Heart Disease Prediction Model
An Analysis of Key Indicators and Metrics

Sanjana Murali Krishna
Biomedical Engineering ScM
https://github.com/sanmkBU/Data1030-Project-repo.git

1. INTRODUCTION

Heart diseases are the leading cause of mortality worldwide, accounting for millions of deaths annually. A significant contributor to these diseases is the narrowing of blood vessels, a condition that can lead to serious heart complications. The ability to predict the likelihood of a person developing heart disease based on this factor is of importance. It can facilitate early intervention, potentially altering the course of the disease and improving patient healthcare and costs.

Therefore, the use of machine learning models can be beneficial. These models can analyze a multitude of factors, like cholesterol, blood sugar levels, types of chest pain, including the diameter narrowing of blood vessels, and many more factors, to predict the risk of heart disease. By training these models on health data, we can equip them to make accurate predictions on new, unseen data.

ABOUT THE DATASET AND TARGET VARIABLE

The dataset used in this study, sourced from Kaggle, was compiled from the medical records of four databases: the Cleveland Clinic Foundation, the Hungarian Institute of Cardiology, the V.A. Medical Center in Long Beach, CA, and the University Hospital in Zurich, Switzerland.

This classification dataset comprises 294 data points and 14 features, providing a comprehensive overview of various health parameters. The features in the dataset include age, sex, chest pain type (cp), resting blood pressure (trestbps), serum cholesterol (chol), fasting blood sugar (fbs), resting electrocardiographic results (restecg), maximum heart rate achieved (thalach), exercise-induced angina (exang), ST depression induced by exercise relative to rest (oldpeak), the slope of the peak exercise ST segment (slope), number of major vessels colored by fluoroscopy (ca), thalassemia (thal), and the diagnosis of heart disease (num).

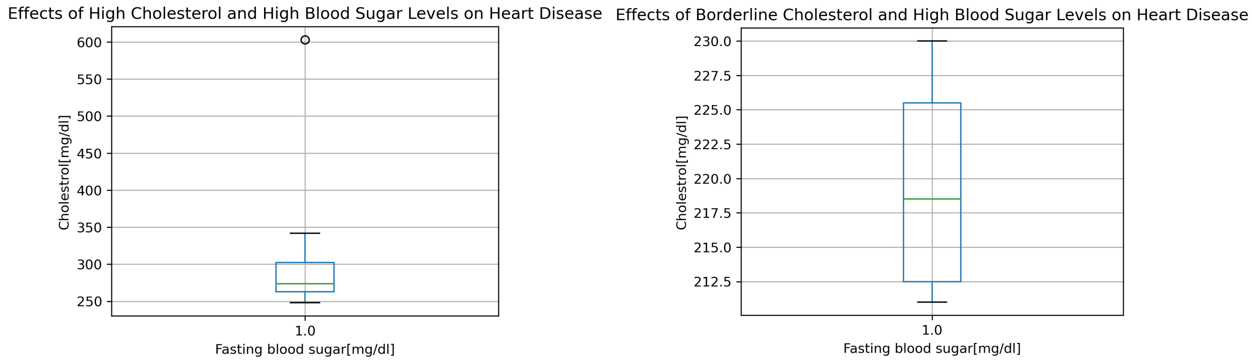
The target variable in this dataset is the diameter narrowing of blood vessels. A value less than 50% (indicated as 1 in the dataset) suggests that the patient is at risk of a heart attack, while a value greater than 50% (indicated as 0 in the dataset) suggests a lower risk.

2. EXPLORATORY DATA ANALYSIS

Exploratory data analysis (EDA) was conducted to gain insights into the dataset. The data types of the features were examined, and summary statistics were calculated to provide a general understanding of the data distribution using the .describe function. The presence of missing values was also assessed, with some features like slope, ca, thal from the dataset containing a large number of missing values that will be addressed during preprocessing.

Additional EDA was conducted to generate different plots incorporating the features within the dataset. These plots provide a comprehensive overview of the relationships between various factors and heart disease status.

Figure.1. shows two box plots that display the effects of high cholesterol and high blood sugar levels on heart disease. These plots can provide insights into the distribution of cholesterol levels among patients with high blood sugar levels, which can be useful in understanding the relationship between these two factors and heart disease.


Effects of High Cholesterol and High Blood Sugar Levels on Heart Disease

Effects of Borderline Cholesterol and High Blood Sugar Levels on Heart Disease

Figure 1. Patients with high blood sugar levels for high cholesterol and borderline cholesterol

Upon further analysis, I determined that out of four patients exhibiting borderline cholesterol and elevated fasting blood sugar levels, two were diagnosed with heart disease. Furthermore, a significant proportion of patients, 11 out of 14, experienced a cardiac attack, which was attributed to high cholesterol and blood sugar levels.

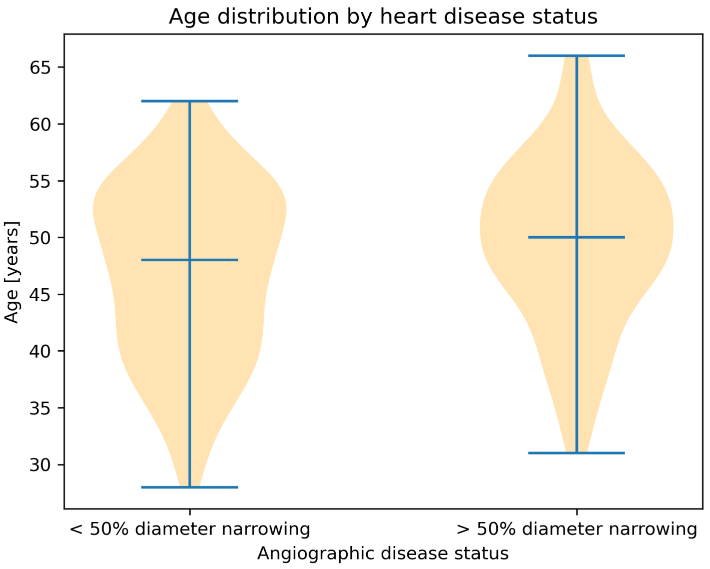


Figure 2. Age distribution of patients based on the Angiographic disease status

Figure.2. shows is a violin plot that shows the age distribution (between 25 and 70) by heart disease status. The x-axis represents the angiographic disease status (i.e. diameter narrowing). This plot can provide insights into whether age influences the likelihood of having heart disease.

After additional study of the dataset, I observed that out of the total number of patients aged 35 and above, 11 out of 76 females and 96 out of 202 males had experienced a heart attack. This data suggests a higher prevalence of heart disease among male patients in this age group.

Figure.3. depicts a stacked bar plot that shows the distribution of chest pain types by heart disease status. The 'cp' variable to represent different types of chest pain namely typical angina, atypical angina, non-anginal pain, and asymptomatic.

Based on the analysis of the dataset, it is evident that patients presenting with atypical angina demonstrate a notably higher tendency for experiencing a heart attack, followed by those with typical angina, non-anginal pain, and asymptomatic cases.

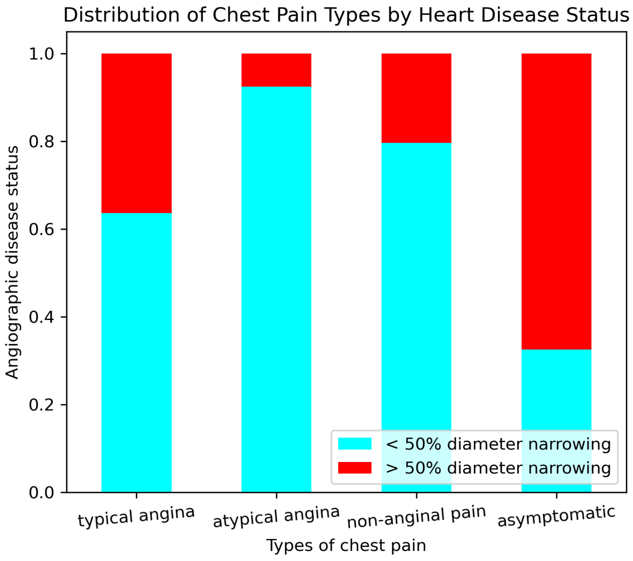


Figure 3. Distribution of patients experiencing a heart attack by the specific type of chest pain they experience.

3. METHODS

DATASET SPLITTING

The target variable (y) and feature matrix (X) were initially derived from the dataset, following which the dataset was split into the X_test set and X_other set using the train_test_split function. Subsequently, the X_other set was further divided into the X_train and X_val sets utilizing the StratifiedK-fold split method with 4 splits. This approach was adopted due to the slight imbalance observed in the target variable. The dataset was apportioned into the three sets - X_train, X_test, and X_val - in a ratio of 6:2:2, signifying 60% for training, and 20% each for testing and validation, respectively. This stratified splitting strategy ensures that the distribution of the target variable is preserved across all subsets, thereby enabling robust model training, evaluation, and validation.

DATASET PREPROCESSING

During preprocessing of the data for model training, it was noted that the dataset consists of only numerical values. To standardize the scale of these numerical features, the standard scaler function from the sklearn library was used within the preprocessing pipeline. This involved the application of the fit transform method to the training set, resulting in the normalization of the dataset. Subsequently, to ensure uniformity in the preprocessing of the data, the same transformation was applied to both the validation and test sets.

After the preprocessing phase, it was observed that the dataset, which initially contained missing values, was successfully transformed, resulting in a reduction to 176 data points and 13 features. The fraction of points with missing values, now standing at 0.9943.

ML PIPELINE

In the course of the analysis, four distinct machine learning techniques were employed to gain insights from the dataset. These techniques included the XGBoost Classifier, RandomForest Classifier, Support Vector Machine (SVM), and Logistic Regression. To optimize the performance of each model, a process of hyperparameter tuning was undertaken.

Table 1. Tuned Hyperparameters for the Machine Learning Models

XGBoost Classifier	'n_estimators': [100, 200, 300] | 'max_depth': [2, 3, 4] | 'learning_rate': [0.01, 0.1, 0.2]
RandomForest Classifier	'max_depth': [1, 3, 10, 30, 100] | 'max_features': [0.5,0.75,1.0]
Support Vector Machine (SVM)	'C': [0.1, 1, 10, 100] | 'gamma': [0.001, 0.01, 0.1, 1] | 'kernel':['rbf']
Logistic Regression	'C': [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0] | 'penalty':['none','l2']

To address uncertainties arising from both data splitting a methodological approach was employed. The model training, testing and validating processes were iteratively executed over 10 random states for all four models, namely the XGBoost Classifier, RandomForest Classifier, Support Vector Machine (SVM), and Logistic Regression. To address the presence of missing values, the Iterative Imputer was employed for the RandomForest Classifier, SVM, and Logistic Regression models. Furthermore, the GridSearchCV library was used to identify the most optimal hyperparameters for each of these three models. The F1_score metric was selected as the scoring metric for these models, given its efficiency in addressing the imbalance observed in the target variable.

4. RESULTS

Table 2. Mean and Standard deviation of the test scores for all the ML models

Models	Mean of Test Scores	Standard Deviation of test scores
XGBoost	0.7048	0.0551
RandomForest	0.7822	0.0708
SVM	0.7519	0.0628
LogisticRegression	0.7640	0.0540

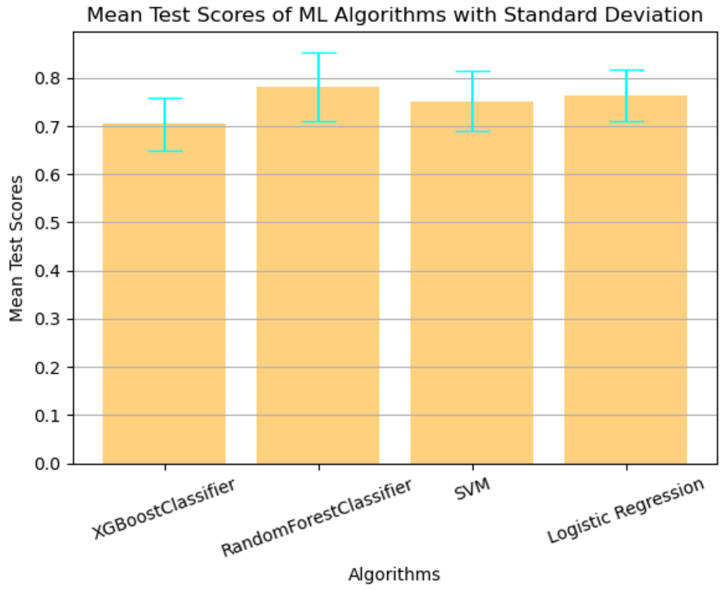
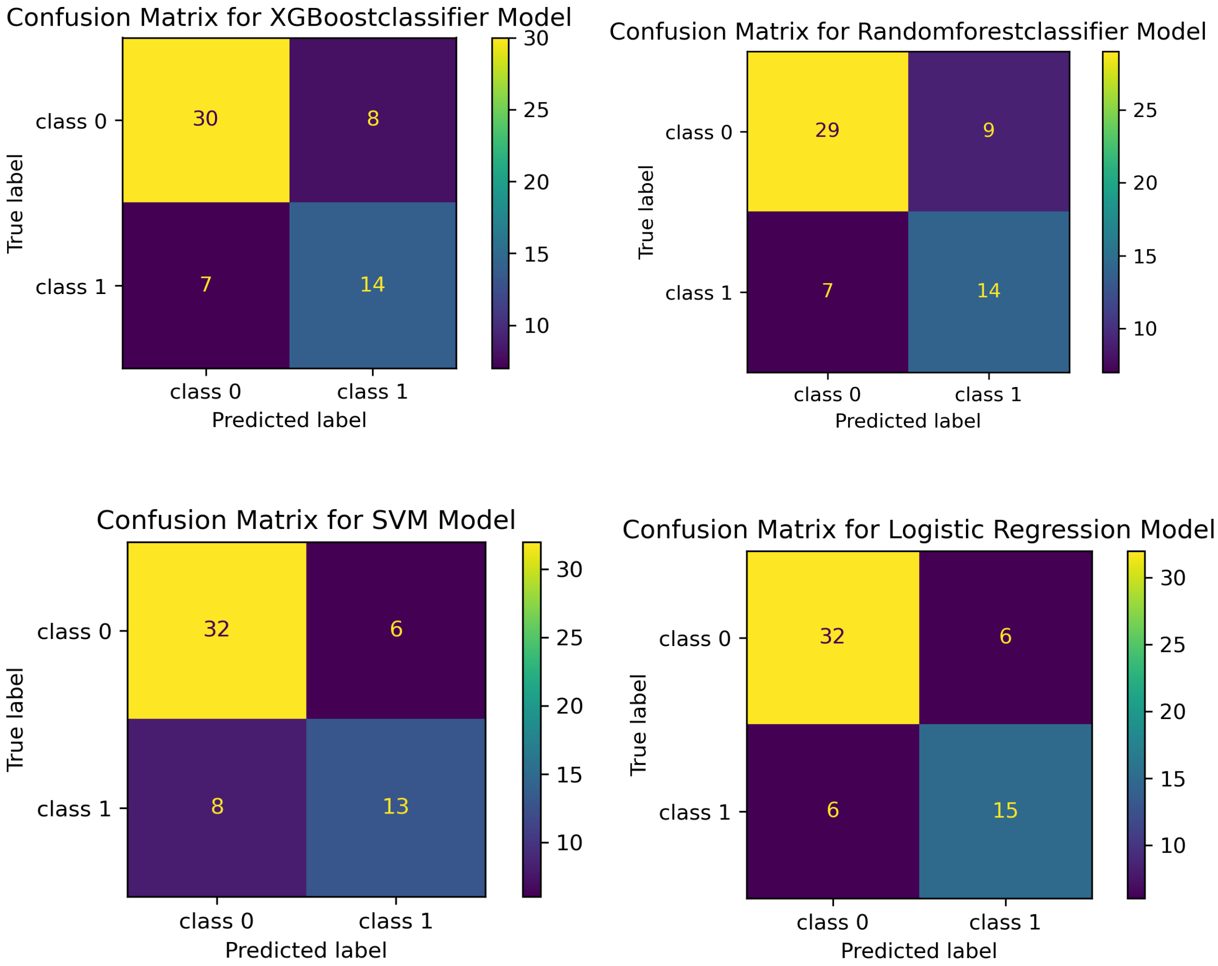


Figure 4. Mean and Standard Deviation for the Machine Learning models.

Figure.4. Shows the mean and standard deviation for the XGBoost classifier, RandomForest classifier, Support Vector Machine (SVM), and Logistic Regression models. The bar height indicates the mean scores, and the error bar indicates the standard deviation scores.



Confusion Matrix for XGBoostclassifier Model

Confusion Matrix for Randomforestclassifier Model

Confusion Matrix for SVM Model

Confusion Matrix for Logistic Regression Model

Figure 5. Confusion matrices for all the machine learning models

Figure.5. Displays the Confusion matrices for XGBoost Classifier, RandomForest Classifier, Support Vector Machine (SVM), and Logistic Regression models, it is evident that the Logistic Regression model demonstrates a relatively lower frequency of instances for false positives and false negatives. This observation suggests that the Logistic Regression model exhibits a more favorable performance in the context of this specific dataset.

Table 3. Accuracy, Precision, Recall and F1 scores for all the ML models

Models	Accuracy score	Precision score	Recall score	F1 score
XGBoost	0.7457	0.6364	0.6666	0.6512
RandomForest	0.7288	0.6087	0.6667	0.6364
SVM	0.7627	0.6842	0.61904	0.65
LogisticRegression	0.7966	0.7142	0.7142	0.7142

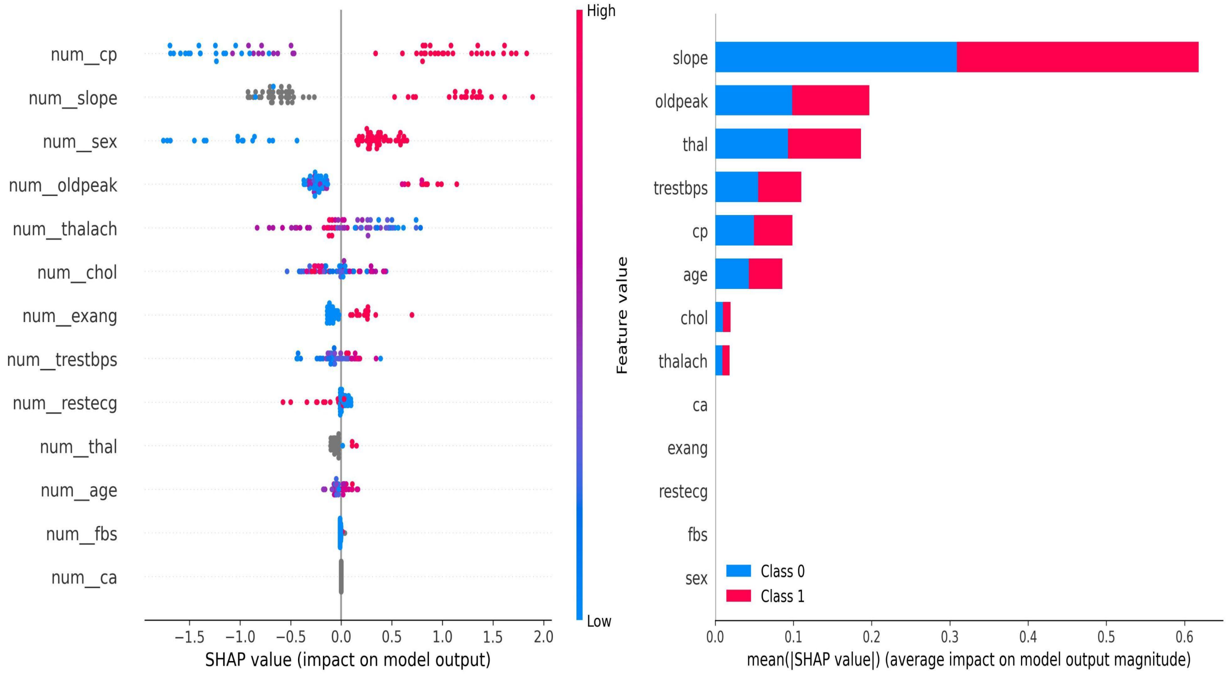


Figure 6. Feature Importance for the machine learning models XGBoost Classifier and RandomForest Classifier

Figure.6. Shows the shap plots for the feature importance of the XGBoost Classifier and RandomForest Classifier models reveal that the features "slope," "cp," and "oldpeak" emerge as the most prominent features in both models. These plots provide valuable insights into the most influential features within the models and the impact of each feature on the model's output.

Based on the results obtained from the analysis, it can be concluded that the Logistic Regression model demonstrates the most favorable performance when applied to the provided dataset. This determination is based on a comprehensive assessment of the model's ability to effectively minimize instances of false positives and false negatives. Iterative imputer help with the performance of the Logistic regression model.

5. OUTLOOK

- Additional Hyperparameter Tuning:

Conducting more extensive hyperparameter tuning, especially for models like XGBoost and Random Forest.

- Advanced Imputation Techniques:

The utilization of advanced imputation techniques beyond iterative imputation, such as K-nearest neighbors imputation, should be considered. These techniques can offer more sophisticated approaches to handling missing data, thereby potentially improving the overall quality of the dataset.

- Model-Specific Improvements:

Experimenting with different tree architectures, such as tree depth and number of estimators.
Explore the use of non-linear kernels for better capturing complex relationships.

6. REFERENCE

1. https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset/data

2. https://rstudio-pubs-static.s3.amazonaws.com/27345_04739111ff384b909a67d8b018f6911c.html#1

3. https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.annotate.html