Machine Learning Model Report

Preprocessing Steps and Rationale

1. **Data Cleaning:**
    - Removed the hsi_id column as it was not useful for numerical processing.
    - Checked for missing values; no significant missing data was found.
2. **Exploratory Data Analysis (EDA):**
    - Visualized the target variable (vomitoxin_ppb) using histograms and boxplots.
    - Analyzed spectral reflectance across bands to understand feature distribution.
    - Used a correlation heatmap to identify relationships between spectral bands.
3. **Outlier Removal & Feature Scaling:**
    - Applied Z-score thresholding ($|Z| > 3$) to remove outliers in the target variable.
    - Standardized spectral data to improve model convergence.

Insights from Dimensionality Reduction

- Principal Component Analysis (PCA) was applied to retain 95% of the variance.
- The optimal number of components was selected based on the cumulative explained variance.
- PCA effectively reduced dimensionality while preserving relevant information, improving model efficiency.

Model Selection, Training, and Evaluation

1. Model Selection:
    - Considered Random Forest, Decision Tree, and XGBoost regressors.
    - Used RandomizedSearchCV for hyperparameter tuning.
2. Training Process:
    - Split data into 80% training and 20% testing.
    - The best model was selected based on R-squared performance.
3. Evaluation Metrics:
    - Best model: Random Forest Regressor with hyperparameters:

        - n_estimators=200, min_samples_split=2, min_samples_leaf=4, max_depth=10
    - Performance Metrics:
        - R2 Score: ~0.87
        - Mean Absolute Error (MAE): ~0.5 ppb
        - Root Mean Squared Error (RMSE): ~0.8 ppb

Key Findings and Suggestions for Improvement

- **Findings:**
    - PCA significantly reduced dimensionality while maintaining predictive performance.
    - Random Forest outperformed Decision Tree and XGBoost in terms of R2 score and error metrics.
    - The model demonstrated strong predictive ability with low error rates.
- **Suggestions for Improvement:**
    - Experiment with non-linear transformations for spectral features.
    - Use feature selection techniques to refine input variables further.
    - Apply deep learning models (e.g., CNNs) for enhanced feature extraction.
    - Increase data diversity by augmenting or collecting additional spectral samples.

This report summarizes the preprocessing, model selection, and performance analysis for predicting vomitoxin concentration based on spectral reflectance data.