# Recursive Feature Elimination (RFE) with Linear Regression

1. **Introduction**: This report summarizes the findings from the Recursive Feature Elimination (RFE) analysis performed on the diabetes dataset available from scikit-learn. The objective of this analysis was to identify the optimal number of features that best predict the progression of diabetes and to evaluate the significance of each feature in contributing to the model's performance

   .

2. **Dataset Description:** The dataset used in this analysis is the Diabetes dataset from scikit-learn, which consists of 442 samples with 10 numerical features. These features represent various physiological and medical measurements that are potentially relevant to diabetes progression. The target variable represents a quantitative measure of disease progression one year after baseline measurements.

The dataset features include:

- Age
- Sex
- Body Mass Index (BMI)
- Blood pressure (BP)
- Serum measurements (S1, S2, S3, S4, S5, S6)

The dataset was loaded and converted into a Pandas DataFrame, with basic statistics computed to understand its distribution. It was then split into training (80%) and testing (20%) sets.

3. **Methodology:** A Linear Regression model was used as the base estimator for the RFE process. The steps involved in the implementation are as follows:

### 3.1 Data Loading and Exploration

- The dataset was loaded and converted into a Pandas DataFrame.
- Descriptive statistics were computed to understand feature distributions.
- The data was split into training (80%) and testing (20%) sets.

### 3.2 Train Linear Regression Model

- A LinearRegression model was trained on the training dataset.
- The model's performance was evaluated using the R² score on the test set.
- R² Score of Linear Regression Model: 0.4526

### 3.3 Implement Recursive Feature Elimination (RFE)

- RFE was applied with LinearRegression as the estimator.
- The model iteratively eliminated the least important features based on coefficients.
- The R² score was recorded at each step to determine the optimal number of features.

#### Optimal Number of Features:

The optimal number of features was identified as 2, where the R² score stabilized, and further reduction negatively impacted model performance.

#### Top 3 Most Important Features (Based on Coefficients):

- **S1**: 931.49
- **S5**: 736.20
- **BMI**: 542.43
- **Visualization**
- A plot illustrating the relationship between the R² score and the number of retained features is generated.
- The optimal feature count is identified based on an R² improvement threshold of 0.01.
- The analysis determines that the optimal number of features is 2.

- A comparison between initial and final feature rankings is performed:
  - **Initial Ranking:** BMI, S5, S1, S2, BP, Sex, S4, S3, S6, Age
  - **Final Selected Features:** The most significant features retained in the final iteration.

## 6. Results

- The optimal number of features is identified as 2.
- The key predictors of diabetes progression are **S1, S5, and BMI**.
- The elimination process enhances model interpretability while preserving predictive accuracy.

**5. Conclusion** This RFE analysis effectively identified the most critical features influencing diabetes progression. The process highlighted that reducing the feature set to the optimal number improves model performance without overfitting. S1, S5, and BMI were the key drivers of diabetes progression predictions, emphasizing the importance of metabolic health indicators in disease management.Overall, RFE proved to be a robust feature selection technique, offering clear insights into the dataset while maintaining high model accuracy.