# HEART DISEASE PREDICTION

A Course Project report submitted

in partial fulfillment of requirement for the award of degree

### BACHELOR OF TECHNOLOGY

in

### COMPUTER SCIENCE & ENGINEERING

by

### R.SRIKAR (2003A51650)

**M.SATWIK CHANDRA (2003A51645)**

### SHAIK TAUSEEF ALI (2003A51618)

Under the guidance of

### Mr. D. Ramesh sir

Assistant Professor, Department of CSE.



**Department of Computer Science and Artificial Intelligence**



**Department of Computer Science and Artificial Intelligence**

**CERTIFICATE**

This is to certify that this project entitled **“HEART DISEASE PREDICTION**" is the bonafied work carried out by **R.SRIKAR,M.SATWIK CHANDRA, SHAIK TAUSEEF ALI, G.VIGNESH** as a Course Project for the partial fulfilment to award the degree **BACHELOR OF TECHNOLOGY** in **COMPUTER SCIENCE & ENGINEERING** during

the academic year2022-2026 under our guidance and Supervision.

# ABSTRACT

World Health Organization has estimated 12 million deaths occur world- wide, every year due to Heart diseases. Machine learning is one of the trending technologies which used in many spheres around the world including healthcare industry for predicting diseases. The aim of our project is to identify the most significant factors of heart diseases and predicting the overall risks by using logistic regression. Thus, binary logistic model which is one of the classification algorithms in machine learning is used in this study to identify the predicators. Further, data analysis is carried out in Python using Jupyter Lab in order to validate the logistic regression. The dataset is taken from Kaggle is filtered and the features are selected based on high positive correlation values with the target and used random order of data (without sorting). The performance of the model is evaluated by different training and testing ratio of dataset.

To check the behaviour of the model with low to high training and testing data. After carrying out the detailed study on our dataset we got the accuracy of 88% on analysing the female dataset and the accuracy of 82% when done on both genders.

# Table of Contents

## Chapter No. Title Page No.

1. Introduction
   1. [Overview 1](#_bookmark0)
   2. [Problem Statement 1-2](#_bookmark1)
   3. Existing system 2
   4. [Proposed system 3](#_bookmark2)
   5. [Define Objectives 3](#_bookmark3)
   6. [Overall architecture 4](#_bookmark4)
2. [Literature survey](#_bookmark5)
   1. Research and survey 5-7
3. Data pre-processing

|  |  |
| --- | --- |
| 3.1. Describe dataset | 8-11 |
| 3.2. Data cleaning | 11 |
| 3.3. Data Visualization | 12-16 |

1. Methodology
   1. Procedure to solve the given problem 17-19
   2. Model architecture 20
   3. Software description 20
2. Results and discussion 21
3. Conclusion and future scope 22
4. References 23

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **S.no** | **Figure name** | **Page.no** |
| 1.6.1 | Architecture of project | 4 |
| 3.1.1 | Describing data set code | 9 |
| 3.1.2 | Information about dataset code | 10 |
| 3.1.3 | Head and tail parts of dataset code | 10 |
| 3.1.4 | Shape() of dataset code | 11 |
| 3.2.1 | data cleaning code | 11 |
| 3.3.1 | Age vs count (female dataset) | 12 |
| 3.3.2 | Histograms on female dataset | 13 |
| 3.3.3 | Pair plot | 14 |
| 3.3.4 | Sex vs count (combined dataset) | 14 |
| 3.3.5 | Scatter plot between age and rest ecg | 15 |
| 3.3.6 | Scatter plot between age and trestbps(hue= target) | 15 |
| 3.3.7 | Scatter plot between age vs cp(hue-age) | 16 |
| 3.3.8 | Age vs target count (cat plot) | 16 |
| 3.3.9 | Grender vs target count(cat plot) | 16 |
| 4.1.1 | Logistic regression s-curve | 17 |

|  |  |  |
| --- | --- | --- |
| 4.1.2 | Sigmoid probability curve | 18 |
| 4.1.3 | Correlation matrix(female dataset) | 19 |
| 4.2.1 | Model architecture | 20 |
| 5.1 | Accuracy of female dataset | 21 |
| 5.2 | Accuracy of combined dataset | 21 |
| 5.3 | Final prediction for test data | 21 |

# 1 . INTRODUCTION

## OVERVIEW

Today the greatest challenge to medical industry to provide higher level facility to health infrastructure to diagnose the disease in the initial day and give timely treatment to improve the quality of life through quality of service. Around 31% of mortality occurs world due to cardiac disease. The developing and under developing countries lacks in infrastructure and technologies, infrastructure and doctors to predict the disease in early stage to avoid complications reduce mortality. In this project, we will be closely working with the heart disease prediction and for that, we will be looking into the heart disease dataset from that dataset we will derive various insights that help us know the weightage of each feature and how they are interrelated to each other but this time our sole aim is to detect the probability of person that will be affected by a savior heart problem or not. In this study, Logistic Regression classifier model are applied. Results are compared with existing studies.

## PROBLEM STATEMENT

Presently, the major challenge of the medical industry is to predict the cardio vascular disease with less expensive and more reliable method. The early detection not only reduce the cost but also improves the quality of life. The major challenge in heart disease is its detection. There are instruments available which can predict heart disease but either they are expensive. Since, we have a good amount of data in today’s world, we can use various machine learning algorithms to analyse the data for hidden patterns. The hidden patterns can be used for health diagnosis in medicinal data.

Based on the given information about each individual we have to calculate that whether that individual will suffer from heart disease or not.

Research&aim:

The aim of this research is to develop an efficient way to predict the presence of the cardiovascular disease. The steps as mentioned below. The Taken dataset is used to predict the disease. The Features are selected based on high positive correlation values with the target and used random order of data (without sorting).The performance of the model is evaluated by different training and testing ratio of dataset. To check the behaviour of the model with low to high training and testing data.

## EXISTING SYSTEMS

Heart disease is even being highlighted as a silent killer which leads to the death of a person without obvious symptoms. The nature of the disease is the cause of growing anxiety about the disease & its consequences. Hence continued efforts are being done to predict the possibility of this deadly disease in prior. So that various tools & techniques are regularly being experimented with to suit the present-day health needs. Machine Learning techniques can be a boon in this regard. Even though heart disease can occur in different forms, there is a common set of core risk factors that influence whether someone will ultimately be at risk for heart disease or not. By collecting the data from various sources, classifying them under suitable headings & finally analysing to extract the desired data we can conclude. This technique can be very well adapted to the do the prediction of heart disease.

## PROPOSED SYSTEM

The working of the system starts with the collection of data and selecting the important attributes. Then the required data is pre-processed into the required format. The data is then divided into two parts training and testing data. The logistic regression algorithm is applied and the model is trained using the training data. The accuracy of the system is obtained by testing the system using the testing data.

This system is implemented using the following modules.

* + 1. Collection of Dataset
    2. Selection of attributes
    3. Data Pre-Processing
    4. Balancing of Data
    5. Disease Prediction

## DEFINE OBJECTIVES

* The main motivation of doing this research is to present a heart disease prediction model for the prediction of occurrence of heart disease.
* As the well-known quote says “Prevention is better than cure”, early prediction & its control can be helpful to prevent & decrease the death rates due to heart disease.
* Further, this research work is aimed towards identifying the best classification algorithm for identifying the possibility of heart disease in a patient.

## OVERALL ARCHITECTURE

Dataset collection is collecting data which contains patient details. Attributes selection process selects the useful attributes for the prediction of heart disease. After identifying the available data resources, they are further selected, cleaned, made into the desired form. Then using logistic classification technique as stated will be applied on pre-processed data to predict the accuracy of heart-disease**.**

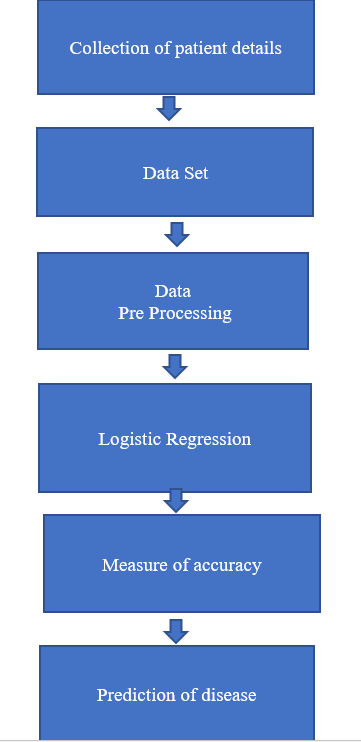


Fig.1.6.1. Model architecture

## 2.LITERATURE SURVEY

* 1. **.** Purushottam, et , al proposed a paper “Efficient Heart Disease Prediction System” using hill climbing and decision tree algorithms .They used Cleveland dataset and pre-processing of data is performed before using classification algorithms. The Knowledge Extraction is done based on Evolutionary Learning (KEEL), an opensource data mining tool that fills the missing values in the data set. A decision tree follows top-down order. For each actual node selected by hill-climbing algorithm a node is selected by a test at each level. The parameters and their values used are confidence. Its minimum confidence value is 0.25. The accuracy of the system is about 82.7%.
  2. **.** Santhana Krishnan. J ,et ,al proposed a paper “Prediction of Heart Disease Using Machine Learning Algorithms” using decision tree and Naive Bayes algorithm for prediction of heart disease. In decision tree algorithm the tree is built using certain conditions which gives True or False decisions. But decision tree for a tree like structure having root node, leaves and branches base on the decision made in each of tree Decision tree also help in the understating the importance of the attributes in the dataset. They have also used Cleveland data set. Dataset splits in 70% training and 30% testing by using some methods. This algorithm gives training 91% accuracy. The second algorithm is Naive Bayes, which is used for classification. It can handle complicated, nonlinear, dependent data so it is found suitable for heart disease dataset as this dataset is also complicated, dependent and nonlinear in nature. This algorithm gives an 80% accuracy.
  3. **.** Sonam Nikhar et al proposed paper “Prediction of Heart Disease Using Machine Learning Algorithms” their research gives point to point explanation of Naïve Bayes and decision tree classifier that are used especially in the prediction of Heart Disease. 3 Some analysis has been led to think about the execution of prescient data mining strategy on the same dataset, and the result decided that Decision Tree has highest accuracy than Bayesian classifier.
  4. **.**Aditi Gavhane et al proposed a paper “Prediction of Heart Disease Using Machine Learning”, in which training and testing of dataset is performed

by using neural network algorithm multi-layer perceptron. In this algorithm there will be one input layer and one output layer and one or more layers are hidden layers between these two input and output layers. Through hidden layers each input node is connected to output layer. This connection is assigned with some random weights. The other input is called bias which is assigned with weight based on requirement the connection between the nodes can be feedforwarded or feedback.

**[5].** Avinash Golande et al, proposed “Heart Disease Prediction Using Effective Machine Learning Techniques” in which few data mining techniques are used that support the doctors to differentiate the heart disease. Usually utilized methodologies are k-nearest neighbour, Decision tree and Naïve Bayes. Other unique characterization-based strategies utilized are packing calculation, Part thickness, consecutive negligible streamlining and neural systems, straight Kernel selfarranging guide and SVM (Bolster Vector Machine).

**[6]** .Lakshmana Rao et al,proposed “Machine Learning Techniques for Heart Disease Prediction” in which the contributing elements for heart disease are more. So, it is difficult to distinguish heart disease. To find the seriousness of the heart disease among people different neural systems and data mining techniques are used.

**[7].** Abhay Kishore et al proposed “Heart Attack Prediction Using Deep Learning” in which heart attack prediction system by using Deep learning techniques and to predict the probable aspects of heart related infections of the patient Recurrent Neural System is used. This model uses deep learning and data mining to give the best precise model and least blunders. This paper acts as strong reference model for another type of heart attack prediction models.

**[8]**. Senthil Kumar Mohan et al, proposed “Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques” in which their main objective is to improve 4 exactness in cardiovascular problems. The algorithms used are KNN, LR, SVM, NN to produce an improved exhibition level with a precision level of 88.7% through the prediction model for heart disease with hybrid random forest with linear model also known as HRFLM.

**[9].** Anjan N. Repaka et al, proposed a model stated the performance of prediction for two classification models, which is analyzed and compared to previous work. The experimental results show that accuracy is improved in finding the percentage of risk prediction of our proposed method in comparison with other models.

**[10]** .Aakash Chauhan et al, proposed “Heart Disease Prediction using Evolutionary Rule Learning”. Data is directly retrieved from electronic records that reduce the manual tasks. The amount of services are decreased and shown major number of rules helps within the best prediction of heart disease. Frequent pattern growth association mining is performed on patient’s dataset to generate strong association. The accuracy of the model is about around 84%.

The aforementioned are few registered model works done by using machine learning algorithms to contribute in health industry specifically for heart disease prediction .The mean percentile accuracy of all these models is about 82%.More works are consistently being done (mostly using advanced neural networks) to improve this in order to achieve more accuracy in predictions and build efficient diagnosis systems for automated monitoring.

# 3.DATA PRE-PROCESSNIG

## DESCRIBE DATASET

The data set is taken from [www.kaggle.com.](http://www.kaggle.com/) It contains 14 attributes including the predicting attribute. The "target" field refers to the presence of heart disease in the patient. It is integer valued 0 = no disease and 1 = disease.

### Attribute information:

1. Age:

Age of a person in years (to be collected from patient’s record - independent variable).

1. Sex:

Gender of a person (male -1 and female – 0) 3.chest pain type (4 values)

There are four types of pain

* 1. Typical angina pain – represented by value 1 ii)Atypical anginal pain-represented by value 2 iii)non-anginal pain-represented by value 3 iv)asymptomatic-represented by value 4

1. resting blood pressure

For the average person who is not at high risk of hyper tension, we aim for a total blood pressure of 120/80 or lower if it is more than 120/80 you could be at risk of heart disease.

1. serum cholestoral in mg/dl

High levels of cholesterol can increase your risk of heart disease. There is a sharp increase in the risk for cardio vascular disease when total cholesterol levels are 240mg/dl and above.

1. fasting blood sugar

Fasting glucose levels <70mg/dL were associated with increased risk of all stroke(hazard ratio 1.06, 95% Cl 1.01-1.11) in men and (hazard ratio 1.11,1.05-1.17) in women.

Both low glucose level and impaired fasting glucose should be considered as predictors of risk for stroke and coronary heart disease.

1. resting electrocardiographic results

Rest ecg varies between 0,1,2. 0 represents normal,1represents having ST- T, 2 represents hypertrophy.

1. maximum heart rate achieve

It is the value of highest heart beat rate recorded by machine of a person.Maximum heart rate achieved should be between 140 and 173.

1. exercise induced angina

Anigma appearing due to physical activity,emotional stress ,or due to exposure to cold temperatures. 1 represents yes and 0 represents no exang.

1. oldpeak

It is ST depression induced by exercise relative to rest.it is a numeric

value.

1. the slope of the peak exercise ST segment

Slope is calculated of the peak exercise st depression curve.1 represents unsloping, 2 represents flat, 3 represents down sloping, it can have any one of the above value.

1. ca

number of major vessels (0-3) coloured by fluoroscopy. It can be any value in the range of 0-3

1. thal:

Commonly known as thalassemia. It is a deficiency of oxygen in blood.

Thalessemia in dataset is represented with following values: 0 = normal; 1 = fixed defect; 2 = reversable defect

1. target –

It is a final predicted value by system or model. If the target has 1 it means the person is suffering from heart disease and 0 represents no heart disease.

(1 = died; 0 = lived)

We have taken only the values of female from the data set to find heart diseases in female.

### Import dataset

After downloading the dataset from Kaggle, I saved it to my working directory with the name female.csv. Next, I used read\_csv() to read the dataset and save it to the dataset variable.Before any analysis, I just wanted to take a look at the data. So, I used the info() method.

### describe()

The describe() method is used for calculating some statistical data like **percentile, mean** and **std** of the numerical values of the Series or DataFrame. It analyzes both numeric and object series and also the DataFrame column sets of mixed data types.

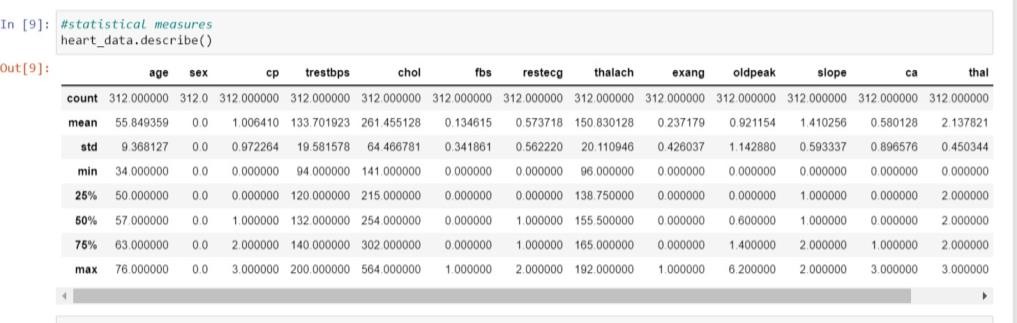


Fig.3.1.1. describing dataset code

### info()

The info() method prints information about the DataFrame.The information contains the number of columns, column labels, column data types, memory usage, range index, and the number of cells in each column (non-null values).

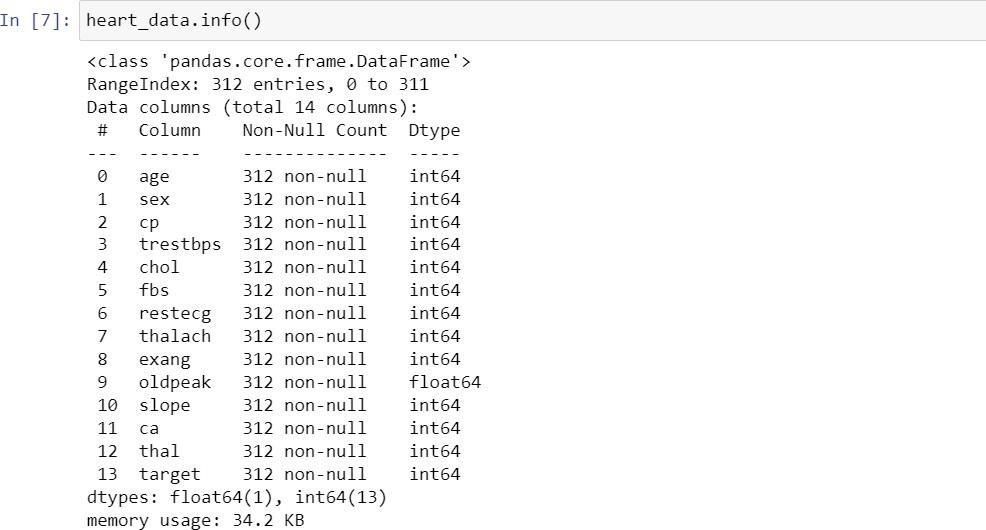


Fig.3.1.2.Info about dataset

### head():

The **head function** in Python displays the first five rows of the dataframe by default. It takes in a single parameter: the number of rows. We can use this parameter to display the number of rows of our choice.

**Tail():**

The **tail function** in Python displays the last five rows of the dataframe by default. It takes in a single parameter: the number of rows. We can use this parameter to display the number of rows of our choice.

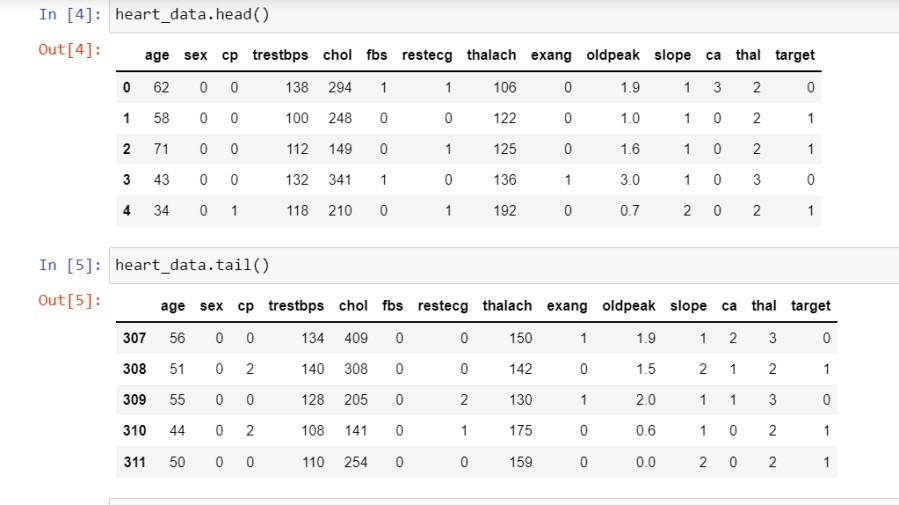


Fig.3.1.3.head and tail of dataset

### shape():

pandas.DataFrame

shape

The

attribute of

stores the number of rows and

columns as a tuple .

(number of rows, number of columns)

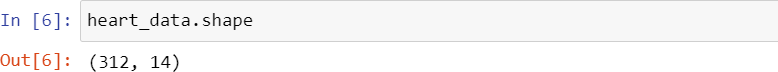


Fig.3.1.4.shape () of dataset

## DATA CLEANING

We need to check whether any value in the data set is null and remove or replace null with mean or minimum or maximum or desired value.

### Isnull():

isnull() function to detect missing values in the given series object.

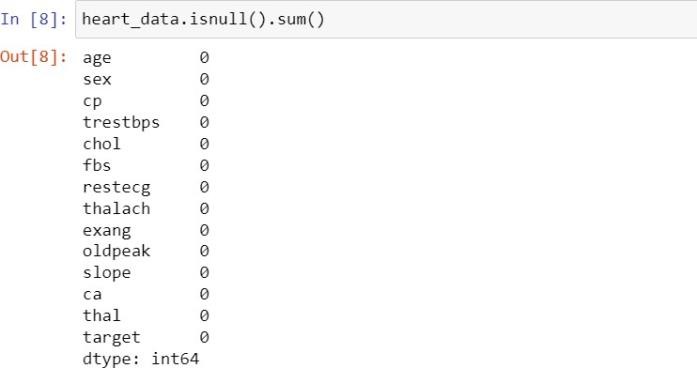


Fig.3.2.1.data cleaning code

## DATA VISUALISATION

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, maps. data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

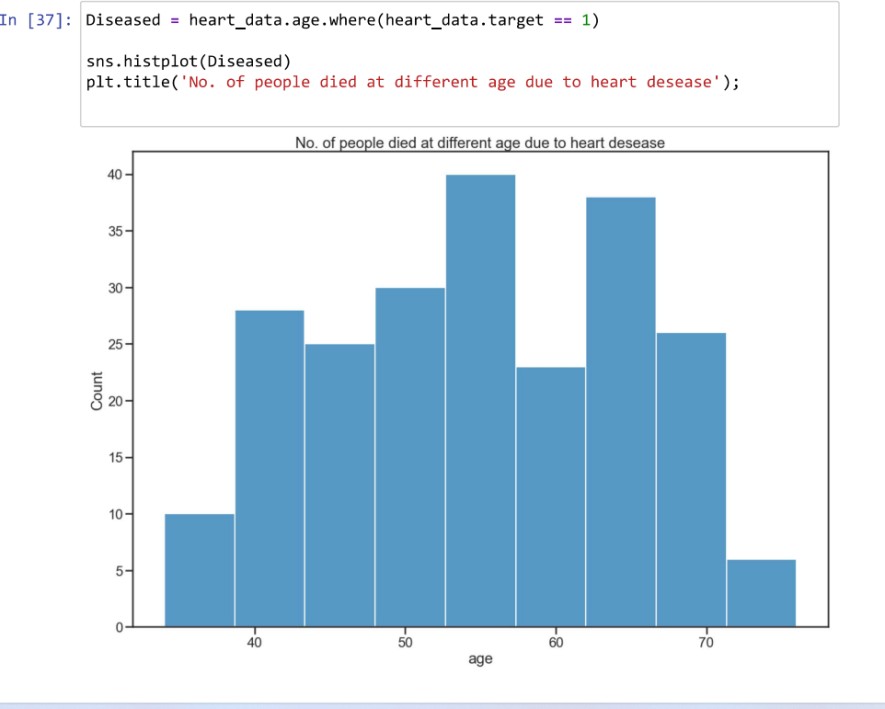
**Graphs for female data set:**

Fig.3.3.1.Age vs count(female dataset)

This plot shows that count of people died at different age, and people with age between 52-58 has highest count.

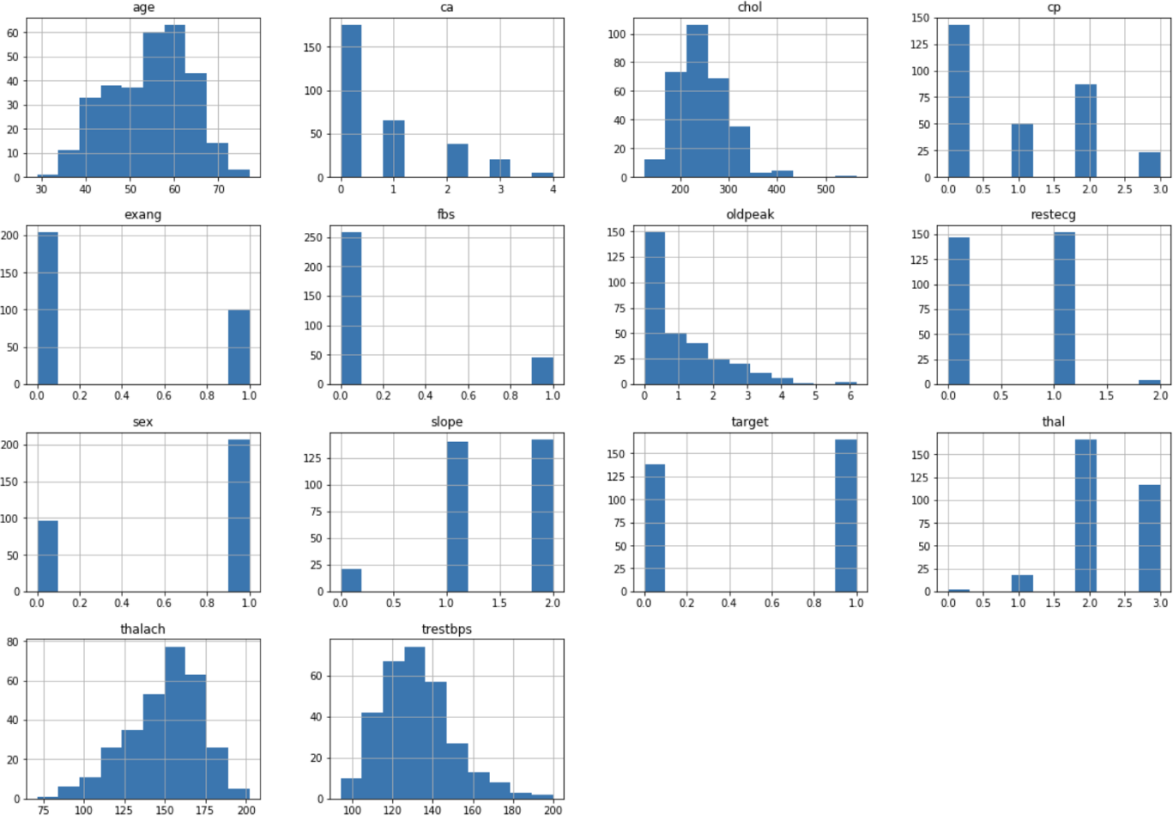


Fig.3.3.2 .Histograms

Histogram which gives more information about dataset.



Fig.3.3.3.pair plot

Pair plot to visualise relationship between each variable.

### Data visualisation on both male and female data set:

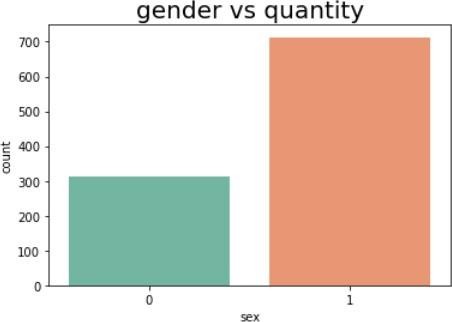


Fig.3.3.4.sex vs count

Plot between number of diseased and gender.

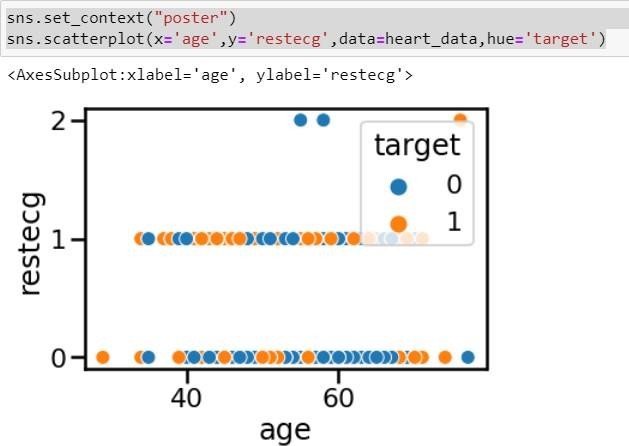


Fig.3.3.5

Scatter plot between age and restecg

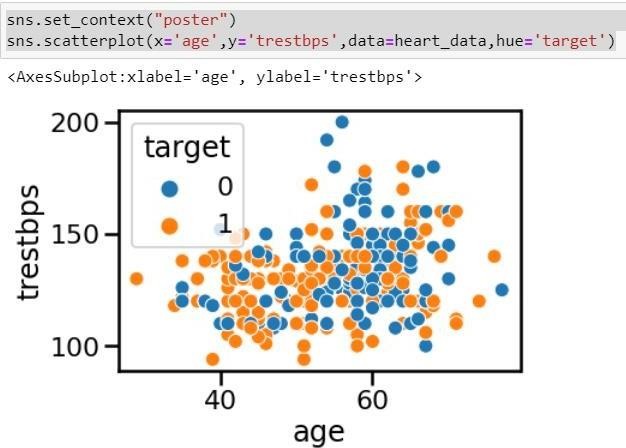


Fig.3.3.6

Scatter plot between age and trestbps taking target as hue.



Fig.3.3.7

Plot between age and cp with hue as age.

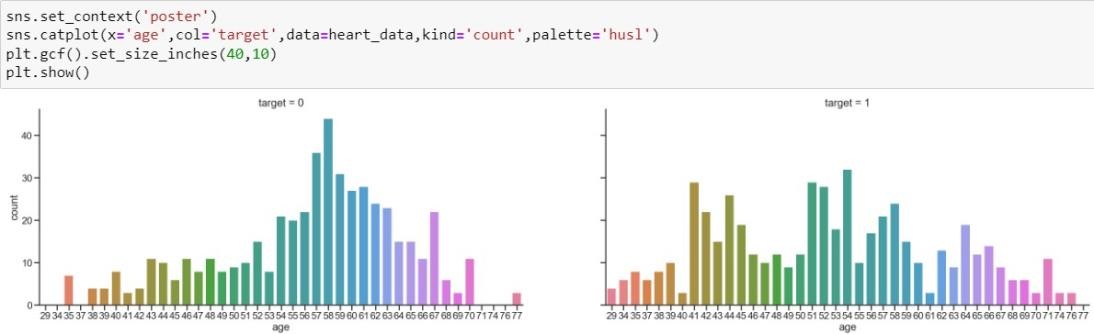


Fig.3.3.8

Cat plot between age and target count.

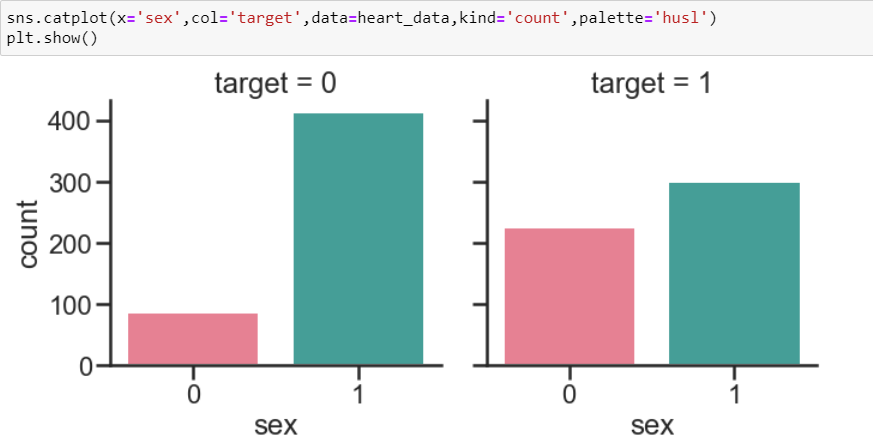


Fig. 3.3.9. Cat plot between gender and target count.

# 4.METHODOLOGY

## PROCEDURE TO SOLVE THE GIVEN PROBLEM Logistic regression

Logistic Regression is widely used to predict binary outcomes for a given set of independent variables. The dependent variable’s outcome is discrete such as y ϵ {0, 1} A binary dependent variable can have only two values such as 0 or 1, win or lose, pass or fail, healthy or sick.

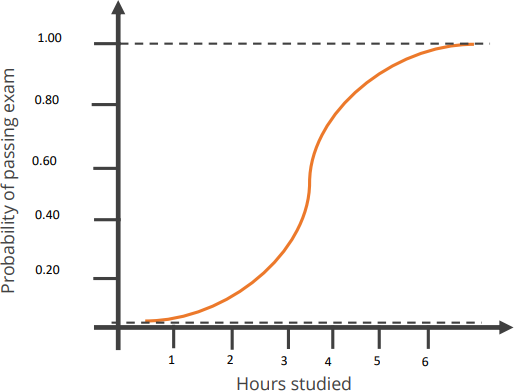


Fig.4.1.1.logistic regression s-curve

The probability distribution of output y is restricted to 1 or 0. This is called as sigmoid probability (σ). If σ (θTx) > 0.5, set y = 1, else set y = 0. Unlike Linear Regression (its Normal Equation solution), there is no closed form solution for finding optimal weights of Logistic Regression. Instead, you must solve this with maximum likelihood estimation (a probability model to detect maximum likelihood of something happening).

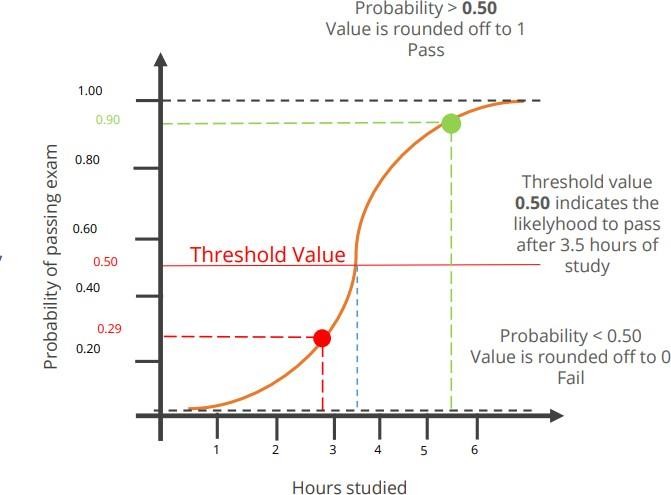
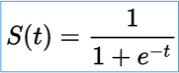


Fig.4.1.2.sigmoid probability curve

### Sigmoid probability

The probability in the logistic regression is represented by the Sigmoid function (logistic function or the S-curve).

t represents data values \* number of hours studied S(t) represents the probability of passing the exam.

The sigmoid function gives an ‘S’ shaped curve.

This curve has a finite limit that is Y can only be 0 or 1.

0 as x approaches to −∞ 1 as x approaches to +∞

Importing libraries

1. **numpy** : To work with arrays
2. **pandas**: To work with csv files and data frames
3. **matplotlib**: To create charts using pyplot, define parameters using rcParams and colour them with cm.rainbow
4. **warnings**: To ignore all warnings which might be showing up in the notebook due to past/future depreciation of a feature
5. **train\_test\_split**: To split the dataset into training and testing data
6. **StandardScaler**: To scale all the features, so that the Machine Learning model better adapts to the dataset

Next, we imported the necessary Machine Learning algorithm Correlation matrix.

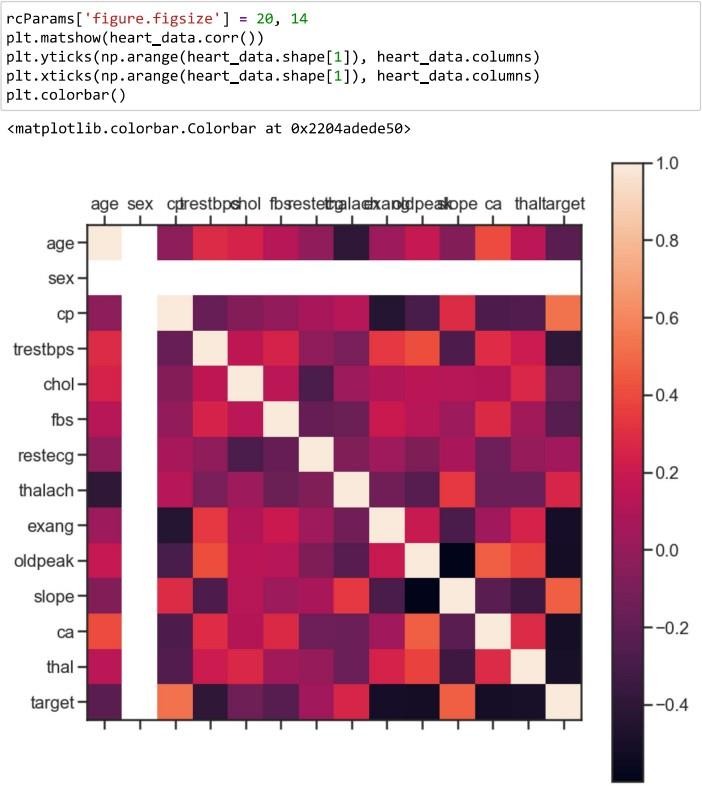
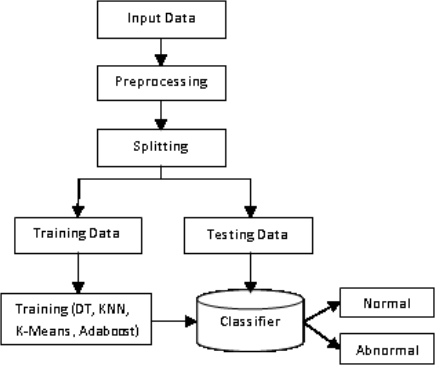


Fig.4.1.3 correlation matrix(female dataset)

It’s easy to see that there is no single feature that has a very high correlation with our target value. Also, some of the features have a negative correlation with the target value and some have positive.

## MODEL ARCHITECCTURE



**Training(logitsic regression)**

Fig.4.2.1.model architecture

## SOFTWARE DESCRIPTION

JUPYTER- The Jupyter Notebook is an open-source web application that you can use to create and share documents that contain live code, equations, visualizations, and text. Jupyter Notebook is maintained by the people at Project Jupyter. Jupyter notebooks are used for a variety of purposes. A notebook is an interactive computational environment in which users can execute a particular piece of code and observe the output and make changes to the code to drive it to the desired output or explore more. Jupyter notebooks are heavily used for data exploration purposes as it involves a lot of reiterations. It is also used in other data science workflows such as machine learning experimentations and modelling.

# RESULTS AND DISCUSSION

* After performing the logistic regression approach for training data we find that accuracy of the training data of female is 0.9076 .
* Accuracy of the testing data of female is found to be 0.8888.
* It means that the efficiency of the model is 88%.
* After performing the logistic regression approach for training data we find that accuracy of the training data of both genders is 0.85243.
* Accuracy of the testing data of both genders is found to be 0.8048





Fig.5.1.accuracy of female dataset

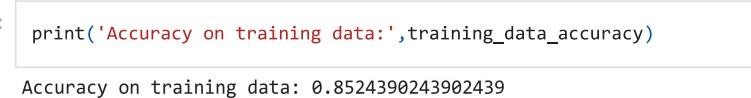


Fig.5.2.accuracy of combined dataset

From existing models it is found that accuracy is 83% by KNN algorithm, 80% by SVM algorithm and 78 % by decision tree.

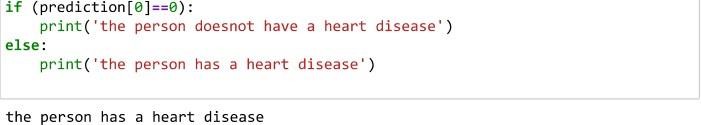


Fig.5.3.final prediction of test data

The model finally predicts whether the patient may get heart disease or not as a binary value true or false.

# CONCLUSION AND FUTURE SCOPE

Heart disease is soon becoming one of the major and common concern for society today . we have tried to build a using the supervised learning logistic classfication model to predict the outcome when given the necessary data .The accuracy of the trained model when tested for only female data is 88%.

The accuracy of the trained model when tested for combined dataset of both genders is about 80%

Today’s, world most of the data is computerized, the data is distributed and it is not utilizing properly. By analysing the available data set we can also use for unknown patterns. The primary motive of this research is the prediction of heart diseases with high rate of accuracy.

For predicting the heart-disease we can use logistic regression algorithm, naïve bayes, sklearn in machine learning. The future scope of the paper is the prediction of heart diseases by using advanced techniques and algorithms in less time complexity.

# REFERENCES

* [www.kaggle.com](http://www.kaggle.com/)
* [https://www.kaggle.com/datasets/dileep070/heart-disease-prediction-using-](https://www.kaggle.com/datasets/dileep070/heart-disease-prediction-using-logistic-regression) [logistic-regression](https://www.kaggle.com/datasets/dileep070/heart-disease-prediction-using-logistic-regression)
* <https://www.ijeat.org/wp-content/uploads/papers/v8i3S/C11410283S19.pdf>
* [https://towardsdatascience.com/predicting-presence-of-heart-diseases-using-](https://towardsdatascience.com/predicting-presence-of-heart-diseases-using-machine-learning-36f00f3edb2c) [machine-learning-36f00f3edb2c](https://towardsdatascience.com/predicting-presence-of-heart-diseases-using-machine-learning-36f00f3edb2c)
* [1] Purushottam,Et,Al(2015). An intelligent heart disease prediction system using k-means clustering and Naïve Bayes algorithm. International Journal of Computer Science and Information Technologies, 6(1), 637-9.
* [2] Santhana Krishnan.J,Et,Al(2014, November). An ensemble-based decision support framework for intelligent heart disease diagnosis. In International Conference on Information Society (i-Society 2014) (pp. 259-64). IEEE. ICCRDA 2020 IOP Conf. Series: Materials Science and Engineering 1022 (2021) 012072 IOP Publishing doi:10.1088/1757-899X/1022/1/012072 9
* [3] Sonam nikhar,Ed,Al (2011, September). HDPS: Heart disease prediction system. In 2011 Computing in Cardiology (pp. 557- 60). IEEE.
* [4]Adithi gavhane "Intelligent heart disease prediction system using CANFIS and genetic algorithm." International Journal of Biological, Biomedical and Medical Sciences 3.3 (2008).
* [4] Avinash golande,Et,Al(2016). Wireless body area network for heart attack detection [Education Corner]. IEEE antennas and propagation magazine, 58(5), 84-92.
* [5] Laxmana rao,Et, Al(2014). Heart attack detection and medical attention using motion sensing device -kinect. International Journal of Scientific and Research Publications, 4(1), 1-4.
* [6] Aakash chaujan (2012). Improved study of heart disease prediction system using data mining classification techniques. International Journal of Computer Applications, 47(10), 44-8.