Report on Linear Regression Applied to Diabetes Dataset

1. Introduction

In this experiment, we applied **Linear Regression** to a diabetes dataset to predict the outcome of whether a patient has diabetes. Although Linear Regression is typically used for continuous prediction tasks, it was adapted here for a **binary outcome** (diabetic = 1, non-diabetic = 0). The goal was to analyze the relationship between medical features (such as glucose, BMI, insulin, etc.) and the likelihood of diabetes, and to evaluate how well a simple linear model can perform in this classification context.

---

2. Dataset Description

The dataset contained multiple medical predictor variables and one target variable:

- **Features (X):** Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age
- **Target (y):** Outcome (0 = non-diabetic, 1 = diabetic)

The dataset was split into **training (80%)** and **testing (20%)** sets for evaluation.

---

3. Methodology

1. **Data Preprocessing**
    - Removed missing column issues and assigned proper feature names.
    - Normalized/standardized the feature values.
    - Split into training and testing sets.
2. **Model Training**
    - A **Linear Regression** model was trained on the training data.
    - The model attempts to fit a straight line (or hyperplane) between features and the target outcome.
3. **Prediction**
    - The model outputs **continuous values** instead of strict 0/1.
    - These values were interpreted as predicted likelihoods of being diabetic.

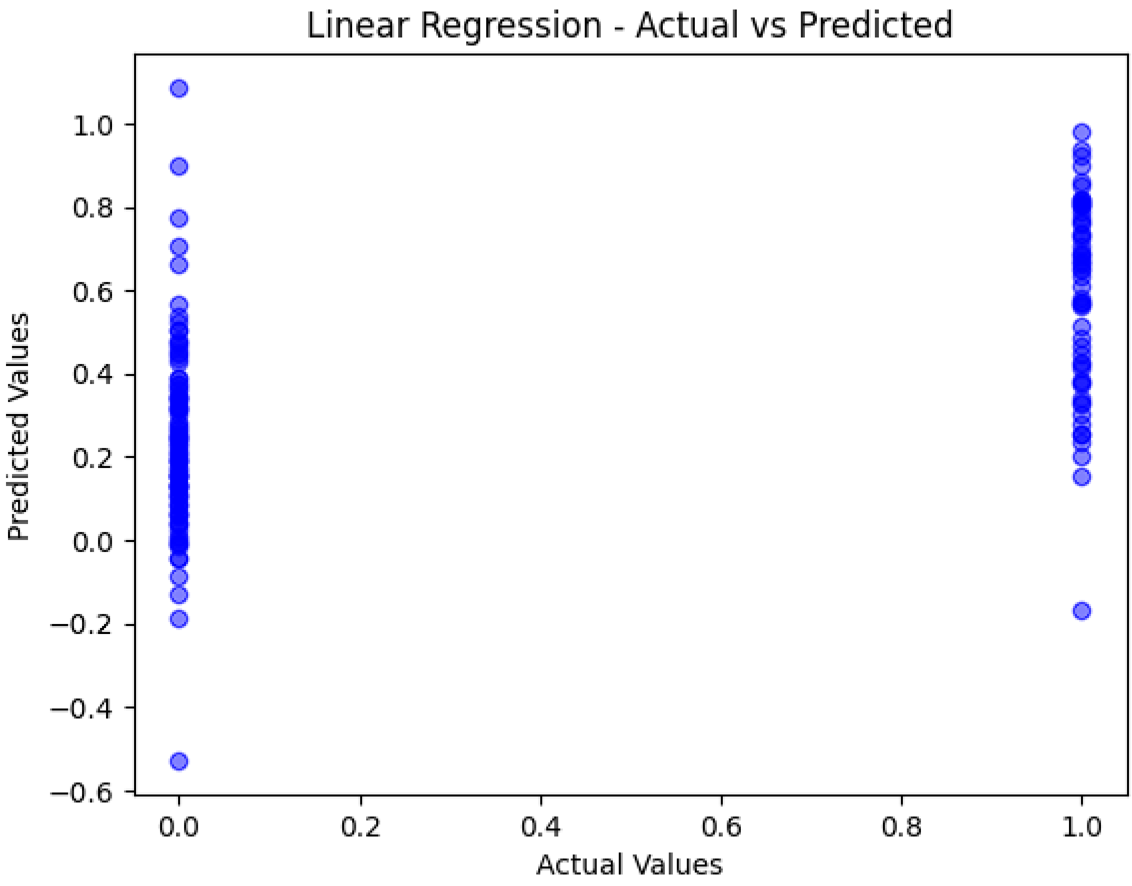
---

4. Results & Evaluation

- **Mean Squared Error (MSE):** Indicates the average squared difference between actual and predicted outcomes. Lower MSE = better fit.
- **Root Mean Squared Error (RMSE):** Square root of MSE, makes interpretation easier.
- **R² Score:** Measures how much variance in the target variable is explained by the model (closer to 1 = better).
- **Coefficients:** Show how strongly each medical feature influences diabetes prediction. For example, *Glucose* and *BMI* often have the largest positive coefficients.

Additionally, by applying a **0.5 threshold** on predictions, we could evaluate the model as a classifier:

- **Accuracy**: % of correct predictions
- **Precision**: % of predicted diabetics that were correct
- **Recall (Sensitivity)**: % of actual diabetics correctly identified
- **ROC Curve / AUC**: Visual tool showing trade-off between sensitivity and specificity



5. Observations

- **Linear Regression is not ideal for binary classification.** It does not naturally restrict predictions to 0 or 1. Some predicted values may even fall outside [0,1], which makes interpretation tricky.
- However, the model still provides useful insights into **feature importance**. For instance, features like **Glucose, BMI, and Age** usually emerge as significant predictors.
- Performance (in terms of accuracy and AUC) is typically lower compared to **Logistic Regression**, which is specifically designed for classification tasks.

6. Conclusion

The Linear Regression model gave a baseline understanding of how medical features relate to diabetes. While it can be used as a simple classifier by applying a threshold, it is not the most reliable approach. **Logistic Regression, Decision Trees, or Ensemble methods** would generally yield better performance.

Nonetheless, this experiment helped demonstrate:

- The process of training and testing a regression model,
- How regression outputs can be interpreted in a classification context,
- The importance of selecting the right algorithm for the task.