Name: Faiyam Islam COMP5310 – Principles of Data Science

Project Proposal – Stage 0

Research Question

How can we predict and classify cardiovascular diseases using machine learning techniques?

Motivation for dataset

According to the 2017 Global Burden Disease research, cardiovascular disease is responsible for 43% fatalities (Bhatt, 2023). A plethora of these diseases refer to conditions such as blocked or narrowed blood vessels, resulting in stroke, chest pain, angina and heart attack. Machine learning techniques are a crucial part of the principles in Data Science. In the medical field there are various areas such as detecting and classifying cardiovascular diseases can help cardiologists provide the appropriate treatment to patients. What is more beneficial than merely detecting the presence or absence of heart diseases is to classify these diseases and understand the relationship between patients from this dataset on a molecular level, including their cholesterol and blood sugar levels. By performing this analysis, misdiagnosis of these diseases can be reduced through these vastly accurate methods using several classification and clustering methods in machine learning. My personal choice for a data science project in heart disease classification stems from previous experience in university projects and my dedication to keep myself fit and maintain my heart health.

Dataset

Link- https://www.kaggle.com/datasets/cherngs/heart-disease-cleveland-uci?datasetId=576697&sortBy=voteCount
This dataset was compiled in the Cleveland database from the following creators:

- Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.
- University Hospital, Zurich, Switzerland: William Steinbr
- University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.
- V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.
- Donor: David W. Aha (aha '@' ics.uci.edu) (714) 856-8779

The dataset contains 13 variables including age, sex and various symptoms of the patient and attributes that could affect their heart condition such as blood sugar levels. It is important to note that there is a target variable, called condition. Prior to our analysis, exploratory data analysis will be conducted to understand the data followed by a classification and potentially clustering methods.

Reference

Bhatt, C. M., Patel, P., Ghetia, T., & Mazzeo, P. L. (2023). Effective Heart Disease Prediction Using Machine Learning Techniques. *Algorithms*, *16*(2), 88. https://doi.org/10.3390/a16020088

Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J., Sandhu, S., Guppy, K., Lee, S., & Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. American Journal of Cardiology, 64,304--310.

David W. Aha & Dennis Kibler. "Instance-based prediction of heart-disease presence with the Cleveland database."

Gennari, J.H., Langley, P, & Fisher, D. (1989). Models of incremental concept formation. Artificial Intelligence, 40, 11--61.