Project Proposal – Predicting the Risk of Heart Disease

Introduction and Background

The goal of this project is to build a machine learning model to predict which patients will develop cardiovascular disease. Cardiovascular disease, or heart disease, is the leading cause of death in the USA, resulting in 659,000 deaths each year (over 1 million including coronary heart disease). Hospitals spend more than $363 billion each year on heart disease treatment. There are certain factors that increase the risk of heart disease, such as smoking, drinking, age, and underlying conditions. Based on such factors, a logistic regression model can be constructed to predict which patients are likely to develop heart disease within a certain time period. Such a model will allow hospitals to focus their treatments and resources on those patients, saving millions of lives and millions of dollars.

Dataset

A dataset of hospital patients was compiled on Kaggle from the US Behavioral Risk Factor Surveillance System (BRFSS); this was used to build the model. The dataset is a CSV file where each row is a patient and each column is a health factor of the patient such as smoking status, BMI, daily sleep hours, etc. The original BRFSS dataset contained 300 such factors but were reduced to only the most important 20 factors. The target variable is binary (will or will not develop heart disease in 5 years), which is why logistic regression is being used.

Procedure

Exploratory data analysis (EDA) will first be conducted on the dataset to better understand the statistics of the dataset and determine if resampling is required. This will comprise a correlation matrix, histograms, and boxplots. Once insight is gained from the EDA step, the logistic regression model dataset can be trained on the data. The dataset must be split into training and testing sets, then the model can be trained on the data. Since this is a binary regression model, its performance will be tested by running predictions on the testing set and calculating accuracy. The more accurate the model, the better it can classify people at risk of developing heart disease.

Ethical Concerns

While this project's main goal was to predict heart disease and save lives, there are two concerns. The first concern is that the accuracy of this project's model must be at least 90% or it risks misidentifying heart disease. A diagnosis of heart disease is a life-changing event that causes many reactions. A false positive (diagnosis without having the disease) would result in many angry patients while a false negative (no diagnosis while having the disease) would focus treatment on the wrong people. The second concern is that insurance companies could use the model's predictions to increase premiums for the at-risk patients. This would disproportionately affect poorer demographics, especially since income and heart disease risk are inversely correlated. Great care must be taken when proceeding with this model.

References

Antipolis, S. (2021). *Machine learning predicts risk of death in patients with suspected or known heart disease.* European Society of Cardiology. Retrieved on April 16, 2022 from https://www.escardio.org/The-ESC/Press-Office/Press-releases/Machine-learning-predicts-risk-of-death-in-patients-with-suspected-or-known-heart-disease#:~:text=The%20machine%20learning%20score%20was,Pezel.

Badr, W. (2019). *Having an Imbalanced Dataset? Here Is How You Can Fix It.* Towards Data Science. Retrieved on April 16, 2022 from https://towardsdatascience.com/having-an-imbalanced-dataset-here-is-how-you-can-solve-it-1640568947eb.

Lemstra, M., Rogers, M., & Moraros, J. (2015). *Income and Heart Disease.* National Library of Medicine. Retrieved on April 16, 2022 from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4541436/#:~:text=For%20example%2C%2010.6%25%20of%20those,heart%20disease%20(Table%202).

Walensky, R. (2021). *Heart Disease Facts.* US Centers for Disease Control and Prevention. Retrieved on April 16, 2022 from https://www.cdc.gov/heartdisease/facts.htm#:~:text=One%20person%20dies%20every%2036,United%20States%20from%20cardiovascular%20disease.&text=About%20659%2C000%20people%20in%20the,1%20in%20every%204%20deaths.