Evaluation of Supervised Learning Algorithms on the Breast Cancer Dataset

Overview:

The goal of this assessment is to apply supervised learning techniques to classify breast cancer data into malignant or benign. The dataset is from sklearn and contains features like mean radius, texture, and perimeter, which are useful for diagnosis.

1. Loading & Pre-processing

- The dataset is loaded, checked for missing values, and scaled.
- **Why scaling?**
  Algorithms like SVM and k-NN are distance-based, so all features must contribute equally.
- **Why check for missing values?**
  Missing data could skew results or cause errors.

2. Implementing Classification Algorithms:

- **Logistic Regression:** Predicts probabilities; great for interpretable binary classification.
- **Decision Tree:** Splits data based on feature thresholds; prone to overfitting.
- **Random Forest:** Combines many decision trees; reduces overfitting and increases accuracy.
- **SVM:** Finds the best hyperplane to separate classes; works well in high-dimensional space.
- **k-NN:** Classifies based on the nearest data points; simple but computationally heavy for large datasets.

Each model is trained, tested, and evaluated on accuracy.

3. Model Comparison:

- **Accuracy is used** as the performance metric.
- Results are plotted for a clear visual comparison.

- The best and worst-performing models are identified, with insights into why they perform that way (e.g., robustness vs sensitivity to data).

Each algorithm has unique strengths and weaknesses. Comparing them ensures you choose the best fit for the data and task, providing insights into the model's suitability based on performance metrics.

This structured approach helps demonstrate practical machine learning skills and an understanding of algorithmic behavior.