Hypothetical Study: Predicting Heart Disease Using Machine Learning

For this study, we will use the Cleveland Heart Disease dataset, which contains patients' clinical and non-clinical features and a target variable indicating whether or not a patient has heart disease. Our goal is to predict the presence of heart disease based on the patient's features.

Unit I: Statistics

Strategy #2: Hypothesis Testing for Inference

To assess the significance of each feature in predicting heart disease, we can perform a hypothesis test using logistic regression. Logistic regression will allow us to evaluate the importance of each feature in predicting heart disease. Specifically, we can use the Wald test to test the null hypothesis that the coefficient of a given feature is equal to zero against the alternative that it is not zero. The Wald test allows us to test the null hypothesis that the coefficient of a given feature is equal to zero, meaning the element does not affect the outcome. If the p-value of the test is less than selected significance level (0.05), we can reject the null hypothesis and conclude that the feature is significant in predicting heart disease. We can also run the hypothesis test with larger significance levels (0.10 and 0.25) to see if a wider range changes the outcome. It is crucial to remember that at higher significance levels, the risk of making a Type I error increases and should be done cautiously. As an example, this strategy is like the study conducted by East West Institute of Technology in India. They described using logistic regression to predict cardiovascular disease, and the steps they would take to train the model and test it (G et al., 2022).

Unit II: Artificial Intelligence

Strategy #1: Classification using machine learning.

We can use different machine learning algorithms to classify patients into those with heart disease and those without. First, we will split the data into training and testing sets and then train the different algorithms to determine the best-performing algorithm. The algorithms we use in the implementation are LDA, Logistic Regression, Linear Regression, Ridge Regression, Decision Tree, Random Forest, and AdaBoost. The results from implementing the algorithm tell us that the best-performing algorithm was Linear Regression, with an accuracy of 0.5708.

Unit III: Data Mining

Strategy #2: Web Scraping using Scrapy.

We can use the Scrapy Python library to scrape data from PubMed and retrieve relevant research articles on heart disease. We can then analyze the abstracts and full texts of the articles to gain insights into the latest research findings on heart disease and potential risk factors. A peer reviewed article found that web scraping, even though an old technique, is still one of the best tools to use when extracting data for a wide range of applications (Glez-Peña et al., 2014).

This hypothetical study combines statistical inference, machine learning, and web scraping to gain insights into predicting heart disease. Using the Cleveland Heart Disease dataset, we can assess the significance of different features using hypothesis testing, classify patients into those with and without heart disease, and use web scraping to retrieve and analyze the latest research on heart disease.

G, A., Ganesh, B., Ganesh, A., Srinivas, C., Dhanraj, & Mensinkal, K. (2022). Logistic regression technique for prediction of cardiovascular disease. *Global Transitions Proceedings*, *3*(1), 127–130. https://doi.org/10.1016/j.gltp.2022.04.008

Glez-Peña, D., Lourenço, A., López-Fernández, H., Reboiro-Jato, M., & Fdez-Riverola, F. (2014). Web scraping technologies in an API world. *Briefings in Bioinformatics*, *15*(5), 788–797. https://doi.org/10.1093/bib/bbt026