MLDL Practical 1

Name: Aryan Sanjay Kale

Class: D15C

Roll No: 28

Batch: B

Aim: Implement Linear and Logistic Regression on real-world datasets

1. Dataset Source

- **Dataset Name:** Heart Disease UCI Dataset
- **Source Platform:** Kaggle
- **Dataset Description:** This dataset contains diagnostic measurements used to predict the presence of heart disease in patients based on several medical attributes.
- **Total Number of Instances:** 1,025
- **Number of Input Features:** 13 (all numerical)
- **Target Variable:** target
    - **1** indicates presence of heart disease
    - **0** indicates absence of heart disease

2. Mathematical Formulation of the Algorithms

Linear Regression

Linear Regression models the relationship between independent variables and a continuous dependent variable using a linear equation.

- **Model Equation:** $\hat{y}=\beta0+\beta1x1+\beta2x2+...+\beta nxn$
- **Cost Function (Mean Squared Error):** $MSE = (1/n) \times \sum (yi – ŷi)^2$

Logistic Regression

Logistic Regression is a supervised learning algorithm used for binary classification. It applies a sigmoid function to map linear predictions to probabilities between 0 and 1.

- **Logistic (Sigmoid) Function:** $\sigma(z)=1/(1+e^{-z})$
- **Decision Rule:**
    - If $\sigma(z) \ge 0.5 \rightarrow$ Class 1 (Disease)
    - If $\sigma(z) < 0.5 \rightarrow$ Class 0 (Healthy)

3. Methodology / Workflow

The experiment follows a structured machine learning pipeline:

1. Dataset acquisition from Kaggle
2. Data exploration and understanding
3. Data preprocessing and Feature normalization using StandardScaler
4. Splitting dataset into training and testing sets (80:20)
5. Model training (Linear and Logistic Regression)
6. Model evaluation and Hyperparameter tuning

4. Performance Analysis

Logistic Regression Results

Metric	**Value**
Accuracy	80.98%
Precision (Class 1)	76.0%
Recall (Class 1)	91.0%
F1-Score (Class 1)	83.0%

High recall indicates that the model successfully identifies patients with heart disease, which is critical in medical diagnosis.

Linear Regression Results

Metric	**Value**
MSE	0.124

R^2	0.506

Hyperparameter Tuning

The best performance was achieved with the following tuned parameters:

- **C (Regularization Strength):** 1
- **Solver:** liblinear

5. Output

Confusion Matrix (Tuned):

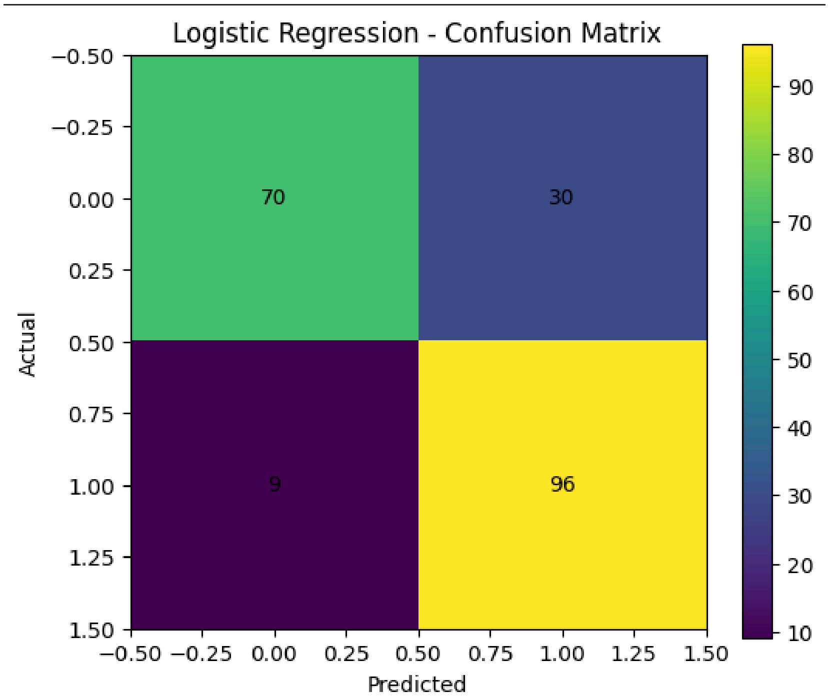
| | Predicted Healthy | Predicted Disease |

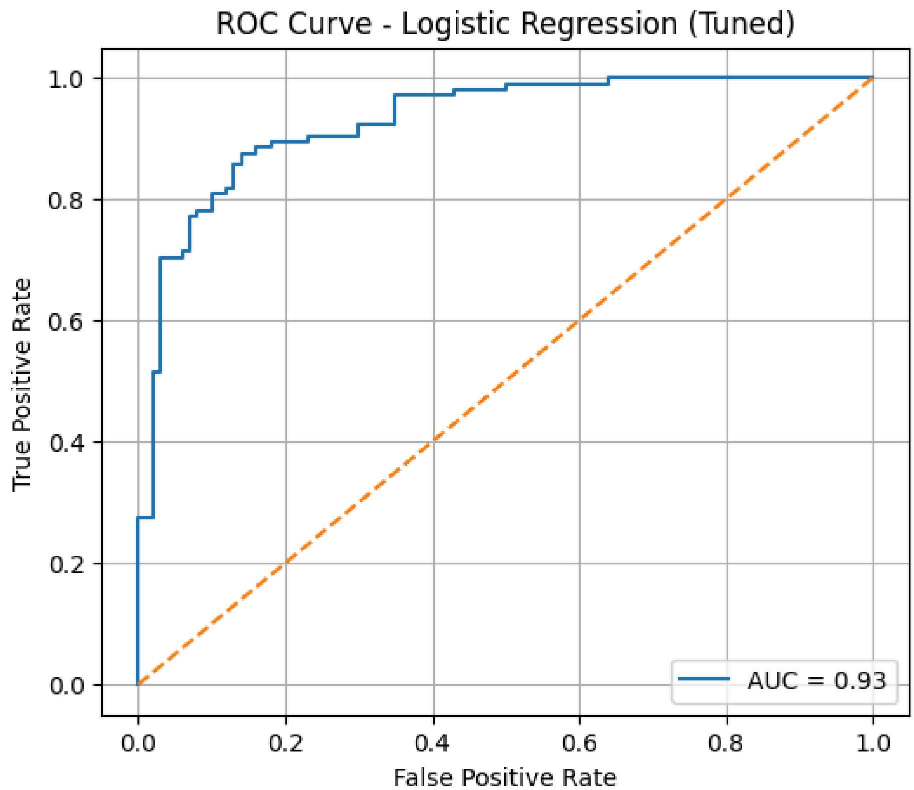
| **Actual Healthy** | 70 (True Negative) | 30 (False Positive) |

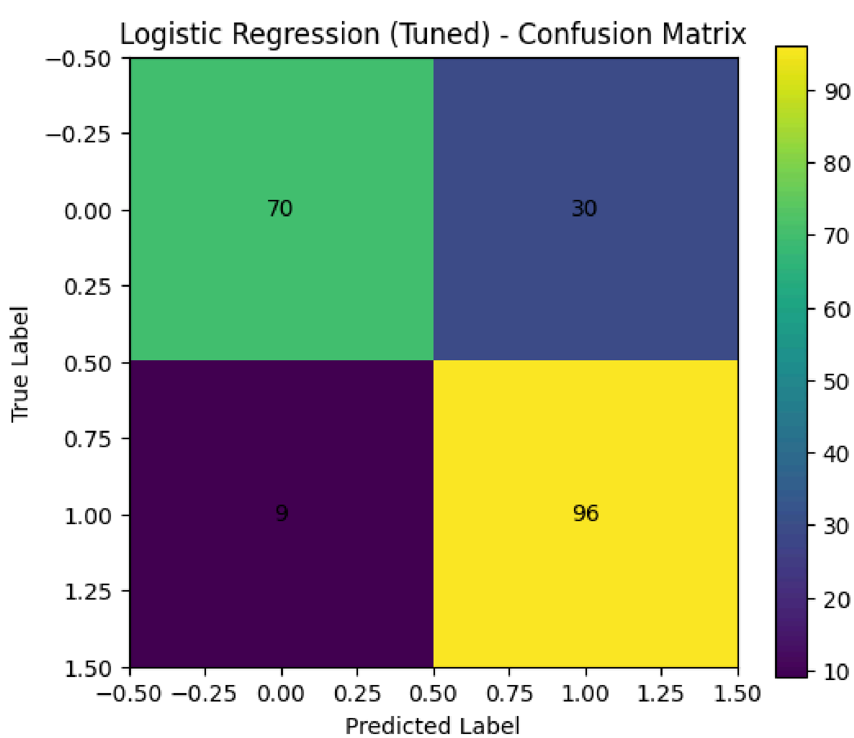
| **Actual Disease** | 9 (False Negative) | 96 (True Positive) |

ROC Curve Performance:

- **AUC Score:** 0.93




ROC Curve - Logistic Regression (Tuned)
True Positive Rate
False Positive Rate
AUC = 0.93
0.0
0.2
0.4
0.6
0.8
1.0


Logistic Regression (Tuned) - Confusion Matrix
True Label
Predicted Label
70
30
9
96

6. Conclusion

In this experiment, Linear and Logistic Regression were successfully implemented on a real-world heart disease dataset. Logistic Regression proved to be an effective and interpretable model, achieving a strong AUC of 0.93 and high recall. This highlights the importance of data preprocessing and hyperparameter tuning in developing reliable models for healthcare applications.