Zain Ul Abedin

BS Data Science

Fast NUCES Lahore

+91 310 4906552

Heart Disease Prediction — Model Evaluation Report

Dataset Information

Dataset source: Fede Soriano/heart-failure-prediction (via KaggleHub)
Initial dataset size: 918 rows (after loading)
Missing values: None found in any column
Outlier treatment: Outliers in numeric columns (Age, RestingBP, Cholesterol, MaxHR, Oldpeak) were removed using the Interquartile Range (IQR) method
Categorical features: Sex, ChestPainType, RestingECG, ExerciseAngina, ST_Slope
Encoding method: Label Encoding applied to categorical columns to convert text labels into numeric form
Feature scaling: StandardScaler applied to numeric features to normalize data (mean=0, std=1)

Data Preparation

Features selected: All columns except HeartDisease used as input features
Target variable: HeartDisease (binary classification)
Dataset split: 80% training data, 20% test data with random state=42 for reproducibility

Models and Hyperparameter Tuning

A unified pipeline with data scaling and a placeholder classifier was used
GridSearchCV was applied with 5-fold cross-validation to tune hyperparameters for the following classifiers:

1. Logistic Regression
   - Parameters tuned: C (0.1, 1, 10), solver ('liblinear', 'lbfgs')
2. Random Forest Classifier
   - Parameters tuned: number of trees (n_estimators 100, 200), max depth (None, 10, 20)
3. Support Vector Machine (SVM)

- Parameters tuned: C (0.1, 1, 10), kernel ('linear', 'rbf'), gamma ('scale', 'auto')

4. XGBoost Classifier
   - Parameters tuned: n_estimators (100, 200), max_depth (3, 5), learning_rate (0.01, 0.1)
5. Multi-layer Perceptron (Neural Network)
   - Parameters tuned: hidden layer sizes ((32,), (64,), (32, 32)), activation ('relu', 'tanh'), alpha (0.0001, 0.001), learning rate ('constant', 'adaptive')

Model Performance Summary (Top Results)

Model	Mean CV Accuracy	Best Hyperparameters
XGBoost Classifier	0.89	learning_rate=0.1, max_depth=3, n_estimators=100
Random Forest	0.87	max_depth=10, n_estimators=200
Support Vector Machine	0.86	C=1, kernel='rbf', gamma='scale'
Logistic Regression	0.85	C=1, solver='liblinear'
Neural Network (MLP)	0.84	hidden_layer_sizes=(64,), activation='relu', alpha=0.0001, learning_rate='constant'

Best Model

The best performing model selected by GridSearchCV was the XGBoost Classifier with the following best parameters:

learning_rate = 0.1
max_depth = 3
n_estimators = 100

This model was trained using scaled input features within a pipeline.

Test Set Evaluation

Classification report on the test data:

	precision	recall	f1-score	support
0	0.88	0.90	0.89	67
1	0.90	0.89	0.90	74

accuracy			0.89	141
macro avg	0.89	0.89	0.89	141
weighted avg	0.89	0.89	0.89	141

The overall accuracy of the model on unseen test data is approximately 89%, showing a balanced precision and recall for both classes.

Model Saving

The best model pipeline including preprocessing steps and the XGBoost classifier was saved as a pickle file named 'best_heart_model.pkl' using joblib for later reuse without retraining.

Code Summary

The entire pipeline and model selection process was implemented in Python using libraries such as pandas, scikit-learn, xgboost, and kagglehub for dataset access.
Preprocessing included null value check, outlier removal with IQR, label encoding, feature scaling.
Hyperparameter tuning was performed with GridSearchCV for multiple models in a single unified pipeline.
Evaluation was done on a separate test set with classification metrics reported.
The final best model was saved to disk for deployment or further inference.