

# VISUALIZING AND PREDICTING HEART DISEASES WITH AN INTERACTIVE DASH BOARD

Team Id: PNT2022TMID44038

Project Design phase-1

Solution Architecture

Faculty Mentor:

Dhrisya.S

Team Leader:Roshini.K

Team Member:Sinthiya.M

Team Member:Sugantharani.F

Team Member:Swetha.K

## Collection of dataset:

Initially, we collect a dataset for our heart disease prediction system. After the collection of the dataset, we split the dataset into training data and testing data. The training dataset is used for prediction model learning and testing data is used for evaluating the prediction model. For this project, 70% of training data is used and 30% of data is used for testing. The dataset used for this project is Heart Disease UCI. The dataset consists of 76 attributes; out of which, 14 attributes are used for the system.

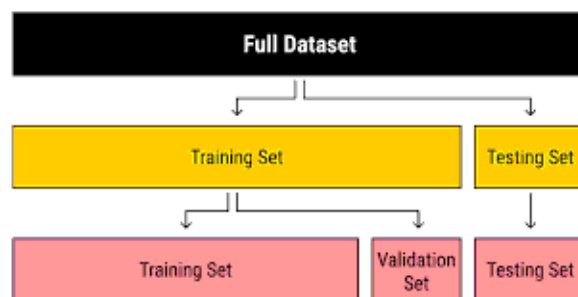


Figure: Collection of dataset

## Selection of attributes:

Attribute or Feature selection includes the selection of appropriate attributes for the prediction system. This is used to increase the efficiency of the system. Various attributes of the patient like gender, chest pain type, fasting blood pressure, serum cholesterol, exang, etc are selected for the prediction. The Correlation matrix is used for attribute selection for this model.

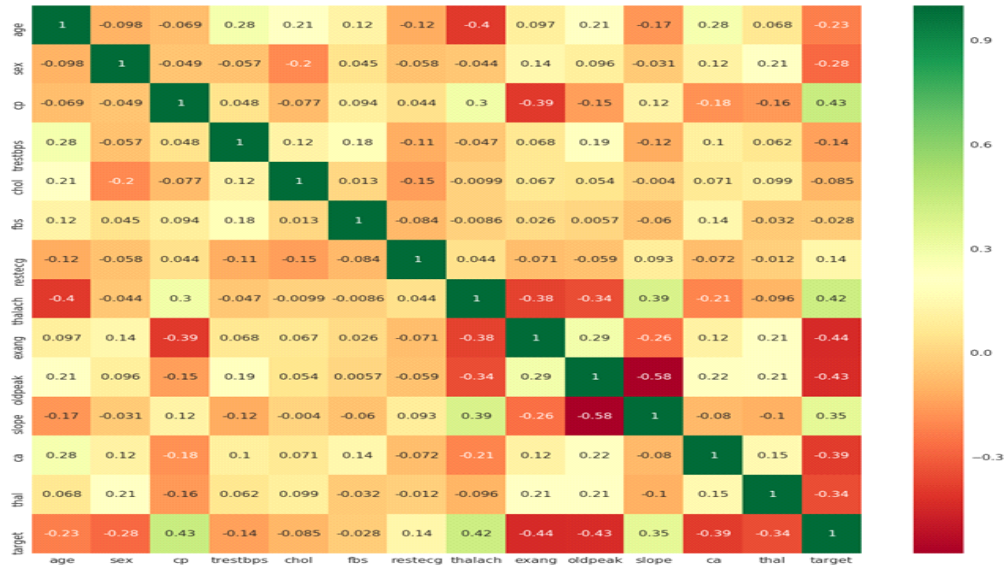


Figure: Correlation matrix

## Pre-processing of Data:

Data pre-processing is an important step for the creation of a machine learning model. Initially, data may not be clean or in the required format for the model which can cause misleading outcomes. In pre-processing of data, we transform data into our required format. It is used to deal with noises, duplicates, and missing values of the dataset. Data pre-processing has the activities like importing datasets, splitting datasets, attribute scaling, etc. Preprocessing of data is required for improving the accuracy of the model.

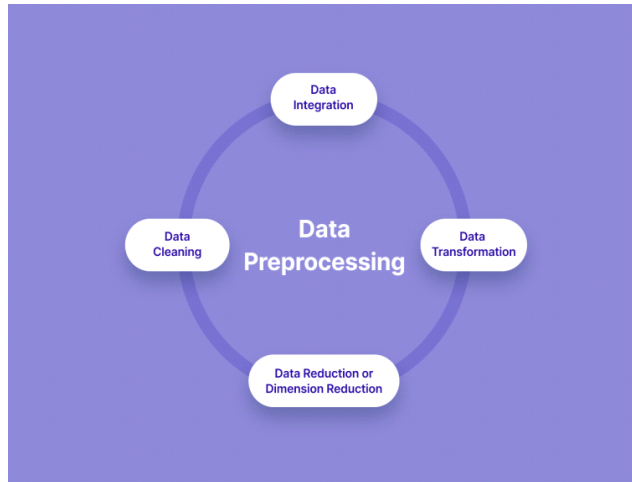


Figure: Data pre-processing

### Balancing of Data:

Imbalanced datasets can be balanced in two ways. They are Under Sampling and Over Sampling

(a) **Under Sampling:** In Under Sampling, dataset balance is done by the reduction of the size of the ample class. This process is considered when the amount of data is adequate.

(b) **Over Sampling:** In Over Sampling, dataset balance is done by increasing the size of the scarce samples. This process is considered when the amount of data is inadequate.

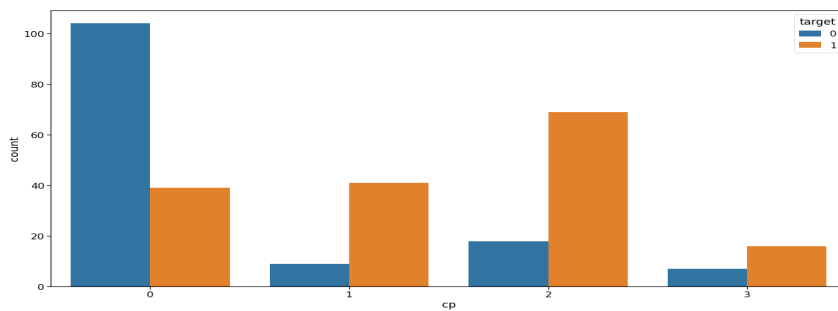


Figure: Data Balancing

### Prediction of Disease:

Various machine learning algorithms like SVM, Naive Bayes, Decision Tree, Random Tree, Logistic Regression, Ada-boost, Xg-boost are used for classification. Comparative analysis is performed

among algorithms and the algorithm that gives the highest accuracy is used for heart disease prediction.

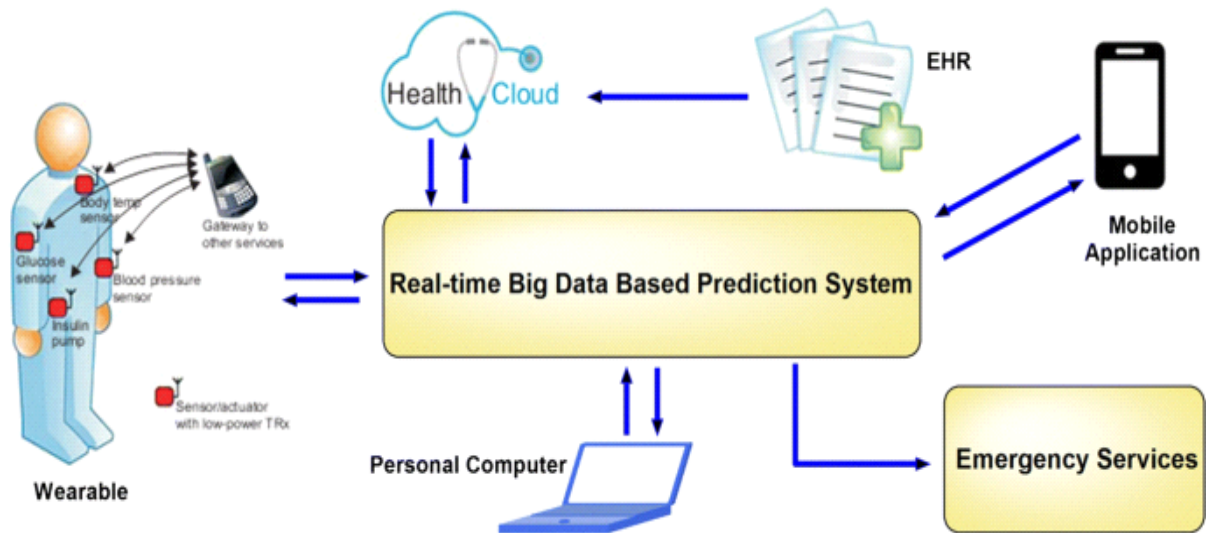


Figure: Prediction of Disease

### Architecture Diagram:

