DA5401 A8: Ensemble Learning for Complex Regression Modeling on Bike Share Data

Objective: This assignment will challenge you to apply and compare three primary ensemble techniques (**Bagging, Boosting, and Stacking**) to solve a complex, time-series-based regression problem. You will demonstrate your understanding of how these methods address model variance and bias, and how a diverse stack of models can yield superior performance to any single model.

---

1. Problem Statement

You are a data scientist for a city's bike-sharing program. Accurate forecasting of bike rentals is critical for managing inventory and logistics. Predicting the total count of rented bikes (cnt) is a complex regression task influenced by multiple factors, such as weather, time of day, and season, with non-linear relationships and high variability.

You will use the **Bike Sharing Demand Dataset**, which contains over **17,000 hourly samples**. Your task is to implement three distinct ensemble strategies and evaluate their effectiveness in minimizing the prediction error (RMSE).

You will submit a Jupyter Notebook with your complete code, visualizations, and a plausible story that explains your findings. The notebook should be well-commented, reproducible, and easy to follow.

Dataset:

- **Bike Sharing Demand Dataset (Hourly Data):** Over 17,000 samples.
    - **Citation:** Fanaee-T, Hadi, and Gamper, H. (2014). *Bikeshare Data Set*. UCI Machine Learning Repository.
    - **Download Link (UCI ML Repository):** [UCI Machine Learning Repository - Bike Sharing Dataset]

---

2. Tasks

Part A: Data Preprocessing and Baseline [10 points]

1. **Data Loading and Feature Engineering:**
    - Load the hour.csv file. The target variable is the **total count of bike rentals** (cnt).
    - Drop irrelevant columns like instant, dteday, casual, and registered.
    - Convert categorical features (e.g., season, weathersit, mnth, hr) into a numerical format suitable for regression models (e.g., **One-Hot Encoding**).
2. **Train/Test Split:** Split the preprocessed data into training and testing sets.
3. **Baseline Model (Single Regressor):**

- Train a single **Decision Tree Regressor** (use a max depth of 6) and a single **Linear Regression** model on the training data.
- Evaluate both models on the test set using the **Root Mean Squared Error (RMSE)**. Use the better of the two single models as your **baseline performance metric**.

Part B: Ensemble Techniques for Bias and Variance Reduction [20 points]

1. **Bagging (Variance Reduction):**
   - **Hypothesis:** Bagging primarily targets **variance reduction**.
   - Implement a **Bagging Regressor** using the **Decision Tree Regressor** (from the baseline) as the base estimator. Use at least 50 estimators.
   - Calculate and report the **RMSE** on the test set. Discuss whether the bagging technique effectively reduced variance compared to the single Decision Tree baseline.
2. **Boosting (Bias Reduction):**
   - **Hypothesis:** Boosting primarily targets **bias reduction**.
   - Implement a **Gradient Boosting Regressor** (a robust, widely used boosting technique).
   - Calculate and report the **RMSE** on the test set. Discuss whether boosting achieved a better result than both the single model and the bagging ensemble, supporting the idea of bias reduction.

Part C: Stacking for Optimal Performance [10 points]

1. **Stacking Implementation:**
   - Explain the principle of **Stacking** and how the Meta-Learner learns to combine the predictions of diverse Base Learners optimally.
   - Define the following three **Base Learners** (Level-0):
     - **K-Nearest Neighbors Regressor** (KNeighborsRegressor)
     - **Bagging Regressor** (from Part B)
     - **Gradient Boosting Regressor** (from Part B)
   - Define the **Meta-Learner** (Level-1): Use a simple **Ridge Regression** model.
   - Implement a **Stacking Regressor** combining these base and meta learners.
2. **Final Evaluation:** Calculate and report the **RMSE** for the Stacking Regressor on the test set.

Part D: Final Analysis [10 points]

1. **Comparative Table:** Create a clear table summarizing the RMSE of all five models:
   - Baseline Single Model (Best of DT/Linear)

    - Bagging Regressor
    - Gradient Boosting Regressor
    - Stacking Regressor
2. **Conclusion:** Based on your results:
    - Identify the **best-performing model**.
    - Explain *why* the Stacking Regressor (or the best ensemble) outperformed the single model baseline, referencing the concepts of the **bias-variance trade-off** and **model diversity**.

---

3. Submission Guidelines

- The assignment is due on November 10th, 2025 (4 pm for on-prem; 2 pm for absentia).
- Submit a single Jupyter Notebook with all your code, analysis, and answers to the conceptual questions in markdown cells.
- Ensure all code is clean, readable, and reproducible.

Evaluation Criteria:

- Correct implementation and evaluation of Bagging, Boosting, and Stacking, including necessary data preprocessing.
- Clear discussion and demonstration of bias vs. variance reduction.
- Accurate calculation and comparison of **RMSE** across all models.
- Insightful final analysis and conclusion regarding ensemble effectiveness.

Good luck!