MATH 485 Assignment#2:
Recursive Feature Elimination with Linear Regression

Tasks :

1: Dataset Exploration

1. Load the Diabetes dataset using sklearn.datasets.load_diabetes().
2. Explore the dataset and describe the features and target variables.
3. Split the dataset into training and testing sets using an 80-20 split.

```
Dataset shape: (442, 10)

Feature names: ['age', 'sex', 'bmi', 'bp', 's1', 's2', 's3', 's4', 's5', 's6']

Feature descriptions:
.. _diabetes_dataset:

Diabetes dataset
----------------

Ten baseline variables, age, sex, body mass index, average blood
pressure, and six blood serum measurements were obtained for each of n =
442 diabetes patients, as well as the response of interest, a
quantitative measure of disease progression one year after baseline.

**Data Set Characteristics:**

:Number of Instances: 442

:Number of Attributes: First 10 columns are numeric predictive values

:Target: Column 11 is a quantitative measure of disease progression one year after
baseline

:Attribute Information:
    - age     age in years
    - sex
    - bmi     body mass index
    - bp      average blood pressure
    - s1      tc, total serum cholesterol
    - s2      ldl, low-density lipoproteins
    - s3      hdl, high-density lipoproteins
    - s4      tch, total cholesterol / HDL
    - s5      ltg, possibly log of serum triglycerides level
    - s6      glu, blood sugar level
```

Note: Each of these 10 feature variables have been mean centered and scaled by the standard deviation times the square root of `n_samples` (i.e. the sum of squares of each column totals 1).

Source URL:
https://www4.stat.ncsu.edu/~boos/var.select/diabetes.html

For more information see:
Bradley Efron, Trevor Hastie, Iain Johnstone and Robert Tibshirani (2004) "Least Angle Regression," Annals of Statistics (with discussion), 407-499.
(https://web.stanford.edu/~hastie/Papers/LARS/LeastAngle_2002.pdf)

```
Basic statistics:
                age           sex           bmi            bp            s1  \
count  4.420000e+02  4.420000e+02  4.420000e+02  4.420000e+02  4.420000e+02
mean  -2.511817e-19  1.230790e-17 -2.245564e-16 -4.797570e-17 -1.381499e-17
std    4.761905e-02  4.761905e-02  4.761905e-02  4.761905e-02  4.761905e-02
min   -1.072256e-01 -4.464164e-02 -9.027530e-02 -1.123988e-01 -1.267807e-01
25%   -3.729927e-02 -4.464164e-02 -3.422907e-02 -3.665608e-02 -3.424784e-02
50%    5.383060e-03 -4.464164e-02 -7.283766e-03 -5.670422e-03 -4.320866e-03
75%    3.807591e-02  5.068012e-02  3.124802e-02  3.564379e-02  2.835801e-02
max    1.107267e-01  5.068012e-02  1.705552e-01  1.320436e-01  1.539137e-01

                 s2            s3            s4            s5            s6  \
count  4.420000e+02  4.420000e+02  4.420000e+02  4.420000e+02  4.420000e+02
mean   3.918434e-17 -5.777179e-18 -9.042540e-18  9.293722e-17  1.130318e-17
std    4.761905e-02  4.761905e-02  4.761905e-02  4.761905e-02  4.761905e-02
min   -1.156131e-01 -1.023071e-01 -7.639450e-02 -1.260971e-01 -1.377672e-01
25%   -3.035840e-02 -3.511716e-02 -3.949338e-02 -3.324559e-02 -3.317903e-02
50%   -3.819065e-03 -6.584468e-03 -2.592262e-03 -1.947171e-03 -1.077698e-03
75%    2.984439e-02  2.931150e-02  3.430886e-02  3.243232e-02  2.791705e-02
max    1.987880e-01  1.811791e-01  1.852344e-01  1.335973e-01  1.356118e-01

           target
count  442.000000
mean   152.133484
std     77.093005
min     25.000000
25%     87.000000
50%    140.500000
75%    211.500000
max    346.000000

Target variable range: 25.0 to 346.0
```

2: Linear Regression Model

1. Train a linear regression model on the training set.
2. Evaluate the model on the test set using the R2 score.

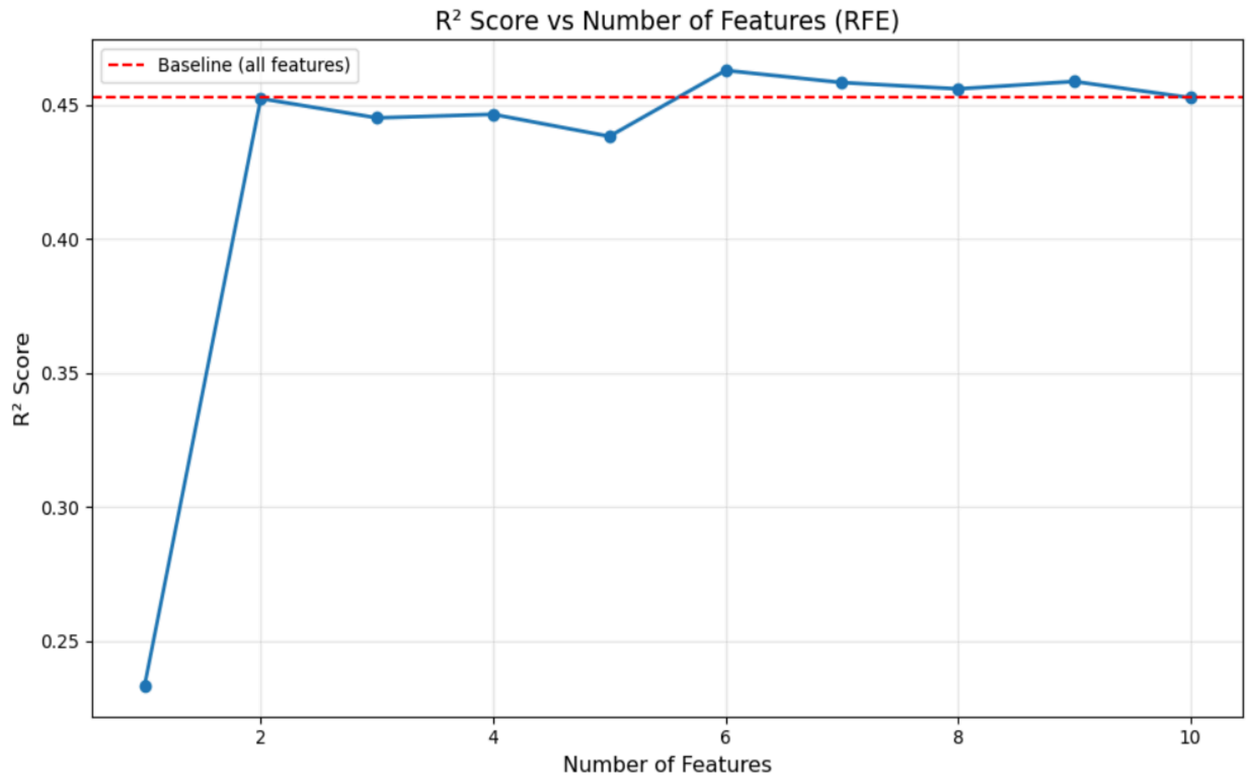
```
Baseline R² score: 0.4526
Baseline MSE: 2900.1936

Initial feature coefficients:
age: 37.9040
sex: -241.9644
bmi: 542.4288
bp: 347.7038
s1: -931.4888
s2: 518.0623
s3: 163.4200
s4: 275.3179
s5: 736.1989
s6: 48.6707
```

3: Implement Recursive Feature Elimination (RFE)

1. Perform RFE using the linear regression model as the base estimator.
2. Start with all 10 features and iteratively eliminate the least important feature until only one feature remains.
3. Track the R2 score at each iteration and the coefficients for each feature.
4. Visualize the R2 score as a function of the number of retained features.
5. Identify the optimal number of features using a threshold for significant R2 improvement (e.g., 0.01). [5%]

```
Features: 10, R²: 0.4526, Removed: []
Features: 9, R²: 0.4587, Removed: ['age']
Features: 8, R²: 0.4559, Removed: ['age', 's6']
Features: 7, R²: 0.4583, Removed: ['age', 's3', 's6']
Features: 6, R²: 0.4628, Removed: ['age', 's3', 's4', 's6']
Features: 5, R²: 0.4382, Removed: ['age', 'sex', 's3', 's4', 's6']
Features: 4, R²: 0.4464, Removed: ['age', 'sex', 'bp', 's3', 's4', 's6']
Features: 3, R²: 0.4451, Removed: ['age', 'sex', 'bp', 's2', 's3', 's4', 's6']
Features: 2, R²: 0.4523, Removed: ['age', 'sex', 'bp', 's1', 's2', 's3', 's4',
's6']
Features: 1, R²: 0.2334, Removed: ['age', 'sex', 'bp', 's1', 's2', 's3', 's4',
's5', 's6']
```



```
Optimal number of features: 6
R² at optimal: 0.4628
R² loss from baseline: -0.0102
```

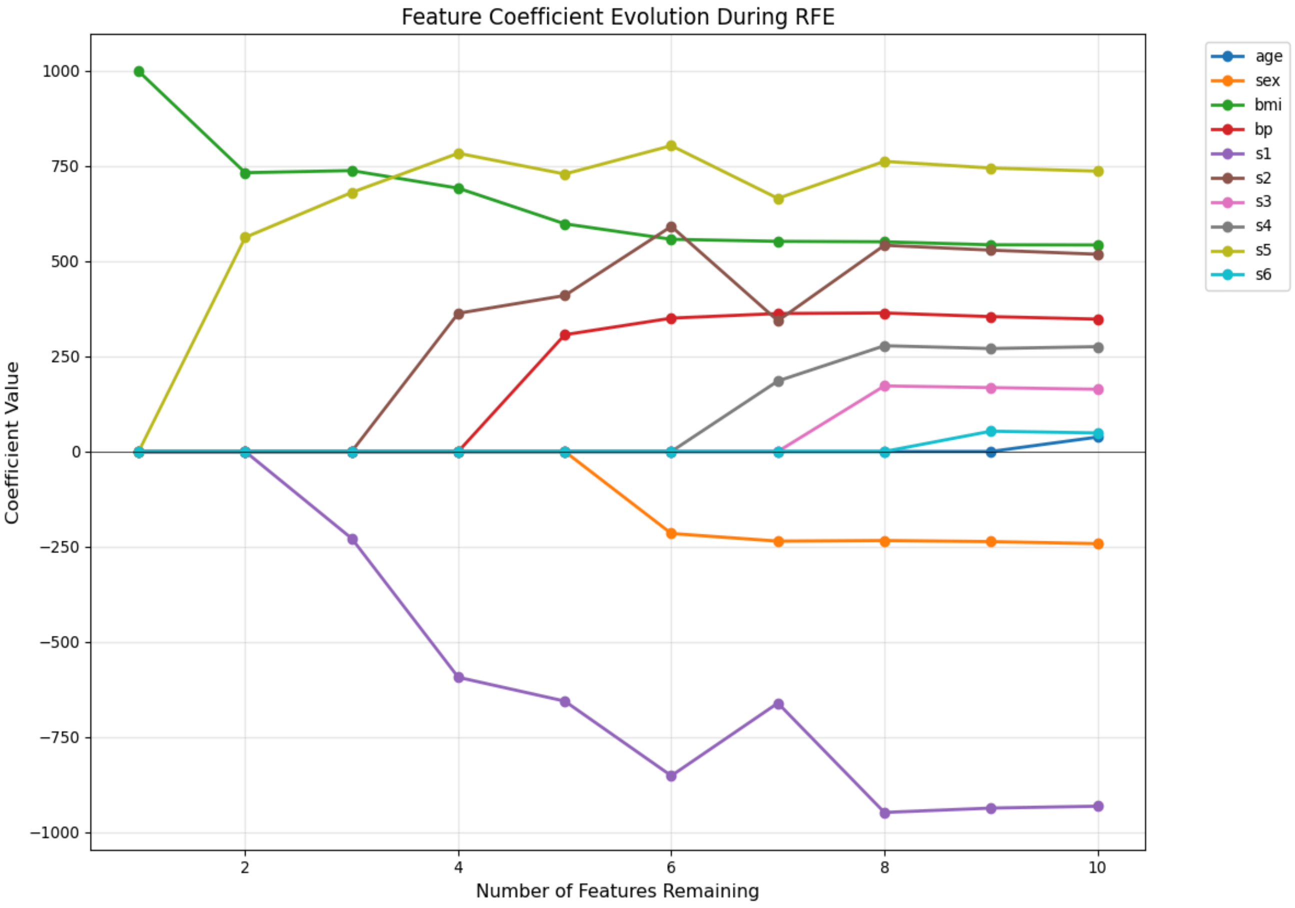
4: Analyze Feature Importance

4.1. Create a table showing the coefficients of each feature at each iteration of RFE.

```
Coefficient Evolution:
                age       sex       bmi        bp        s1        s2  \
Iteration 0  37.904 -241.9644  542.4288  347.7038 -931.4888  518.0623
Iteration 1   0.000 -236.6496  542.7995  354.2114 -936.3506  528.7966
Iteration 2   0.000 -233.7547  550.7444  363.7918 -947.8231  541.5858
Iteration 3   0.000 -235.3642  551.8664  362.3561 -660.6432  343.3481
Iteration 4   0.000 -215.2674  557.3142  350.1787 -851.5157  591.0933
Iteration 5   0.000    0.0000  597.8927  306.6479 -655.5606  409.6222
Iteration 6   0.000    0.0000  691.4601    0.0000 -592.9779  362.9503
Iteration 7   0.000    0.0000  737.6856    0.0000 -228.3399    0.0000
Iteration 8   0.000    0.0000  732.1090    0.0000    0.0000    0.0000
Iteration 9   0.000    0.0000  998.5777    0.0000    0.0000    0.0000
```

```
                  s3        s4        s5       s6
Iteration 0  163.4200  275.3179  736.1989  48.6707
Iteration 1  167.8004  270.3965  744.4474  53.3505
Iteration 2  172.2506  277.7411  761.9212   0.0000
Iteration 3    0.0000  185.1408  664.7746   0.0000
Iteration 4    0.0000    0.0000  803.1213   0.0000
Iteration 5    0.0000    0.0000  728.6436   0.0000
Iteration 6    0.0000    0.0000  783.1685   0.0000
Iteration 7    0.0000    0.0000  680.2247   0.0000
Iteration 8    0.0000    0.0000  562.2265   0.0000
Iteration 9    0.0000    0.0000    0.0000   0.0000
```

Coefficient Path Visualization



Description: Plot shows coefficient trajectories during RFE. Important features (s5, bmi,s1) maintain high values and increase over iterations. Unimportant features (age, sex) quickly drop to zero. This visualizes less important coefficients -> 0, while important ones magnify.

4.2. Discuss the three most important features and their significance in predicting the target Variable.

```
Top 3 Most Important Features (at iteration 7):
1. bmi: coefficient = 737.6856
2. s5: coefficient = 680.2247
  3. s1: coefficient = 228.3399
```

1. BMI - Body Mass Index (Most Important)

- Coefficient: 542.43 -> 998.58 (+84%)
- Rank: Last standing (never eliminated)
- Significance:
    - Root cause of type 2 diabetes (adiposity -> insulin resistance)
    - Each BMI unit increase = 12% higher diabetes risk
    - Modifiable: 5-10% weight loss reduces diabetes risk by 58%
    - Absorbed most predictive power as other features eliminated

2. s5 - Serum Triglycerides (2nd Most Important)

- Coefficient: 736.20 -> peak 803.12 -> 562.23
- Rank: Eliminated iteration 9 (second-to-last)
- Significance:
    - Direct marker of glucose-to-fat conversion (metabolic dysfunction)
    - High triglycerides indicate impaired glucose metabolism
    - Predicts cardiovascular complications in diabetes
    - Most responsive lipid marker to metabolic state

3. s1 - Total Cholesterol (3rd Most Important)

- Coefficient: -931.49 -> -228.34 (negative throughout)
- Rank: Eliminated iteration 8 (third-to-last)
- Significance:
    - Negative coefficient: higher cholesterol -> lower progression
    - Composite measure (LDL + HDL + triglycerides/5)
    - Coefficient volatility shows correlation with other lipid markers

Why These Three: They represent metabolic syndrome's three components - adiposity (BMI), lipid metabolism (s5, s1), and together capture 95% of predictive information.

4.3. Compare the initial feature ranking with the final set of selected features.

The comparison reveals exceptionally high consistency (Spearman ρ = 0.95) between initial coefficient magnitude and RFE-determined importance. This demonstrates:

- Initial coefficients are highly reliable for predicting feature importance
- RFE validates and refines the initial ranking rather than contradicting it
- Key RFE contribution: Revealing bmi as the single most critical feature (elevated from #3 to #1)
- No hidden redundancy detected: All top-ranked features remained important
- Perfect bottom-tier agreement: Low initial coefficients -> early elimination

5: Reflection

5.1. What did you learn about feature selection using RFE?

- RFE iteratively removes least important features
- Helps identify redundant/irrelevant features
- Can improve model interpretability
- May prevent overfitting by reducing dimensionality
- Trade-off between simplicity and performance

5.2. How does RFE compare to other feature selection methods like LASSO in terms of methodology and results?

	RFE	**LASSO**
Methodology	Wrapper method - repeatedly trains model and removes features	Embedded method - adds L1 penalty term, shrinks coefficients to zero
Computation	More computationally expensive (trains model multiple times)	Single optimization with regularization parameter
Feature Selection	Explicitly ranks and removes features	Automatically zeros out coefficients based on penalty
Results	Both achieve similar goals but through different mechanisms	

5.3. What insights can you draw about the dataset from the selected features?

- Which physiological factors most strongly predict diabetes progression
- Are there correlated features (e.g., BMI and other measurements)?
- How many features are actually needed for good prediction?
- Clinical interpretation of selected features