PREDICTION OF HEART DISEASE USING MACHINE LEARNING TECHNIQUES

INTRODUCTION

Cardiovascular diseases or heart diseases are a range of conditions that affect the structures or function of heart and include: coronary artery disease, arrhythmias, congenital heart defects, heart valve disease, disease of the heart muscle, and heart infection. Heart diseases is one of the most prominent cause of death all around the world. According to World Health Organization, annually 17.9 million people die of heart related diseases world-wide (World Health Organization, 2017). In the United States alone, about 655,000 people die from heart disease each year (CDC, 2020). To this date, several risk factors associated with heart diseases have been identified that includes high blood pressure, diabetes, smoking/tobacco use, obesity, physical inactivity, high cholesterol level and other lipids, and kidney disease. Since the **heart diseases are associated with several contributory risk factors, it is very difficult to diagnose the heart disease on time**. In addition, the diagnosis of heart disease involves a complex combination of clinical and pathological data resulting a very high medical cost.

PROBLEM STATEMENT

Medical professionals in hospitals and various institutions, all around the world, collect data on various health related issues including heart disease. These institutions have massive medical records which are often very noisy. As a result, these huge datasets are almost impossible for human mind to comprehend. Therefore, these huge datasets with a lot of hidden information are mostly ignored and clinical decisions are made based on doctors' intuition and experience. Nevertheless, these datasets can be analyzed using various machine learning techniques that allow to develop predictive models for the presence or absence of heart related diseases accurately. Therefore**, the objective of this project is to develop a predictive model of heart disease using machine learning based on the demographic, co-morbidity, vital sign and some laboratory investigation data**. Such studies, if able to predict heart disease on time, will be **useful for healthcare professionals for accurate heart disease diagnosis and treatment**.

DATASET

For this project, I will use the original cohort dataset from Framingham heart study (FHS) which is publicly available in Kaggle (https://www.kaggle.com/amanajmera1/framingham-heart-study-dataset) in CSV format. The goal of the FHS was to identify the characteristics that are responsible for cardiovascular diseases in human. The dataset includes 4,240 records and 16 attributes. The attributes include our target response feature- presence of heart disease (TenYearCHD), demographic information such as gender, age and education level, co-morbidity such as blood pressure, stroke, hypertension, diabetes, smoking habit, vital statistics such as systolic blood pressure, diastolic blood pressure, Body Mass Index, heart rate, and other lab investigation such as total cholesterol level and glucose level.

METHODOLOGY

In this project, I will employ various machine learning algorithms such as linear regression, logistic regression, decision tree and random forest model to identify the best predictive model for the presence or likelihood of getting heart diseases. Such model will be useful for healthcare professionals for accurate heart disease diagnosis. A GitHub repo containing the report on the finding of this study along with the code will be created and shared.

REFERENCES

CDC (2020). Heart Disease in the United States. https://www.cdc.gov/heartdisease/facts.htm. Retrieved on 1/30/2021.

World Health Organization (2017) Cardiovascular diseases (CVDs). https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds). Retrieved on 1/30/2021.