Fatima Alamgir Apurba
Programming in Data Science and Artificial Intelligence

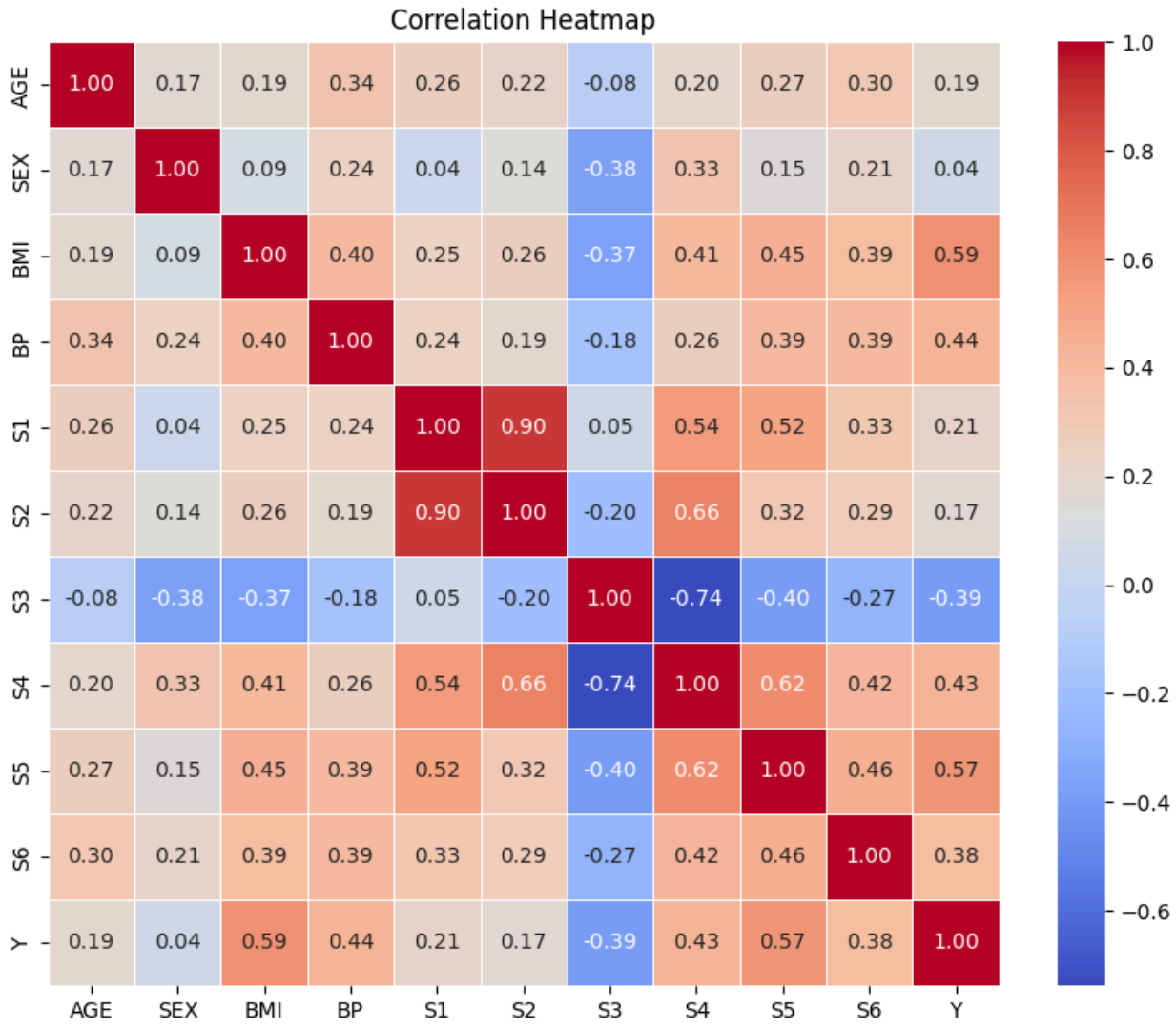
Report: Predicting Diabetes Outcome

Introduction

The goal of this analysis was to predict the target variable `Y` (diabetes outcome) using clinical and demographic features. We aimed to build two models—a **Linear Regression model** and a **Random Forest model**—and compare their performance in terms of predictive accuracy. Correlation analysis was performed to understand relationships within the dataset and guide model development.

Methods

1. **Correlation Analysis**:
    - Explored relationships between features and the target variable (`Y`) using a correlation matrix.
    - Identified features with high positive or negative correlations to prioritize for modeling.



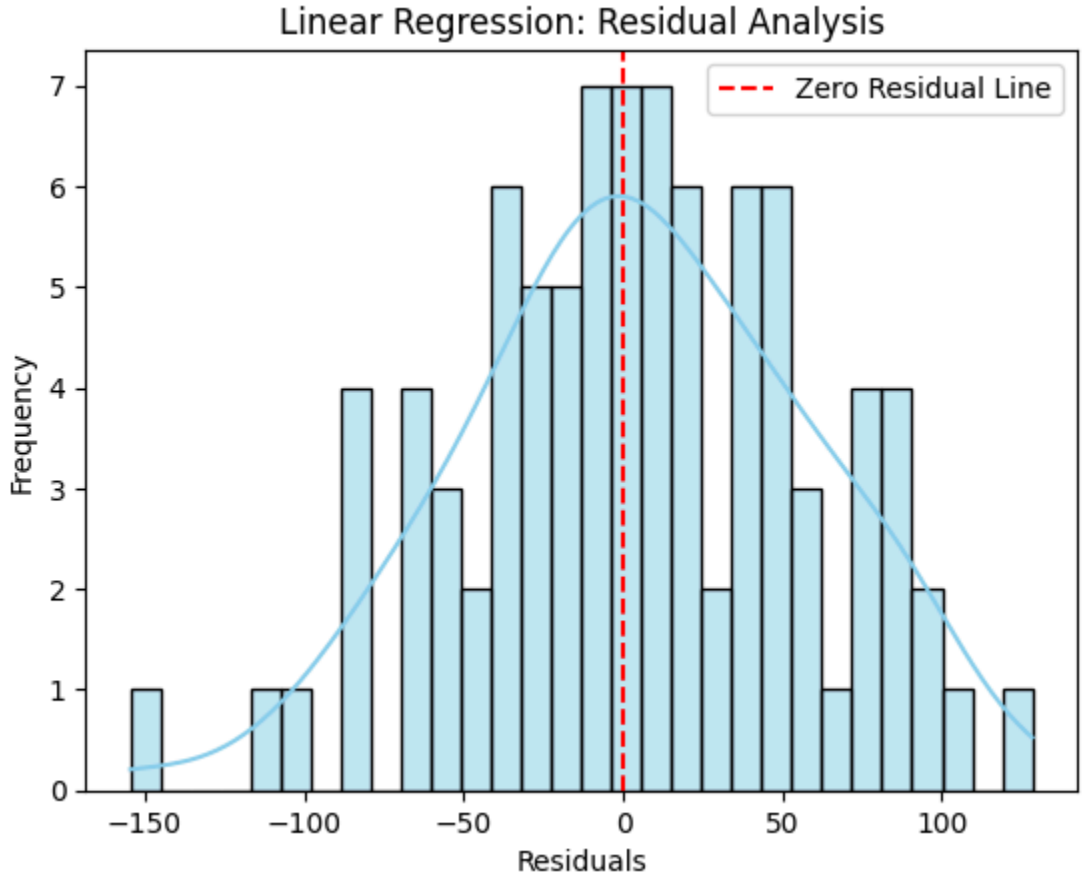
2. **Data Preprocessing**:
    - Features were standardized using `StandardScaler` for consistency.
    - The dataset was split into **training (80%)** and **testing (20%)** subsets to evaluate model performance.
3. **Modeling**:
    - **Model 1**: A **Linear Regression** model was built to establish a simple, interpretable benchmark.
    - **Model 2**: A **Random Forest Regressor** was trained to capture potential non-linear relationships in the data.

Results

Model 1: Linear Regression

Performance:

- **Mean Absolute Error (MAE):** 42.79
- **Mean Squared Error (MSE):** 2900.19
- **Root Mean Squared Error (RMSE):** 53.85
- **R² (Coefficient of Determination):** 0.45



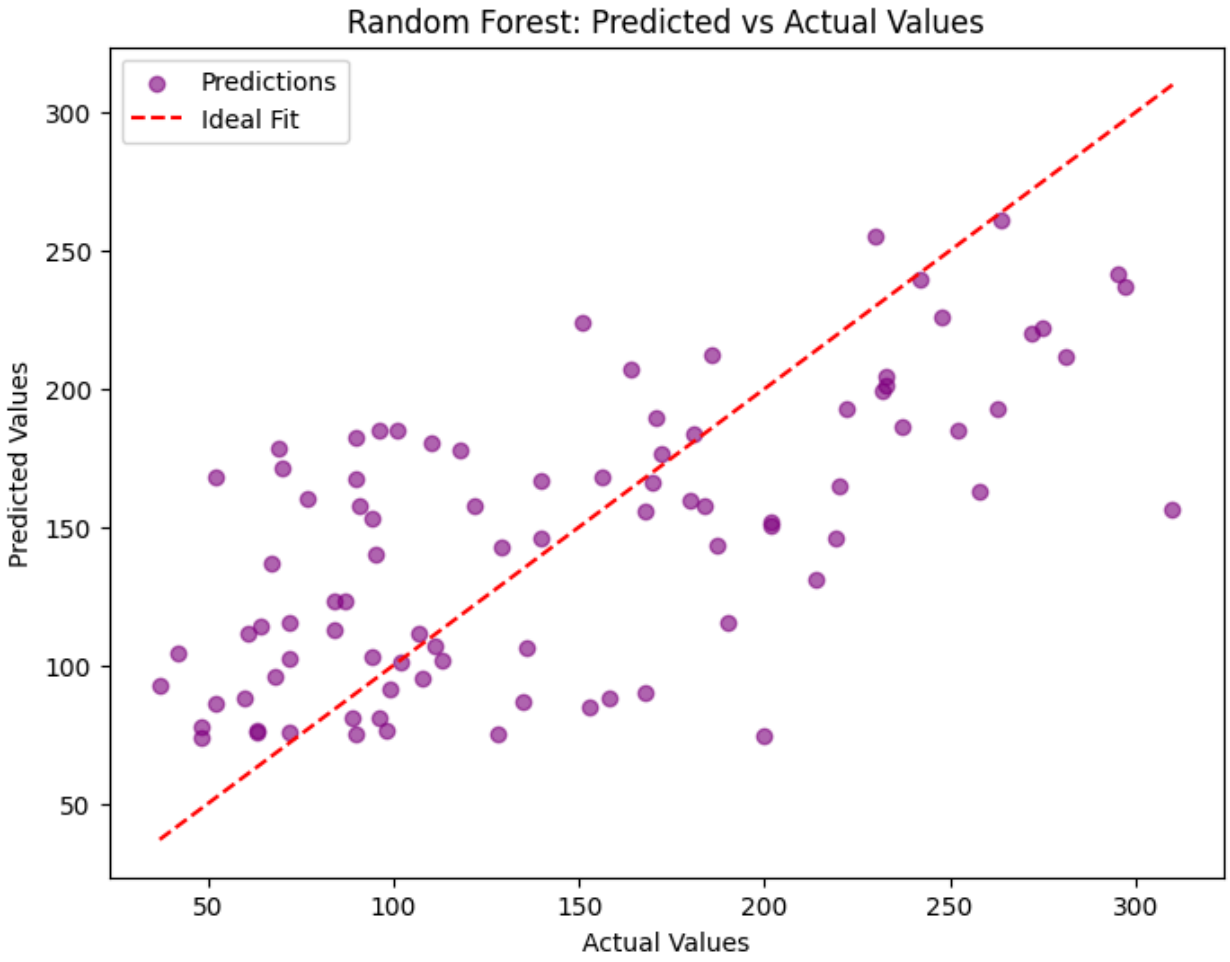
The residual analysis for the Linear Regression model shows that the errors (residuals) are symmetrically distributed around zero. This suggests that the model's predictions are unbiased, as it does not systematically overpredict or underpredict. However, the spread of residuals indicates that the model struggles to fully capture the variability in the data, particularly for cases where relationships may not be purely linear. While interpretable and straightforward, the Linear Regression model is less effective in handling more complex patterns

Model 2: Random Forest

Performance:

- **Mean Absolute Error (MAE):** 39.21

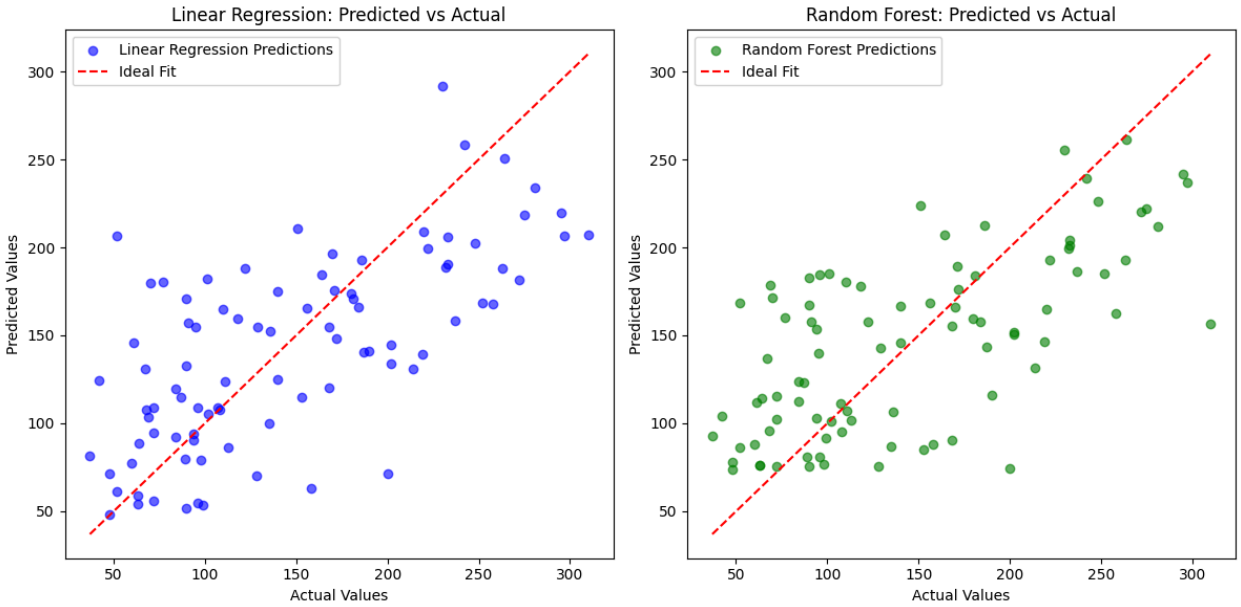
- **Mean Squared Error (MSE):** 2410.35
- **Root Mean Squared Error (RMSE):** 49.10
- **R² (Coefficient of Determination):** 0.52



This model compares predicted values to actual outcomes. Most data points closely align with the red dashed line, which represents perfect predictions. This indicates that the model performs well in capturing the patterns of the dataset. However, a few points deviate from the line, suggesting the model struggles with certain extreme or complex cases. Overall, the Random Forest model shows strong predictive performance, making it a reliable choice for this task.

Comparison Visualization: Predicted vs Actual Values

This side-by-side comparison highlights how the Linear Regression and Random Forest models perform when predicting the target variable:



1. **Linear Regression**:

    - The scatterplot shows that while some predictions align closely with the ideal fit line (red dashed line), many points deviate significantly.
    - This indicates that Linear Regression struggles to fully capture the underlying relationships in the data, especially for complex cases.

2. Random Forest:

    - The predictions from the Random Forest model are much closer to the ideal fit line, with fewer deviations.
    - This suggests that Random Forest is better at capturing the non-linear relationships within the dataset, leading to improved accuracy.

Conclusion:

These visualizations reveal the strengths and limitations of the models:

- The **Random Forest model** demonstrates better overall predictive accuracy, particularly for complex or non-linear relationships.
- The **Linear Regression model**, while interpretable and unbiased, may be less effective in handling complex patterns, leading to higher residual errors for some predictions.

These findings justify the preference for Random Forest in scenarios where accuracy is a priority, whereas Linear Regression remains valuable for interpretability and simplicity.