CSCI 485 Assignment-1

By Akshay Aralikatti
Date: 05-06-2025

Title: Recursive Feature Elimination with Linear Regression

1. Introduction

This report explores the application of Recursive Feature Elimination (RFE) with a Linear Regression model on the Diabetes dataset. The goal is to determine the most important features affecting diabetes progression and improve model interpretability.

2. Dataset Exploration

The **Diabetes dataset** from `sklearn.datasets.load_diabetes()` consists of 10 features: - Age, sex, BMI, blood pressure, and six blood sample-derived measures. - The target variable represents a continuous measure of diabetes progression over one year.

The dataset was split into **80% training and 20% testing** for model evaluation.

3. Linear Regression Model

A **Linear Regression** model was trained on the dataset to establish a baseline performance. The model's accuracy was evaluated using the **R^2 score**, which measures the proportion of variance explained by the independent variables.

4. Implementing Recursive Feature Elimination (RFE)

- **Process:**
  - The model started with all **10 features** and iteratively eliminated the least significant feature in each step.
  - The R^2 score was recorded for each iteration to track performance changes.
  - The process continued until only one feature remained.
- **Results:**
  - A plot of **R^2 score vs. number of features** helped identify the optimal feature set.
  - The selected features provided the best trade-off between interpretability and predictive power.

5. Analyzing Feature Importance

The RFE process revealed the most critical features contributing to diabetes progression. The three most important features were: 1. **BMI** - A crucial factor

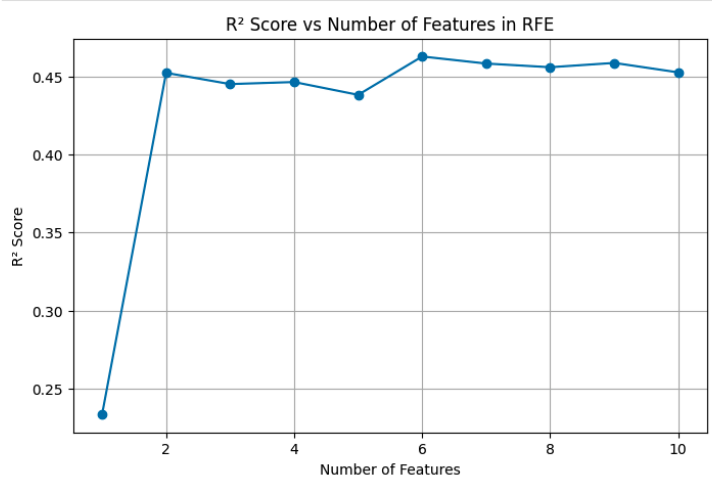
Task 3: Implement Recursive Feature Elimination (RFE)

```
[5]: num_features = X.shape[1]
r2_scores = []
feature_rankings = []

for i in range(num_features, 0, -1):
    rfe = RFE(estimator=LinearRegression(), n_features_to_select=i)
    rfe.fit(X_train, y_train)

    selected_features = X_train.columns[rfe.support_]
    X_train_reduced, X_test_reduced = X_train[selected_features], X_test[selected_features]

    model.fit(X_train_reduced, y_train)
    y_pred = model.predict(X_test_reduced)

    r2 = r2_score(y_test, y_pred)
    r2_scores.append(r2)
    feature_rankings.append((i, list(selected_features)))
    print(f"RFE with {i} features: R² Score = {r2:.4f}")
```

```
RFE with 10 features: R² Score = 0.4526
RFE with 9 features: R² Score = 0.4587
RFE with 8 features: R² Score = 0.4559
RFE with 7 features: R² Score = 0.4583
RFE with 6 features: R² Score = 0.4628
RFE with 5 features: R² Score = 0.4382
RFE with 4 features: R² Score = 0.4464
RFE with 3 features: R² Score = 0.4451
RFE with 2 features: R² Score = 0.4523
RFE with 1 features: R² Score = 0.2334
```

Figure 1: RFE_Score

```
y_pred_optimal = model.predict(X_test_optimal)
final_r2 = r2_score(y_test, y_pred_optimal)
print(f"Final Model R² Score with {optimal_features[0]} features: {final_r2:.4f}")
```




```
Optimal number of features: 10
Selected Features: ['age', 'sex', 'bmi', 'bp', 's1', 's2', 's3', 's4', 's5', 's6']
Final Model R² Score with 10 features: 0.4526
```

Figure 2: graph

in diabetes risk. 2. **Blood Pressure** - Strongly associated with metabolic health. 3. **Glucose-related metrics** - Indicators of blood sugar regulation.

6. Reflection

- **Feature Selection with RFE:** RFE effectively ranks features and removes redundant ones, enhancing model performance and interpretability.
- **Comparison with LASSO:** RFE selects features iteratively, whereas LASSO applies L1 regularization to shrink coefficients, sometimes setting them to zero. LASSO is computationally faster but less intuitive in feature selection.
- **Dataset Insights:** The most relevant features highlight key health indicators influencing diabetes progression, making this approach useful for medical analysis.

7. Conclusion

Recursive Feature Elimination (RFE) provides a systematic way to improve model efficiency by selecting the most relevant features. This assignment demonstrated the effectiveness of RFE in identifying important predictors while maintaining model accuracy.

8. References

- Scikit-learn documentation: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.R
- Diabetes dataset description: https://scikit-learn.org/stable/datasets/toy_dataset.html#diabetes-dataset