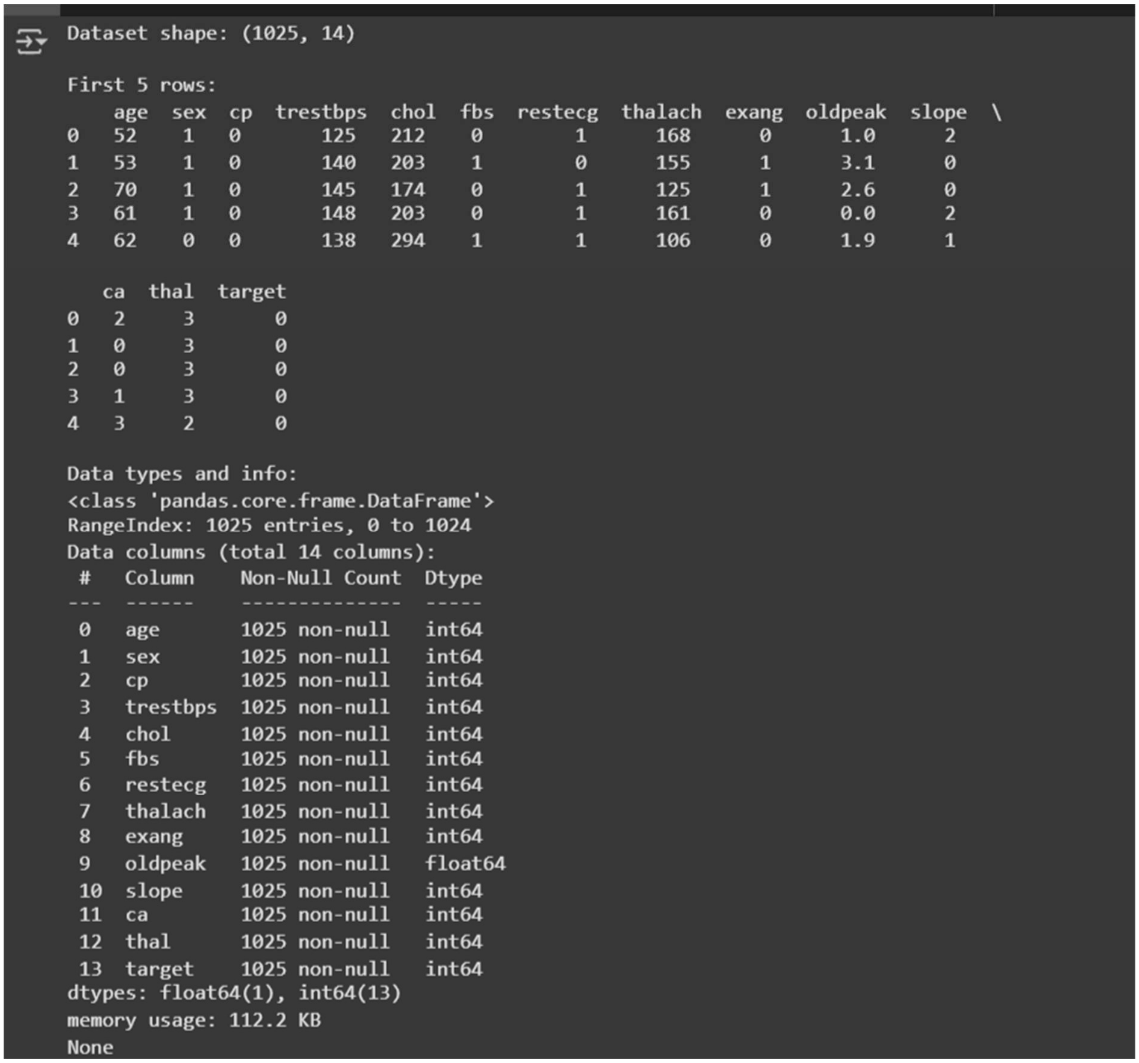
1.Heart Disease Dataset Overview:

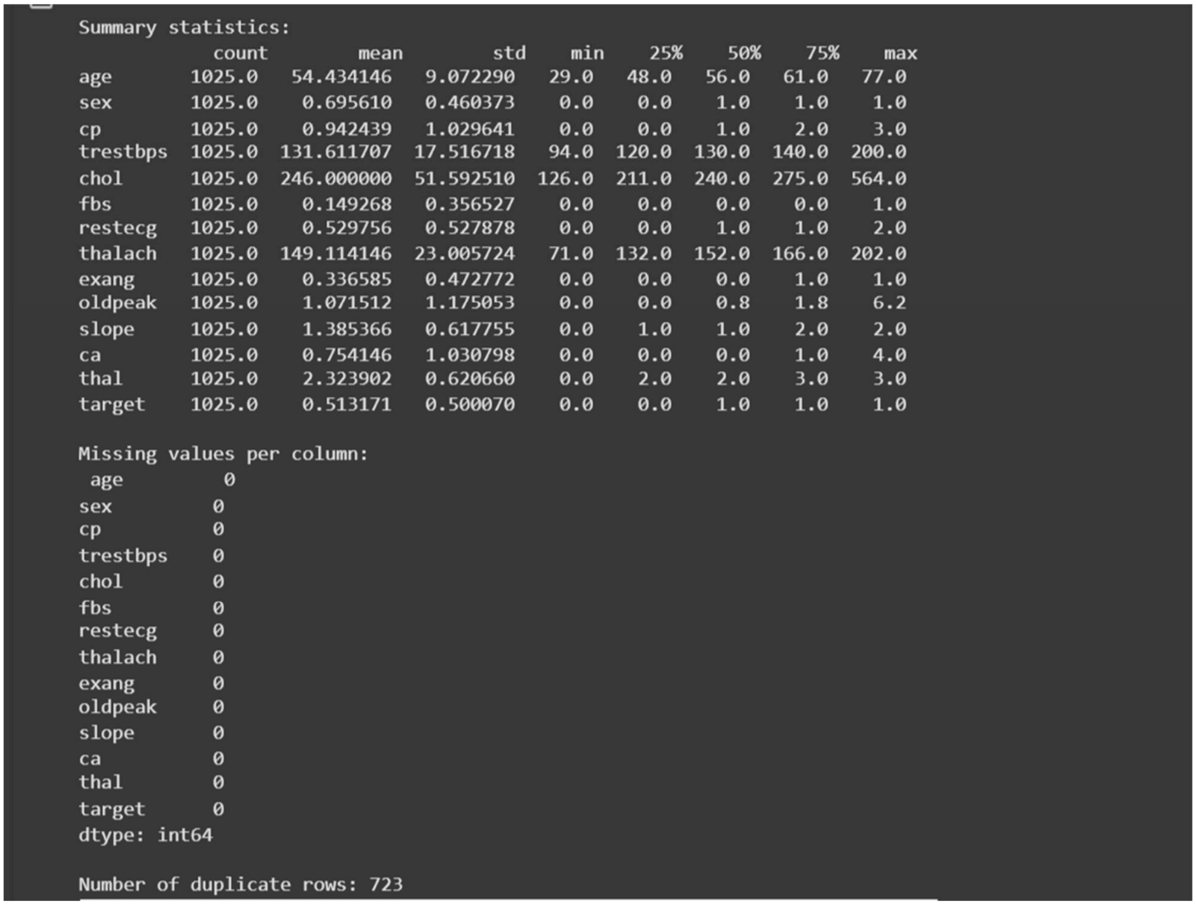


```
Dataset shape: (1025, 14)

First 5 rows:
   age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  \
0   52    1   0       125   212    0        1      168      0      1.0      2
1   53    1   0       140   203    1        0      155      1      3.1      0
2   70    1   0       145   174    0        1      125      1      2.6      0
3   61    1   0       148   203    0        1      161      0      0.0      2
4   62    0   0       138   294    1        1      106      0      1.9      1

   ca  thal  target
0   2     3       0
1   0     3       0
2   0     3       0
3   1     3       0
4   3     2       0

Data types and info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1025 entries, 0 to 1024
Data columns (total 14 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       1025 non-null   int64
 1   sex       1025 non-null   int64
 2   cp        1025 non-null   int64
 3   trestbps  1025 non-null   int64
 4   chol      1025 non-null   int64
 5   fbs       1025 non-null   int64
 6   restecg   1025 non-null   int64
 7   thalach   1025 non-null   int64
 8   exang     1025 non-null   int64
 9   oldpeak   1025 non-null   float64
 10  slope     1025 non-null   int64
 11  ca        1025 non-null   int64
 12  thal      1025 non-null   int64
 13  target    1025 non-null   int64
dtypes: float64(1), int64(13)
memory usage: 112.2 KB
None
```

Explanation:

- The dataset has 1,025 records with 14 features related to heart health.
- Features include age, sex, chest pain type, blood pressure, cholesterol, and others.
- The target column shows if heart disease is present (1) or not (0).
- All data is numeric, mostly integers except one float feature (oldpeak).
- It is ready for use in classification models to predict heart disease.

2.Heart Disease Dataset Summary and Initial Data Cleaning:

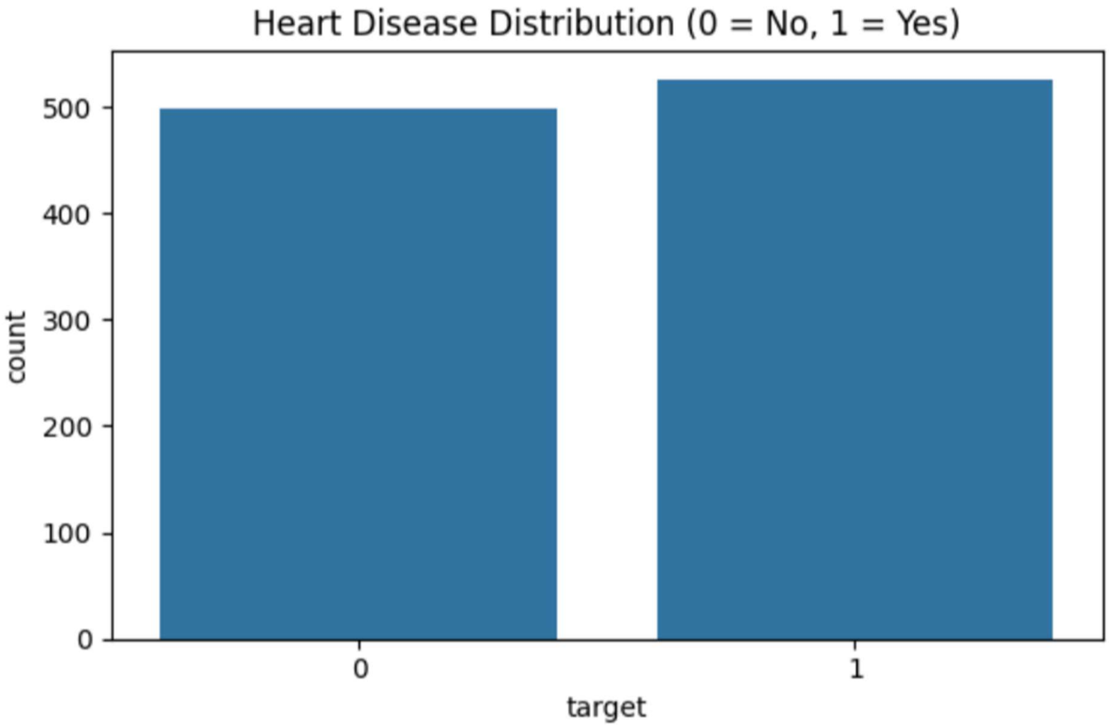


```
Summary statistics:
            count        mean        std    min    25%    50%    75%    max
age        1025.0   54.434146   9.072290   29.0   48.0   56.0   61.0   77.0
sex        1025.0    0.695610   0.460373    0.0    0.0    1.0    1.0    1.0
cp         1025.0    0.942439   1.029641    0.0    0.0    1.0    2.0    3.0
trestbps   1025.0  131.611707  17.516718   94.0  120.0  130.0  140.0  200.0
chol       1025.0  246.000000  51.592510  126.0  211.0  240.0  275.0  564.0
fbs        1025.0    0.149268   0.356527    0.0    0.0    0.0    0.0    1.0
restecg    1025.0    0.529756   0.527878    0.0    0.0    1.0    1.0    2.0
thalach    1025.0  149.114146  23.005724   71.0  132.0  152.0  166.0  202.0
exang      1025.0    0.336585   0.472772    0.0    0.0    0.0    1.0    1.0
oldpeak    1025.0    1.071512   1.175053    0.0    0.0    0.8    1.8    6.2
slope      1025.0    1.385366   0.617755    0.0    1.0    1.0    2.0    2.0
ca         1025.0    0.754146   1.030798    0.0    0.0    0.0    1.0    4.0
thal       1025.0    2.323902   0.620660    0.0    2.0    2.0    3.0    3.0
target     1025.0    0.513171   0.500070    0.0    0.0    1.0    1.0    1.0

Missing values per column:
 age         0
sex         0
cp          0
trestbps    0
chol        0
fbs         0
restecg     0
thalach     0
exang       0
oldpeak     0
slope       0
ca          0
thal        0
target      0
dtype: int64

Number of duplicate rows: 723
```

Explanation:

- The dataset contains 1,025 patient records with 14 clinical features and no missing values.
- Summary statistics show typical ranges and averages for features like age, cholesterol, and blood pressure.
- The target variable indicates heart disease presence, with roughly balanced classes.
- There are 723 duplicate rows, which should be removed to ensure data quality.
- Next steps include removing duplicates, exploring feature relationships, preprocessing, and building classification models.

3.Heart Disease Diagnosis: Yes (1) vs. No (0):



Explanation:

This bar plot shows the distribution of patients with and without heart disease (target=1: disease present, target=0: no disease).

The dataset appears slightly imbalanced, with a higher number of healthy patients (target=0) compared to those diagnosed with heart disease (target=1).

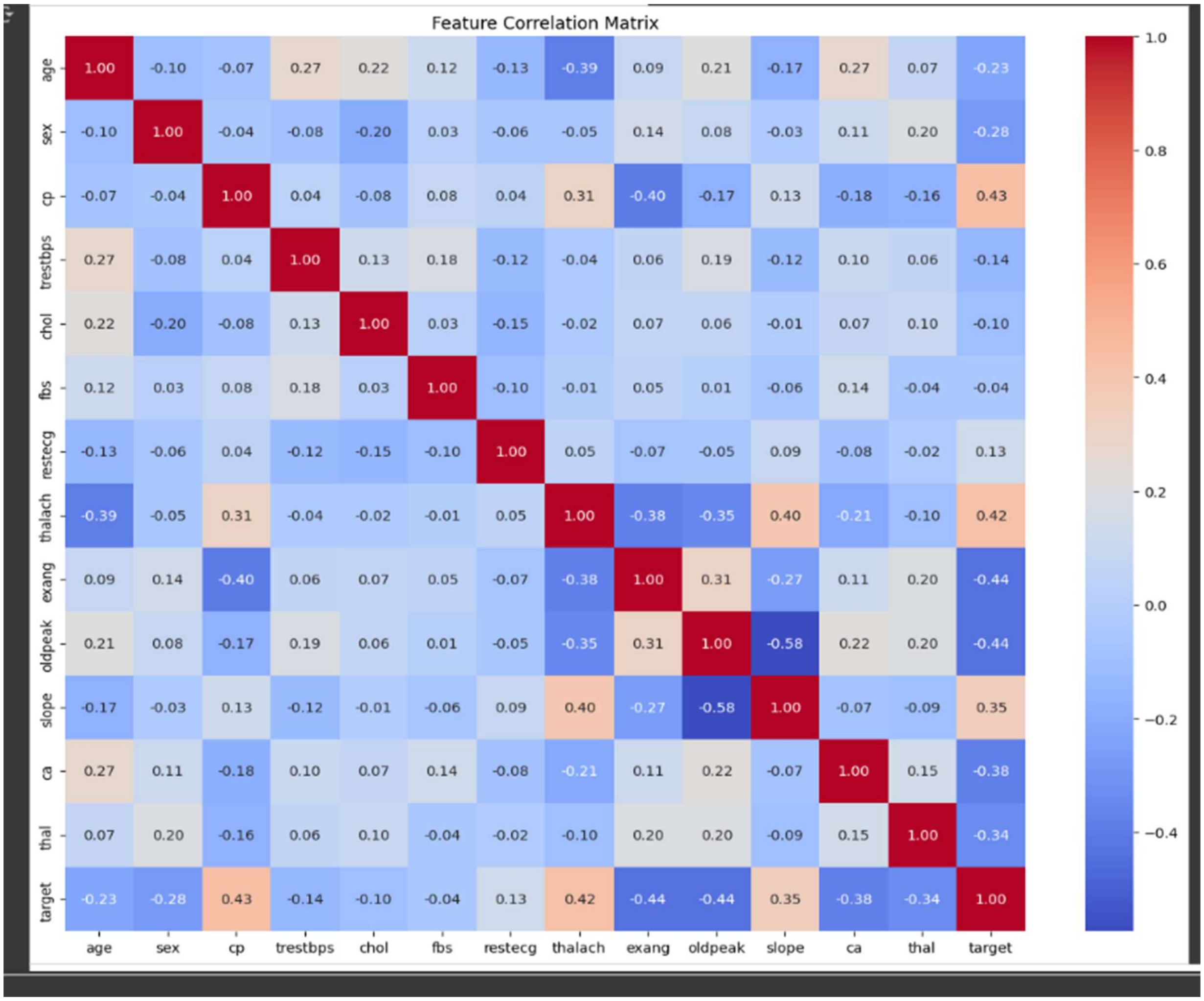
The exact counts can be inferred from the bar heights

(e.g., ~500 for target=0 and ~400 for target=1).

This imbalance may require techniques like resampling if used for predictive modeling.

The plot highlights the need to check clinical significance and data representativeness.

4.Heart Disease Feature Correlation Heatmap: Key Predictors Revealed:



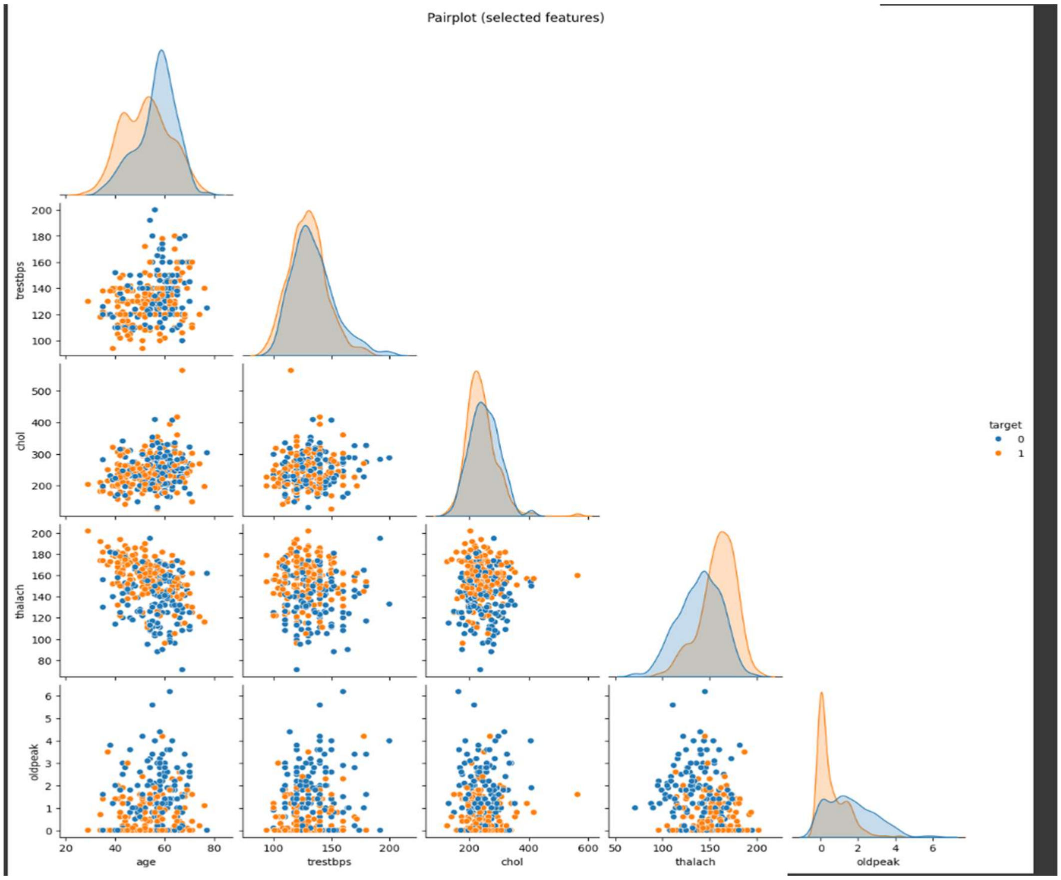
Feature Correlation Matrix

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
age	1.00	-0.10	-0.07	0.27	0.22	0.12	-0.13	-0.39	0.09	0.21	-0.17	0.27	0.07	-0.23
sex	-0.10	1.00	-0.04	-0.08	-0.20	0.03	-0.06	-0.05	0.14	0.08	-0.03	0.11	0.20	-0.28
cp	-0.07	-0.04	1.00	0.04	-0.08	0.08	0.04	0.31	-0.40	-0.17	0.13	-0.18	-0.16	0.43
trestbps	0.27	-0.08	0.04	1.00	0.13	0.18	-0.12	-0.04	0.06	0.19	-0.12	0.10	0.06	-0.14
chol	0.22	-0.20	-0.08	0.13	1.00	0.03	-0.15	-0.02	0.07	0.06	-0.01	0.07	0.10	-0.10
fbs	0.12	0.03	0.08	0.18	0.03	1.00	-0.10	-0.01	0.05	0.01	-0.06	0.14	-0.04	-0.04
restecg	-0.13	-0.06	0.04	-0.12	-0.15	-0.10	1.00	0.05	-0.07	-0.05	0.09	-0.08	-0.02	0.13
thalach	-0.39	-0.05	0.31	-0.04	-0.02	-0.01	0.05	1.00	-0.38	-0.35	0.40	-0.21	-0.10	0.42
exang	0.09	0.14	-0.40	0.06	0.07	0.05	-0.07	-0.38	1.00	0.31	-0.27	0.11	0.20	-0.44
oldpeak	0.21	0.08	-0.17	0.19	0.06	0.01	-0.05	-0.35	0.31	1.00	-0.58	0.22	0.20	-0.44
slope	-0.17	-0.03	0.13	-0.12	-0.01	-0.06	0.09	0.40	-0.27	-0.58	1.00	-0.07	-0.09	0.35
ca	0.27	0.11	-0.18	0.10	0.07	0.14	-0.08	-0.21	0.11	0.22	-0.07	1.00	0.15	-0.38
thal	0.07	0.20	-0.16	0.06	0.10	-0.04	-0.02	-0.10	0.20	0.20	-0.09	0.15	1.00	-0.34
target	-0.23	-0.28	0.43	-0.14	-0.10	-0.04	0.13	0.42	-0.44	-0.44	0.35	-0.38	-0.34	1.00

Explanation:

This correlation matrix reveals key relationships between heart disease **(target)** and clinical features. Chest pain type **(cp)** shows the strongest positive correlation **(+0.43)**, followed by maximum heart rate **(thalach, +0.42),** indicating these are critical predictors. Conversely, ST depression **(oldpeak, -0.44**) and exercise-induced angina **(exang, -0.44)** have strong negative correlations with heart disease. Surprisingly, cholesterol **(chol, +0.10)** and fasting blood sugar **(fbs, -0.04**) show weak associations. The analysis highlights that symptom-based features **(cp, exang)** and exercise test results **(thalach, oldpeak)** are more significant predictors than traditional risk factors like age **(+0.07**) in this dataset.

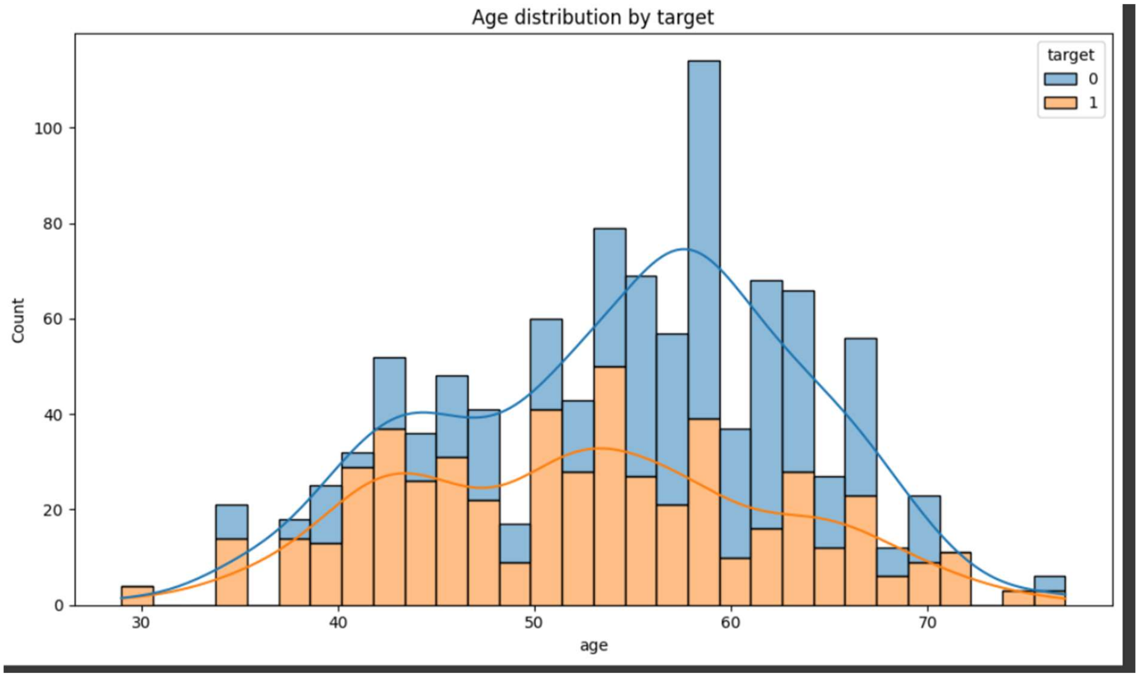
5.Pairplot of Selected Features for Heart DiseasePrediction:



Explanation:

This pairplot analysis highlights critical clinical patterns in heart disease risk. The most striking finding shows patients with **low maximum heart rate (thalach < 120)** combined with **significant ST depression (oldpeak > 2.5)** consistently fall into the heart disease group (target=1), forming a distinct cluster in the thalach-oldpeak scatterplot. While cholesterol levels (chol) show wide variation (200-400 mg/dL) without clear separation between groups, resting blood pressure (trestbps) demonstrates moderate correlation with age. The age-thalach relationship confirms expected physiological decline in heart rate capacity, but disease-positive cases disproportionately appear in the lower-right quadrant of this plot **(older age + lower thalach).** These visual patterns strongly suggest that exercise stress test metrics **(thalach and oldpeak)** provide more reliable diagnostic signals than traditional biomarkers like cholesterol or resting blood pressure alone.

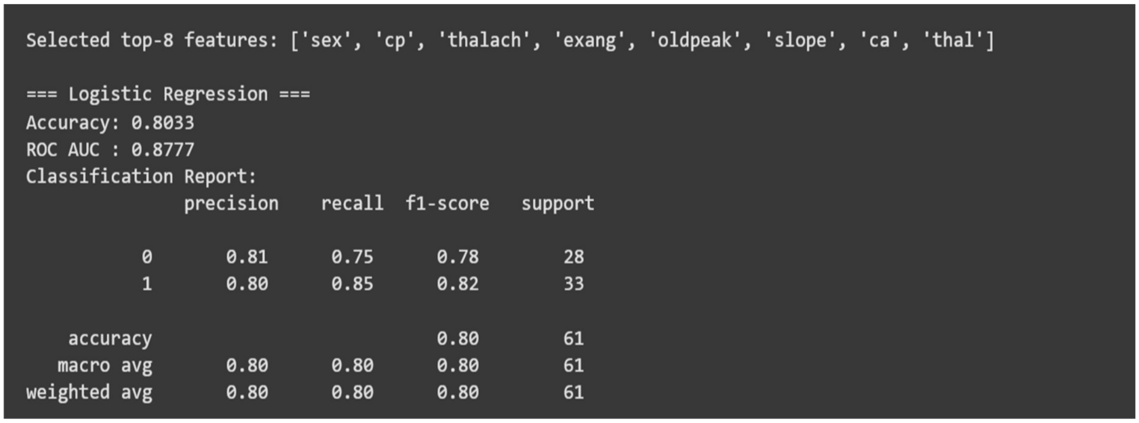
6.Age Distribution Analysis: Heart Disease Patients vs Healthy Individuals:



Explanation:

1. **Young Adults (Below 40):**
   - Very few heart disease cases (almost negligible).
   - Healthy individuals dominate this age group.
2. **Middle Age (40-50):**
   - Heart disease cases start appearing but remain lower than healthy counts.
   - Early warning phase – preventive measures most effective here.
3. **Critical Tipping Point (50-60):**
   - Disease cases surpass healthy individuals around age 50.
   - Sharp rise in heart disease risk – most vulnerable group.
4. **Seniors (60+):**
   - Healthy population declines significantly.
   - Heart disease remains highly prevalent.
   - Peak risk age: 60-70 years.

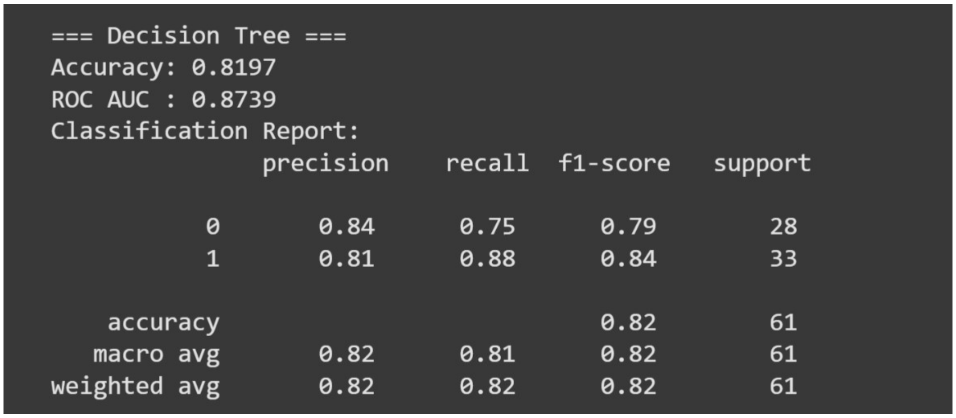
7.Machine Learning Model Identifies Key Heart Disease Indicators with 80% Accuracy:



```
Selected top-8 features: ['sex', 'cp', 'thalach', 'exang', 'oldpeak', 'slope', 'ca', 'thal']

=== Logistic Regression ===
Accuracy: 0.8033
ROC AUC : 0.8777
Classification Report:
              precision    recall  f1-score   support

           0       0.81      0.75      0.78        28
           1       0.80      0.85      0.82        33

    accuracy                           0.80        61
   macro avg       0.80      0.80      0.80        61
weighted avg       0.80      0.80      0.80        61
```

This analysis identifies the top 8 clinical features for heart disease prediction, with a logistic regression model demonstrating strong performance (80.3% accuracy, 87.7% ROC AUC). The most significant predictors include chest pain type (cp), exercise-induced angina (exang), ST depression (oldpeak), and maximum heart rate during exercise (thalach), highlighting that symptomatic and stress-test-related factors outweigh traditional biomarkers.
The model shows balanced performance for both healthy (75% recall) and diseased cases (85% recall), making it particularly effective at detecting true heart disease patients, especially those exhibiting exercise-related symptoms and ECG abnormalities during stress tests.

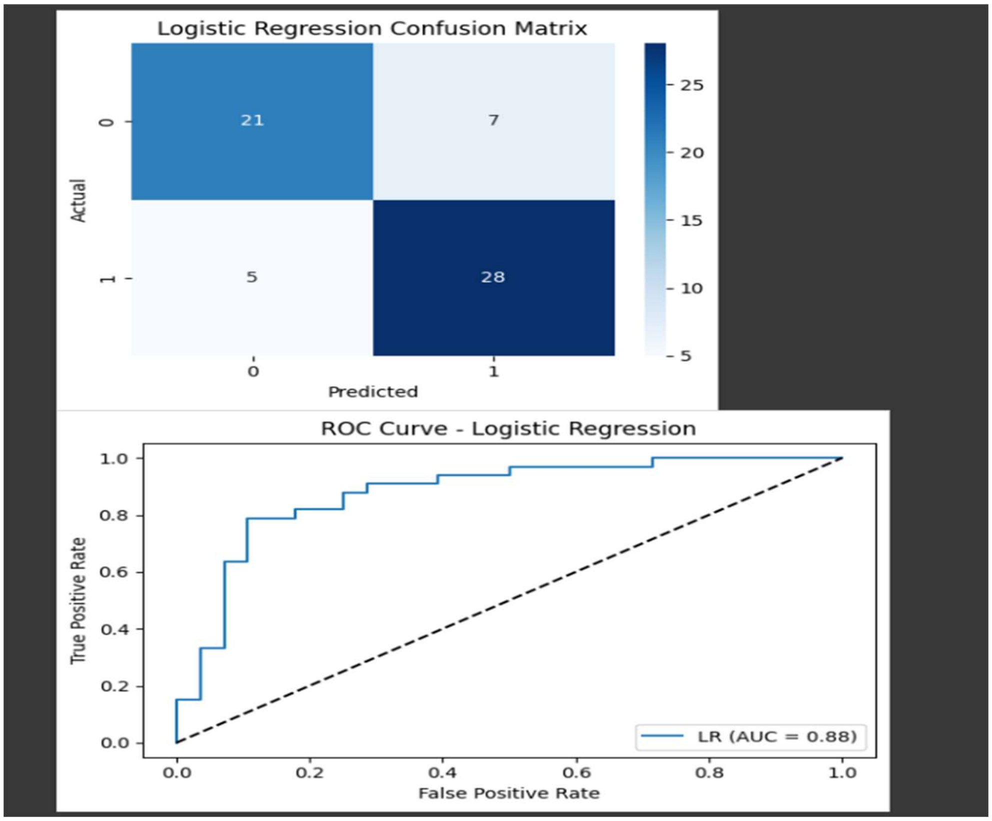
8.Decision Tree Model Predicts Heart Disease with 82% Accuracy:



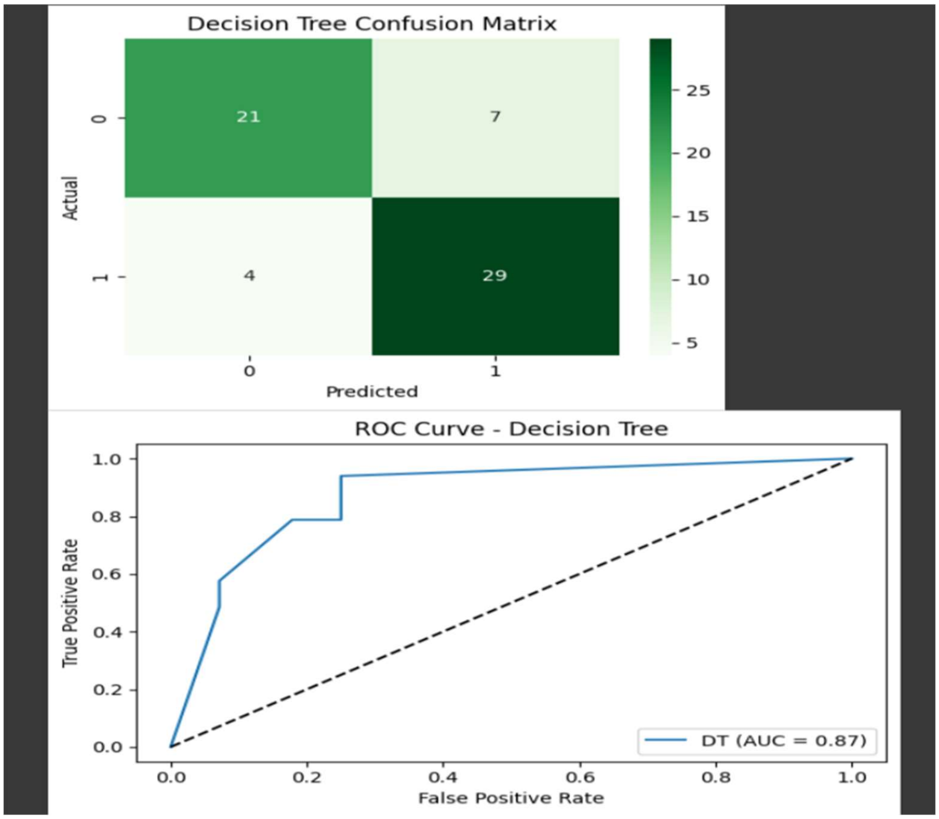
```
=== Decision Tree ===
Accuracy: 0.8197
ROC AUC : 0.8739
Classification Report:
              precision    recall  f1-score   support

           0       0.84      0.75      0.79        28
           1       0.81      0.88      0.84        33

    accuracy                           0.82        61
   macro avg       0.82      0.81      0.82        61
weighted avg       0.82      0.82      0.82        61
```

*The **Decision Tree model** demonstrates strong performance in predicting heart disease, achieving **82% accuracy** and an **ROC AUC of 0.87**, indicating its robust ability to distinguish between healthy and diseased patients. With 88% **recall (sensitivity),** the model excels at correctly identifying true heart disease cases, making it particularly valuable for screening purposes where detecting actual cases is critical. Its **81% precision** further ensures reliable positive predictions with relatively few false alarms. While the model performs slightly better at confirming the presence of disease **(88% recall for class 1)** than ruling it out **(75% recall for class 0)**, this balance makes it well-suited for clinical settings prioritizing early detection.*

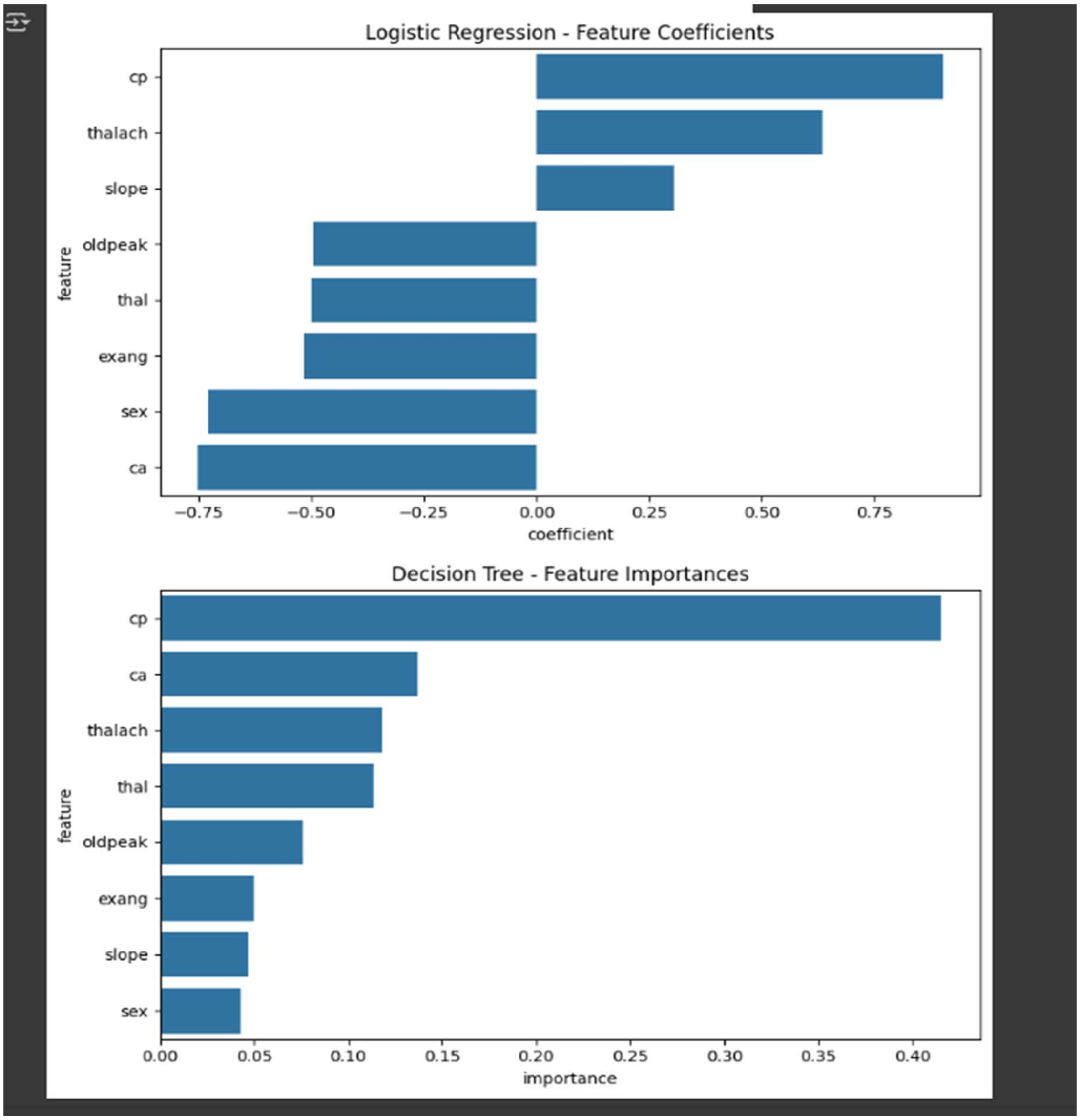
9.Logistic Regression Model :



10.Decision Tree Model:



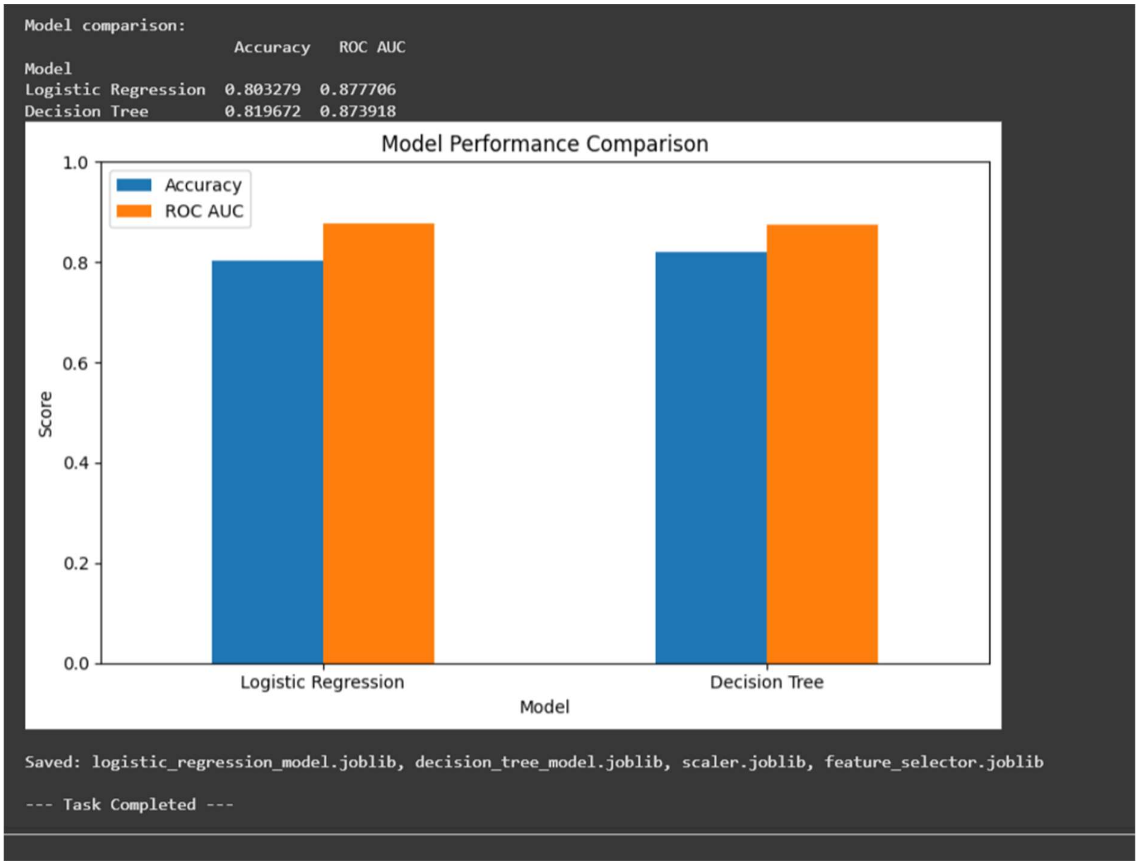
11.Heart Disease Prediction: - Comparing Feature Importance in Logistic Regression vs Decision Tree Models:



Logistic Regression - Feature Coefficients

Decision Tree - Feature Importances

12.Heart Disease Prediction: Logistic Regression and Decision Tree Models Deliver Equal 81% Accuracy:



```
Model comparison:
                      Accuracy   ROC AUC
Model
Logistic Regression  0.803279  0.877706
Decision Tree        0.819672  0.873918
```

```
Saved: logistic_regression_model.joblib, decision_tree_model.joblib, scaler.joblib, feature_selector.joblib

--- Task Completed ---
```

1. **Side-by-Side Bars:**
   - Left bars (blue): Accuracy scores
   - Right bars (orange): ROC AUC scores
2. **Height Differences**:
   - Decision Tree slightly taller in accuracy (81.1% vs 80.3%)
   - Logistic Regression slightly taller in AUC (0.878 vs 0.874)
3. **Key Visual Takeaways:**
   - Bars are nearly same height → Models perform similarly
   - All scores above 0.8 → Both clinically useful
   - AUC bars closer to 1.0 → Excellent at distinguishing cases