HEART FAILURE PREDICTION

|  |  |
| --- | --- |
| Name | Student ID |
| Kamalpreet kaur | 0756411 |
| Bhawna Bawa | 0756420 |
| Renu Bala | 0765207 |
| Baljinder Kaur | 0765237 |
| Supreet Kaur | 0771080 |

Submitted to:

Ms. Savita Sherawat

Table of Contents

Abstract………………………………………………………………………………………...3

Detail Data Dictonary………………………………………………………………………….4-6

Research Questions…………………………………………………………………………….6

Github Link……………………………………………………………………………………6

History of Dataset..……………………………………………………………………………7

Initial Data Analysis…………………………………………………………………………..8-21

Refrences……………………………………………………………………………………..22

# Abstract

In this project (DATA SET: HEART FAILURE PREDICTION), we will explore the clinical records of heart failure patients, the causes of their heart attacks. There are many other datasets available, but we chose this one because we want to know why they had a heart attack. Once we know why, we can simply solve the problem.

Cardiovascular diseases (CVDs) are the leading cause of death worldwide, killing an estimated 17.9 million people each year, accounting for 31% of all deaths. Heart attacks and strokes account for four out of every five CVD deaths, with one-third of these deaths occurring before the age of 70. CVDs are a common cause of heart failure, and this dataset contains 11 variables that can be used to predict heart disease.

People who have cardiovascular disease or are at high cardiovascular risk (due to one or more risk factors such as hypertension, diabetes, hyperlipidemia, or previously established disease) require early detection and management. In this dataset the group decide to present the information of cardiac disease among people like which factor is more contributing to heart failure.

**Research Questions:**

1. Which Age group is more affected by the heart disease?
2. Which attribute is more contributing in CVDs problem?
3. Heart Disease ratio between males and females?
4. Which type of chest pain cause heart disease?

**Keywords:** Linear Regression, Logistic Regression, KNN Model, Decision Tree, Kmeans Clustering, Confusion Metrix.

Tools: R studio, MS-Word

# Detail Data Dictionary

Dataset Contains 918 observations and 12 Attributes

|  |  |  |  |
| --- | --- | --- | --- |
| S.no | Attribute | Type | Description |
| 1 | Age | integer | Age of the patient in Years. |
| 2 | Sex | character | Gender of the patient  M: male  F: female |
| 3 | Chest Pain Type | character | Different types of chest pain  TA: Typical Angina,  ATA: Atypical Angina,  NAP: Non-Anginal Pain,  ASY: Asymptomatic |
| 4 | Resting BP | integer | Resting blood pressure [mm Hg] High blood pressure can damage the arteries that supply your heart over time. High blood pressure combined with other health problems, such as obesity, high cholesterol, or diabetes, raises the risk even more |
| 5 | Cholesterol | integer | Serum cholesterol (mm/dl). This attribute shows the cholesterol level of the patient. |
| 6 | Fasting BS | integer | fasting blood sugar (1: if fasting > 120 mg/dl, 0: otherwise) Blood sugar levels rise when the pancreas does not produce enough insulin or when the body does not respond to insulin adequately, increasing the risk of a heart attack |
| 7 | Resting ECG | character | Displays resting electrocardiographic results  0=normal  1=having ST-T wave abnormality  2 = left ventricular hypertrophy |
| 8 | Max HR | integer | Maximum heart rate achieved [Numeric value between 60 and 202] The increase in cardiovascular risk related with heart rate acceleration was comparable to the risk associated with high blood pressure. |
| 9 | Exercise Angina | character | Exercise-induced angina  Y: yes  N: No  Angina pain or discomfort is usually tight, gripping, or squeezing, and it can range from mild to severe. |
| 10 | Old peak | numeric | ST [Numeric value measured in depression The duration of ST-segment depression is also important, as prolonged recovery after peak stress is consistent with a positive treadmill ECG stress test |
| 11 | ST\_Slope | character | The slope of the peak exercise ST segment (Up: upsloping, Flat: flat, down: down sloping) When the ST-segment depression is less than 1 mm at 60–80 ms following the J point on a treadmill ECG stress test, it is considered abnormal |
| 12 | Heart Disease | integer | 1: heart disease  0: Normal |

**GitHub Link**

[Baljinderkaur29/DAB402 (github.com)](https://github.com/Baljinderkaur29/DAB402)

**History of dataset**

This data comes from the University of California Irvine's Machine Learning Repository at <https://archive.ics.uci.edu/ml/datasets/Heart+Disease>. This directory contains 4 databases concerning heart disease diagnosis. All attributes are numeric-valued. The data was collected from the four following locations:

Creators:

1. Cleveland Clinic Foundation (Cleveland. Data)

2. Hungarian Institute of Cardiology, Budapest (Hungarian. Data)

3. V.A. Medical Center, Long Beach, CA (long-beach-va.data)

4. University Hospital, Zurich, Switzerland (Switzerland. Data)

Prasanta Kumar Sahoo and Pravalika Jeripothula have been researched well and proposed different Classification and prediction algorithms but each one has its own limitations. The main objective of this paper is to overcome the limitations and to design a robust system which works efficiently and will be able to predict the possibility of heart failure accurately. This paper uses the data set from the UCI repository and having 13 important attributes. This work is implemented using many algorithms such as SVM, Naïve Bayes, Logistic Regression, Decision Tree and KNN. It is found that SVM gave the best result with accuracy up to 85.2%.

Devansh Shah, Samir Patel & Santosh Kumar Bharti applied four data mining classification techniques, K-nearest neighbor, Naive Bayes, decision tree, and random forest. The data were pre-processed and then used in the model. K-nearest neighbor, Naïve Bayes, and random forest are the algorithms showing the best results in this model. I found the accuracy after implementing four algorithms to be highest in K-nearest neighbors.

**Initial Data Analysis**

**Histogram**: It means a statistical view of our dataset. **The histogram is a pictorial representation of a dataset distribution with which we could easily analyze which factor has a higher amount of data and the least data.**

Chart, histogram

Description automatically generated

Resting BP histogram shows that the blood pressure of CVD’s patient mostly in between 120 to 140.

Chart, histogram

Description automatically generated

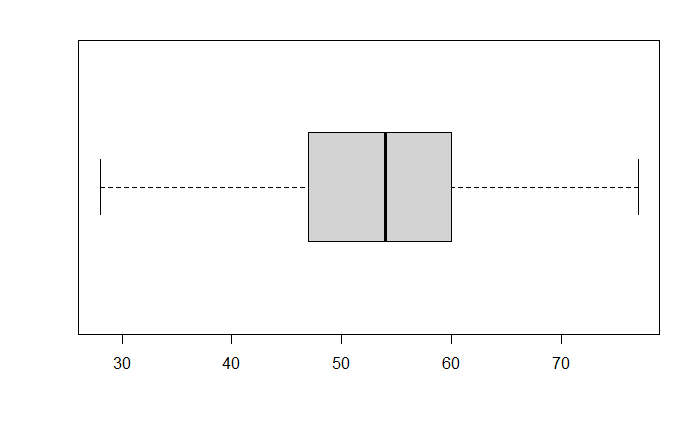
The nature of this histogram is symmetrical. And it shows that max heart rate of patient is 120.

Chart, histogram

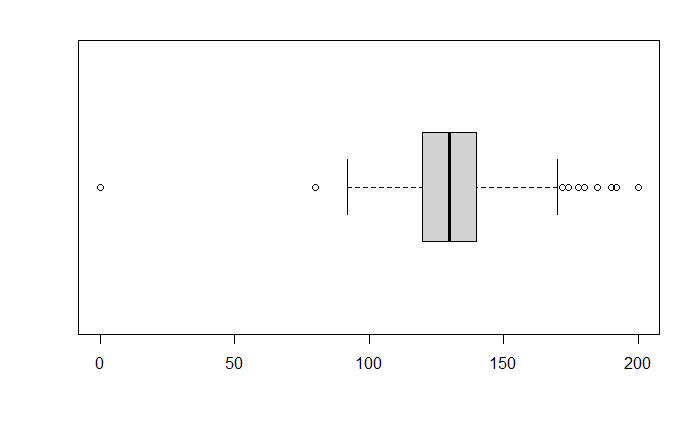
Description automatically generated

From the above histogram I analyzed that person who have heart diseases reported cholesterol level between 200 and 300.

**Boxplot**: It divides the dataset into maximum, minimum and median form. Boxplots are a popular type of graphic that visualize the minimum non-outlier, the first quartile, the median, the third quartile, and the maximum non-outlier of numeric data in a single plot



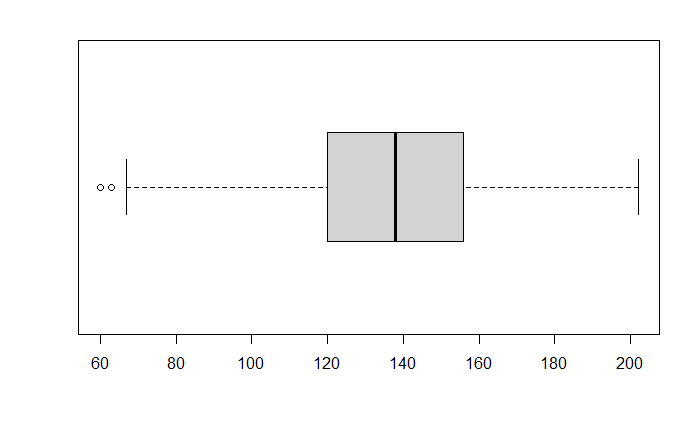
Boxplot of Age attribute



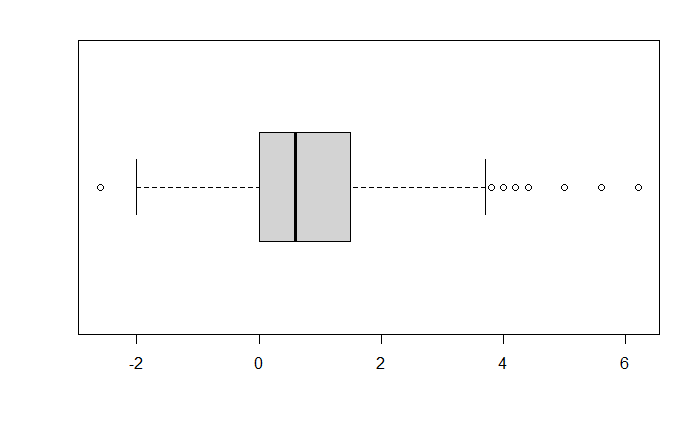
Boxplot of RestingBP attribute



Boxplot of Cholesterol attribute



Boxplot of MaxHR attribute



Boxplot of Oldpeak attribute

**Removing outliers:**

In this dataset “Age” attribute does not have any outlier.

To remove Outlier from the RestingBP, MaxHR and Oldpeak attribute we use Replace method we find the average of remaining values and replace the outliers with the average value. On the other hand, to remove outliers from the Cholesterol attribute we used remove method that we delete the outlier records.

**Barplot**: Barplot is used to display the relationship between numeric and categorical attributes. Each entity of the categorical variable is represented as a bar.

Chart, bar chart, histogram

Description automatically generated

In the above bar plot I analyzed that male have more heart problems as compared to females.

Chart, bar chart

Description automatically generated

Barplot of chestpain

After analysing the above barplot I noticed that out of the 4 chest pain types ASY (Asymptomatic) chest pain is more responsible for heart diseases.

**Correlation:** It is a mutual relationship or connection between two or more things. To find Correlation we used ‘cor’ function. Useful to highlight the most correlated variables in a data table. A heatmap (or heat map) is another way to visualize correlation. It’s also called a false colored image, where data values are transformed to color scale.

Chart, box and whisker chart

Description automatically generated

**Logistic Regression:** It is nothing but generalization of liner regression model, but the condition is target variable has two outputs. Logistic regression in R is defined as the binary classification problem in the field of statistic measuring. The difference between a dependent and independent variable with the guide of logistic function by estimating the different occurrence of the probabilities, i.e., it is used to predict the outcome of the independent variable

Chart, line chart, scatter chart

Description automatically generated

**Classification**

Classification is the process of predicting values based on previous values. The possible classes are already known and based on the known values we can predict new values.

To classify our data firstly we divide our dataset into training and test set. I divide 60% of data in training set and rest 40% is in test set.

**KNN Model**: K-nearest neighbors (KNN) algorithm is a type of supervised ML algorithm which can be used for both classification as well as regression predictive problems. However, it is mainly used for classification predictive problems

Graphical user interface, application, Teams

Description automatically generated

KNN model on Chest pain Type Attribute

**Accuracy: 0.4172414**

I found KNN model is 42% accurate to predict chest Pain

**Decision Tree Model**: Decision tree is a type of supervised learning algorithm that can be used in both regression and classification problems. It works for both categorical and continuous input and output variables

Diagram

Description automatically generated

Decision tree for chest pain type

**Accuracy: 0.5586207**

I found Decision tree model is 56% accurate to predict chest Pain

**Random forest Model**: It is a supervised learning algorithm that randomly creates and merge multiple decision trees into one forest. It also tells the error rate in prediction

Chart, histogram

Description automatically generated

Random forest on chest Pain

**Accuracy: 0.5827586**

I found random forest is 58% accurate to predict chest Pain

Model Accuracy comparison for chest pain Type

Table

Description automatically generated

Above table compares the accuracies of three classification models that is 0.4172414,0.5586207 and 0.5827586 for KNN, Decision tree and Random Forest respectively.

Classification for Sex attribute

Graphical user interface, application, Teams

Description automatically generated

KNN model on Sex Attribute

**Accuracy: 0.6793103**

I found KNN model is 67% accurate to predict Gender

Diagram

Description automatically generated

Decision tree on gender

**Accuracy: 0.7551724**

I found Decision tree model is 75% accurate to predict Gender

**Random Forest**

Chart, histogram

Description automatically generated

Random forest for Sex

**Accuracy: 0.7551724**

I found random forest is 75% accurate to predict chest Pain

Model Accuracy comparison for Sex attribute

Graphical user interface, application

Description automatically generated

Above table compares the accuracies of three classification models that is 0.6793103,0.7551724and 0.7551724 for KNN, Decision tree and Random Forest respectively. Here we notice that Accuracy of Decision tree and Random Forest model is exactly same that means we can use any of these models for prediction.

# References

* <https://www.kaggle.com/fedesoriano/heart-failure-prediction>
* <https://towardsdatascience.com/heart-disease-prediction-73468d630cfc>
* Prasanta Kumar Sahoo (Sreenidhi Institute of Science and Technology, December 15, 2020)
* Pravalika Jeripothula (Sreenidhi Institute of Science & Technology (SNIST), December 15, 2020)
* Devansh Shah, Samir Patel & Santosh Kumar Bharti (Computer Engineering Department, School of Technology, India) 16 October 2020