This is dataset from UCI for binary classification. This dataset contains 1025 rows and 14 columns.

The Features included are as follows:

|  |  |
| --- | --- |
| **Column** | **Description** |
| age | Age in years |
| sex | Sex (1 = male, 0 = female) |
| cp | Chest pain type (0–3) |
| trestbps | Resting blood pressure (mm Hg) |
| chol | Serum cholesterol (mg/dL) |
| fbs | Fasting blood sugar > 120 mg/dL (1 = true, 0 = false) |
| restecg | Resting ECG results (0–2) |
| thalach | Maximum heart rate achieved |
| exang | Exercise-induced angina (1 = yes, 0 = no) |
| oldpeak | ST depression induced by exercise |
| slope | Slope of the peak exercise ST segment (0–2) |
| ca | Number of major vessels (0–3) |
| thal | Thalassemia (3 = normal, 6 = fixed defect, 7 = reversible defect) |
| target | Presence of heart disease (1 = disease, 0 = no disease) |

Project Workflow:

**Data Loading**

**Exploratory Data Analysis (EDA)**

* Explore presence of Missing Value
* Explore if dataset is balance for Target
* Explore data distribution and Category
* Explore presence of outliers
* Cap the outliers and find Pearson Correlation

**Modeling**

* Train/test split
* Determine outliers based on IQR in train dataset and clip the outliers beyond outliers (Note this only cap the data at the lower or upper limit but does not remove the data).
* Standardization of the data by standard scalar
* Test Random Forest, Gradient Boost, Support Vector Machine and Logistic regression to find base model. All models are trained for Accuracy, F1-Score and AUC.
* Hyperparameter tuning using RandomsearchCV for best performance for the selected model.

**Evaluation**

* Accuracy, Precision, Recall, F1-score, Classification Metrics
* ROC-AUC curves
* Confusion matrix

**Explainability**

* Feature importance by gini purity
* Permutation importance

Result:

The model initially outperformed with accuracy at 98%, F1-Score at 97% and AUC at about 100%. This model was fitted using default scoring metrics for screening.

Afterwards in hyperparameter tuning with 5-fold cross validation the accuracy dropped to 84%.  This may be against the expectation, but in first place model overfits because no cross validation was included, and default fitting is done by gini, with max depth possible for the trees (n\_sample\_split=2, and n\_sample\_leafs=1). This initial model was intentionally done to screen the best model. Under hyperparameter tuning the overfitting issue was resolved.

In test dataset after hyper parameter tuning initially AUC and accuracy were 99% and 84%, respectively. Now the question is why accuracy and AUC are so different?

One of the reasons could be un-balanced sample cohort but that option rules out since in this case both categories are almost close to 50%. The second option could be poor optimization threshold used for calculating accuracy. By default, the model uses 0.5 as threshold, which may not be correct in this case. Thus, I used Youden's J statistic to improve accuracy, so AUC and Accuracy go hands in hand.

To determine the feature importance both feature importance from RandomForest and permutation\_importance were used. Both the Feature\_importance and Permutation\_importnce showed that cp and age are the most contributing factors.

Requirements:

Python 3.11.5

Packages: numpy, pandas, matplotlib.pyplot,seaborn,scipy,sklearn

References:

* UCI Machine Learning Repository
* Kaggle Heart Disease Data

Acknowledgement:

UCI for uploading and providing free access to the data.