A. Introduction

In this project, we aim to predict a health-related numerical value using patient medical data from the Diabetes dataset. Accurate predictions can assist in medical decision-making and improving treatment planning.

B. Data Exploration (EDA)

Dataset Description:

- Number of samples: 768
- Features: Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age
- Target: Glucose (for regression task)

Key Visualizations:

- Correlation Heatmap
- Histograms
- Scatter plot of actual vs predicted values

Key Findings:

- Glucose and BMI have high correlation with the outcome.
- Some features such as Insulin and SkinThickness have skewed distributions.

C. Methodology

Preprocessing Steps:

- Standardized features using StandardScaler
- No missing values present in the dataset

Target Variable:

- Glucose (continuous numeric value)

Algorithm Used:

- Linear Regression
- Chosen for its simplicity and interpretability as a baseline model

D. Model Training

Train/Test Split:

- 80% training data, 20% testing data
- Used random_state=42 for reproducibility

Model:

- LinearRegression() from scikit-learn
- No hyperparameter tuning applied

E. Results and Evaluation

Evaluation Metrics:

- RMSE: 27.12
- MAE: 21.35
- R^2 Score: 0.47

Error Analysis:

- The model underperforms on outlier samples.
- Residuals indicate some degree of heteroscedasticity.

F. Conclusion

Summary:

- The linear regression model gave a moderate baseline performance on the regression task.
- Further improvements can be achieved with more advanced models.

Challenges Faced:

- Skewed distributions and possible zero-values in some features
- Limited interpretability of certain features without domain expertise

Future Improvements:

- Use ensemble models such as Random Forest or Gradient Boosting
- Perform feature engineering and possibly dimensionality reduction