Overview

This report summarizes the methodology, key insights, and findings from the analysis of hyperspectral imaging data to predict vomitoxin_ppb levels in corn samples. The process involves data preprocessing, dimensionality reduction, CNN model development with hyperparameter tuning, and performance evaluation.

Preprocessing Steps and Rationale

- **Data Cleaning:**
  The dataset was loaded and inspected. The non-numeric column (hsi_id) was dropped to prevent issues with numerical computations. Missing values were filled using the column mean, preserving the dataset size while ensuring data consistency.
- **Feature Standardization:**
  Spectral reflectance features were standardized using StandardScaler so that each feature has a mean of 0 and a standard deviation of 1. Standardization is crucial for both dimensionality reduction (using PCA) and training deep learning models.

Dimensionality Reduction Insights

- **Principal Component Analysis (PCA):**
  PCA was applied to the standardized data to explore the variance distribution across spectral bands. Retaining 95% of the variance required a reduced number of components. Visualization of the explained variance ratio and a 2D PCA plot revealed clustering patterns and provided insights into the data structure relative to vomitoxin_ppb levels.

Model Selection, Training, and Evaluation Details

- **Model Choice:**
  A 1D Convolutional Neural Network (CNN) was selected due to its effectiveness in capturing local spectral features from hyperspectral data. The architecture includes:
  - Two convolutional blocks (Conv1D layers with ReLU activations, followed by MaxPooling1D and Dropout for regularization).
  - A Flatten layer leading into Dense layers, culminating in a single neuron output for regression.
- **Hyperparameter Tuning:**
  The CNN model was wrapped using KerasRegressor from scikeras and tuned using GridSearchCV. The hyperparameter grid explored variations in:
  - Number of filters and kernel sizes.
  - Dropout rates and dense layer units.
  - Batch sizes and training epochs.
    This tuning process helped identify the optimal configuration for the best validation performance.
- **Model Evaluation:**
  The final model achieved the following evaluation metrics on the test set:
  - **MAE:** 3170.00

- **RMSE:** 9138.95
- **R^2:** 0.7012
  A scatter plot of actual versus predicted vomitoxin_ppb values illustrated that the model captures the overall trend well, though there remains some variability in individual predictions.

Key Findings and Suggestions for Improvement

- **Key Findings:**
  - The CNN effectively extracts local spectral features and captures the overall trend in vomitoxin_ppb.
  - Hyperparameter tuning improved model performance, but the performance is still limited by the small dataset size.
- **Suggestions for Improvement:**
  - **Advanced Architectures:** Exploring attention mechanisms or transformer-based models could help capture long-range dependencies across spectral bands.
  - **Data Augmentation:** Employing data augmentation or transfer learning might improve model generalizability.
  - **Alternative Tuning Strategies:** Using RandomizedSearchCV or Bayesian optimization can reduce computational overhead during hyperparameter tuning.
  - **Ensemble Methods:** Combining multiple models through ensemble techniques may enhance robustness and overall performance.