Research Article

Predicting Heart Disease Risk: Analysing Health and Socio-Economic Indicators

Aisha Mohammad1

1 amohammad4@saintpeters.edu

**Abstract:** Heart disease remains one of the leading global health challenges, with its prevalence steadily increasing Studies highlight that individuals with diabetes are significantly more prone to developing cardiovascular conditions compared to those without diabetes. Alongside clinical indicators like physical activity and cholesterol levels, socioeconomic factors such as income, education, race, and gender also play a crucial role in influencing heart disease risk. This project aims to conduct an in-depth analysis and prediction of heart disease risk by examining both clinical health indicators and socioeconomic determinants. Techniques such as ML models (Logistic regression, Random Forest, SVM and gradient boosting), EDA and visualizations are used to conduct this research. Two datasets were utilized and combined to integrate diverse features for comprehensive analysis. Results: \_\_\_\_ Conclusions: \_\_\_\_\_

**Keywords:** Machine learning; heart disease prediction; Socio-economic indicators; cardiovascular disease

1. Introduction

Cardiovascular diseases (CVDs) are a leading cause of mortality worldwide, accounting for over 43% of all deaths globally, according to the Global Burden of Disease Study 2017.[[1](file:///Users/aishamohammad/Downloads/1.%09Bhatt,%20C.M.%3B%20Patel,%20P.%3B%20%20Ghetia,%20T.%20&%20Mazzeo,%20P.L.%20(2023).%20Effective%20%20Heart%20Disease%20Prediction%20Using%20%20Machine%20Learning%20Techniques.%20%20Algorithms,%2016%20https:/www.mdpi.com/1999-4893/16/2/88)] In wealthier nations like the United States, the prevalence of CVD is projected to grow by 10% from 2010 to 2030, with socioeconomic and racial disparities playing a significant role in this increase. The factors driving these changes have been studied, with causes such as an increase in unhealthy diets and reduced physical activity being identified as key contributors. Yuan Zhao and colleagues (2021), highlight that the impact is particularly severe in low- and middle-income countries, where 75% of CVD-related fatalities occur. [2] Social determinants of health (SDoH) which influence these changes is defined by WHO as the “conditions in which people are born, grow, live, work and age”.[2] This increasing global rise of CVD emphasizes an urgency to develop innovative solution, particularly predictive models that enable early detection and targeted interventions.

With patient data now being well-documented, there has been a surge in medical data available for early disease detection. Machine learning (ML) and data mining have become areas of significant interest for researchers, as the growing volume of data offers immense potential to develop improved models for predicting heart diseases and preventing it early on. As mentioned in a literature study by Deepa et al. (2024) early identification of individuals at risk enables focused preventive actions, promoting a shift from reactive to proactive healthcare.[3] Data driven techniques and ML are already being used to automate diabetes screening, detect complications in blood vessels, and help create personalized prevention and treatment plans.[4]

In recent years, studies show that machine learning models, such as supervised learning techniques like logistic regression, SVM, random forest, gradient boosting, and deep learning neural networks, have been widely applied. One study conducted an exploratory data analysis (EDA) using K-means clustering to identify patterns, and visualizations were also created to provide clearer insights into the data. These techniques help in improving the accuracy of predictions and understanding the underlying structures in the data.

2. Materials and Methods

The aim of this study is to predict heart disease based on diverse features for a better understanding of who are the most affected by heart diseases. By identifying high-risk individuals and vulnerable groups, healthcare organizations can implement proactive care strategies, optimize treatment plans, and improve accessibility to underserved communities, ultimately enhancing patient outcomes and reducing the financial burden on the healthcare system. Two datasets have been used for this study; both are available on free dataset repositories. This is done to integrate socio-economic factors as well, since the first dataset had only clinical data. These two datasets are processed for standard scaling, missing values, Resampling techniques are used, then they are combined together and analysis is done on the new dataset. Fig1. Shows the detailed High level architecture diagram of the model.

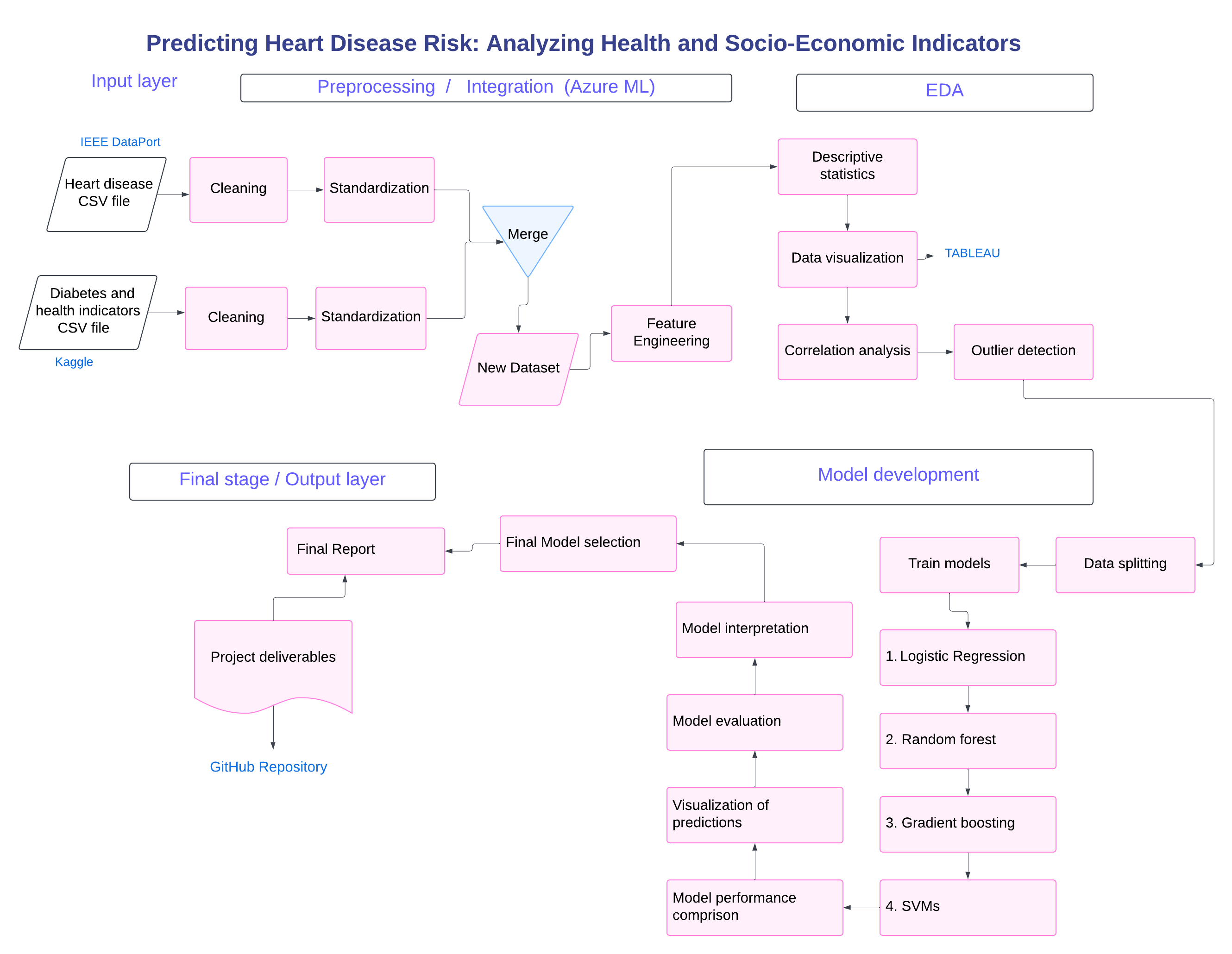


Figure 1. High level architecture diagram of the model.

* 1. Data source

The study utilizes datasets freely available on IEEE DataPort and Kaggle. First dataset, Heart Disease Comprehensive Dataset (referred to as ‘Heart’ in the code); retrieved from IEEE DataPort, this dataset was created by Manu Siddhartha by combining five popular heart disease datasets from the UCI ML repository. It integrates 11 common features, making it the largest heart disease dataset currently available for research, with a focus on cardiovascular conditions. Second dataset, Diabetes Binary Health Indicators Dataset (referred to as ‘lifestyle’ in the code); sourced from Kaggle, this dataset contains 253,680 survey responses from the 2015 Behavioral Risk Factor Surveillance System (BRFSS) conducted by the CDC. It was cleaned and reduced by Alex Teboul to 21 feature variables from the original dataset of 441,455 responses with 330 features. The dataset includes health-related risk behaviors, chronic conditions, and preventive service usage but is imbalanced, offering opportunities for data pre-processing.

Using Azure ML, these datasets are merged to provide a unified source for an in-depth analysis of the relationship between diabetes, socio-economic factors, and heart disease. This combination leverages the strengths of both datasets to support comprehensive analysis and actionable insights.

* 1. Data preprocessing

Once the datasets are loaded, they were studied individually. The numerical features were standardized using Standard Scaler. All features were assessed for null values. No missing values were identified in either dataset, eliminating the need for imputation. From feature engineering techniques, encoding on categorical variables was attempted but after some study, it was found that the categorical variables are ordinal already follow a structured order, eliminating the need for encoding. After this the preprocessed datasets, were assessed for their shape size. The first dataset, Heart Disease Comprehensive Dataset, had (1190, 12) rows and columns and the second dataset’s size (253680, 22). but overfitting of models was observed. So, using resampling technique, the second dataset was down sampled to match the first dataset. Indexes were reset for both datasets because after operations like up sampling, the index may become misaligned or non-sequential**.** Then, both were merged using the ‘pd.concat’. Data quality of the combined dataset was checked by checking for null values, duplicates and dropping one of the similar columns that were being repeated due to joining two datasets. Next, the two target variables, 'target' and 'HeartDiseaseorAttack' were combined, which are similar and drop 'HeartDiseaseorAttack'**.** Some feature engineering techniques were applied and ‘Target’ and ‘HeartDiseaseorAttack’ were merged into a new target variable**.** Then SMOTE was applied to balance the target variable. Lastly, 4 methods for feature selection were used to identify key features: Correlation Matrix- this removes highly correlated features to reduce multicollinearity, Point-Biserial Correlation**-** measures the strength and direction of the association between continuousfeature and a binary target variable**,** Random Forest Classifier- an easy way to measure feature importance and Recursive Feature Elimination (RFE)- using Logistic Regression as the estimator to select the top 20 features. From above 4 feature selection techniques, results from Random forest classifier alligns with the analysis goals. So features with threshold equal to and above 0.02 were selected. Selected features aligned with socio-economic factorsfrom Random Forest Classifier results, while irrelevant ones were dropped. The same feature selection was applied on test feature matrix as well.

* 1. Model development

Using the above selected features, we will be training 4 models using Function based coding. For model development use, X\_train\_resamp\_selected, X\_test\_selected, y\_train\_resampled, y\_test. The shape of the train- test columns are X\_train\_resamp\_selected (1453, 20) y\_train\_resampled (1453,) X\_test\_selected (238, 20) and lastly y\_test (238,). Using function- based coding 4 models were developed:

**Logistic Regression**:

A type of regression analysis used for prediction of outcome variables that are binary.

Efficient and requires less computational power. Can perform well with linearly separable data.

**Random Forest (RF):**

A decision-tree-based ensemble algorithm that combines multiple decision trees to make predictions. It handles large datasets and missing values well. Also is robust to overfitting and works well with a mix of continuous and categorical data.

**Gradient Boosting :**

An algorithm that involves sequentially training weak models (typically decision trees) where each new model corrects the errors made by the previous one. It can achieve high performance on both classification and regression tasks. Also, handles imbalanced datasets better with tuning.

**SVM :**

A powerful classifier that aims to find a hyperplane in an n-dimensional space that separates the data points of different classes. It is effective in high-dimensional spaces, especially useful in cases where the number of dimensions exceeds the number of samples.

After evaluating all models using Evaluation metrics: Accuracy, Precision, AUC/ROC, MAE, MSE, RMSE, MAPE, the best model was decided.

2.4 Model Saving, Deployment and feature importance

To keep the process simple and organized, all trained models were saved into a single pickle file (all\_models.pkl) for easy loading during interface development. After loading, a mapping was created to correctly match each model with its name, ensuring clear identification.

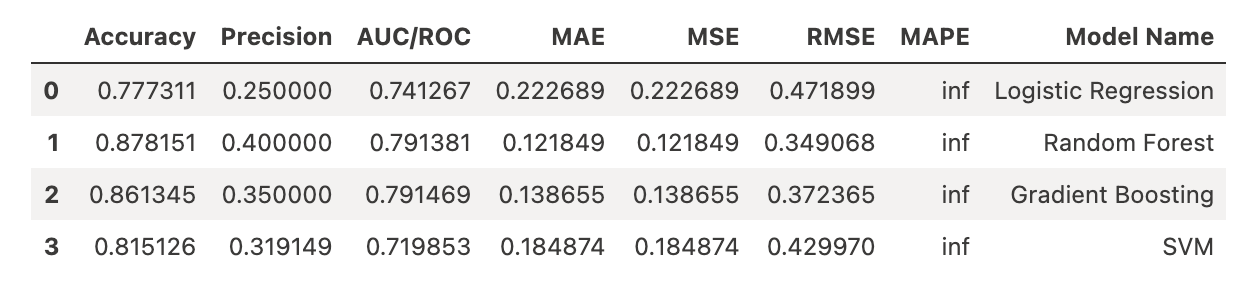
The models were then sorted based on accuracy scores to automatically select the best-performing one for final use. The best model was then saved into another pickle file (best\_heart\_model\_new.pkl). The StandardScaler, which was fitted only on the training data for selected numerical features such as age, blood pressure, cholesterol, max heart rate, BMI, and physical health, was saved as a pickle file (scaler\_.pkl) to be reused later in the Streamlit application for preprocessing user input.

Feature importance was calculated using the best-performing model to identify which variables contributed most to heart disease prediction. The plot clearly highlights the top features, allowing for better interpretation and understanding of the model’s decision-making process.

Using the saved pickle files, a streamlit app was deployed which provided users with a questionnaire about clinical and social aspects and this would generate a prediction of heart disease or no heart disease.

3. Results

The four models were trained on selected features, resampled training dataset and the results are as follows:



Best Model is the Random Forest with highest accuracy: 87.81% and best AUC/ROC: 79.13%. Gradient Boosting performs similarly (86.13% accuracy) but has a slightly lower AUC/ROC. Logistic Regression & SVM underperform, with lower accuracy and precision. Error Metrics (MAE, MSE, RMSE) are lowest for Random Forest, confirming its reliability. MAPE is infinite (∞) due to potential zero values in the denominator.

The feature importance done on the best model as seen in Figure 2, observed that:

**Top Factors:** High Cholesterol, High Blood Pressure, Smoking, and General Health are the strongest predictors.

**Moderate Impact:** Diabetes, BMI, Difficulty Walking, Age, and Resting Blood Pressure contribute significantly.

**Lower Impact:** **Income and Education** have relatively low importance, suggesting socioeconomic factors may not be direct indicators of heart disease risk.

**Least Important:** Sex, Alcohol Consumption, and Access to Healthcare show minimal influence.

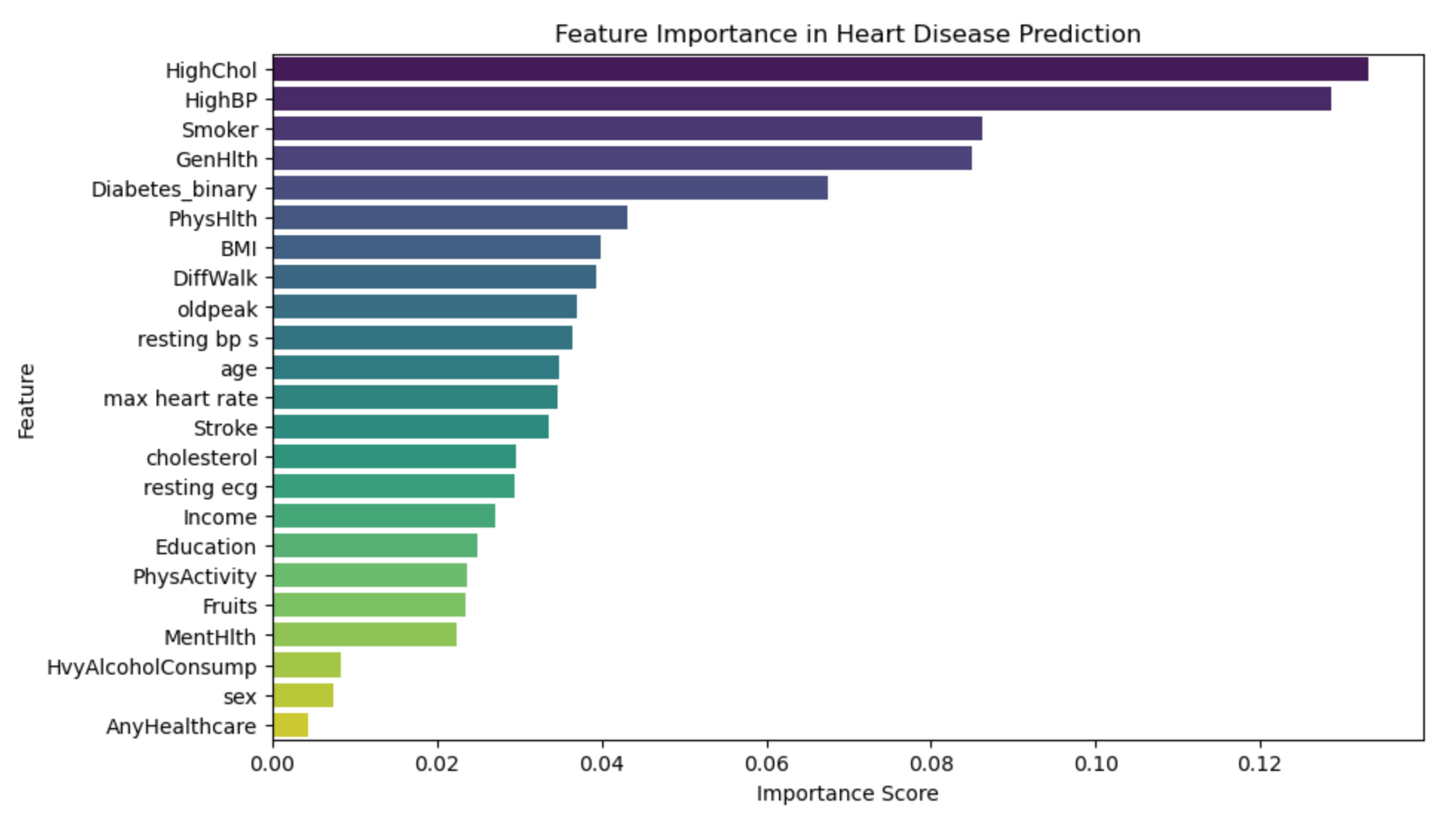


Figure 2. Feature Importance of the best model

A correlation matrix between socio economic factors, ‘education’ and ‘income’ and the top factors, revealed that Income and General Health have a negative correlation (-0.34), suggesting lower income is associated with poorer health. Education and General Health also show a negative correlation (-0.32). This suggests an indirect influence of socio-economic factors on heart disease prediction.

4. Challenges and limitations

4.1 Model Bias in Predictions

Despite extensive efforts, including SMOTE for balancing, multiple feature selection techniques, and different pre-processing methods, the model consistently predicts a higher bias for no heart disease (0). Also, despite being the best-performing model, the confusion matrix for random forest reveals Low recall for positive class (1), indicating bias as shown below in Figure 3.

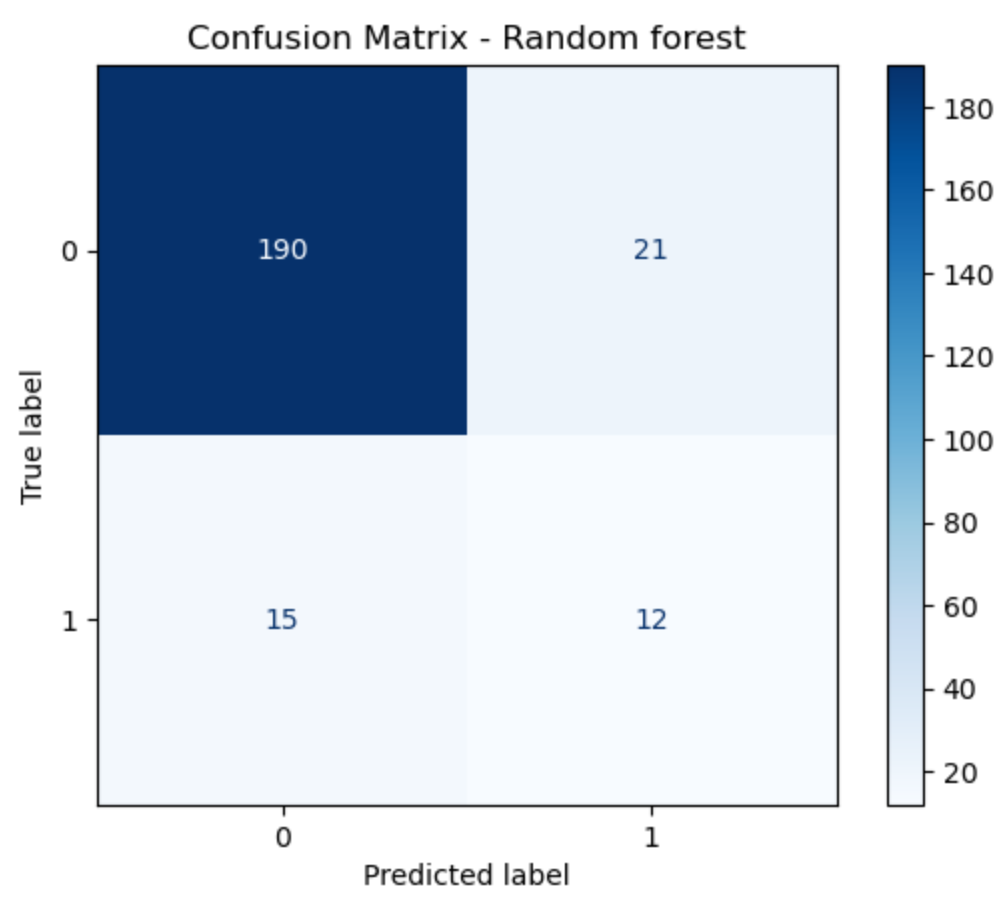


Figure 3. Confusion matrix for Random Forest [Best model].

4.2 Streamlit App Prediction Issue

The Streamlit app was successfully built, but despite multiple debugging attempts, it consistently predicts "Low risk of heart disease" (0). Even though the training data is balanced with equal representation of 0 and 1, the deployed model remains biased, likely due to issues with the decision boundary or the influence of certain features.

5. Future work

Future work will focus on improving dataset concatenation to ensure more seamless and effective data integration. Potential issues arising from the merging of target variables will be thoroughly investigated to prevent negative impacts on model performance. Advanced analytical techniques will be applied to uncover deeper relationships and emerging trends within the data. Additional feature selection methods will be explored to enhance model accuracy, and efforts will be made to address any remaining bias or imbalance in the model's predictions.

**References:**

1. Bhatt, C.M.; Patel, P.; Ghetia, T. & Mazzeo, P.L. (2023). *Effective Heart Disease Prediction Using Machine Learning Techniques*. Algorithms, 16 <https://www.mdpi.com/1999-4893/16/2/88>
2. Zhao, Y.; Wood, E.P.; Mirin, N.; Cook, S.H. &Chunara, R. (2021) *Social Determinants in Machine Learning Cardiovascular Disease Prediction Models: A Systematic Review*. American Journal of Preventive Medicine, 61(4).<https://www.sciencedirect.com/science/article/pii/S0749379721002749>
3. Deepa, Dr. R.; Sadu, V.B.; C, Prashant. G. & Sivasamy, Dr. A. (2024). *Early prediction of cardiovascular disease using machine learning: Unveiling risk factors from health records*. AIP Advances, 14(3) <https://doi.org/10.1063/5.0191990>
4. Oikonomou, E.K., & Khera, R. (2023). *Machine learning in precision diabetes care and cardiovascular risk prediction*. Cardiovascular Diabetology, 22(1). <https://doi.org/10.1186/s12933-023-01985-3>