Cardiovascular Disease Dataset: The Scope

Our group project aims to apply Data Science and Machine Learning techniques to an open cardiovascular disease dataset sourced from Kaggle:

<https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset>

This project will involve:

* Data exploration, cleaning and preprocessing
* Feature engineering to enhance predictive power
* Development and evaluation off classification models (Logistic Regression, Random Forest, and other advanced methods)
* Statistical analysis and visualization to highlight relationships in the data
* A discussion of ethical, social, and professional considerations in using medical datasets and deploying predictive models in healthcare contexts

The final output will be a 3000-word IEEE-formatted research paper co-authored by the project team. Additionally, each group member will present their individual contributions via a recorded technical presentation.

Key Objectives:

* Identify and understand key features contributing to cardiovascular disease
* Train machine learning models to predict disease presence
* Evaluate model performance using appropriate metrics and visualizations
* Reflect on limitations, ethics, and implications of deploying AI in healthcare

Introduction

Cardiovascular Disease (CVD) is one of the leading causes of death worldwide, contributing to significant health and economic burdens. The ability to accurately predict individuals at risk of developing cardiovascular disease can enable earlier interventions and improve health outcomes with the rise of data availability and machine learning techniques, predictive models can now be developed using large datasets containing medical, demographic, and lifestyle factors.

This study presents a data-driven approach to predicting cardiovascular disease using an open-source dataset sourced from Kaggle. The dataset includes over 70,000 records with variables such as age, blood pressure, cholesterol levels, smoking habits, and physical activity.

The aim of this project is to build and evaluate several machine learning models to classify individuals based on their risk of cardiovascular disease. In doing so, we will follow the CRISP-DM data science lifecycle, which will include data cleaning, exploratory data analysis, feature engineering, model development, and evaluation.

Additionally, this study will consider ethical, legal, and professional issues surrounding the use of personal health data and predictive analytics in healthcare. We will explore issues of bias, fairness, and the societal impact of deploying machine learning models for medical decision-making.

This paper is structured as follows: Section II discusses related works in cardiovascular disease prediction: Section III outlines the dataset, data processing methods, and modeling techniques; Section IV presents the results and evaluation; Section V discusses findings, limitations, and ethical considerations; and Section VI concludes the study and proposes future work.