Hyperspectral Imaging for Mycotoxin Prediction in Corn

1. Introduction:

This report outlines the steps taken to preprocess hyperspectral imaging data, perform dimensionality reduction, train a machine learning model, and evaluate its performance in predicting DON concentration in corn samples.

2. Data Preprocessing:

2.1 Handling Missing Values

- The dataset was inspected for missing values and inconsistencies.
- Missing values in numerical columns were replaced with their mean to maintain data integrity.

2.2 Normalization

- StandardScaler was used to normalize the spectral reflectance values, ensuring that all features were on a similar scale to improve model performance.

2.3 Data Visualization

- Spectral reflectance was plotted to analyze the patterns across wavelength bands.
- This helped in understanding data distributions and variations across samples.

3. Dimensionality Reduction:

3.1 Principal Component Analysis (PCA)

- PCA was implemented to reduce feature dimensionality while retaining significant variance.
- The cumulative explained variance plot was used to determine the optimal number of principal components.
- A scatter plot of the first two principal components was generated to visualize data distribution.

4. Model Training:

4.1 Model Selection

- XGBoost Regressor was chosen for its efficiency in handling structured data and its ability to capture complex patterns.
- The dataset was split into 80% training and 20% testing.

4.2 Model Training and Hyperparameter Tuning

- The model was trained with 100 estimators and a learning rate of 0.1.

- Default hyperparameters were used for simplicity, but further tuning could improve performance.

5. Model Evaluation:

5.1 Performance Metrics

- Mean Absolute Error (MAE): Measures the average magnitude of errors.
- Root Mean Squared Error (RMSE): Evaluates error magnitude with greater sensitivity to large deviations.
- R^2 Score: Determines how well the model explains variance in the data.

5.2 Results

- The model's predictions were compared with actual values using a scatter plot.
- The performance metrics were calculated and analyzed.

6. Key Findings and Suggestions for Improvement:

6.1 Findings

- PCA effectively reduced dimensionality while preserving essential variance.
- XGBoost provided reasonable predictive performance, but some errors remain.

6.2 Suggested Improvements

- Further hyperparameter tuning (e.g., Grid Search, Random Search) to optimize performance.
- Experimenting with other models like Random Forest or Neural Networks.
- Exploring additional feature engineering techniques to enhance predictive power.

7. Conclusion:

This project successfully demonstrated data preprocessing, dimensionality reduction, machine learning model training, and evaluation. Future improvements could enhance accuracy and generalizability.