

# CONTENT TABLE

## **Chapter 1 - INTRODUCTION**

1. Introduction
2. The Data
3. Why these Parameters

## **Chapter 2- Data Analysis**

1. HeatMap
2. Age Distribution
3. Gender Distribution
4. Chest Pain
5. Fasting Sugar

## **Chapter 3 - Models**

1. K Nearest Neighbours
2. Logistic Regression
3. Decision Trees
4. Neural Network

## **Chapter 4 - Result**

## **Chapter 5 - PREVIOUS WORK**

## **Chapter 6 - FUTURE WORK**

## **Chapter 7 - REFERENCES**

## 1.1 Introduction

---

**Heart disease** describes a range of conditions that affect your heart. Diseases under the heart disease umbrella include blood vessel diseases, such as coronary artery disease, heart rhythm problems (arrhythmias) and heart defects you're born with (congenital heart defects), among others.

The term "heart disease" is often used interchangeably with the term "cardiovascular disease". Cardiovascular disease generally refers to conditions that involve narrowed or blocked blood vessels that can lead to a heart attack, chest pain (angina) or stroke. Other heart conditions, such as those that affect your heart's muscle, valves or rhythm, also are considered forms of heart disease.

Heart disease is one of the biggest causes of morbidity and mortality among the population of the world. Prediction of cardiovascular disease is regarded as one of the most important subjects in the section of clinical data analysis. The amount of data in the healthcare industry is huge. Data mining turns the large collection of raw healthcare data into information that can help to make informed decisions and predictions.

This makes heart disease a major concern to be dealt with. But it is difficult to identify heart disease because of several contributory risk factors such as diabetes, high blood pressure, high cholesterol, abnormal pulse rate, and many other factors. Due to such constraints, scientists have turned towards modern approaches like Data Mining and Machine Learning for predicting the disease.

Machine learning (ML) proves to be effective in assisting in making decisions and predictions from the large quantity of data produced by the healthcare industry.

We will be applying Machine Learning approaches (and eventually comparing them) for classifying whether a person is suffering from heart disease or not, using one of the most used dataset — Cleveland Heart Disease dataset from the UCI Repository.

## 1.2 The Data

---

The dataset used in this article is the Cleveland Heart Disease dataset taken from the UCI repository.

The dataset consists of 303 individuals' data. There are 14 columns in the dataset, which are described below.

Serial No.	Column	Name
1	3	Age
2	4	Sex
3	9	Chest Pain Type
4	10	Resting Blood Pressure
5	12	Serum Cholesterol
6	16	Fasting Blood Sugar
7	19	Resting ECG
8	32	Max Heart Rate Achieved
9	38	Exercise-Induced Angina
10	40	ST depression induced by exercise relative to rest
11	41	Peak exercise ST segment
12	44	Number of major vessels (0–3) colored by fluoroscopy
13	51	Thal
14	58	Diagnosis of heart disease

## 1.3 Why these parameters:

---


In the actual dataset, we had 76 features but for our study, we chose only the above 14 because :

**Age:** Age is the most important risk factor in developing cardiovascular or heart diseases, with approximately a tripling of risk with each decade of life. Coronary fatty streaks can begin to form in adolescence. It is estimated that 82 percent of people who die of coronary heart disease are 65 and older. Simultaneously, the risk of stroke doubles every decade after age 55.

**Sex:** Men are at greater risk of heart disease than pre-menopausal women. Once past menopause, it has been argued that a woman's risk is similar to a man's, although more recent data from the WHO and UN disputes this. If a female has diabetes, she is more likely to develop heart disease than a male with diabetes.

**Angina (Chest Pain):** Angina is chest pain or discomfort caused when your heart muscle doesn't get enough oxygen-rich blood. It may feel like pressure or squeezing in your chest. The discomfort also can occur in your shoulders, arms, neck, jaw, or back. Angina pain may even feel like indigestion.

**Resting Blood Pressure:** Over time, high blood pressure can damage arteries that feed your heart. High blood pressure that occurs with other conditions, such as obesity, high cholesterol, or diabetes, increases your risk even more.



**Serum Cholesterol:** A high level of low-density lipoprotein (LDL) cholesterol (the “bad” cholesterol) is most likely to narrow arteries. A high level of triglycerides, a type of blood fat related to your diet, also ups your risk of a heart attack. However, a high level of high-density lipoprotein (HDL) cholesterol (the “good” cholesterol) lowers your risk of a heart attack.

**Fasting Blood Sugar:** Not producing enough of a hormone secreted by your pancreas (insulin) or not responding to insulin properly causes your body's blood sugar levels to rise, increasing your risk of a heart attack.


**Resting ECG:** For people at low risk of cardiovascular disease, the USPSTF concludes with moderate certainty that the potential harms of screening with resting or exercise ECG equal or exceed the potential benefits. For people at intermediate to high risk, current evidence is insufficient to assess the balance of benefits and harms of screening.

**Max heart rate achieved:** The increase in cardiovascular risk, associated with the acceleration of heart rate, was comparable to the increase in risk observed with high blood pressure. It has been shown that an increase in heart rate by 10 beats per minute was associated with an increase in the risk of cardiac death by at least 20%, and this increase in the risk is similar to the one observed with an increase in systolic blood pressure by 10 mm Hg.

**Exercise-induced angina:** The pain or discomfort associated with angina usually feels tight, gripping or squeezing, and can vary from mild to severe. Angina is usually felt in the center of your chest but may spread to either or both of your shoulders, or your back, neck, jaw or arm. It can even be felt in your hands.

o Types of Angina

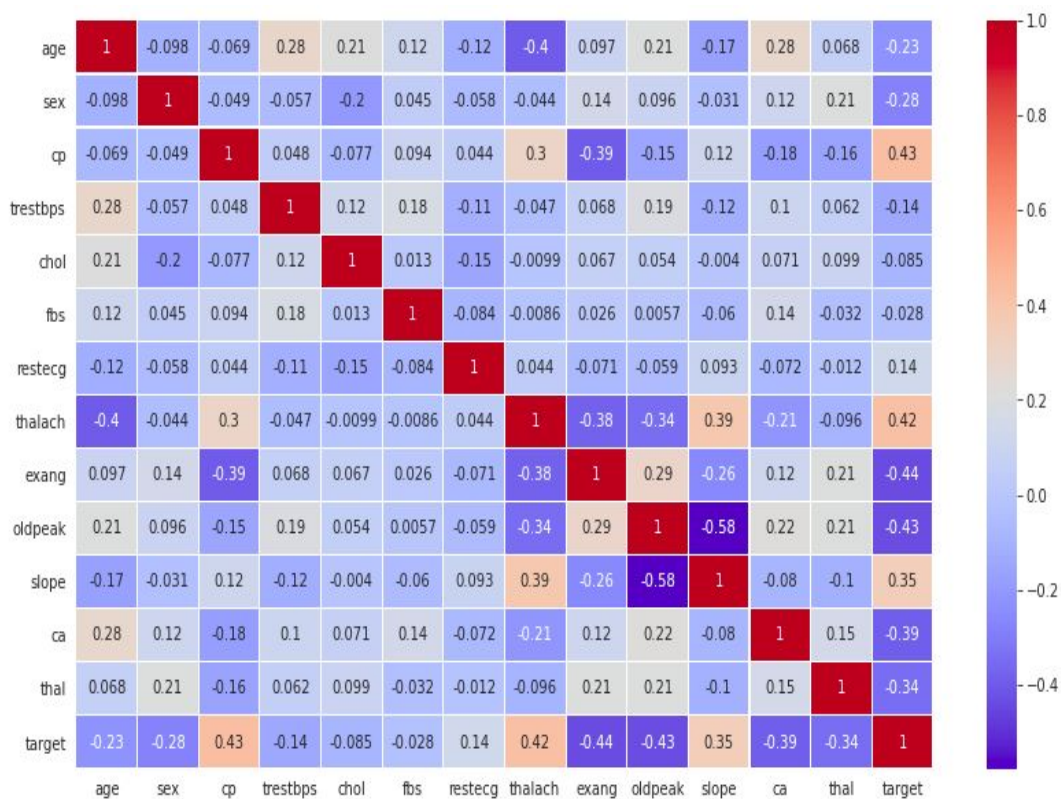
- a. Stable Angina / Angina Pectoris
- b. Unstable Angina
- c. Variant (Prinzmetal) Angina
- d. Microvascular Angina.



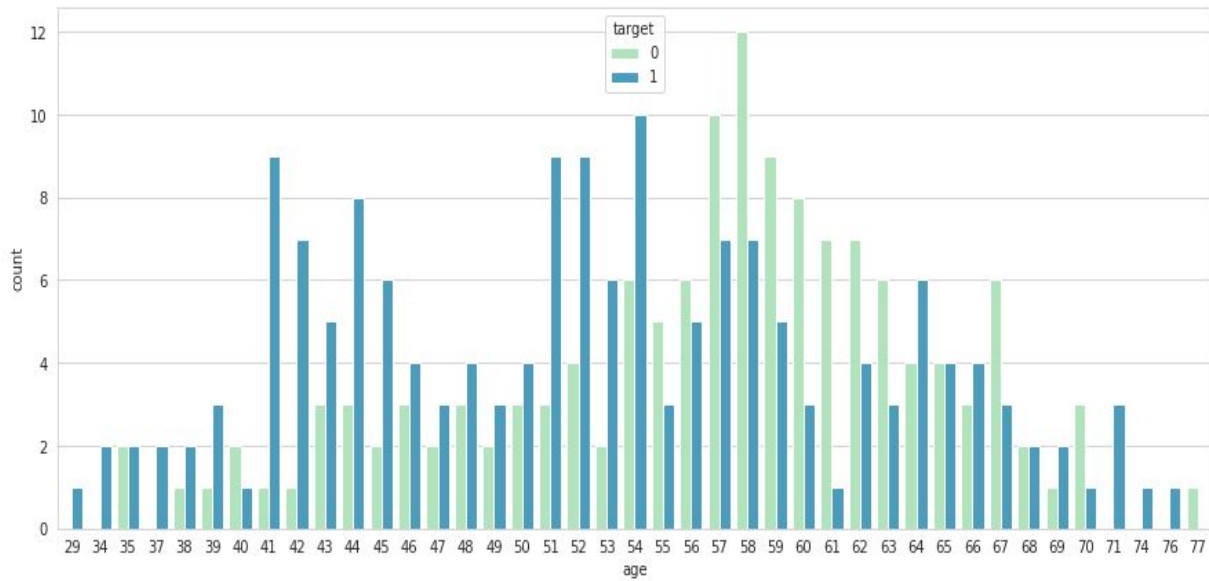
**Peak exercise ST segment:** A treadmill ECG stress test is considered abnormal when there is a horizontal or down-sloping ST-segment depression  $\geq 1$  mm at 60–80 ms after the J point. Exercise ECGs with up-sloping ST-segment depressions are typically reported as an 'equivocal' test. In general, the occurrence of horizontal or down-sloping ST-segment depression at a lower workload (calculated in METs) or heart rate indicates a worse prognosis and higher likelihood of multi-vessel disease. The duration of ST-segment depression is also important, as prolonged recovery after peak stress is consistent with a positive treadmill ECG stress test. Another finding that is highly indicative of significant CAD is the occurrence of ST-segment elevation  $> 1$  mm (often suggesting transmural ischemia); these patients are frequently referred urgently for coronary angiography.

## 2. Data Analysis

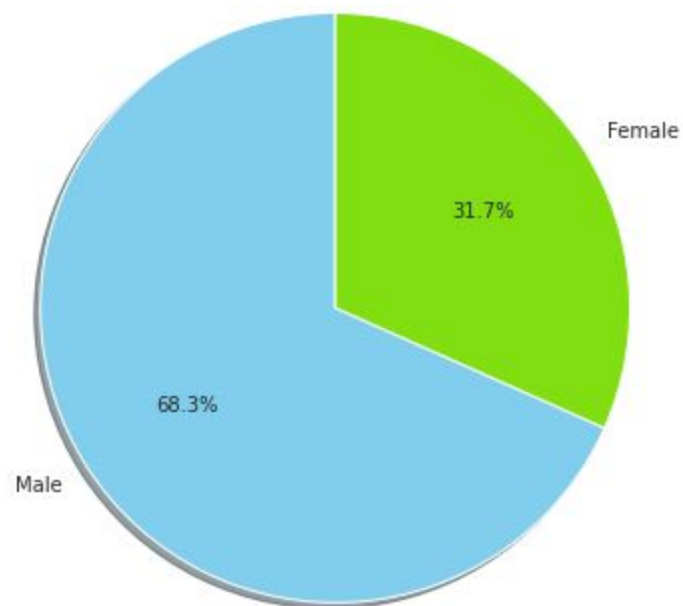
### 2.1 Heatmap



## 2.2 Age Distribution

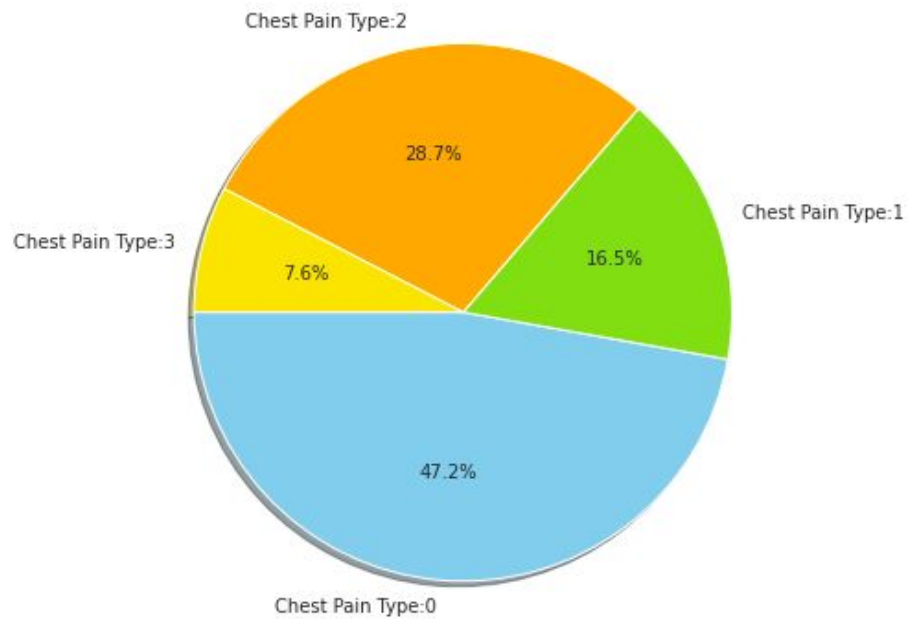


## 2.3 Gender Distribution

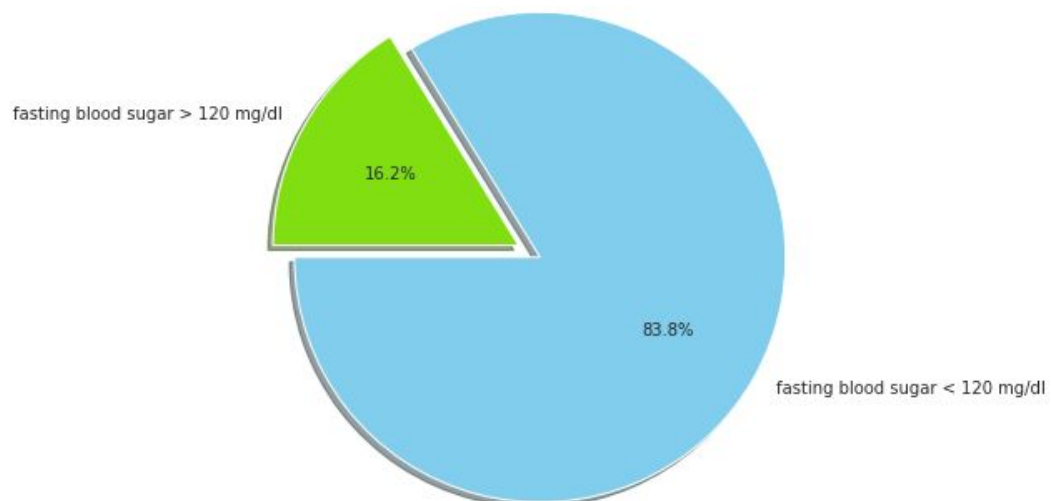




## 2.4 Chest Pain



## 2.5 Fasting Sugar



## 3. Models

---

### 3.1 K Nearest Neighbours

In pattern recognition, the k-nearest neighbors algorithm (k-NN) is a non-parametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression:

In k-NN classification, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor.

### 3.2 Logistic Regression

The logistic function, also called the sigmoid function was developed by statisticians to describe properties of population growth in ecology, rising quickly and maxing out at the carrying capacity of the environment. It's an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1.

$$1 / (1 + e^{-\text{value}})$$

Where e is the base of the natural logarithms and value is the actual numerical value that you want to transform. Below is a plot of the

numbers between -5 and 5 transformed into the range 0 and 1 using the logistic function.

### 3.3 Decision Trees

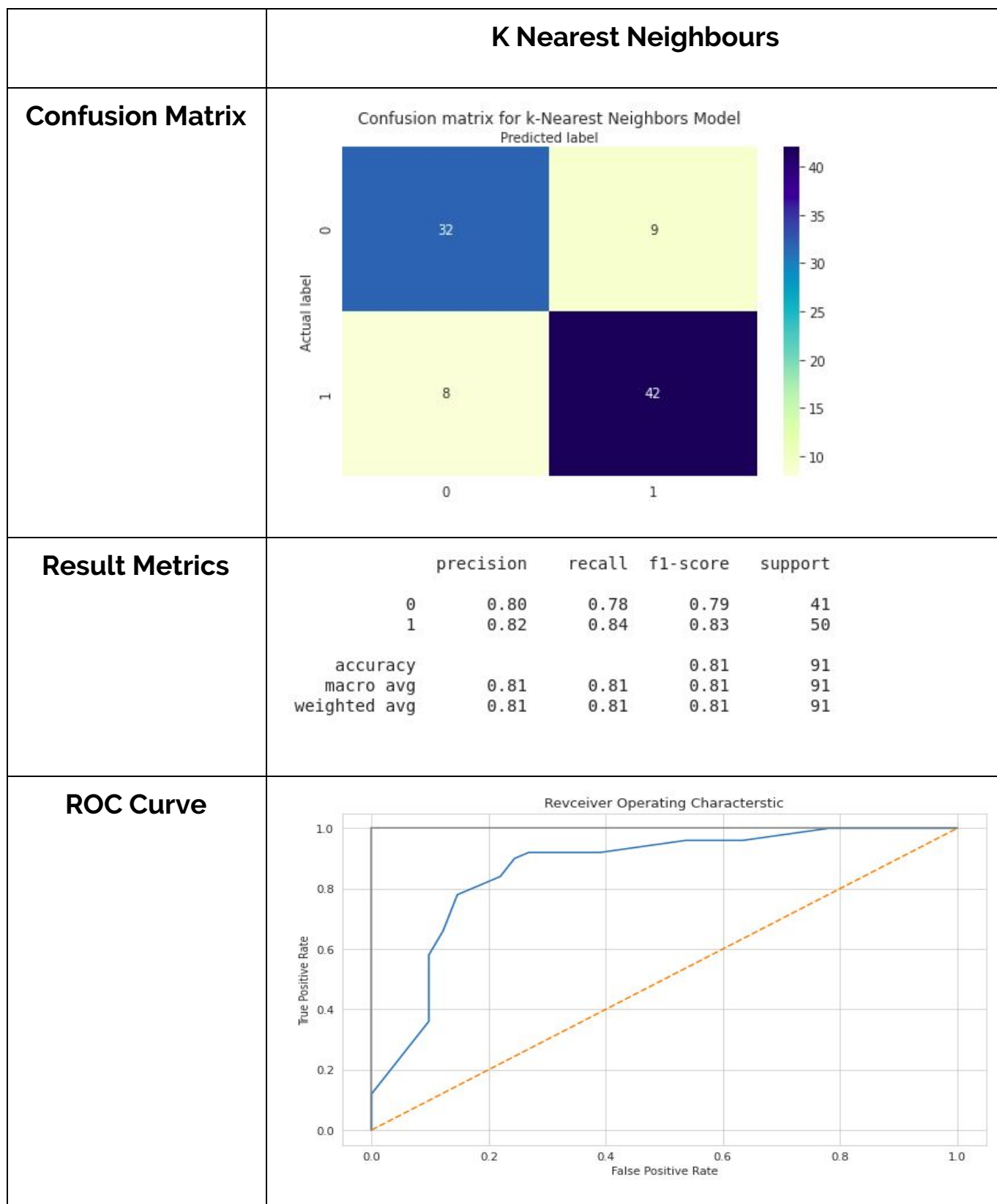
A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.

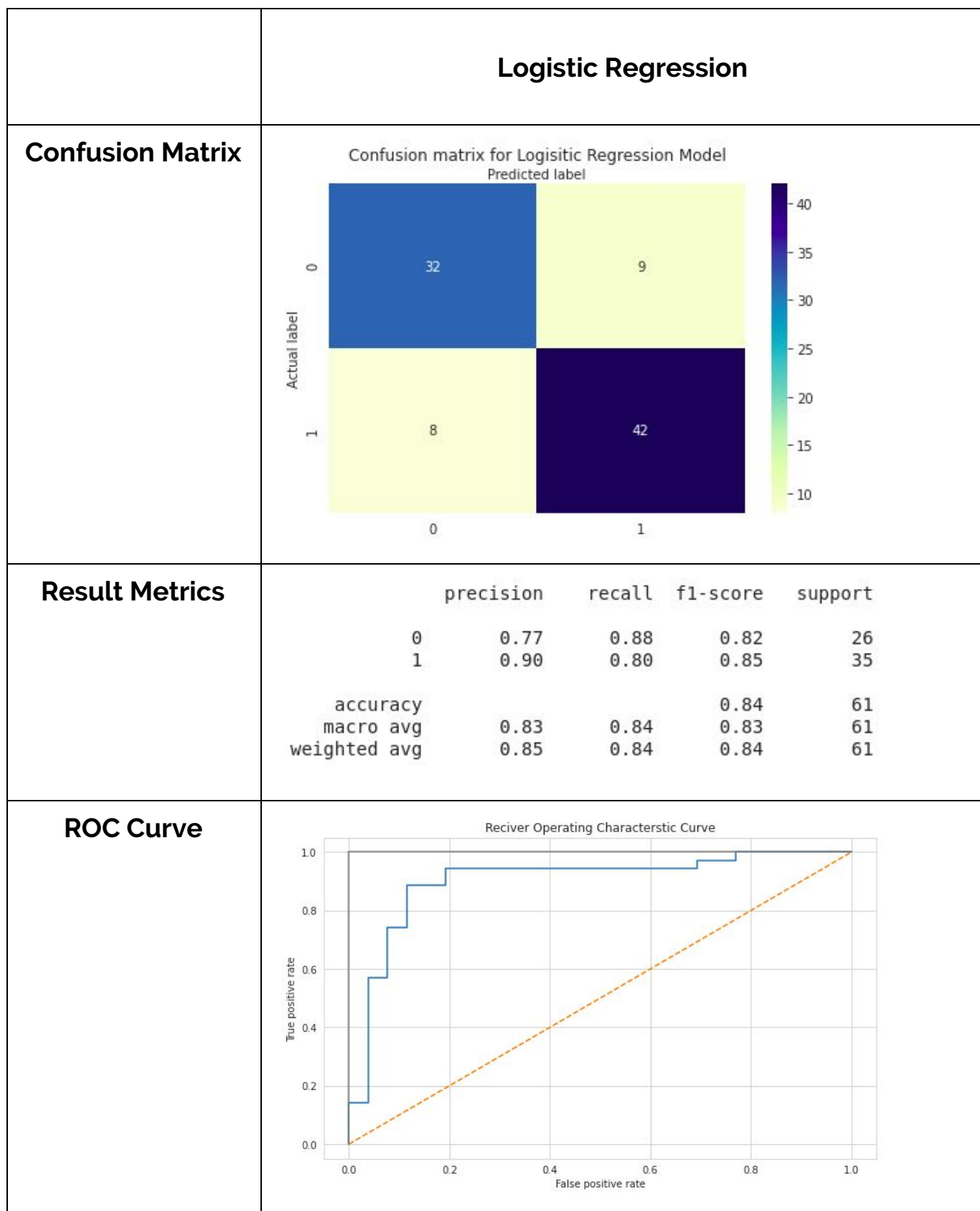
A decision tree is a flowchart-like structure in which each internal node represents a “test” on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). The paths from the root to the leaf represent classification rules.

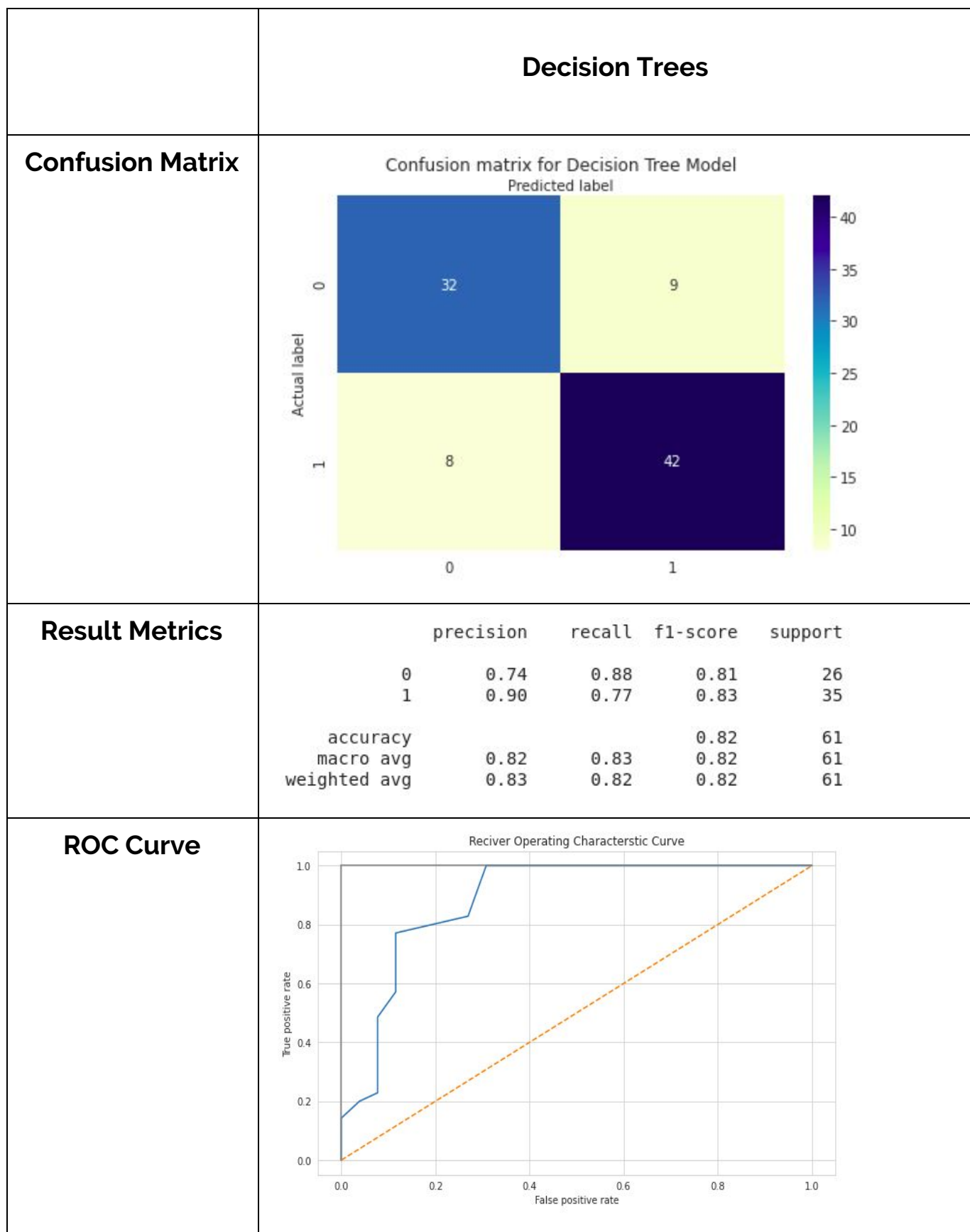
### 3.4 Neural Network


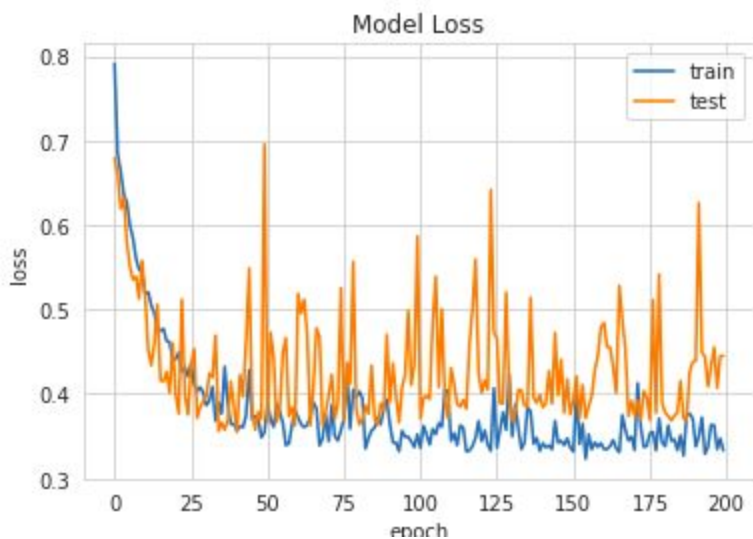
Layer ( type )	Output Shape
Dense	16
Dense	8
Dense	2

Epochs	200
Batch Size	10
Verbose	10







	<h2>Neural Networks</h2>																														
<b>Model Accuracy</b>	 <p>The graph titled "Model Accuracy" displays the performance of a neural network over 200 epochs. The y-axis represents accuracy, ranging from 0.5 to 0.9. The x-axis represents the epoch number, ranging from 0 to 200. Two lines are plotted: a blue line for training accuracy and an orange line for test accuracy. Both lines show a rapid increase in accuracy during the first 25 epochs, after which they fluctuate between 0.8 and 0.9. The training accuracy is generally slightly higher than the test accuracy, though they are very close for much of the training process.</p> <table><tr><th>epoch</th><th>train</th><th>test</th></tr><tr><td>0</td><td>0.45</td><td>0.55</td></tr><tr><td>25</td><td>0.80</td><td>0.85</td></tr><tr><td>50</td><td>0.85</td><td>0.85</td></tr><tr><td>75</td><td>0.85</td><td>0.85</td></tr><tr><td>100</td><td>0.85</td><td>0.85</td></tr><tr><td>125</td><td>0.85</td><td>0.85</td></tr><tr><td>150</td><td>0.85</td><td>0.85</td></tr><tr><td>175</td><td>0.85</td><td>0.85</td></tr><tr><td>200</td><td>0.85</td><td>0.85</td></tr></table>	epoch	train	test	0	0.45	0.55	25	0.80	0.85	50	0.85	0.85	75	0.85	0.85	100	0.85	0.85	125	0.85	0.85	150	0.85	0.85	175	0.85	0.85	200	0.85	0.85
epoch	train	test																													
0	0.45	0.55																													
25	0.80	0.85																													
50	0.85	0.85																													
75	0.85	0.85																													
100	0.85	0.85																													
125	0.85	0.85																													
150	0.85	0.85																													
175	0.85	0.85																													
200	0.85	0.85																													
<b>Model Loss</b>	 <p>The graph titled "Model Loss" displays the loss function values for the neural network over 200 epochs. The y-axis represents the loss, ranging from 0.3 to 0.8. The x-axis represents the epoch number, ranging from 0 to 200. Two lines are plotted: a blue line for training loss and an orange line for test loss. Both lines show a general downward trend. The training loss decreases from approximately 0.8 to 0.35. The test loss decreases from approximately 0.7 to 0.4, but it exhibits much higher volatility than the training loss, with several sharp spikes reaching up to 0.6 or 0.7.</p> <table><tr><th>epoch</th><th>train</th><th>test</th></tr><tr><td>0</td><td>0.80</td><td>0.70</td></tr><tr><td>25</td><td>0.45</td><td>0.45</td></tr><tr><td>50</td><td>0.40</td><td>0.40</td></tr><tr><td>75</td><td>0.38</td><td>0.40</td></tr><tr><td>100</td><td>0.35</td><td>0.40</td></tr><tr><td>125</td><td>0.35</td><td>0.40</td></tr><tr><td>150</td><td>0.35</td><td>0.40</td></tr><tr><td>175</td><td>0.35</td><td>0.40</td></tr><tr><td>200</td><td>0.35</td><td>0.40</td></tr></table>	epoch	train	test	0	0.80	0.70	25	0.45	0.45	50	0.40	0.40	75	0.38	0.40	100	0.35	0.40	125	0.35	0.40	150	0.35	0.40	175	0.35	0.40	200	0.35	0.40
epoch	train	test																													
0	0.80	0.70																													
25	0.45	0.45																													
50	0.40	0.40																													
75	0.38	0.40																													
100	0.35	0.40																													
125	0.35	0.40																													
150	0.35	0.40																													
175	0.35	0.40																													
200	0.35	0.40																													

## 4. Result

---


All the models that we used produced satisfactory results. The result of all the models is summarised in the table below

Model	Accuracy
KNN	81.319 %
Logistic Regression	90.109 %
Decision Tree	88.571 %
Neural Network	87.371 %

## 5. Previous work done in this field

For providing appropriate results and making effective decisions on data, some advanced data mining techniques are used. The five data mining classifying algorithms, with large datasets, have been utilized to assess and analyze the risk factors statistically related to heart diseases in order to compare the performance of the implemented classifiers (e.g., Naive Bayes, Decision Tree, Discriminant, Random Forest, and Support Vector Machine). An effective heart disease prediction system (EHDPS) is developed using a neural network for predicting the risk level of heart disease. The system uses 15 medical parameters such as age, sex, blood pressure, cholesterol, and obesity for prediction. The EHDPS predicts the likelihood of patients getting heart disease. It enables significant





knowledge, eg, relationships between medical factors related to heart disease and patterns, to be established.

The experiment was carried out on a publicly available database for heart disease. The dataset contains a total of 303 records that were divided into two sets, training set (40%) and testing set (60%). A data mining tool named Weka 3.6.11 was used for the experiment. Additionally, a multilayer perceptron neural network (MLPNN) with backpropagation (BP) was used as the training algorithm. MLPNN is one of the most significant models in artificial neural networks. The MLPNN consists of one input layer, one or more hidden layers and one output layer.<sup>3</sup> In MLPNN, the input nodes pass values to the first hidden layer, and then nodes of first hidden layer pass values to the second and so on till producing outputs.

The data are collected from a standard dataset that contains 303 records. The 15 parameters, such as age, sex, chest pain type (CP), and cholesterol (chol), with some domain values associated with them, considered to predict the probability of heart disease. In this study, an EHDPS has been presented using data mining techniques. From ANN, an MLPNN together with BP algorithm is used to develop the system. The MLPNN model proves the better results and assists the domain experts and even the person related to the medical field to plan for a better and early diagnosis for the patient.


Numerous data mining & machine learning methods have been discussed and examined for predicting heart disease. It analyses and concludes that the data mining algorithm i.e. ANN and SVM that used UCI Heart disease is a chronic disease due to which a number of people are suffering which has become a great deal of attention. It analyses and concludes that the data mining algorithm i.e. ANN and SVM that used UCI Repository dataset perform better for heart disease prediction than the remaining algorithms.

## **6. Future Work**

In today's modern world, cardiovascular disease is the most lethal one. This disease attacks a person so instantly that it hardly gets any time to get treated with. So diagnosing patients correctly on a timely basis is the most challenging task for the medical fraternity. A wrong diagnosis by the hospital leads to earning a bad name and losing reputation. At the same time treatment of the said disease is quite high and not affordable by most of the patients particularly in India. The purpose is to develop a cost effective treatment using data mining technologies for facilitating database decision support systems.

Almost all the hospitals use some hospital management system to manage healthcare in patients. Unfortunately most of the systems rarely use the huge clinical data where vital information is hidden. As these systems create huge amounts of data in varied forms but this data is seldom visited and remains untapped. So, in this direction lots of efforts are required to make intelligent decisions. The diagnosis of this disease using different features or symptoms is a complex activity.

Future work will involve the amalgamation of the various specified algorithms to augment the accuracy so that the diagnosis can develop into more accurate in case of imperceptibly identified data sets. The machine learns patterns from the existing dataset, and then applies them to an unknown dataset in order to predict the outcome. Classification is a powerful machine learning technique that is commonly used for prediction.



Some classification algorithms predict with satisfactory accuracy, whereas others exhibit a limited accuracy. A method termed ensemble classification, which is used for improving the accuracy of weak algorithms by combining multiple classifiers. Experiments with this tool were performed using a heart disease dataset. A comparative analytical approach was done to determine how the ensemble technique can be applied for improving prediction accuracy in heart disease. The focus is not only on increasing the accuracy of weak classification algorithms, but also on the implementation of the algorithm with a medical dataset, to show its utility to predict disease at an early stage.

The results of the study indicate that ensemble techniques, such as bagging and boosting, are effective in improving the prediction accuracy of weak classifiers, and exhibit satisfactory performance in identifying risk of heart disease. A maximum increase of 7% accuracy for weak classifiers was achieved with the help of ensemble classification. The performance of the process was further enhanced with a feature selection implementation, and the results showed significant improvement in prediction accuracy.

The echocardiography offers two-dimensional images during examination. Unfortunately, the images generated from each examination are not stored by the hospital instead; they are discarded as soon as the examination is over.

The hospital should find a way to store the image so that they can be used to extract relevant information related to the disease using intelligent image recognition systems.

As a future work, the planning is to perform additional experiments with more dataset and algorithms to improve the classification accuracy and to build a model that can predict specific heart disease types.

## **7. REFERENCES**

1. **Towards Data Science** - <https://towardsdatascience.com/heart-disease-prediction-73468d630cfc>
2. **NCBI** - <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5863635/>
3. **Kaggle** - <https://www.kaggle.com/c/heart-disease>
4. **DovePress** - <https://www.dovepress.com/ensemble-approach-for-developing-a-smart-heart-disease-prediction-syst-peer-reviewed-fulltext-article-RRCC>
5. **AlliedAcademics** - <https://www.alliedacademies.org/articles/prediction-of-heart-disease-using-knearest-neighbor-and-particle-swarm-optimization.html>