FTMK

BITI2513 INTRODUCTION TO DATA SCIENCE

PROJECT TITLE: CLUSTERING HEART DISEASE PATIENT DATA

LECTURER: AP DR SHARIFAH SAKINAH SYED AHMAD

GROUP: FABULOUS FOUR

JEYSHALINI TEVOSHA	**B031810246**
PREVINA MUNUGANAN	**B031810286**
VISHWAREETA VANOO	**B031810196**
ZAITI AKTA BINTI ZAHARUDDIN	**B031810365**

Introduction

Heart disease has been around as early as the 1900s and is the number one killer of men and women. There are several factors can increase the risk of developing heart disease. Some of these factors are beyond your control like family history, but deciding to lead a healthier lifestyle will prevent this problem most of the time. Getting a regular exercise, reducing smoking and eating healthier foods are ways of reducing heart diseases.

In this project, our problem domain is Healthcare. Doctors research previous cases to figure out how to better handle their patients. A patient with a similar history of health or symptoms to a prior patient will benefit from being treated the same way. This project examines whether we could group patients together to target treatments using some machine learning techniques.

Objective

1. To cluster heart disease patients.
2. To learn the feature importance
3. To visualise the dataset
4. To explore machine learning techniques
    a) K-means clustering
    b) Hierarchical clustering
    c) Mean-shift clustering
5. To produce a clustering model with a better performance

Goal

The main goal of this project is to ensure healthy lives and promote well-being for all at all ages. ‘Good health and well-being’ is the 3rd goal of Sustainable Development Goals. As a developing computer scientist, while advancing computer science as a discipline, we can and should play a key role in helping to address societal and environmental challenges in pursuit of a sustainable future. The goal of this project is to be a part of such a responsibility.

Besides that, our goal is also to contribute the knowledge of Artificial Intelligence in Healthcare. Although Artificial Intelligence has reached higher places, medical field still does not give a warming welcome for it in our country. It is mainly because of the concern on the medical errors it might cause. The goal of our project is to produce a better model that would reduce the errors and give confidence to the healthcare.

Questions

1. What are the benefits of this project?
2. Who does the project help?
3. How clean is the data?
4. What are the important variables in the dataset that could help the clustering?
5. What is the suitable data type for each technique?
6. Do I need to standardize the variable?
7. How to increase the performance of each technique used?

Success and measurements

The success of this project is to produce a better performing model in clustering heart disease. The performance of the model is measured in following ways:

a) Model Accuracy-

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

b) Model Precision-

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$= \frac{True\ Positive}{Total\ Predicted\ Positive}$$

c) Model Recall-

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$= \frac{True\ Positive}{Total\ Actual\ Positive}$$

d) Model F1 score

$$\text{F1} = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

Data source

The dataset used of this project is “Heart Disease Data Set” from UCI Machine Learning Repository. This dataset contains 14 variables, which are:

1. Age:
2. Sex
3. Chest pain type
4. Resting blood pressure
5. Serum Cholesterol
6. Fasting blood sugar
7. Resting electrocardiographic results
8. Maximum heart rate achieved
9. Exercise induced angina
10. ST depression induced by exercise relative to rest
11. The slope of the peak exercise ST segment
12. Number of major vessels (0-3) coloured by fluoroscopy
13. thal
14. Diagnosis of heart disease

Link : https://archive.ics.uci.edu/ml/datasets/heart+Disease