Train-Test Split & Evaluation Metrics

Dataset: Heart Disease Dataset
Model: Logistic Regression
Train-Test Split: 80% Train / 20% Test
Train Samples: 820
Test Samples: 205

---

1. Objective

The goal of this task is to build a baseline classification model using **Logistic Regression** to predict whether a patient has heart disease (1) or not (0), and evaluate the model using key performance metrics such as **Accuracy, Precision, Recall**, and **Confusion Matrix**.

---

2. Methodology

Train vs Test Split

The dataset was divided into:

- **Training set (80%)**: Used to train the model and learn patterns.
- **Testing set (20%)**: Used to evaluate the model on unseen data.

This split ensures the evaluation is realistic and helps detect issues like **overfitting** (good performance on training data but poor performance on new data).

Model Used

A **Logistic Regression** classifier was trained because it is a strong and interpretable baseline model for binary classification.

---

3. Model Results (Actual Output)

Evaluation Metrics

- **Accuracy: 0.7951** (79.51%)
- **Precision: 0.7563** (75.63%)
- **Recall: 0.8738** (87.38%)

---

4. Confusion Matrix Analysis (TP/TN/FP/FN)

Confusion Matrix:

```
[[73 29]
 [13 90]]
```

From the matrix:

- **TN (True Negatives) = 73**
  → Model correctly predicted **no heart disease**.
- **FP (False Positives) = 29**
  → Model predicted heart disease, but actually no disease.
- **FN (False Negatives) = 13**
  → Model predicted no disease, but actually heart disease.
- **TP (True Positives) = 90**
  → Model correctly predicted **heart disease**.

---

5. Interpretation of Results

Accuracy (79.51%)

The model correctly classified **163 out of 205** test samples overall. This indicates decent baseline performance.

Precision (75.63%)
Out of all patients predicted as heart disease, **75.63% were actually positive**.
This means there are some **false alarms (FP = 29)**, where the model incorrectly predicts disease.

Recall (87.38%)
The model successfully identified **87.38% of actual heart disease cases**, which is very important in healthcare.
Here, **FN = 13** indicates the model missed 13 real heart disease cases, but overall recall is strong.

Key Insight (Medical Relevance)
In medical diagnosis tasks, **False Negatives (FN)** are more dangerous than False Positives because missing a heart disease case can delay treatment.
This model performs well because it has **high recall**, meaning it detects most disease cases.

---

6. Conclusion

The Logistic Regression model achieved **79.51% accuracy**, with a strong **recall of 87.38%**, making it effective at identifying heart disease cases. The confusion matrix shows the model correctly detects many positive cases (**TP = 90**) and misses relatively few (**FN = 13**). This makes it a good baseline model for heart disease prediction. Further improvement can be achieved through feature scaling, hyperparameter tuning, and advanced models like Random Forest or XGBoost.