Multivariate Analysis and Prediction of Heart Disease

Mengdi Chai (mchai@hsph.harvard.edu), Zhongling Tang (ztang@hsph.harvard.edu), Tingting Yan (tingtingyan@college.harvard.edu), Rajeev K Pillai (rkrishn1@bidmc.harvard.edu)

Background and Motivation:

The project group formed has a background and interest in the medical field and public health domain and wanted to explore projects relevant to the field. The group has interests in applying the data science knowledge acquired during this class to apply it in real world clinical and research settings.All of us have computational background and hope to use the deeper understanding of ML in enhancing patient care, enhancing clinical decision making and optimizing healthcare operations.We further hope to use the knowledge gained from this project to identify new research opportunities.

The topic we select is heart diseases. Heart diseases, also called Cardiovascular diseases (CVDs), are the leading cause of death and disabilities in the world. In 2019, the World Health Organization (WHO) reported that an estimated 17.9 million people died from CVDs, occupying 32% of all global deaths (2021). Cardiovascular diseases happen commonly in developed countries and become an increasingly serious problem in developing countries. Thus far, lots of reviews and research tried to explore the CVD risk factors including unhealthful dietary intake, physical inactivity, diabetes, high blood pressure, and other related causes. However, there is a lack of comprehensive analysis on how various combinations of these risk factors interact to contribute to the development of cardiovascular diseases because not all heart diseases for individuals share a single or similar etiology. Our study will fill the gap in the study of Cardiovascular diseases by highlighting the need for multivariate data analysis approaches to consider not only individual factors but also their interactions in predicting heart disease. In the study, we will aim to draw attention to finding potential causes for heart disease to provide early prevention for reducing the burden of cardiovascular disease morbidity and mortality.

Data:

The data is from the UCI Machine Learning Heart Disease Repository. It is publicly available on both Kaggle (https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data) and the UCI website (https://archive.ics.uci.edu/dataset/45/heart+disease). The dataset is multivariate, consisting of 14 variables such as age, sex, chest pain type, and more. These features will help us understand the patients' characteristics and aid in model building.

In the dataset, it contains many missing values, so we need to deal with missing value issues before we proceed to model building. In addition, some features may need to be converted to factor or numeric levels. Exploratory data analysis will be required to examine the distribution of each variable and check for outliers. A heatmap should also be constructed to avoid collinearity. For model building, we do not aim to simply implement a basic model and perform classification tasks. Instead, we aim to develop a more advanced algorithm to determine whether a patient has heart disease based on the various dimensions of data they provide. ?

Scope:

The primary objective for this project is to analyze and predict the presence of heart disease using the published dataset. In this project we aim to demonstrate the application of data preprocessing, exploratory data analysis, and multiple machine learning model build. We plan to further apply appropriate model evaluation, model selection, and cross-validation, to arrive at a robust model to deploy. The models are expected to predict the individual's likelihood of heart disease based on the data attributes (features) used for analysis.

Key Goals

Data Exploration and Preprocessing:

- Understanding the dataset structure, including features, target variables and missing values.
- Cleaning the dataset by handling missing values, outliers and incorrect data entries
- Encoding categorical variables and scaling numerical features as necessary

Exploratory Data Analysis:

- Visualizing the distribution of features and the correlation between variables
- Identifying key attributes that influence the presence of heart disease.
- Analyzing patterns and trends in patient demographics and clinical data

Model Development and Evaluation:

- Implementing classification models such as logistic regression, decision trees, random forest and any additional techniques we plan to learn during the class which is appropriate for this project.
- Comparing model performance using metrics like accuracy, precision, recall and F1 score.

Predictive Analysis and Insights:

- Identifying key factors contributing to heart disease risk based on model interpretation.
- Using model results to provide recommendations for further research or clinical decision making.

Expected Outcomes

- A comprehensive understanding of the factors contributing to heart disease in the dataset.
- Development of a reliable machine learning model that can predict heart disease presence with high accuracy.
- Insights into the importance of various medical and demographic attributes in predicting heart disease.
- A detailed report and presentation summarizing the findings, methodology, and conclusions of the project.

Relevant Papers using Same Dataset:

1. Early prediction of heart disease using deep learning approach - ScienceDirect
2. Deep Learning to Improve Heart Disease Risk Prediction | SpringerLink

Citations:

Cardiovascular diseases (CVDs). (2021). World Health Organization. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds). Accessed 25 September 2024