CSCI 485 Assignment#1: Recursive Feature Elimination with Linear Regression

Objective

The goal of this assignment is to understand the concept of Recursive Feature Elimination (RFE) and its application in feature selection. You will explore how RFE works in conjunction with linear regression to determine the most important features of a dataset and improve model interpretability.

Dataset

We will use the **Diabetes dataset** available in `scikit-learn`. This dataset contains 10 features and a target variable representing a quantitative measure of disease progression one year after baseline.

Dataset Description

- Features:
  - `age`: Age of the patient.
  - `sex`: Gender of the patient.
  - `bmi`: Body mass index.
  - `bp`: Average blood pressure.
  - Six other quantitative variables derived from blood samples.
- Target: A continuous variable measuring the disease progression within a one-year period. Larger values generally correspond to more severe progression of diabetes.

Tasks

Task 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.

Task 2: Linear Regression Model

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

Task 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 R^2 score at each iteration and the coefficients for each feature.
4. Visualize the R^2 score as a function of the number of retained features.
5. Identify the optimal number of features using a threshold for significant R^2 improvement (e.g., 0.01).

Task 4: Analyze Feature Importance

1. Create a table showing the coefficients of each feature at each iteration of RFE.
2. Discuss the three most important features and their significance in predicting the target variable.
3. Compare the initial feature ranking with the final set of selected features.

Task 5: Reflection

Answer the following questions:

1. What did you learn about feature selection using RFE?
2. How does RFE compare to other feature selection methods like LASSO in terms of methodology and results?
3. What insights can you draw about the dataset from the selected features?

Deliverables

1. Python code or Jupyter Notebook with the complete implementation in a Github project (more in the Submission Requirement Section)
2. A brief report (1-2 pages) in PDF summarizing your findings, including visualizations and tables.

Hints

- Use `sklearn.feature_selection.RFE` for implementing Recursive Feature Elimination.
- Use `matplotlib` for visualizations.

- Pay attention to the interpretation of the R^2 score and the significance of feature coefficients.

Grading Rubric

1. **Dataset Exploration (10%)**: Completeness and clarity in exploring and describing the dataset.
2. **Linear Regression (20%)**: Correct implementation and evaluation of the baseline model.
3. **RFE Implementation (30%)**: Proper implementation of RFE and visualization of results.
4. **Analysis (30%)**: Insightful discussion of feature importance and comparison of results.
5. **Report (10%)**: Clarity, conciseness, and presentation of findings.

Submission Requirements

- Create a Github project that includes the following:
    - Name of your repository: MATH/CSCI485_Spring25_<Firstname>_<Lastname>
    - Within your repository, create project: Assignment_1
    - Within your project, include the following:
        - Source code (either python code, or python code in jupyter notebook)
        - A readme file describes how to use your code
        - A screenshot of how your code is successfully executed and generating required outputs, graphs and tables etc.
        - A PDF file of the report
- Submission:
    - On Canvas Assignments, PDF report and link to your Github repo/projects
    - Make sure you have invited me to your github repo/projects

Additional Notes:

For submission of github links, you need to either make your github repo public, or add me as a collaborator. Here is an instruction on adding people as collaborators.

https://docs.github.com/en/account-and-profile/setting-up-and-managing-your-personal-account-on-github/managing-access-to-your-personal-repositories/inviting-collaborators-to-a-personal-repository

You can find me on github using my email: bshen2@csuchico.edu or my github account name: boshen-csuchico.