CS334
Dr. Joyce Ho
October 21st, 2023

Project Proposal

Title: Predicting the Risk of Heart Disease Using Key Features: A Comparative Analysis of Machine Learning Models

Group Members: Lupin Cai, Leo Dai

Description of Problem:
Heart disease is still a leading cause of death across various races in the United States that almost half of the American population contains at least one major risk factor. Although traditional factors such as high blood pressure, cholesterol, and smoking are sound, other factors like diabetes, obesity, physical inactivity, and alcohol consumption are now recognized as important indicators. With modern computational tools and techniques, there is a need to analyze and understand these risk factors better and predict heart disease earlier.

Description of Dataset:
The dataset, titled "Indicators of Heart Disease (2022 UPDATE)," is a source from CDC's annual Behaviour Risk Factor Surveillance System survey, presenting over 400000 adult responses related to health status from 2022. For this analysis, the dataset has been selected with 300 variables to 40 key indicators of heart disease, including both direct and indirect influences. The target variable "HadHeartAttack" will be treated as a binary response "yes" or "no".

References/Work Done So Far:
A logistic regression model has been develop and deployed in an application that assesses heart conditions. This application can be accessed at the link provided.
https://share.streamlit.io/kamilpytlak/heart-condition-checker/main/app.py.

Description of Tentative Plan:

Data Understanding and Exploration: Begin with an extensive exploratory data analysis to understand the variables, their distribution and correlation.

Data Preprocessing: Fill in empty slots, eliminate outliers, drop features that are closely related to each other. If possible, modifying indicator weights for machine learning.

Model Development and Evaluation: Multiple machine learning algorithms, such as knn, decision tree and neuron networks, will be evaluated. Model performance will be assessed using accuracy, precision, auc-roc curve.

Bias Analysis: Examine any inherent biases in the prediction models.

Model Deployment: If we have time, we will build an interactive web application, otherwise, we will run the scripts as examples for features input.

Conclusion:Summarize the findings, highlight the best-performing model, and provide recommendations for healthcare providers and patients based on the insights from the study.
By the end of this project, we aim to provide a comprehensive understanding of the factors leading to heart disease and to create an accurate predictive model to aid in its early diagnosis and prevention.