National University
of computer and emerging sciences

Project Report

Project title: Heart Disease Prediction

Couse: Data Mining

Names/Roll no./section.

Nayyar Shoaib Malik/19i-0528/A

Ahsan Qamar/19i-2048/A

Tabarak Sikandar/19i-0479/A

Submitted to: Sir Wassem Shahzad

Due date: 15th Dec, 2021

Abstract

The correct prediction of heart disease can prevent life threats, and incorrect prediction can prove to be fatal at the same time. In this project different machine learning algorithms are applied to compare the results and analysis of heart disease dataset (based on the blood pressure). Various promising results are achieved and are validated using accuracy and confusion matrix. The dataset consists of some irrelevant features which are handled in preprocessing, and data are also normalized for getting better results.

Introduction

Heart disease describes a range of conditions that affect heart. Today, cardiovascular diseases are the leading cause of death worldwide with 17.9 million deaths annually, as per the World Health Organization reports. Various unhealthy activities are the reason for the increase in the risk of heart disease like high cholesterol, obesity, increase in triglycerides levels, hypertension, etc. There are certain signs like the persons having sleep issues, a certain increase and decrease in heart rate (irregular heartbeat), swollen legs. All these symptoms resemble different diseases also like it occurs in the aging persons, so it becomes a difficult task to get a correct diagnosis, which results in fatality in near future.

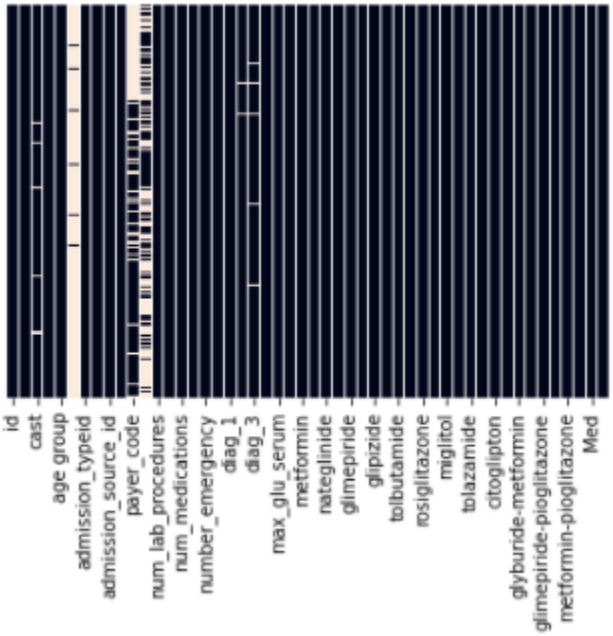
Methodology

1. **Description of dataset**

   The dataset used for this project is publicly available on Kaggle. This dataset contain almost 50 attributes including the labels as well which can be map using according to following criteria {'NO': 0, '>5' : 1, '<30' : 2}.

2. **Preprocessing of dataset**

   The dataset used contain many null values and some attributes that are not necessary for analysis so we need to preprocess the dataset. Before preprocessing the dataset following is the heatmap which shows the NULL values presents in dataset



   We filled the NULL values with mode where the datatype of attributes is object.

   Mapping:

Replace: Unknown/Invalid: mode(gender)

Replace: 'Female': 0, 'Male': 1

Mapping weight and age group column according to mean values

We did the same process for other columns and finally map the label column according to {'NO': 0, '>5' : 1, '<30' : 2}.

The last step is to remove the columns that are not important for analysis.

After all the preprocessing there is no NULL values or inconsistent data in dataset

```
copy_df.isnull().values.any()
```

```
False
```

3. Machine learning algorithms

The first step is to convert the dataset into X and Y and then into train-test. All the columns other than label column is consider as X and the label column is considered as Y. The dataset is now converted to the train test split where the test size is 30% and train size is 70%. Following is the shape of train-test split.

```
Shape of X_train =  (71236, 39)
Shape of y_train =  (71236,)
Shape of X_test =  (30530, 39)
Shape of y_test =  (30530,)
```

Standard scaler

Standard scaler is used to resize the dataset so that the mean of the observed values is 0 and standard deviation is 1 for better performance.

Logistic regression

Following is the results obtain from logistic regression model.

```
Accuracy :  57.517196200458564
Report :
              precision    recall  f1-score   support

           0       0.59      0.90      0.71     16389
           1       0.52      0.26      0.35     10758
           2       0.41      0.00      0.01      3383

    accuracy                           0.58     30530
   macro avg       0.50      0.39      0.35     30530
weighted avg       0.54      0.58      0.50     30530

F1 Score :  35.45682418312022
```

Random forest

Following is the result obtain from random forest.

```
Accuracy :  59.45627251883393
Report :
              precision    recall  f1-score   support

           0       0.62      0.83      0.71     16389
           1       0.53      0.42      0.47     10758
           2       0.44      0.02      0.04      3383

    accuracy                           0.59     30530
   macro avg       0.53      0.42      0.41     30530
weighted avg       0.57      0.59      0.55     30530

F1 Score :  40.72535089685296
```

Decision tree classifier

Following is the result obtain from decision tree classifier.

```
Accuracy :  58.244349819849326
Report :
              precision    recall  f1-score   support

           0       0.61      0.86      0.71     16389
           1       0.50      0.35      0.41     10758
           2       0.00      0.00      0.00      3383

    accuracy                           0.58     30530
   macro avg       0.37      0.40      0.37     30530
weighted avg       0.50      0.58      0.53     30530

F1 Score :  37.4335359308653
```

K Nearest Neighbor

Following is the result obtain from KNN.

```
Accuracy :  50.24238453979692
Report :
              precision    recall  f1-score   support

           0       0.55      0.74      0.63     16389
           1       0.38      0.29      0.33     10758
           2       0.14      0.02      0.03      3383

    accuracy                           0.50     30530
   macro avg       0.36      0.35      0.33     30530
weighted avg       0.45      0.50      0.46     30530

F1 Score :  33.17389546880466
```

Majority voting

Finally majority voting is applied to obtain final result and below is the final result.

```
Accuracy is: 0.5939731411726171
F1 Score is: 0.3829390227399765
```

It indicates that our model testing accuracy is 59%.