Milestone 5 Final

Brian Reppeto

Bellevue University

DSC 630 T303 2251

Frank Neigebauer

November 7, 2024

Milestone 5 Final

**Introduction**

**Problem Statement:** Cardiovascular disease, particularly coronary heart disease (CHD), remains one of the leading causes of death worldwide, accounting for millions of fatalities each year. The burden of CHD not only affects individuals but also places an immense strain on healthcare systems due to the costs associated with treatment and long-term care. This project aims to address the critical issue of predicting CHD using the Framingham heart disease dataset (Ashish Bhardwaj, 2022). The goal is to develop a predictive model that can accurately identify individuals at high risk, enabling early intervention strategies that can mitigate the risk of CHD and improve patient outcomes.

**Importance of the Problem:** The ability to predict the likelihood of CHD is crucial for several reasons. Firstly, early detection allows for timely preventive measures, such as lifestyle modifications or pharmacological interventions, which can significantly reduce the risk of CHD. Secondly, accurate risk prediction models can help prioritize healthcare resources, ensuring that high-risk individuals receive the attention and care they need. Moreover, understanding the factors that contribute to CHD can help inform the public, and help create health strategies and policies aimed at reducing the incidence of heart disease at the population level.

**Target Audience:** The outcome of this project is relevant to a wide range of stakeholders. Healthcare providers, cardiologists and primary care physicians, can use the predictive model to identify patients who are at an elevated risk of CHD and tailor their treatment plans accordingly. Public health officials may also find the model useful for designing targeted interventions and educational campaigns that address the risk factors associated with CHD. Insurance companies could leverage the model to better assess the health risks of their clients, leading to more accurate premium calculations and risk management strategies.

**Data Source:** The Framingham heart disease dataset used in this project originates from the Framingham Heart Study, a study that began in 1948 in Framingham, Massachusetts. The study's primary objective was to identify the common factors or characteristics that contribute to cardiovascular disease (CVD) (Ashish Bhardwaj, 2022). Over the years, the study has collected extensive data on several generations of participants, making it one of the most comprehensive sources of data on cardiovascular health. The dataset used in this project includes over 4,000 records with 15 attributes, such as age, cholesterol levels, systolic and diastolic blood pressure, smoking status, diabetes status, and family history of heart disease. These variables are crucial for assessing an individual's risk of developing CHD, and their inclusion in the dataset makes it an invaluable resource for predictive modeling.

**Data Relevance:** The Framingham dataset provides a rich source of data on the risk factors associated with CHD, many of which are well documented within medical literature. The dataset's lengthy 10-year study also allows for the analysis of trends and changes in risk factors over time, which is essential for accurately predicting the risk of CHD. Furthermore, the dataset includes a diverse range of participants, which helps ensure that the predictive model developed in this project will be applicable to a broader population. Moreover, the use of this dataset allows for the creation of a model that is not only accurate but also grounded in real-world clinical data.

**Methods/Results**: Several insights were gained from the feature relationships. Some of the features, such as systolic and diastolic blood pressure, had moderate to high correlations. This indicates interconnected health indicators, which is expected in medical data but may introduce multicollinearity in models. Additionally, the dataset showed a significant imbalance in the target variable, with a larger proportion of individuals not experiencing coronary heart disease (CHD) within the ten-year window. This required special handling to ensure fair model training.

**Visualizations:** There were several visualizations used to help understand the data. A pie chart showing the distribution of CHD cases highlighted the class imbalance, prompting the use of SMOTE to balance the dataset. Moreover, aa heatmap visualizing the feature correlations offered insight into how variables like blood pressure and BMI interact, informing feature selection and model design.

**Data Cleaning**: Missing values were imputed using the mean values, and unnecessary columns were dropped. SelectKBest was used to identify the five most predictive features for efficiency and model interpretability. The data was then divided into training and testing sets in an 80-20 ratio. To address the class imbalance, SMOTE was applied to create a balanced training dataset, generating synthetic samples for the minority class.

**Model Selection:** Two types of machine learning models were used on the data.

Logistic Regression: Chosen for its simplicity and interpretability, and suitable for binary classification (Khan et al., 2024).

Random Forest Classifier: A more complex ensemble model known for its accuracy and robustness, especially effective in handling a mixture of numerical and categorical data (Khan et al., 2024).

**Measure the Results:** ROC-AUC Score was used to measure the model's ability to distinguish between positive and negative CHD cases. It’s particularly useful for imbalanced data. Precision, Recall, F1-Score: These metrics assess the model’s performance on each class, providing insight into its handling of both positive and negative outcomes. Accuracy offers a general measure of correct predictions but is less informative on imbalanced data without supporting metrics. ROC-AUC was selected because it provides a balanced measure of model performance across thresholds, making it ideal for imbalanced datasets. Precision and Recall gave insight into the model’s accuracy per class, highlighting its effectiveness in detecting CHD cases versus non-cases. F1-Score combines precision and recall, offering a balanced metric especially useful when both false positives and false negatives have consequences.

**Conclusion:** The analysis showed that Random Forest, combined with SMOTE achieved a high accuracy of 87% and a ROC-AUC score of 0.8706, indicating strong performance in identifying CHD cases. Logistic Regression, while less accurate, provided consistent results and could be beneficial where model interpretability is prioritized.The best performing model Random Forest is recommended due to its high predictive accuracy and effectiveness with SMOTE.Additional feature engineering and hyperparameter tuning could further enhance the accuracy, especially for the Logistic Regression model.For real-world deployment, SMOTE or other balancing techniques should be integrated to maintain model fairness.The Random Forest model with SMOTE is close to deployment readiness, demonstrating strong accuracy and balanced class handling. However, further validation on external data and calibration may be necessary to ensure robust performance across diverse populations.Testing on new datasets would confirm the model’s generalizability. Moreover, fine-tuning may be needed to optimize thresholds for high-stakes applications.Additionally, exploring new or combined features may help improve the model’s predictive accuracy.

**Ethical Concerns:** Developing a predictive model for coronary heart disease (CHD) involves several important ethical considerations, particularly around data privacy, model fairness, and real-world implications. Protecting data privacy is essential, as health information is sensitive and must comply with healthcare regulations (HIPPA) to prevent unauthorized access or misuse. Fairness and bias are also key concerns; if the model performs unevenly across different demographic groups, it could exacerbate existing healthcare disparities. Regular audits are crucial to ensure fair and accurate predictions for all populations. Additionally, \*transparency and interpretability are vital, especially for complex models like Random Forest. Utilizing explainable AI techniques, such as SHAP values, can help clarify how the model makes predictions, which fosters trust among healthcare providers (Basu et al., 2020). The ethical implications of deploying such a model include the potential for discrimination if certain groups are systematically misclassified, as well as psychological impacts on patients, such as anxiety from false positives or inadequate care from false negatives. To address these concerns, it’s important to conduct fairness audits, apply explainable AI for greater transparency, involve human oversight to ensure the model supports rather than dictates clinical decisions, and clearly communicate the model’s limitations. These steps collectively promote responsible, equitable, and effective use of the CHD prediction model in healthcare, ultimately enhancing patient care while upholding ethical standards.

References

Ashish Bhardwaj. (2022). Framingham heart study dataset [Data set]. Kaggle. <https://doi.org/10.34740/kaggle/dsv/3493583>

Basu, T., Menzer, O., & Engel-Wolf, S. (2020). The ethics of machine learning in medical sciences: Where do we stand today? Indian Journal of Dermatology, 65(5), 358. Retrieved November 10, 2024, from <https://doi.org/10.4103/ijd.ijd_419_20>

Khan, H., Bilal, A., Aslam, M., & Mustafa, H. (2024). Heart disease detection: A comprehensive analysis of machine learning, ensemble learning, and deep learning algorithms. Nano Biomedicine and Engineering. Retrieved November 10, 2024, from <https://doi.org/10.26599/nbe.2024.9290087>