Dataset: https://www.kaggle.com/datasets/mathchi/diabetes-data-set

Task 1 – Load and Inspect the Dataset

1. Load the CSV file into a pandas DataFrame.
2. Print / Display:
   - Shape of the dataset
   - Column names
   - First 5 rows
   - Descriptive statistics (describe)
   - Info (data types & non-null counts)

Task 2 – Missing Value Analysis

1. Check the number of missing or zero values per column.
2. Decide how missing values should be handled.

Task 3 – Exploratory Data Analysis (EDA)

Target Variable Analysis

1. Plot the class distribution (Outcome = 0 vs Outcome = 1).
2. Check whether the dataset is balanced or imbalanced.
3. Visualize class frequencies with a bar chart.

Feature Relationships

1. Compute the correlation matrix for numerical features.
2. Visualize correlations using a heatmap.
3. Identify the top features most correlated with the Outcome.

Feature-Level Analysis

1. Visualize distributions of Glucose, BMI, and Age for diabetic vs non-diabetic groups.
2. Compare the average glucose and BMI values between classes using boxplots.

Task 4 – Data Preprocessing

Feature and Target Separation

1. Separate dataset into features (X) and target (y = Outcome).

Feature Type Separation

1. Identify numerical features (all in this dataset are numeric).

Preprocessing Pipeline

1. Create a preprocessing pipeline that includes:
    - Imputation of missing values (mean/median)
    - Feature scaling
2. (If any feature encoding is required later) create pipeline for categorical features – but here all are numerical.

Task 5 – Train-Test Split

1. Split the data into training and test sets with:
    - 80% training data
    - 20% testing data
2. Set random_state=42.

Task 6 – Logistic Regression Model

1. Create a pipeline that includes:
    - Preprocessing
    - Logistic Regression classifier
2. Train the model on the training set.
3. Predict the target values for the test set.

Task 7 – Model Evaluation

Evaluate the model using:

1. Accuracy
2. Confusion Matrix
3. Precision, Recall, F1-Score
4. Log Loss (Binary Cross-Entropy)

Analyze whether the model:

- Fits well
- Shows evidence of overfitting or underfitting

Task 8 – Regularization

Train and compare the following logistic regression variants:

1. L2 (Ridge) regularization
2. L1 (Lasso) regularization
3. ElasticNet

Analyze:

- Effect on coefficients
- Effect on model performance

Task 9 – Cross-Validation

1. Apply k-fold cross-validation with k = 5.
2. Evaluate model stability using accuracy and log loss.
3. Compare cross-validation results with test set results.

Which model performed best across all evaluation metrics?
Support your answer with results from accuracy..