**Heart Disease Prediction**

*A project report submitted to ICT Academy of Kerala*

*in partial fulfillment of the requirements*

*for the certification of*

**CERTIFIED SPECIALIST**

**IN**

**DATA SCIENCE & ANALYTICS**

submitted by

**Nayana J B**

**Nidhi Anil**

**Nimmi Nelson**

**Sreelakshmi Vijayan C**

**Geena Tharian**



**ICT ACADEMY OF KERALA**

**Table of Contents**

**Abstract**

**1.Problem statement**

1.1 Overview

1.2 Problem Statement

**2. Introduction**

**3. Literature Survey**

**4.Methods and algorithms used**

4.1 Collection of dataset

4.2 Exploratory data analysis

4.2.1 Data visualization techniques

4.3 Preprocessing

4.3.1 Null Value handling

4.3.2 Outlier detection and handling

4.4 Algorithms used

4.4.1 Logistic Regression

4.4.2 K-NN algorithm

**5. CODING**

5.1 Software requirements

5.1 Dataset

5.2 Null value handling

5.3 Outlier detection and handling

5.4 Encoding

5.5Splitting the dataset

5.6 Modeling

5.7 Website hosting

**6. CONCLUSION AND FUTURE SCOPE**

**Abstract**

Machine Learning is used across many ranges around the world. The healthcare industry is no exception. Machine Learning can play an essential role in predicting presence/absence of locomotors disorders, Heart diseases and more.According to the [CDC](https://www.cdc.gov/heartdisease/risk_factors.htm), heart disease is one of the leading causes of death for people of most races in the US (African Americans, American Indians and Alaska Natives, and white people). About half of all Americans (47%) have at least 1 of 3 key risk factors for heart disease: high blood pressure, high cholesterol, and smoking. Other key indicators include diabetic status, obesity (high BMI), not getting enough physical activity or drinking too much alcohol.Detecting and preventing the factors that have the greatest impact on heart disease is very important in healthcare. . We work on predicting possible Heart Diseases in people using Machine Learning algorithms. In this project we perform the different data analysis steps including exploratory data analysis ,preprocessing ,predictive modeling ,fine tuning,and web hosting to detect "patterns" from the data that can predict a patient's condition.

**1. Problem Definition**

**1.1 Overview**

The main motivation of doing this research is to present a heart disease prediction model for the prediction of occurrence of heart disease. Further, this research work is aimed towards identifying the best classification algorithm for identifying the possibility of heart disease in a patient. This work is justified by performing a comparative study and analysis using three classification algorithms

**1.2 Problem Statement**

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 the chance of heart disease in humans. Early detection of cardiac diseases can decrease the mortality rate and overall complications. However, it is not possible to monitor patients everyday in all cases accurately and consultation of a patient for 24 hours by a doctor is not available since it requires more patience, 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.

**2. Introduction**

According to the World Health Organization, every year 12 million deaths occur worldwide due to 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 data analysis. The load of cardiovascular disease is rapidly increasing all over the world from the past few years. Many researches have been conducted in an 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 reduces the complications. Machine learning proves to be effective in assisting in making decisions and predictions from the large quantity of data produced by the healthcare industry. This project aims to predict future Heart Disease by analyzing data of patients which classifies whether they have heart disease or not using a machine-learning algorithm. 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 analyzing to extract the desired data we can say that this technique can be very well adapted to do the prediction of heart disease.

**3. Literature Survey**

With growing development in the field of medical science alongside machine learning various experiments and research has been carried out in recent years releasing the relevant significant papers.

[1] Purushottam ,et al proposed a paper “Efficient Heart Disease Prediction System” using hill climbing and decision tree algorithms .They used Cleveland dataset and preprocessing of data is performed before using classification algorithms. The Knowledge Extraction is done based on Evolutionary Learning (KEEL), an open source 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 86.7%.

[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 the decision tree algorithm the tree is built using certain conditions which give True or False decisions. The algorithms like SVM, KNN are results based on vertical or horizontal split conditions depending on dependent variables. But a decision tree for a tree-like structure having root nodes, leaves and branches based on the decision made in each of tree Decision tree also helps in the understanding of the importance of the attributes in the dataset. They have also used the Cleveland data set. Dataset splits in 70% training and 30% testing by using some methods. This algorithm gives 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 87% 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 classifiers that are used especially in the prediction of Heart Disease. 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(HRFM).

[4] Anjan N. Repaka et al, proposed a model stating 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.

[5] 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 a major number of rules helps within the best prediction of heart disease. Frequent pattern growth association mining is performed on the patient's dataset to generate strong associations.

**4.Methods And Algorithms used**

**4.1 Collection of dataset**

Initially, we collected a dataset for our heart disease prediction system. The dataset

used for this project Heart Disease comes from the CDC and is a major part of the Behavioral Risk Factor Surveillance System (BRFSS), which conducts annual telephone surveys to gather data on the health status of U.S. residents. The dataset consists of 19 columns and 319795 rows.

**4.2 Exploratory data analysis**

Exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods.

The main underlying principles of an EDA are-

* The aim should be to uncover information that should lead to showing patterns and trends.
* Missing values and outliers need to be given proper consideration
* The relationship between different variables must be established.
* A suitable technique of variate analysis should be chosen for the target to be achieved

**4.2.1 Data visualization techniques**

The different data visualization techniques used are:

* Countplot
* Pie Chart
* Histogram
* Bar Graph

**4.3 Preprocessing**

Pre-processing refers to the transformations applied to our data before feeding it to the algorithm. Data preprocessing is a technique that is used to convert the raw data into a clean data set.

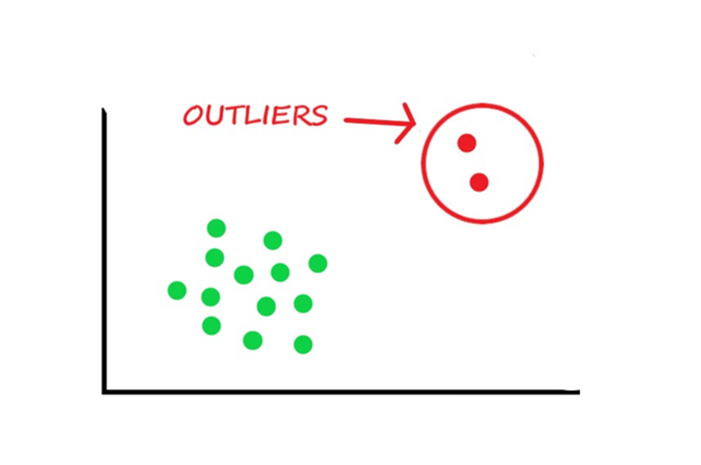
Different preprocessing include

**4.3.1 Null Value handling**

NULL is the term used to represent a missing value.There are many ways you can remove or impute missing values in a dataset. How you handle missing values also depends on the type of variable, i.e., numerical or categorical, as some methods work only on one kind of variable.It is important to handle missing values as they can lead to inaccurate conclusions about the data, which can significantly impact the accuracy of the analysis.There are several methods available to handle missing values, such as removal, imputation, flagging, etc. You can use libraries available in Python language to handle missing values in your dataset.

**4.3.2 Outlier detection and handling**

Outliers are those data points that are significantly different from the rest of the dataset. They are often abnormal observations that skew the data distribution, and arise due to inconsistent data entry, or erroneous observations. Can use the box plot, or the box and whisker plot, to explore the dataset and visualize the presence of outliers. The points that lie beyond the whiskers are detected as outliers.Outliers can be treated in different ways, such as trimming, capping, discretization,droppingor by treating them as missing values. Empirical relations are used to detect outliers in normal distributions, and Interquartile Range (IQR) is used to do so in skewed distributions.



**4.3.3 Encoding**

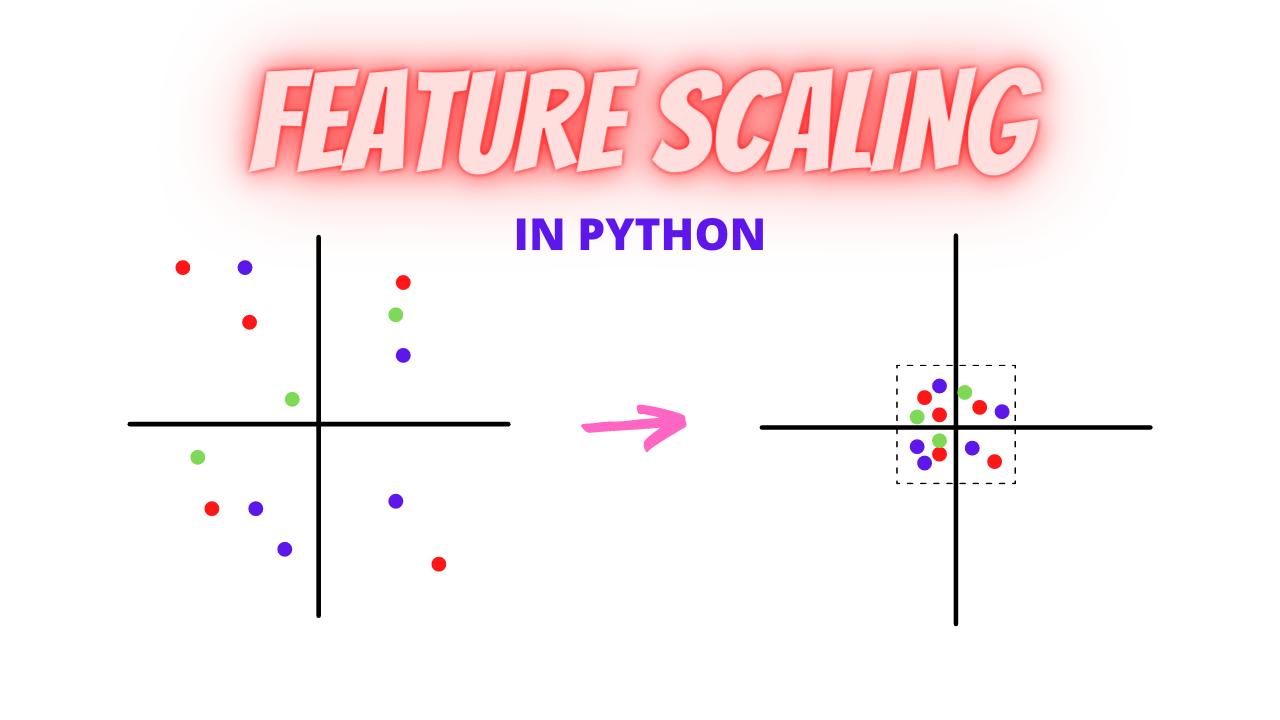
Machine learning models can only work with numerical values. For this reason, it is necessary to transform the categorical values of the relevant features into numerical ones. This process is called feature encoding.

Ways to Encode Features for Machine Learning Algorithms

* One-hot/dummy encoding.
* Label / Ordinal encoding.
* Target encoding.
* Frequency / count encoding.
* Binary encoding.
* Feature Hashing.

**4.3.4 Scaling**

Feature scaling is the process of normalizing the range of features in a dataset. Real-world datasets often contain features that are varying in degrees of magnitude, range, and units. Therefore, in order for machine learning models to interpret these features on the same scale, we need to perform feature scaling.

****

**4.4 Algorithms used**

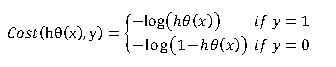
The main purpose of designing this system is to predict the ten-year risk of future heart disease.We have used Logistic regression as a machine-learning algorithm to train our system and various feature selection algorithms like Backward elimination and Recursive feature elimination. These algorithms are discussed below in detail.

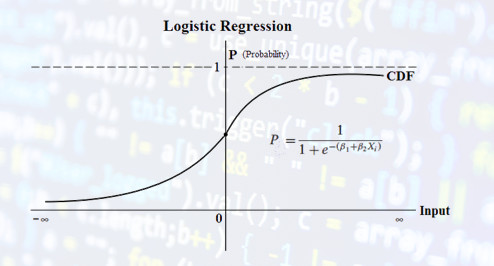
**4.4.1 Logistic Regression**

Logistic Regression is a supervised classification algorithm. It is a predictive analysis algorithm based on the concept of probability. It measures the relationship between the dependent variable.(TenyearCHD) and the one or more independent variables (risk factors) by estimating probabilities using the underlying logistic function (sigmoid function). Sigmoid function is used as a cost function to limit the hypothesis of logistic regression between 0 and 1 (squashing)

i.e. 0 ≤ hθ(x) ≤ 1.

In logistic regression cost function is defined as:

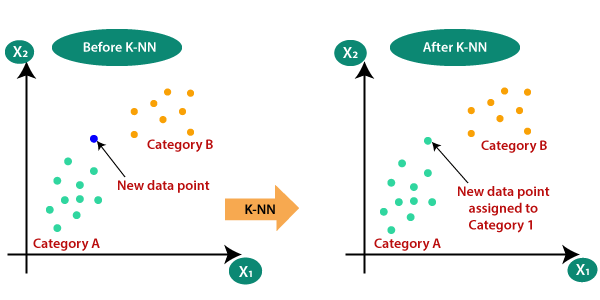




Logistic Regression relies highly on the proper presentation of data. So, to make the model more powerful, important features from the available data set are selected using Backward elimination and recursive elimination techniques.

**4.4.2 K-NN algorithm**

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique. The K-NN algorithm assumes the similarity between the new case/data and available cases and puts the new case into the category that is most similar to the available categories.

.

**5. CODING**

**5.1 Software Requirements**

**1. Anaconda-**

It is an open source software available to us which enables us to easily code using python or R on different operating systems such as Windows, Linux, and Mac OS. It has millions of users worldwide, and is well known as the industry which helps us in developing systems, testing them, and training the machines. This further enables us to:

• Manage all the imported libraries, their dependencies, and the environments of developing with Anaconda.

• In developing techniques to train our machine with TensorFlow, scikit-learn, etc.

• Analyze the datasets and manipulate the with Dask, NumPy, pandas, and Numba

• Visualize or plot the results with Matplotlib, Holoviews, Bokeh, and Datashader.

It also provides us with jupyter notebooks which have all the in-built libraries embedded in the already. This eases our coding stress, and also helps us code with more efficiency.

**2. Python**

The most abundantly used general level programming language. It is used for both a small scale and big scale systems. It can easily be interpreted. It is said to support multiple programming paradigms. It includes features of procedural, object-oriented, and functional programming together. It is already garbage-collected which makes it more efficient.

**3. Numpy**

It is a python programming library. This basically helps us deal with large datasets, matrices, and multi-dimensional arrays. It also provides us with a number of mathematical functions which help us and ease the calculations. It is an open-source software available to all.

**4. Pandas**

- It is a library that is written in python language. It helps us with the analysis of data. It also provides us with tools and functions to manipulate a large amount of data. 25 5. Sklearn- It is a library used in machine learning in the python programming language. It mainly helps us with the classification of data, regression of models, and in clustering algorithms. These algorithms include SVM, random forest etc.

**5. Sklearn-**

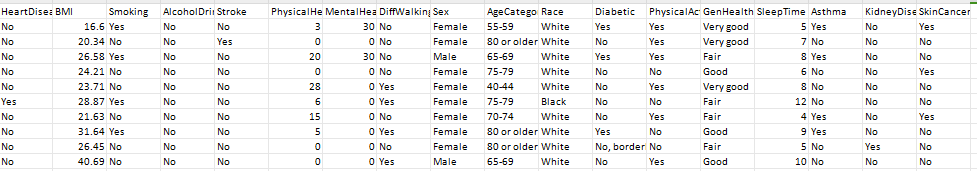
It is a library used in machine learning in the python programming language. It mainly helps us with the classification of data, regression of models, and in clustering algorithms. These algorithms include SVM, random forest etc.

**6.spyder**

Spyder, the Scientific Python Development Environment, is a free integrated development environment (IDE) that is included with Anaconda. It includes editing, interactive testing, debugging, and introspection features

**5.2 Dataset**

Originally, the dataset comes from the CDC and is a major part of the Behavioral Risk Factor Surveillance System (BRFSS), which conducts annual telephone surveys to gather data on the health status of U.S. residents. As the [CDC](https://www.cdc.gov/heartdisease/risk_factors.htm) describes: "Established in 1984 with 15 states, BRFSS now collects data in all 50 states as well as the District of Columbia and three U.S. territories.

****

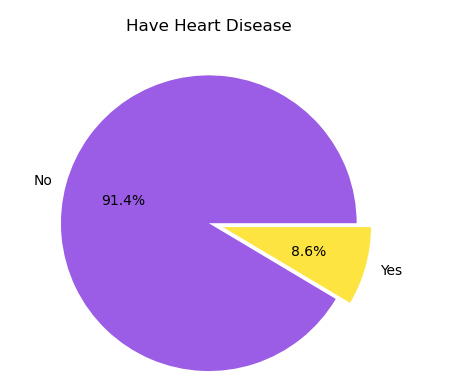
**Dataset**

It provides patient information which includes over 319,795 records and 18 attributes. The attributes include: The attributes include: HeartDisease, BMI, Smoking, AlcoholDrinking, Stroke,

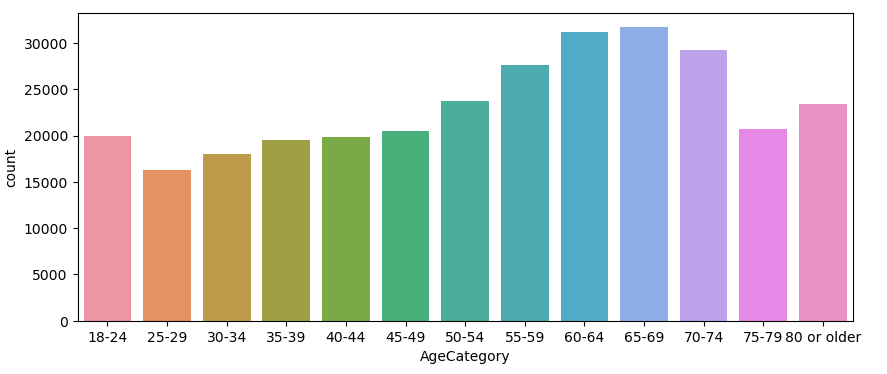
PhysicalHealt, MentalHealth, DiffWalking, Sex, AgeCategory, Race, Diabetic, PhysicalActivity, GenHealth, SleepTime,Asthma, KidneyDisease, SkinCancer.

**5.3 Exploratory data analysis**

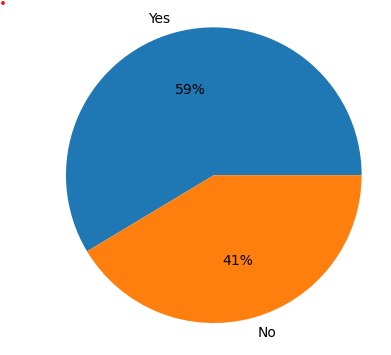
* 91.4% of patients have no heart disease and only 8.6% have heart disease.



* Most patients' ages are between 65 to 70 and least patients' ages are 25 to 29.



* 59% of patients that have heart disease have smoked at least 100 cigarettes in their entire life (5 packs = 100 cigarettes.



**5.4 Null Value handling**

HeartDisease 0

BMI 0

Smoking 0

AlcoholDrinking 0

Stroke 0

PhysicalHealth 0

MentalHealth 0

DiffWalking 776

Sex 360

AgeCategory 0

Race 1141

Diabetic 0

PhysicalActivity 0

GenHealth 796

SleepTime 0

Asthma 0

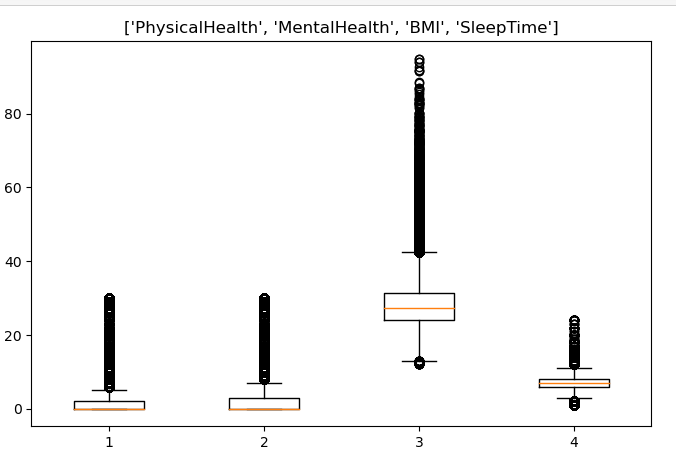
KidneyDisease 0

SkinCancer 0

There are null values in Difficulty in walking,sex,Race,General health.Updated null values with mode of values of each column because they all are object type.



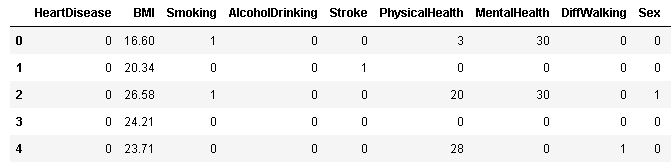
**5.5 Outlier detection and handling**

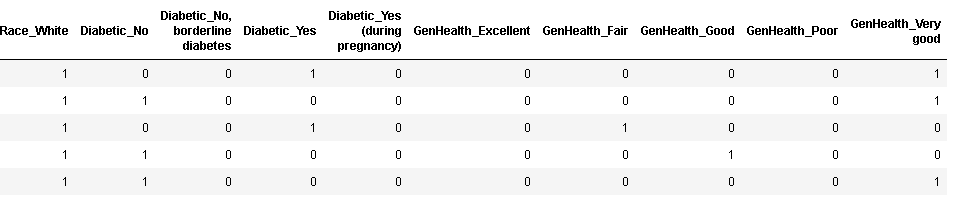
****

Using Boxplot, find outliers and apply the IQR method to handle outliers.

**5.6 Encoding**

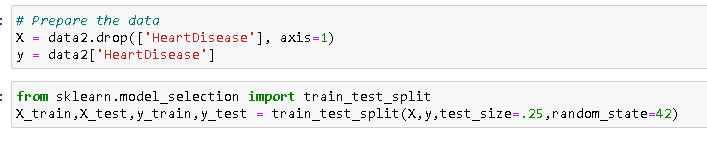
one−hot encoding and label encoding are two extensively used strategies in machine learning for encoding categorical data. One hot encoding generates binary columns for each category, whereas label encoding provides each category a unique numeric label.





**5.7 Splitting the dataset**

The train-test split is a technique for evaluating the performance of a machine learning algorithm. It can be used for classification or regression problems and can be used for any supervised learning algorithm.



**5.8 Modeling and prediction**

Machine learning models are computer programs that are used to recognize patterns in data or make predictions. Machine learning models are created from machine learning algorithms.After splitting the data we apply different classification algorithms such as logistic regression, KNN,SVM,Random Forest etc**.**Inorder to know if we really have the best model, we need to look at the distribution of the accuracy measures. It is similar to the idea of marginal interpretation vs conditional interpretation: is the model good on average or is it always performing well, no matter the data.We find logistic regression model has more accuracy and it is best fit to this data.

**5.9 Website hosting**

Now, Flask is a Python-based micro framework used for developing small-scale websites. Flask is very easy to make Restful APIs using python. As of now, we have developed a model i.e model.pkl , which can predict a class of the data based on various attributes of the data. Now we will design a web application where the user will input all the attribute values and the data will be given to the model. Based on the training given to the model, the model will predict what should be the heart disease state of the person whose details have been fed.

In order to collect the data, we create an HTML form which would contain all the different options to select from each attribute. Here, we have created a simple form using HTML only. If you want to make the form more interactive you can do so as well.

**6. Conclusion And Future scope**

Heart diseases are a major killer in India and throughout the world. Application of promising technology like machine learning to the initial prediction of heart diseases will have a profound impact on society. The early prognosis of heart disease can aid in making decisions on lifestyle changes in high-risk patients and in turn reduce the complications, which can be a great milestone in the field of medicine. The number of people facing heart diseases is on a raise each year. This prompts for its early diagnosis and treatment. The utilization of suitable technology support in this regard can prove to be highly beneficial to the medical fraternity and patients. In this paper, the several different machine learning algorithms used to measure the performance are SVM, Random Forest,, Logistic Regression,KNN applied on the dataset. The expected attributes leading to heart disease in patients are available in the dataset which contains 17 features. and that are useful to evaluate the system are selected among them. Using these attributes we predicted the heart disease chance of a patient.Along with it we developed a website which can use patients . its user friendly and can help to change patients' lifestyles.