A ES
ISSN: 2096-3246
Volume 54, Issue 02, December, 2022
Special Issue

# DIAGNOSING CARDIOVASCULAR DISEASE: AN ASSOCIATIVE CLASSIFICATION USING MACHINE LEARNING APPROACH

**M.T. Beig**
SGT University, Gurgram, India
mirzatanweer@gmail.com

**Nitesh**
SGT University, Gurgram, India

**Keshav Yadav**
SGT University, Gurgram, India

**Rajiv**
SGT University, Gurgram, India

**Ramesh Kumar Pandey**
Babu Banarasi Das University, Lucknow (U.P.) India

**Abstract**
In the case of the human body, the central aspect of metabolism is the demand for oxygenated blood to complete the process. Wastes and deoxygenated blood must be pushed out of the many organs in the body due to metabolism for the human body to continue functioning. As the body's pumping organ, the heart pumps oxygenated blood to all body regions and eliminates wastes and deoxygenated blood. As a result, it is critical to have frequent check-ups to ensure optimal cardiac care. There are a variety of causes for this, including genetic background and some acquired habits that can hurt the heart. Several research studies have been published investigating human heart health prediction. In this study, we looked at 304 patient cases and attempted to identify the significant risk factors that could cause heart difficulties. This study aims to present a work that can be used as a first step in determining a probability score for a cardiac condition. The various risk factors can be the primary cause of a heart problem. This study looked at different categorization models to pinpoint cardiac conditions. The fundamental goal of such an activity is to give a straightforward solution that can assist the patient in determining whether there is a statistical chance that a heart problem will occur. This solution is not intended to replace a medical practitioner but to assist any doctor in their diagnosis process. This procedure ensures that the therapy between a doctor and a patient is transparent. The number of False Positives and False Negatives produced by the model is verified, and the optimum algorithm for prediction is chosen based on that.

**KEYWORDS:** Heart Disease, Machine learning (ML), Artificial Intelligence (AI), Data processing, Algorithm

### 1.1. Introduction

The human body's sustenance requires the circulation of vital nutrients that allow it to function properly and live. The circulation of blood, which delivers necessary nutrients for survival, is one of the most critical parts of the human system. The heart, the pumping system that causes blood to flow to each area of the human body, is thus the most crucial organ. As a result, monitoring the heart has become a vital aspect of monitoring the human body's state. The data is one of the most significant components in performing a proper heart diagnosis. The researchers are conducting a considerable study in this area to use machine learning to forecast the status of the heart. The following are the primary objectives that we will be addressing to analyse this research:

1. To discuss the various machine learning algorithms to identify the optimal method for predicting cardiac problems with a greater probability and fewer false predictions.

2. To identify key risk factors or elements that contribute to heart disease and cardiac problems.

3. To monitor the heart's condition and predict heart failure. Heart failure, caused by various factors affecting the heart, is one of the leading causes of natural death worldwide. Some of the reasons may be based on the person's family genetic background, while others may be found in the various elements he considers in his life. This requires a frequent check-up of the heart to ensure its health.

In the human heart, there are four chambers in total, each with its function. The heart's principal purpose is to transport oxygenated blood to various body regions, aiding metabolism, and then return deoxygenated blood to the heart. The primary goal of this project is to propose a solution that allows a doctor to check for a heart condition in a patient with certainty. The human body is exposed to high-frequency radiation during coronary tomography, which is used to determine whether or not there is a heart disease [8][9]. The impact of this radiation on an average human body will be negative. In addition, the expense of coronary tomography is highly significant. This study will consider various risk factors such as Family History, Fasting Glucose, Smoking Habits, Hypertension, Dyslipidaemia, Obesity, Sedentary Lifestyle, CABG, and High Serum to model a solution that can predict whether or not a heart problem will occur. Aside from these risk variables, some of the characteristics of the patients, such as their demographic details, can be factored into the modelling. Such features as Gender, location, age, etc., can be considered. As the pumping organ, the heart is subjected to various blockages that can cause blood to flow improperly through the blood vessels. According to a survey conducted by the World Health Organization, 17.5 million people die each year. It is expected to reach 75 million by 2030. Medical practitioners specialising in cardiac disease have their own set of parameters and can predict a heart attack risk of up to 67 percent. With the emerging epidemic condition, clinicians require a support network to better precisely forecast cardiac disease. Using the Machine Algorithm and Deep learning opens up new possibilities for effectively predicting heart attacks. Journals provide a wealth of up-to-date knowledge on computing methods and deep learning in computers. An empirical comparison was provided to facilitate the conduct of a new study in this field.

### 1.2. Literature Review

Healthcare is one of the industries with a massive volume of data. As a result, one of the critical goals in data science is to use the features and hidden insights within the data to forecast the heart attack with a high degree of certainty. Some of the datasets utilised in one of the research domains are categorised in terms of medical parameters [1]. Decision Trees and Naive Bayes are the algorithms used to determine the level of accuracy in the case of cardiac problem prediction. In most research studies, Naive Bayes has been utilised as the dominant algorithm for heart attack prediction [2][3]. Many risk factors are examined in another study for a heart problem prediction system [4] to see if there is a correlation between individual features and the patient's heart condition. Then classification algorithms are examined to see which performs best in predicting the heart problem with the lowest false-negative rate. In this study, a new metric known as the selection value is selected and established, which considers the accuracy and the false-negative speed when determining the optimum method in this context. The top performers were the Random Forest Classifier and Support Vector Machines using Radial Basis Function as the kernel. A standard test, which comprises an ECG and a CT scan to determine whether or not the patient has a heart condition, is time-consuming and expensive [3]. Furthermore, if an average human body is subjected to high-frequency radioactive waves, the body may experience some negative impacts [4]. Ensemble approaches are utilised in another study paper [5] to improve heart disease risk prediction. This technique aims not only to enhance the accuracy of the model's prediction abilities but also to investigate if machine learning models can be used in a medical setting to make early predictions about the severity of human diseases. Using the Boosting and Bagging techniques, an increase in accuracy of 7% was achieved, which was further strengthened by introducing feature choices to show considerable improvement. Many risk factors, such as gender, age, cholesterol level, smoking, hypertension, eclampsia, and others, are employed in another study to evaluate risk using fuzzy logic, with decisions based on fuzzy weight rules.

Some non-invasive approaches have been used, and they are pretty effective in diagnosing heart failure and preventing it [6]. A comprehensive analysis of the numerous machine learning algorithms being used, as well as several feature reductions and extraction strategies that have been used in this particular research, is performed. Age, family history, diabetes, hypertension, high cholesterol, cigarette smoking, and other characteristics provide various notifications [7]. Because the preceding hazards cannot be used to forecast heart failure, conditional monitoring can be used to take preventive steps. The system as a whole predicted with an accuracy of 89%. Even neural networks were utilised in certain studies to forecast the severity of heart disease [7]. Endovascular aneurysm repair (EVAR) is a new minimally invasive surgical treatment that helps patients recover faster while reducing postoperative morbidity and mortality [10]. This study presents an ensemble approach for predicting postoperative morbidity following EVAR. The ensemble model was built using a training set of consecutive patients who underwent EVAR between 2000 and 2009. All data needed for modelling predictions were collected prospectively and recorded into a clinical database, including patient demographics, preoperative, comorbidity, and complication as outcome variables. A discretization

strategy was used to categorise numerical values into informative feature spaces. The Bayesian Network (BN), Artificial Neural Network (ANN), and Support Vector Machine (SVM) were used as basic models in this study, along with other stacking models.

Among the study's findings was an ensemble model for predicting postoperative morbidity following EVAR, the prevalence of prospectively reported postoperative problems, and BNs' understanding of causal effect with Markov's blanket principle. The work is a bit hazy in a study about the growth of Decision Support Systems for coronary artery diagnostic and evidential disorder [11]. Information on coronary artery disease (CAD). Sets from the University of California, Irvine (UCI) are included. This uses the Fuzzy Decision Support Program's information base. Using a tool based on Rough Set Theory to extract rules. The rules are then chosen and blended based on discrete numerical attribute knowledge. The details from the extracted help Regulations are intended to be used in fuzzy weight rules. Data sets were gathered from heart disease patients in the United States. The proposed programme will be tested in UCI, Switzerland, and Hungary, as well as at Ipoh Specialist Hospital in Malaysia. Evidence suggests the gadget can provide a higher percentage of coronary artery-blocking than cardiologists and angiography. Three cardiologists evaluated and validated the findings of the proposed program, and others were described as straightforward and useful. The training collection for the framework is made up of incomplete sets of CAD data. RST on ANN is used to infer this set of training. The assumed training collection is then constructed. There are 358 things in this training collection (patients). The usage of ROSETTA software to create RST rules yielded 3881 rules [12]. The proposed RST-based rule selection method only allows for the selection of 27 rules. The method of selection described in this study improves accuracy and coverage efficiency. Evidence that artificial intelligence (AI) is beneficial for detecting and administering risk factors for hypertension was found in a study aimed at determining the influence of hypertension on heart disease [13]. But I'm still a long way from being able to use cutting-edge AI approaches to detect hypertension risk factors and apply them to individualised treatment. This paper reviews recent advances in the fields of computer science and medicine, with a focus on ground-breaking AI strategies for the early identification of hypertension. I've also evaluated current research and potential implications of AI in hypertension care and clinical trials, focusing on customised medicine. Although recent studies suggest that AI is practical and potentially valuable in hypertension research, AI-assisted therapy has yet to change blood pressure regulation (BP). This is largely owing to a scarcity of data on the AI's consistency, precision, and dependability in the BP space. Nonetheless, various factors, including heredity, climate, and lifestyle, contribute to poorly controlled blood pressure. AI can help prescribers and patients understand how to extrapolate data analytics to alert them about specific issues that could affect their blood pressure control. To present, AI has mostly been used to investigate hypertension risk factors. Still, it has not yet been employed to control hypertension because of constraints in study design and physician engagement in computer science literature. Incorporating biological, behavioural, and environmental aspects into the decision-making process for successful medicine administration to monitor blood pressure, AI's future of more flexible architecture employing multi-o mics methodologies and wearable technology will surely be a

significant resource. Artificial intelligence and machine learning have the potential to impact practically every aspect of human life, including cardiology [14]. This paper guides clinicians on key features of artificial intelligence and machine learning evaluates chosen implementations of these approaches in cardiology to date and discusses how artificial intelligence might be integrated into cardiovascular medicine. The study covers predictive modelling techniques important to cardiology, such as feature selection and common problems like inappropriate dichotomization. Second, common supervised learning methods are explained, and selected applications in cardiology and related areas are analysed. Third, it explains how deep learning and related approaches, collectively known as unsupervised learning, came to be, provides context examples in general and cardiovascular medicine, and then demonstrates how these methods might be used to improve cardiac accuracy and patient outcomes.

AI approaches have been used in cardiovascular medicine to discover novel genotypes and phenotypes in established conditions, enhance patient care quality, enable cost-effectiveness, and reduce readmission and mortality rates [15]. For the diagnosis and prediction of cardiovascular illnesses, several machine-learning algorithms were applied. Each challenge necessitates awareness of the problem, both in terms of cardiovascular medicine and statistics, to use the best machine-learning method. AI may soon contribute to a paradigm shift toward cardiovascular precision medicine. In cardiovascular medicine, AI's potential is enormous; however, a lack of understanding of the problems can obscure its therapeutic implications. This study gives an overview of the use of artificial intelligence (AI) in cardiovascular clinical care and examines its potential role in precision cardiovascular medicine. For stroke prediction, a study compared multiple methodologies with the Cardiovascular Health Study (CHS) data set approach [16]. The decision tree approach is used for feature selection, the major component analysis algorithm is used to minimise dimension, and the neural network classification algorithm with backpropagation is used to build a classification model. After analysing and comparing classification efficiency with different approaches and variation models' accuracy, this work has the best predictive model for stroke disease with 97.7% accuracy. Patients were not required to undergo any diagnostic tests, and clinical diagnosis is primarily made through the skill of a doctor [17]. However, not all studies can result in a correct condition diagnosis. Selecting a subset of features is a preprocessing technique that reduces dimensionality and eliminates extraneous data. In this research, we describe a classification approach employing ANN and subset characteristics to classify heart disease. PCA is used to pre-process data and reduce the number of variables. A set of characteristics that lower the number of diagnostic tests a patient must undergo indirectly. Our method was tested using a database on cardiac disease in Andhra Pradesh. Our results show that classification accuracy has improved over time compared to traditional methods. This technology is viable, faster, and more precise for treating cardiac disease. Artificial Neural Networks (ANN) methods were developed in a study employing an artificial intelligence system to diagnose heart illness from phonocardiogram (PCG) signals [18].

Four new signal parameters are used as inputs to the neural network: operation, intensity, mobility, and spectral peaks from power spectral density plots. In this study, 94 PCG signals were used to test

the neural networks' accuracy for three different heart conditions. The neural networks give the features after filtering the signals and extracting the feature attributes. The Radial Base Function (RBF) network and the Back Propagation Network are used to classify the data in this case (BPN). The receiver's Operating Characteristic (ROC) is used to assess the consistency of both systems. The results demonstrate that, compared to 90.8 percent for BPN, RBF predicted the disease with 98 percent accuracy. They created an artificial intelligence algorithm that successfully detects heart illness using PCG data. Cardiac disease is the most serious of the conditions. This disease is very popular these days, so we've used a variety of variables that may be applied to these heart disorders to identify the best approach to predict them. We've also employed predictive algorithms in this study [19]. On the dataset, the Naive Bayes method, which is based on risk variables, is tested. Researchers employed decision trees and a combination of algorithms in this study to predict heart disease based on the characteristics mentioned above. The results indicated that when the dataset is small, the naive Bayes method gives the proper answers, and when the dataset is huge, decision trees produce the same results. The effectiveness and robustness of the hybrid technique proposed in Data analysis of diverse forms for the diagnosis of heart illness [20] were found in a study suggesting a mixed solution, which was used in a data set for heart disease. As a result, this thesis explores various machine training methods and compares outcomes using performance metrics, such as precision, accuracy, recall, rate f1, etc. Using the optimised FCBF, PSO model, and ACO, the total classification accuracy was 99.65%. The results suggest that the proposed system outperforms the previous classic classification performance techniques.

### 1.3. Methodology

#### 1.3.1 Steps

In this section, I will be discussing the methods that can be applied to achieve the solution for predicting the heart problem in the human body.

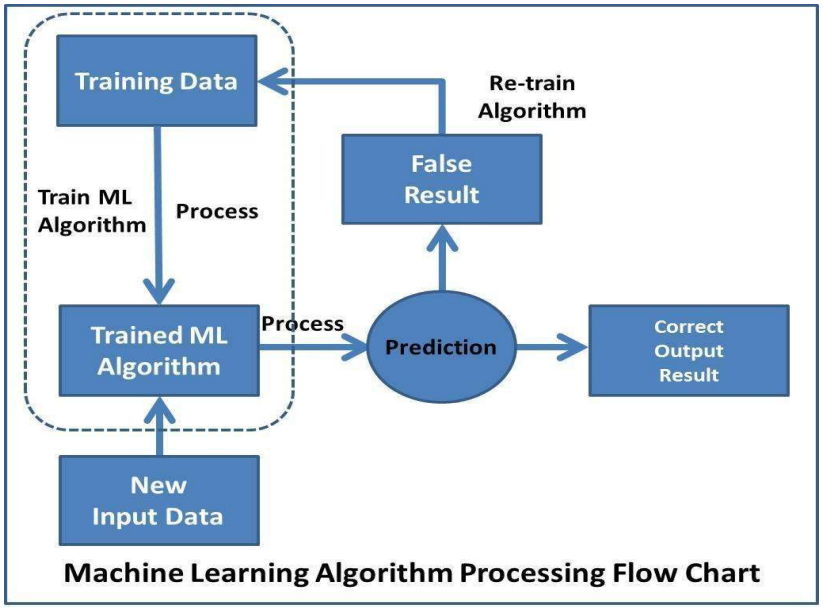



Figure 1.1 Proposed architecture for determining a patient's heart issue based on the patient's medical

history and available data.

### 1.3.2 Algorithms Description

**Step 1**: The patient's medical history is saved in a database readily available for training the model and obtaining the best outcomes.

**Step 2**: There can be no second thoughts in terms of model performance to predict in the healthcare system. As a result, the best-performing model for prediction should be hired. As a result, to fix the model, a large number of algorithms must be passed through this framework to assess the model's performance and check it.

**Step 3**: The training weights are then saved and can be used to evaluate the model's performance during the real-time analysis phase.

**Step 4**: Data from actual patients is analysed in real-time. The resulting error should be handed to the training stage, which will fine-tune the model parameters for more accurate outcomes.

To train and evaluate the data, I'll use the classification algorithms listed below:

1. **Logistic Regression:** When there are several explaining variables, logistic regression is used to calculate the odds ratio. Except for the binomial response parameter, the procedure is comparable to multiple linear regression. The impact of each variable on the odds ratio of the event of interest is the result. The key benefit is that all variables are analysed simultaneously, eliminating any potential for confusion. This post will show you how to utilise drawings to explain the logistic regression method. Once the approach is developed, it emphasises a fundamental understanding of the findings before addressing several specific issues.

2. **KNN (K-Nearest Neighbors):** KNN can be used to solve classification and regression problems in statistics. However, it is more extensively used in the industry for classification issues.

3. **Naïve Bayes:** It's a simple but effective predictive modelling algorithm. It's written as P(h|d) = (P(d|h) * P(h)) / P(d|h) (d). The probability of hypothesis h given the data d is P(h|d). The posterior probability is the term for this. P(d|h) is the probability that data d is correct if hypothesis h is correct. P(h) is the chance that hypothesis h is correct (regardless of the data). This is referred to as the h prior probability. The probability of the data is P(d) (regardless of the hypothesis).

4. **Linear SVC (Linear Support Vector Classifier):** A linear SVC aims to adapt to your input data and provide a hyperplane "best match" that divides or categorises it. After obtaining the hyperplane, you may input some features into your classifier to see the "predicted" class.

5. **Decision Tree Classifier:** A decision tree is a tree structure with a fluctuation chart, with each node indicating the outcome, the internal node representing the function(s), and the branch representing the decision rule. The decision-making process A decision tree's uppermost node is the core node. The attribute value can be used to partition. Recursive partitioning is when the tree is partitioned repeatedly. This fluid diagram will assist you in making decisions.

6. **Random Forest Classifier:** In order to increase the dataset's predictive accuracy, a classifier called Random Forest uses many decision trees on different subsets of the input data. Instead, of relying on a single decision tree, the random forest gathers predictions from each tree and predicts the final result based on the predictions that received the most votes.
7. **XGBoost Classifier:** A machine learning method called the XGBoost classifier. It can be used to categorise both structured and tabular data. Gradient-boosted decision trees are implemented using XGBoost, a fast and efficient method. It is therefore a substantial Machine Learning method with lots of moving parts. It can handle enormous, complicated datasets.
8. **Neural Network Classifier:** A neural network classifier transforms an input vector into output by using units (neurons) stacked in layers. Each unit receives an input, processes it using an (often nonlinear) function, and then transfers the results to the following layer. In general, networks are specified as feed-forward: there is no feedback to the previous layer; instead, each unit feeds its output to all the units on the layer above it. Signals travelling from one unit to another are given weightings, and these weightings are changed throughout the training process to adapt a neural network to the specific issue at hand.

### 1.3.2 Dataset

The data for this project will be taken from the git hub. The data set contains information on a patient's medical history and demographic information about the patient, such as gender, age, blood pressure, and other characteristics. This data will be treated as out-of-sample data, and the entire system will be corrected using open-source data.

### 1.3.3 Data Pre-Processing Techniques

To prepare the dataset for inclusion in the modelling and, afterwards, to conduct the study, various data prepossessing and imputation processes must be used. Some of the most prevalent are missing values imputation, outliers' treatment, data scaling, pine-hotel features generation, one hot encoding, and other approaches. Because the information is real-time, and in the healthcare sector, the most significant component of data imputation is consulting a domain expert for assistance rather than imputing it in terms of statistical metrics.

### 1.3.4 Modelling

Some of the most effective modelling approaches are SVM, Decision Trees, Logistic Regression, KNN, Gaussian Naive Bayes, Random Forest, XGBoost, Neural Network, and Decision Tree Classifier utilised for the analysis and getting the top methods out of the packs. In this scenario, the model with various hyper-parameter settings was fed through hyper-parameter optimization techniques such as the Grid Search and Random Search Algorithms. The ideal algorithm setting is achieved with the help of the selected hyperparameters and the optimum hyperparameter optimization setting. The best setting models are then put through the prediction process, with the best model being chosen based on the outcomes. The best model with hyperparameters is saved and

placed in the actual test, where real-time out-of-sample data is run through them to verify accuracy and performance. The analysis and suggestion models can aid doctors in making better diagnoses. Unlike many of the papers and studies available online, we will not be competing with doctors; instead, we will be assisting doctors in making appropriate diagnoses while reducing the turnaround time for each patient's diagnosis.

### 1.4. Implementation

#### 1.4.1. The Data Set

This study paper's data set is a patient's data set from various locations with 304 records. Approximately 304 patients were investigated in this study, and critical characteristics that may be the underlying cause of heart attacks were discovered. This research intends to present a job that could be an immediate step toward resolving a heart problem. The specific risk factors are the primary causes of heart disease. The core issue in this study has been described using a variety of classification frameworks. The records are binary, with one indicating yes and 0 indicating no. In addition, there are more data columns: age and gender.

#### 1.4.2. Data Processing

To clean and explain this investigation's results, the author used various data analytic methods. Specific approaches, such as pre-possession and data imputation, would ensure that the data collected was accurate enough to be used in the modelling and investigation. Missing values are imputed, outliers are processed, results are scaled, polynomial functions are generated, and one hot encoding is used. Once the data is imputed, it's critical to consult a domain authority instead of relying on mathematical approaches to support your input, mainly if the dataset comprises real-time data and healthcare context. Data frames are now categorised in features X and Y in Data Processing. Information categorised has been added to the location column. Information has been organised from the location column. I also made a categorical data conversion for the Sex column.

#### 1.4.2. Model

Given the data supplied, I experimented with various algorithms to see which one was the most accurate at detecting a cardiac condition. Logistic regression, KNN, XGBoost, SVM, Naive Bayes, Decision Tree Classifier, Random Forest, and Neutral Network are examples of algorithms. The goal was to determine whether the prediction was more accurate because the heart is one of the human body’s most vital parts, and a poor prognosis can result in the patient's death. I discovered a correlation between the features in the first section of model coding. I've also added a function that returns the model's accuracy, which helps us distinguish between different techniques. When using a logistic algorithm, there are a few things to remember.

The accuracy of the logistic algorithm implementation is 85.25 percent.

The following is the output of logistic regression:

```
The accuracy score achieved using Logistic Regression is: 85.25 %
```

KNN is the second algorithm that I have implemented. As discussed in the methodology segment, this algorithm returned 67.21% accuracy, as displayed in the output image below.

```
The accuracy score achieved using KNN is: 67.21 %
```

One of the classification models I've tested to see if this is more accurate is the Gaussian naive Bayes model. Following implementation, I discovered that the accuracy rate of this technique is 85.25, as seen in the output window below.

```
The accuracy score achieved using Naive Bayes is: 85.25 %
```

### 1.4.3. Hardware and Software

I used Jupyter Notebook as the platform for this research paper, an open-source web program allowing you to create and share documents. Data cleaning and transformation, numerical simulation, statistical modelling, data visualisation, machine learning, and many other applications are possible. This is a cloud-based platform that requires no setup to execute your task. To write code, a developer does not need to install anything. In comparison to other local platforms, I found this to be quite quick and straightforward to use and import, to the point that I was able to work on my project without having to use my laptop. Because your code is kept in the cloud, you can access it from anywhere you are, whenever you need it.

## 1.5. Results And Analysis

Various metrics can be used to better analyse different machine learning models and algorithms. ROC curves, Confusion matrices, and accuracy are a few examples. The model fixing algorithm that is the most stable and performs the best in this procedure can be used.
The following are the outcomes of various data processing code segments and algorithms.

### 1.5.1. Data Processing

We started by importing the necessary libraries to execute the algorithms.

```
In [1]: import numpy as np
        import pandas as pd
        import matplotlib.pyplot as plt
        import seaborn as sns

        %matplotlib inline

        import os
        print(os.listdir())

        import warnings
        warnings.filterwarnings('ignore')
```

```
['.conda', '.condarc', '.ipynb_checkpoints', '.ipython', '.jupyter', '.matplotlib', '.spyder-py3', '3D Objects
', 'anaconda3', 'AppData', 'Application Data', 'Contacts', 'Cookies', 'Desktop', 'Documents', 'Downloads', 'Fa
vorites', 'IntelGraphicsProfiles', 'Jedi', 'Links', 'Local Settings', 'main script of project.ipynb', 'Music',
'My Documents', 'NetHood', 'new one.ipynb', 'New project.ipynb', 'NTUSER.DAT', 'ntuser.dat.LOG1', 'ntuser.dat.
LOG2', 'NTUSER.DAT{ae2cbebd-7c14-11ec-8b05-9d995dc45230}.TM.blf', 'NTUSER.DAT{ae2cbebd-7c14-11ec-8b05-9d995dc4
5230}.TMContainer00000000000000000001.regtrans-ms', 'NTUSER.DAT{ae2cbebd-7c14-11ec-8b05-9d995dc45230}.TMContai
ner00000000000000000002.regtrans-ms', 'ntuser.ini', 'OneDrive', 'OneDrive - SGT University', 'PrintHood', 'pro
ject workshop.ipynb', 'PycharmProjects', 'Recent', 'Saved Games', 'Searches', 'SendTo', 'Start Menu', 'Templat
es', 'Untitled.ipynb', 'Untitled1.ipynb', 'Untitled14.ipynb', 'Untitled2.ipynb', 'Untitled3.ipynb', 'Untitled
4.ipynb', 'Untitled5.ipynb', 'Untitled6.ipynb', 'Untitled7.ipynb', 'Untitled8.ipynb', 'Untitled9.ipynb', 'Vide
os', 'Zotero']
```

Figure 1.2 Shows about libraries we used in the code.

Then we import the data files that include the patient's information.

```
In [2]: dataset = pd.read_csv("F:\Heart\heart11.csv")

In [3]: type(dataset)

Out[3]: pandas.core.frame.DataFrame
```

Figure 1.3 About the way to load the data set in memory

Then give a summary of the information's data shape and content.

```
In [4]: dataset.shape
Out[4]: (303, 14)

In [5]: dataset.head(5)
Out[5]:
```

| | age | sex | cp | trestbps | chol | fbs | restecg | thalach | exang | oldpeak | slope | ca | thal | target |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 63 | 1 | 3 | 145 | 233 | 1 | 0 | 150 | 0 | 2.3 | 0 | 0 | 1 | 1 |
| 1 | 37 | 1 | 2 | 130 | 250 | 0 | 1 | 187 | 0 | 3.5 | 0 | 0 | 2 | 1 |
| 2 | 41 | 0 | 1 | 130 | 204 | 0 | 0 | 172 | 0 | 1.4 | 2 | 0 | 2 | 1 |
| 3 | 56 | 1 | 1 | 120 | 236 | 0 | 1 | 178 | 0 | 0.8 | 2 | 0 | 2 | 1 |
| 4 | 57 | 0 | 0 | 120 | 354 | 0 | 1 | 163 | 1 | 0.6 | 2 | 0 | 2 | 1 |

```
In [8]: dataset.sample(5)
Out[8]:
```

| | age | sex | cp | trestbps | chol | fbs | restecg | thalach | exang | oldpeak | slope | ca | thal | target |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 26 | 59 | 1 | 2 | 150 | 212 | 1 | 1 | 157 | 0 | 1.6 | 2 | 0 | 2 | 1 |
| 17 | 66 | 0 | 3 | 150 | 226 | 0 | 1 | 114 | 0 | 2.6 | 0 | 0 | 2 | 1 |
| 4 | 57 | 0 | 0 | 120 | 354 | 0 | 1 | 163 | 1 | 0.6 | 2 | 0 | 2 | 1 |
| 211 | 61 | 1 | 0 | 120 | 260 | 0 | 1 | 140 | 1 | 3.6 | 1 | 1 | 3 | 0 |
| 249 | 69 | 1 | 2 | 140 | 254 | 0 | 0 | 146 | 0 | 2.0 | 1 | 3 | 3 | 0 |

Figure 1.4 Details of data have information

Information about datatype:

```
In [11]: info = ["age","1: male, 0: female","chest pain type, 1: typical angina, 2: atypical angina, 3: non-anginal pain,


for i in range(len(info)):
    print(dataset.columns[i]+":\t\t\t"+info[i])
```

```
age:                    age
sex:                    1: male, 0: female
cp:                     chest pain type, 1: typical angina, 2: atypical angina, 3: non-anginal pain, 4: asympt
omatic
trestbps:                       resting blood pressure
chol:                    serum cholestoral in mg/dl
fbs:                    fasting blood sugar > 120 mg/dl
restecg:                        resting electrocardiographic results (values 0,1,2)
thalach:                         maximum heart rate achieved
exang:                  exercise induced angina
oldpeak:                        oldpeak = ST depression induced by exercise relative to rest
slope:                  the slope of the peak exercise ST segment
ca:                     number of major vessels (0-3) colored by flourosopy
thal:                   thal: 3 = normal; 6 = fixed defect; 7 = reversable defect
```

Figure 1.5 Information about data which shows brief detail of the dataset

Let's discuss the target

```
In [12]: dataset["target"].describe()
```

```
Out[12]: count    303.000000
         mean       0.544554
         std        0.498835
         min        0.000000
         25%        0.000000
         50%        1.000000
         75%        1.000000
         max        1.000000
         Name: target, dtype: float64
```

```
In [13]: dataset["target"].unique()
```

```
Out[13]: array([1, 0], dtype=int64)
```

```
In [14]: print(dataset.corr()["target"].abs().sort_values(ascending=False))
```

```
target      1.000000
exang       0.436757
cp          0.433798
oldpeak     0.430696
thalach     0.421741
ca          0.391724
slope       0.345877
thal        0.344029
sex         0.280937
age         0.225439
trestbps    0.144931
restecg     0.137230
chol        0.085239
fbs         0.028046
Name: target, dtype: float64
```

The value to target 0 represents a person with no heart disease, while the value to target 1 represents a person with heart disease.

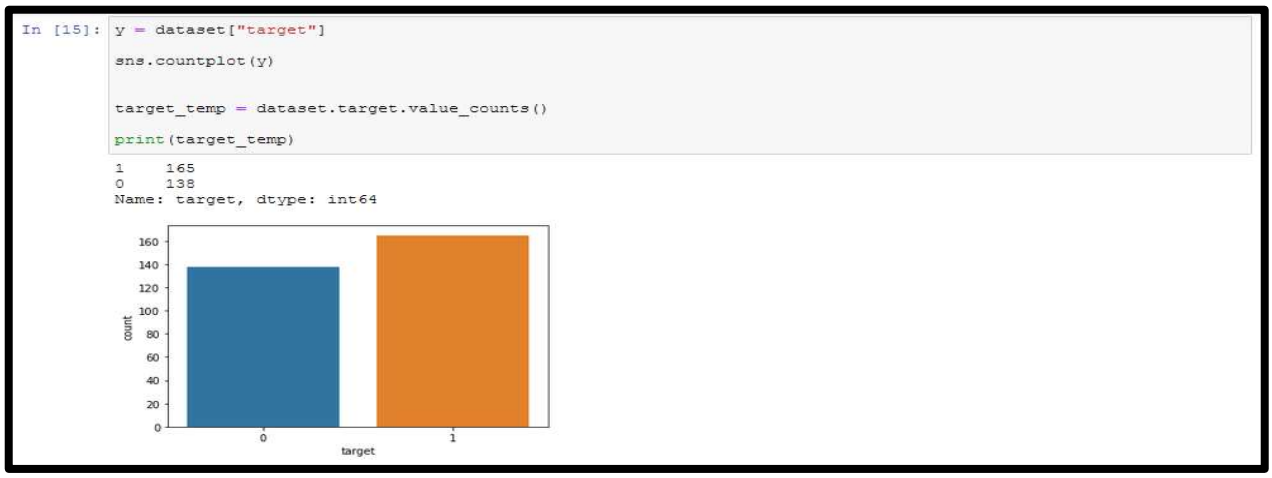
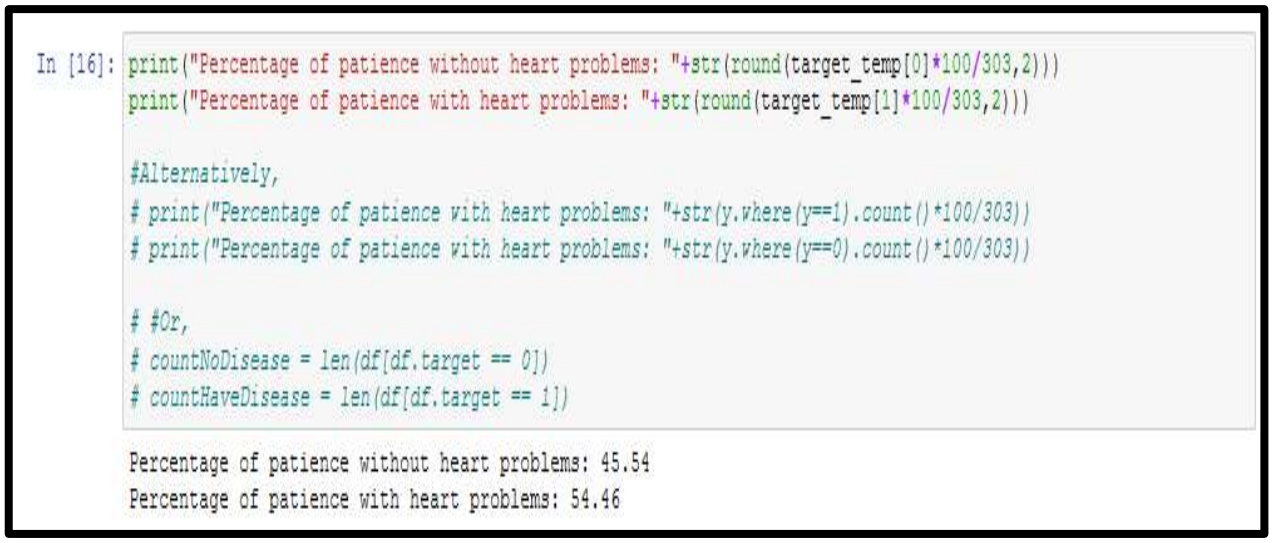



Figure 1.6 Shows the target graph.




It displays the number of males and females in a data list, with 0 denoting females and 1 indicating males.

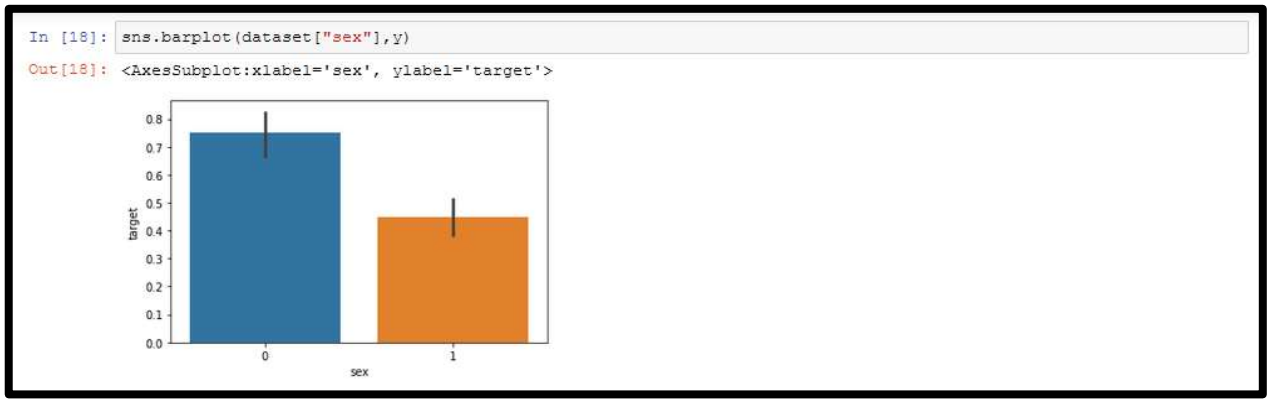



Figure 1.7 Shows a distribution graph between male and female

Type of Chest pain:

0 – Typical Angina 1 – Atypical Angina 2 – Non-Anginal Pain
3 - Asymptomatic

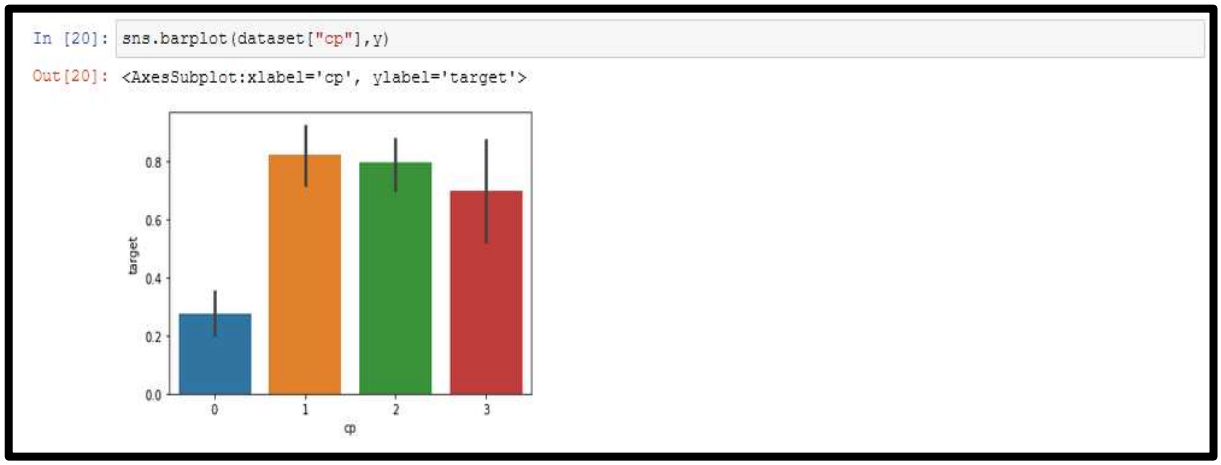



Figure 1.8 Shows a distribution graph between chest pains

Shows Fbs – Fasting Blood Sugar (> 120 mg/dl, 1 = true; 0 = false)

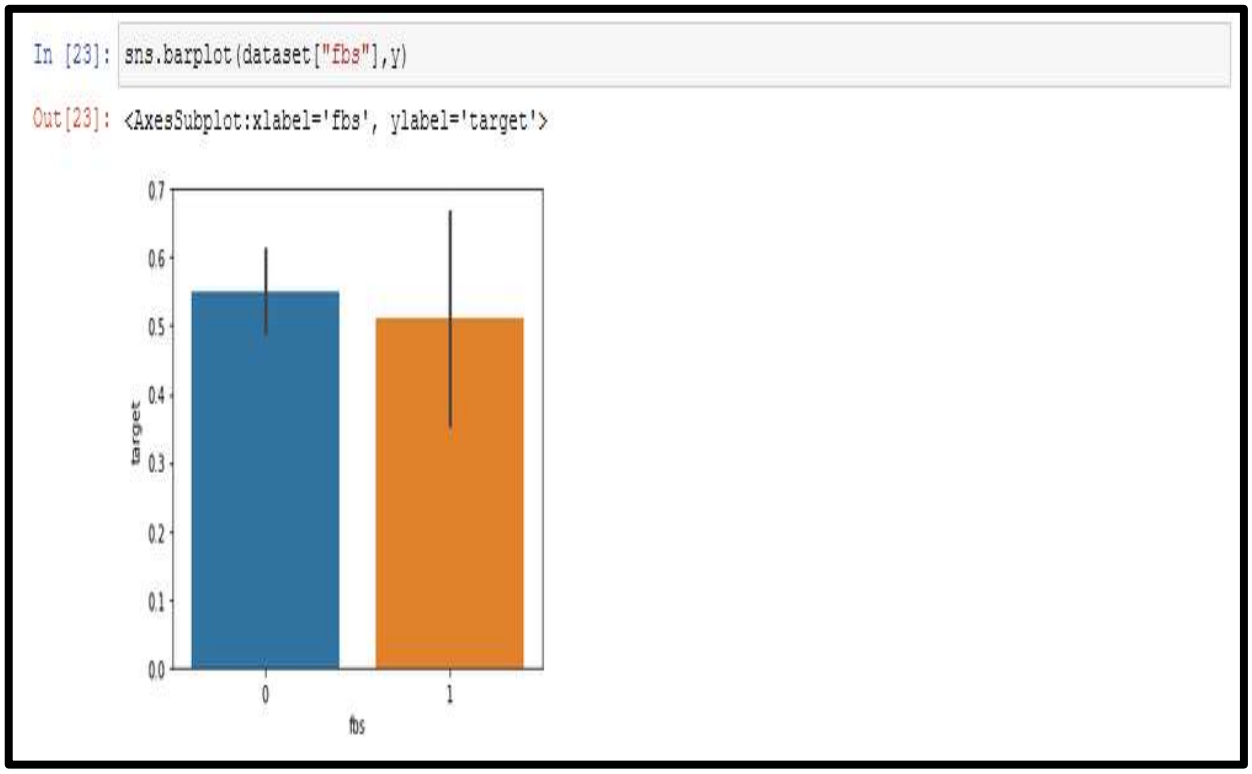



Figure 1.9 Shows the distribution graph of Fasting Blood Sugar.

The above result shows restecg -Resting Electrocardiographic results:

Value 0: showing probable or definite left ventricular hypertrophy by Estes' criteria

Value 1: normal

Value 2: having ST-T wave abnormality (T wave inversions and ST elevation or depression of > 0.05 mV)

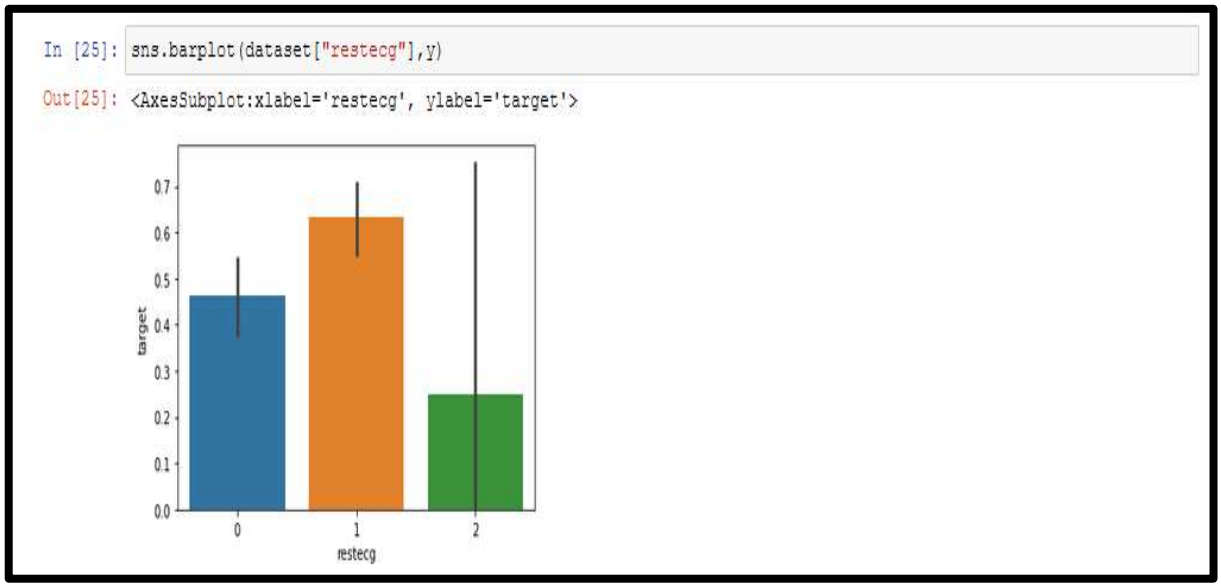



Figure 1.10 Shows a distribution graph between different types of restecg

The above figure shows exang - Exercise-induced angina (1 = yes; 0 = no)

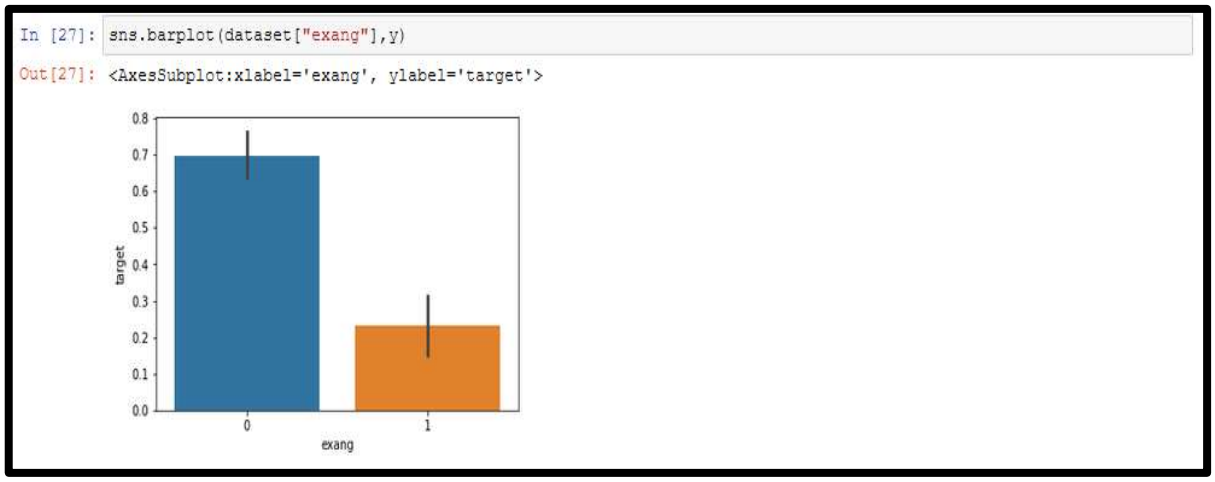



Figure 1.11 Shows the distribution graph of Exercise-induced angina

The above figure shows the slope of the peak exercise ST segment

0: downsloping;

1: flat;

2: upsloping

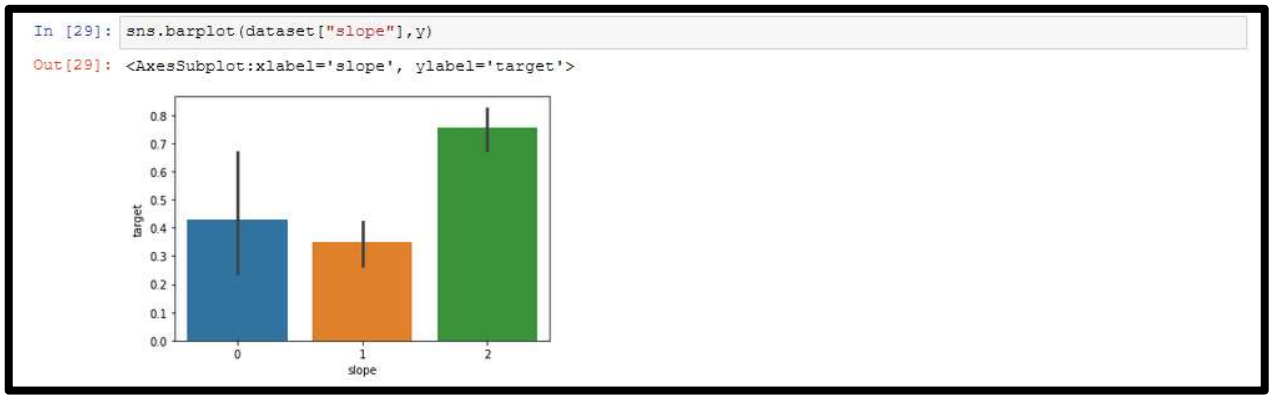



Figure 1.12 shows the distribution graph of the Slope

The above figure shows ca - the number of major vessels (0-3) coloured by fluoroscopy

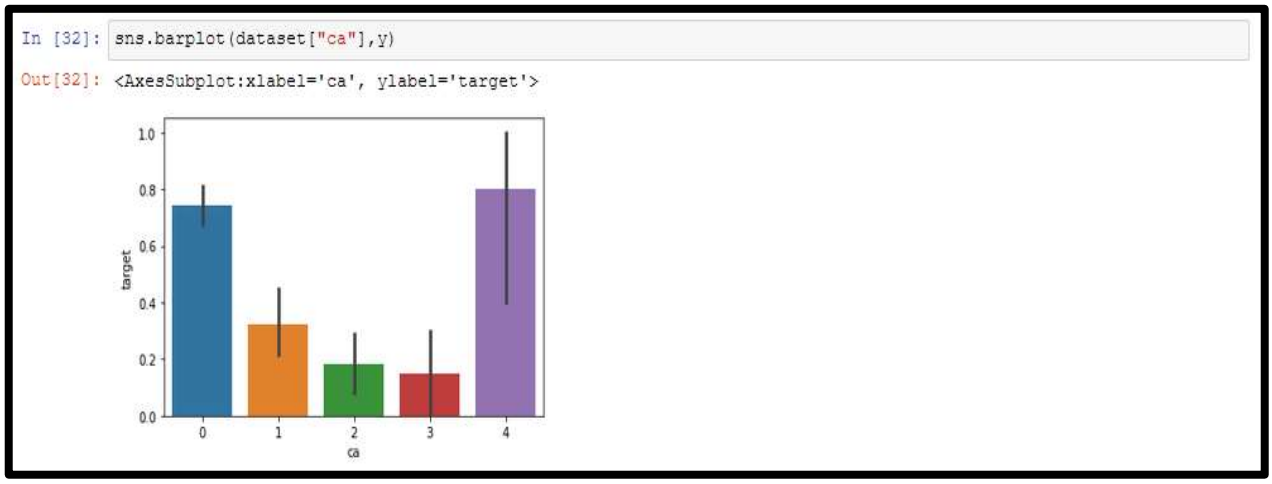



Figure 1.13 shows the distribution graph of the CA

The above figure shows Shows thal – A blood disorder called thalassemia.

Value 1: fixed defect (no blood flow in some part of the heart)

Value 2: normal blood flow

Value 3: reversible defect (a blood flow is observed, but it is not normal)

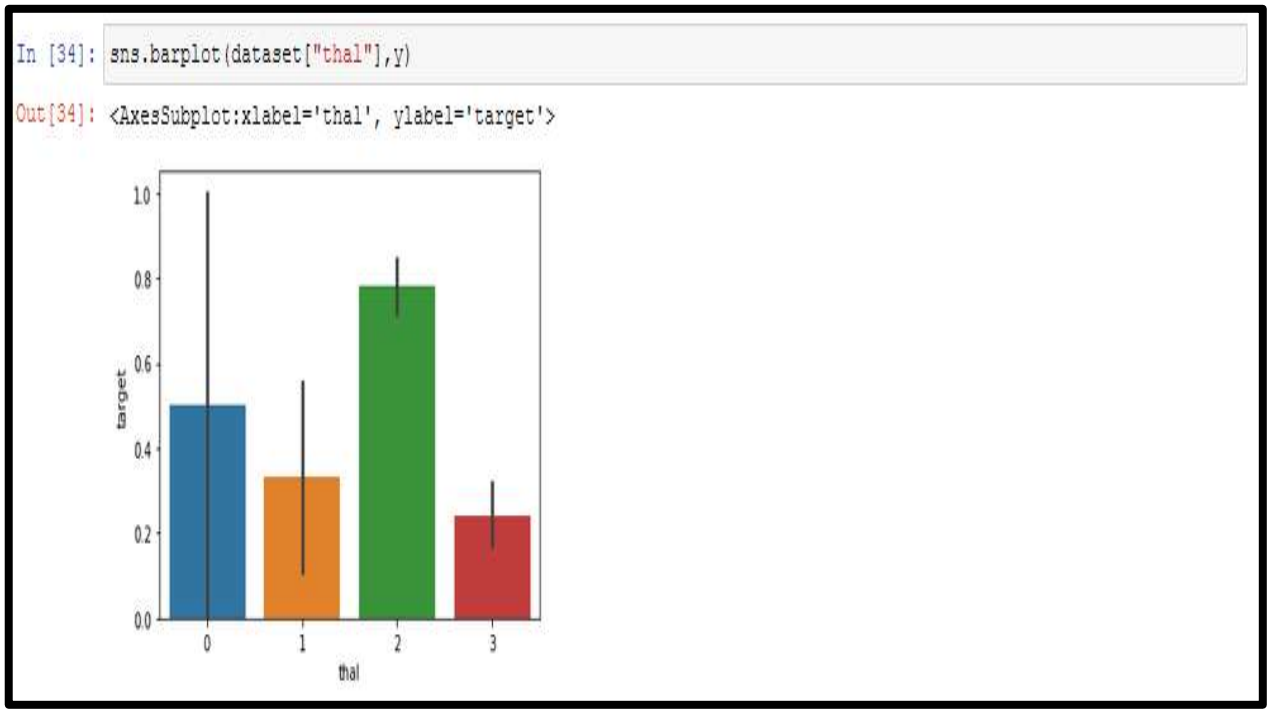



Figure 1.14 The distribution graph of Thal

### 1.5.2 Implemented Algorithms

As previously stated in the methodology section, the algorithms that I utilised to conduct this research and the outputs of those algorithms are described in this section.

### 1.6 Conclusion

One of the leading causes of death is heart disease. According to the World Health Organisation, 17.9 million people die yearly, with heart disease accounting for 80% of deaths. This study aims to employ techniques such as machine learning to discover a heart-related issue early on so that it may be cured and a person can have a second chance at life. From a technological standpoint, machine learning assisted me in determining which algorithm is best suitable for usage in the pharmaceutical business to design a device or software system that can detect heart-related issues at an early stage. The author proved that **Random Forest** is one of the algorithms that may be used to assess a human body for heart-related issues based on the risk variables presented. I found that Random Forest is a significantly more accurate algorithm in my research. It's not that those other algorithms aren't helpful; they do. They help me understand the likelihood of whether to use them or not and since health is something we can't bargain with, having a correct method or process that produces the most accurate results is critical. Based on my findings, I believe the **Random Forest** algorithm or model can be utilised to diagnose a heart-related condition using the parameters provided.

**References**

[1] *Chen, Austin H. et al. "HDPS: Heart disease prediction system." 2011 Computing in Cardiology (2011): 557-560.*

[2] Pahulpreet Singh Kohli and Shriya Arora, "Application of Machine Learning in Disease Prediction", International Conference on Computing Communication and Automation (I.C.C.C.A.), IEEE, 2018.

[3] Nikhar, S. and Aarti Karandikar. "Prediction of Heart Disease Using Machine Learning Algorithms." International Journal of Advanced Engineering, Management and Science 2 (2016): 239484.

[4] Rodgers, Anthony, et al." Blood pressure and risk of stroke in patients with cerebrovascular disease." Bmj 313.7050 (1996): 147.

[5] V. Krishnaiah, M. Srinivas, G. Narsimha and N. S. Chandra, "Diagnosis of heart disease patients using fuzzy classification technique," International Conference on Computing and Communication Technologies, 2014, pp. 1-7, doi: 10.1109/ICCCT2.2014.7066746.

[6] Humar Kahramanli; Novruz Allahverdi (2008). Design of a hybrid system for diabetes and heart diseases., 35(1-2), 82–89. doi: 10.1016/j.eswa.2007.06.004.

[7] Das, Resul, Ibrahim Turkoglu, and Abdulkadir Sengur. "Effective diagnosis of heart disease through neural networks ensembles." Expert systems with applications 36.4 (2009): 7675-7680.

[8] Verlangieri AJ, Kapeghian JC, el-Dean S, Bush M. Fruit and vegetable consumption and cardiovascular disease mortality. Med Hypotheses 1985; 16:7-15.

[9] US Office on Smoking and Health. The Health Consequences of Smoking: Cardiovascular Diseases: A Report of the Surgeon General. Washington, DC: US Government Printing Office; 1989:179-203.

[10] Jan G. Bazan. A comparison of dynamic and non-dynamic rough set methods for extracting laws from decision tables. In Polkowski and Skowron [169], chapter 17, pages 321–365.

[11] Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. (1984), Classification and Regression Trees, Wadsworth International Group.

[12] Leung AA, Daskalopoulou SS, Dasgupta K, McBrien K, Butalia S, Zarnke KB, et al. Hypertension Canada's 2017 guidelines for diagnosis, risk assessment, prevention, and treatment of hypertension in adults. Can J Cardiol. 2017; 33:557–76.

[13] Giovanni Jacopo Ughi, Tom Adriaenssens, Peter Sinnaeve, Walter Desmet, and Jan D'hooge, "Automated tissue characterization of in vivo atherosclerotic plaques by intravascular optical coherence tomography images," Biomed. Opt. Express 4, 1014-1030 (2013).

[14] Kansadub, Teerapat, et al. "Stroke risk prediction model based on demographic data." 2015 8th Biomedical Engineering International Conference (BMEiCON). IEEE, 2015.

[15] MA. Jabbar, B.L Deekshatulu, Priti Chandra, "heart disease prediction system using associative classification", ICECIT 2012, Elsevier vol. no. 1 pp183-192.

[16] F. Yaghouby, A. Ayatollahi and R. Soleimani, Classification of Cardiac Abnormalities Using Reduced Features of Heart Rate Variability Signal". World Applied Sciences Journal 6 (11), 2009,

pp. 1547-1554.

[17] H. Sharma, “Prediction of Heart Disease using Machine Learning Algorithms: A Survey,” Int. J. Recent Innov.Trends Comput. Commun., vol. 5, no. 8, pp. 99–104,2017.
[18] Jyoti Rohilla, Preeti Gulia, ―Analysis of Data Mining Techniques for Diagnosing Heart Disease”, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 7, ISSN: 2277 128X, July 2015.
[19] M. Akhil, B. L. Deekshatulu, and P. Chandra, “Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm,” ProcediaTechnol., vol. 10, pp. 85–94, 2013.
[20] M. Fatima and M. Pasha, "Survey of machine learning algorithms for disease diagnostic. Journal of Intelligent Learning Systems and Applications", Journal of Intelligent Learning Systems and Applications, Vol.9, No.01, pp.1,2017.