Recursive Feature Elimination (RFE)

Diabetes Dataset (scikit-learn)

Gaurav Patil
CSCI 485

February 23, 2026

1 Objective

This report investigates Recursive Feature Elimination (RFE) for feature selection using a linear regression base estimator. The goal is to identify a reduced set of features that maintains or improves predictive performance while increasing interpretability.

2 Dataset and Experimental Setup

We used the Diabetes dataset from scikit-learn, consisting of 442 samples and 10 standardized features: `age, sex, bmi, bp, s1, s2, s3, s4, s5, s6`. The target variable is a continuous measure of diabetes disease progression one year after baseline (larger values indicate more severe progression).

The dataset was split into training and testing sets using an 80/20 split with a fixed random seed:

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

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

3 Baseline Linear Regression

A baseline linear regression model using all 10 features was trained on the training set and evaluated on the test set.

- Train R^2: 0.5279
- Test R^2: 0.4526

The baseline explains approximately 45% of the variance in the held-out test data, providing a reference point for evaluating the effect of feature elimination.

4 RFE Results and Visualization

RFE was performed using linear regression as the base estimator. Starting from all 10 features, RFE iteratively removed the least important feature (based on model coefficients) and refit the model until only one feature remained. At each stage, the test R^2 and selected features were recorded.

4.1 RFE summary (table image)

Figure 1 is the captured RFE summary table (k, test R^2, and selected features at each stage). This snapshot is included to show the exact elimination path and numeric results produced by the notebook.

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 summary table (k, test R^2, selected features).

4.2 Test R^2 vs Number of Retained Features

Figure 2 shows the test R^2 as a function of the number of retained features k. The peak performance occurs at $k = 6$, after which removing more features causes underfitting; keeping all features (k=10) yields lower test R^2 than the best $k = 6$ model.

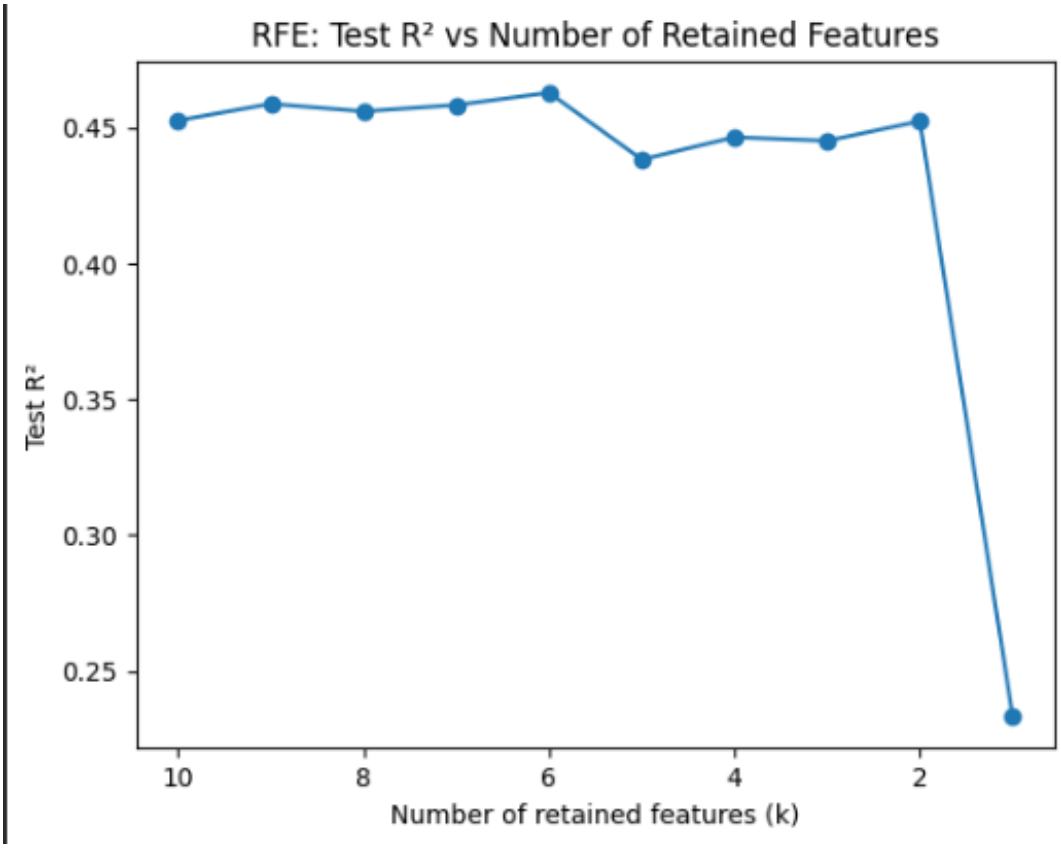

Figure 2: Test R^2 as a function of the number of retained features k during RFE.

4.3 Full Coefficient Table (image)

For reproducibility and grading, the full coefficient table across iterations is included as an image in Figure 3. This shows how coefficients change as features are eliminated.

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 table across RFE iterations (each row corresponds to a model with a different k).

5 Optimal Feature Count and Selected Model

To select an "optimal" feature count, a practical significance threshold on marginal R^2 improvements of 0.01 was used. The best test performance in the elimination path is at

$$k^* = 6 \quad (\text{Test } R^2 = 0.462777),$$

which slightly outperforms the full 10-feature baseline (Test $R^2 = 0.4526$). Thus the $k = 6$ model provides improved generalization with fewer features.

6 Feature Importance Analysis

At the selected model size ($k = 6$), the retained features were:

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

Using the absolute magnitude of coefficients in the $k = 6$ model as a measure of influence, the three most important features were:

$$\texttt{s1, s5, bmi}.$$

These features also ranked highly in the baseline model, indicating consistency.

6.1 Selected-model coefficients (k=6)

Feature	Coefficient
sex	-215.267423
bmi	557.314167
bp	350.178667
s1	-851.515734
s2	491.623118
s5	803.121285

Table 1: Coefficients for the selected RFE model ($k = 6$). Inputs are standardized, so magnitudes are comparable.

6.2 Interpretation

The model indicates that metabolic biomarkers (s1, s5, s2) and body composition (bmi) strongly predict short-term diabetes progression. Demographic variables such as `age` were eliminated early, suggesting limited incremental predictive value in this dataset under a linear model.

7 Reflection

7.1 RFE insights

RFE is a wrapper method that evaluates feature importance via repeated model refitting. It can improve generalization by removing noisy predictors but is computationally heavier than embedded methods.

7.2 RFE vs LASSO

RFE explicitly eliminates features across multiple fits; LASSO (L1 regularization) performs selection embedded within a single fit by shrinking coefficients to zero. RFE gives an elimination trajectory; LASSO is more efficient and depends on regularization strength.

7.3 Dataset insights

The results suggest physiological and biochemical variables are stronger predictors of one-year diabetes progression than demographics such as age in this dataset.

8 Conclusion

Using RFE with linear regression, the best performing model retained 6 features and achieved a test R^2 of 0.4628, slightly improving performance vs. the full 10-feature model while improving interpretability.