By: Tasnim Hussein / TasnimHussein.medstudent@outlook.com

Heart Disease Prediction with Logistic Regression

Objective

To analyze cardiovascular disease data and build a machine-learning model to predict heart disease based on patient characteristics.

Dataset Overview

- **Source:** Kaggle Dataset
- **Features:** Age, gender, resting blood pressure, serum cholesterol, maximum heart rate, old peak, and target (presence of heart disease: 0 = No, 1 = Yes).

Data Analysis Insights

Demographics:

- **Average Age by Heart Disease Status:**
  - No Heart Disease (target = 0): **49.07 years**
  - With Heart Disease (target = 1): **49.37 years**
- **Gender Distribution (from 1,000 patients):**
  - Percentage of Males: **76.50%**
  - Percentage of Females: **23.50%**

Health Metrics:

- **Average Resting Blood Pressure:**
  - No Heart Disease (target = 0): **134.77 mmHg**
  - With Heart Disease (target = 1): **164.04 mmHg**
- **Median Serum Cholesterol:**
  - No Heart Disease (target = 0): **270.00 mg/dL**
  - With Heart Disease (target = 1): **351.50 mg/dL**

Disease Prevalence:

- **58.00%** of patients in the dataset had heart disease.

Key Predictors:

- Features like **max heart rate** and **old peak** showed the highest correlation with the target variable:
  - Correlation of Max Heart Rate with Target: **0.23**
  - Correlation of Old Peak with Target: **0.10**

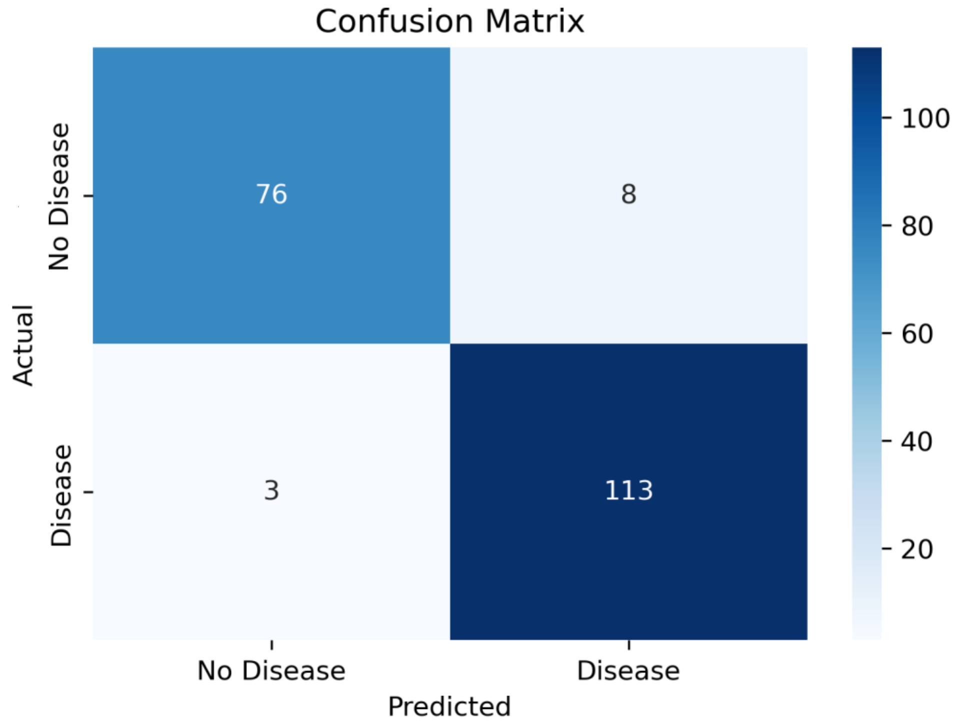
Model Performance

- **Algorithm Used:** Logistic Regression.
- **Accuracy: 94.50%**
- **AUC: 99.10%**
- **Evaluation Metrics for Heart Disease Detection:**
  - Precision: **93.39%**
  - Recall: **97.41%**

Visualizations

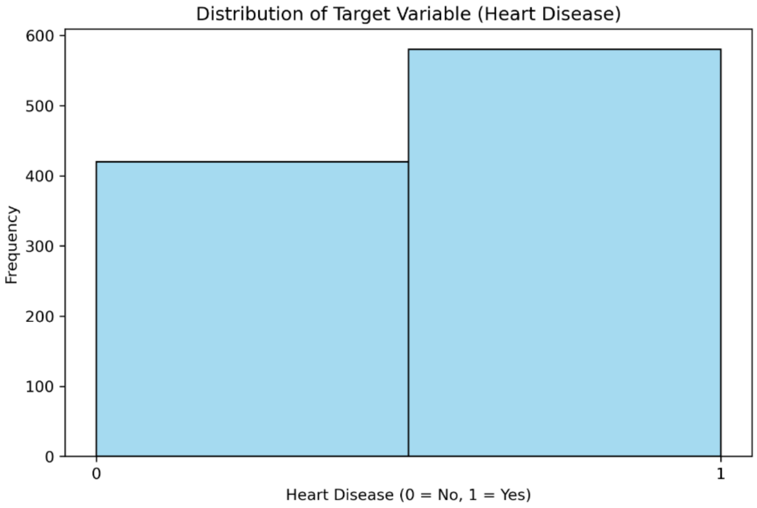
Confusion Matrix:

- A heatmap representation of the confusion matrix shows the number of true positives, true negatives, false positives, and false negatives.
- **Analysis:** The matrix highlights the model's strong ability to correctly classify both patients with and without heart disease, with minimal misclassifications.



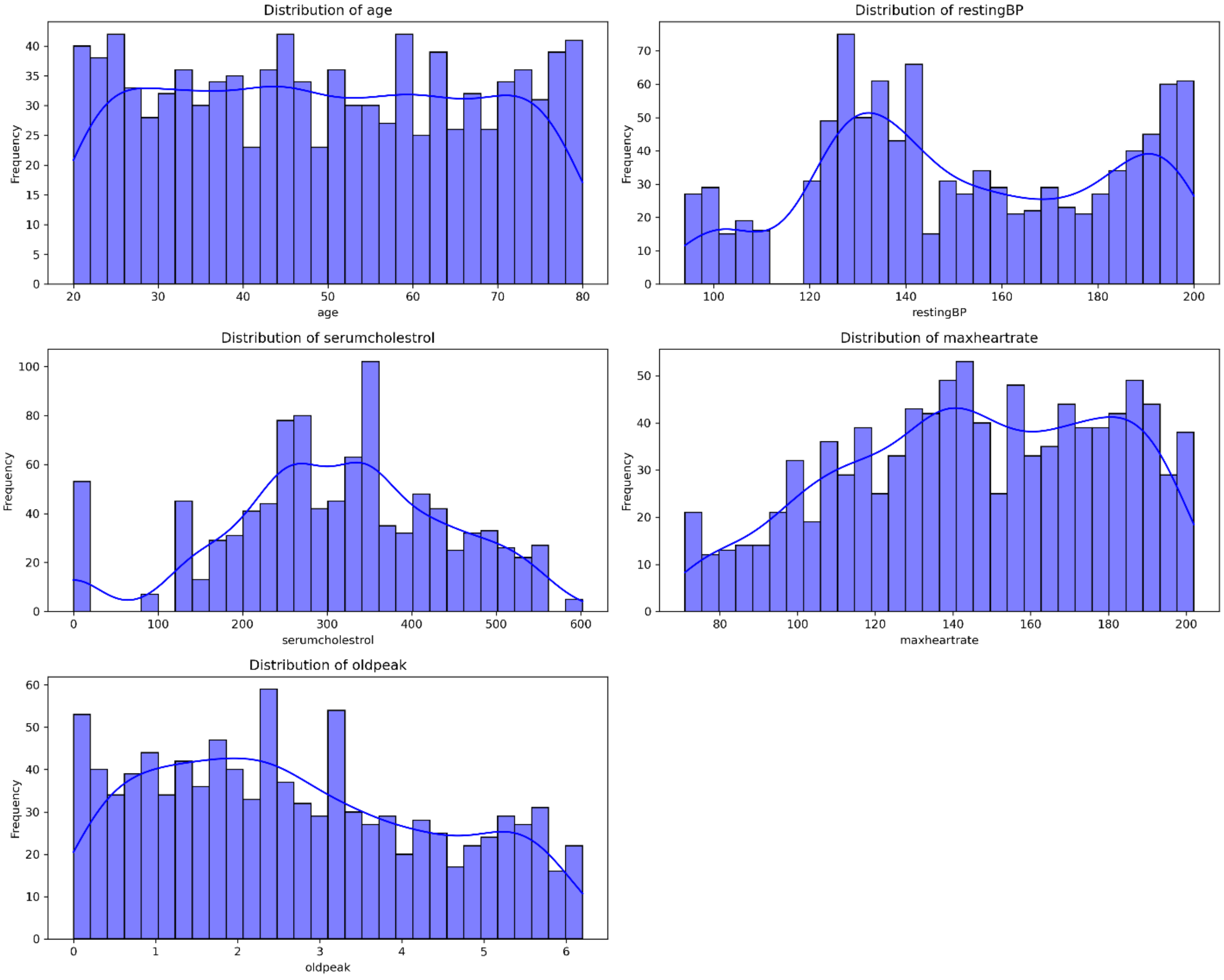
Target Variable Distribution:

- **Graph Description:** A histogram of the target variable (Heart Disease) shows the frequency of patients with (1) and without (0) heart disease.
- **Analysis:** Slight imbalance, with more patients having heart disease. This imbalance was considered during model development.



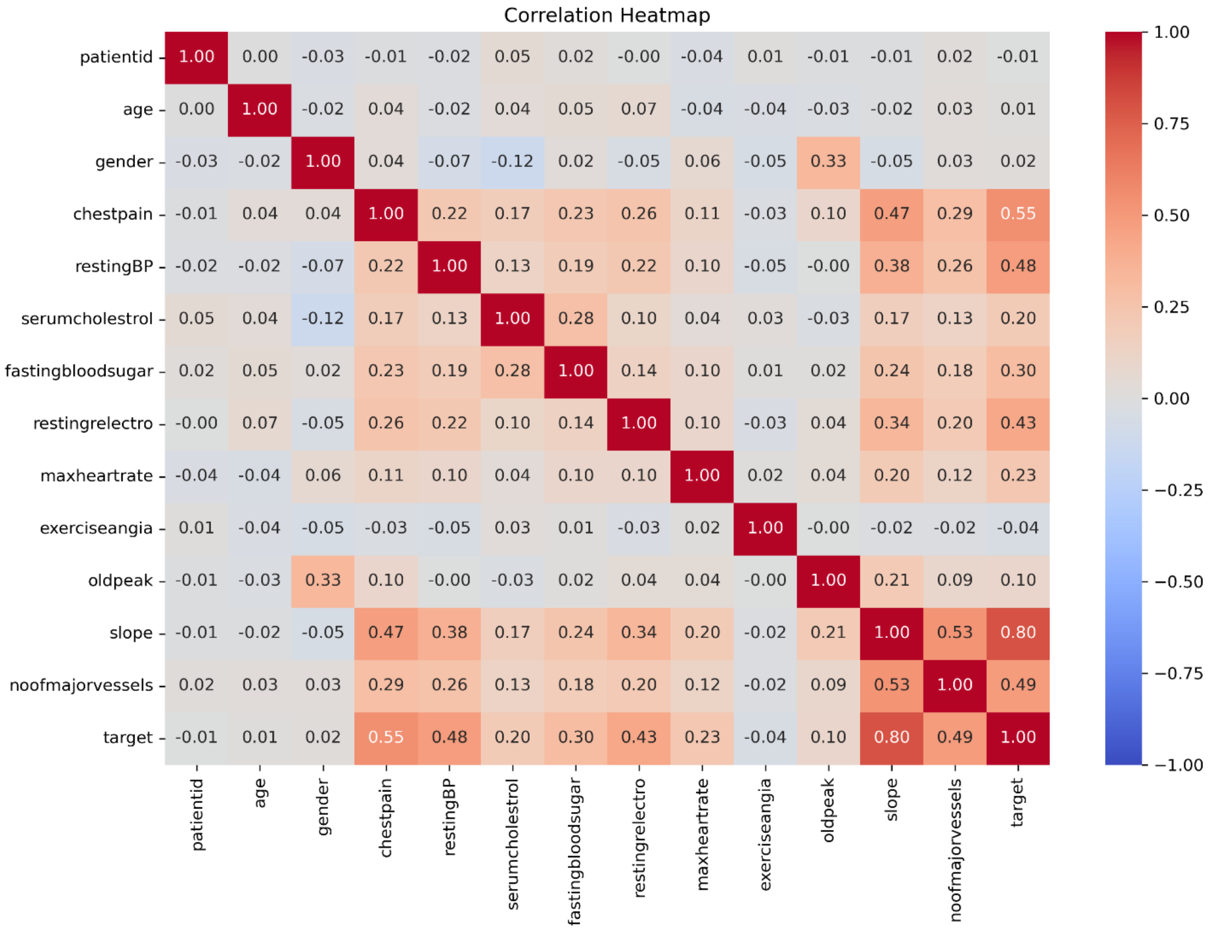
Numerical Feature Distributions:

- Histograms for features like age, restingBP, serumcholestrol, maxheartrate, and oldpeak show the overall distribution of these variables.
- **Analysis:**
    - Most variables display expected distributions, with some outliers in serumcholestrol and restingBP.



Correlation Heatmap:

- **Graph Description:** The heatmap visualizes relationships between numerical features.
- **Analysis:** Features like maxheartrate (positive correlation) and oldpeak (negative correlation) are highly predictive of heart disease.

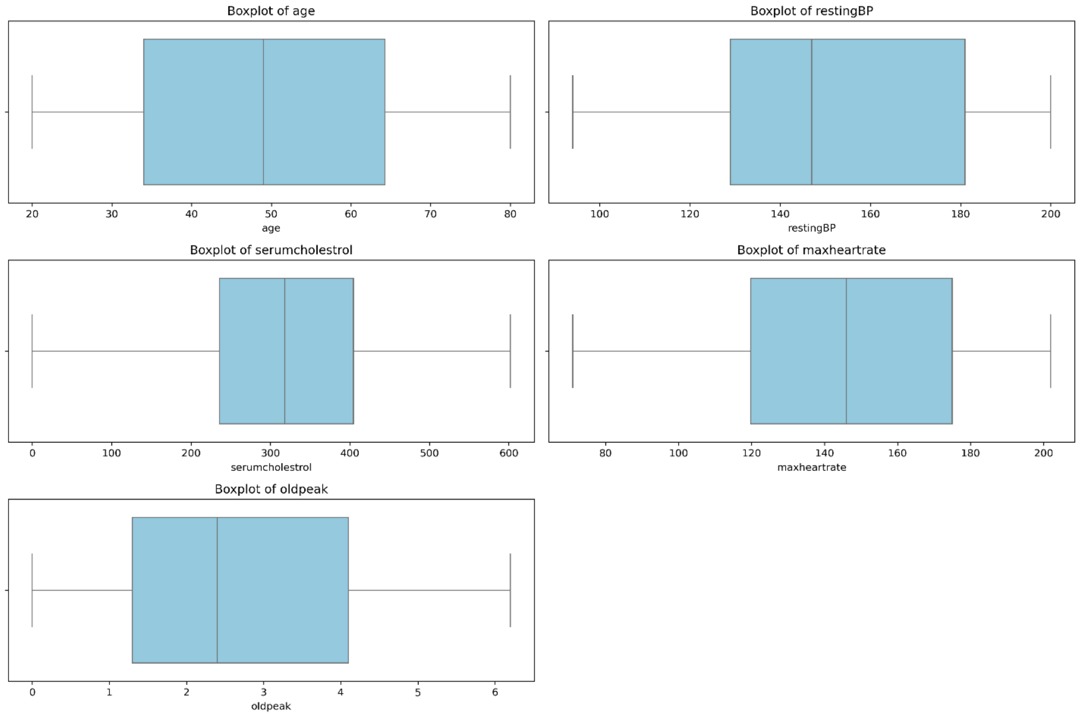


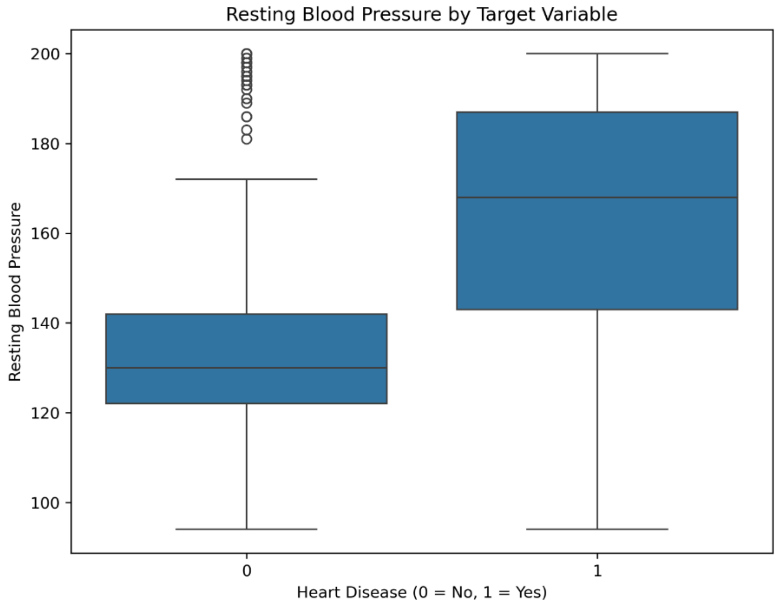
	patientid	age	gender	chestpain	restingBP	serumcholestrol	fastingbloodsugar	restingrelectro	maxheartrate	exerciseangia	oldpeak	slope	noofmajorvessels	target
patientid	1.00	0.00	-0.03	-0.01	-0.02	0.05	0.02	-0.00	-0.04	0.01	-0.01	-0.01	0.02	-0.01
age	0.00	1.00	-0.02	0.04	-0.02	0.04	0.05	0.07	-0.04	-0.04	-0.03	-0.02	0.03	0.01
gender	-0.03	-0.02	1.00	0.04	-0.07	-0.12	0.02	-0.05	0.06	-0.05	0.33	-0.05	0.03	0.02
chestpain	-0.01	0.04	0.04	1.00	0.22	0.17	0.23	0.26	0.11	-0.03	0.10	0.47	0.29	0.55
restingBP	-0.02	-0.02	-0.07	0.22	1.00	0.13	0.19	0.22	0.10	-0.05	-0.00	0.38	0.26	0.48
serumcholestrol	0.05	0.04	-0.12	0.17	0.13	1.00	0.28	0.10	0.04	0.03	-0.03	0.17	0.13	0.20
fastingbloodsugar	0.02	0.05	0.02	0.23	0.19	0.28	1.00	0.14	0.10	0.01	0.02	0.24	0.18	0.30
restingrelectro	-0.00	0.07	-0.05	0.26	0.22	0.10	0.14	1.00	0.10	-0.03	0.04	0.34	0.20	0.43
maxheartrate	-0.04	-0.04	0.06	0.11	0.10	0.04	0.10	0.10	1.00	0.02	0.04	0.20	0.12	0.23
exerciseangia	0.01	-0.04	-0.05	-0.03	-0.05	0.03	0.01	-0.03	0.02	1.00	-0.00	-0.02	-0.02	-0.04
oldpeak	-0.01	-0.03	0.33	0.10	-0.00	-0.03	0.02	0.04	0.04	-0.00	1.00	0.21	0.09	0.10
slope	-0.01	-0.02	-0.05	0.47	0.38	0.17	0.24	0.34	0.20	-0.02	0.21	1.00	0.53	0.80
noofmajorvessels	0.02	0.03	0.03	0.29	0.26	0.13	0.18	0.20	0.12	-0.02	0.09	0.53	1.00	0.49
target	-0.01	0.01	0.02	0.55	0.48	0.20	0.30	0.43	0.23	-0.04	0.10	0.80	0.49	1.00

Boxplots of Features by Target Variable:

- **Graph Description:** Boxplots were created to compare the distributions of key features like age, restingBP, and serumcholestrol across the two classes of the target variable (0 and 1).

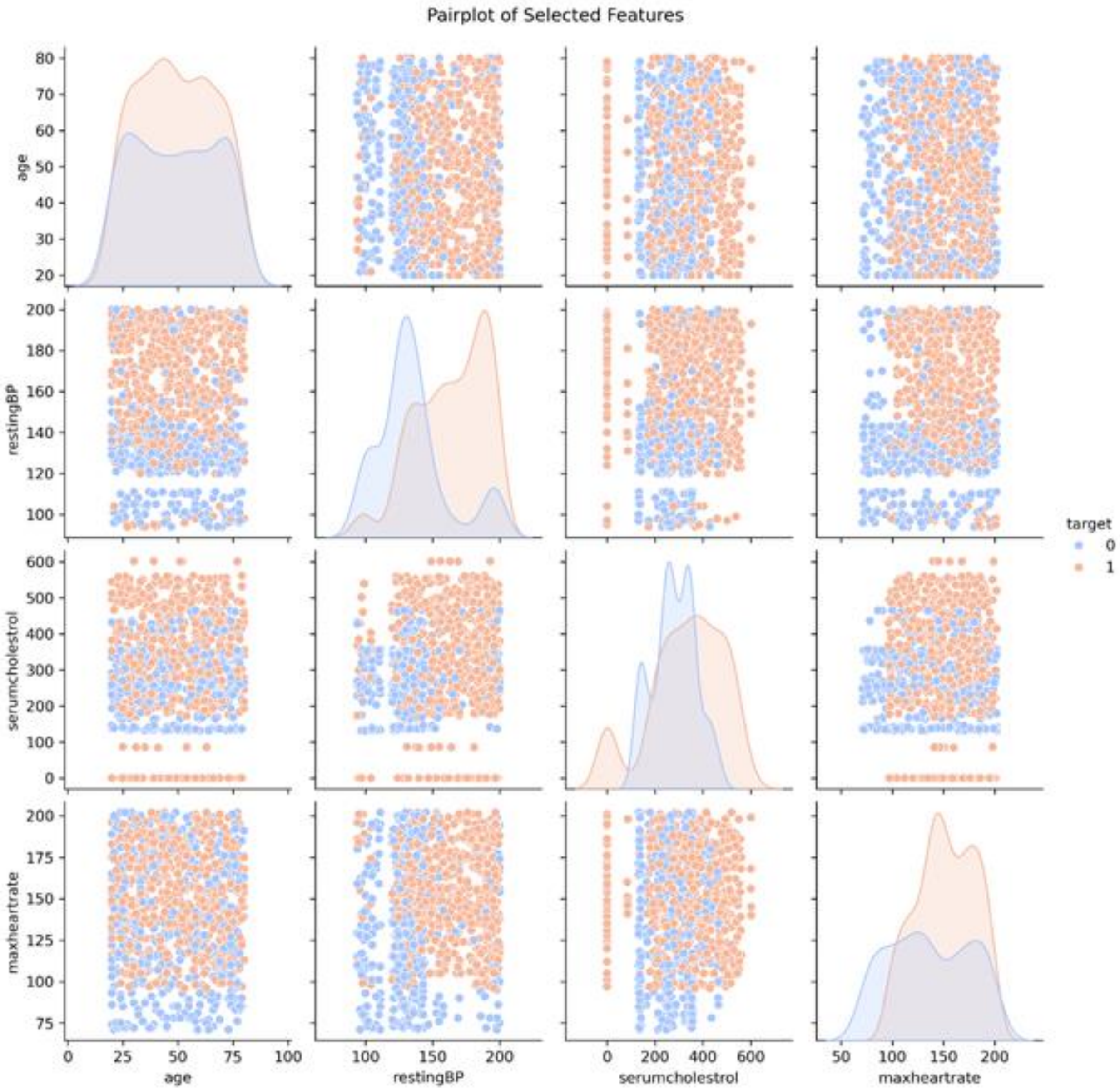
- **Analysis:**
  - **Age:** No significant difference in age distribution between the two groups.
  - **RestingBP:** Patients with heart disease tend to have higher resting blood pressure on average.
  - **Serum Cholesterol:** Patients with heart disease generally have higher cholesterol levels than those without.





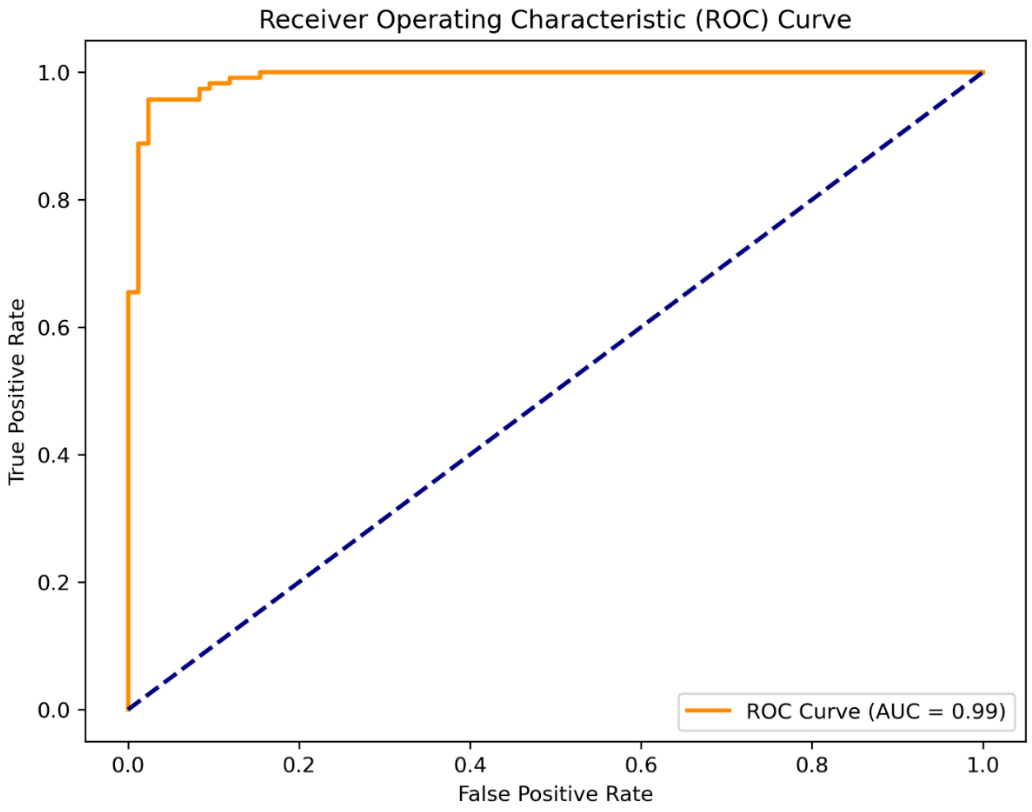
Pairplot of Selected Features:

- **Graph Description:** Pairwise relationships between selected features and the target variable were analyzed.
- **Analysis:**
  - Overlapping distributions for some variables indicate difficulty in separability.
  - Trends, such as lower max heart rates and higher old peak values for patients with heart disease, stand out.



ROC Curve:

- **Graph Description:** The curve plots the true positive rate against the false positive rate for the Logistic Regression model.
- **Analysis:** With an AUC of **99.10%**, the model demonstrates excellent predictive ability.



General Observations

1. Symmetric distributions and lack of significant outliers enhance the model's reliability.
2. Features like age, restingBP, and maxheartrate show meaningful differences across classes, supporting their inclusion in the model.

Conclusions and Recommendations

1. The Logistic Regression model is highly effective at identifying patients at risk of heart disease.
2. Key predictors, including **resting blood pressure** and **max heart rate**, provide valuable insights for healthcare interventions.
3. **Future Work:**
   - Experiment with advanced models (e.g., Random Forest, Gradient Boosting).
   - Include additional features, such as family history and lifestyle factors, to enhance predictive accuracy.