BIOS 611 Final Project

Shuqi Zhang

December 2, 2024

1 Introduction

Global death counts due to cardiovascular disease (CVD) increased from 12.4 million in 1990 to 19.8 million in 2022 reflecting global population growth, aging and the contributions from preventable metabolic, environmental, and behavioral risks. Heart attack is one of the life-threatening coronary events with sudden cardiac death and the most severe clinical presentation of coronary artery disease. Data science and machine learning can be used for diagnosing and predicting heart attack to save lives, improves health outcomes, and allocates healthcare resources efficiently.

2 Problem and Database Description

The Kaggle dataset I selected is the "Heart Attack Risk Analysis Competition" which is ChatGPT generated. The competition aims to train a model to predict an individual is at high risk or low risk of a heart attack based on the input parameters including patient's socio-demographics, clinical characteristics, lifestyle, and comorbidities. The outcome of this study is the occurrence of heart attack (1: Yes, 0: No). The other features include age, sex, country, continent where the patient resides, hemisphere where the patient resides, income, Body Mass Index (BMI), cholesterol, blood pressure, heart rate, triglycerides, smoking, alcohol consumption, exercise hours/week, diet, sedentary hours/day, physical activity/week, sleep/day, obesity, previous heart problems, medication use, stress level, diabetes, and family history. I mainly aim to solve the following two questions:

2.1 To compare the model performance of Extra-trees classifier and Random Forest in predicting a participant had heart attack or not.

2.2 To identify the top 10 important features to prevent heart attack

3 Exploratory Data Analysis and Visualization

Two datasets were provided by the competition. The training set includes 7,010 participants and the testing set includes 1,753 participants with undisclosed heart attack outcomes. In the training set, 30.23% participants were female, with a mean age of 53.51 (SD=21.29). The BMI ranged from 18 to 40 Kg/m2 and the sleep duration ranged from 4 to 10 hours per day. The patients were from all around the world and very diversified, of whom 28.9% were from Asia and 25.5% were from Europe. Overall, 35.72% participants had heart attack.

3.1 Visualization of participants' features

The distributions of the input features by the heart attack outcome are presented in figures 1-4.

3.2 Correlation of participants' features

From the heatmap of the correlation matrix of the features (figures 5), we found age and sex are highly associated with smoking.

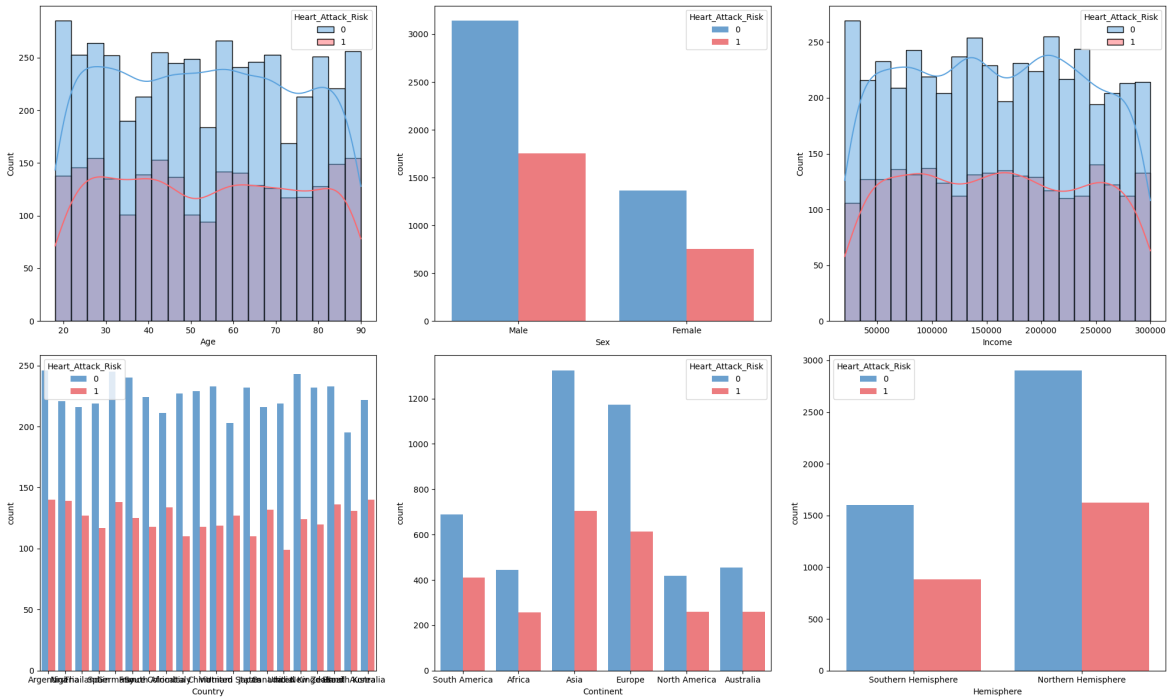

Figure 1: Distributions of socio-demographic features by heart attack outcome.

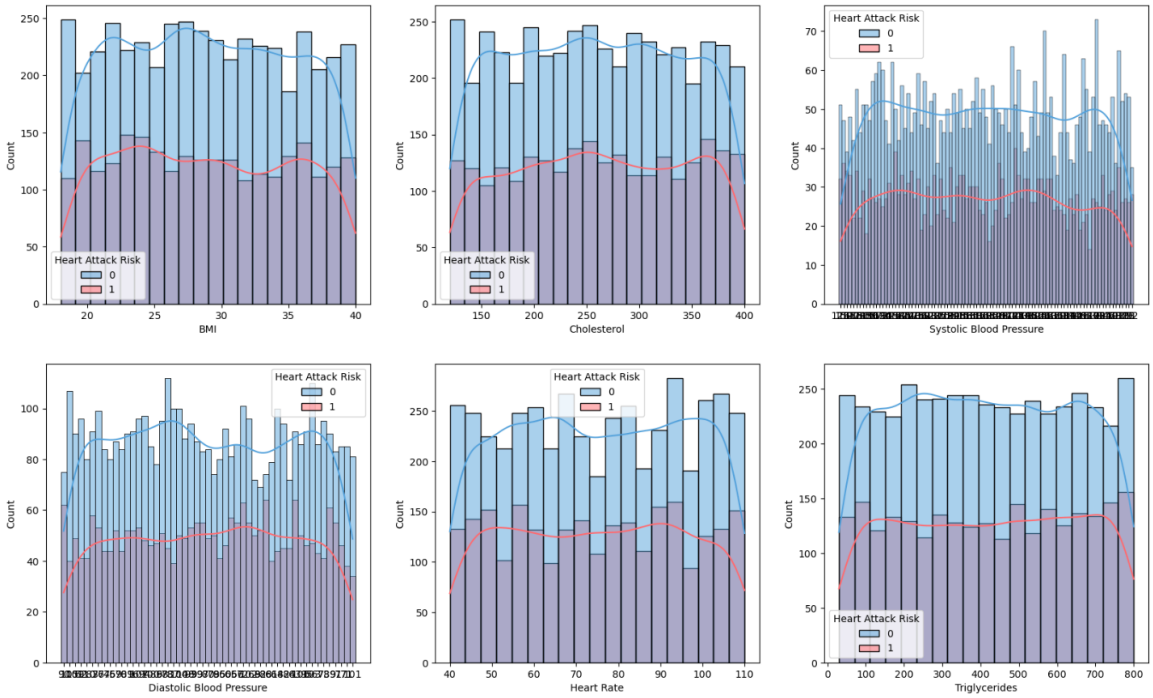

Figure 2: Distributions of clinical characteristics by heart attack outcome.

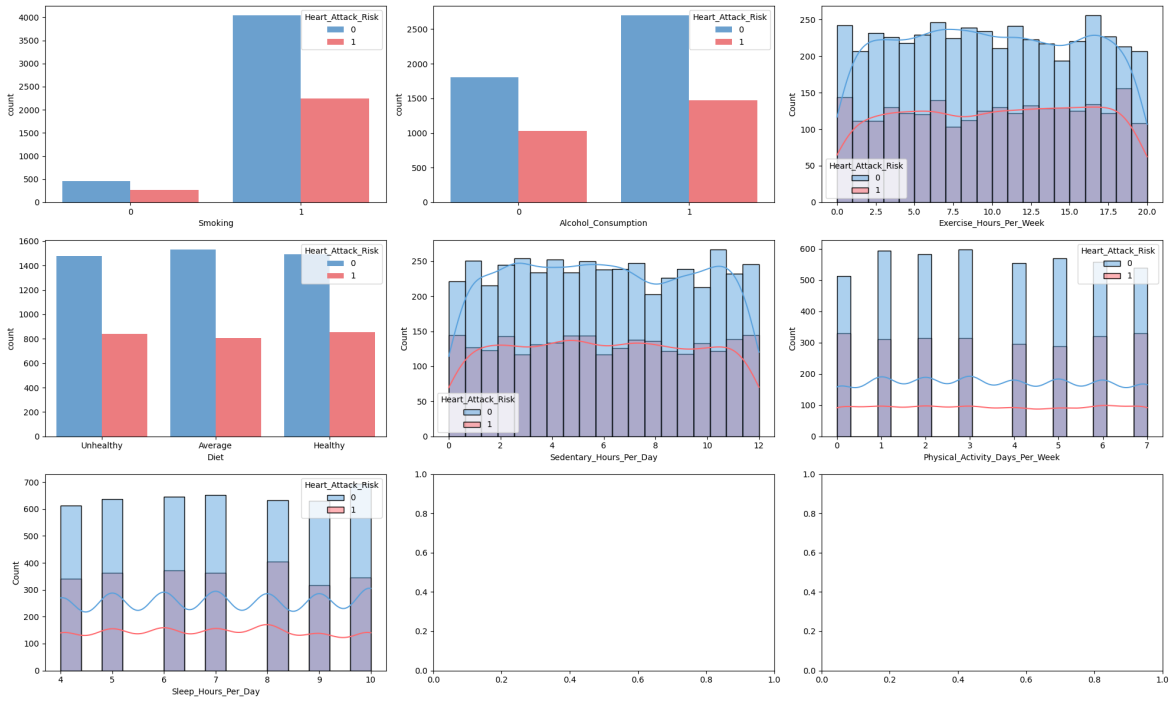

Figure 3: Distributions of lifestyle behaviors by heart attack outcome.

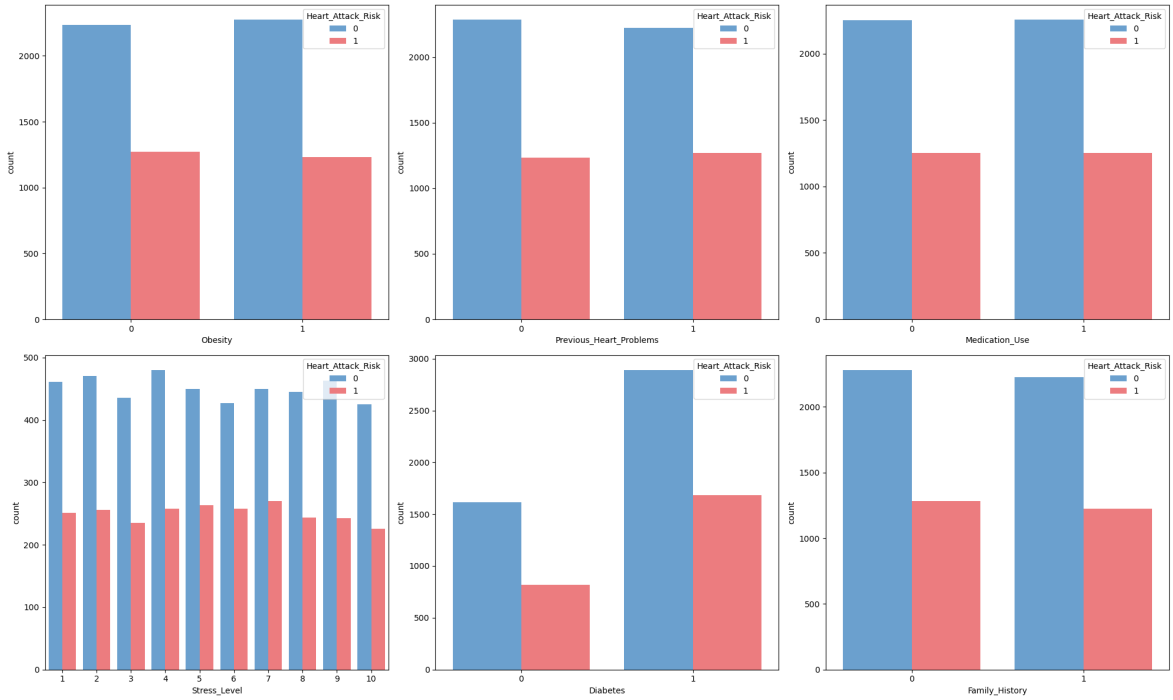

Figure 4: Distributions of comorbidities by heart attack outcome.

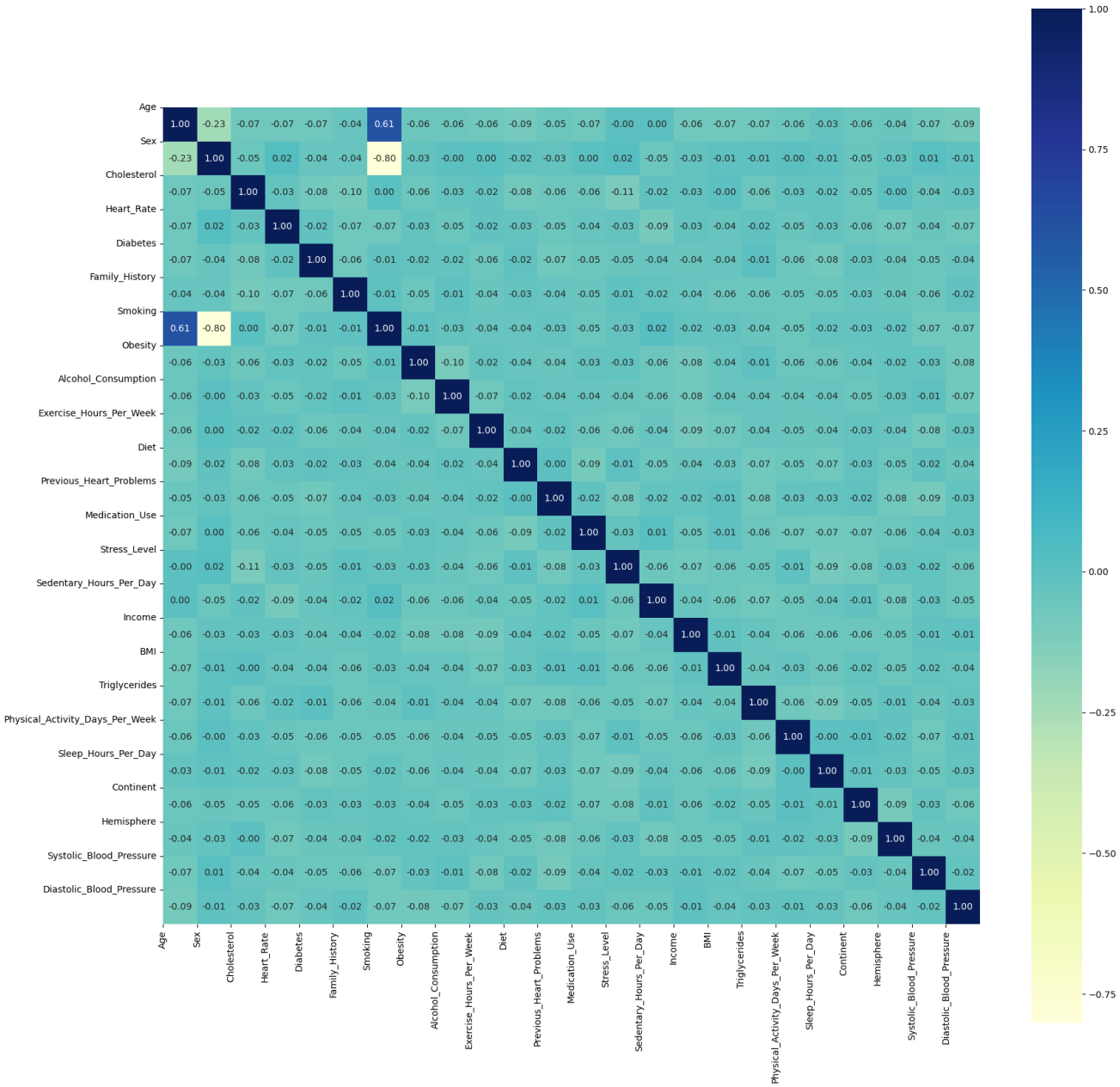

Figure 5: Heatmap of the pairwise correlation matrix of the features.

4 Data Preprocessing

To prepare and process the data for modelling, I divided the combined blood pressure (e.g., 120/80 mmHg) into systolic blood pressure (120 mmHg) and diastolic blood pressure (80 mmHg). Then, I performed one-hot encoding and concerted all the categorical variables into a set of binary dummy variables. All the features were standardized to reduce the models' sensitivity to outlier. No missing value exists in the dataset.

5 Models

Two classification models were selected as candidate models. Some model characteristics and specified hyperparameters are described as fellows,
Extra Trees classifier can build multiple decision trees with high randomness, selecting random thresholds for each feature at each node, leading to reduced overfitting, improved generalization, and easy interpretability. I specified the number of trees in the Ext model as 20 to achieve a balance of the model performance and computational resources.
Random Forest is an ensemble learning approach constructing multiple decision trees with bootstrapped samples and random feature selection, boosting accuracy and resilience to overfitting, and suitable for large datasets with minimal tuning. The turning parameter I specified for the number of trees/bootstrapped samples was also 20 and for the number of predictors considered at each split was as the square root of the total number of features.

5.1 Performance Metrics

The model performance was mainly evaluated based on two metrics, accuracy score and F-1 score. The accuracy score is a metric to compare the number of predictions that match the true labels of the data. The F-1 score is a metric that considers both the precision and recall of the model to compute a single score that balances these two aspects.

5.2 Cross Validation

As common practice in Machine Learning, training dataset was preliminarily split into training set (80%) and validation set (20%). In addition, to test the robustness in estimating the performance of a ML model, I further evaluated the model performance through 10-fold cross-validation (CV). Average metrics and their standard deviations were estimated. The 10-fold CV can provide a more reliable assessment of model performance compared to one time train-test split

6 Results

6.1 Model Performance

	ExtraTrees_Accuracy	ExtraTrees_F1	RandomForest_Accuracy	RandomForest_F1
1	0.62	0.04	0.63	0.02
2	0.65	0.05	0.65	0.05
3	0.62	0.05	0.62	0.04
4	0.63	0.05	0.63	0.06
5	0.61	0.04	0.61	0.06
6	0.66	0.09	0.65	0.05
7	0.62	0.03	0.61	0.03
8	0.64	0.05	0.64	0.04
9	0.61	0.05	0.62	0.05
10	0.66	0.05	0.66	0.03
Average	0.63	0.05	0.63	0.04

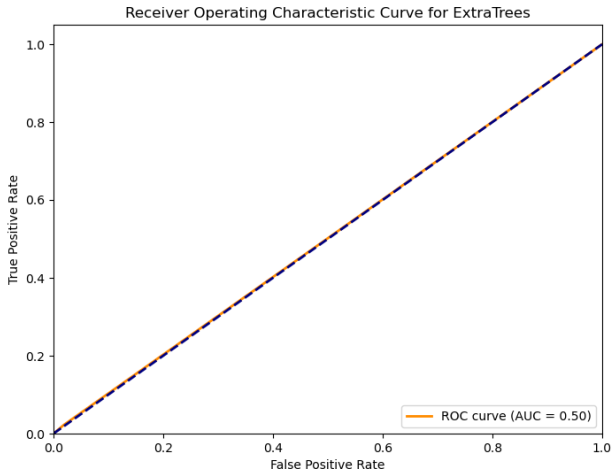

Figure 6: ROC plot for Extra Trees.

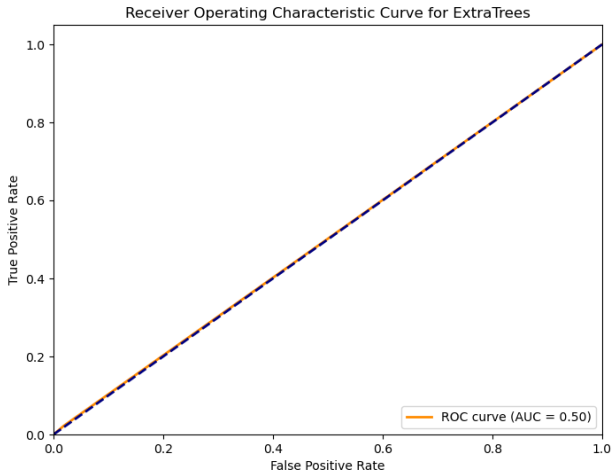

Figure 7: ROC plot for Random Forest.

As we can observe from the table presenting the accuracy and F-1 score of Extra Trees and Random Forest obtained from 10-fold Cross-Validation, the two models have similar performance in accuracy and the Extra Trees did a litter better in F-1 score. Overall, both model had poor performance in predicting heart attack, since the ROC plots show they both got an AUC of around 0.5. The primary reason for the poor prediction performance is that the dataset was generated by AI, resulting in deviations from real-world data and a significant amount of noise.

6.2 Feature importance

Using the Extra Trees model and Random Forest model, we can identify 10 exactly same important features to answer the second question (see Figure 8-9). These features include age, sedentary hours per day, triglycerides, BMI, systolic blood pressure, exercise hours per week, cholesterol, heart rate, income, and diastolic blood pressure. However, the two methods gave slightly different ranks to the selected features. Except for age, We may intervene on those modifiable features to prevent population from getting heart attack.

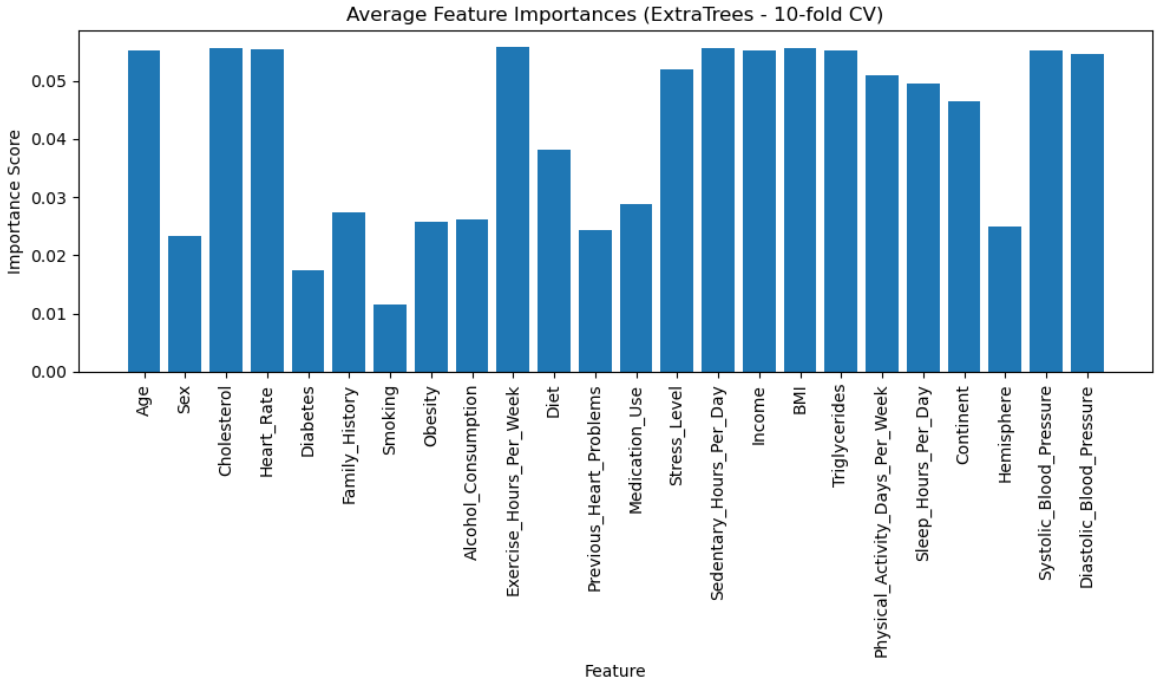

Figure 8: Feature importance identified from Extra Tree model.

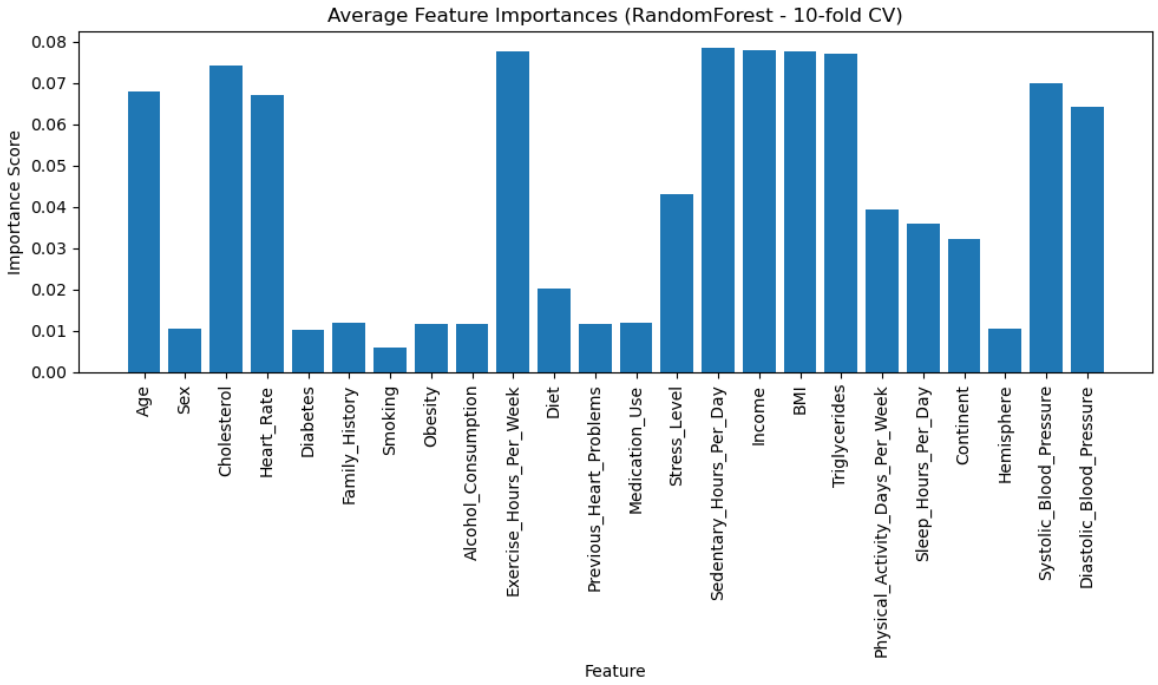

Figure 9: Feature importance identified from Random Forest.