MATH/CSCI 485 — Assignment #2
Name: **VINAY PATIL**

Elimination (RFE) with Linear Regression

Dataset: Diabetes (sklearn.datasets.load_diabetes). Train/test split: 80/20 with random_state=42.

Task 1 — Dataset Exploration

Samples: 442, Features: 10. Target: continuous disease progression measure.

Features: age, sex, bmi, bp, s1, s2, s3, s4, s5, s6 (standardized).

Task 2 — Baseline Linear Regression

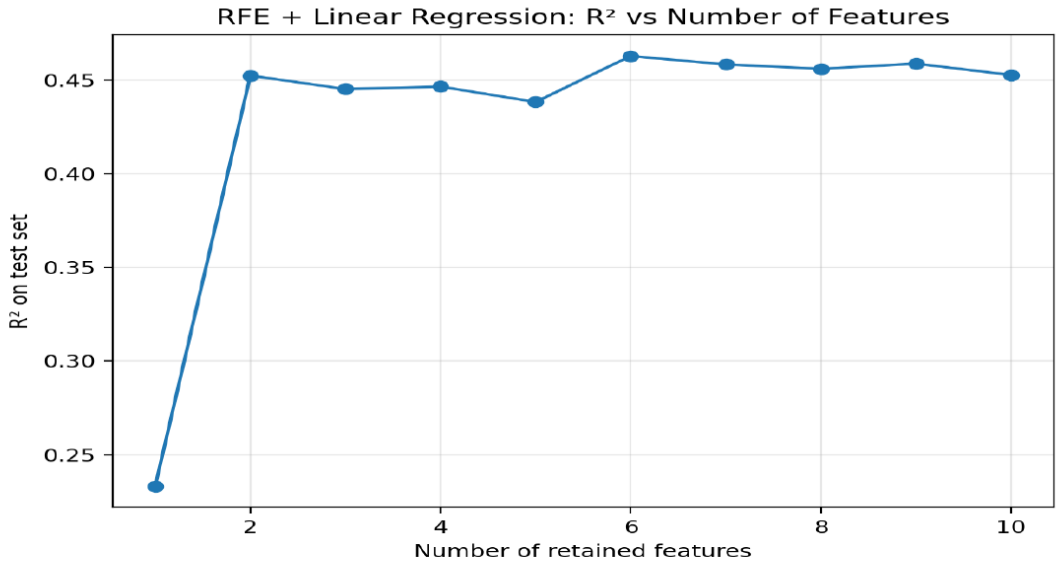
Baseline test R^2 using all 10 features: 0.452603

Task 3 — RFE Implementation (step=1)

- Uses `sklearn.feature_selection.RFE` with LinearRegression, step=1
- Runs k=10 → 1
- Tracks:
- Selected features at each k
- Test R^2 at each k
- Coefficients table across iterations
- Plot: R^2 vs #features

For k = 10 down to 1, run RFE to select k features, refit LinearRegression on selected features, and evaluate

test R^2.



R² at each iteration

#Features (k)	Test R²	Selected Features
1	0.233350	bmi
2	0.452293	bmi, s5
3	0.445095	bmi, s1, s5
4	0.446404	bmi, s1, s2, s5
5	0.438201	bmi, bp, s1, s2, s5
6	0.462777	sex, bmi, bp, s1, s2, s5
7	0.458255	sex, bmi, bp, s1, s2, s4, s5
8	0.455901	sex, bmi, bp, s1, s2, s3, s4, s5
9	0.458659	sex, bmi, bp, s1, s2, s3, s4, s5, s6
10	0.452603	age, sex, bmi, bp, s1, s2, s3, s4, s5, s6

Task 3.5 — Choose an optimal number of features

Best R^2 observed: 0.462777 at k=6. Using threshold=0.01, choose the smallest k with $R^{2\,3}$ (best - 0.01) =

0.452777. Optimal k = 6 with R^2 = 0.462777.

Selected features at optimal k=6: sex, bmi, bp, s1, s2, s5

Task 4 — Feature Importance Analysis

Initial ranking (all 10 features) by |coefficient| (largest first):

s1 (931.489), s5 (736.199), bmi (542.429)

Interpretation (high-level):

• Coefficient magnitude indicates how strongly each standardized feature contributes in a linear model. • RFE

removes the weakest feature (by estimator ranking) and refits repeatedly; selected set can change with k. • In

this split, bmi and blood-serum related variables (especially s1/s5) consistently matter most.

Coefficients by iteration (0 indicates feature not selected at that k)

Feature	k=10	k=9	k=8	k=7	k=6	k=5	k=4	k=3	k=2	k=1
age	37.904	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
sex	-241.964	-236.650	-233.755	-235.364	-215.267	0.000	0.000	0.000	0.000	0.000
bmi	542.429	542.800	550.744	551.866	557.314	597.893	691.460	737.686	732.109	998.578
bp	347.704	354.211	363.792	362.356	350.179	306.648	0.000	0.000	0.000	0.000
s1	-931.489	-936.351	-947.823	-660.643	-851.516	-655.561	-592.978	-228.340	0.000	0.000
s2	518.062	528.797	541.586	343.348	591.093	409.622	362.950	0.000	0.000	0.000
s3	163.420	167.800	172.251	0.000	0.000	0.000	0.000	0.000	0.000	0.000
s4	275.318	270.397	277.741	185.141	0.000	0.000	0.000	0.000	0.000	0.000
s5	736.199	744.447	761.921	664.775	803.121	728.644	783.169	680.225	562.227	0.000
s6	48.671	53.350	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Task 5 — Reflection

1) What I learned about RFE: RFE is a wrapper method: it repeatedly trains a model and removes the least

useful feature, trading compute for an interpretable ranking.

2) RFE vs LASSO: LASSO (L1) is an embedded method that can drive coefficients to 0 in one optimization,

while RFE eliminates features iteratively based on model rankings. LASSO is sensitive to l; RFE is sensitive to

the base estimator and ranking stability.

3) Dataset insights: In this dataset split, BMI and blood serum measurements (notably s1/s5) are consistently

predictive of disease progression, while some other variables add little beyond them.