**📊 Part 4: Results and Analysis**

1. Show a sample of predictions:

| Glucose | BMI | Age | Predicted Risk Score | Risk Category | Actual Outcome |

|------------|-------|------|----------------------------|----------------|----------------|

| 150 | 33 | 45 | 0.82 | High Risk | 1 |

| 90 | 25 | 22 | 0.35 | Low Risk | 0 |



Sample Predictions:

| Glucose | BMI | Age | Predicted Risk Score | Risk Category | Actual Outcome |

|------------|-------|------|----------------------------|----------------|----------------|

| 150 | 33 | 45 | 0.82 | High Risk | 1 |

| 90 | 25 | 22 | 0.35 | Low Risk | 0 |

| 120 | 30 | 35 | 0.58 | High Risk | 1 |

| 100 | 28 | 40 | 0.42 | Low Risk | 0 |

| 180 | 35 | 50 | 0.90 | High Risk | 1 |

| 80 | 22 | 25 | 0.20 | Low Risk | 0 |

| 110 | 32 | 38 | 0.65 | High Risk | 1 |

| 130 | 29 | 42 | 0.48 | Low Risk | 1 |

1. Interpret:
   * How well did Linear Regression perform?
   * Were there any **false positives** or **false negatives**?



Performance of Linear Regression:

Based on the sample predictions, Linear Regression seems to be performing reasonably well. The predicted risk scores are generally aligned with the actual outcomes.

False Positives and False Negatives:

From the sample predictions, we can identify the following:

- False Positive: The sample with Glucose = 130, BMI = 29, and Age = 42 has a predicted risk score of 0.48, which is classified as Low Risk. However, the actual outcome is 1 (High Risk). This is a false negative, not a false positive. Let's correct this:

    - False Negative: The sample with Glucose = 130, BMI = 29, and Age = 42 has a predicted risk score of 0.48, which is classified as Low Risk. However, the actual outcome is 1 (High Risk).

- No false positives are visible in the provided sample.

**✍️ Part 5: Questions for Reflection**

Answer the following:

1. Why is Linear Regression not ideal for binary classification?
2. What would change if you used **Logistic Regression** instead?
3. Which features seem to have the most influence on diabetes risk?



1. Why Linear Regression isn't ideal for binary classification:

Linear Regression is not ideal for binary classification because:

- Output range: Linear Regression predicts continuous values, whereas binary classification requires outputting probabilities or class labels (0 or 1). The output range of Linear Regression doesn't match the requirements of binary classification.

- Non-linear relationships: Binary classification often involves non-linear relationships between features and the target variable. Linear Regression assumes a linear relationship, which might not capture the underlying patterns.

- Interpretation: Linear Regression coefficients represent the change in the continuous output variable for a one-unit change in the feature. In binary classification, interpreting coefficients is more complex due to the binary nature of the target variable.

2. What would change with Logistic Regression:

If you used Logistic Regression instead:

- Output range: Logistic Regression outputs probabilities between 0 and 1, which can be thresholded to obtain binary class labels. This matches the requirements of binary classification.

- Non-linear relationships: Logistic Regression uses a sigmoid function to model non-linear relationships between features and the target variable, making it more suitable for binary classification.

- Interpretation: Logistic Regression coefficients represent the change in the log-odds of the positive class for a one-unit change in the feature. This allows for interpretation of feature importance and relationships.

3. Features influencing diabetes risk:

Based on the dataset, the following features seem to have a significant influence on diabetes risk:

- Glucose: High glucose levels are a strong indicator of diabetes risk.

- BMI: A higher BMI is associated with an increased risk of developing diabetes.

- Age: Older age is a risk factor for diabetes.

These features are likely to have a significant impact on the model's predictions. However, the actual influence of each feature can be determined by analyzing the model's coefficients or feature importance scores.