Assignment 2

Fumiya Nagatomo

February 20, 2026

1 Overview

This report explores feature selection techniques, specifically Recursive Feature Elimination (RFE), applied to the diabetes dataset. The goal is to identify the most important features and determine the optimal number of features based on model performance.

2 Dataset Description

The diabetes dataset from scikit-learn contains 442 samples and 10 numerical features. The features include demographic variables (age and sex), clinical measurements such as BMI and blood pressure, and six blood samples variables. All features are standardized with means close to zero.

The target variable is continuous and represents disease progression after one year, making this a regression problem.

3 Methodology

RFE was applied using Linear Regression as the base model. At each step, the least important feature (based on coefficient magnitude) was removed and the model was refitted by using the remaining features. Model performance was evaluated using the R^2 score.

A threshold of 0.01 was used to determine significant improvement in R^2.

4 Results

The highest R^2 value was achieved when six features were selected. Removing additional features beyond this point led to a noticeable decrease in performance. The top three most important features identified were bmi, s1, and s5, with bmi remaining as the final selected feature in RFE.

5 Discussion

The results indicate that clinical features play a more important role in predicting diabetes progression than demographic variables (age and sex). BMI consistently showed strong predictive power.