**HEART FAILURE PREDICTION**

Project submitted to the

SRM University – AP, Andhra Pradesh

for the partial fulfillment of the requirements to award the degree of

**Bachelor of Technology/Master of Technology**

In

**Computer Science and Engineering**

**School of Engineering and Sciences**

Submitted by

**S.Faiyajuddin K. Sai Teja G. Chandra Shekar K. Jayaprakash D. Nithin**

**AP20110010363 AP20110010745 AP20110010410 AP2010010403 AP20110010362**

**A picture containing text

Description automatically generated**

Under the Guidance of

**Prof. / Dr. Saleti Sumalatha Madem**

**SRM University–AP**

**Neerukonda, Mangalagiri, Guntur**

**Andhra Pradesh – 522 240**

**December, 2022**

# Certificate

Date: 17-Dec-22

This is to certify that the work present in this Project entitled “**Heart Failure Prediction**” has been carried out by **Shaik.Faiyajuddin, Kadiyala.Jayaprakash, G Chandra Shekar, Sai Teja Kambhampati, Desu.Nithin Datta** under my/our supervision. The work is genuine, original, and suitable for submission to the SRM University – AP for the award of Bachelor of Technology/Master of Technology in **School of Engineering and Sciences**.

**Supervisor**

(Signature)

Prof. / Dr. **Saleti Sumalatha Madem**

Department of CSE, SRM University -AP

Andhra Pradesh.

# Acknowledgements

We thank Saleti Sumalatha Madem, our professor in charge, for the help and direction in finishing our project on this subject. This work has given us an opportunity to explore new topics which we are curious about. Your insightful counsel and recommendations were quite beneficial to us as we finished the project. we will be always grateful to you for this.

# Table of Contents

[Certificate i](#_Toc105799088)

[Acknowledgements iii](#_Toc105799089)

[Table of Contents v](#_Toc105799090)

[Abstract vii](#_Toc105799091)

[List of Figures xix](#_Toc105799097)

[1. Introduction 1](#_Toc105799099)

[1.1. Problem Definition 1](#_Toc105799100)

[1.2. Objectives 3 1](#_Toc105799101)

1.3. Data Set Descripton…………………………………………………………….....2

1.3.1. Data Set Attributes…………………………………………………………..2

1.3.2 Data Set Info…………………………………………………………………..2

2. Methodology……………………………………………………………………………4

2.1. Label Encoding………………………………………………………………….………….4

2.2. Data Scaling……………………………………………………………...………………….4

2.3. Min-Max Scaling……………………………………………………………………4

2.4. Standard Scaling……………………………………………………………………5

2.5. Chi-Square Test……………………………………………………………………..5

2.6. Anova Test………………………………………………………………………….6

2.7. Logistic Regression……………………………………………………………..…6

2.7.1. Advantages of the logistic regression algorithm………………………….6

2.7.2. Mathematical representation of Logistic algorithm…………………..….7

3. Results and Discussion………………………………………………………………....8

3.1. Exploratory Data Analysis………………………………………………………8

4. Concluding Remarks………………………………………………………………….11

5. Future Works…………………………………………………………………………..13

References…………………………………………………………………………………..15

# Abstract

Cardiovascular diseases (CVDs) are the number 1 cause of death globally, taking an estimated 17.9 million lives each year, which accounts for 31% of all deaths worldwide. Four out of 5CVD deaths are due to heart attacks and strokes, and one-third of these deaths occur prematurely in people under 70 years of age. Heart failure is a common event caused by CVDs and this dataset contains 11 features that can be used to predict a possible heart disease.

People with cardiovascular disease or who are at high cardiovascular risk (due to the presence of one or more risk factors such as hypertension, diabetes, hyperlipidaemia or already established disease) need early detection and management wherein a machine learning model can be of great help. In this way, we try to solve automate another problem that occurs in the nature with a view to counter it and focus on to the next problem with the help of AI techniques.

# List of Figures

Figure 1. Min-Max Scaling……………………………...……………………….4

Figure 2. Standard Scaling……………………………...……………………….5

Figure 3. Chi Square Test……………………………...………………………...5

Figure 4. Anova Test……..……………………………...……………………….6

Figure 5. Mathematical Representation of Logistic Regression……………..7

Figure 6. Catogorial features Chi Square Test results…….………………….9

Figure 7. Numerical features Chi Square Test results………………………..9

Figure 8. Logistic Regression...………………………...……………………….10

# Introduction

According to the World Health Organization, every year 12 million deaths occur worldwide due to Heart Diseases. The load of cardiovascular disease is rapidly increasing all over the world from the past few years. Many researches have been conducted in attempt to pinpoint the most influential factors of heart disease as well as accurately predict the overall risk. Heart Disease is even highlighted as a silent killer which leads to the death of the person without obvious symptoms. The early diagnosis of heart disease plays a vital role in making decisions on lifestyle changes in high-risk patients and in turn reduce the complications. This project aims to predict future Heart Disease by analyzing the data of patients which classifies whether they have heart disease or not using machine-learning algorithms

## 1.1 Problem Definition

The major challenge in heart disease is its detection. There are instruments available which can predict heart disease but either they are expensive or are not efficient to calculate chance of heart disease in human. Early detection of cardiac diseases can decrease the mortality rate and overall complications. However, it is not possible to monitor patients every day in all cases accurately and consultation of a patient for 24 hours by a doctor is not available since it requires more sapience ,time and expertise. Since we have a good amount of data in today’s world, we can use various

machine learning algorithms to analyze the data for hidden patterns. The hidden patterns can be used for health diagnosis in medicinal data.

### 1.2 Objectives

The main objective of developing this project are:

1. To develop machine learning model to predict future possibility of heart disease by implementing **Logistic Regression**.

2.To determine significant risk factors based on medical dataset which may lead to heart disease.

1.3 Data Set Discription**:**

**1.3.1 Dataset Attributes:**

**Age** : age of the patient [years]

**Sex** : sex of the patient [M: Male, F: Female]

**ChestPainType** : chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]

**RestingBP** : resting blood pressure [mm Hg]

**Cholesterol** : serum cholesterol [mm/dl]

**FastingBS** : fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]

**RestingECG** : resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]

**MaxHR** : maximum heart rate achieved [Numeric value between 60 and 202]

**ExerciseAngina** : exercise-induced angina [Y: Yes, N: No]

**Oldpeak** : = ST [Numeric value measured in depression]

**ST\_Slope** : the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]

**HeartDisease** : output class [1: heart disease, 0: Normal]

**1.3.2 Dataset info:**

RangeIndex: 918 entries, 0 to 917

Data columns (total 12 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 Age 918 non-null int64

1 Sex 918 non-null object

2 ChestPainType 918 non-null object

3 RestingBP 918 non-null int64

4 Cholesterol 918 non-null int64

5 FastingBS 918 non-null int64

6 RestingECG 918 non-null object

7 MaxHR 918 non-null int64

8 ExerciseAngina 918 non-null object

9 Oldpeak 918 non-null float64

10 ST\_Slope 918 non-null object

11 HeartDisease 918 non-null int64

dtypes: float64(1), int64(6), object(5)

memory usage: 86.2+ KB

# 2. Methodology

The main purpose of designing this system is to predict the ten-year risk of future heart disease. We have used Logistic regression as a machine-learning algorithm to train our system and various techniques like lable encoding, data scaling, min-max scaling, standard scaling, chi-square test and anova test. These techniques are discussed below in detail.

### 2.1 Label Encoding:

Label Encoding is a popular encoding technique for handling categorical variables. In this technique, each label is assigned a unique integer based on alphabetical ordering. This approach is very simple and it involves converting each value in a column to a number.

**2.2 Data Scaling:**

Scaling of the data comes under the set of steps of data pre-processing when we are performing machine learning algorithms in the data set. As we know most of the supervised and unsupervised learning methods make decisions according to the data sets applied to them and often the algorithms calculate the distance between the data points to make better inferences out of the data.

**2.3 Min-Max Scaling:**

Min-Max scaling is a normalization technique that enables us to scale data in a dataset to a specific range using each feature's minimum and maximum value.

n this approach, the data is scaled to a fixed range - usually 0 to 1.  
The cost of having this bounded range - in contrast to standardization - is that we will end up with smaller standard deviations, which can suppress the effect of outliers.

A Min-Max scaling is typically done via the following equation:

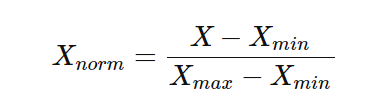


Figure:-1

**2.4 Standard Scaling:**

Standardization is a useful method to scales independent variables so that it has a distribution with 0 mean value and variance equals 1. However, Standard Scaler is not a good option if our datapoints aren’t normally distributed i.e they do not follow Gaussian distribution.

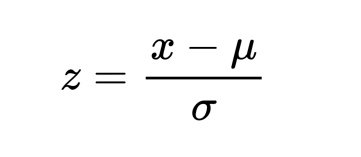


Figure:-2

# 2.5 Chi-Square Test:

A chi-square test is used in statistics to test the independence of two events. Given the data of two variables, we can get observed count O and expected count E. Chi-Square measures how expected count E and observed count O deviates each other.

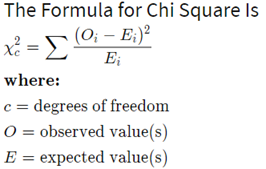


Figure:-3

Let’s consider a scenario where we need to determine the relationship between the independent category feature (predictor) and dependent category feature (response). In feature selection, we aim to select the features which are highly dependent on the response. When two features are independent, the observed count is close to the expected count, thus we will have smaller Chi-Square value. So high Chi-Square value indicates that the hypothesis of independence is incorrect. In simple words, higher the Chi-Square value the feature is more dependent on the response and it can be selected for model training.

**2.6 Anova Test:**

ANOVA stands for Analysis of variance. As the name, suggests it uses variance as its parameter to compare multiple independent groups. ANOVA can be one-way ANOVA or two-way ANOVA. One-way ANOVA is applied when there are three or more independent groups of a variable. We will implement the same in python.

F-Statistic can be calculated by

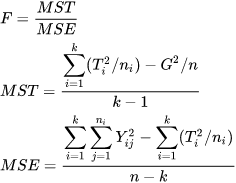


Figure:-4



## 2.7 Logistic Regression:

Logistic regression is a statistical method that is used for building machine learning models where the dependent variable is dichotomous: i.e. binary. Logistic regression is used to describe data and the relationship between one dependent variable and one or more independent variables. The independent variables can be nominal, ordinal, or of interval type.

The name “logistic regression” is derived from the concept of the logistic function that it uses. The logistic function is also known as the sigmoid function. The value of this logistic function lies between zero and one.

## 2.7.1 Advantages of the Logistic Regression Algorithm:

* Logistic regression performs better when the data is linearly separable
* It does not require too many computational resources as it’s highly interpretable
* There is no problem scaling the input features—It does not require tuning
* It is easy to implement and train a model using logistic regression
* It gives a measure of how relevant a predictor (coefficient size) is, and its direction of association (positive or negative)

## 2.7.2 Mathematical representation of logistic regression:

Probability always ranges between 0 (does not happen) and 1 (happens). Using our Covid-19 example, in the case of binary classification, the probability of testing positive and not testing positive will sum up to 1. We use [logistic function or sigmoid function](https://en.wikipedia.org/wiki/Logistic_function) to calculate probability in logistic regression. The logistic function is a simple S-shaped curve used to convert data into a value between 0 and 1.

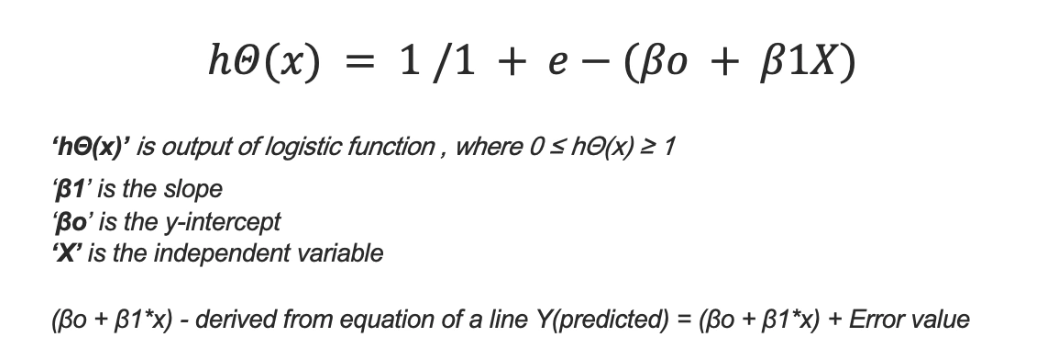


Figure:-5

# 3.Result and Discussion

**3.1 EXPLORATORY DATA ANALYSIS:**

**3.1.1 Dividing features into Numerical and Categorical :**

Here, categorical features are defined if the the attribute has less than 6 unique elements else it is a numerical feature. Typical approach for this division of features can also be based on the datatypes of the elements of the respective attribute.

**Example:**

datatype = integer, attribute = numerical feature ; datatype = string, attribute = categorical feature.

For this dataset, as the number of features are less, we can manually check the dataset as well.

Order of features for positive cases of heart disease :

**Categorical Features (Order) :**

* Sex : Male > Female
* ChestPainType : ASY > NAP > ATA > TA
* FastingBS : ( FBS < 120 mg/dl ) > ( FBS > 120 mg/dl)
* RestingECG : Normal > ST > LVH
* ExerciseAngina : Angina > No Angina
* ST\_Slope : Flat > Up > Down

**Numerical Features (Range) :**

* Age : 50+
* RestingBP : 95 - 170
* Cholesterol : 160 - 340
* MaxHR : 70 - 180
* Oldpeak : 0 - 4

Now that we have understood the typical values of the features, we move on to the next step where we select the appropriate features for modeling!

**Feature Selection for Categorical Features :**

For categorical feature we implement Chi square test and result ,

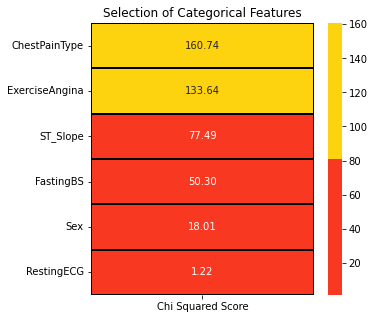


Figure:-6

**Feature Selection for Numerical Features :**

For Numerical feature we implement Chi square test and result ,

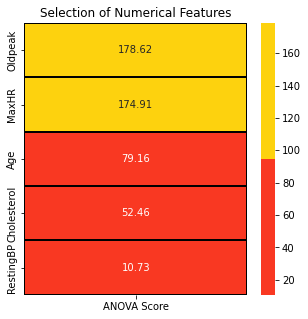


Figure:-7

Then the implementation of logistic regression.

**Result:**

**Model Evaluation**

Accuracy : 87.50%

Cross Validation Score : 91.12%

ROC\_AUC Score : 87.43%

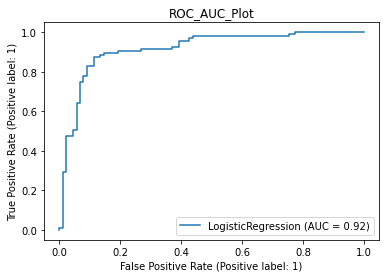


Figure:-8

# Concluding Remarks

* This dataset is great for understanding how to handle binary classification problems with the combination of numerical and categorical features.
* Subject matter experts, in this case doctors or nurses, can be assisted by providing insights that enables them to take the next line of action.
* For feature engineering, it might feel confusing about the order of the processes. In this case, data scaling was executed before the feature selection test. We might feel like we are tampering the data before passing it to the tests but the results are same irrespective of the order of the process. (Try it out!)
* For this problem, outlier detection was not done as I was not able to read any papers about heart diseases. It becomes a pivotal part to understand the subject before removing outliers even though the outlier detection tests come out positive.
* Visualization is key. It makes the data talkative. Displaying the present information and results of any tests or output through visualization becomes crucial as it makes the understanding easy.
* For modeling, hyperparameter tuning is not done. It can push the performances of the algorithms. Overall the algorithm performance is good.

# Future Work

In the above regression modeling, hyperparameter tuning is not done. It can push the performances of the algorithms better. Overall the algorithm performance is good. So, we are willing to work on the hyperparameter tuning in the coming days to obtain more accuracy and for the better performance of the algorithm.

# References

1. A. H. M. S. U. Marjia Sultana, "Analysis of Data Mining Techniques for Heart Disease Prediction," 2018.

2. M. I. K. ,. A. I. ,. S. Musfiq Ali, "Heart Disease Prediction Using Machine Learning Algorithms".

3.K.Bhanot,"towarddatascience.com,"13-Feb-2019. Available:https://towardsdatascience.com/predicting-presence-of-heart-diseases-using-machine-learning-36f00f3edb2c. [Accessed 2 March 2020].

4. https://www.kaggle.com/ronitf/heart-disease-uci#heart.csv.

5.M.A.K.S.H.K.M.a.V.P.M Marimuthu, "A Review on Heart Disease Prediction using Machine Learning and Data Analytics Approach".