Diabetes Prediction using Logistic Regression (Pima Indians Dataset)

■ Overview
This project predicts whether a patient is diabetic based on diagnostic measurements using **Logistic Regression**. The model is trained on the **Pima Indians Diabetes Dataset** from Kaggle and aims to assist in early diagnosis and health risk assessment through machine learning techniques.

■ Dataset Information
- **Source:** [Kaggle - Pima Indians Diabetes Dataset](https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database)
- **Attributes:**
- Pregnancies
- Glucose
- Blood Pressure
- Skin Thickness
- Insulin
- BMI
- DiabetesPedigreeFunction
- Age
- Outcome (Target: 1 = Diabetic, 0 = Non-Diabetic)
- **Size:** 768 samples, 9 columns

■■ Workflow
1. **Data Loading and Cleaning**
- Import dataset and handle missing or zero values.
- Perform data standardization and normalization.

2. **Exploratory Data Analysis (EDA)**
- Visualize distributions using histograms and boxplots.
- Identify correlations using a heatmap.

3. **Feature Engineering**
- Scale data using StandardScaler to improve model convergence.

4. **Model Training**
- Split data into training and testing sets (80:20).
- Train a **Logistic Regression** classifier using Scikit-learn.

5. **Model Evaluation**
- Evaluate performance using Accuracy, Confusion Matrix, Precision, Recall, and F1-Score.
- Generate ROC Curve and AUC for classification quality.

■ Mathematical Explanation
Sigmoid Function:
\[h_θ(x) = \frac{1}{1 + e^{-θ^T x}} \]

Cost Function (Binary Cross-Entropy):
\[J(θ) = -\frac{1}{m} \sum_{i=1}^{m} [y^{(i)} \log(h_θ(x^{(i)})) + (1 - y^{(i)}) \log(1 - h_θ(x^{(i)}))] \]

Gradient Descent:
\[θ := θ - α \frac{∂}{∂θ}J(θ) \]

The model learns the optimal parameters (θ) that minimize the cost function.

■ Results and Analysis
- **Accuracy:** ~78–82%

- **Precision & Recall:** Balanced performance indicating effective classification.
- **Confusion Matrix:** Displays true vs. predicted outcomes.
- **AUC-ROC Curve:** Demonstrates good model discrimination ability.

Visualization and metric evaluation confirm that Logistic Regression is well-suited for binary classification problems like diabetes prediction.

■ Future Scope
- Implement advanced models like Random Forest or XGBoost for higher accuracy.
- Deploy the model as a web app using Streamlit or Flask for real-time prediction.
- Expand the dataset with demographic or lifestyle data to improve robustness.

■■■ Author
Elavarasi Chinnadurai
Department of Agriculture Engineering
Passionate about Machine Learning, Data Analytics, and AI-driven Healthcare Solutions.