AutoML Final Report

AutoML Final Report

Generated: 2025-12-20 10:39:17

1. Dataset Overview

Original Dataset:
• Rows: 1025
• Columns: 14
• Missing Values: 0 (0.00%)
• Duplicate Rows: 723

Column Summary:
• Numeric Columns (14): age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal, target
• Categorical Columns (0):

After Preprocessing:
• Rows: 1025
• Columns: 14
• Rows Removed: 0

2. EDA Findings

Numeric Features Statistics:

3. Data Quality Issues Detected

Issues Summary:
• High Severity: 3
• Medium Severity: 2
• Low Severity: 3

Issues Detected:

• [HIGH] DUPLICATE_ROWS in 'N/A': 723 (70.54%)
• [MEDIUM] OUTLIERS in 'trestbps': 30 (2.93%)

• [MEDIUM] OUTLIERS in 'chol': 16 (1.56%)
• [HIGH] OUTLIERS in 'fbs': 153 (14.93%)
• [LOW] OUTLIERS in 'thalach': 4 (0.39%)
• [LOW] OUTLIERS in 'oldpeak': 7 (0.68%)
• [HIGH] OUTLIERS in 'ca': 87 (8.49%)
• [LOW] OUTLIERS in 'thal': 7 (0.68%)

Recommendations:
• Remove duplicate rows from dataset
• Consider removing or capping outliers using IQR method

4. Preprocessing Decisions

Methods Applied:
• Missing Values Strategy: Median
• Outlier Handling: Remove
• Scaling Method: Standard
• Encoding Method: OneHot
• Test Size: 0.2

Impact Summary:
• Rows Removed: 0
• Missing Values Reduced: 0 → 0
• Features Modified: 14 (from 14)

5. Model Configurations & Hyperparameters

Training Configuration:
• Test Size: 0.2
• Random State: 42
• Hyperparameter Tuning: Yes

Models Trained:
• Logistic Regression
• K-Neighbors Classifier
• Decision Tree Classifier
• Gaussian Naive Bayes
• Random Forest
• Support Vector Machine
• Decision Tree Rule-based

• **Logistic Regression (Tuned)** (Tuned)
Best Params: {'C': 1, 'max_iter': 100, 'penalty': 'l2', 'solver': 'lbfgs'}

• **K-Neighbors Classifier (Tuned)** (Tuned)
Best Params: {'metric': 'manhattan', 'n_neighbors': 5, 'weights': 'distance'}

• **Decision Tree Classifier (Tuned)** (Tuned)
Best Params: {'criterion': 'gini', 'max_depth': 10, 'min_samples_leaf': 1, 'min_samples_split': 2}

• **Gaussian Naive Bayes (Tuned)** (Tuned)
Best Params: {'var_smoothing': 0.012915496650148827}

• **Random Forest (Tuned)** (Tuned)
Best Params: {'criterion': 'gini', 'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 50}

• **Support Vector Machine (Tuned)** (Tuned)
Best Params: {'C': 100, 'gamma': 'scale', 'kernel': 'rbf'}

• **Decision Tree Rule-based (Tuned)** (Tuned)
Best Params: {'criterion': 'gini', 'max_depth': 10, 'min_samples_leaf': 1, 'min_samples_split': 2}

6. Model Performance Comparison

7. Best Model Summary & Justification

Selected Model: K-Neighbors Classifier (Tuned)

Reason: Best F1 score: 1.0000

Performance Metrics:
• Accuracy: 1.0
• Precision: 1.0
• Recall: 1.0
• F1-Score: 1.0
• ROC-AUC: 1.0
• Training Time: 1.5126826763153076s

Hyperparameters:
• metric: manhattan
• n_neighbors: 5
• weights: distance