Heart Disease Prediction

Summary:

As the number of heart disease related cases are rising each year [1], it is important that there needs to be some action taken to prevent or slow down the growth. Heart Disease Prediction is an end-to-end Python project which aims to save lives by predicting the presence of a heart disease. This information can be conveyed to the patients early on and thus help them take preventive/reversive action sooner before the disease gets severe. This will be done by implementing a machine learning model which is trained on medical data of heart disease patients.

Introduction:

The term “heart disease” refers to several types of heart conditions. The most common type of heart disease in the United States is coronary artery disease (CAD), which affects the blood flow to the heart. Decreased blood flow can cause a heart attack [2].

17.9 million people die each year from CVDs, an estimated 32% of all deaths worldwide [3]. Despite the high number of fatalities, heart diseases are preventable and, in some cases, reversible. A person’s negligence can cost them and hence this project aims to predict the presence of heart disease and inform patients early on. CDC ranks it as the number 1 cause of death [4].

Questions:

The questions that we hope to answer from this project are as follows:

1. Does the patient gender have any relation with their chance of developing a heart disease?
2. Does age play a role in the development of heart disease?
3. What are the most important factors according to machine learning algorithms?
4. Which Machine learning algorithm performs the best at predicting heart disease existence?

Data and Methodology:

The dataset which we used in the project was a Kaggle dataset on heart disease: [LINK](https://www.kaggle.com/johnsmith88/heart-disease-dataset). This dataset was relatively clean with very few corrections needed. The only column where the corrections were made was ‘chol’ which is short for Cholesterol. This column has outliers which were cleaned. This dataset has 1025 rows in total with 13 independent features and 1 dependent target variable, making it a total of 14 columns. The quality of the data was good.

Methodology used for the development of this project was the typical ML development methodology which includes the below steps:

* Data collection
* EDA/ Data Analysis
* Data cleaning
* Scaling data
* Building models
* Evaluating models
* Deploying the best performing model

Outcomes:

We found out that count of female patients in the dataset was almost half of the male patients. However, the percentage of female patients with heart disease was much higher than its counterpart. We also found out that heart disease was most frequently observed in patients in the ages between 40-55. This was found out by analyzing the data and representing it in a visual format using graphs and charts. Later, 6 models were built to analyze and train on the data. The models were:

1. Logistic Regression
2. KNN Classification
3. Support Vector Classification
4. Naïve Bayes Classification
5. Decision Tree Classification
6. Random Forest Classification

Out of the models above, Random Forest Classification and Decision Tree Classification models both displayed 100% accuracy on the test/validation set. Decision Tree models are prone to overfitting. Hence, we went ahead with Random Forest model.

Model Deployment:

Models were deployed with the help of Pickle and Flask libraries. Pickle is an inbuilt library which allows us to convert Python objects to byte-stream data while preserving their state and vice versa. This allows us to fit models and scalers on training data and transfer the same pre-trained model & scaler to other programs.

Flask provides an API like structure to interact with the model through a webpage. The inputs can be accepted from an HTML form. These inputs are scaled and provided to the pre-trained model for predicting the output. The output is later displayed back on the HTML page.

Conclusion:

All the objectives set at the beginning of the project were achieved. All the questions were answered. We now have a trained Random Forest classification model which gave 100% accuracy on the validation/testing set. I hope this model will be useful in predicting heart diseases in patients and save lives.

Sources:

[1] - <https://www.acc.org/about-acc/press-releases/2020/12/09/18/30/cvd-burden-and-deaths-rising-around-the-world>

[2] - <https://www.cdc.gov/heartdisease/about.htm>

[3] - <https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1>

[4] - <https://www.cdc.gov/nchs/fastats/heart-disease.htm>