Group Members: Ishayu Das, Daniel Jeun Valenzuela

Introduction

Cardiovascular diseases, particularly heart disease, are among the leading causes of death worldwide. According to the World Health Organization, millions of people die from heart-related ailments every year, making early detection and prevention crucial. Factors such as high cholesterol, hypertension, smoking, obesity, and diabetes contribute significantly to the risk of heart disease. Predictive models and machine learning techniques offer valuable tools in identifying at-risk individuals and enabling early interventions, potentially reducing mortality rates and healthcare burdens.

Advancements in artificial intelligence and machine learning have enhanced healthcare professionals' ability to assess heart disease risks using large-scale data analysis. These models can recognize hidden patterns in medical data, improving diagnostic accuracy and aiding in preventive healthcare.

The dataset used for this project, sourced from Kaggle (UCI Heart Disease Data), provides a detailed collection of medical and lifestyle attributes linked to heart disease. It includes patient-specific details such as age, gender, cholesterol levels, blood pressure, resting electrocardiogram (ECG) results, and key indicators like smoking habits, fasting blood sugar, maximum heart rate, and exercise-induced angina. Additionally, the dataset contains information on chest pain type and the presence of major vessels detected by fluoroscopy, making it a comprehensive resource for predictive analysis.

By analyzing this dataset, we aim to develop a predictive model that can accurately classify patients at risk of heart disease. This will provide valuable insights to healthcare professionals and contribute to developing preventive strategies. Additionally, machine learning techniques can explore which features contribute most significantly to heart disease prediction, offering new avenues for medical research and early intervention policies.

Dataset: https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data

Related Works

Different studies have explored different machine-learning approaches to improve heart disease prediction by utilizing this dataset. These are some key contributors:

Heart Disease Prediction Binary Classification, Muhammad Abdullah, 01/04/2024
This study focuses on heart disease prediction using this dataset. The main objectives of the study include Exploratory Data Analysis (EDA) to uncover insights, data preprocessing to

handle missing values and scale features, and model training using Random Forest and XGBoost classification. The aim is to enhance prediction accuracy and derive important insights that help to heart disease diagnosis. The research highlights data quality, feature engineering, and model optimization to improve predictive performance and assist in medical decision-making.

International application of a new probability algorithm for the diagnosis of coronary artery disease, Detrano et al. (1989)- Research Paper

This paper presents a probability-based algorithm for diagnosing coronary artery disease (CAD) and evaluates its effectiveness across different international medical institutions. The study assesses the predictive performance of the model using clinical and angiographic data from various patient populations, demonstrating its potential for improving diagnostic accuracy.

Instance-based prediction of heart-disease presence with Cleveland database, Aha & Kliber

This study explores instance-based learning methods to predict the presence of heart disease using the Cleveland heart disease dataset. The authors analyze the effectiveness of these methods compared to traditional statistical approaches, highlighting the advantages of instance-based models in handling complex medical data.

Data Preprocessing

The dataset has several data quality issues that must be addressed before the analysis. A key concern is missing values, which affect different columns, including *trestbps* (resting blood pressure), *chol* (cholesterol levels), *thalch* (maximum heart rate), and *oldpeak* (ST depression). Some categorical variables, such as *slope* (slope of peak exercise ST segment), *ca* (number of major vessels, colored by fluoroscopy), and *thal* (thalassemia test result), have an important amount of missing values, requiring careful deletion strategies. Furthermore, inconsistencies in data types are evident, as categorical variables such as *sex*, *fbs* (fasting blood sugar), and *exang* (exercise-induced angina) are stored as object types instead of numerical representations. In addition, continuous variables such as cholesterol and heart rate must be examined for potential outliers that could skew the results. Addressing these data quality issues is important to ensure the reliability of predictive modeling efforts.

To prepare the dataset, different preprocessing steps are required. First, missing values in numerical columns like *trestbps*, *chol*, and *thalch* will be deleted using the median, as these values may not follow a normal distribution. For categorical variables such as *fbs*, *restecg*, and *exang*, missing values can be replaced with the most frequently occurring category to maintain consistency. However, columns like *ca* and *thal*, which contain a large percentage of missing values, require a more detailed examination before deciding on a deletion strategy. Next, normalization is required for continuous variables to ensure that all features contribute equally to machine learning models. Methods like Min-Max Scaling or Z-score Standardization will be

applied to *age*, *trestbps*, *chol*, *thalch*, and *oldpeak* to bring their values into a comparable range.

Furthermore, categorical variables such as *sex*, *fbs*, *exang*, *slope*, and *thal* must be encoded for machine learning algorithms to process them effectively. Label encoding or one-hot encoding will be applied depending on the nature of the variable. For example, binary categories such as *sex* and *exang* can be label-encoded, while multi-class variables like *thal* require one-hot encoding to prevent ordinal confusion. In addition, outlier detection is necessary to prevent extreme values from modifying model predictions. Techniques such as the interquartile range method or Z-score analysis will be used to identify outliers in variables like cholesterol and heart rate, and adjustments will be made by either capping extreme values or removing outliers. By executing these preprocessing steps, the dataset will be transformed into a cleaner and more structured format, enhancing the accuracy of predictive models for heart disease diagnosis.