**Mycotoxin Prediction Using Hyperspectral Imaging Data**

**1. Introduction**

The primary objective of this project is to predict mycotoxin levels (specifically DON/vomitoxin) in corn samples using hyperspectral imaging data. Each corn sample has spectral reflectance values across multiple wavelength bands. We aim to preprocess and visualize the data, reduce its dimensionality, and train a machine learning model (in this case, a 1D Convolutional Neural Network) to predict the target variable.

**2. Data Exploration and Preprocessing**

1. **Data Loading:**
   * The dataset (TASK-ML-INTERN.csv) was loaded and inspected for missing values.
   * Column names were cleaned to remove trailing/leading spaces.
2. **Handling Missing Values & Inconsistencies:**
   * Any missing or invalid entries were identified using isnull().sum().
   * For this dataset, no significant missing-value issues were observed. If there were, appropriate imputation strategies would be applied.
3. **Feature Standardization:**
   * The spectral features were standardized using StandardScaler to ensure all bands contribute equally to the model training process.
4. **Data Visualization:**
   * A line plot of the **average spectral reflectance** was created to observe general trends across bands.
   * A **heatmap** of correlations among spectral bands provided insights into feature redundancy or grouping patterns.

**3. Dimensionality Reduction (PCA)**

1. **PCA Implementation:**
   * Principal Component Analysis (PCA) was applied to the standardized spectral data.
   * The top two principal components were used for visualization, and the **explained variance ratio** was reported.
2. **Insights:**
   * The scatter plot of the PCA-transformed data (colored by the target variable) revealed whether samples with similar toxin levels cluster together in the reduced space.
   * The explained variance helped identify how many components might be necessary to capture a significant portion of the data’s variability.

**4. Model Training**

1. **Model Selection:**
   * A **1D Convolutional Neural Network (CNN)** was chosen for its ability to capture local patterns in spectral data.
2. **Data Split & Reshaping:**
   * The dataset was split into **training (80%)** and **testing (20%)** sets.
   * The input features were reshaped to (samples, spectral\_bands, 1) to accommodate the 1D CNN architecture.
3. **CNN Architecture:**
   * Two convolutional layers (with Conv1D), each followed by max pooling, were used to extract local features from the spectral bands.
   * A fully connected layer with dropout (for regularization) was added before the final regression output neuron.
4. **Training Procedure:**
   * The model was compiled with the **Adam optimizer** and a **mean squared error (MSE)** loss function.
   * Training was conducted for a specified number of epochs (e.g., 50) with a validation split to monitor overfitting.

**5. Model Evaluation**

1. **Regression Metrics:**
   * **Mean Absolute Error (MAE)**
   * **Root Mean Squared Error (RMSE)**
   * **R² Score**
2. **Results Visualization:**
   * A scatter plot comparing **actual vs. predicted** mycotoxin levels helped illustrate the model’s predictive accuracy.
   * (Optional) If adapted for classification (by thresholding the continuous target), metrics such as Accuracy, Precision, Recall, F1-Score, and a confusion matrix could be examined.
3. **Performance Summary:**
   * The CNN captured the underlying relationships in the spectral data reasonably well, with moderate errors on the test set.
   * Minor discrepancies were noted in certain samples, suggesting room for further model tuning.

**6. Key Findings and Suggestions for Improvement**

1. **Findings:**
   * **PCA** helped visualize the data’s intrinsic structure and revealed that a few principal components capture a substantial fraction of variance.
   * The **1D CNN** effectively learned local spectral patterns but may benefit from further hyperparameter tuning (e.g., number of filters, kernel sizes, dropout rate).
2. **Limitations:**
   * Deep learning models can be **data-hungry**; more extensive datasets might further improve performance.
   * **Interpretability** is limited, as CNNs are often considered “black box” models.
3. **Future Work:**
   * **Hyperparameter Optimization:** Employ systematic search methods (grid search, random search) or advanced optimization algorithms to refine the CNN architecture.
   * **Alternative Architectures:** Explore transformers or attention mechanisms for better sequence modeling.
   * **Data Augmentation:** If data collection is expensive, synthetic or augmented spectral data could enhance model robustness.

**7. Conclusion**

This project demonstrates a complete pipeline for predicting mycotoxin levels from hyperspectral data: from preprocessing and PCA-based exploration to training and evaluating a CNN for regression. While the model shows promising results, further tuning and experimentation could yield more accurate and interpretable predictions. Overall, the approach provides a strong foundation for advanced spectral analysis in agricultural applications.