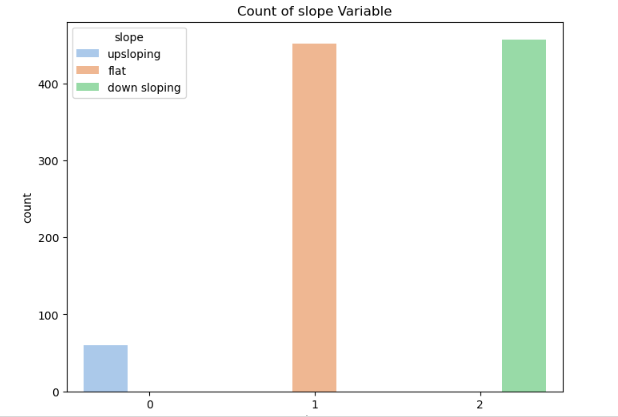


Figure 23 barchart of slope variable

Slope=1 and slope=2 are having a similar distribution and slope=0 is a small distribution

according to others. That can be seen in the above Figure 23.

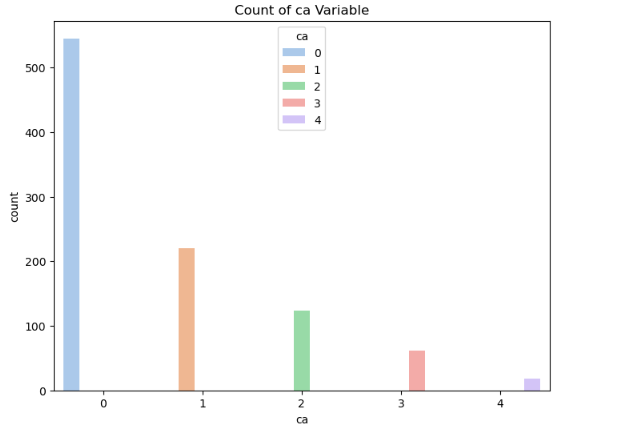


Figure 24 barchart of ca variable

This Figure 24 shows the the distribution of the ca variable. This is about the number of major

vessels colored by fluoroscopy. Ca =0 has the highest distribution and the ca=4 has the lowest

distribution.

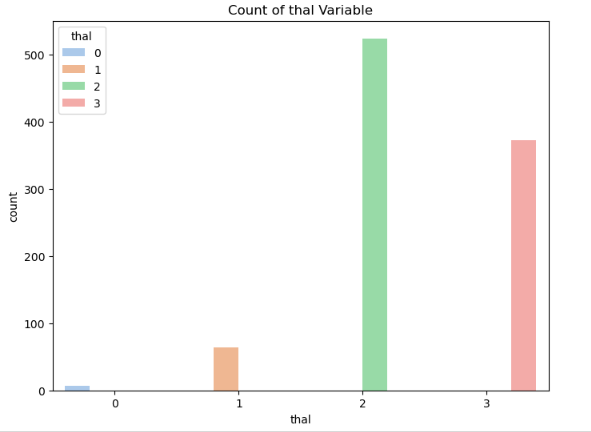


Figure 25 barchart of thal variable

According to the above Figure 25 shows the distribution of the thal variable. Thal=0 means

normal thal=2 means reversable defect and that have the highest distribution.

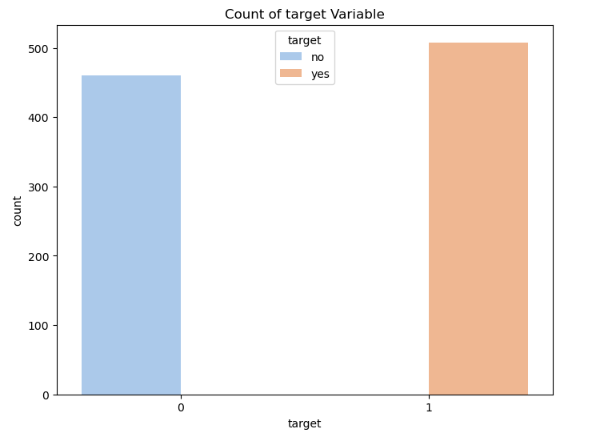


Figure 26 barchart of target variable

This above Figure 26 shows that there are more people having heart disease.

### **4.3.2 Bivariate Analysis**

In this part I have done analysis of numerical vs numerical variables and categorical vs

categorical variables and numerical vs categorical variables.

#### **4.3.2.1 Numerical variables vs Numerical variables**

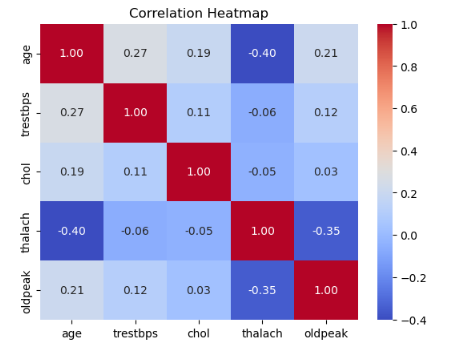


Figure 27 heatmap of numerical variables

The above Figure 27 shows the correlation plot of the numerical variables. There is a higher

correlation between the age and the thalach variable and that is a negative relationship and lower

correlation between the oldpeak and chol variables and that is a positive relationship. Thalach

variable have negative relationship with chol variable and the trestbps variable. Other variables

are having positive relationships.

#### **4.3.2.2 Numerical variables vs Categorical variables**

In this part I have used bar charts to plot the numerical and categorical variables.

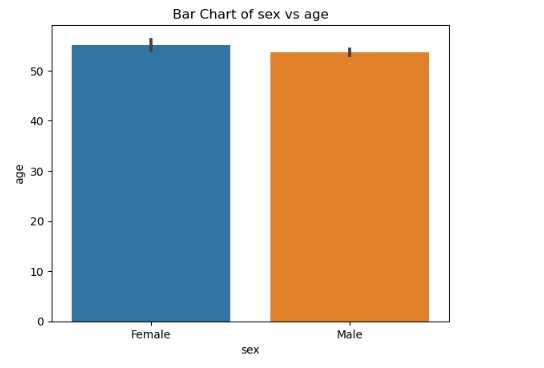
****

Figure 28 barchart of sex vs age

The above Figure 28 shows the bar chart of the sex variable with the age variable. In this dataset

there are more females than the males and that can be seen above.

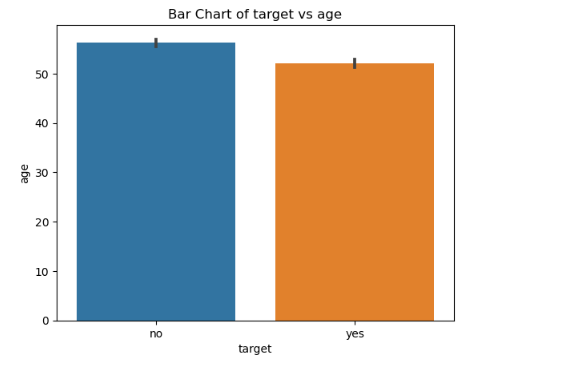


Figure 29 barchart of target vs age

The above Figure 29 shows the distribution of the age and the target variable.

#### **4.3.2.3 Categorical variables vs Categorical variables**

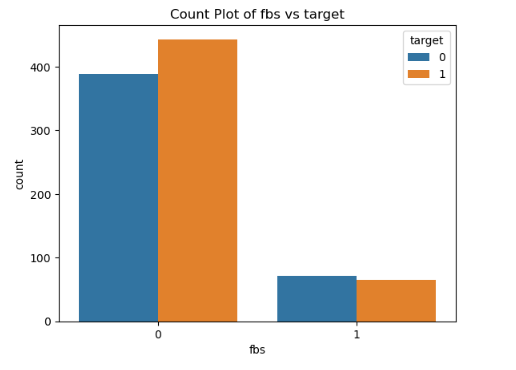


Figure 30 barchart of fbs vs target

The above Figure 30 shows that the people not having fbs is high and having fbs is very low. The

people having fbs and having heart disease is low than the people with fbs and not having heart

disease. So it seems like fbs is not seriously affecting the heart disease.

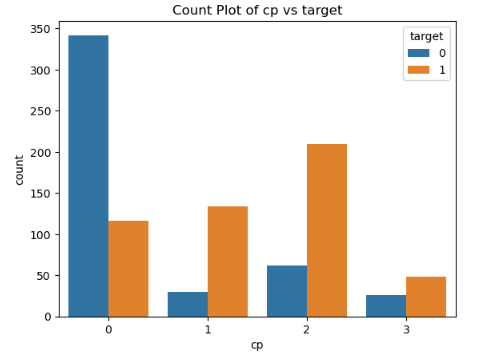


Figure 31 barchart of cp vs target

The above Figure 31 shows how cp variable affects the target variable. When cp=0, a lot of

people don’t have heart disease. When cp=2, a lot of people are having heart disease. Cp=2

means non -angina pain so that is not a good sign for the people because that is a risk. When

cp=1, also lot of people are having heart disease. So cp=1,cp=2,cp=3 are risky for the people

and they can be affected by heart disease.

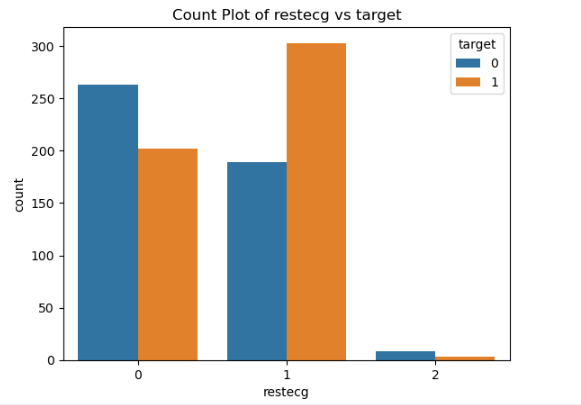


Figure 32 barchart of restecg vs target

This Figure 32 shows that restecg=1 are having a risk of heart disease. restecg=1 means having

ST\_T wave abnormal. So people with restecg=1 have to take care of them. Restecg=2 have a

small distribution in this dataset. In restecg=0 not having heart disease people are higher than the

people having heart disease.

### **4.3.3 Statistical Analysis**

In this study, I have used t-test to identify the relationship between numerical variables and

anova tests to identify the relationships between categorical and numerical variables and chi

-squared tests to identify the relationships between categorical and categorical.

I have used t-tests for age and thalach, age and chol, age and trestbps, chol and trestbps. The

results of the t-tests says that there is a relationship between those variables. The anova tests have

been used for age and fbs, age and target, trestbps and target, chol and target. All these variable

are having relationships with each other according to the anova tests. The chi-squared tests have

been used for exang and ca, sex and target, cp and target, fbs and target, restecg and target, exang

and target, slope and target, ca and target, thal and target. So theare are relationships between the

variables except the fbs and target because there is no relationship between them. So I have

droped the fbs column because that is not affecting to heart disease. After that I have saved the

dataset to a csv file for the training process.

# **Chapter 5: Training and Testing the Model**

First, the saved dataset has been imported to a new jupyter notebook. That can be seen in the

following Figure 33.

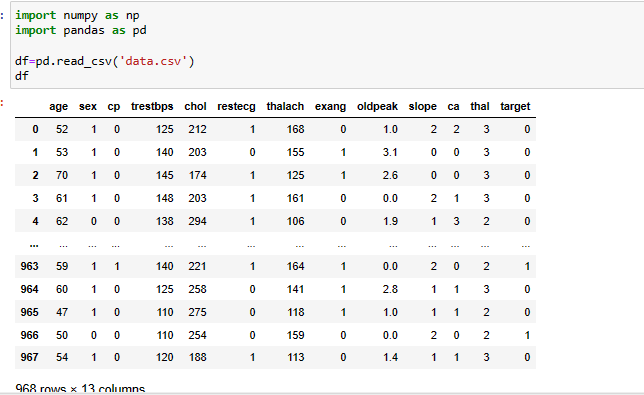
****

Figure 33 importing the dataset to training

Then the dataset has been converted into a numpy array by using the following code in the

Figure 34.

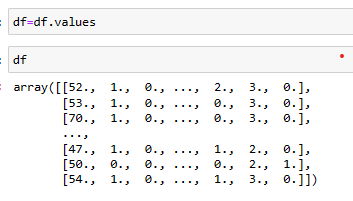


Figure 34 converted the dataset into numpy array

After that the dataset has been divided into data and target to separate the features and labels.

That can be seen in the following Figure 35.

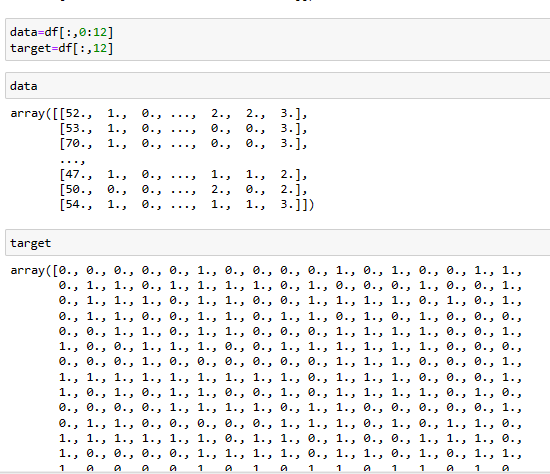


Figure 35 dividing the dataset into data and target

After that normalizing the data and the target have been done using the MinMaxScaler function

in the sklearn library. Scaling the data is very important because it will give faster convergence in

the training. Also it will improve the performance of the model. The normalization part can be

seen in the following Figure 36.

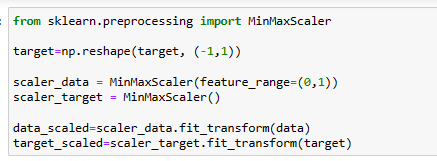


Figure 36 normalizing the data and target

After the normalization part the the scaled data and the scaled target have been splited into

training and testing using the train\_test\_split function from the sklearn library. I have used 90%

for training and 10% for testing. That can be seen in the following Figure 37.

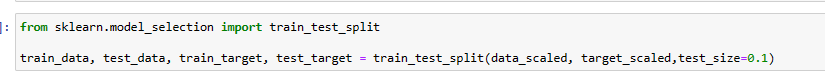


Figure 37 dividing the dataset into training and testing

After this the ANN model has been created. The following figure is about the ANN model.

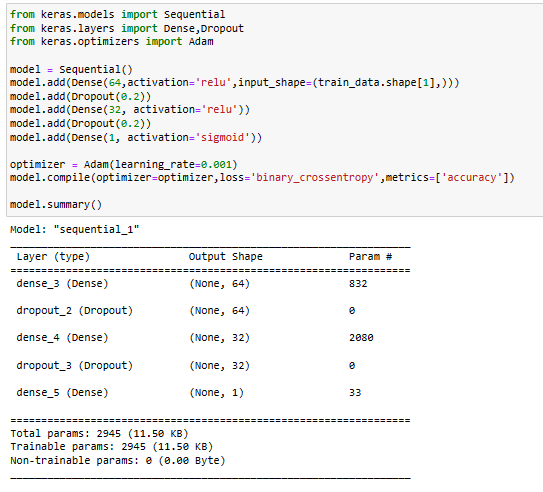


Figure 38 creating the model

According to the above Figure 38 I have used Dense layers, Dropout layers and relu activation

functions, sigmoid activation function. For the learning rate I have used Adam optimizer and the

staring learning rate has been scheduled to 0.001. The binary\_crossentropy loss function has

been used because this is a binary classification problem. The output of the code can be seen

according to the above figure. After that I have trained the model and saved the best model that

can be viewed from the following Figure 39.

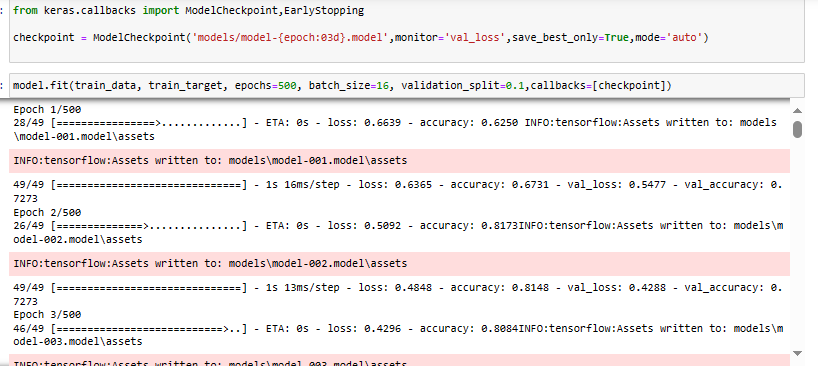


Figure 39 training the model

In the training I have used batch\_size of 16 and 500 epochs and validation\_split of 0.1.

Then the training loss and the validation loss have been plotted.

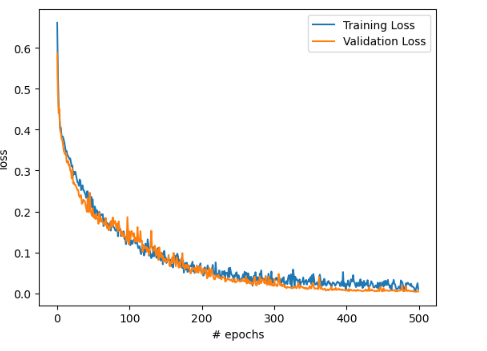


Figure 40 loss functions of the model

According to this Figure 40 the model is giving extremely good results on the training and the

validation. So it is very important when predicting to the unseen data and this model is very good

at it because the validation loss curve is excellent in this case.

After that the accuracy curves have been plotted.

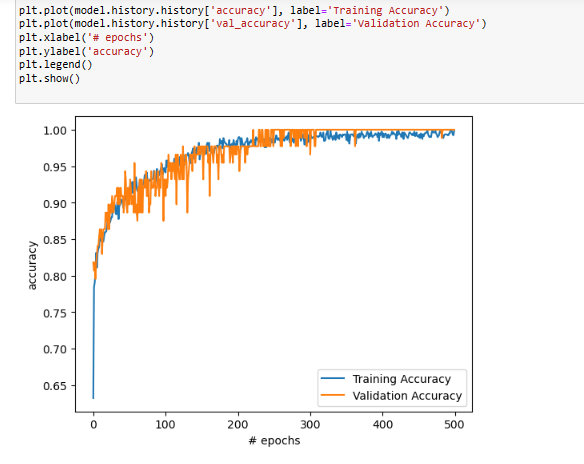


Figure 41 accuracy functions of the model

According to the above Figure 41 shows that this is giving good results for the training and

validation by looking at the accuracy. The validation accuracy has been converged and I think

that will give extreamly good results for the unseen data.

The following Figure 42 shows predicting the target using test data.

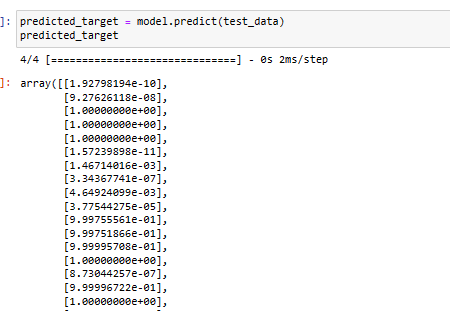


Figure 42 predicting using test data

After this the accuracy of the model has been found and that is very excellent. That can be seen

in the following Figure 43.

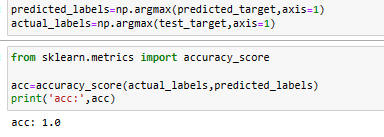


Figure 43 accuracy of the model

After that I have tested whether the model is giving correct results or what by the following

Figure 44.

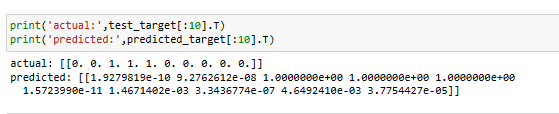


Figure 44 testing the model

After training the model then I have saved the scaled data and scaled target into dump files from

the joblib library and I have saved the model too. Those things can be viewed in the following

Figure 45.

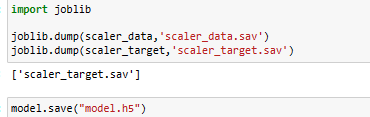


Figure 45 saving the model and scaled data and scaled target

Then I have imported the model into another jupyter notebook and then tested the model using

some random data. The following Figure 46 shows the code and the output of that.



Figure 46 testing the model with random data

# **Chapter 6: Creating the web application**

In addition to developing a heart disease detection model using Artificial Neural Networks

(ANN), a web application was created using Flask to facilitate the interaction with the model.

This web application comprises two distinct sections represented by separate HTML templates:

one for inputting patient details and another for displaying the results of the heart disease

prediction. The first template allows users to input relevant patient information, such as age,

gender, cholesterol levels, and blood pressure, which are essential features for predicting heart

disease risk. Upon submission, the input data is passed through the trained ANN model for

prediction. The second template serves as the output interface, presenting the prediction results

obtained from the model. This approach enhances accessibility and usability by providing a user

-friendly interface for both inputting patient details and visualizing the corresponding heart

disease prediction outcomes. Integrating these web application functionalities into the report

serves to showcase the practical implementation and usability of the heart disease detection

model developed in this project, emphasizing its potential real-world applications and impact.

The following Figure 47 and 48 will show the interface and the results of the web application.

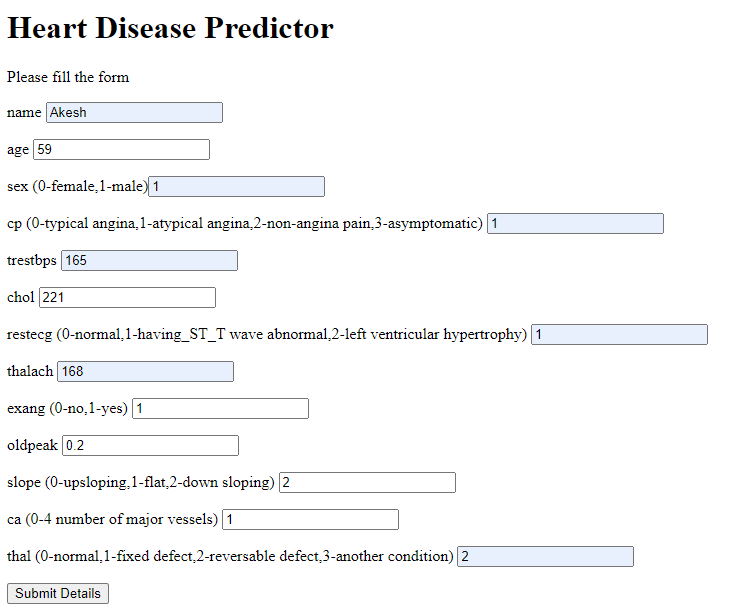


Figure 47 picture of the web application

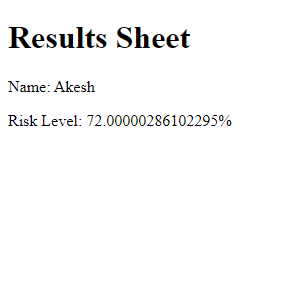


Figure 48 picture of results of the web app

# **Chapter 7: Conclusion, Discussion and Recommendation**

## **7.1 Conclusion**

In conclusion, this project has successfully completed a comprehensive analysis of heart disease

prediction using a dataset obtained and preprocessed in Jupyter Notebook. The initial steps

involved importing the dataset followed by preprocessing to ensure data quality and consistency.

Subsequently, an Exploratory Data Analysis (EDA) was conducted to gain insights into the

dataset's characteristics, including summarizing variables, analyzing relationships, and

visualizing data distributions. The statistical analysis performed during EDA further enhanced

our understanding of the dataset and identified potential risk factors associated with heart

disease. Building on these insights, an Artificial Neural Network (ANN) model was trained using

Deep Learning techniques to predict heart disease based on input variables. The model's

performance was evaluated, demonstrating its capability to accurately predict heart disease

outcomes. Furthermore, the development of a web application utilizing this trained model

enhances accessibility and usability, allowing for real-time predictions based on individual input.

Overall, this project not only showcases the technical proficiency in data analysis and model

training but also highlights the practical application of these findings in healthcare decision

-making and risk assessment for heart disease prevention.

## **7.2 Discussion**

In this study, we have successfully achieved our main objective of detecting heart disease by

developing an Artificial Neural Network (ANN) model that accurately predicts heart disease

based on several key risk factors. Through comprehensive data analysis, we identified significant

risk factors such as age, sex, chest pain type (cp), cholesterol levels (chol), resting blood pressure

(trestbps), and resting electrocardiographic results (restecg). Notably, we found that fasting blood

sugar (fbs) did not exhibit a relationship with heart disease and was consequently dropped from

our analysis. Leveraging statistical analysis tests, including ANOVA and chi-squared tests, we

elucidated the relationships between various variables and heart disease. Specifically, using a

heatmap, we visualized the correlations between numerical variables, revealing notable

associations between age, trestbps, chol, and restecg with the target variable. Further, ANOVA

tests confirmed significant relationships between age, trestbps, and chol with heart disease, while

chi-squared tests revealed associations between categorical variables such as sex, chest pain type,

and exercise-induced angina (exang) with the target variable. The accuracy of the ANN model is

1 and that is very good in making predictions. Moreover, we successfully implemented a web

application to facilitate real-time heart disease predictions based on our ANN model. Overall, our

findings contribute to a deeper understanding of heart disease risk factors and provide valuable

insights for personalized prevention and treatment strategies.Top of Form

## **7.3 Recommendation**

Based on the findings of this project, it is recommended to further explore and validate the

identified risk factors associated with the early diagnosis of heart disease through larger-scale

studies and diverse populations. Additionally, incorporating advanced machine learning

techniques and incorporating more comprehensive datasets may enhance the accuracy and

generalizability of heart disease prediction models. Furthermore, it is crucial to prioritize public

health initiatives aimed at promoting lifestyle modifications, such as healthy diet and regular

physical activity, to reduce the burden of heart disease. Collaborative efforts between healthcare

professionals, researchers, policymakers, and community stakeholders are essential to implement

effective prevention and treatment strategies tailored to individual risk profiles. Moreover,

continued research into the interplay between genetic predisposition, environmental factors, and

lifestyle behaviors is warranted to inform the development of personalized interventions for heart

disease prevention. Overall, leveraging multidisciplinary approaches and harnessing emerging

technologies hold promise in addressing the complex challenges associated with heart disease

and improving cardiovascular health outcomes on a global scale.

# **Appendices**

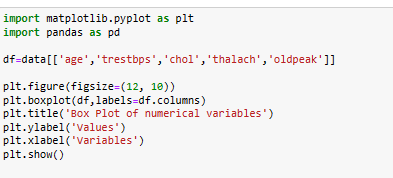
****

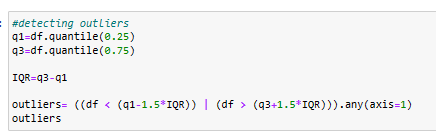
****

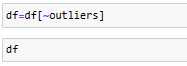
****

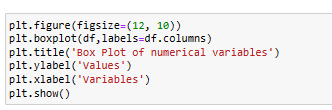
****

****

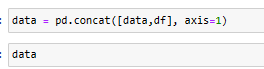
****

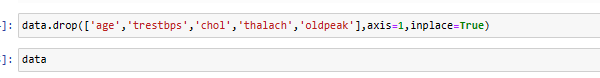
****

****

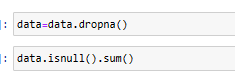
****

****

****

****

****

****

****

****

****

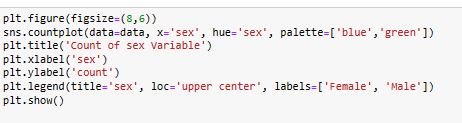
****

****

****

****

****

****

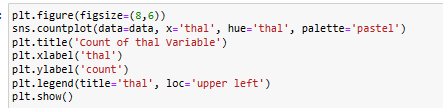
****

****

****

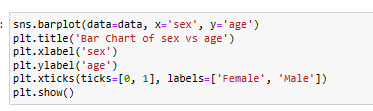
****

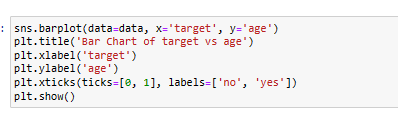
****

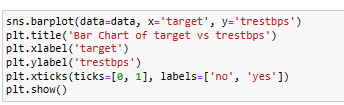
****

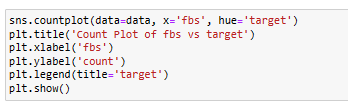
****

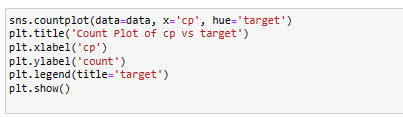
****

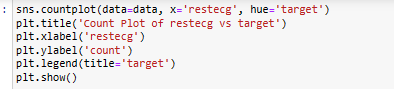
****

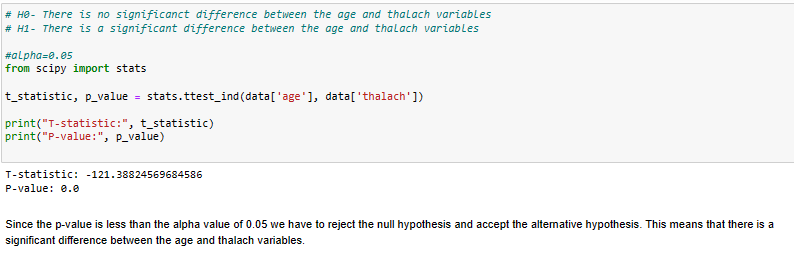
****

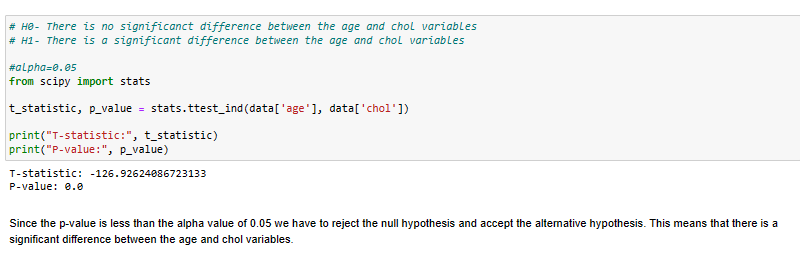
****

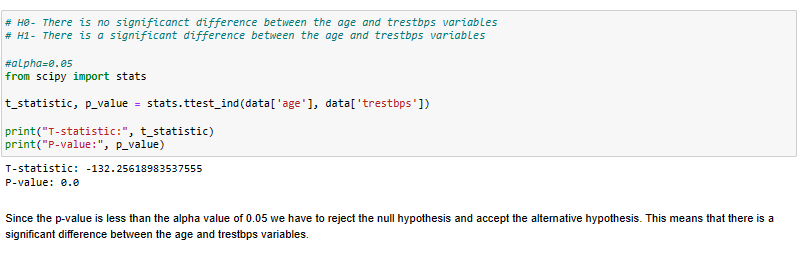
****

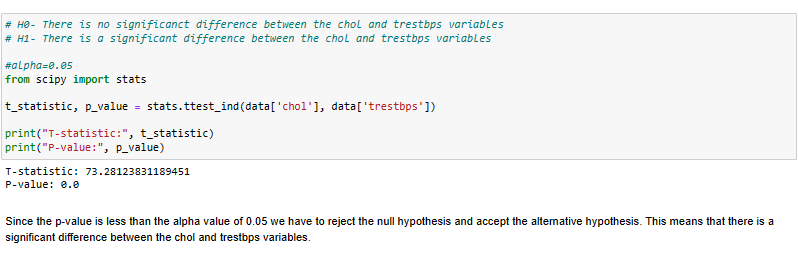
****

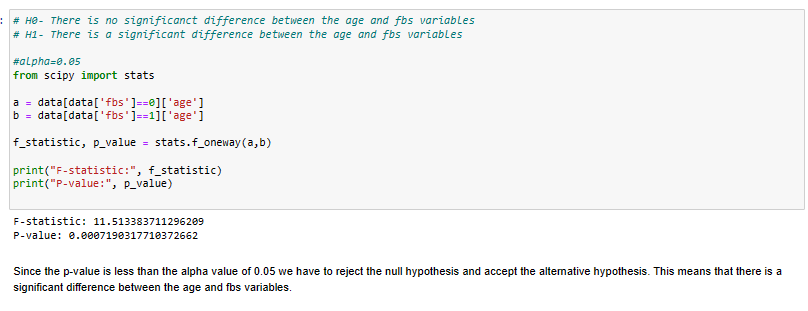
****

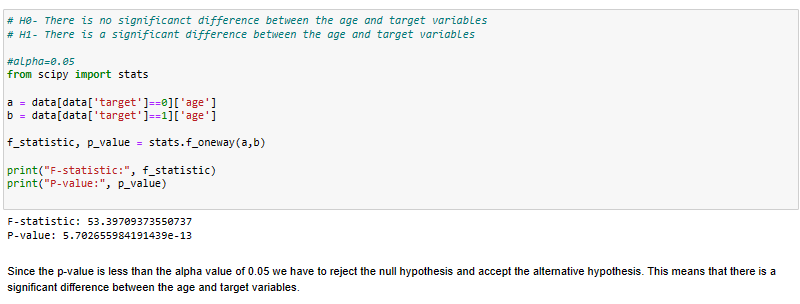
****

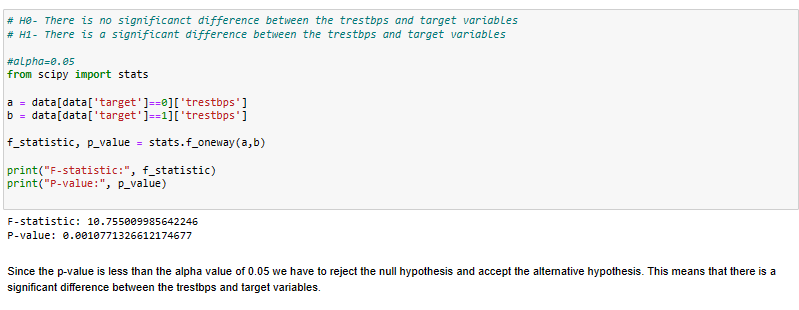
****

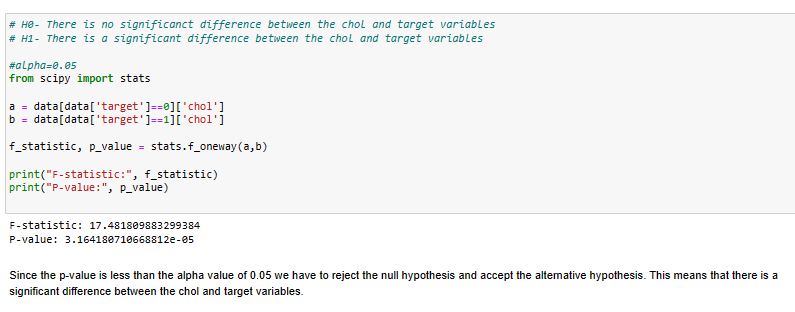
****

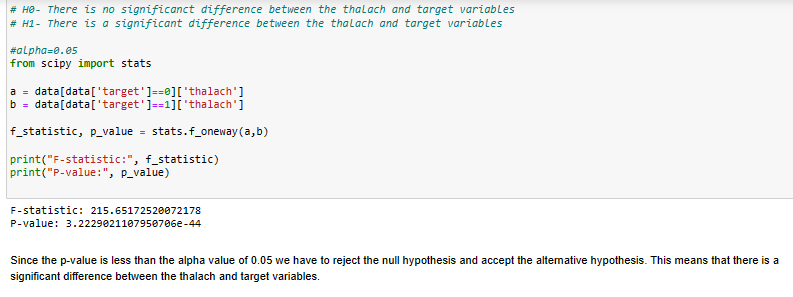
****

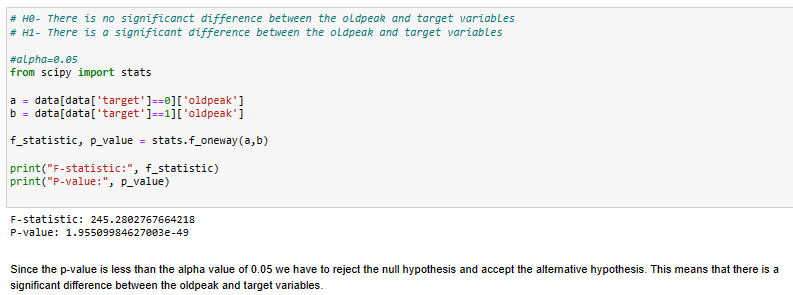
****

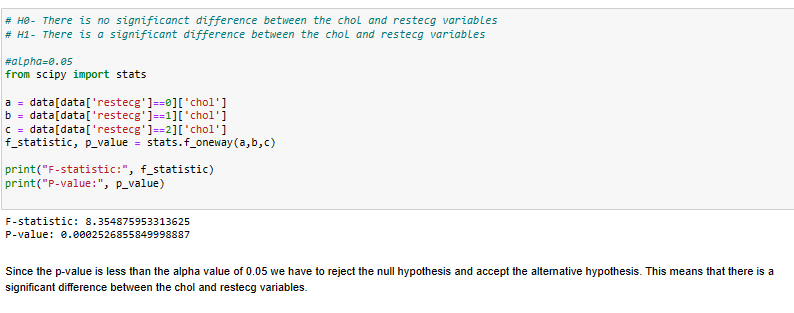
****

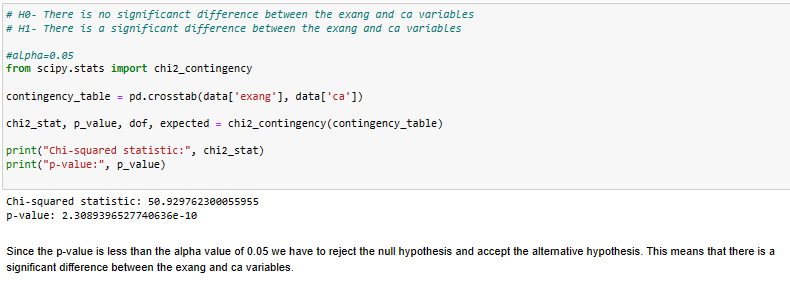
****

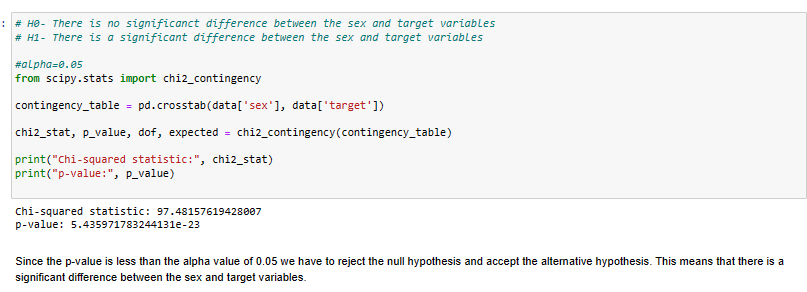
****

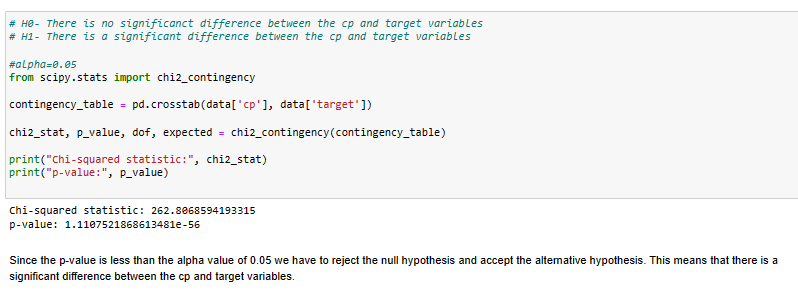
****

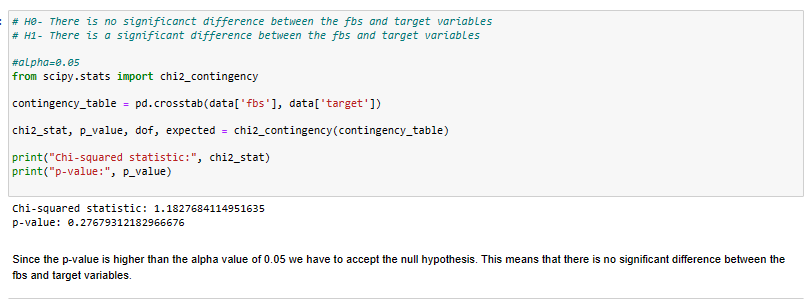
****

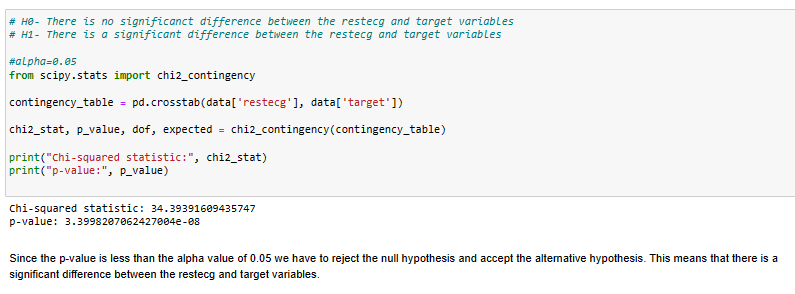
****

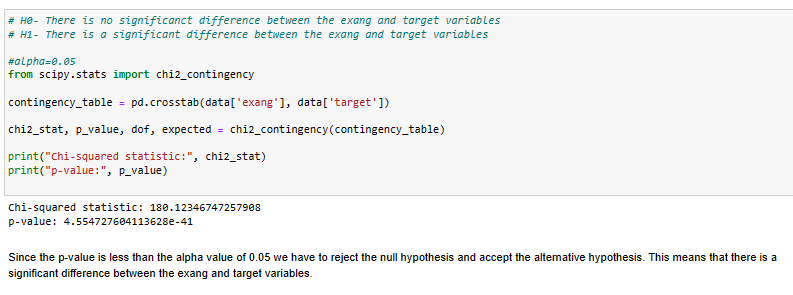
****

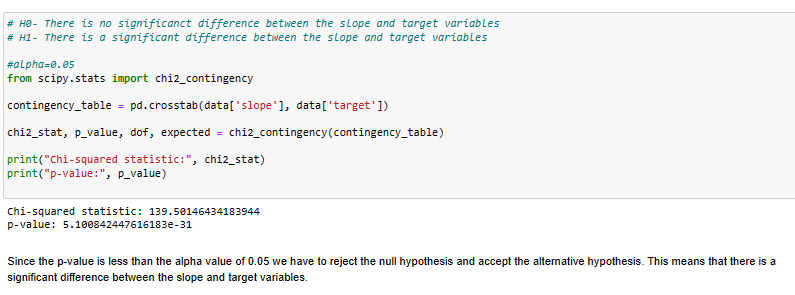
****

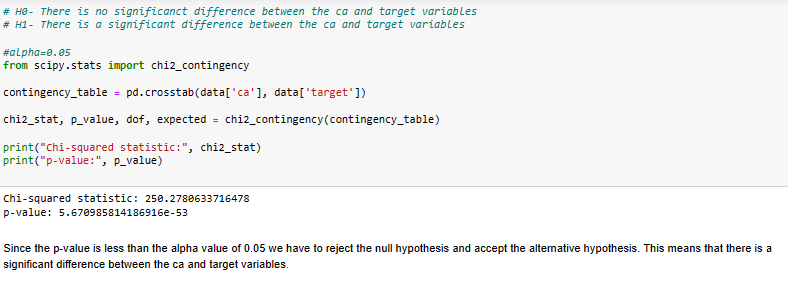
****

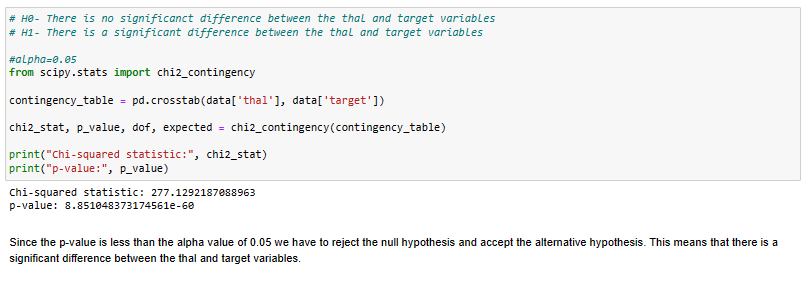
****

****

****

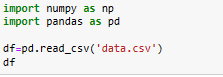
****

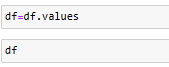
****

****

****

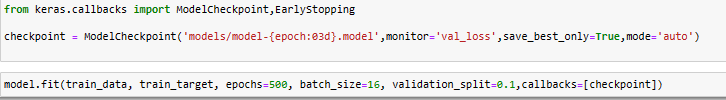
****

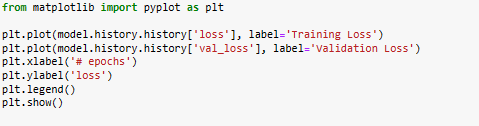
****

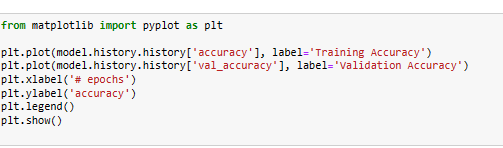
****

****

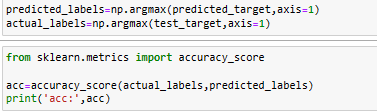


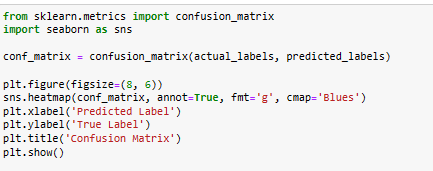




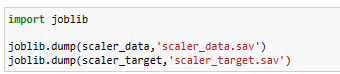


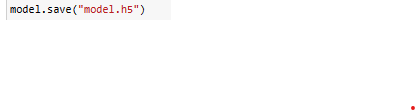


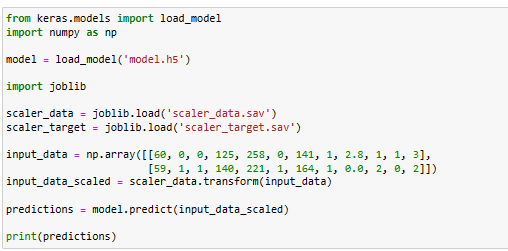
****

****

****

****

****

****

# **List of references**

<https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset>

<https://www.cdc.gov/nchs/data/hus/2017/019.pdf>

<https://www.ijedr.org/papers/IJEDR1704226.pdf>

<https://www.sciencedirect.com/science/article/pii/S187705091630638X>

<http://noiselab.ucsd.edu/ECE228-2020/projects/Report/72Report.pdf>

<https://www.researchgate.net/publication/364949647_Enhanced_accuracy_for_heart_disease_prediction_using_artificial_neural_network>

<https://iopscience.iop.org/article/10.1088/1742-6596/1997/1/012022>