Recursive Feature Elimination (RFE) Analysis Report

1. Introduction This report presents an analysis of the Diabetes dataset using Recursive Feature Elimination (RFE) for feature selection. The objective is to identify the most important features that contribute to predicting disease progression.

---

2. Dataset Exploration

- The dataset consists of **442 instances** and **10 numerical features**.
- The target variable represents **disease progression one year after baseline**.
- Features include **age, sex, body mass index (BMI), blood pressure (BP), and six blood serum measurements**.

---

3. Linear Regression Model

- A **linear regression model** was trained on an 80-20 train-test split.

---

4. Recursive Feature Elimination (RFE) Process

- RFE was applied using the linear regression model as the base estimator.
- Features were iteratively eliminated until only one remained.
- The R^2 score was tracked at each iteration to assess the impact of feature selection.
- A visualization of R^2 scores as a function of the number of retained features is shown below:

---

5. Feature Importance Analysis

A DataFrame is created to show the ranking and coefficient values at each iteration.

The top three most important features are identified and analyzed:

s1 (Serum Cholesterol Level): 931.49

s5 (Log Serum Triglycerides Level): 736.20

BMI (Body Mass Index): 542.43

Initial and final feature rankings are compared:

Initial Ranking: BMI, s5, s1, s2, bp, sex, s4, s3, s6, age

Final Selected Features: All features retained after final iteration

---

6. Comparison with Other Feature Selection Methods

- **RFE** systematically eliminates features, making it useful for identifying optimal subsets.
- Compared to **LASSO**, which applies L1 regularization to shrink coefficients to zero, RFE allows direct ranking and stepwise elimination.
- **Key Insight:** RFE provides a clear, interpretable ranking of feature importance, while LASSO is more efficient for high-dimensional datasets.

---

7. Conclusion

- **RFE helped identify the most predictive features** for diabetes progression.
- **The optimal number of features was determined** by tracking R^2 score improvements.
- **Feature importance analysis provided insights** into how medical attributes contribute to disease progression.
- Future work could involve testing other models such as **Ridge regression** or **Random Forest** for comparison.