

Maxine Baghdadi
Data Science Institute
Github Link: <https://github.com/maxinebaghdadi12/HeartHealth>

Exploring Machine Learning Techniques for Heart Disease Prediction

Introduction

Heart disease is a leading cause of mortality worldwide, with approximately 805,000 Americans experiencing a heart attack each year.¹ Early detection and risk prediction are essential for improving patient outcomes and efficiently allocating healthcare resources. In recent years, machine learning (ML) has emerged as a powerful tool for analyzing healthcare data and predicting disease risk. Studies such as Dubey et al., which achieved 89% accuracy using Logistic Regression and SVM on the Cleveland dataset,² Li et al., which reported 90.47% accuracy using Random Forest,³ and Dixit et al., which achieved 92.6% accuracy with hybrid ML approaches,⁴ highlight the potential of ML in heart disease prediction. However, these studies often rely on accuracy as the primary metric, which can be misleading in the context of imbalanced datasets and fails to capture model performance nuances when false negatives—patients misclassified as low risk—carry severe consequences.

This study applies several ML techniques, selecting an f-beta score greater than 1 as the primary evaluation metric to emphasize recall over precision. This approach ensures the models are tuned to minimize false negatives, reducing the likelihood of overlooking at-risk patients. The dataset for this research is a subset of the Behavioral Risk Factor Surveillance System (BRFSS), an annual survey conducted by the Centers for Disease Control and Prevention⁵. It includes 21 behavioral, demographic, and clinical health indicators, such as age, BMI, blood pressure, and glucose levels, collected from over 250,000 individuals. The target variable is whether a patient has experienced a heart attack or disease, framing the problem as a binary classification task. By focusing on recall and examining critical health indicators, this study enhances early detection efforts and identifies key predictors of heart disease risk.

EDA

Initial exploratory data analysis is used to analyze the relationships and shapes of the variables and dataset. Both demographic data and clinical data was explored.

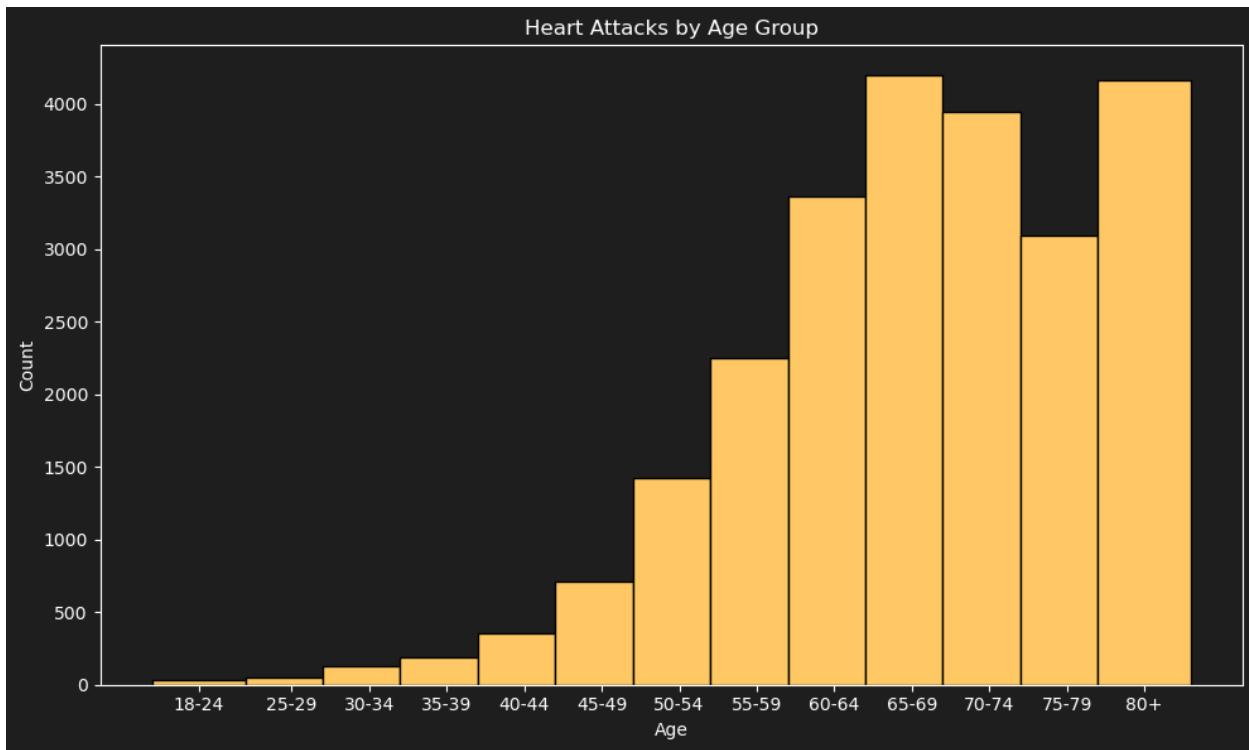


Fig. 1 Histogram displaying distribution of heart attacks by age group

As demonstrated, the number of people who have experienced a heart attack or disease significantly increases with age. In this dataset, there is a sharp rise starting after the age of 50 and peaking between 60 to 69 years old. The highest count of those who have experienced a heart attack or disease is individuals who are 65+.

To analyze the most important clinical data variables, the below correlation maps the relationship between critical health conditions and the target variable. As seen below, the strongest correlations are high blood pressure, diabetes and physical health with correlations of 0.21, 0.18 and 0.18 respectively.

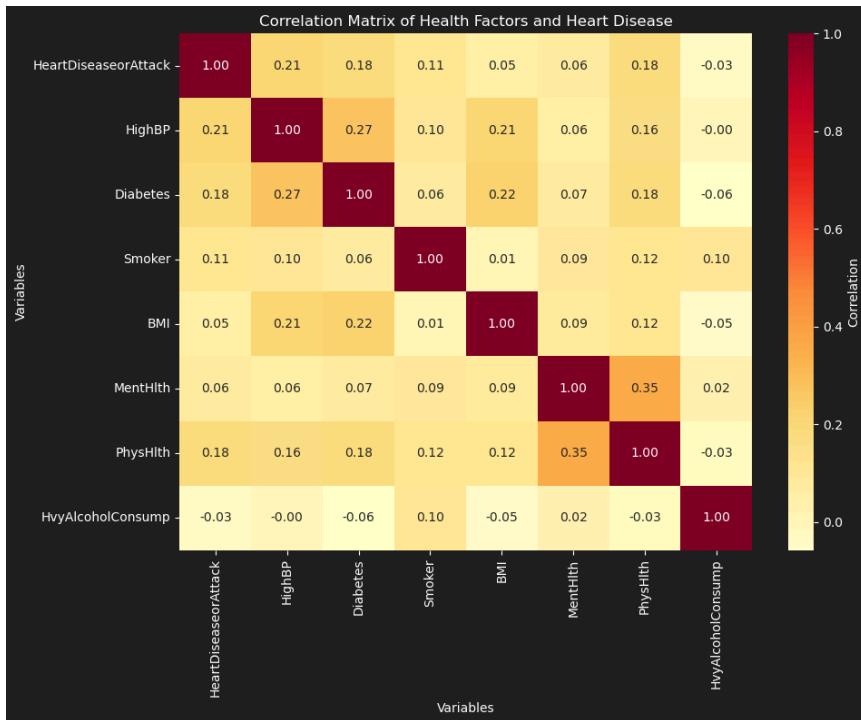


Fig. 2 Correlation Matrix relating clinical data to target variable

Looking deeper to understand the relationship between these variables, a consistent trend emerges: A higher proportion of those who have experienced severe health conditions have experienced a heart attack. The below figures demonstrate this trend for both “High BP” and “Diabetes.”

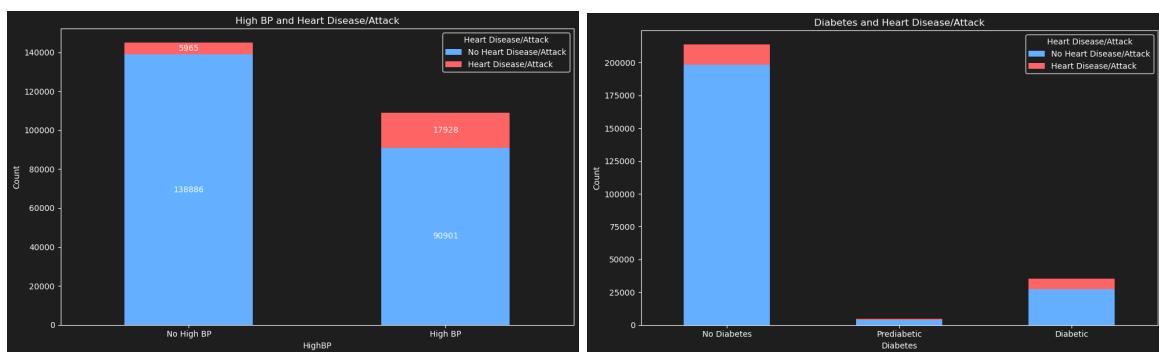


Fig 3 Proportion of those with diabetes and High BP.

After general EDA was conducted to understand the variables and their relationships, the dataset was explored to understand the balance of data.

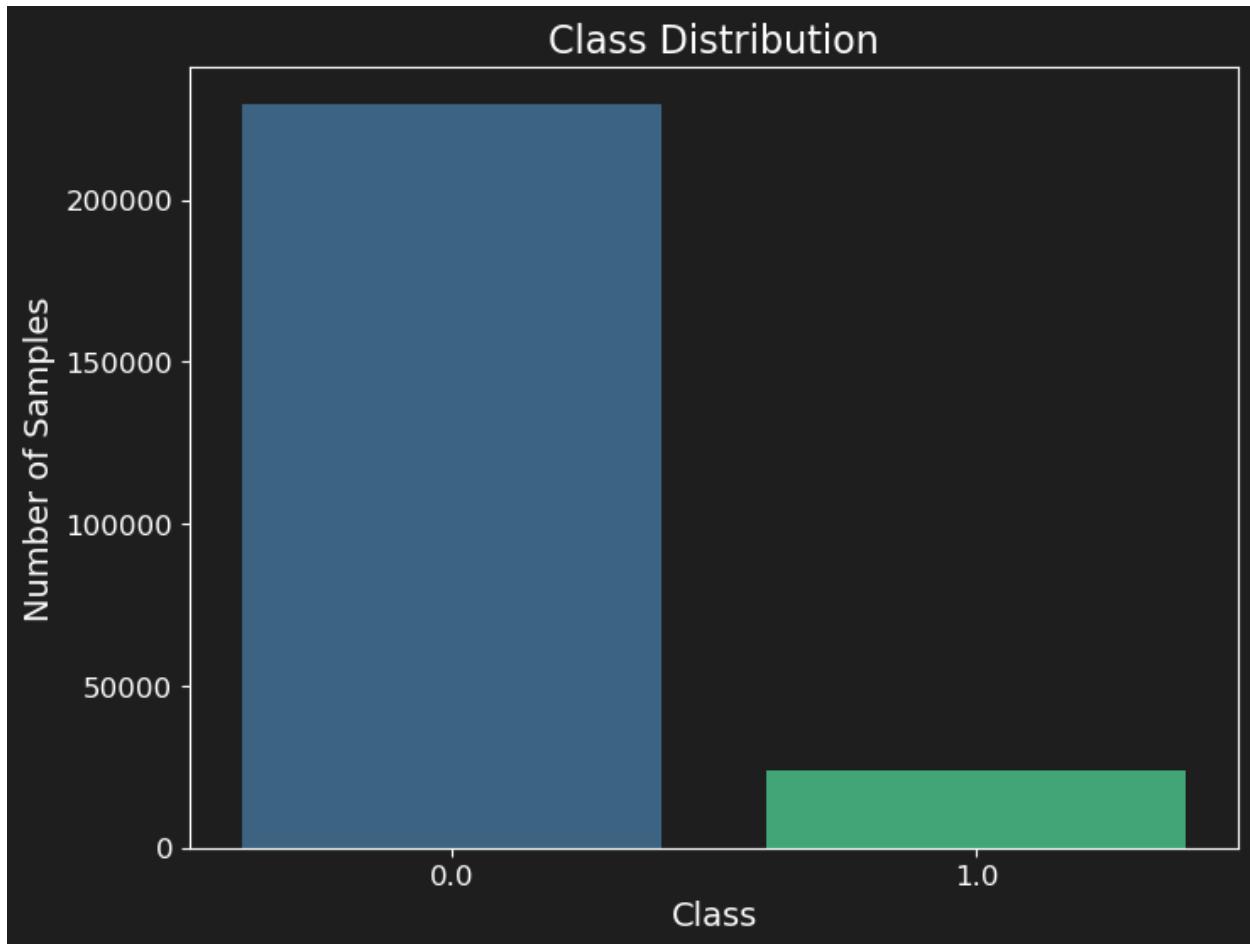


Fig 4 Class Distribution

The dataset is highly imbalanced with only 9% of the data actually in class 1.

Methods

The next step of model development deals with splitting, feature preprocessing, and hyperparameter tuning. This paper explores four models: Logistic Regression, Support Vector Classifier (SVC), Random Forest and XGBoost evaluated on their F5 scores.

Splitting Strategy

As previously demonstrated, only 9% of the data is in Class 1, lending to a highly imbalanced dataset. Therefore, our splitting strategy is a stratified split, using 20% of the data for testing. The remaining 80% is further divided into training and validation sets using stratified sampling. Additionally, a 4-fold cross-validation is employed within the training set to evaluate models and select hyper parameters. Below is a figure demonstrating the splitting of the data:

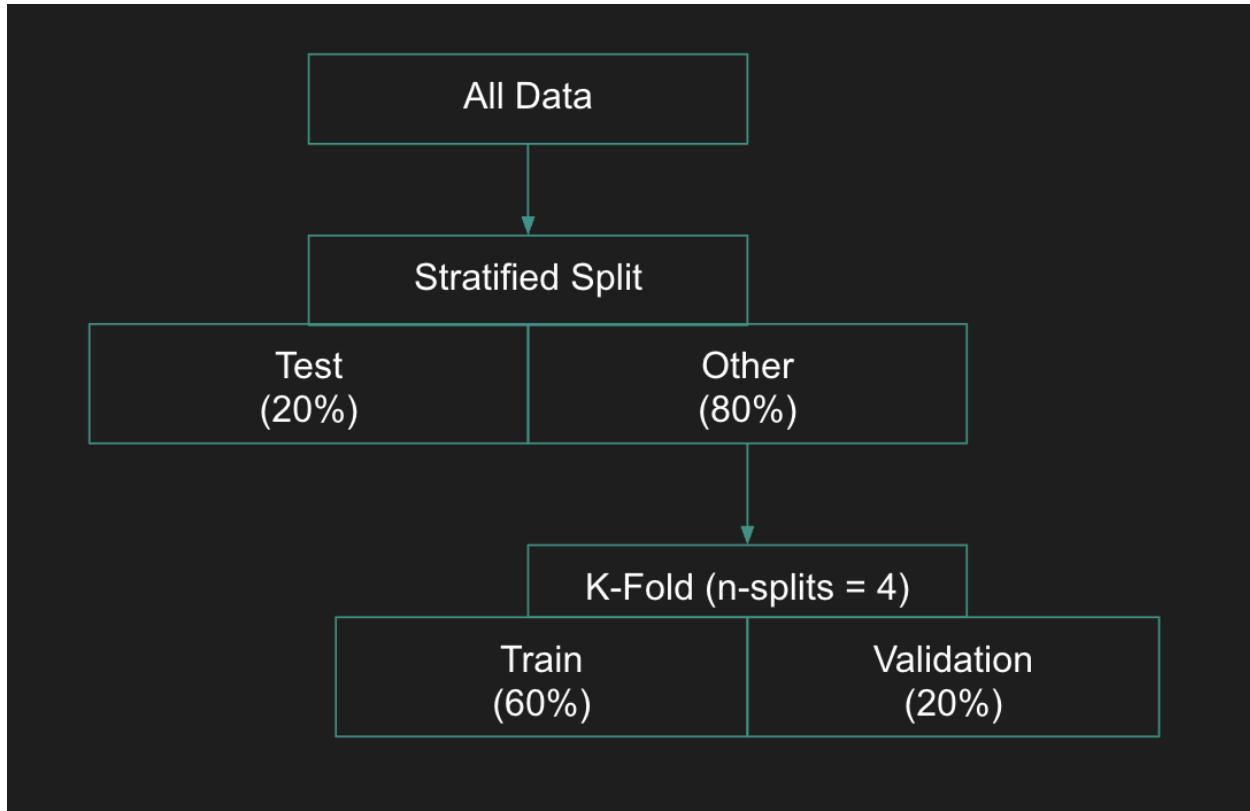


Fig 5 Data splitting strategy

Data Preprocessing

In order to preprocess the data dynamically in each model, a preprocessor pipeline was developed to later be applied to our training sets. The dataset contained no missing values and all strings were already converted to numbers. The preprocessor focuses on scaling the features to ensure a mean around 1 with a standard deviation around 0 by leveraging a `StandardScaler()`.

Hyperparameter Tuning

The purpose of hyperparameter tuning is to identify the parameters that optimize the model's performance with respect to the evaluation metric. This study utilizes a grid search to iterate through several parameter configurations for each model, train the model on a portion of the training data and evaluate the performance against a validation set. The validation sets were built using a 4-fold cross-validation to ensure robustness in identifying the hyperparameters. Below are the hyperparameters and values tuned for each model:

Model	Hyperparameters
Random Forest	max_depth: [10, 30, 100, 300] min_samples_split: [16, 32, 64, 128, 256]
Logistic Regression	c: [0.01, 0.1, 1, 10, 100] penalty: ['l1', 'l2'] solver: ['liblinear', 'saga']
XGBoost	n_estimators: [100, 300, 500] gamma: [0, 1, 5, 100] max_depth: [10, 30, 100, 300] min_child_weight: [5, 10, 20]
SVC	c: [0.1, 1, 10, 100] gamma: ['scale', 'auto']

Table 1 Summary of Hyperparameters

The best hyperparameters were chosen based on the highest validation F5 score, prioritizing recall over precision to identify as many true positives (TPs) as possible while minimizing false negatives (FNs). A FN, in this case, would mean failing to alert a patient at risk of heart disease or a heart attack. By setting $\beta = 5$, the model emphasizes recall, though it may increase false positives (FPs). However, the assumed cost of FPs, such as mistakenly warning patients, is considered low compared to the potential benefits of early intervention.

Once the hyperparameters were selected, a model with those specified parameters was fitted onto the test data. This entire process was conducted over five random states, meaning five separate test scores were generated. This approach allowed us to measure the uncertainties of our evaluation metric by capturing the variability introduced by data splitting and the inherent randomness of non-deterministic ML methods. By averaging the F5 scores across these random states, we obtained a more robust and reliable estimate of the model's predictive power.

One last consideration while tuning the model was the imbalance of the data. It is imperative to ensure that no bias is given towards the majority class while actually running the models. To account for that, a balanced class weighting was used for all models except for XGBoost, which used a scale_pos_weight to ensure that this imbalance was accounted for.

Results

A baseline score was calculated so as to compare to the results of all the models. Below is a graph that captures the mean test score and their standard deviations. The baseline test score was 0.52 and is displayed with the purple line.

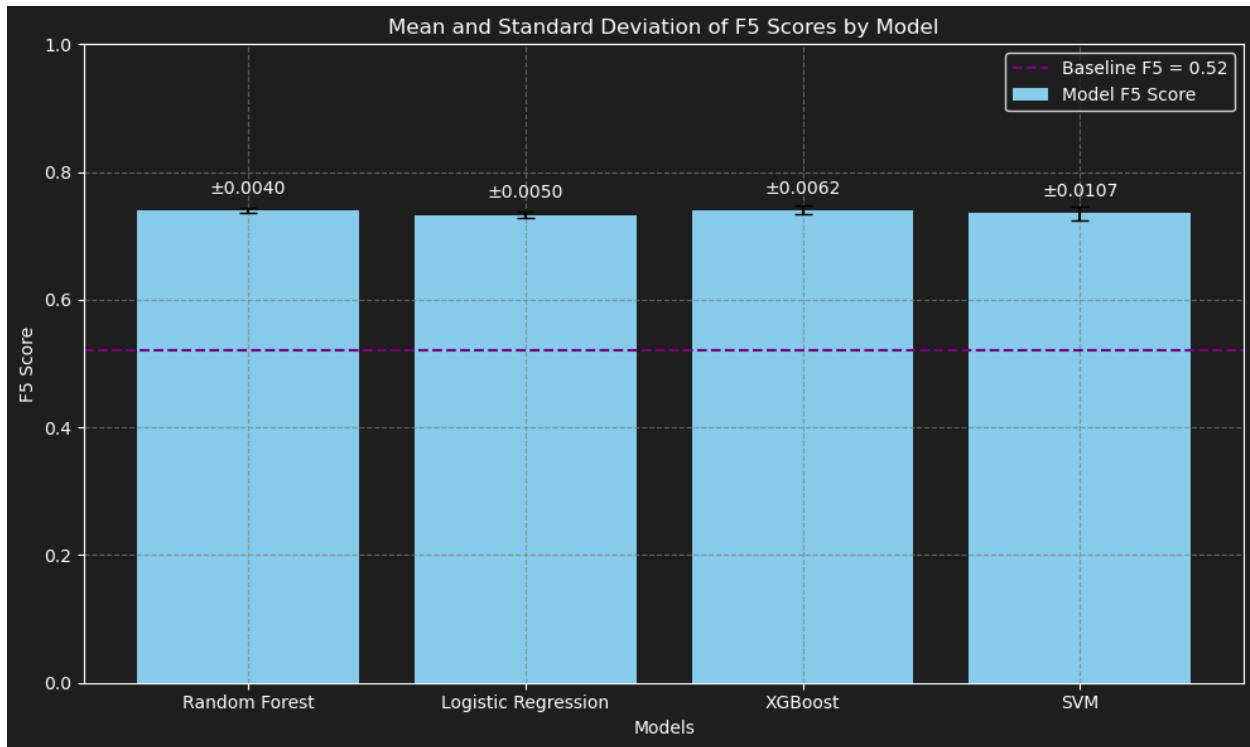


Fig 6 Mean test core for every model benchmarked with the baseline F5 score

All models performed better than the baseline and the best performing model was the Random Forest Classifier with an average F-5 score of 0.7430 across the five test sets. The Random Forest Classifier in itself was approximately 57 standard deviations above the baseline.

The best hyperparameters for the Random Forest Model were a `max_depth` of 10 and a `min_samples_split` of 256.

Using the prediction values from the test set that returned the highest test score, here is a confusion matrix:

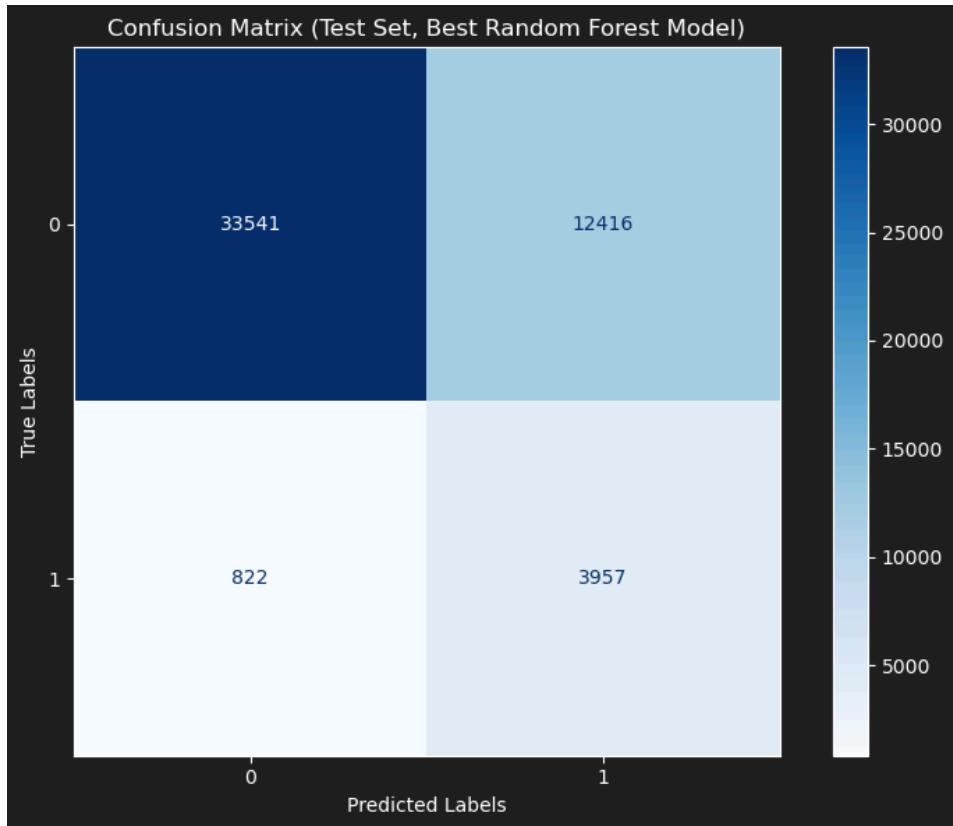


Fig 7 Confusion Matrix with Best Test Set using Random Forest Classifier

It is evident that the Random Forest Classifier successfully kept a small number of people in the FN box and captured a significant number of TPs. However, this came at the expense of an increased number of FPs. By using an F-5 score as the evaluation metric, this was expected.

Global Feature Importance

In order to better understand and interpret the model, global feature importance metrics help gauge the contribution of a feature to the predictability of the model. This report generated three different methods of doing so.

Permutation feature importance showed the strong impact of health-related features to the predictability of a person having a heart attack or disease is clearly demonstrated. Having a previous stroke, general health, history of diabetes and BMI are all predictors of a person's probability of developing a heart disease or attack. Age, surprisingly, does not appear in the top 10 most important features for this model, even though it is a well-known risk factor for heart disease and heart attacks. Although it may be surprising, sex emerges as one of the top features

in the model, which aligns with general existing research that shows that biological differences between males and females can influence heart disease risk.

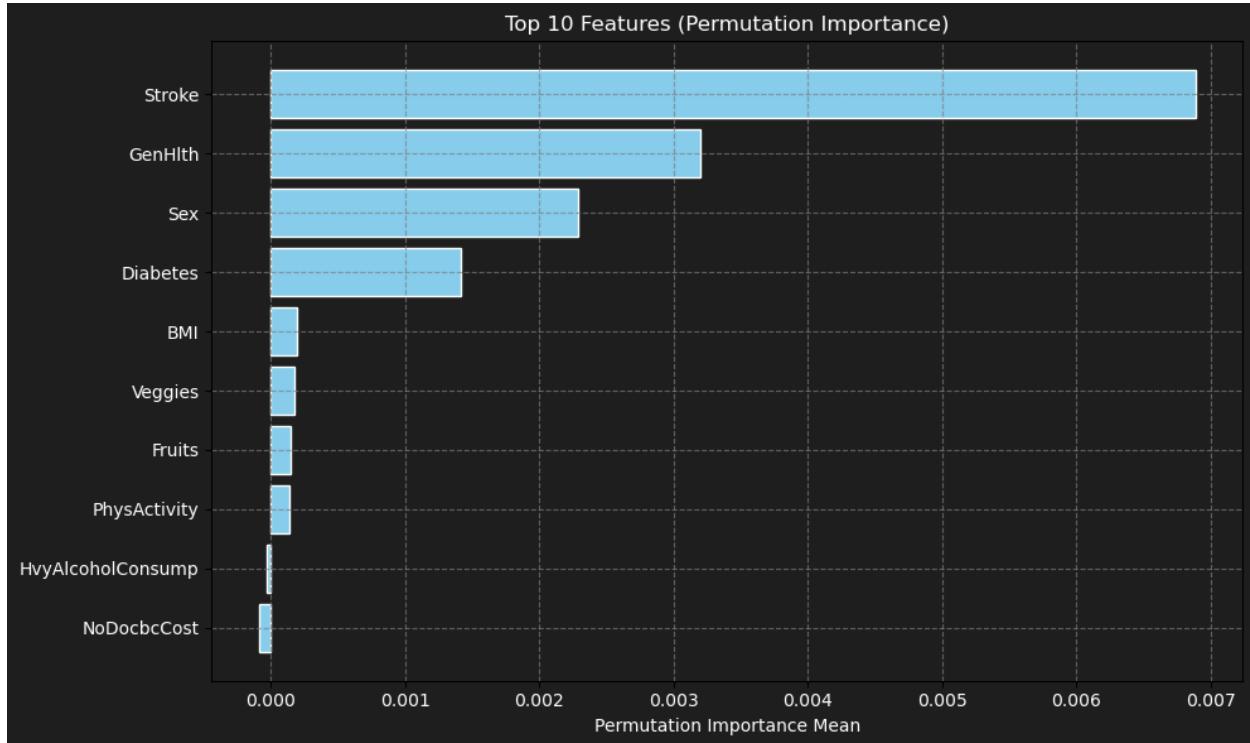


Fig 8 Top 10 Features calculated using Permutation Importance

A global SHAP Value plot can also demonstrate the most important features. There is generally a lot of overlap between this and the permutation feature importance. Namely, general health, BP, Cholesterol. However, this plot brings to life what was seen during EDA: the importance of age on developing a heart attack or disease. According to this calculation, age is *the* most important feature in determining that.

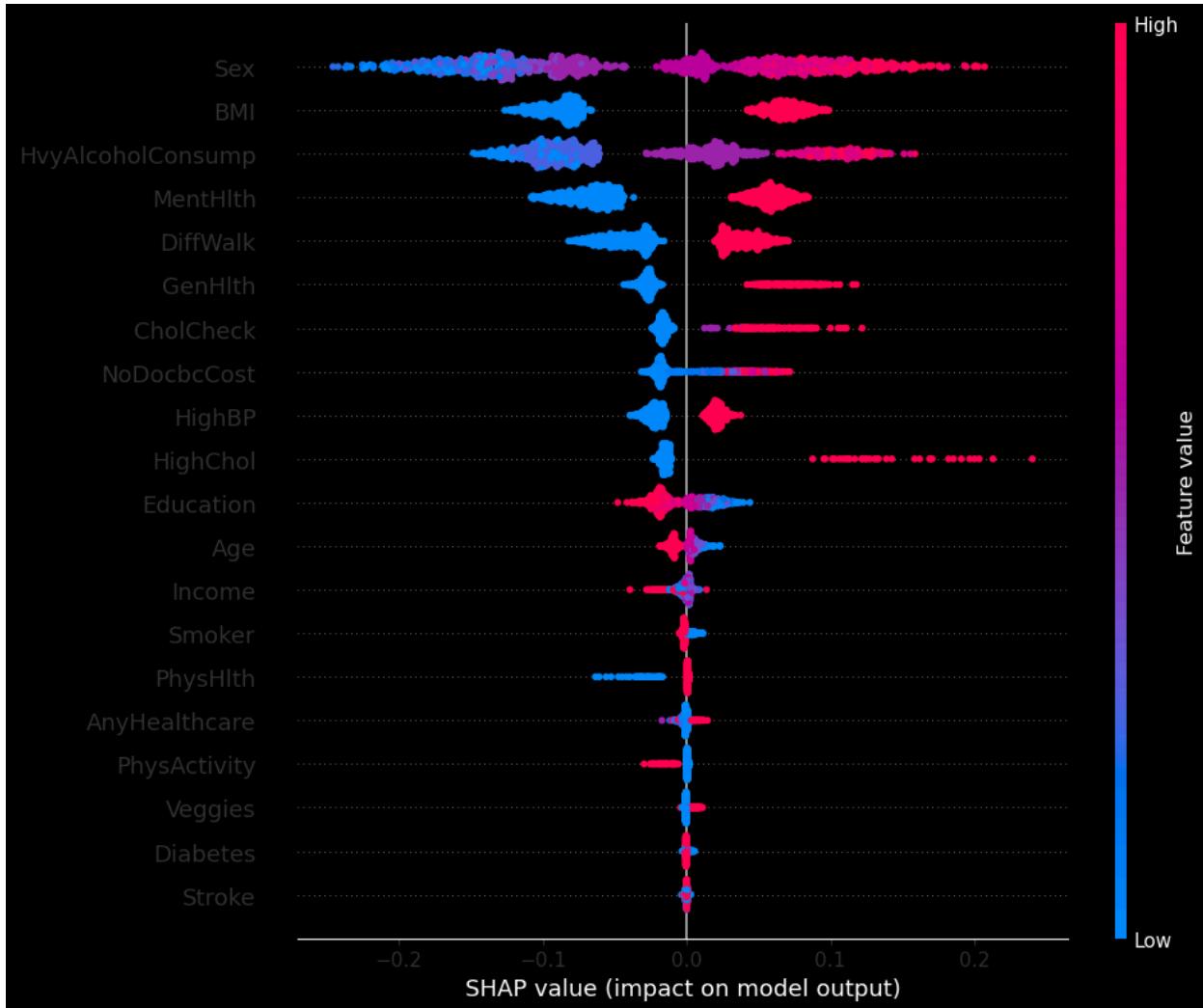


Fig 9 SHAP Global Value Plot

This bar chart, based on Random Forest feature importances, identifies Sex, HvyAlcoholConsump, and BMI as the top predictors by mean decrease in impurity. Unlike the permutation importance plot, where Stroke ranked highly, it is absent here, highlighting how Random Forest emphasizes features influencing tree splits, while permutation importance focuses on predictive performance.

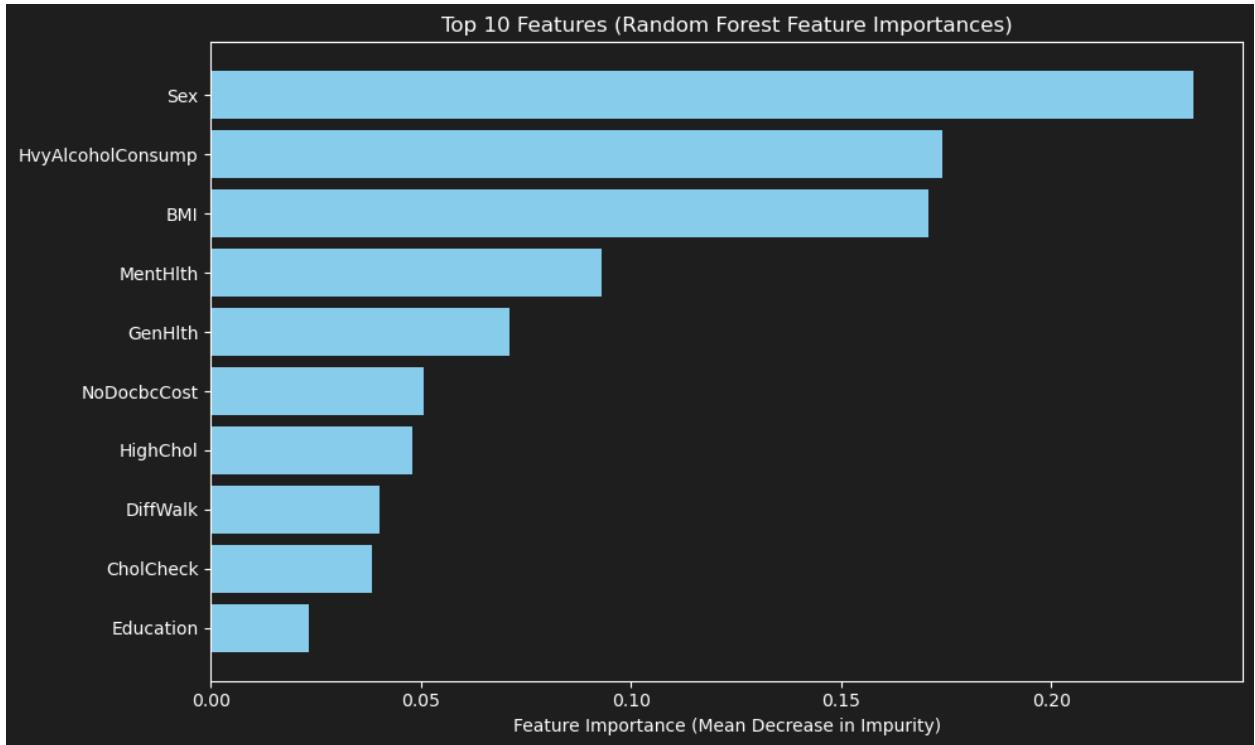


Fig 10 Feature Importance Using Impurity

Local Feature Importance

Local feature importance was also studied following the development of the global SHAP Value plot. Below is a local SHAP plot for a randomly indexed row of class 0 and of class 1:



Fig 11 Class 0 SHAP Plot



Fig 12 Class 1 SHAP Plot

The SHAP force plots highlight how health-related features influence predictions for specific observations. For Class 0, features like high BP = 0 and high cholesterol = 0 decrease the prediction probability, while heavy alcohol consumption = 1 slightly increases it. For Class 1, you see nearly the exact reverse, showcasing the model's sensitivity to health and lifestyle factors.

Overall, it is evident that the most important features in determining whether a person will develop a heart attack or disease is their age, general health and presence of high cholesterol, high BP or diabetes.

Outlook

The outlook is the place to describe what else you could do to improve the model or the interpretability, and what are the weak spots of your modeling approach. How would you improve this model? What additional techniques could you have used? What additional data could you collect to improve model performance?

There are numerous ways to improve this model and its interpretability. One key step involves exploring other potential models or fine-tuning the parameters of those already used. While the SVC model demonstrated strong performance, it was trained on only 20% of the dataset due to computational limitations. Running the SVC model on the full dataset would be a crucial next step, as it might outperform the Random Forest model by better capturing the nuances in the data. Additionally, experimenting with ensemble techniques, such as combining predictions from different models, could lead to more robust and accurate results.

The Random Forest model, while effective, can also be refined further. Testing various F-Beta values to identify the optimal trade-off between precision and recall is essential for ensuring the model aligns with real-world priorities. Collaboration with domain experts in healthcare could help determine the acceptable costs of false positives versus false negatives, as preventative care for heart disease or heart attacks can involve significant financial or personal implications. Finally, the dataset could benefit from temporal data, such as tracking variables over time, which would allow the model to account for changes in health patterns or risk factors and potentially provide more actionable predictions.

References

- 1https://www.cdc.gov/heart-disease/data-research/facts-stats/?CDC_AAref_Val=https://www.cdc.gov/heartdisease/facts.htm
- 2<https://www.techscience.com/iasc/v30n3/44095/html>
- 3https://www.researchgate.net/publication/342058675_Heart_Disease_Identification_Method_Using_Machine_Learning_Classification_in_E-Healthcare
- 4<https://link.springer.com/article/10.1007/s11042-021-11083-9>
- 5<https://www.kaggle.com/code/alextreboul/heart-disease-health-indicators-dataset-notebook>