CIS 635 Project Proposal - Predicting Cardiovascular Disease

Team members: Alyssa Adamczak, Leah Bishop, Laxmi Sowjanya Doddi, & Jonathan Kivuva

Dataset: Cardiovascular Disease dataset - kaggle.com

Overview of Project

Cardiovascular disease (CVD) is considered the leading cause of death globally. CVD refers to diseases involving the heart and blood vessels, including stroke, heart failure, hypertensive heart disease, rheumatic heart disease, and peripheral arterial disease among others (Adhikary et al., 2022; Amini et al., 2021). Because of the high prevalence and mortality rates of CVD worldwide, this study will aim to uncover risk factors that predict the presence of disease to aid in the development of clinical screening tools.

This project will analyze the *Cardiovascular Disease dataset* published by kaggle.com. The dataset contains data from 70,000 patient records, eleven features, and a target feature. Specifically, this project aims to build a model that predicts the presence or absence of CVD by assessing which features are significant risk factors for CVD. After an initial exploratory data analysis, supervised learning techniques such as logistic regression and decision trees will be implemented. Afterward, performance metrics will be utilized to evaluate the model's performance on unseen data.

Data Plan

This project will utilize the *Cardiovascular Disease dataset* from Kaggle, which includes one target feature (CVD presence) and eleven categorical and numerical features. Prior to analysis, standard data mining procedures will be applied: handling missing values, removing duplicates, and detecting outliers using the 1.5x IQR method. Numerical data will be scaled using Z-score for normally distributed features and min-max scaling for non-normal data. This will prepare the dataset for building a model to predict the presence of CVD and assess key risk factors.

Related Work

The field of machine learning (ML) has already demonstrated its positive impact on many areas of healthcare including radiology, genetics, electronic health records, and neuroimaging (Habehh & Gohel, 2021). Because of its capability to uncover concealed patterns, ML holds great potential to aid in clinical decision-making and improve evidence-based practices. Early detection of diseases like CVD through ML can also reduce "the need for extensive and expensive clinical and laboratory investigations, resulting in a reduction of the financial burden on both the healthcare system and individuals" (Baghdadi et al., 2023).

A quick Google search reveals the vast amount of research already conducted on ML and CVD. One such study, "Advanced machine learning techniques for cardiovascular disease early detection and diagnosis" by Baghdadi et al., proposes a classification scheme for predicting CVD through data analysis, preprocessing, feature selection, model training, and model validation.

While this project shares several key aspects with Baghdadi et al.'s study, there are also significant differences. This project will detect outliers using the 1.5x IQR method and remove them, whereas Baghdadi et al. did not explicitly address how outliers were handled. Additionally, this project plans to apply Z-score and min-max scaling for the appropriate data, while Baghdadi et al.lacked clear insight into how the distribution of data was addressed prior to model training.

Perhaps the most notable difference is the simplicity and straightforwardness of this project's approach. Baghdadi et al. employed advanced models and techniques such as XGBoost, AdaBoost, and hyperparameter tuning with the Optuna library (Baghdadi et al., 2023). While their study may be more powerful due to its complexity and advanced techniques, this could pose challenges for replication and interpretability. Due to limitations in knowledge and expertise, time, and computational resources, this project will instead focus on the fundamental concepts of ML to analyze a dataset. Despite these constraints, this project still has the potential to uncover valuable insights into CVD diagnosis and contribute to the application of ML in healthcare.

Implementation Plan

Data collection & preprocessing: The *Cardiovascular Disease dataset* will be uploaded into Google Colab for preprocessing. The dataset will be analyzed to remove outliers, missing values, and duplicate records.

Exploratory data analysis: Descriptive statistics will be used to analyze and normalize the dataset utilizing Python libraries like Pandas and NumPy. Data will be visualized using techniques such as histograms, box plots, and scatter plots from Matplotlib.

Modeling: Logistic regression and decision trees will be used to build a model that predicts the presence or absence of CVD along with assessing the most significant risk factors for predicting CVD. Scikit-learn will be used for these modeling tasks.

Evaluation Plan

The dataset will be split into a training dataset and a test dataset. Unseen data will be input into the model to determine the presence or absence of CVD. A confusion matrix will be constructed to compare the predicted values with the actual values. Metrics including accuracy, precision, sensitivity, and specificity will be measured to evaluate the model's performance.

Plan for Group Collaboration

Communication: Slack will be used to communicate and coordinate asynchronously throughout the project.

Meetings: In-person meetings will be held before or after class and virtual meetings will be held via Zoom if determined necessary.

Data handling: Google Colab and GitHub will be used to manage code and data.

Timeline

Week	Task	Due Date
Week 1	Begin project proposal	
Week 2	Finalize and submit project proposal	10/14/24
Weeks 3-4	Begin developing pipeline	
Weeks 5-6	Write and submit project report	11/08/24
Weeks 7-8	Finalize pipeline	
Weeks 9-10	Write and submit final report	12/04/24

References

Adhikary D, Barman S, Ranjan R, Stone H. A Systematic Review of Major Cardiovascular Risk Factors: A Growing Global Health Concern. Cureus. 2022 Oct 10;14(10):e30119. Doi: 10.7759/cureus.30119. PMID: 36381818; PMCID: PMC9644238.

Amini, M., Zayeri, F. & Salehi, M. Trend analysis of cardiovascular disease mortality, incidence, and mortality-to-incidence ratio: results from global burden of disease study 2017. *BMC Public Health* 21, 401 (2021). https://doi.org/10.1186/s12889-021-10429-0

Baghdadi, N. A., Abdelaliem, S. M. F., Malki, A., Gad, I., Ewis, A., & Atlam, E. (2023). Advanced machine learning techniques for cardiovascular disease early detection and diagnosis. *Journal of Big Data*, *10*(1). https://doi.org/10.1186/s40537-023-00817-1

Cardiovascular Disease dataset. (2019, January 20). Kaggle. https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset/data

Habehh, H., & Gohel, S. (2021). Machine Learning in Healthcare. *Current genomics*, *22*(4), 291–300. https://doi.org/10.2174/1389202922666210705124359