DS Lab Assignment 2

Group members: Md Masum Billah, Vinay Sanga, Ammara Asif

Title: Comparative analysis of Linear Regression vs. Random Forest on different feature selection methods on Boston housing dataset.

Features Selection Methods Employed:

1. **Original Dataset**: Using all features.
2. **Feature Importance Filtering**: Removing features based on their importance scores from the Random Forest model.
3. **Correlation-Based Filtering**: Removing features based on their correlation with the target variable and inter-feature correlation.

Results:

- **Original Dataset**: Linear Regression achieved R^2 of 68.94% while Random Forest achieved R^2 of 89.08%.
- **Feature Importance Filtering**: Linear Regression achieved R^2 of 63.98%, and Random Forest achieved R^2 of 88.82%.
- **Correlation-Based Filtering**: Linear Regression achieved R^2 of 61.90%, and Random Forest achieved R^2 of 87.21%.

A comparative result table is given below:

Features selection method	**The accuracy (as measured by the coefficient of determination R^2)**	
	Linear Regression	**Random Forest**
Original Dataset	68.94%	89.08%
Feature Importance Filtering	63.98%	88.82%
Correlation-Based Filtering	61.90%	87.21%

Discussion and Comparison:

- Across all data filtering methods, the Random Forest algorithm consistently outperformed Linear Regression.
- Both algorithms experienced a decline in performance when features were removed, but the decline was sharper for Linear Regression.
-

Pros and Cons:

- **Linear Regression**:
  - **Pros**: Simpler model, easier to interpret, and faster to train.
  - **Cons**: Lower accuracy compared to Random Forest across all feature sets. More sensitive to feature removal.

- **Random Forest**:
    - **Pros**: Higher accuracy across all feature sets. Can capture complex non-linear relationships.
    - **Cons**: More complex model, potentially harder to interpret, and requires more computational resources.

Overall Choice:
Considering both accuracy and model simplicity, the **Random Forest algorithm on the original dataset** is the preferred choice. It provided the highest accuracy and, despite its complexity, its ability to capture intricate patterns in the data justifies its use.

Brief Analysis and Overall Results:
The exploration underscored the robustness of the Random Forest algorithm in handling various feature sets. While Linear Regression offers simplicity, its performance, especially after feature removal, was notably inferior to Random Forest. Feature selection is a critical aspect of modeling, but it's evident that the choice of algorithm can play a substantial role in how much information is retained or lost during this process. For the Boston Housing dataset, Random Forest on the full set of features emerged as the most potent combination for predictive accuracy.