Hyperspectral Data Analysis for Vomitoxin Prediction in Corn Samples

1. Introduction

The objective of this study is to analyze hyperspectral imaging (HSI) data to predict the concentration of vomitoxin (in ppb) in corn samples. The dataset contains 455 samples with 449 spectral bands and one target column for vomitoxin concentration. The analysis involves data preprocessing, outlier treatment, data visualization, dimensionality reduction, and model training.

2. Dataset Information

The dataset consists of 500 rows and 450 columns. Each row represents a corn sample with 449 spectral band reflectance values and a target column 'vomitoxin_ppb'.

Summary Statistics:

- **Mean Reflectance:** Ranges from 0.44 to 0.74 across bands.
- **Standard Deviation:** Highest standard deviation observed in higher bands, indicating variability.
- **Minimum Reflectance:** Around 0.25.
- **Maximum Reflectance:** Around 0.94.
- **Vomitoxin_ppb:** Ranges from 0 to approximately 13000 ppb.

3. Outlier Treatment

To ensure data quality, outliers were treated using the Interquartile Range (IQR) method for each spectral band. Any values below Q1 - 1.5*IQR or above Q3 + 1.5*IQR were removed. This significantly improved the data distribution and reduced extreme reflectance values.

4. Data Visualization

4.1 Average Spectral Reflectance Across Bands

A line plot was created to observe the average spectral reflectance across all bands. The plot shows a steady increase in reflectance from lower bands, reaching a peak, and then gradually declining.

4.2 Heatmap of Spectral Reflectance

A heatmap was generated to visualize the spectral reflectance across all samples. The color intensity reflects the reflectance values, allowing us to identify patterns of high and low reflectance.

5. Data Preprocessing

The data was standardized using the StandardScaler from sklearn to ensure uniform scale across spectral bands. This step was essential for accurate PCA and model training.

6. Dimensionality Reduction using PCA

Principal Component Analysis (PCA) was performed to reduce the dimensionality of the data. Different component counts were tested, and the best results were obtained with 50 components, preserving maximum variance.

7. Model Training

A Random Forest Regressor model was trained using the PCA-reduced data. Hyperparameter tuning was conducted using GridSearchCV with parameters:

- **n_estimators:** [100, 200, 300]
- **max_depth:** [10, 20, 30]
- **min_samples_split:** [2, 5, 10]
- **min_samples_leaf:** [1, 2, 4]

The best model was found to be:

RandomForestRegressor(max_depth=30, random_state=42)

Additionally, an LSTM model was trained with the following architecture:

- **Input Layer:** Bidirectional LSTM with 256 units and LayerNormalization
- **Hidden Layer 1:** Bidirectional LSTM with 128 units and LayerNormalization
- **Hidden Layer 2:** Bidirectional LSTM with 64 units and LayerNormalization
- **Output Layer:** Dense with 1 neuron
- **Optimizer:** Adam
- **Loss Function:** Mean Absolute Error (MAE)

Training results for the LSTM model:

- **Final Train Loss:** 2303.34
- **Final Validation Loss:** 3647.39
- **Mean Absolute Error (MAE):** 3647.31
- **Root Mean Squared Error (RMSE):** 13541.86
- **R-squared (R2):** -0.07

8. Model Performance

For the optimal Random Forest model (with 50 PCA components):

- **Mean Absolute Error (MAE):** 2685.54
- **Root Mean Squared Error (RMSE):** 6706.52
- **R-squared (R2):** 0.74

For the LSTM model:

- **Mean Absolute Error (MAE):** 3647.31
- **Root Mean Squared Error (RMSE):** 13541.86
- **R-squared (R2):** -0.07

9. Conclusion

The analysis successfully demonstrated that hyperspectral imaging data can be effectively used to predict vomitoxin concentration in corn samples. By applying PCA for dimensionality reduction and Random Forest for regression, we achieved a reasonably high R2 score. The LSTM model, however, showed poor performance, indicating that it might not be suitable for this type of data without further feature engineering or architecture optimization.

Best Model: The Random Forest model with PCA (50 components) achieved the best performance with:

- **R-squared (R2):** 0.74
- **Mean Absolute Error (MAE):** 2685.54

Future work could explore advanced deep learning models, Transformer architectures, or ensemble learning techniques for improved prediction accuracy.