Heart Disease Prediction With Machine Learning

Introduction

Heart disease is a major health issue that affects millions of people worldwide. In this report, we will describe our approach to predicting the presence of heart disease using machine learning algorithms. We will explain the preprocessing techniques we applied to the dataset, perform analysis and visualizations on the dataset, state the models and hyperparameters used in the modeling step, and mention any other techniques that we used to enhance the results.

Dataset Description

The dataset Contains 270 row each row consists of 16 independent features (id -> smoking_status) and 1 dependent feature (heart disease) - Target column (heart disease) has 2 classes (Yes (Has heart disease), No (Hasn’t heart disease)).

Preprocessing:

- Checking for missing values: The dataset has been checked for missing values using the “isnull().sum()” method, which showed that there are missing values in the dataset.

- Checking for duplicates: The dataset has been checked for duplicates using the duplicated().sum() method, which showed that there are no duplicates in the dataset

- Encoding categorical features: Categorical features such as “Gender”, “Heart Disease”, “smoking_status”, and “work_type” have been encoded using the “LabelEncoder()” method.

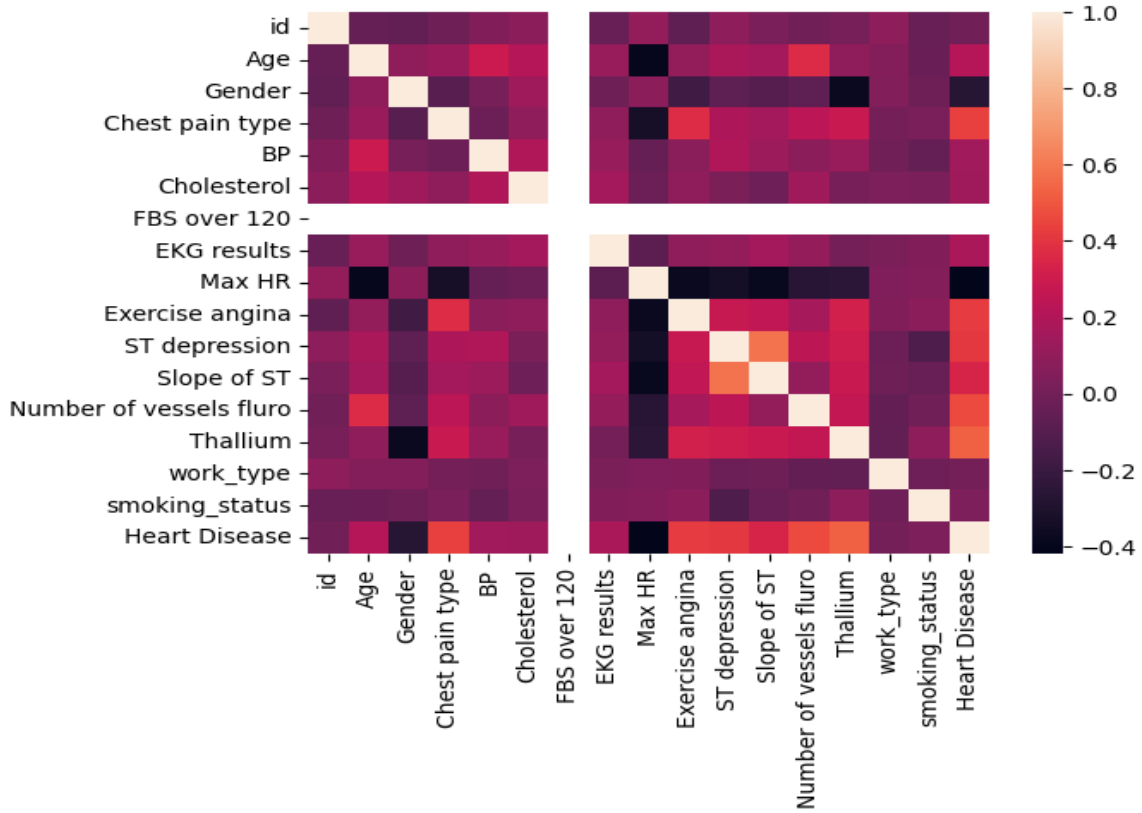
- Imputing missing values: Missing values have been imputed using the KNN imputer from the “fancyimpute” package.

- Outlier Removal: Outliers have been removed using a function that identifies the lower and upper bounds of each feature using the IQR method.

- Standard Scaler: The data has been scaled using “StandardScaler()” to standardize the features.

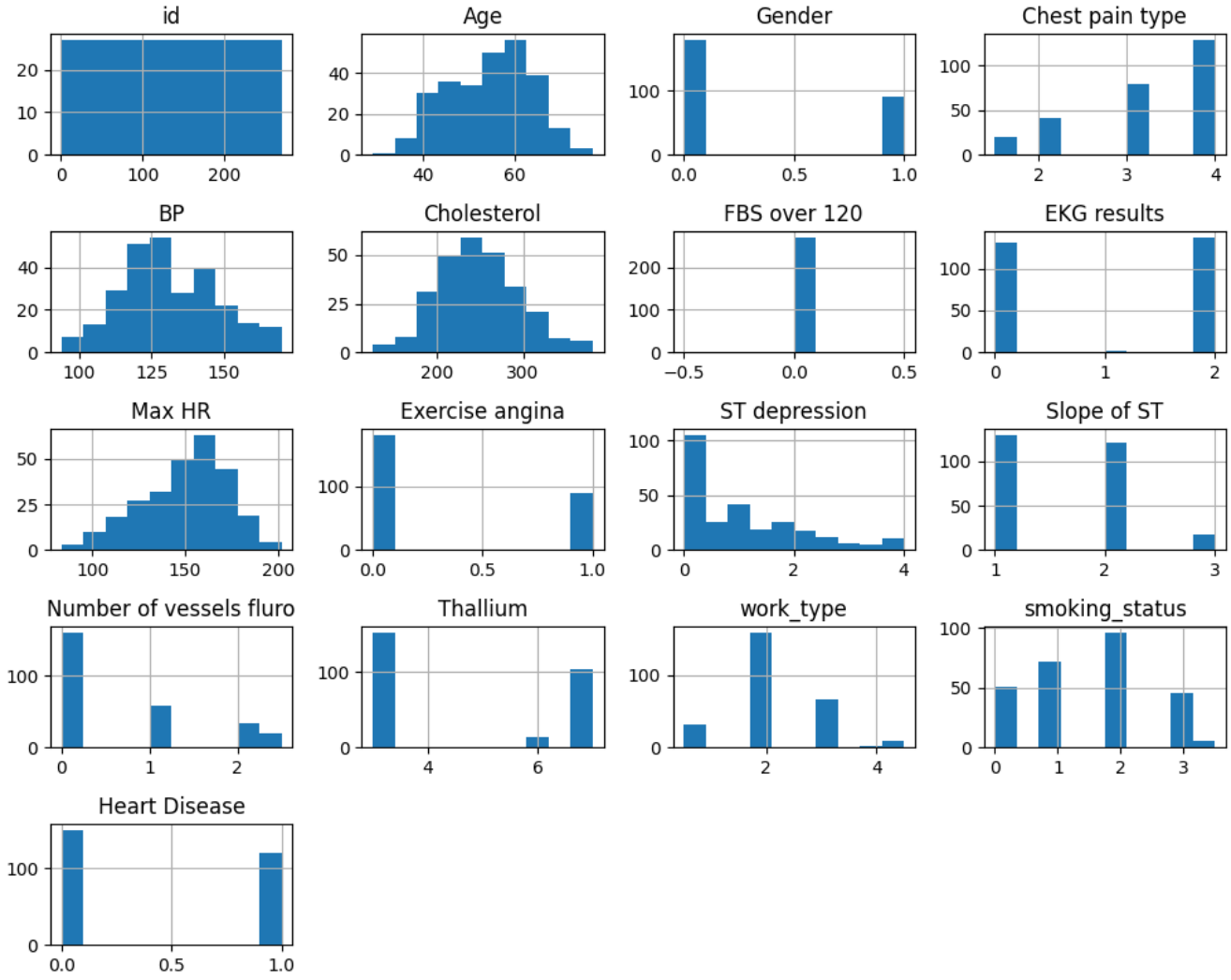
Analysis and Visualizations:

- Correlation matrix: A correlation matrix has been created using the encoded data, which showed that the Age, BP, and Cholesterol features have a strong positive correlation with the presence of heart disease.



There are several features that may have an impact on each other.
For example, Age and Hypertension ”BP” have a positive correlation, which means that as “Age” increases, the likelihood of having hypertension also increases. Similarly,” Smoking_Status” and” Work_Type” have a negative correlation, which means that people who work in private jobs are less likely to smoke than those who work in “self-employed” or “government” jobs.

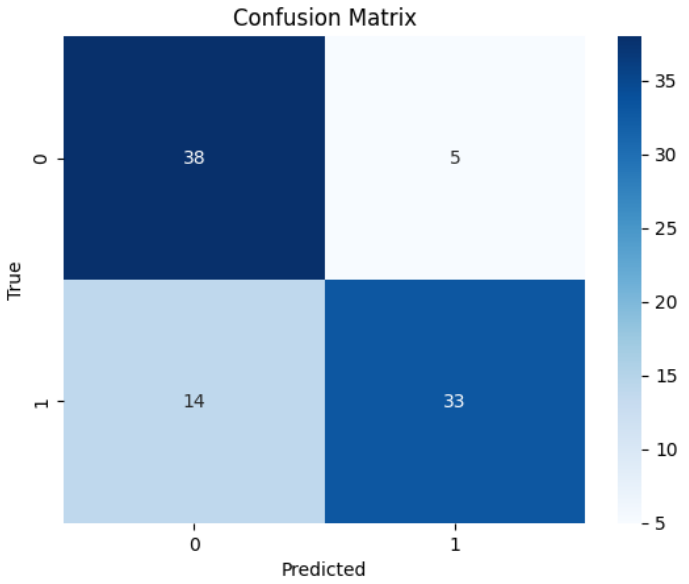
- Histograms: Histograms have been plotted to visualize the distribution of each feature in the dataset.



Models and Hyperparameters:

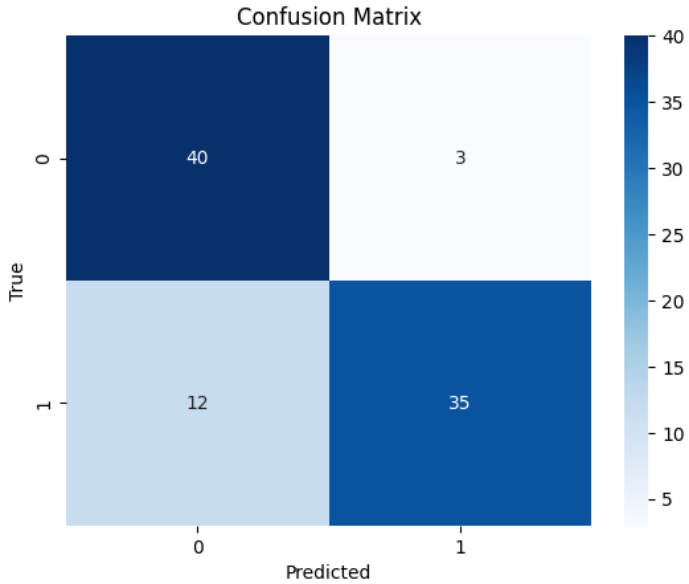
- XGBoost: “xgb.XGBClassifier()” with default hyperparameters has been used.

  The Confusion Matrix of XGBoost:



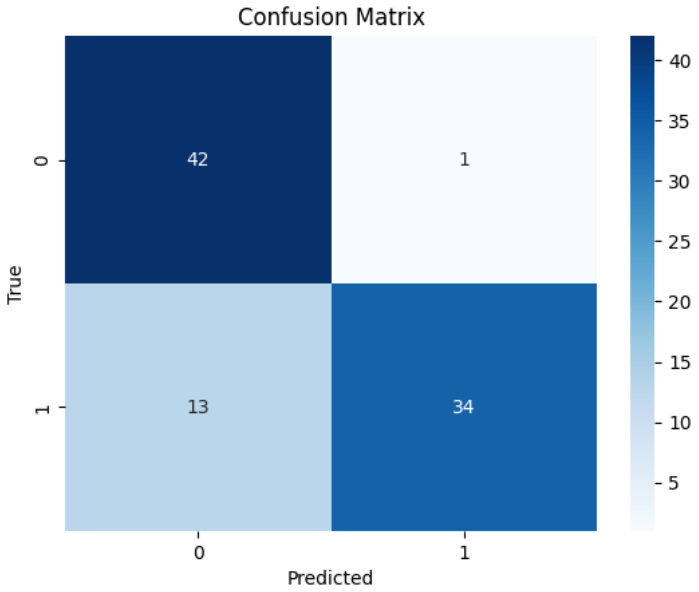
- Random Forest: “RandomForestClassifier(n_estimators=100, random_state=42)” has been used with 100 trees and random state 42.

  The Confusion Matrix of Random Forest:



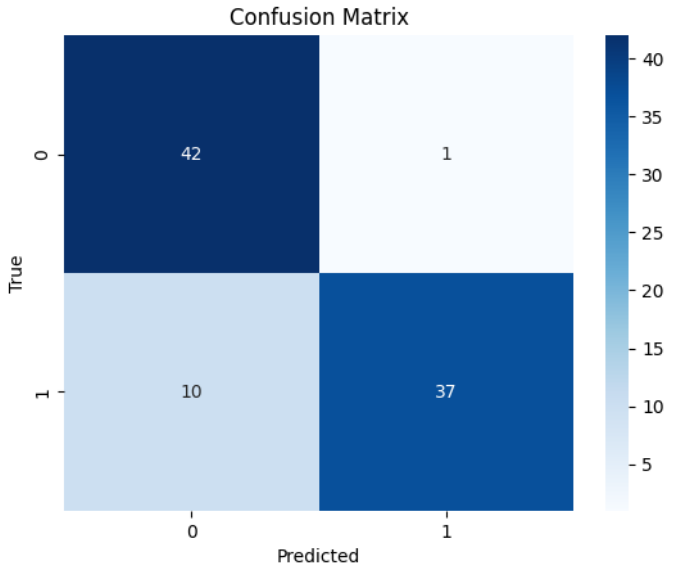
- Logistic Regression: “LogisticRegression(random_state=4)” has been used with a random state of 4.

The Confusion Matrix of Logistic Regression:



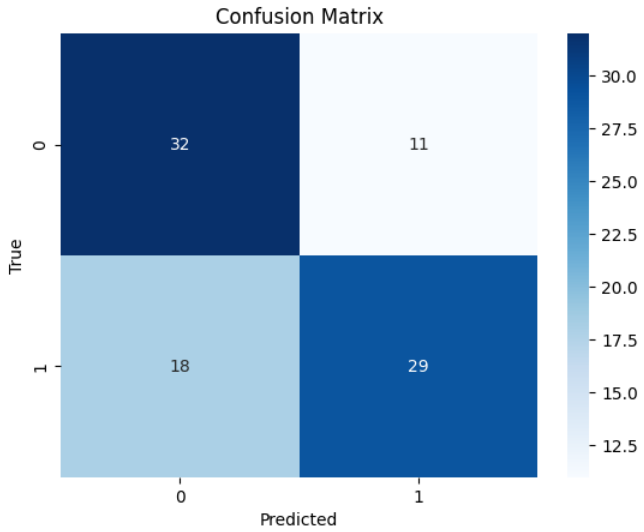
- Support Vector Machine (SVM): “SVC(kernel='linear')” has been used with a linear kernel.

The Confusion Matrix of SVM:



- Decision Tree: “tree.DecisionTreeClassifier()” has been used with default hyperparameters.

The Confusion Matrix of Decision Tree:



Conclusion:

In conclusion, several machine learning models have been trained and evaluated on the heart disease dataset to predict the presence of heart disease. The “SVC” and “LogisticRegression” models performed the best with an accuracy score of 0.87 and 0.84, respectively. The “Thallium”, “Number of vessels fluro”, and “Chest pain type” features have a strong positive correlation with the presence of heart disease. The dataset was preprocessed using various techniques such as encoding categorical features, imputing missing values, and removing outliers. The results of this analysis can be used to predict the presence of heart disease in patients and help in the diagnosis and treatment of heart disease.