A
Major Project Report
on

Heart Disease Predictor

Submitted for the partial fulfillment of the requirement for the award of the degree of

Bachelor of Technology
in
COMPUTER SCIENCE & ENGINEERING

Submitted to:
Prof. Sanjeev Khambra
Deptt. of CSE
GJUS&T, Hisar

Submitted by:
Name- Daksh Manu Arya
Roll no- 190010130031
B.Tech (IT) – 8th Sem

Department of Computer Science & Engineering
Guru Jambheshwar University of Science & Technology, Hisar
'A' Grade NAAC Accredited
2019-2023

CANDIDATE'S DECLARATION

I/we, hereby declare that the project work entitled " Heart Disease predictor" is an authentic work carried out by me/us under the guidance of Prof. Sanjeev Khambra , Department of Computer Science & Engineering in partial fulfillment of the requirement for the award of the degree of Bachelor of Technology in Computer Science & Engineering and this has not been submitted anywhere else for any other degree.

Date: Signature:

Daksh Manu Arya

(190010130031)

PLAGIARISM CERTIFICATE

This is to certify that *Daksh Manu Arya* (*190010130031*) is a student of B.Tech (CSE), Department of Computer Science & Engineering, Guru Jambheshwar University of Science & Technology, Hisar has completed the project entitled "*Heart Disease Predictor*".

His complete project report has been checked by Turntin Software and the similarity index is _____ i.e. the accepted norms of the university. The project report may be considered for the award of the degree.

Supervisor

Signature
Daksh Manu Arya
(190010130031)

Contents

Topic of Major Project (Centered)

Page No

Introduction

Heart disease describes a range of conditions that affect your heart. Heart Diseases 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 called as "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 the 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.

According to a news article, heart disease proves to be the leading cause of death for both women and men. The article states the following :

About 610,000 people die of heart disease in the United States every year–that's 1 in every 4 deaths.1

Heart disease is the leading cause of death for both men and women. More than half of the deaths due to heart disease in 2009 were in men.1

Coronary Heart Disease(CHD) is the most common type of heart disease, killing over 370,000 people annually.

Every year about 735,000 Americans have a heart attack. Of these, 525,000 are a first heart attack and 210,000 happen in people who have already had a heart attack.

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.

DATASET

Dataset by Heart Disease UCI:

Dataset source: https://archive.ics.uci.edu/ml/datasets/Heart+Disease

Dataset columns:

- **age:** The person's age in years
- **sex:** The person's sex (1 = male, 0 = female)
- **cp:** chest pain type

  — Value 0: asymptomatic

  — Value 1: atypical angina

  — Value 2: non-anginal pain

  — Value 3: typical angina
- **trestbps:** The person's resting blood pressure (mm Hg on admission to the hospital)
- **chol:** The person's cholesterol measurement in mg/dl
- **fbs:** The person's fasting blood sugar (> 120 mg/dl, 1 = true; 0 = false)
- **restecg:** resting electrocardiographic results

  — Value 0: showing probable or definite left ventricular hypertrophy by Estes' criteria

  — Value 1: normal

  — Value 2: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
- **thalach:** The person's maximum heart rate achieved
- **exang:** Exercise induced angina (1 = yes; 0 = no)
- **oldpeak:** ST depression induced by exercise relative to rest ('ST' relates to positions on the ECG plot. See more here)
- **slope:** the slope of the peak exercise ST segment — 0: downsloping; 1: flat; 2: upsloping

  0: downsloping; 1: flat; 2: upsloping
- **ca:** The number of major vessels (0–3)
- **thal:** A blood disorder called thalassemia Value 0: NULL (dropped from the dataset previously

  Value 1: fixed defect (no blood flow in some part of the heart)

Value 2: normal blood flow

Value 3: reversible defect (a blood flow is observed but it is not normal)

- **target:** Heart disease (1 = no, 0= yes)

LIBRARIES USED

```
import tkinter as tk
from tkinter import *
import numpy as np
import pandas as pd
from pandas import *
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import RandomizedSearchCV
from sklearn.metrics import accuracy_score,recall_score,precision_score,f1_score
import warnings
warnings.filterwarnings("ignore")
from sklearn.metrics import confusion_matrix
from sklearn.model_selection import cross_val_score
from sklearn.metrics import classification_report
```

Data Visualization

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. Additionally, it provides an excellent way for employees or business owners to present data to non-technical audiences without confusion.

1. Countplot

```
df2['target'] = df2['target'].apply(chng2)
sns.countplot(data= df2, x='sex',hue='target')
plt.title('Gender v/s target\n')
```

There are four types of chest pain, asymptomatic, atypical angina, non-anginal pain and typical angina. Most of the Heart Disease patients are found to have asymptomatic chest pain. These factors include:

· Age

· Diabetes

· Excess weight

· Family History of Heart Disease

· High Blood Pressure

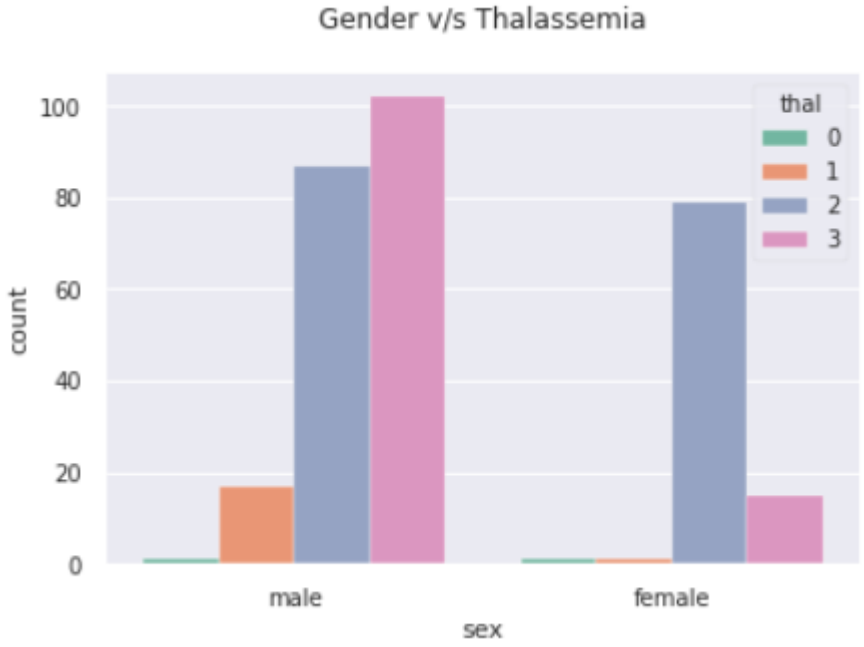
· High cholesterol

· Lack of exercise

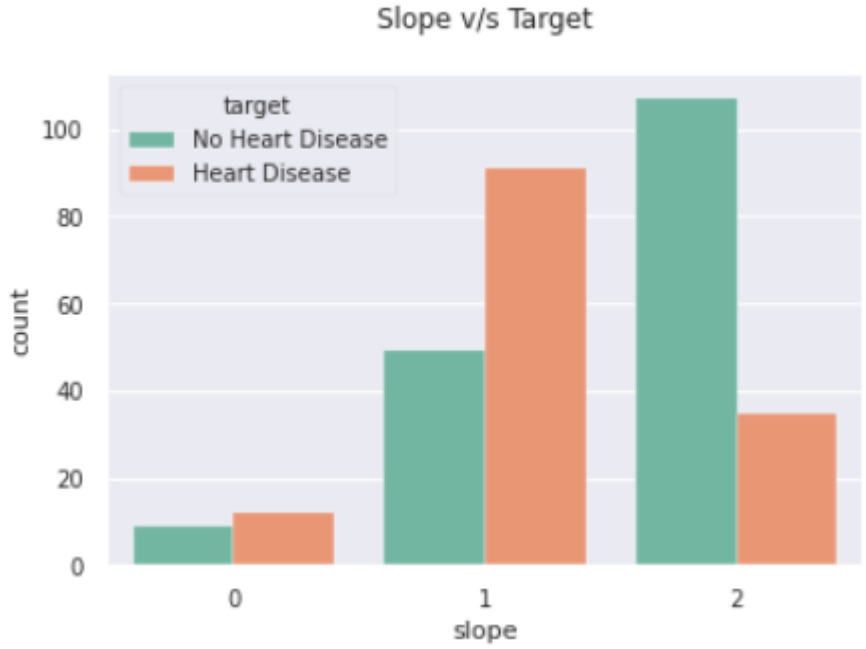
· Prior Heart attack

· Tobacco use

```
sns.countplot(data= df2, x='sex',hue='thal')
plt.title('Gender v/s Thalassemia\n')
print('Thalassemia (thal-uh-SEE-me-uh) is an inherited blood disorder that causes your body to have less hemoglobin than normal. Hemoglobin enables red blood cells to carry oxygen')
```




```
sns.countplot(data= df2, x='slope',hue='target')
plt.title('Slope v/s Target\n')
```

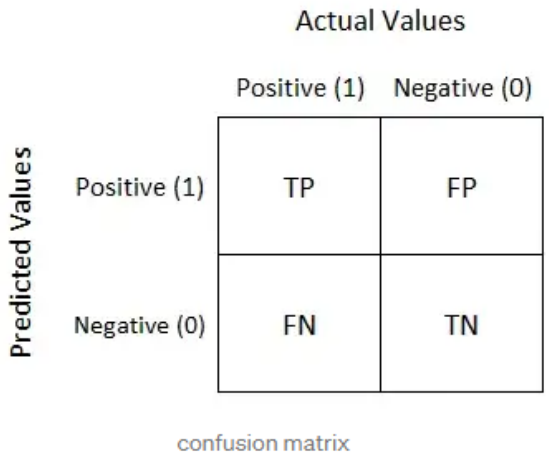


Training Data

In machine learning, training data is the data you use to train a machine learning algorithm or model. Training data requires some human involvement to analyze or process the data for machine learning use

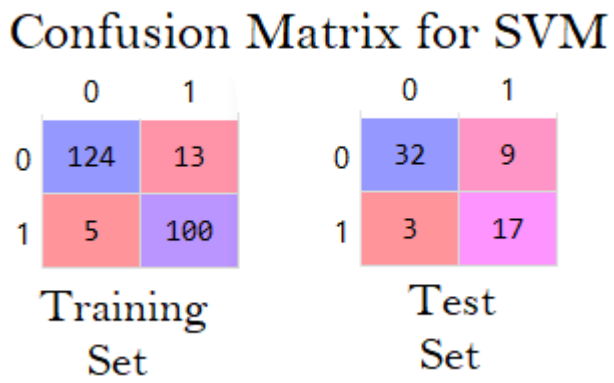
I will be using the following classification models for classification :

- SVM
- Naive Bayes
- Logistic Regression
- Decision Tree
- Random Forest
- LightGBM
- XGboost



confusion matrix

SVM

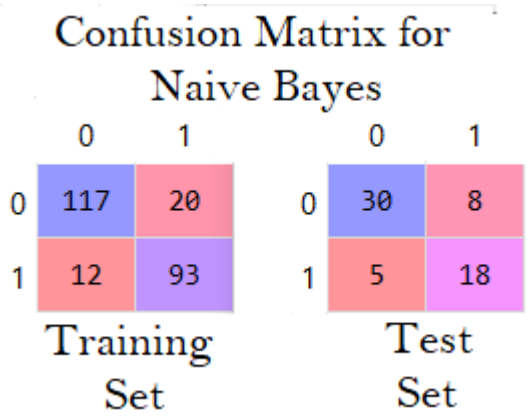


Accuracy for SVM for training set = ((124+100)/(5+13+124+100))*100 = 92.51%

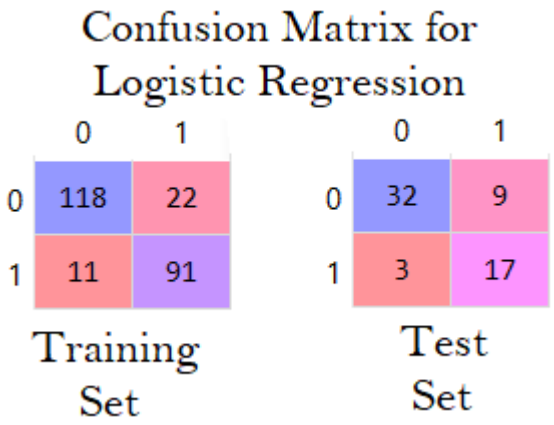
Accuracy for SVM for test set = 80.32%

Similarly let us look at all the confusion matrices for each classifier.

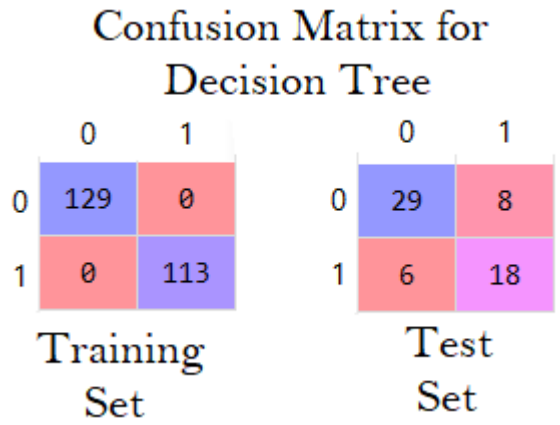
Naive Bayes



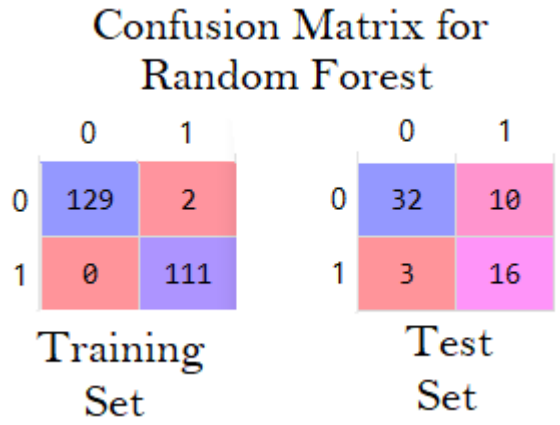
Logistic Regression



Decision Tree

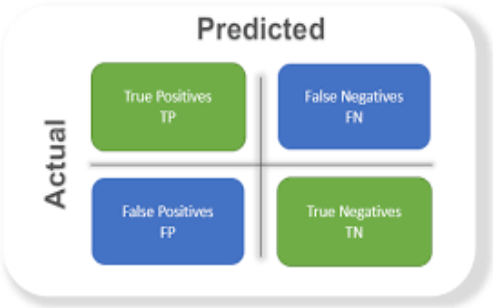


Random Forest



CHECKING ACCURACY OF MODELS

Model accuracy is defined as the number of classifications a model correctly predicts divided by the total number of predictions made. It's a way of assessing performance of model.



```
Accuracy for training set for svm = 0.9256198347107438
Accuracy for test set for svm = 0.8032786885245902

Accuracy for training set for Naive Bayes = 0.8677685950413223
Accuracy for test set for Naive Bayes = 0.7868852459016393

Accuracy for training set for Logistic Regression = 0.8636363636363636
Accuracy for test set for Logistic Regression = 0.8032786885245902

Accuracy for training set for Decision Tree = 1.0
Accuracy for test set for Decision Tree = 0.7704918032786885

Accuracy for training set for Random Forest = 0.987603305785124
Accuracy for test set for Random Forest = 0.7540983606557377

Accuracy for training set for LightGBM = 0.9958677685950413
Accuracy for test set for LightGBM = 0.7704918032786885

Accuracy for training set for XGBoost = 0.987603305785124
Accuracy for test set for XGBoost = 0.7540983606557377
```

The highest accuracy for the test set is achieved by Logistic Regression and SVM which is equal to 80.32%.
The highest accuracy for the training set is 100% achieved by Decision Tree.

OUTPUT


tk
Heart Diesease Predcition
Age: 63
Sex: 1
CP: 3
Trestbps: 145
Chol: 233
Fbs: 1
Restecg: 0
Thalach: 150
Exang: 0
Oldpeak: 2.3
Slope: 0
CA: 0
Thal: 1
submit
Result=1
Heart disease detected !!!
-PLEASE VISIT NEAREST CARDIOLOGIST AT THE EARLIEST-

Conclusion

Heart Disease is one of the major concerns for society today.

It is difficult to manually determine the odds of getting heart disease based on risk factors.

However, machine learning techniques are useful to predict the output from existing data.

Reference

- https://www.kaggle.com/ronitf/heart-disease-uci/kernels
- https://towardsdatascience.com/predicting-presence-of-heart-diseases-using-machine-learning-36f00f3edb2c#:~:text=Machine%20Learning%20can%20play%20an,and%20treatment%20per%20patient%20basis.