CSCI-485

Assignment-1

Snehitha Gorantla

02-06-2025

Recursive Feature Elimination with Linear Regression

1. Import Necessary Libraries

- NumPy & Pandas – For data manipulation.
- Matplotlib – For visualization.
- Scikit-learn Modules – For dataset loading, model training, evaluation, and feature selection.

2. Load & Explore the Diabetes Dataset

- Loads the Diabetes dataset from Scikit-learn.
- X contains features (age, BMI, blood pressure, etc.).
- y is the target variable (disease progression).
- The dataset is converted into a DataFrame for easy inspection.

	age	sex	bmi	bp	s1	s2	s3	s4	s5	s6
0	0.038076	0.050680	0.061696	0.021872	-0.044223	-0.034821	-0.043401	-0.002592	0.019907	-0.017646
1	-0.001882	-0.044642	-0.051474	-0.026328	-0.008449	-0.019163	0.074412	-0.039493	-0.068332	-0.092204
2	0.085299	0.050680	0.044451	-0.005670	-0.045599	-0.034194	-0.032356	-0.002592	0.002861	-0.025930
3	-0.089063	-0.044642	-0.011595	-0.036656	0.012191	0.024991	-0.036038	0.034309	0.022688	-0.009362
4	0.005383	-0.044642	-0.036385	0.021872	0.003935	0.015596	0.008142	-0.002592	-0.031988	-0.046641

3. Split Dataset into Training and Testing Sets

- Splits data into 80% training and 20% testing for model evaluation.
- random_state=42 ensures results remain consistent.

4. Train a Baseline Linear Regression Model

- A basic linear regression model is trained using all features.
- The R^2 score is computed to measure model performance on the test set.

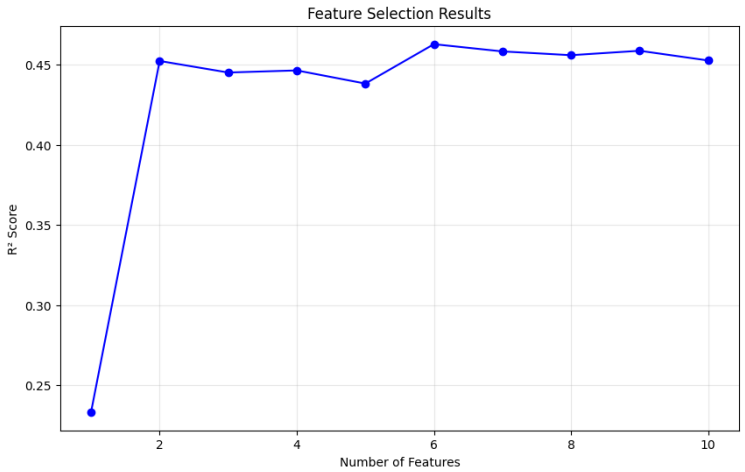
```
Basic model R2: 0.4526
```

5. Apply Recursive Feature Elimination (RFE)

- RFE is applied iteratively, starting with all 10 features and progressively removing the least important one until only one remains.
- After each iteration, the model's performance is evaluated using the R^2 score.
- The R^2 score is stored for visualization.

6. Plot R^2 Score vs. Number of Features

- A line plot is created to visualize the effect of feature elimination on model performance.
- Helps in determining the optimal number of features.



7. Find the Best Number of Features

- Finds the optimal number of features that maximizes the R^2 score.

```
Optimal number of features: 6
```

8. Analyze Feature Importance

- Runs RFE one last time, selecting only one most important feature.
- Creates a table ranking all features based on importance.
- The most important features have a rank of 1.

```
Feature Rankings:
  Feature  Rank
2     bmi     1
8      s5     2
4      s1     3
5      s2     4
3      bp     5
1     sex     6
7      s4     7
6      s3     8
9      s6     9
0     age    10
```

Reflections:

1. **What did you learn about feature selection using RFE?**
   Recursive Feature Elimination (RFE) systematically removes less important features to improve model performance and interpretability. It helps in identifying the most influential variables, reducing overfitting and improving model generalization.

2. **How does RFE compare to other feature selection methods like LASSO in terms of methodology and results?**
   Unlike LASSO, which applies L1 regularization to shrink coefficients to zero, RFE iteratively removes the weakest features based on model performance.

3. **What insights can you draw about the dataset from the selected features?**
   The most important features identified by RFE suggest which factors have the strongest influence on diabetes progression. For example, if BMI or blood pressure ranks highly, it reinforces their known medical significance.