CS 414: Artificial Intelligence: Logistic Regression Assignment
HEART DISEASE PREDICTION

The World Health Organization reports that heart disease claims 12 million lives annually worldwide, with half of deaths in developed nations like the United States attributed to cardiovascular diseases. Early detection of these conditions can empower high-risk individuals to make lifestyle changes and potentially reduce complications. This research aims to identify the most significant risk factors for heart disease and predict overall risk using logistic regression. The dataset, sourced from the Kaggle website, originates from an ongoing cardiovascular study in Framingham, Massachusetts. It comprises over 4,000 records with 15 attributes, including demographic, behavioral, and medical risk factors. Demographic factors encompass sex and age, while behavioral factors include smoking status and cigarette consumption. Medical history factors consider blood pressure medication use, stroke history, hypertension, and diabetes. Current medical factors involve total cholesterol, systolic and diastolic blood pressure, BMI, heart rate, and glucose levels. The primary goal is to predict the 10-year risk of coronary heart disease, a binary classification task.

For the preprocessing features such as sex and age, while behavioral factors include smoking status, cigarette consumption, blood pressure medication use, stroke history, hypertension, diabetes, total cholesterol, systolic and diastolic blood pressure, BMI, heart rate, and glucose levels are correlated and combined into one feature. For the analysis, the data is divided into training data, which is 80% of the complete data set, and testing data, this is the remaining 20% of the complete data set and the model learns the relationship of the variables from the training data only without considering the testing data. The training data is used to build

the logistic regression model, allowing it to learn the relationship between the correlated variables and 10 year risk of coronary heart disease CHD. Once the model is trained, we use the testing data to evaluate its performance and assess how well it can predict the 10-year risk of coronary heart disease based on the correlated variables.

Reference:

https://www.kaggle.com/datasets/dileep070/heart-disease-prediction-using-logistic-regression/data