

Heart Disease

**Project Description**: 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. Many people suffer from heart disease, which even cost people their lives all around the world. Machine learning can be used to detect whether a person is suffering from a heart disease by considering certain attributes like chest pain, cholesterol level, age of the person and some other attributes. Such information, if predicted well in advance, can provide important insights to doctors who can then adapt their diagnosis and treatment per patient basis.

**Objective:** Predict if the patient has heart disease or not from the given features**.**

**1- Steps of Data Preprocessing:**

* **Data collection:** we have collected the datausing “ Heart\_Disease.csv ”sent by the college to predict the patient’s having heart disease or not.
* **Data cleansing:** data cleansing went through some steps **:**

Categorical: filled with mode**.**

**1 - Checking Nulls:** Data found to have nulls

Numerical: filled with mean.

“Approximate the decimal numbers of numerical data into integers”

**2 - Checking Outliers:** Data found to have outliers 🡪 check interquartile range “**IQR**” and replace the outliers with the mean.

**3 -** **Checking Duplicates:** No Duplicates found**.**

**4 – Converting Categorical data into Numerical data.**

**And by this step data preprocessing is Done!**

**Risk factors**

**2- Steps of feature extraction:** There is two types of columns found

**Indicators.**

**Risk factors:** Having risk factors doesn’t guarantee the development of heart disease but increases the likelihood**.**

**Ex**: “id, age, gender, etc…”

**Indicators:** Indicators in healthcare are signs or measurements that provide information about a specific condition.

**Ex**:” chest pain type, EKG results, Max HR, etc…”

**From feature extraction 🡪 Risk factor was dropped.**

**From feature extraction 🡪 Visualization of correlation.**

**And by this step Feature extraction is Done!**

**3- Steps of Model Training:** First of all data was split into 4 variables: (x\_train ” 80 % ” , x\_test ” 20 % ” , y\_train ” 80 % ” , y\_test ” 20 % ” ) **; where x : data & y : Heart disease :**

1. **Train using Logistic regression 🡪 with accuracy percentage of “92.5 %”.**

1. **Train using SVM 🡪 with accuracy percentage of “92.5 %”.**
2. **Train using Decision Tree 🡪 with accuracy percentage of “81.5 %”.**

**4- Steps of Model evaluation: Logistic regression** and **SVM** was found with the best accuracy with percentage of **“92.5 %”**

• *Confusion matrix*: **26 true positive & 0 false negative & 4 false positive & 24 true negative**.

*• Classification report*: Explained the confusion matrix.

*• Mean square error*: The closer Mean Square Error (MSE) is to 0, the better the model can fit the training data.

**And our mean square error is “0”.**

**Our Project: https://colab.research.google.com/drive/1uM7hf\_m-l1f\_ltzD3RIjiKsUwG6GZC3i?usp=sharing**