Assignment 1: Recursive Feature Elimination with Linear Regression

Objective

This report explores the application of Recursive Feature Elimination (RFE) for feature selection in a linear regression model. The goal is to identify the most influential features in predicting diabetes progression using the Diabetes dataset from scikit-learn.

Dataset Overview

The Diabetes dataset consists of 10 numerical features and a target variable representing disease progression after one year. The features include:

- **age**: Patient's age
- **sex**: Patient's gender
- **bmi**: Body mass index
- **bp**: Average blood pressure
- **s1-s6**: Six blood serum measurements

Methodology

Data Exploration

- Loaded the dataset using sklearn.datasets.load_diabetes().
- Examined dataset statistics using describe() and info().
- Split the data into 80% training and 20% testing sets.

Linear Regression Model

- Trained a linear regression model on the training set.
- Evaluated the model using the R^2 score.
- **Initial R^2 Score**: 0.4523.

Recursive Feature Elimination (RFE)

- Implemented RFE with linear regression as the base estimator.
- Iteratively removed the least important feature, tracking R^2 score.
- Identified the optimal number of features using an R^2 improvement threshold (0.01).

- **Optimal Number of Features**: 10.

Results

Visualization and Findings

- A graph was generated to show the relationship between R^2 score and the number of retained features.
- The R^2 score remained relatively stable as features were eliminated, indicating that some features had little impact on model performance.
- The optimal number of features was determined to be **10**, as the R^2 improvement threshold (0.01) did not justify removing additional features.
- The graph suggests that the model can maintain predictive accuracy with all features included, supporting their collective significance.

Feature Importance Analysis

- Ranked features based on importance in each iteration.
- Identified the top three most important features:
    1. **bmi** (542.428759)
    2. **bp** (347.703844)
    3. **s5** (736.198859)
- Compared the initial feature ranking with the final selected features.

Reflection

Key Takeaways

- RFE effectively identifies the most relevant features, enhancing model interpretability.
- **bmi, bp, and s5** were found to be the most significant predictors of diabetes progression, aligning with medical insights on metabolic and cardiovascular factors.
- Unlike LASSO, which reduces coefficients to zero, RFE explicitly removes less important features.
- The findings from the plot suggest that feature elimination does not drastically improve R^2 score, reinforcing the importance of all features.

Conclusion

- The model achieved optimal performance with **10 features**.
- The selected features provided better interpretability and efficiency compared to using a reduced set of features.

Supporting Materials

- **Visualization of R^2 Score vs. Number of Features**
- **Feature Ranking Table**
- **Final Selected Features List**