Heart Disease Prediction using Logistic Regression with GridSearchCV

I. Introduction

- Heart disease is one of the leading causes of death globally. Early prediction using machine learning can help in timely diagnosis and treatment.
- This project uses the Cardiovascular Disease Dataset to build a predictive model with Logistic Regression and GridSearchCV for hyperparameter tuning.

II. Dataset Overview

- **Dataset:** Cardiovascular_Disease_Dataset.csv
- **Rows:** 1000 (approx.)
- **Columns:** 14
- **Target variable:** target
    1. 0 → No Heart Disease
    2. 1 → Heart Disease
- **Features used:**
- Age
- Gender (0 = Female, 1 = Male)
- Chest Pain Type
- Resting Blood Pressure
- Serum Cholesterol
- Fasting Blood Sugar
- Resting ECG Results
- Maximum Heart Rate Achieved
- Exercise-Induced Angina
- Oldpeak (ST depression)
- Slope of ST segment
- Number of Major Vessels

The patientid column was removed since it does not contribute to prediction.

III. Methodology

Data Preprocessing:

- No missing values were found.
- Removed unnecessary columns (patientid).
- Split dataset into 75% training and 25% testing.

Model Training:

- Algorithm: Logistic Regression
- Hyperparameters tuned using GridSearchCV with 5-fold cross-validation.
- **Parameters tested:**

1. Regularization strength C = [0.01, 0.1, 1, 10]
2. Solvers = ["liblinear", "saga"]
3. Penalties = ["l1", "l2"]

IV. Results

Best Model Found:

- C = 10
- Penalty = l1
- Solver = liblinear
- Best Cross-validation Accuracy: 95.8%

Accuracy:

- Training Accuracy: 97.2%
- Testing Accuracy: 96.4%

Confusion Matrix (Test Data):

	Predicted 0	Predicted 1
Actual 0	108	4
Actual 1	5	133

Classification Report (Test Data):

- Precision (No Disease): 0.96
- Precision (Disease): 0.97
- Recall: 0.96 for both classes
- Overall Accuracy: 96%

Correlation Heatmap:
A heatmap was plotted to visualize relationships between features.
Strong correlations were observed between:

- Age and Resting BP
- Cholesterol and Max Heart Rate
- Exercise Angina and Oldpeak

V. New Patient Prediction

A test case was given for a 55-year-old male with the following parameters:

- Resting BP: 140
- Cholesterol: 250
- Max Heart Rate: 160
- Chest Pain Type: 2
- Oldpeak: 1.2
- No major vessels blocked

Prediction: ✅ Heart Disease

VI. Conclusion

- Logistic Regression with tuned hyperparameters performed very well (96% accuracy) in predicting heart disease.
- The model showed balanced precision and recall, indicating reliability in both detecting disease and avoiding false alarms.
- This approach can be integrated into a clinical decision-support system for early detection of cardiovascular disease.