AutoML Final Report

AutoML Final Report

Generated: 2025-12-24 22:02:44

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: 302
• Columns: 14
• Rows Removed: 723

2. EDA Findings

Numeric Features Statistics:

Feature	Mean	Median	Std	Min	Max
age	54.43	56.00	9.07	29.00	77.00
sex	0.70	1.00	0.46	0.00	1.00
cp	0.94	1.00	1.03	0.00	3.00
trestbps	131.61	130.00	17.52	94.00	200.00
chol	246.00	240.00	51.59	126.00	564.00
fbs	0.15	0.00	0.36	0.00	1.00

restecg	0.53	1.00	0.53	0.00	2.00
thalach	149.11	152.00	23.01	71.00	202.00
exang	0.34	0.00	0.47	0.00	1.00
oldpeak	1.07	0.80	1.18	0.00	6.20
slope	1.39	1.00	0.62	0.00	2.00
ca	0.75	0.00	1.03	0.00	4.00
thal	2.32	2.00	0.62	0.00	3.00
target	0.51	1.00	0.50	0.00	1.00

3. Data Quality Issues Detected

Issues Summary:

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

Issues Detected:

• [HIGH] DUPLICATE_ROWS in 'Dataset': 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%)

4. Preprocessing Decisions

Methods Applied:

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

Impact Summary:

• Rows Removed: 723
• 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: No

Models Trained: 7 total

Non-Tuned Models (7):

• Logistic Regression (default parameters)
• K-Neighbors Classifier (default parameters)
• Decision Tree Classifier (default parameters)
• Gaussian Naive Bayes (default parameters)
• Random Forest (default parameters)
• Support Vector Machine (default parameters)
• Decision Tree Rule-based (default parameters)

6. Model Performance Comparison

Model	Accuracy	Precision	Recall	F1-Score	Training Time
Logistic Regression	0.7705	0.7931	0.7705	0.7680	0.08s
K-Neighbors Classifier	0.7377	0.7497	0.7377	0.7364	0.00s
Decision Tree Classifier	0.7213	0.7221	0.7213	0.7200	0.01s
Gaussian Naive Bayes	0.8525	0.8531	0.8525	0.8525	0.01s
Random Forest	0.8361	0.8429	0.8361	0.8359	0.31s
Support Vector Machine	0.7541	0.7578	0.7541	0.7541	0.03s
Decision Tree Rule-based	0.7377	0.7379	0.7377	0.7370	0.01s

7. Best Model Summary & Justification

Selected Model: Gaussian Naive Bayes

Reason: Best F1 score: 0.8525

Performance Metrics:

• Accuracy: 0.8524590163934426
• Precision: 0.8530759739115107
• Recall: 0.8524590163934426
• F1-Score: 0.8525384035247887
• ROC-AUC: 0.8556034482758621
• Training Time: 0.005033254623413086s