CIS 635 Final Report: Predicting Cardiovascular Disease

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

Introduction

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). The motivation behind this project was to build a model that predicts the presence or absence of CVD and to assess which features are significant risk factors for CVD.

Using the *Cardiovascular Disease dataset* published by Kaggle.com, this project was divided into four sections: data collection and preprocessing, exploratory data analysis, modeling, and modeling evaluation. Logistic regression, decision tree, and random forest models were constructed and tested on the data via supervised learning methods. The models were evaluated against one another based on their performance metrics, with the logistic regression model performing the overall best. While slight differences were observed between the models regarding feature importance, systolic blood pressure, age, and cholesterol levels were consistently identified as the top three most significant predictors of CVD.

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. (Elhaddad & Hamam, 2024). 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 significant differences. This project detects and removes outliers using the 1.5x IQR method, whereas Baghdadi et al. did not explicitly address how outliers were handled. Additionally, while this project anticipated needing to apply Z-score and min-max scaling for the appropriate data, Baghdadi et al. lacked clear insight into how the distribution of data was addressed before 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 instead focused on the fundamental concepts of ML to analyze a dataset. Despite these constraints, this project was still able to uncover valuable insights into the detection of CVD and significant risk factors.

Methods

Data Collection & Preprocessing

For this project, we utilized the *Cardiovascular Disease dataset* published by Kaggle.com. This dataset includes over 70,000 records of patient data, eleven features (age, height, weight, gender, systolic blood pressure, diastolic blood pressure, cholesterol, glucose, smoking, alcohol intake, and physical activity), and one target variable (the presence or absence of CVD). While an initial investigation of the data confirmed no missing or duplicate records, descriptive statistics for the 'age' feature gave seemingly unrealistic results (ex. mean value of 19468.87). However, after referencing the data source, we realized that 'age' was given in units of days. To improve the interpretability of our analysis, we decided to convert all values of 'age' from units of days to years using simple transformation methods.

The most significant challenge we faced during data preprocessing was the inclusion of unrealistic or biologically impossible values for some numerical features. For example, a brief survey of minimum values revealed a minimum height of 55 cm, weight of 10 kg, systolic blood pressure of -150 mmHg, and diastolic blood pressure of -70 mmHg. Conversely, the maximum values of these same features revealed a height of 250 cm, weight of 200 kg, systolic blood pressure of 16,020 mmHg, and diastolic blood pressure of 11,000 mmHg. To build more accurate models, the dataset was filtered to include records that made biological sense for adult females and males. Height was filtered to contain values between 149 cm and 171 cm (*What Is the Average Height for Women?*, 2024, *The Average Height for Men, and What It Means for Men's Health*, 2024), weight for values between 38 kg and 125 kg (*Ideal Weight Chart | Staying Well*, n.d.), systolic blood pressures between 90 mmHg and 180 mmHg (*Understanding Blood Pressure Readings*, 2024), and diastolic blood pressures between 60 mmHg and 120 mmHg (*Low Blood Pressure*, 2024). After this filtering, any remaining outliers were identified and removed via the 1.5x IQR method.

Exploratory Data Analysis

To begin our exploratory data analysis, we first defined the numerical variables: age, height, weight, systolic blood pressure, and diastolic blood pressure. We then calculated descriptive statistics for each variable, including the mean, median, maximum, minimum, and standard deviation. Next, we defined the categorical variables: gender, cholesterol, glucose, smoking, alcohol intake, physical activity, and the presence or absence of CVD. We calculated counts for each category and included a key for each variable. Finally, we conducted a visual inspection of the data by constructing a histogram and Q-Q plot for each of the numeric variables (Fig. 1).

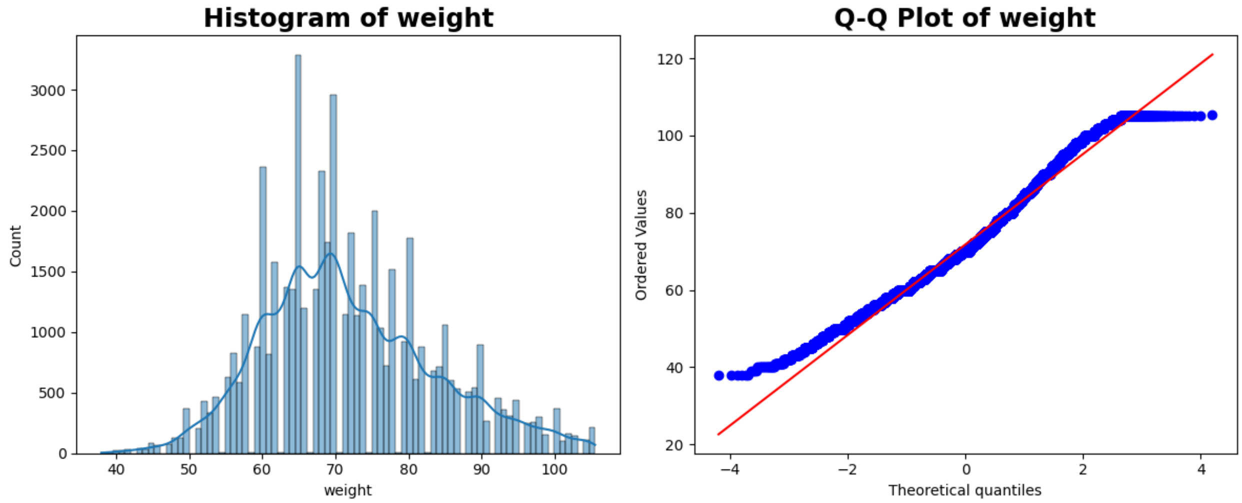

Figure 1: Example histogram and Q-Q plot for ‘weight’ feature

Modeling & Evaluation

In this section, we built three supervised learning models—logistic regression, decision tree, and random forest—to predict whether new samples would be classified as CVD present or absent. We first constructed the logistic regression model using the scikit-learn library in Python. We split our data using 70% as the training set and 30% as the test set. We used all eleven predictor variables to train the model. Hyperparameter tuning was performed using GridSearchCV to find the best combination of hyperparameters to optimize the model’s performance. This included implementing a 5-fold cross-validation to evaluate performance for each combination of hyperparameters. We then selected the best model after hyperparameter tuning for evaluation. After predicting probabilities and classes, the predicted outcome was displayed alongside the actual outcome for a sample of the data. We further visualized the model’s performance by constructing a confusion matrix. Accuracy, precision, recall, and F1 score metrics were calculated to measure the model’s performance, and the ROC-AUC curve was plotted to visualize the model’s performance. Finally, we calculated the coefficients of feature performance to determine the most influential predictors of CVD. A decision tree and random forest model were constructed using the same process for model training, hyperparameter tuning, cross-validation, selecting the best model, calculating performance metrics, and visualizing results. Instead of a 5-fold cross validation, 3-fold cross validation was used for the random forest due to computational limitations.

Modeling Comparison

In the last section of our pipeline, we compared the performance of the logistic regression, decision tree, and random forest models against one another. We created a new data frame containing the accuracy, precision, recall, F1, and AUC-ROC scores, which were displayed in a table. These same metrics were then plotted onto a heatmap for visualization (Fig. 2). Finally, we constructed an AUC-ROC curve overlay plot to compare the curves against one another.

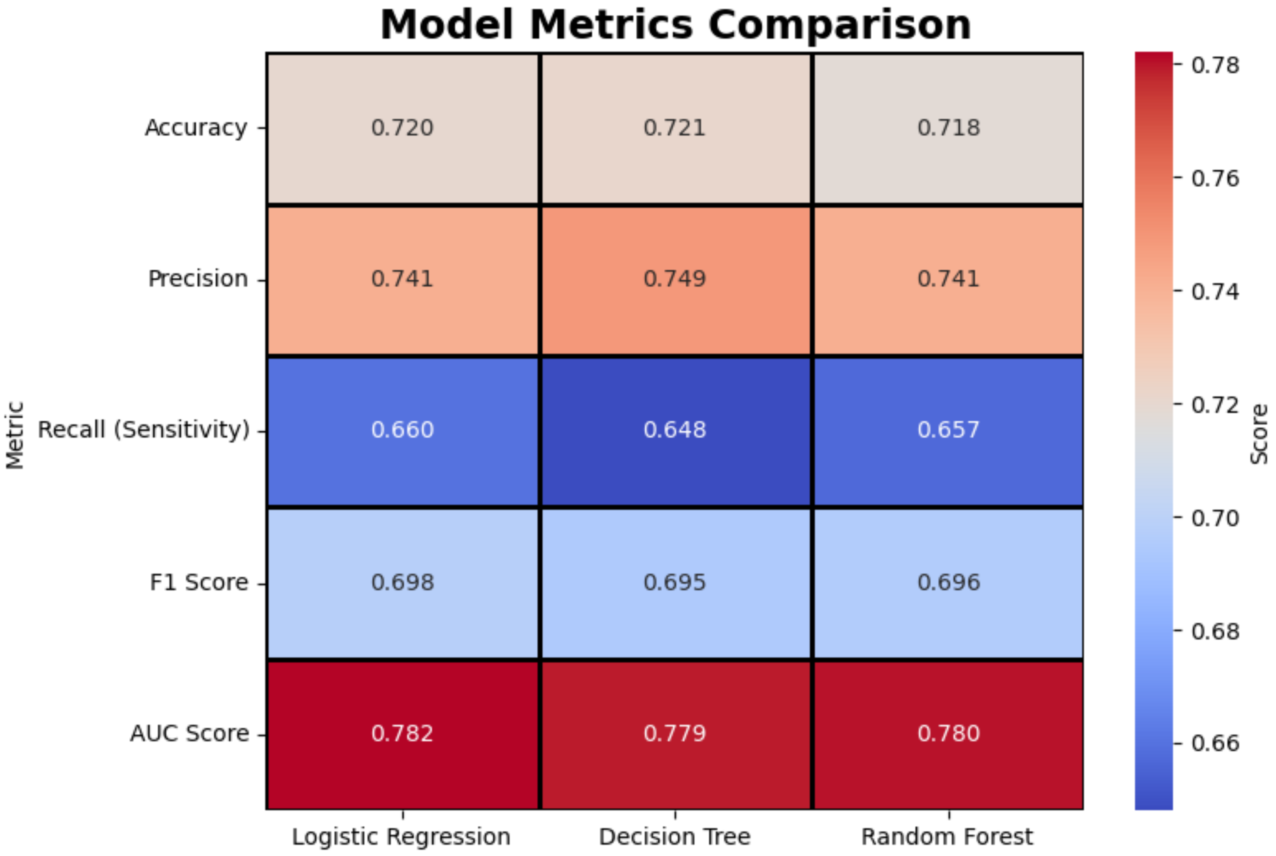

Figure 2: Heatmap of Model Performance Metrics

Experiments, Results, and Discussion

The goal of this project was to use supervised learning techniques to predict the presence or absence of CVD along with determining the most significant risk factors of the disease. Utilizing the *Cardiovascular Disease dataset* from kaggle.com, we built logistic regression, decision tree, and random forest models in Google Colab using the sci-kit library. After model training and hyperparameter tuning, we calculated evaluation metrics and visualized them using Matplotlib. We used these metrics to determine how accurate each model was in predicting the presence or absence of CVD. Additionally, we used the feature importance coefficients of each feature to determine the influence of each feature on the models.

The implementation of hyperparameter tuning using GridSearchCV revealed the best parameters for the training of each model. The best parameters for the logistic regression model were a ‘C’ of 0.1, ‘penalty’ of ‘l2’, and 'solver’ of liblinear. For the decision tree model, the best parameters were found to be ‘criterion’ of ‘entropy', 'max_depth’ of 7, 'max_features’ of “None”, 'min_samples_leaf’ of 4, and 'min_samples_split’ of 2. Finally, the best parameters for the random forest model were 10 for 'max_depth', 'sqrt’ for' max_features’, 1 for 'min_samples_leaf', 2 for 'min_samples_split’, and 100 for ‘n_estimators’.

In terms of predicting the presence or absence of CVD correctly, we found that each of the models performed very similarly. The F1 scores of logistic regression, decision tree, and random forest models were 0.698, 0.695, and 0.696, respectively. The AUC-ROC scores, which specifically measure the model’s ability to predict classes, were 0.782 for the logistic regression model, 0.779 for the decision tree model, and 0.780 for the random forest model. Based on the AUC-ROC scores, logistic regression was the best model for predicting the presence or absence of CVD, followed by the random forest model and then the decision tree model. The F1 scores confirm these findings. It is worth noting that the evaluation metrics of all three models are nearly identical, as the AUC-ROC scores are the same when using two significant figures instead of three. This can be visualized using the overlay plot of AUC-ROC scores (Fig. 3).

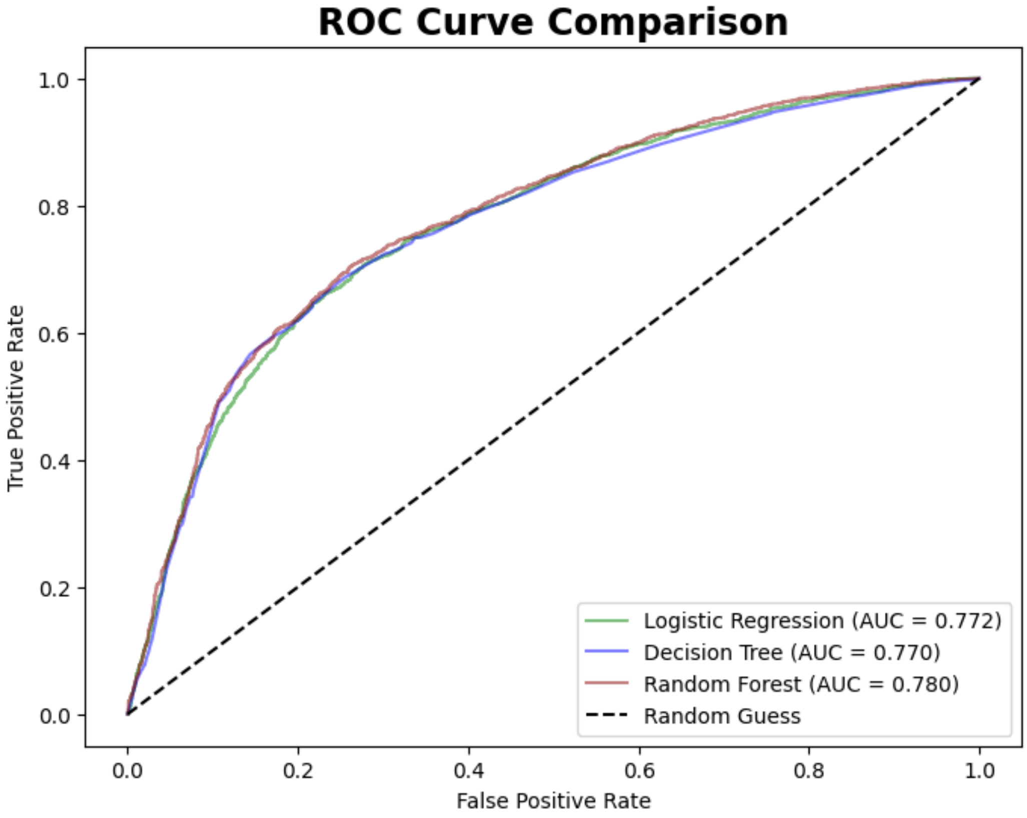

Figure 3: AUC-ROC Curve Comparison for all models

In our analysis, the logistic regression and decision tree models identified systolic blood pressure, age, and cholesterol level as the top three most significant predictors of CVD, although not in the same order. Alternatively, the random forest model determined systolic blood pressure, age, and diastolic blood pressure as its top three features, with cholesterol level being the fourth. The feature importance coefficients varied greatly between the models. For example, the logistic regression revealed a coefficient of 0.90 for systolic blood pressure, while the decision tree showed a coefficient of 0.69 and the random forest a coefficient of 0.08. All of the models calculated the absolute coefficients of gender, height, smoking, and alcohol intake to be less than 0.05. The logistic regression model consistently gave higher coefficients than the other models.

Conclusion

This project aimed to predict the presence or absence of CVD using machine learning models and to identify the most significant risk factors for developing the disease. Utilizing the *Cardiovascular Disease dataset* from Kaggle.com, we went through data cleaning and preprocessing, exploratory data analysis, modeling, and model evaluation. Our logistic regression, decision tree, and random forest models were exposed to training and test set data and evaluated on their accuracy, precision, recall, F1, and AUC-ROC scores. While all three models performed similarly, the logistic regression model yielded the best performance results in predicting the presence or absence of CVD, followed closely behind by the random forest and then decision tree models. Although slight differences were observed in feature performance rankings between the models, they consistently identified systolic blood pressure, age, and cholesterol levels as the most significant risk factors for CVD.

Despite these results, there were several limitations to our analysis. First, our dataset required extensive cleaning and preprocessing to remove a large portion of records that contained unrealistic or biologically impossible values for numerical features. After completing this process, we were able to retain 51,187 patient records from the original dataset. While this was still a substantial sample size to ensure robust model training and testing, we recognize that this could have led to a potential loss of information. However, by basing our data preprocessing decisions on biological plausibility and real-world context, we believed we improved the reliability and relevance of our analysis, highlighting the importance of domain knowledge when building machine learning models.

We also faced limitations regarding computational resources and model complexity. Due to the lack of in-depth knowledge, hyperparameter tuning of the random forest model involved a great deal of trial and error to identify a combination of parameters that would allow the model to run in a reasonable amount of time. Several hyperparameter searches took over seven minutes before being manually stopped. Given these constraints, the refinement of these parameters was limited, impacting the model's efficiency and possibly the true ability to predict the presence or absence of CVD.

Future work in this area would likely benefit from the use of more advanced models or more powerful computational resources to better predict the presence or absence of CVD and its greatest risk factors. By leveraging these resources along with advanced optimization tools, the full range of hyperparameter tuning could be explored to discover better-performing models. Additionally, incorporating more features like clinical data and medical history might provide deeper insights into the risk factors for developing CVD.

Data and Software Availability

Dataset: Cardiovascular Disease dataset - kaggle.com

GitHub repository: Team 2 GitHub

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

Elhaddad, M., & Hamam, S. (2024). AI-Driven Clinical Decision Support Systems: An Ongoing Pursuit of Potential. *Cureus*, *16*(4), e57728. https://doi.org/10.7759/cureus.57728

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

Ideal Weight Chart | Staying well. (n.d.). https://www.bannerhealth.com/staying-well/health-and-wellness/fitness-nutrition/ideal-weight

Low blood pressure. (2024, October 4). Cleveland Clinic. https://my.clevelandclinic.org/health/diseases/21156-low-blood-pressure-hypotension

The average height for men, and what it means for men's health. (2024, October 17). Cleveland Clinic. https://health.clevelandclinic.org/what-is-the-average-height-for-a-man

Understanding Blood Pressure Readings. (2024, May 17). Heart.org. https://www.heart.org/en/health-topics/high-blood-pressure/understanding-blood-pressure-readings

What is the average height for women? (2024, October 16). Cleveland Clinic. https://health.clevelandclinic.org/what-is-the-average-height-for-women