Week 3: Baseline Model Training and Cross-Validation

Tasks Completed:

1. **Model Implementation**:

- Trained 5 regression models using Scikit-learn and XGBoost:
  - **Linear Regression**: Default parameters.
  - **AdaBoost**: `n_estimators=50`, `learning_rate=0.1`.
  - **XGBoost**: `max_depth=3`, `learning_rate=0.2`.
  - **Gradient Boosting**: `n_estimators=100`, `max_depth=3`.
  - **Random Forest**: `n_estimators=100`, `max_depth=10`.

2. **Performance Evaluation**:
   - **Random Forest** achieved the highest test R^2 (**0.81**) and lowest MSE (**10.1**).
   - **Gradient Boosting** showed strong results ($R^2 = 0.87$) on merged dataset (Table 2).
   - **Linear Regression** underperformed ($R^2 = 0.54$) due to non-linear relationships.

3. **Cross-Validation**:
   - 5-fold cross-validation confirmed Random Forest's stability (CV $R^2 = 0.88$, CV MSE = 35).

Challenges Faced:

- Overfitting in XGBoost (training $R^2 = 1.0$, test $R^2 = 0.65$).
- High computational time for Gradient Boosting.

Outcomes:

- Identified **Random Forest** and **Gradient Boosting** as top candidates for tuning.
- Drafted **ML Model Evaluation** section with comparative tables (Tables 1–2).

Tasks Planned for Week 4:

1. Optimize models using GridSearchCV and RandomizedSearchCV.
2. Interpret model decisions via SHAP and permutation importance.
3. Develop an interactive UI for real-time predictions.