# HEART DISEASE PREDICTION USING LOGISTIC REGRESSION IN COMPARISON WITH KNN ALGORITHM

**A PROJECT REPORT**

*Submitted by*

**Sahil Anwar (RA2011042010131) Uday Kumar (RA2011042010128)**

*Under the guidance of* **Dr.S Jeeva (Assistant Professor)**

## in partial fulfillment for the award of the degree of

**BACHELOR OF TECHNOLOGY**

in

**COMPUTER SCIENCE & ENGINEERING**

of

**FACULTY OF ENGINEERING AND TECHNOLOGY**



S.R.M. Nagar, Kattankulathur, Chengalpattu District

**NOVEMBER 2022**

# SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

(Under Section3 of UGC Act, 1956)

# BONAFIDE CERTIFICATE

Certified that 18CSP361L mini project report titled “**Heart Disease Prediction using Logistic Regression in comparison with KNN Algorithm**” is the bonafide work of “Sahil Anwar **(**RA2011042010131**)**, Uday Kumar **(**RA2011042010128**)**” worked on the small project under my direction. Furthermore, I certify that, to the best of my knowledge, the work described here does not constitute any other research proposal or doctorate on the basis of which this candidate or any other candidate has previously received a degree or prize.

|  |  |
| --- | --- |
| Dr.S Jeeva  **SUPERVISOR**  Assistant Professor  Department of Data Science and Business System  Signature of Internal Examiner | Dr.M.Lakshmi  **HEAD OF DEPARTMENT**  Department of Data Science and Business System  Signature of External Examiner |

2

4

Own Work Declaration

Department Of Computer Science and Engineering

**SRM INSTITUTE OF SCIENCE AND TECHNOLOGY**

**Own Work Declaration Form**

This sheet must be filled in (each box ticked to show that the condition has been met). It must be signed and dated along with your student registration number and included with all assignments you submit – work will not be marked unless this is done.

To be completed by the student for all assessmentsDegree/ Course

: B.TECH

**Student Name :** Sahil Anwar, Uday Kumar

**Registration Number :** RA2011042010131, RA2011042010128]

**Title of Work :** HEART DISEASE PREDICTION USING ML

I / We hereby certify that this assessment compiles with the University’s Rules and Regulations relating to Academic misconduct and plagiarism\*\*, as listed in the University Website, Regulations, and the Education Committee guidelines.

I / We confirm that all the work contained in this assessment is my / our own except where indicated, and that I

/ We have met the following conditions:

* Clearly references / listed all sources as appropriate
* Referenced and put in inverted commas all quoted text (from books, web, etc)
* Given the sources of all pictures, data etc. that are not my own
* Not made any use of the report(s) or essay(s) of any other student(s) either past or present
* Acknowledged in appropriate places any help that I have received from others (e.g. fellowstudents, technicians, statisticians, external sources)
* Compiled with any other plagiarism criteria specified in the Course handbook / University website

I understand that any false claim for this work will be penalised in accordance with the University policies and regulations.

|  |
| --- |
| **DECLARATION:** |
| I am aware of and understand the University’s policy on Academic misconduct and plagiarism and I certify that this assessment is my / our own work, except where indicated by referring, and that I have followed the good academic practices noted above. |
| If you are working in a group, please write your registration numbers and sign with the date for every student inyour group. |

Sahil Anwar Uday Kumar

# ACKNOWLEDGEMENT

We express our heartfelt thanks to our honorable Vice Chancellor Dr. C. MUTHAMIZHCHELVAN, for being the beacon in all our endeavors.

We would like to express my warmth of gratitude to our Registrar Dr. S. Ponnusamy, for his encouragement.

We express our profound gratitude to our Dean of College of Engineering and Technology, Dr.

T. V.Gopal, for bringing out novelty in all executions.

We would like to express my heartfelt thanks to the Chairperson, School of Computing Dr. Revathi Venkataraman, for imparting confidence to complete my course project.

We are highly thankful to our course project faculty Dr. Jeeva S, Assistant Professor, Department of DSBS for his assistance, timely suggestion and guidance throughout the duration of this course project.

We are highly thankful to our mentor Mrs. M. Sangeetha, Assistant Professor, Department of DSBS for his assistance, timely suggestion and guidance throughout the duration of this course project.

We extend my gratitude to our HoD, Professor Dr. M. Lakshmi , Department of DSBS and my Departmental colleagues for their Support.

Finally, we thank our parents and friends near and dear ones who directly and indirectly contributed to the successful completion of our project. Above all, I thank the almighty for showering his blessings on me to complete my Course project.

Sahil Anwar (RA2011042010131) Uday Kumar (RA2011042010128)

# ABSTRACT

The most of the time, a difficult mix of pathological and clinical evidence is utilized to arrive at the diagnosis of heart disease. Because of this complexity, clinical practitioners and researchers are very eager to create a reliable method for predicting heart disease. In this work, we develop a method for determining the presence of heart disease that really can help doctors identify a patient's cardiovascular disease status based on their health records. Our approaches are divided into three phases. Firstly, we choose 14 significant clinical parameters, comprising sex, age, angina caused by activity, cholesrol, blood sugar, blood pressure, resting electrocardiogram, maximum heart rate, slope, old peak, aggregate of vessels, and thal. Secondly we need to process our data that we have collected because we cannot put our raw data directly into our machine learning algorithm. After we process our data we will split our data into two parts training data and testing data which will be used to train our machine learning model and evaluate the performance of our model respectively.

In the third step we will pass data into our trained logistic regression model and identify whether the person has a defective heart or not. Here we will be using various other machine learing algorithms also like K nearest neighbour, Support Vector Machine, Gausiaan Naive Bais and Gradient Bossting. In order to predict heart disease, the algorithms are applied based on the characteristics. This study aims to compare the accuracy of different machine learning methods. The important moto for our project will be to precisely forecast the risk of heart disease. Tests are primarily based on the features taken into account, and the findings are generally reliable. There are aggregate of input that can be utilized, but our aim is to predict the risk of heart disease with maximum speed. This approach results in unwanted biases, mistakes, and exorbitant medical expenses, that have an impact on the quality of care provided to patients.

# TABLE OF CONTENTS

[ABSTRACT 5](#_bookmark0)

[LIST OF FIGURES 9](#_bookmark1)

[ACRONYMS AND ABBREVIATIONS 12](#_bookmark2)

[INTRODUCTION 13](#_bookmark3)

* 1. [Project Description 13](#_bookmark4)
  2. [Motivation 13](#_bookmark5)
  3. [Problem Statement 13](#_bookmark6)
  4. [Software Requirements Specifications 14](#_bookmark7)

[LITERATURE SURVEY 15](#_bookmark8)

[SYSTEM ARCHITECTURE AND DESIGN 18](#_bookmark9)

* 1. [Machine learning 19](#_bookmark11)
  2. [Algorithms 20](#_bookmark14)
     1. [Logistic Regression Algorithm 20](#_bookmark15)
     2. [Naive Bayes Algorithm 21](#_bookmark17)
     3. [Support Vector Machine (SVM) 22](#_bookmark19)
     4. [K Nearest Neigbour Algorithm 23](#_bookmark21)
     5. [Gradient Boosting Algorithm 24](#_bookmark23)

[METHODOLOGY 27](#_bookmark25)

* 1. [Existing System 27](#_bookmark26)
  2. [Proposed System 27](#_bookmark27)
     1. [Dataset collection 27](#_bookmark28)
     2. [Attributes selection 28](#_bookmark30)
     3. [Exploratory data analysis 29](#_bookmark32)
     4. [Data Preprocessing 32](#_bookmark39)
     5. [Prediction of heart disease 33](#_bookmark41)

[CODING AND TESTING 35](#_bookmark43)

* 1. [Importing the Libraries 35](#_bookmark44)
  2. [Importing the Dataset 36](#_bookmark47)
  3. [Fixing the data types 37](#_bookmark49)
  4. [Exploratory data analysis 38](#_bookmark52)
  5. [Creating dummy variables 41](#_bookmark59)
  6. [Dropping unnecessary Variables 42](#_bookmark61)
  7. [Feature Separating 42](#_bookmark62)
  8. [Data Normalization 42](#_bookmark64)
  9. [Splitting the dataset 43](#_bookmark67)
  10. [Model Implementation 43](#_bookmark69)
      1. [Logistic regression Model 43](#_bookmark70)
      2. [K Nearest Neighbor Algorithm 44](#_bookmark72)
      3. [Support Vector Machine(SVM) Model 45](#_bookmark74)
      4. [Gaussian Naïve Bayes Model 46](#_bookmark76)
      5. [Gradient boosting Model 47](#_bookmark78)
      6. [Creating Accuracy comparison table 48](#_bookmark80)
      7. [Predicting case 48](#_bookmark82)

[Results and discussions 50](#_bookmark84)

[Conclusion and Future enhancements 52](#_bookmark87)

[REFERENCES 53](#_bookmark88)

[APPENDIX 56](#_bookmark89)

[APPENDIX A 56](#_bookmark90)

# LIST OF FIGURES

[Figure 1: system architecture 18](#_bookmark10)

[Figure 2 : graph for logistic regression 21](#_bookmark16)

[Figure 3: naive bayes classifier 22](#_bookmark18)

[Figure 4: support vector machine 23](#_bookmark20)

[Figure 5: KNN algorithm 24](#_bookmark22)

[Figure 6: boosting algorithm 26](#_bookmark24)

[Figure 7: collection of data 28](#_bookmark29)

[Figure 8: correlation matrix 29](#_bookmark31)

[Figure 9: distribution of heart disease on gender 29](#_bookmark33)

[Figure 10: distribution of heart disease on major vessels 30](#_bookmark34)

[Figure 11: heart disease scatter plot based on age 30](#_bookmark35)

[Figure 12: chest pain type distribution based on age 31](#_bookmark36)

[Figure 13: distribution of heart disease based on blood sugar 31](#_bookmark37)

[Figure 14: correlation map of numerical variables 32](#_bookmark38)

[Figure 15: data preprocessing 33](#_bookmark40)

[Figure 16: prediction of disease 34](#_bookmark42)

[Figure 17: importing library 35](#_bookmark45)

[Figure 18: creation of color palattes 36](#_bookmark46)

[Figure 19: importing dataset 37](#_bookmark48)

[Figure 20: datatype fixing 37](#_bookmark50)

[Figure 21: dataset information and detail 38](#_bookmark51)

[Figure 22: code part A 39](#_bookmark53)

[Figure 23: code part B 39](#_bookmark54)

[Figure 24: code part C 40](#_bookmark55)

[Figure 25: code part D 40](#_bookmark56)

[Figure 26: code part E 41](#_bookmark57)

[Figure 27: code part F 41](#_bookmark58)

[Figure 28: creation of dummy variables 42](#_bookmark60)

[Figure 29: code for dropping variables 42](#_bookmark63)

[Figure 30: code for feature seperating 42](#_bookmark65)

[Figure 31: normalization of data 43](#_bookmark66)

[Figure 32: splitting of dataset 43](#_bookmark68)

[Figure 33: code for LR model 44](#_bookmark71)

[Figure 34: code for KNN model 45](#_bookmark73)

[Figure 35: code for SVM model 46](#_bookmark75)

[Figure 36: code for gaussian naive bayes model 47](#_bookmark77)

[Figure 37: code for gradient boosting model 48](#_bookmark79)

[Figure 38: code for model comparison 48](#_bookmark81)

[Figure 39: code for prediction 49](#_bookmark83)

[Figure 40: dataset 51](#_bookmark85)

[Figure 41: comparison of different models 51](#_bookmark86)

# LIST OF TABLES

[Table 1: Functional Requirements 20](#_bookmark12)

[Table 2: Non Functional Requirement 20](#_bookmark13)

# ACRONYMS AND ABBREVIATIONS

|  |  |
| --- | --- |
| LR | Logistic Regression |
| SVM | Support Vector Machine |
| KNN | K Nearest Neighbor |
| EDA | Exploratory Data Analysis |
| NB | Naïve Bayes |
| GB | Gradient Boosting |
| FR | Functional requirement |
| NFR | Nonfunctional requirement |

**Chapter 1****INTRODUCTION**

# Project Description

Since the prime cause of death in India and today’s world is rising case of heart disease but application of proper machine learning techniques would be of great help to today’s society. Early detection of heart disease will be very helpful in the healthcare industry as well as in the area of cardiology. It would help patients also they can adjust their lifestyles accordingly. Every year a lot of people are dying due to heart disease. This can only be controlled by its early detection and immediate necessary treatment. The use proper technology will help both medical researchers as well as patients. Hence it is very imperative to create a method to precisely and effectively anticipate heart diseases given the rise in fatalities from cardiovascular causes. The aim of our study and this project was to identify an effective Machine Learning technique for heart disease identification. This study’s findings show that the gradient-boosting technique was more effective when compared to other algorithms [1].

# Motivation

The major objective for performing this research is to construct a system for detecting the heart disease. Three classification algorithms are applied at different levels of evaluations in a comparative study and analysis to corroborate this work. Though these machine learning algorithms are widely utilized. Consequently, a range of levels and evaluation strategy types are used to analyze the five algorithms namely logistic regression, KNN, SVM, Gaussian naive bayes and Gradient boosting. This will empower medical professional and scientist to build a better.

# Problem Statement

There are techniques which can predict heart disease, but it is costly or useless in estimating the chance of heart disease in a person. Early heart attack detection can decrease overall problems and number of deaths. However, it is impossible to accurate daily patient surveillance in all circumstances [2]. Since it takes more knowledge, time, and expertise, hours by such a doctor are not accessible. World has a lot of data; therefore, we could utilize a variety of machine developing

algorithms to seek for underlying patterns among the data. The underpinning patterns may be used in medical information for health diagnosis.

# Software Requirements Specifications

Operating system : Windows 11 Technology : Python 3.10

IDE : Colab

# CHAPTER 2 LITERATURE SURVEY

In 2020 Apurb Rajdhan explains that it is crucial to develop a system that really can identify heart diseases accurately and effectively considering the rise in fatalities caused by cardiovascular diseases. The UCI machine-learning repositories dataset is employed in this work to evaluate the prediction accuracy of the algorithms utilized in our project. The results of this research shows that the Random Forest algorithm was the most efficient algorithm for prediction of heart disease, with an accuracy score of 90.16% [3]. The work might well be upgraded by developing a web application dependent on the Random Forest algorithms and by utilizing a large - scale dataset compared to the one utilized in this research, which also will assist to medical professional in order to predict heart disease.

In 2021 Aadar Pandita explains that the Cleveland Heart Disease Dataset, which is found inside the UCI repository and has 303 instance and 76 attributes but still only 14 of these are utilized because of missing values, has been utilized by the majority of the researchers. By analyzing the relationship between various qualities and their effects on the precision of the models, method of selection of features can be used to decrease the number of attributes required to generate an precise model. Several studies in the field have discovered that KNN and neural networks are usually highly trustworthy in predicting heart disorders.

In 2021 Sudheer Akella explains that the most prevalent type of heart disease, coronary artery disease (CAD), has a variety of risk factor. Accurate diagnosis of CAD enables the management that would assist in minimizing undesirable clinical implications. The UC Irvine Cleveland dataset was utilized in a comparative analysis of six machine learning algorithms to predict the course of the illness. Six ML algorithms were trained before even being applied to the preprocessed dataset. On the basis of accuracy, recall, area under the curve and F1 score, ML algorithms were evaluated. The accuracy of the six ML models was determined to be more than 0.79, signifying good performance [4].

In 2018 Amin UI explains that A hybrid intelligent machine-learning-based predictive system was carried out in this research study for the identification of heart disease. On a dataset of Cleveland

heart disease, the system was examined. Three methods for selecting features were integrated with seven well-known classifiers, comprising logistic regression, ANN, KNN, NB, SVM, random forest, and DT. The crucial features were selected using Relief, LASSO and mRMR. The system's verification process used the K-fold cross-validation approach. Different evaluation measures were also employed to analyze classifier performance. The algorithms for feature selection select critical features that improve classifier performance in terms of accuracy, sensitivity, and responsiveness, MCC, and reduce algorithm computation times. The classifiers' highest accuracy was 89% while using logistic regression with 10-fold cross-validation.

In 2021 Aditya Khamparia explains that We bring up three methods that permitted comprehensive evaluation and provided positive outcomes. We arrived at the conclusion that somehow this analysis was better handled by machine learning methods. It was additionally determined that when a dataset is reviewed, statistical analysis is indeed necessary and that it needs to have a Gaussian distribution. Outlier identification is therefore crucial and is performed using the Isolation Forest methodology. When the dataset size is increased, deep learning can be employed along with a variety of other enhancements to produce more promising outcomes [5].

In 2020 Anna Karen explains that we suggested combining a chi-square (CHI) with PCA to increase machine learning systems' capability for predictions. Predicting if a patient has heart disease was indeed the classifier's primary goal. When system components must be taken account, using all capabilities is not feasible. Chi-square was employed to extract the features from the research that had physiological and anatomical significance, such as cholesterol, the maximal chest pain, heart type, ST anxiety features, and cardiac vessels. To analyse enormous amounts of information and discover the risk factors associated with different diseases, our method can be used in a wide range of real-world situations as well as in other medical diagnoses.

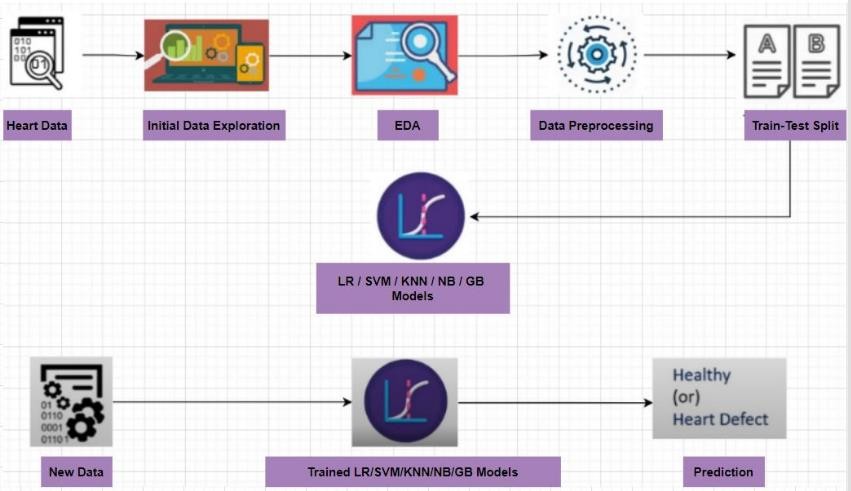
In 2020 Jian Ping explains that for the accurate determination of heart disease, a method that utilizes machine learning has been developed. The system is constructed using classifiers from machine learning, including such LR, SVM, KNN, ANN, DT, NB. In in addition to the four traditional attribute selection methods Relief, MRMR, LLBFS and LASSO. The algorithm determines the optimal hyperparameters using the LOSO cross-validation technique. Heart disease

dataset is utilized to assess the system's effectiveness. Furthermore, quality evaluation measures are used to measure how well the recognition system is working. Thus, the greatest predictive technique for determining optimal persons is ANN with alleviation. The responsiveness of the classifier NB on unique attributes specified by the LASSO FS algorithm likewise produces the greatest result (linear). The classification performance of the Logistic Regression MCC classifier that use the FCMIM FS method is 91% [6].

# CHAPTER 3

# SYSTEM ARCHITECTURE AND DESIGN

The architecture of the system will provide a glance of working of the entire system.



*Figure 1: system architecture*

The following is an explanation of how the system works:

This is a machine learning system which can predict heart disease in person. We collected heart data. The data set consists of several health parameters that correspond to the person's healthiness of the heart. Attribute selection process selects the important attributes for heart disease prediction. So, once we have that dataset, we need to process this data set because we cannot beat this raw data directly into machine learning algorithm therefore, we should process our data. Before splitting our data into two sets we will perform exploratory data analysis to have a more thorough insight in the data and have a clearer understanding of it. After processing it we will have to divide our data into two sets training data and testing data [7].

Once we do that we will feed our training data to our machine learning model in this case we will implement logistic regression model because this particular use case is a binary classification so here we are going to classify whether heart disease is their or not this is either yes or no condition kind of question and in this binary classification logistic regression model is very useful. We are going to perform this with other algorithms also like SVM, KNN, Naïve bayes, Gradient boosting. Finally, we will measure the accuracy of all our models that are being used and we will go the most accurate algorithm.

# Machine learning

In machine learning classification points to a predictive model problem in which we must find based upon the input data that the data belongs to which class label.

* + - Supervised Learning

In supervised learning, training data that is provided to the machines serves as the supervisors, guiding them about how to appropriately determine the output. It implements the very same idea that a learner would learn under a guidance of the teacher. The technique of supervised learning involves providing the machine learning models the relevant input data in addition to the output data [8] .

* + - Unsupervised Learning

We cannot directly apply unsupervised learning to a classification or regression. In this learning we train our machine learning algorithm with unlabeled data. Identifying the internal structure of a dataset, categorizing the data into categories based on their similarities, and summarizing the dataset in a compacted manner are the purposes of unsupervised learning.

* + - Finding meaningful insights from the information is made much easier with the assistance of unsupervised learning.
    - Unsupervised learning is much more like how an individual learns to understand via personal observation, which brings it closer to real-world artificial intelligence.
    - Unsupervised learning is more essential because it operates on uncategorized and unlabeled data.
    - Unsupervised learning is required to deal with circumstances where the input and output are not always identical in the real world.
    - Reinforcement Learning

Machine learning involves the study of reinforcement learning. It means acting effectively to maximize reward in a certain situation [9].

*Table 1: Functional Requirements*

|  |  |
| --- | --- |
| ID | REQUIREMENT |
| FR1.1 | A sample size that is the perfect reflection of  the population. |
| FR1.2 | Computer system compatible with python  module. |

*Table 2: Non Functional Requirement*

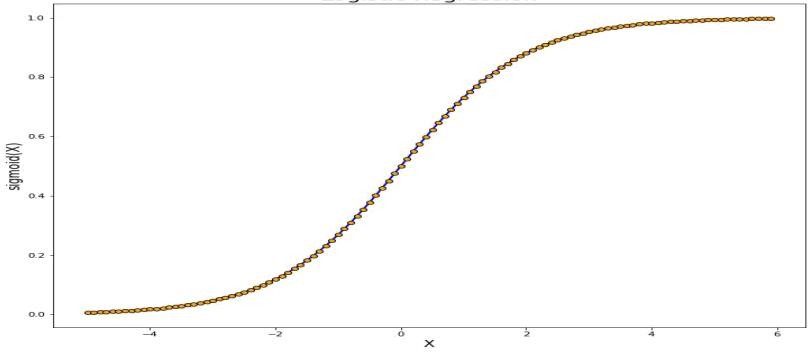
|  |  |
| --- | --- |
| ID | REQUIREMENT |
| NFR1.1 | The model of the research should be secure and  should not have any valuable data. |
| NFR1.2 | The model should be scalable, which means  that it must have some future spoke where it can be expanded. |
| NFR1.3 | The model should be reliable and continue  throughout the stipulated time period. |
| NFR1.4 | The model must be portable so that the research can be implemented on any system  anywhere. |

# Algorithms

## Logistic Regression Algorithm

A commonly used machine learning approach is logistic regression, algorithms, an element of the supervised learning process. Categorical dependent quantity is anticipated utilizing a combination

of factors. The outcome of dependent categorical variable is estimated via logistic regression. The major difference in logistic regression and linear regression is it is used. Problems involving regression are handled using linear regression, however, to overcome categorization issues, logistic regression is used. In logistic regression, we design a "S"-shaped curve instead of a regression line. By using the logistic function, two highest values are predicted (0 or 1). The logistic function's curve indicates the probability of such an event. Dependent on its weight, a mouse is either obese or not, depends on whether the cell is malignant and so on. Because it delivers the following benefits over all other machine learning algorithms. The capacity to evaluate fresh data using discrete and continuous methods and to calculate probability datasets.



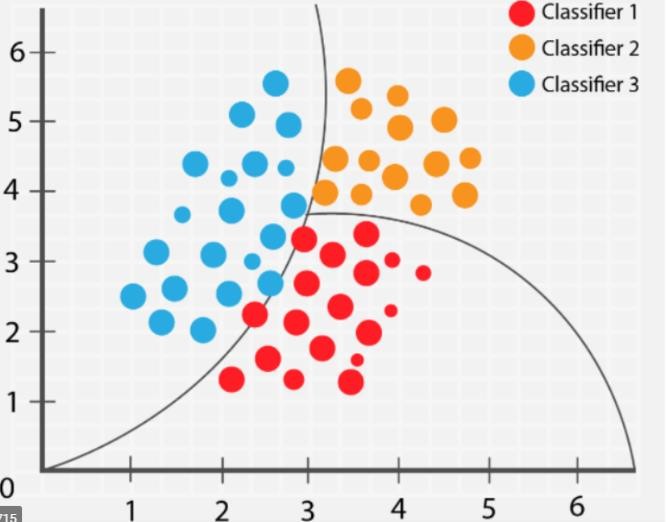
*Figure 2 : graph for logistic regression*

## Naive Bayes Algorithm

The Naive Bayes system is easy to implement and is exceptionally advantageous for the really big data sets. In addition to being uncomplicated, Naive Bayes works better when compared with most complex classification methods.

The naive bayes algorithm mainly comprises of two terms those are Naive and Bayes, which can be elaborated as:

* + - * Naive : As it presumes that the existence of one aspect is irrelevant to the prevalence of many other attributes, it is known as naive. For example, if a fruit is identified as an apple depending on its red, delicious and spherical fruit., flavour and form So, without dependent on one another, each attribute allows us to identify it as an apple.
      * Bayes : It is mainly dependent on the concept of the bayes theorem [10].



*Figure 3: naive bayes classifier*

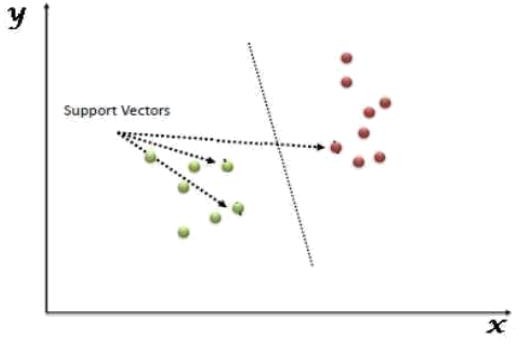
## Support Vector Machine (SVM)

SVM algorithm is used in order to deal with regression as well as classification issues. Important terminologies in SVM are as follows:

Support vectors - those data points which are very near to the hyperplane is known as support vectors.

Hyperplane - Hyperplane is nothing but a decision boundary which separates the objects to different classes.

Margin - it is a gap between the two lines that is drawn on the nearest data points in different classes [11].



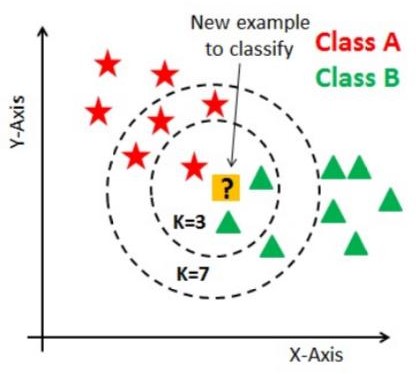
*Figure 4: support vector machine*

## K Nearest Neigbour Algorithm

It is a data classification strategy that evaluates that a data point will join one group or another group based upon which group the data points that really are nearest to it are a member of.

Advantages and disadvantages of KNN algorithm are as follows:

* + - * It is straightforward to put into practice.
      * It can tolerate chaotic training data.
      * When there is a lot of training data, it could work much better.
      * K's value must be constantly determined, and sometimes this can be difficult.
      * The powerful computational cost is caused due to the requirement to determine the difference between every data point for every training sample [12].



*Figure 5: KNN algorithm*

## Gradient Boosting Algorithm

Boosting is a strategy for transforming children with learning difficulties into strong ones. Each new tree in boosting is indeed a fit on an alternate version of the original data set. It is anticipated that when integrated with earlier models, the latest model will produce estimates with reduced error rates. To minimize errors, the basic purpose of this future model is to establish target outcomes. Gradient boosting is a technique for gradually, sequentially, and progressively training various models. Because each case's target results are decided by the gradient's inaccuracy compared to the predictions, the term "gradient boosting" comes into use. Each model builds momentum in the right way by minimizing prediction errors.

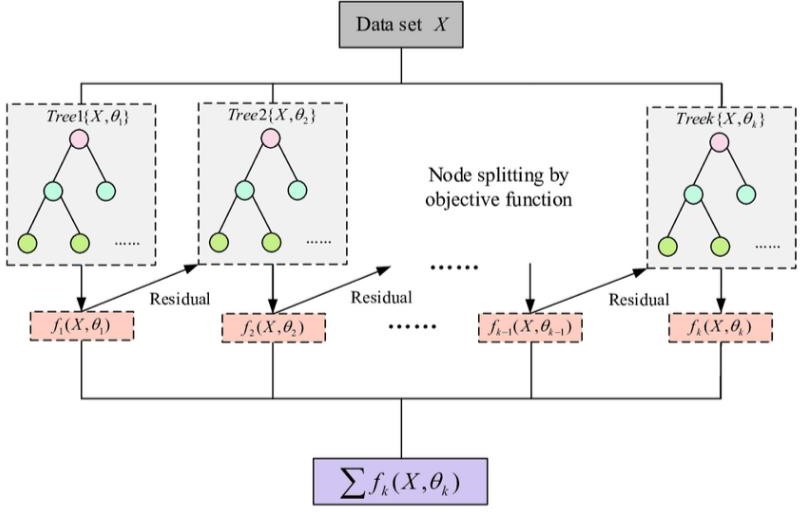
Let's suppose that we have a training data (D). In original training data set let's suppose we have n instances or samples. To each sample we attach their respective weights. initially we take into consideration that all the weights that are attached to their respective samples are equal. Dataset (D1 which is subset of D) Which is going to be used to train our model. In this D1 dataset whatever instances that is going to come from this D dataset will all have equal probability to get selected

into D1 [13]. After model M1 is trained with the help of D1 dataset then we will input all the instances of D to model M1 for its testing. Then M1 is going to tell us that 1,2, 3...n belongs to which class that is either 0 or 1 but there might be some misclassification as well. Now we must train our next model that is M2. Before training it, we need to keep in mind that we have to random sampling in D2 dataset but before random sampling we must update the weights. While doing this we have to focus on those samples weights which were misclassified in the previous model. Increasing the weights implies that probability for those weights will be high to get selected into D2 dataset. This process will be continued until all the model is trained then we will combine all model (weak models) to generate strong model. This model will have higher accuracy, less error when compared to weak models.

Regularization by shrinking is an important element of gradient boosting. The updating changes to the rules due to shrinkage. Put simply, the updating rule is a learning rate. As per observations, generalizing the prediction system is important if learning rates are lower than 0.1. Even though shrinkage is not available, gradient boosting still can drastically enhance if the learning rate is comparable to one. But doing this elongates the computation time. Higher prerequisites for the number of repetitions are present whenever the learning rate is very low.

Below are some of the advantages and disadvantages of gradient boosting algorithm:

* + - * Most of them support managing categorical attributes, a few of them handle missing data natively, and train quicker on bigger dataset.
      * sensitive to overfitting
      * while using L1 and L2 regularization penalties, this could be fixed.
      * We can also explore with a low learning rate; but the model results may be difficult to comprehend and might be operationally time-consuming and costly to develop, specifically on CPUs [14].



*Figure 6: boosting algorithm*

# CHAPTER 4 METHODOLOGY

# Existing System

Heart disease is a health hazard that causes an individual to leave this world without showing any visible symptoms. Increasing concerns regarding the disease and consequences is an outcome of nature of disease. To meet the demands of contemporary health, new tools and techniques are often tested . Regardless of the fact that disease can take many different forms, there are a few common core risk factors that influence a person's eventual chance of developing heart disease. The well- known proverb says Early prognosis & its control can help with the adage that "Prevention is better than cure." Prevent and reduce the number of heart disease-related deaths [15].

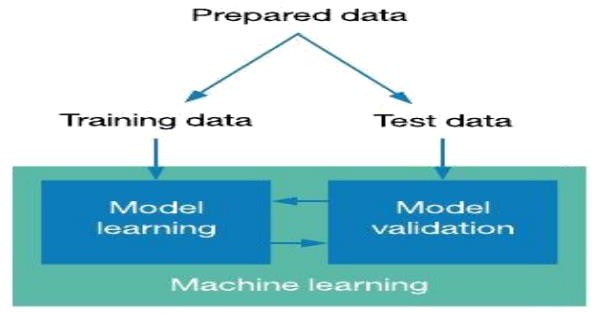
# Proposed System

The working of the system begins with dataset collection. We need to collect a dataset that contains attributes that correspond to the person’s healthiness of the heart. Then we have chosen all the important attributes. After this, we performed exploratory data analysis to have a more thorough insight into the data and a better understanding. Then we have done data preprocessing where we need to process the data before directly feeding it to our prediction model and then we will split the dataset. After this, we will implement our model with five algorithms like logistic regression, SVM, KNN, NB, and, gradient boosting. This model is implemented using the following steps

1. data set collection
2. attribute selection
3. exploratory data analysis
4. data preprocessing
5. prediction of heart disease

## Dataset collection

First we have to collect the data set for our prediction system then after the collecting data set we have to split our data set that we have collected into two sets. For this project we have taken 80% test training data and 20% test testing data the dataset utilized in this project is UCI dataset. The

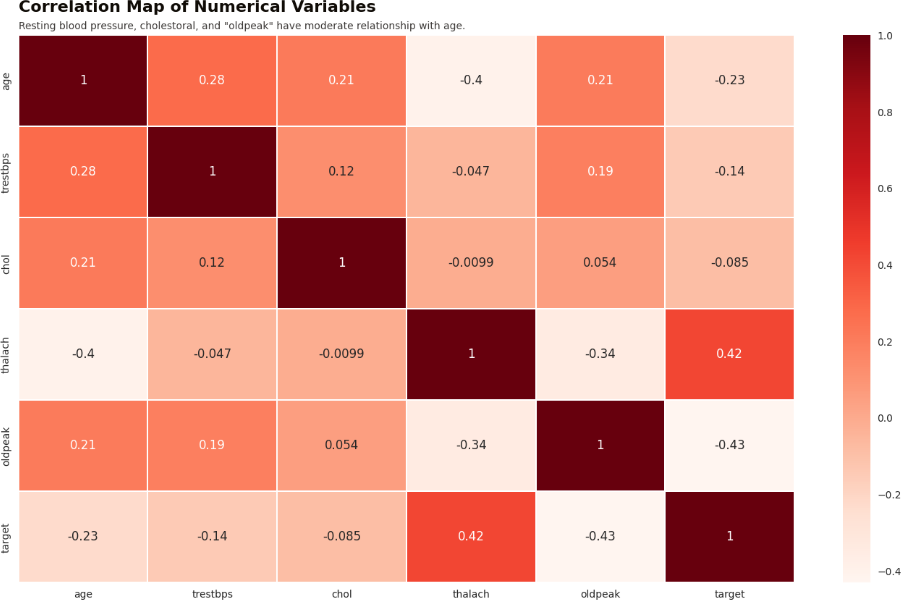
datasets that we have collected totally consist of 14 columns and 303 rows it consists of 13 attributes that is related to the person healthiness of the heart and last column is the target column which consists of 1 and 0 where one depicts that the person has heart disease and zero defects that the person does not have a heart disease [16].

*Figure 7: collection of data*

## Attributes selection

Several attributes of the patient are selected that corresponds to the person healthiness of the heart. The correlation matrix is used for the selection of the attribute for this model.

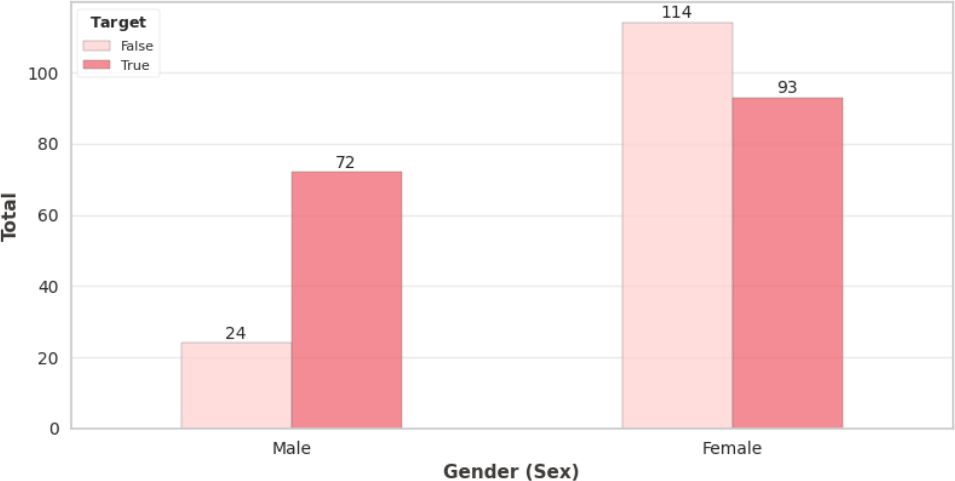
Below is the correlation map or the heat map of the numerical variables in order to show the correlation level or the values for each of the variables with other variables.



*Figure 8: correlation matrix*

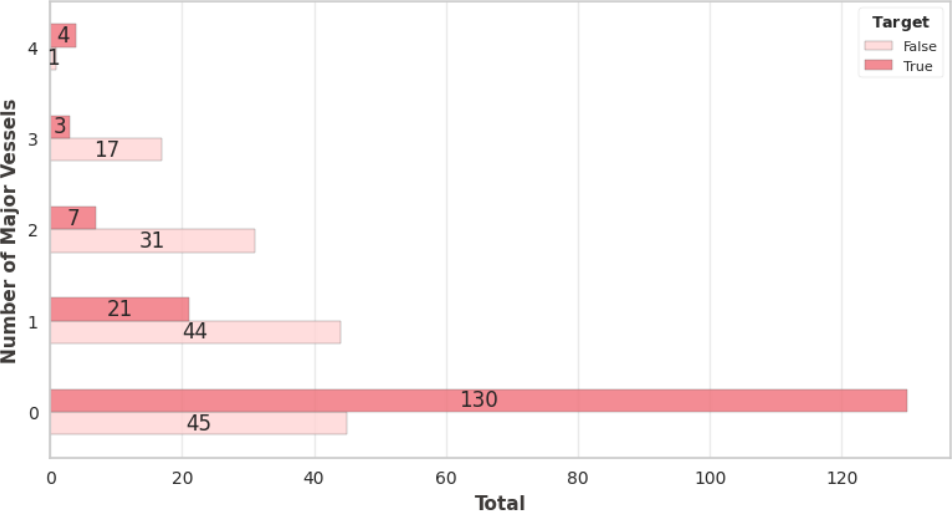
## Exploratory data analysis

After we have collected the information on the data set now we are going to implement the exploratory data analysis EDA this will help us to have a more thorough insight about the data that we have collected and through this we have can also gain a proper understanding of the variables and the relationship among the variables and this section is divided among six parts [17].

a)

*Figure 9: distribution of heart disease on gender*

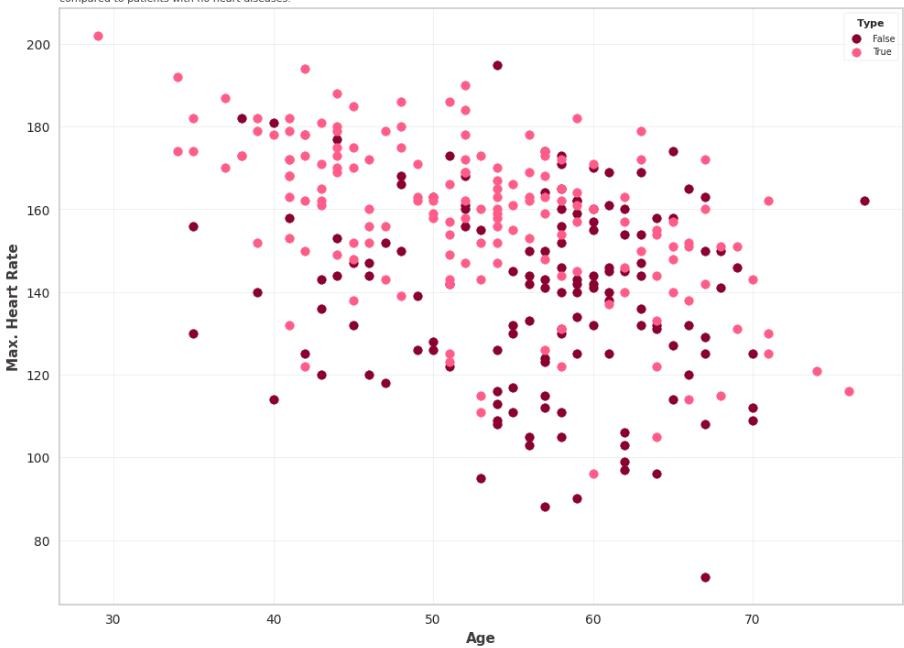
* + - 1. heart disease distribution based on major vessels



*Figure 10: distribution of heart disease on major vessels*

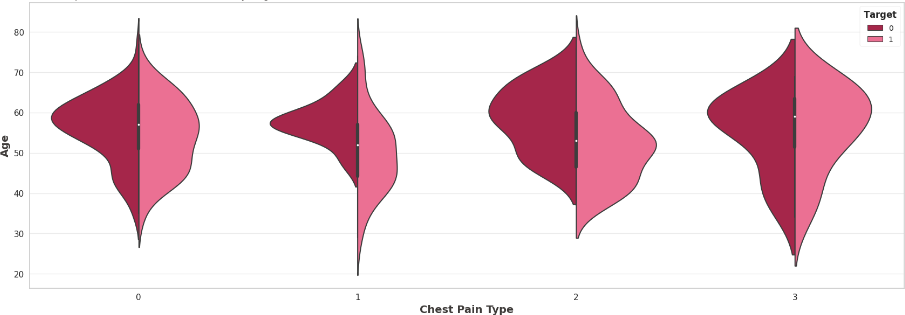
from the above graph we can see that the patients with 0 and 4 major vessels tends to have heart diseases however patients who have a number of vessels 1 to 3 tend not to have heart disease.

* + - 1. heart disease scatterplot based on age - this section will show a space scatter plot of the patients with heart disease based upon the age



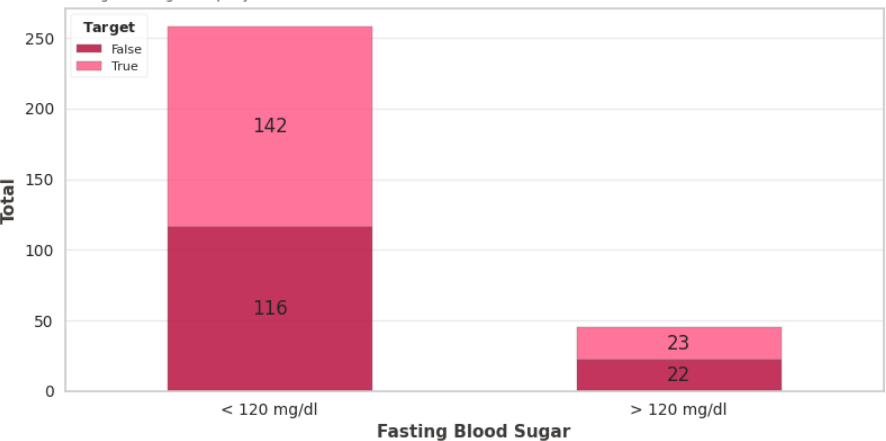
*Figure 11: heart disease scatter plot based on age*

* + - 1. Chest pain type based on age - this section will show us voilin plot about age distribution based on the chest pain type [18].



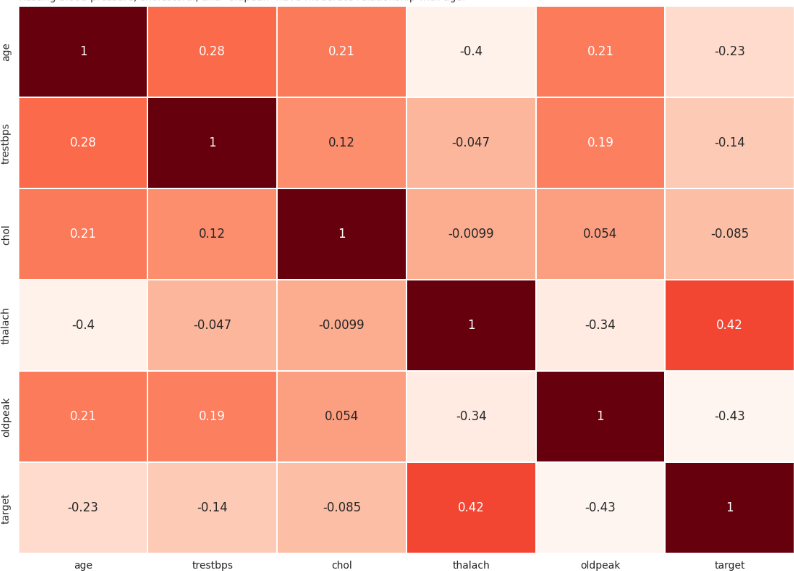
*Figure 12: chest pain type distribution based on age*

* + - 1. Heart disease distribution based on fasting blood sugar - this section will show us the heart disease distribution on stacked bar chart based on fasting blood sugar.



*Figure 13: distribution of heart disease based on blood sugar*

* + - 1. Heatmap - this is the correlation map or the heat map of the numerical variables to show the correlation level or values of each variables with others.



*Figure 14: correlation map of numerical variables*

## Data Preprocessing

We need to process our data in order to make it fit and compatible for our machine learning model because we cannot put directly the raw data into our machine learning algorithm it is used to deal with noises values and duplicates of the datasets and attributing scaling. the processing of the data is health required for enhancing the accuracy and efficiency as well as productivity of the model [19].

In our model we are going to perform four steps in data preprocessing. First step is one hot encoding. In this step we will transform the categorical variables and then we will do one hot encoding as part of the data preprocessing. Second step is dropping unnecessary variable now as there are some unnecessary variable in our data frame these variables will be dropped. Third is feature separating in this section the dependent column will be separated from the independent columns that is the target column will be separated. Fourth is Data normalization in this section data normalization will be performed [20].

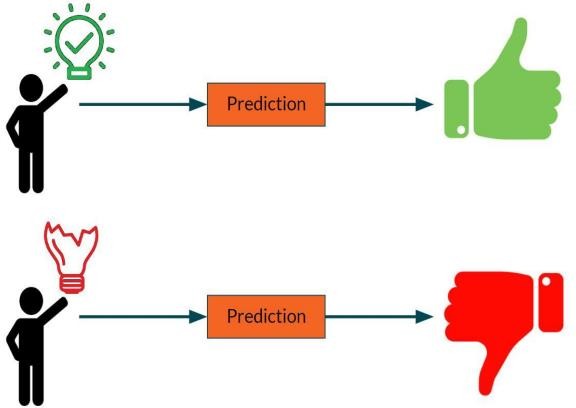


*Figure 15: data preprocessing*

## Prediction of heart disease

Several machine learning techniques are implemented like Logistic regression, SVM, KNN, Gaussian Naive Bayes, Gradient boosting [21]. Comparison between the algorithm is also

performed and the technique with highest accuracy and precision will be taken into consideration.

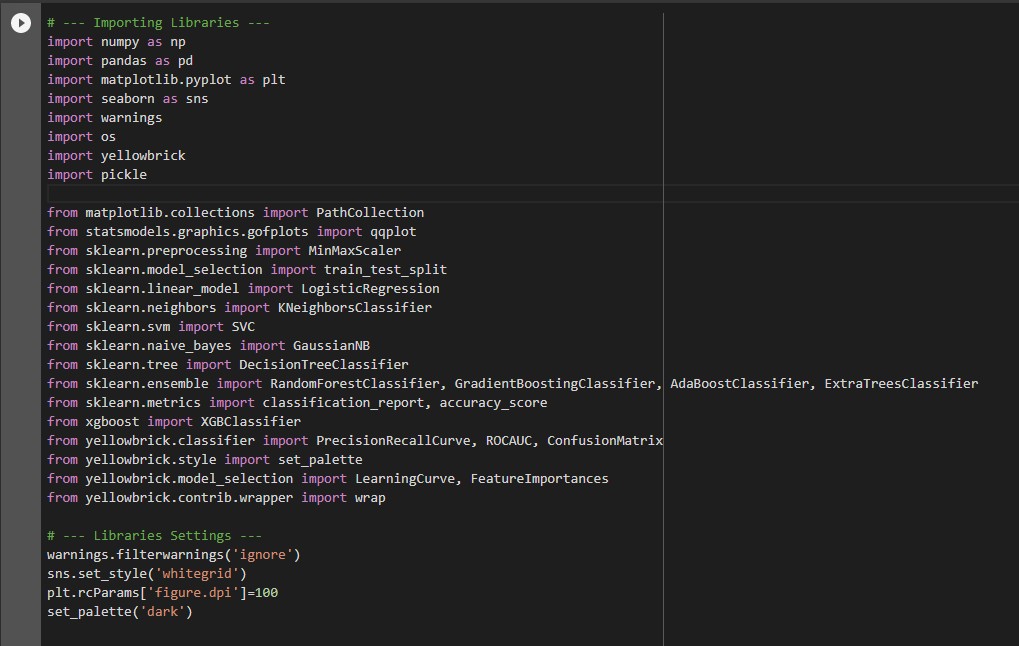


*Figure 16: prediction of disease*

# CHAPTER 5 CODING AND TESTING

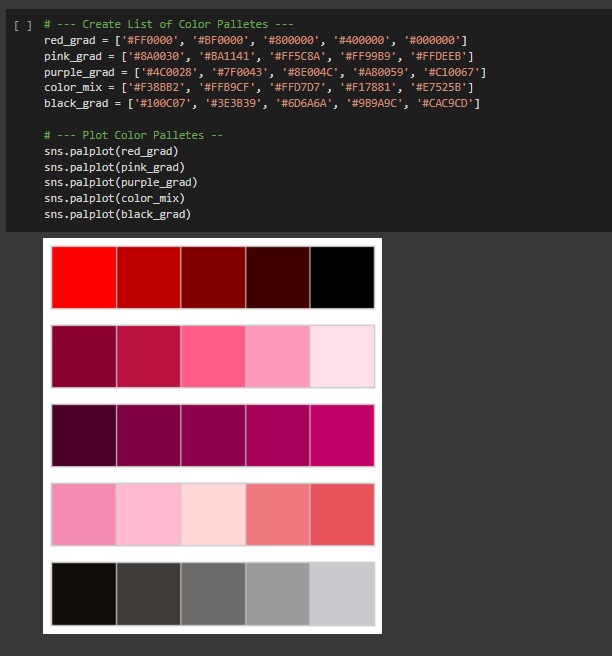
# Importing the Libraries

Here we are importing the libraries which are necessarily needed for the implementation of the models so here we imported libraries like pandas, numpy, yellowbrick, seaborn and the other algorithm related libraries.



*Figure 17: importing library*

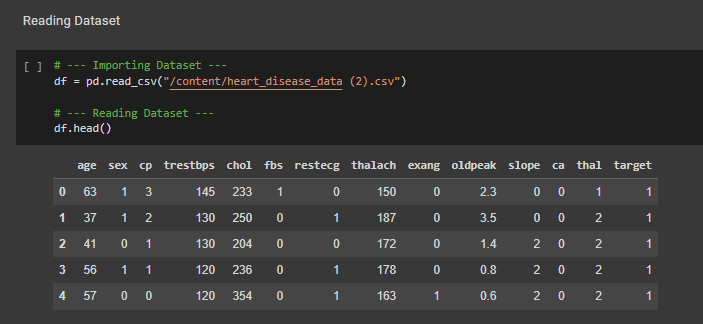
Now we will create the color palates which helps us assisting with the colors furtherly



*Figure 18: creation of color palattes*

# Importing the Dataset

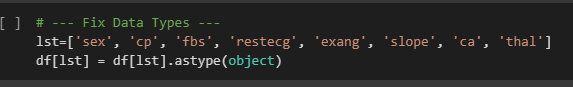
Now we will import the dataset using the below code using pandas through read function and we check the head of the dataset that is the first 5 columns of the dataset to check whether we got the dataset imported properly or not. Now we will print the information of the dataset using info function which gives us the complete details of the dataset that is the tota number of rows the total number of columns, the attributes that are present in dataset, the datatypes of each of the attribute values and the null values and the non null values [22].



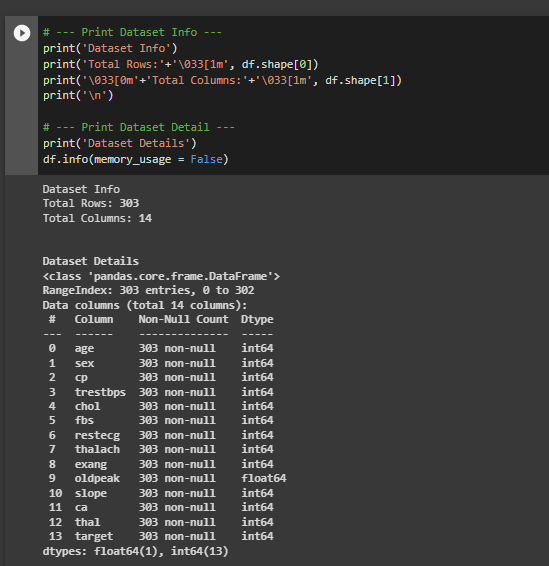
*Figure 19: importing dataset*

# Fixing the data types

In this section for few of the attributes we need to change the data type using astype(object) function we will be changing the datatypes of the following attributes.



*Figure 20: datatype fixing*

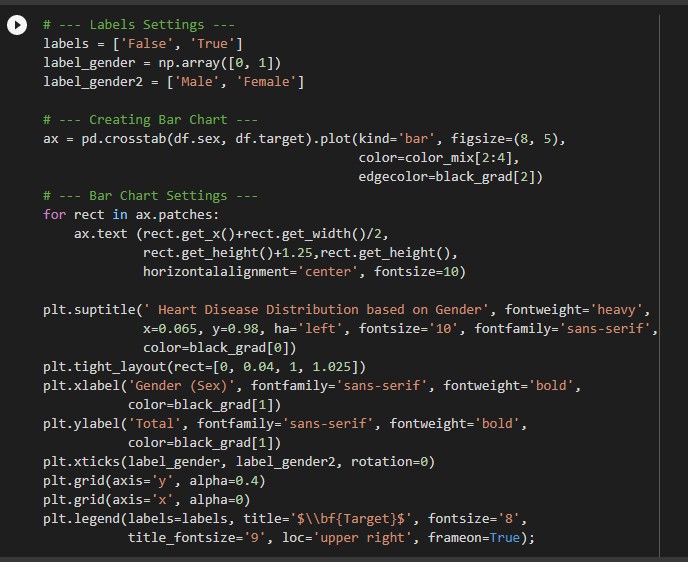


*Figure 21: dataset information and detail*

# Exploratory data analysis

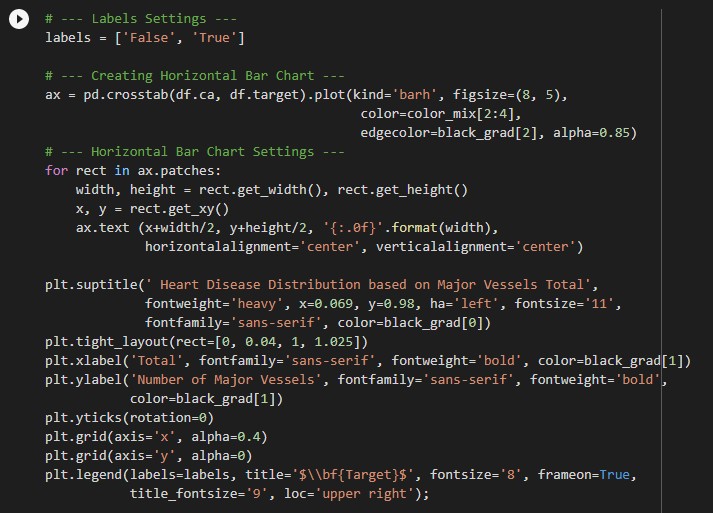
This section will perform some EDA to get more insights about dataset. Here we perform the data analysis to know better insights of the attributes which are having major impact on the heart disease cause. We checked with each attribute like the heart disease with gender and major blood vesssels. By making use of the pandas and plot from matplot library we have drawn the bar graph for the heart disease vs gender.

* + 1. Heart Disease Distribution based on Gender - This section will show heart disease distribution based on gender.



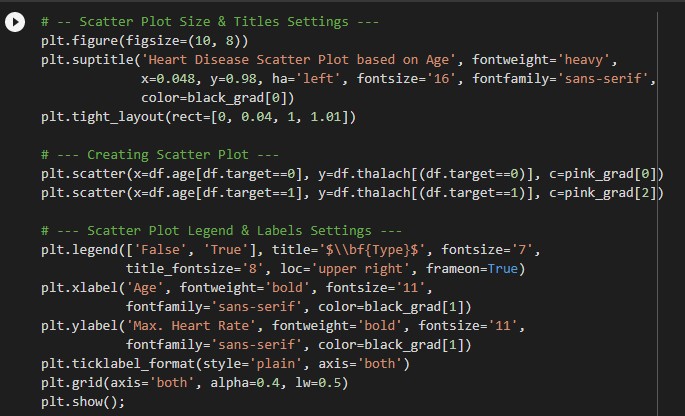
*Figure 22: code part A*

* + 1. Heart Disease Distribution based on Major Vessels Total



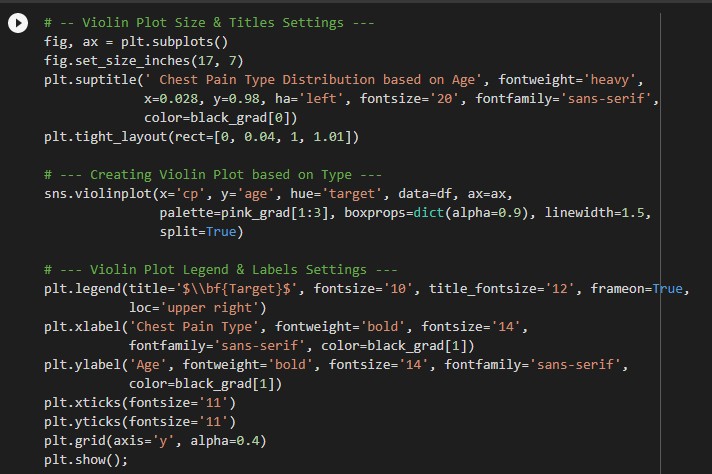
*Figure 23: code part B*

* + 1. Heart Disease Scatter Plot based on Age - This section will show scatter plot of patients with heart diseases based on age.



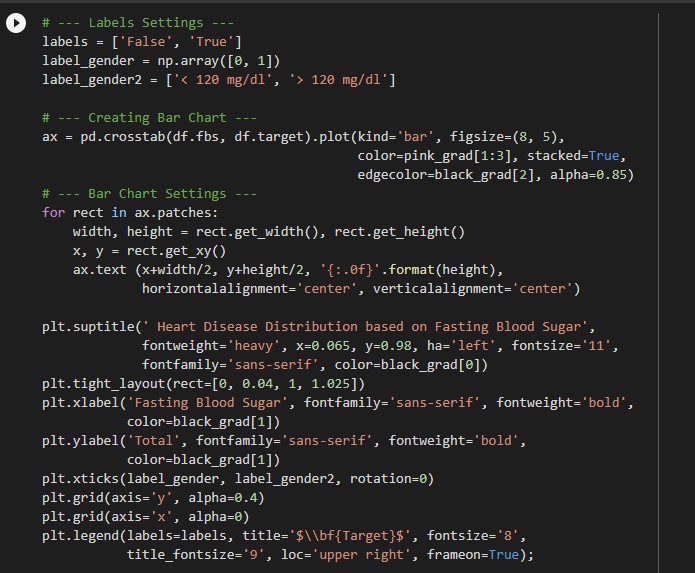
*Figure 24: code part C*

* + 1. Chest Pain Type based on Age - This section will show violin plot about age distribution based on chest pain type.



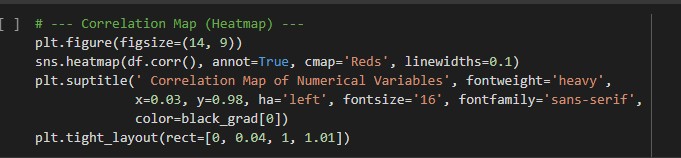
*Figure 25: code part D*

* + 1. Heart Disease Distribution based on Fasting Blood Sugar – This shows heart disease distribution on stacked bar chart based on fasting blood sugar.



*Figure 26: code part E*

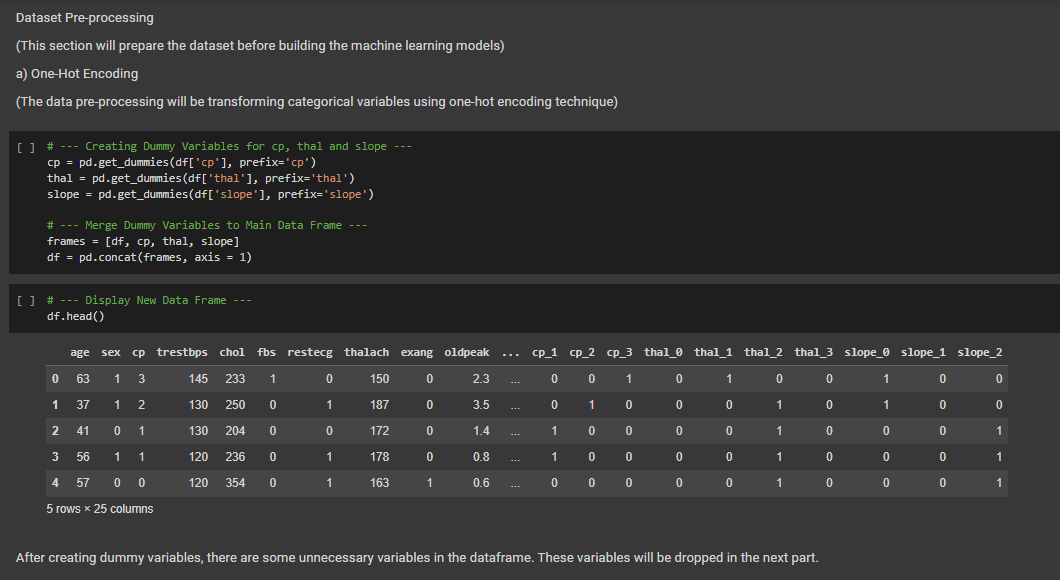
* + 1. Heatmap - This section is correlation map/heatmap of numerical variables to show correlation level/values for each variables with others.



*Figure 27: code part F*

# Creating dummy variables

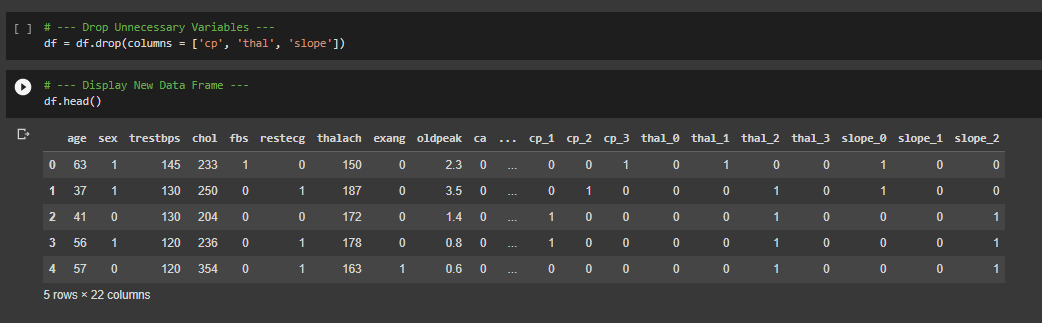
Now we will create dummies for chest pain type, slope and thal because cp has only three types similarly slope and thal. After creating dummy variables, there are some unnecessary variables in the dataframe. These variables will be dropped in the next part [23].



*Figure 28: creation of dummy variables*

# Dropping unnecessary Variables

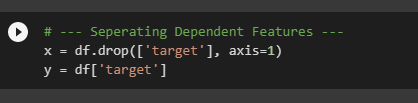
Using drop function we will drop the unwanted varaiables and we will display the dataset after removing the unnecessary variables.



# Feature Separating

*Figure 29: code for dropping variables*

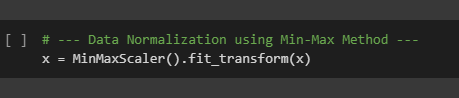
In this section, the 'target' (dependent) column will be seperated from independent columns Here we are dropping the target variable from the dataset [24].



# Data Normalization

*Figure 30: code for feature separating*

In this section, data normalization will be performed to normalize range of the variables which are independent. Normalization of data will use min-max normalization.

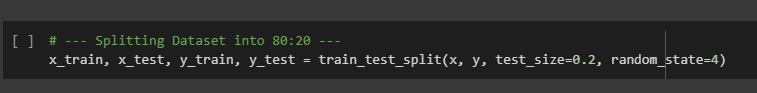


*Figure 31: normalization of data*

# Splitting the dataset

The dataset will be splitted into 80:20 ratio (80% training and 20% testing).

Using train\_test\_split we split the dataset into the above mentioned ratio which will be best suited for the prediction.



*Figure 32: splitting of dataset*

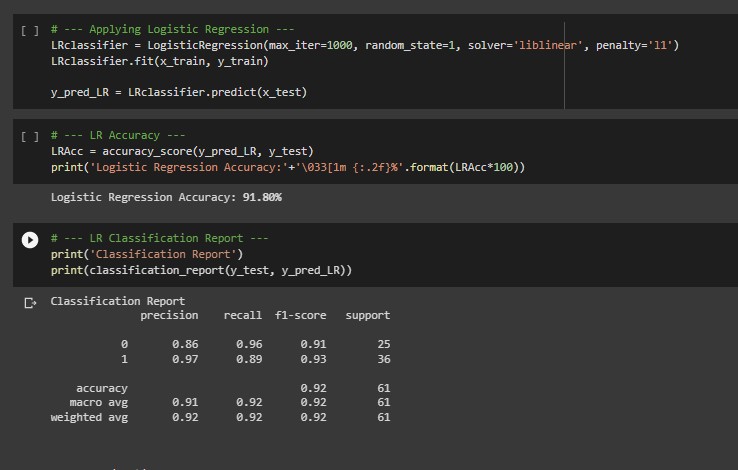
# Model Implementation

## Logistic regression Model

We are performing logistic regression algorithm to predict.

In this module, we create the LR classifier using the Logistic regression function in which we pass max\_iter, solver, penalty as tuning parameters and solver is an algorithm takes care of optimizing the coefficients and penalty is the type of regularization we used L1 regularization which basically reduces the dimensionality. We have to set the random state for reproducibility.

Model is learning the relationship between (x\_train) and (y\_train) through fit function. In LRclassifier variable we are importing model and then we are using fit function. We are passing training data to train the model and in y\_pred\_LR variable we are storing predictions of test data then in lRAcc variable we are storing accuracy of the model then printing it.

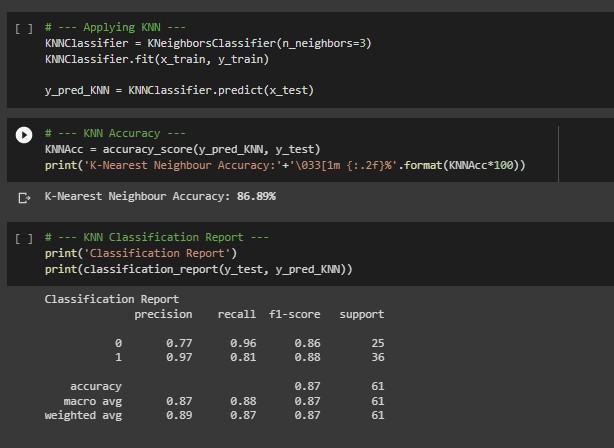


*Figure 33: code for LR model*

## K Nearest Neighbor Algorithm

We create an instance for the KNN Classifier with the name KneighborsClassifier which finds the K-neighbors of a point by passing a single parameter named n\_neighbors where n\_neighbors is an integer value where we assigned 3 which means 3 neighborhood points are required for classifying a given point as this value indicates the number of neighbors required for sample. This value is passed to the constructor.

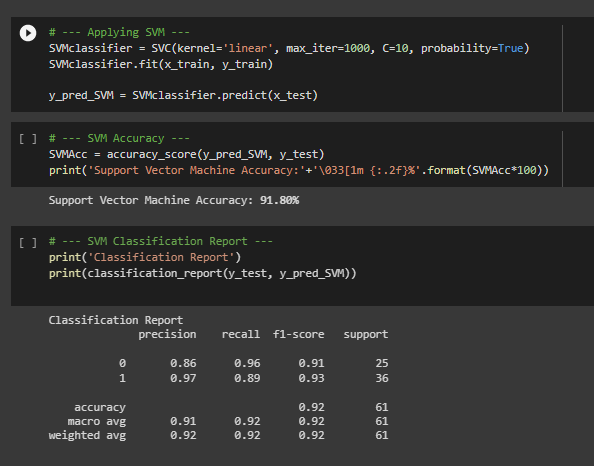
fit(x\_train, y\_train) this function fits the k-nearest neighbors classifier from the training dataset. Predict(x\_test) this function predicts class labels for the data.



*Figure 34: code for KNN model*

## Support Vector Machine(SVM) Model

In this module we implemented the Support Vector Machine Model. We created the SVM Classifier object by creating an instance called SVC which is a function passing the parameters like kernel, max\_iter, C, probability. Here we have given linear as kernel type because the data is linearly seperable which can be separated using a single line and also this is the most used kernel type and best suits for the large datasets having more number of features. C is regularization parameter. We gave 10 as the value which tells the SVM optimization.

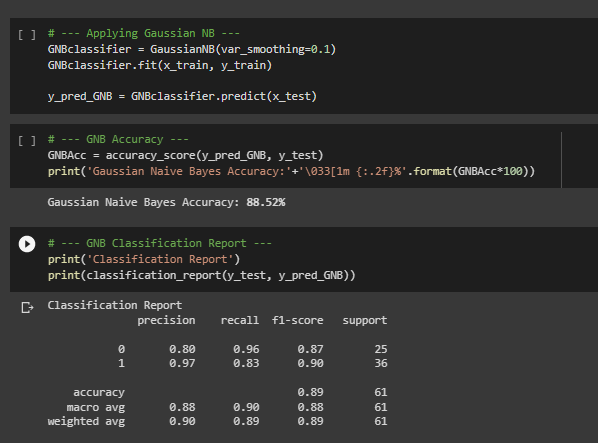


*Figure 35: code for SVM model*

## Gaussian Naïve Bayes Model:

In this module we implemented the gaussian naïve bayes classifier which is well suited for the numerical data. It is trained very efficiently these works very well with small training data. This classifier have a very simple design and hence can be applied to our situation [25].

First we create an instance for the GNBClassifier object with GaussianNB as a function passing the parameter var\_smoothing which basically a stability calculation to widen or smoothen the curve and therefore which will account for more samples that are further away from the distribution mean. This helps to set the selected parameters used to find the optimal combination. This parameter is basically a float value. The fit function is used to fit Gaussian Naïve bayes according to the x\_train, y\_train data.



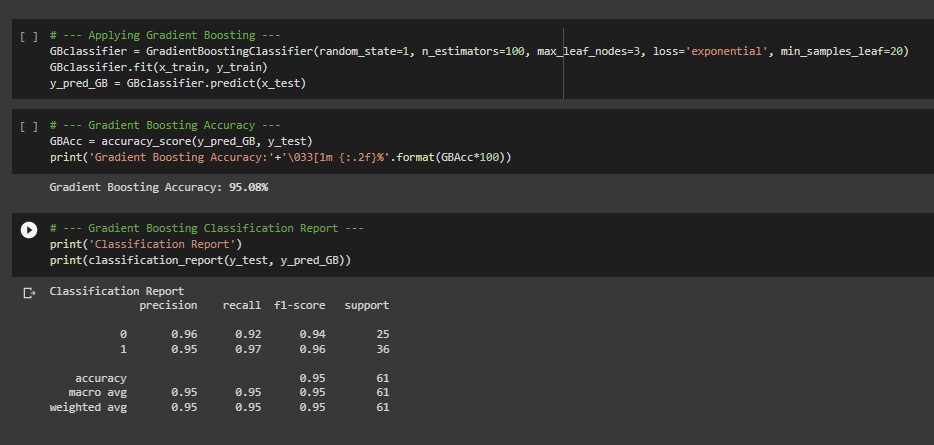
*Figure 36: code for gaussian naive bayes model*

## Gradient boosting Model:

In this module we implemented the gradient boosting algorithm. Here we create a function called Gradient Boosting Classifier with parameters n\_estimators, max\_leaf\_nodes, random\_state, loss, min\_samples\_leaf where n\_estimators is an integer value

Max\_leaf\_nodes is also a integer value which should be minimum 2 we gave 3 as its value which indicates the maximum number of terminal nodes or leaves in a tree it would produce a depth of

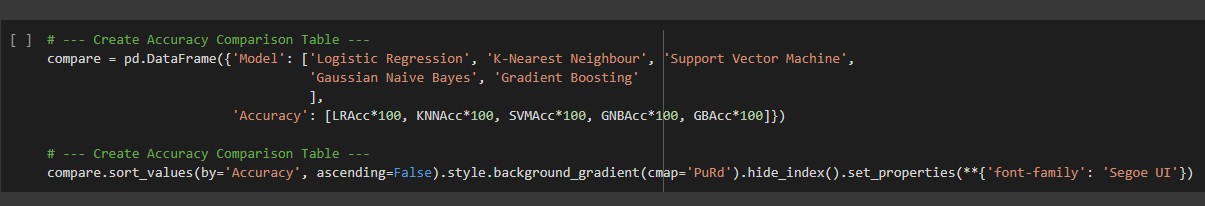
8. min\_sample\_leaf defines the minimum samples or observations required in a terminal node or leaf and this controls the over-fitting problem generally lower values should be choosen we gave 20.



*Figure 37: code for gradient boosting model*

## Creating Accuracy comparison table

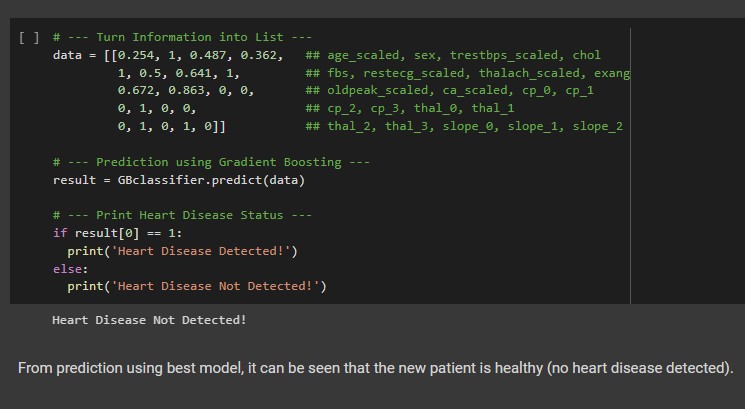
Using compare we will compare the accuracies of all the five algorithms by using sort to sort the values of the accuracies in the required order.



*Figure 38: code for model comparison*

## Predicting case

Predicting the presence of disease taking the values of all the attributes by inputting the values and choosing the gradient boosting which is obviously having the highest accuracy which can be seen from the comparison table.



*Figure 39: code for prediction*

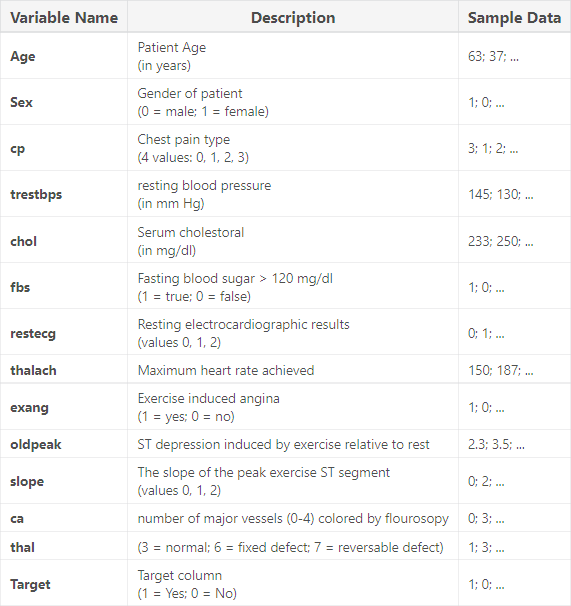
# CHAPTER 6

# Results and discussions

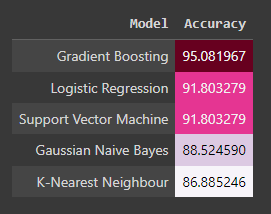
After collection of dataset we have applied exploratory data analysis technique in order to have more thorough understanding of the data and get a more clear picture of it. In this project we have implemented our model with five different techniques like LR, KNN, NB, SVM, GB. All the attributes in the dataset relates to the person healthiness of the heart. By implementing EDA we have obtained the following results

* + - * Female distribution patients are higher when compared to that of with the males.
      * Chest pain type which has the highest number is 0 when compared to other types of chest pain.
      * It can be seen that number of patients with blood pressure less than 120 mg/dl achieves the highest number.
      * Result 1 in restecg has the highest distribution when compared to others.
      * The distribution of slope 1 and 2 is almost the same.
      * Patients having no exercised induced angina are highest when compared to patients who has exercise induced angina.

We have done comparison of all the five algorithms that we have implemented in our project and found that the accuracy percentage for logistic regression is 91.80, for k nearest neighbor algorithm accuracy percentage was 86.88, whereas accuracy percentage for support vector machine, naïve bayes and gradient boosting were 91.80, 88.52 and 95.08 respectively [26]. We can see that gradient boosting algorithm was very accurate and precise when compared to other algorithms.



*Figure 40: dataset*



*Figure 41: comparison of different models*

# CHAPTER 7

# Conclusion and Future enhancements

Since the major causes of death in India and today’s world is the rising case of heart disease and application of proper machine learning techniques would be of great help to today’s society. Early detection of heart disease plays a very crucial role in the healthcare industry as well as in the area of cardiology. It would help patients also they can adjust their lifestyles accordingly. Every year a lot of people are dying due to heart disease. This can only be controlled by its early detection and immediate necessary treatment. The use proper technology will help both medical researchers as well as patients. Finally, we can draw the conclusion that real-time predictors are now crucial in the healthcare industry.

With the help of this project, we will be able to accurately forecast heart risks using machine learning in real-time using patient data from the model created using the Gradient Boosting Algorithm. Hence it is very imperative to create a method to precisely and effectively anticipate heart diseases given the rise in fatalities from cardiovascular causes. The aim of our study and this project was to identify the most effective Machine Learning algorithm for heart disease identification. This study’s findings show that the gradient-boosting technique was more effective when compared to other algorithms.

Future work - A web application built utilizing the Gradient boosting technique and using larger datasets can be developed in the future to improve the work. And also may be considering other factors which are impacting human health majorly the heart like the smoking statistics the aggregate of cigarettes a person smokes every day and the aggregate of people who are becoming fatal because of smoking or the approach combining the datasets may be few important attributes and also making the dataset on our own by doing the proper research and going to hospitals and medical institutions or the doctors and the professors who did research on these and getting the data to make it more realistic and helps in increasing the accuracy of our models.

# REFERENCES

1. Rajdhan, Apurb, Avi Agarwal, Milan Sai, Dundigalla Ravi, and Poonam Ghuli. "Heart disease prediction using machine learning." International Journal of Research and Technology 9, no. 04 (2020): 659-662.
2. Shah, Devansh, Samir Patel, and Santosh Kumar Bharti. "Heart disease prediction using machine learning techniques." SN Computer Science 1, no. 6 (2020): 1-6.
3. Jindal, Harshit, Sarthak Agrawal, Rishabh Khera, Rachna Jain, and Preeti Nagrath. "Heart disease prediction using machine learning algorithms." In IOP conference series: materials science and engineering, vol. 1022, no. 1, p. 012072. IOP Publishing, 2021.
4. Pandita, Aadar, Sarita Yadav, Siddharth Vashisht, and Aryan Tyagi. "Review Paper on Prediction of Heart Disease using Machine Learning Algorithms." International Journal for Research in Applied Science and Engineering Technology 9, no. 6 (2021).
5. Akella, Aravind, and Sudheer Akella. "Machine learning algorithms for predicting coronary artery disease: efforts toward an open source solution." Future science OA 7, no. 6 (2021): FSO698.
6. Haq, Amin Ul, Jian Ping Li, Muhammad Hammad Memon, Shah Nazir, and Ruinan Sun. "A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms." Mobile Information Systems 2018 (2018).
7. Bharti, Rohit, Aditya Khamparia, Mohammad Shabaz, Gaurav Dhiman, Sagar Pande, and Parneet Singh. "Prediction of heart disease using a combination of machine learning and deep learning." Computational intelligence and neuroscience 2021 (2021).
8. Sharma, Himanshu, and M. A. Rizvi. "Prediction of heart disease using machine learning algorithms: A survey." International Journal on Recent and Innovation Trends in Computing and Communication 5, no. 8 (2017): 99-104.
9. Gárate-Escamila, Anna Karen, Amir Hajjam El Hassani, and Emmanuel Andrès. "Classification models for heart disease prediction using feature selection and PCA." Informatics in Medicine Unlocked 19 (2020): 100330.
10. Li, Jian Ping, Amin Ul Haq, Salah Ud Din, Jalaluddin Khan, Asif Khan, and Abdus Saboor. "Heart disease identification method using machine learning classification in e- healthcare." IEEE Access 8 (2020): 107562-107582.
11. Malav, Amita, Kalyani Kadam, and Pooja Kamat. "Prediction of heart disease using k-means and artificial neural network as hybrid approach to improve accuracy." International Journal of Engineering and Technology 9, no. 4 (2017): 3081-3085.
12. Ali, Farman, Shaker El-Sappagh, SM Riazul Islam, Daehan Kwak, Amjad Ali, Muhammad Imran, and Kyung-Sup Kwak. "A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion." Information Fusion 63 (2020): 208-222.
13. Repaka, Anjan Nikhil, Sai Deepak Ravikanti, and Ramya G. Franklin. "Design and implementing heart disease prediction using naives Bayesian." In 2019 3rd International conference on trends in electronics and informatics (ICOEI), pp. 292-297. IEEE, 2019.
14. Safdar, Saima, Saad Zafar, Nadeem Zafar, and Naurin Farooq Khan. "Machine learning based decision support systems (DSS) for heart disease diagnosis: a review." Artificial Intelligence Review 50, no. 4 (2018): 597-623.
15. Saxena, Kanak, and Richa Sharma. "Efficient heart disease prediction system." Procedia Computer Science 85 (2016): 962-969.
16. Marimuthu, M., M. Abinaya, K. S. Hariesh, K. Madhankumar, and V. Pavithra. "A review on heart disease prediction using machine learning and data analytics approach." International Journal of Computer Applications 181, no. 18 (2018): 20-25.
17. Yahaya, Lamido, N. David Oye, and Etemi Joshua Garba. "A comprehensive review on heart disease prediction using data mining and machine learning techniques." American Journal of Artificial Intelligence 4, no. 1 (2020): 20-29.
18. Singh, Archana, and Rakesh Kumar. "Heart disease prediction using machine learning algorithms." In 2020 international conference on electrical and electronics engineering (ICE3), pp. 452-457. IEEE, 2020.
19. Aljanabi, M., Mahmoud H. Qutqut, and Mohammad Hijjawi. "Machine learning classification techniques for heart disease prediction: a review." International Journal of Engineering & Technology 7, no. 4 (2018): 5373-5379.
20. Ghosh, Pronab, Sami Azam, Mirjam Jonkman, Asif Karim, FM Javed Mehedi Shamrat, Eva Ignatious, Shahana Shultana, Abhijith Reddy Beeravolu, and Friso De Boer. "Efficient prediction of cardiovascular disease using machine learning algorithms with relief and LASSO feature selection techniques." IEEE Access 9 (2021): 19304-19326.
21. Dinesh, Kumar G., K. Arumugaraj, Kumar D. Santhosh, and V. Mareeswari. "Prediction of cardiovascular disease using machine learning algorithms." In 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT), pp. 1-7. IEEE, 2018.
22. Gupta, Akansh, Lokesh Kumar, Rachna Jain, and Preeti Nagrath. "Heart disease prediction using classification (naive bayes)." In Proceedings of First International Conference on Computing, Communications, and Cyber-Security (IC4S 2019), pp. 561-

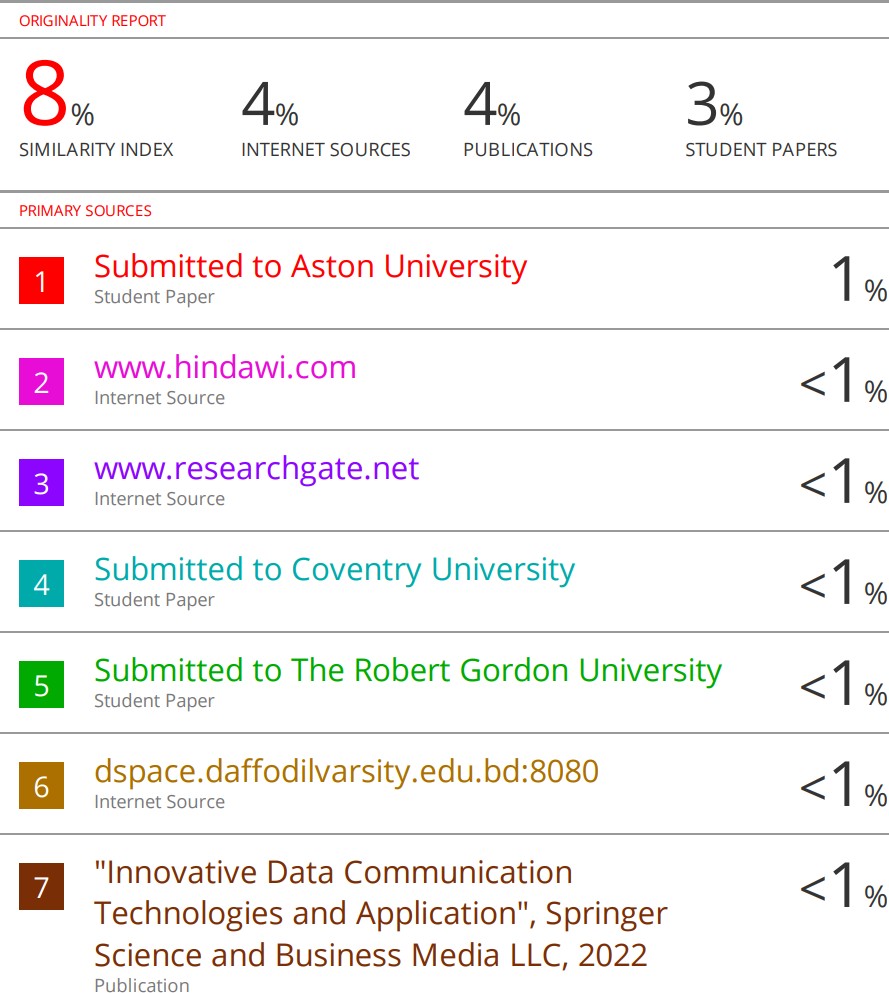
573. Springer, Singapore, 2020.

1. Rani, Pooja, Rajneesh Kumar, Nada MO Ahmed, and Anurag Jain. "A decision support system for heart disease prediction based upon machine learning." Journal of Reliable Intelligent Environments 7, no. 3 (2021): 263-275.
2. Sandhya, Yamala. "Prediction of Heart Diseases using Support Vector Machine." International Journal for Research in Applied Science & Engineering Technology (IJRASET)(ISSN: 2321-9653) Volume 8 (2020).
3. Sun, Weicheng, Ping Zhang, Zilin Wang, and Dongxu Li. "Prediction of cardiovascular diseases based on machine learning." ASP Transactions on Internet of Things 1, no. 1 (2021): 30-35.
4. Ali, Md Mamun, Bikash Kumar Paul, Kawsar Ahmed, Francis M. Bui, Julian MW Quinn, and Mohammad Ali Moni. "Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison." Computers in Biology and Medicine 136 (2021): 104672.

# APPENDIX

# APPENDIX A

PLAGIARISM REPORT

HEART DISEASE PREDICTION USING LOGISTIC REGRESSION IN COMPARISON WITH KNN ALGORITHM