Recursive Feature Elimination (RFE) with Linear Regression

Diabetes Dataset (scikit-learn)

Gaurav Patil
CSCI 485

Feb 23, 2026

1 Objective

This report investigates Recursive Feature Elimination (RFE) using linear regression to perform feature selection on the scikit-learn Diabetes dataset. The goal is to analyze how predictive performance changes as features are iteratively removed and to identify an optimal subset that balances model interpretability and generalization performance.

2 Dataset and Experimental Setup

The Diabetes dataset consists of 442 observations and 10 standardized predictors: `age, sex, bmi, bp, s1--s6`. The target variable measures disease progression one year after baseline.

The dataset was split into training and testing sets using an 80/20 split with `random_state=42` for reproducibility:

$$\text{Train: } (353, 10), \quad \text{Test: } (89, 10).$$

Model performance was evaluated using the coefficient of determination (R^2) on the test set.

3 Baseline Linear Regression

Using all 10 features, the baseline model achieved:

- Train $R^2 = 0.5279$
- Test $R^2 = 0.4526$

The model explains approximately 45% of the variance in unseen data. The largest baseline coefficients (by absolute magnitude) were associated with `s1`, `s5`, `bmi`, `s2`, and `bp`, suggesting that metabolic and physiological variables dominate prediction compared to demographic variables such as `age`.

4 Recursive Feature Elimination

RFE was applied using linear regression as the base estimator. Starting with all 10 features, the algorithm iteratively removed the least important feature (based on coefficient magnitude) and retrained the model until only one feature remained. At each step, the selected feature subset and test R^2 were recorded.

4.1 Elimination Path

RFE summary:

	k_features	r2_test	selected_features
0	10	0.452603	age, sex, bmi, bp, s1, s2, s3, s4, s5, s6
1	9	0.458659	sex, bmi, bp, s1, s2, s3, s4, s5, s6
2	8	0.455901	sex, bmi, bp, s1, s2, s3, s4, s5
3	7	0.458255	sex, bmi, bp, s1, s2, s4, s5
4	6	0.462777	sex, bmi, bp, s1, s2, s5
5	5	0.438201	bmi, bp, s1, s2, s5
6	4	0.446404	bmi, s1, s2, s5
7	3	0.445095	bmi, s1, s5
8	2	0.452293	bmi, s5
9	1	0.233350	bmi

Figure 1: RFE elimination path (k, test R^2, selected features).

4.2 Performance vs. Number of Features

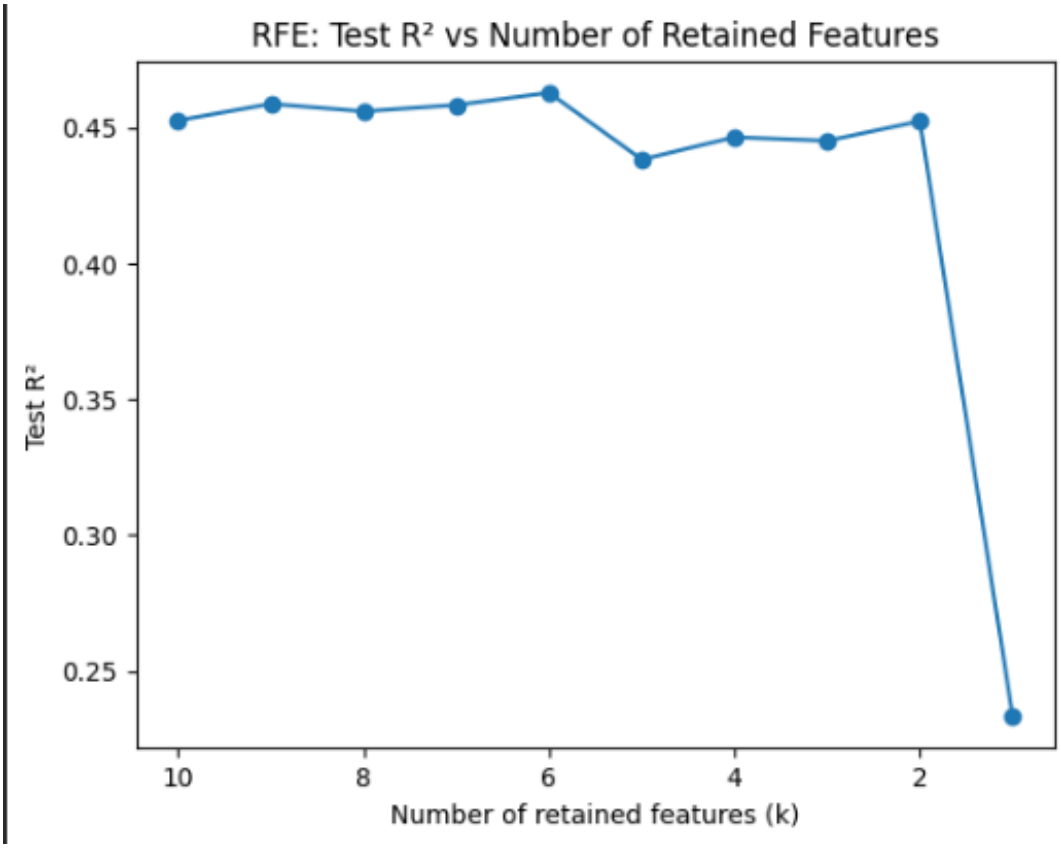

Figure 2: Test R^2 as a function of retained features ($10 \rightarrow 1$).

Test performance peaks at:

$$k^* = 6 \quad (R^2 = 0.4628),$$

which slightly improves upon the full 10-feature model ($R^2 = 0.4526$).

5 Optimal Feature Selection

A practical significance threshold of 0.01 in marginal R^2 improvement was used to guide selection. The elimination path shows that increasing beyond six features does not provide a meaningful gain in predictive performance, while reducing below six features eventually leads to underfitting.

Therefore, the selected model retains:

$$\{\texttt{sex, bmi, bp, s1, s2, s5}\}.$$

This model achieves the highest observed test R^2 while reducing dimensionality by 40%.

6 Coefficient Evolution Across RFE

Coefficient table:

	k_features	age	sex	bmi	bp	s1	s2	s3	s4	s5	s6
0	10	37.904021	-241.964362	542.428759	347.703844	-931.488846	518.062277	163.419983	275.317902	736.198859	48.670657
1	9	0.000000	-236.649588	542.799508	354.211438	-936.350589	528.796592	167.800414	270.396514	744.447429	53.350483
2	8	0.000000	-233.754686	550.744365	363.791753	-947.823133	541.585796	172.250588	277.741072	761.921177	0.000000
3	7	0.000000	-235.364224	551.866448	362.356114	-660.643160	343.348089	0.000000	185.140764	664.774591	0.000000
4	6	0.000000	-215.267423	557.314167	350.178667	-851.515734	591.093315	0.000000	0.000000	803.121285	0.000000
5	5	0.000000	0.000000	597.892739	306.647913	-655.560612	409.622184	0.000000	0.000000	728.643647	0.000000
6	4	0.000000	0.000000	691.460102	0.000000	-592.977874	362.950323	0.000000	0.000000	783.168538	0.000000
7	3	0.000000	0.000000	737.685594	0.000000	-228.339889	0.000000	0.000000	0.000000	680.224653	0.000000
8	2	0.000000	0.000000	732.109021	0.000000	0.000000	0.000000	0.000000	0.000000	562.226535	0.000000
9	1	0.000000	0.000000	998.577689	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000

Figure 3: Coefficient values across RFE iterations (eliminated features set to zero).

The coefficient table illustrates how parameter estimates shift as features are removed. As correlated predictors are eliminated, remaining coefficients often increase in magnitude. This reflects redistribution of explanatory variance among retained variables, a common phenomenon when multicollinearity is present.

7 Feature Importance Analysis

7.1 Selected Model Coefficients ($k = 6$)

Feature	Coefficient
sex	-215.27
bmi	557.31
bp	350.18
s1	-851.52
s2	591.09
s5	803.12

Table 1: Coefficients for the selected $k = 6$ model.

7.2 Three Most Important Features

Based on absolute coefficient magnitude, the three most influential predictors are:

`s1`, `s5`, `bmi`.

BMI (`bmi`). BMI is directly related to obesity and metabolic stress, both strongly associated with diabetes progression. Its large positive coefficient indicates that higher BMI corresponds to increased disease progression within one year.

Serum Measure `s1`. `s1` has the largest absolute coefficient in the selected model. This suggests that this biochemical marker captures substantial information about disease severity. The

sign of the coefficient represents its conditional relationship given the presence of other correlated serum variables.

Serum Measure `s5`. `s5` also exhibits a large magnitude and remains consistently important across RFE iterations. The dominance of serum variables highlights the importance of biochemical markers in predicting short-term disease outcomes.

Why `s2` is not ranked in the top three. Although `s2` has a relatively large coefficient, it ranks fourth by absolute magnitude. Serum variables are often correlated, meaning that predictive information may be shared among them. As features are removed, coefficient values can redistribute. Thus, while `s2` remains important, it does not exceed the magnitude of `s1`, `s5`, or `bmi` in the final model.

7.3 Baseline vs. Final Selection

The strongest predictors in the baseline model (`s1`, `s5`, `bmi`) remain in the optimal subset. In contrast, `age`, `s3`, `s4`, and `s6` are eliminated earlier in the process. This suggests that certain demographic and serum variables contribute less incremental predictive value under a linear model for one-year progression.

8 Reflection

8.1 What Was Learned About RFE

RFE demonstrates that removing weaker or redundant features can slightly improve generalization performance. In this case, reducing from 10 to 6 features increased test R^2, indicating that some predictors introduced noise or redundancy. However, aggressive elimination eventually reduces predictive power, as shown by the significant drop in performance when only one feature remains. This highlights the tradeoff between model simplicity and explanatory capacity.

8.2 RFE vs. LASSO

RFE is a wrapper method that repeatedly fits the model and removes features sequentially. This provides a transparent elimination path but requires multiple model fits. LASSO is an embedded method that performs feature selection through L_1 regularization during a single optimization process. LASSO is computationally more efficient and automatically shrinks some coefficients to zero, but the selected features depend on the chosen regularization parameter. In contrast, RFE explicitly evaluates model performance at each feature count, making the selection process more interpretable.

8.3 Insights About the Dataset

The selected model emphasizes the importance of metabolic and physiological variables in predicting short-term diabetes progression. BMI and blood-based serum markers consistently dominate the model, while demographic features such as age are eliminated early. This suggests that biochemical measurements capture more relevant variation in disease severity than demographic information for the one-year prediction horizon.

9 Conclusion

Using RFE with linear regression, the optimal model retained six features and achieved $R^2 = 0.4628$, slightly outperforming the full model while improving interpretability. The most influential predictors were `s1`, `s5`, and `bmi`, reinforcing the role of metabolic biomarkers and body composition in predicting diabetes progression.

Reproducibility: All results are generated by running `Assignment2_485.ipynb` with `random_state=42`.