

# Heart Disease Prediction

Adimulam Hemalatha<sup>1</sup>, Sai Krishna<sup>2</sup>, Rajesh Reddy<sup>3</sup>,  
Vivekananda Chary<sup>4</sup>, KN Soujanya<sup>5</sup>

1-4 UG Student, Dept. of Computer science Engineering, GITAM UNIVERSITY, Visakhapatnam, Andhra Pradesh, India.

5 Assistant Professor, Dept. of Computer science Engineering, GITAM UNIVERSITY, Visakhapatnam, Andhra Pradesh, India.

Date of Submission: 12-04-2023

Date of Acceptance: 22-04-2023

## ABSTRACT

Cardiovascular diseases are the main cause of death in both the developed and underdeveloped countries as well as emerging worlds. We have two factors that can reduce mortality include clinical staff supervision and early detection of heart conditions.

We require more knowledge, time so it not always possible to effectively more patients throughout the day and doctor cannot consult with a patient for full 24 hours.

In this project, We have Created a model that predicts Whether a patient will have heart disease in ten years based on different features using the logistic regression of patient with dataset made accessible to the public on the kaggle website, utilizing Machine learning techniques including some algorithms

Early detection of this diseases can enable those who having at high risk decide on lifestyle changes that will minimize issues

## I. LITERATURE REVIEW

According to Tom Mitchell, machine learning is "intended to be a computer virus that learns from prior experience and certain tasks and a few performances on, as judged by, increases with experience." Because machine learning involves a combination of correlations and connections, the majority of current machine learning algorithms are focused on identifying and/or exploiting dataset interrelationships.

Arthur Samuel presented a fresh perspective on machine learning in 1959. The focus of machine learning is on the study and development of algorithms that can learn from data and make predictions about data. Procedure statistics, which also concentrates on computing predictions, is closely related to machine learning. The concept of mathematical optimization is

closely related to the industrial processes, theories, and application fields that it supplies.

## II. INTRODUCTION:

According to World Health Organization it estimates that heart disease causes more over 12 million deaths per Each year, Since a few years ago, the cardiovascular disease has been raising very Quickly around the World so we have to identify the most important factors

That is risk factors for the heart disease and to estimate the total risk.

Heart disease is also known as a silent killer since it causes a person to death without any evident signs.

In Order to avoid problems in high-risk individuals and make decisions about lifestyle modifications, early detection of heart disease is crucial

## PROBLEM DEFINATION:

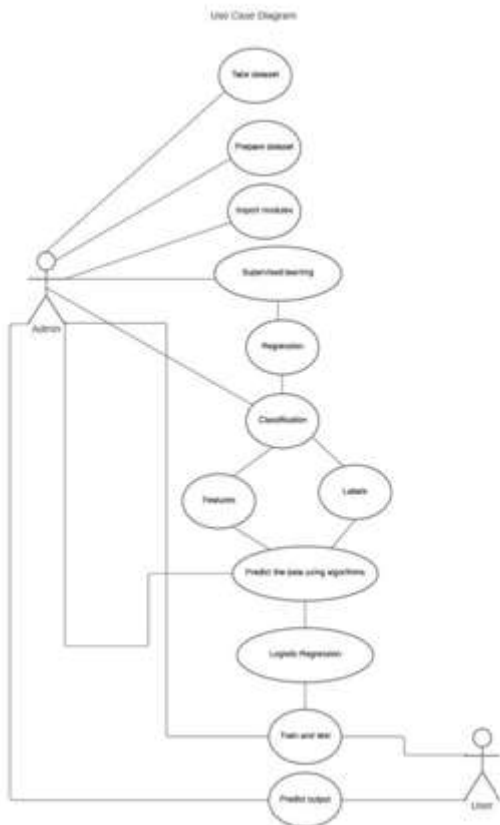
Finding a cardiac problem is the biggest challenge. Although there is technology that can predict heart disease, it is either costly or inefficient for estimating the likelihood of heart disease in humans. Early detection of heart diseases can lower the mortality rate and overall implications. It is not always possible to properly monitor patients every day, and a doctor cannot discuss with a patient for a whole 24 hours because it requires more intellect, time, and knowledge. With the availability of modern data, we may use a variety of machine learning algorithms to look for hidden patterns. The hidden patterns in medical data may be used to make health diagnoses.

## MOTIVATION:

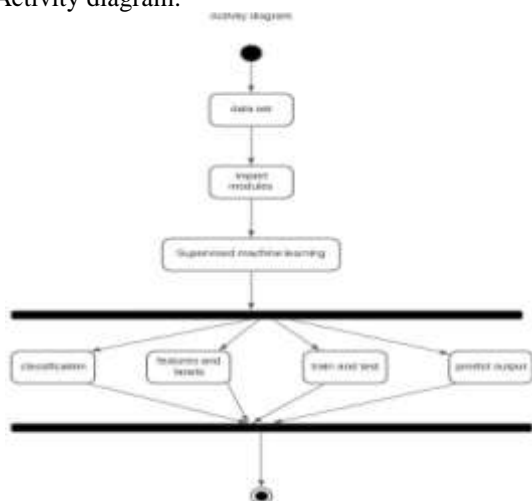
Applications for data science have compared and used machine learning algorithms for analysis in various contexts. The main objective of this research-based study was to examine the

feature selection, data preparation, and processing methods employed in machine learning training models. The challenge we now face with empirical models and libraries is data. In addition to their amount and our cooked models, the accuracy we see throughout training, testing, and actual validation has a higher variation. This is the result look at the models' fundamental presumptions and then use a Logistic Regression model

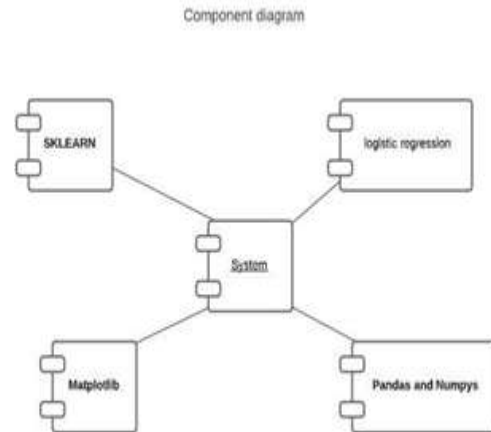
**ARCHITECTURAL DESIGN**



Activity diagram:



Component Diagram:



**III. OVERVIEW OF ALGORITHMS:**

**1. Logistic Regression**

When attempting to estimate the likelihood that an event will occur or not based on a collection of input features, a statistical model known as logistic regression is frequently utilised. Based on a variety of clinical data and risk factors, logistic regression can be used to forecast the likelihood that a patient would develop heart disease. The logistic function's curve shows the likelihood of anything, such whether the cells will develop heart disease or not. The dependent variable in this problem is whether we can predict that the patient has heart disease or not heart disease) by (Independent variables are features in our data set)

**Decision Tree**

Decision tree is a graphical representation like tree we used to calculate entropy in different combinations for our attributes Based on entropy values which have high we consider them and best entropy decision tree is formed with them . Entropy means purity or highest success rate for the single attributes or combination of attributes we have choosed In decision tree entropy is main dependent variable in classifying the tree After finding best decision tree we will train and test the data and find accuracy .

**K-Nearest Neighbor(KNN)**

The K-Nearest Neighbors (KNN) algorithm is a machine learning algorithm used for classification and regression tasks. The classification or regression of a new data point is based on its k-nearest neighbors in the training dataset. Here the value of k is a hyperparameter

that specifies the number of neighbors to consider. KNN calculates the distance between the new data point and all the training data points using a distance metric such as Euclidean distance. Then it will take k nearest neighbors based on distance between them, distance must be less between them. After getting k nearest neighbors then we take most repeated value from them and classify the new data.

### Support Vector Machine

The SVM algorithm aims to construct the optimum decision boundary or line that can divide n-dimensional space into classes so that we can quickly classify fresh data points in the future. A hyperplane is the name given to this optimal decision boundary. SVM selects the extreme vectors and points that aid in the creation of the hyperplane. Support vectors are used in situations like this. The margin is the separation between each class's nearest data points and the hyperplane. The SVM model performs more well in terms of generalisation the wider the margin. The coefficients of the hyperplane are computed using these support vectors, and predictions are then made for brand-new data points.

### FORMULAS:

The correctness of our output from our training data was evaluated using a "Confusion matrix" analysis.

Confusion matrix:

This is an Error matrix, It is used to characterize the model based on the performance. The important thing of this concept is to count both corrects and incorrect guesses, categorizing them rather than just counting errors.

Table 1: Confusion Matrix as a consequence of Data Training (feature selection by backward elimination)

TP=3569	FP=27
FN=599	TN=45

Table 1: Confusion Matrix Obtained after training the data (feature selection by backward elimination)

TP=3582	FP=14
FN=600	TN=44

Table 2: Confusion Matrix Obtained after training the data (feature selection by RFECV method)

The accuracy is calculated as:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

Where:

- True Positive (TP) = Observation is positive and is anticipated to be +ve.
- False Negative (FN) occurs when an observation is +ve but the result is projected to be -ve.
- True Negative (TN) = Prediction is -ve and observation is -ve.
- False Positive (FP) Means observation is -ve yet result is anticipated to be +ve

With backward elimination applied after feature selection, the obtained accuracy was 86% during training and 83% during testing.

After training the data utilising the feature selection process and the REFCV technique, the achieved accuracy was 86% during testing.

Recall:

The proportion of all positively categorised examples that were properly classified to all positively classified examples. High /More Recall means the class has successfully identified (a small number of FN). Recall is determined by:

$$\text{Recall} = \frac{TP}{TP+FN}$$

After utilising backward elimination to pick features, the recall attained during training and testing was 0.99.

Following feature selection using the REFCV approach, the acquired recall during training the data was 1.00, and during testing it was 0.99.

Precision:

To calculate the precision value, we divide the total number of successfully detected positively categorized cases by the total no of positively prediction Examples,

A +ve example will always be +ve, according to high precision. The recipe for accuracy is:

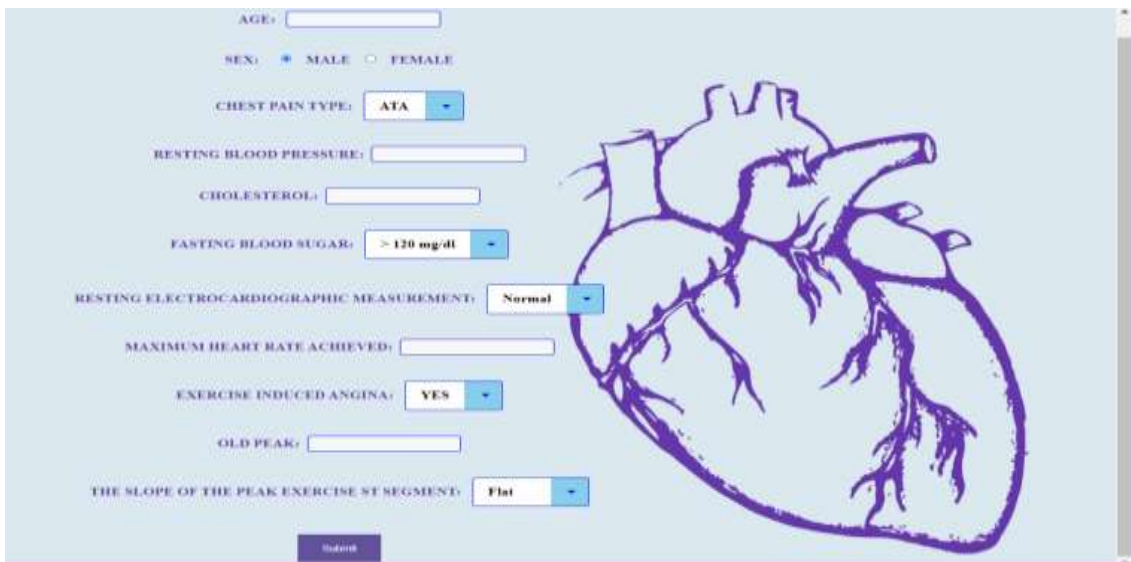
When training the data using backward elimination after feature selection, the gained accuracy is 0.86, and when testing it was 0.84.

After testing the data using the REFCV technique after features selection, the achieved accuracy was 0.86.

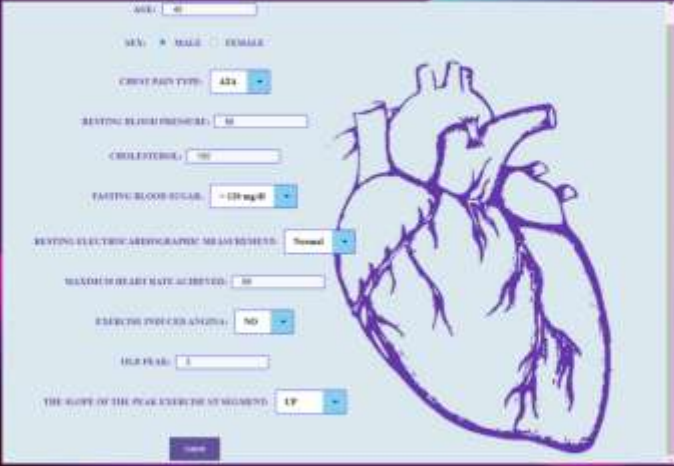
OUTPUT:

We used flask application imported from python library to create our webpage Flask is a popular Python web framework that allows developers to build web applications quickly and easily. Our webpage code is written in HTML &

CSS using tags like title, link, form, input, select, option, bold (b), line break (br) link tag is used to link the css file to html We created a form like structure to take inputs for patient details In form we used input tag to take inputs like age, sex, cholesterol, restingbp etc. Select tag is used to select like cholesterol types to select in different options



## Webpage input & output



You dont have a heart disease

#### **IV. CONCLUSION:**

An important development in medicine is the ability to assist high risk patients in making decisions about lifestyle changes that will minimize issues by detecting cardiovascular abnormalities early .By tackling the features selection problems , or backward elimination and RFECV ,underlying the models , this effort successfully predicted heart disease with 85% accuracy . This model that was used was logistics regression we might use more sophisticated models and train on them to foresee various cardiovascular issues whilealso advising consumers to further enhance it.

#### **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, "towarrrdatascience.com," 13 Feb 2019. [Online]. Available: <https://towardsdatascience.com/predicting-presence-of-heart-diseases-using-machinelearning-36f00f3edb2c>. [Accessed 2 March 2020].
- [4]. [Online]. Available: <https://www.kaggle.com/ronitf/heart-disease-uci#heart.csv>. [Accessed 05 December 2019].
- [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".