**Objective**

The goal of this assignment is to understand Recursive Feature Elimination (RFE) and its role in feature selection. We analyze how RFE, combined with linear regression, helps identify the most significant features in the Diabetes dataset.

---

**1. Dataset Exploration**

We use the Diabetes dataset from sklearn.datasets.load_diabetes(). It consists of 10 features representing patient attributes and a target variable that measures disease progression after one year.

**Features:**

- age: Patient's age
- sex: Gender
- bmi: Body mass index
- bp: Average blood pressure
- Six other blood serum measurements (s1 to s6)

**Target Variable:**

- A continuous variable indicating diabetes progression severity.

We split the dataset into 80% training and 20% testing sets using train_test_split().

---

**2. Linear Regression Model**

A linear regression model was trained on the training set. Model performance was evaluated using the $R^2$ score on the test set.

Baseline Model Performance:

- $R^2$ Score on Test Set: *[0.4526]*

---

**3. Recursive Feature Elimination (RFE)**

We applied RFE using Linear Regression as the base estimator.

- Starting with all 10 features, we iteratively eliminated the least important one at each step.
- At each iteration, we tracked the $R^2$ score and feature coefficients.
- The $R^2$ score vs. number of features retained was visualized.
- We identified the optimal number of features based on significant $R^2$ improvement.

**Key Observations:**

- The $R^2$ score remained stable after removing certain features, indicating they were not highly informative.
- **Top 3 selected features**: **BMI, S1, S5**

---

**4. Feature Importance Analysis**

We created a table tracking feature coefficients at each RFE iteration.

**Most Important Features:**

1. **BMI** (Body Mass Index) – Strongly associated with diabetes progression due to obesity-related insulin resistance.
2. **S1** (Serum Cholesterol) – Indicates metabolic health, linking lipid levels to diabetes.
3. **S5** (LDL Cholesterol) – Correlates with insulin resistance, impacting disease severity.

**Comparison:**

- Initial ranking included features like age and bp, but they were not retained in the final set.
- Final selection prioritized metabolic-related features over demographic factors.

---

**5. Reflection**

1. **What did you learn about feature selection using RFE?**
    - RFE helps remove redundant features while retaining the most predictive ones, improving interpretability.
2. **How does RFE compare to LASSO?**

- **RFE** explicitly removes features based on importance scores.
- **LASSO (L1 Regularization)** shrinks coefficients to zero, inherently selecting features.
- RFE provides stepwise selection, while LASSO enforces sparsity through penalties.

3. **Insights from Selected Features:**
   - Metabolic factors (BMI, S1, S5) are primary diabetes indicators.
   - Age and blood pressure were less significant compared to cholesterol and BMI.
   - Managing weight and lipid levels can reduce diabetes progression risk.

---

**Conclusion**

RFE effectively identified the most influential features, improving model interpretability. The strongest predictors were BMI, cholesterol, and triglycerides, emphasizing the importance of metabolic health in diabetes progression.

---