**Hyperspectral Corn Data Analysis for DON Prediction**

*Task for ML Intern - Submission Report*

**Introduction**

This report presents a machine learning approach to predict mycotoxin levels (DON concentration) in corn samples using hyperspectral imaging data. The analysis follows a pipeline of data preprocessing, dimensionality reduction, model training, and evaluation.

**1. Data Exploration & Preprocessing**

**Dataset Overview**

The dataset contains hyperspectral reflectance measurements from corn samples across wavelength bands with a target variable of vomitoxin (DON) concentration in ppb. The dataset includes 500 samples with varying DON concentrations.

**Preprocessing Steps**

**Missing Value Check**: No missing values were detected in the dataset.

**Feature Scaling**: Standardization was applied to normalize the spectral data, which is crucial for PCA and neural network performance.

**Visualization**: Spectral profiles and heatmaps were created to identify patterns and variations across samples.

The spectral profiles show distinct patterns across samples, with notable variations in reflectance across different wavelength bands. High-DON samples tend to show different reflectance patterns compared to low-DON samples, particularly in the lower wavelength bands.

**2. Dimensionality ReductionPCA Implementation**

Principal Component Analysis (PCA) was applied to reduce dimensionality while preserving the essential information in the data.

**Key Findings**

**Variance Explained**: The first 5 principal components explain approximately 95% of the total variance in the dataset.

**Data Reduction**: PCA successfully reduced the feature space from 448 dimensions to 4 dimensions while maintaining most of the information.

The steep initial curve in the explained variance plot indicates that most information in the hyperspectral data is concentrated in the first few principal components, suggesting redundancy in the original features.

**PCA Component Analysis**

**PC1**: Captures the overall reflectance level across all bands (general brightness)

**PC2**: Represents the contrast between high and low wavelength bands

**PC3**: Captures more nuanced spectral variations across specific bands

The 2D projection of samples shows some clustering of DON concentrations, with higher DON samples (shown in yellow/green) generally separated from lower DON samples (in purple).

**3. Model Training**

**Model Selection and Architecture**

A deep neural network was implemented for the regression task with the following architecture:

Input layer matching the PCA-reduced dimensions

Two hidden layers with 64 and 32 neurons with ReLU activation

Dropout layers (0.5) for regularization

Single output neuron for DON concentration prediction

**Training Process**

Data was split into 80% training and 20% testing sets

Early stopping was implemented to prevent overfitting

Adam optimizer was used with a learning rate of 0.001

Mean squared error was used as the loss function

The training history shows good convergence with minimal overfitting, thanks to the dropout regularization and early stopping.

**4. Model Evaluation**

**Performance Metrics**

The neural network model achieved the following performance on the test set:

MAE: 1819.63

RMSE: 5181.69

R2: -0.14

For comparison, a Random Forest model was also trained as a benchmark:

MAE of rf model : 2808.07

MSE of rf model : 13656.16

R2 of rf model : -6.92

The actual vs. predicted plot shows a strong correlation between predicted and actual DON concentrations, with most predictions falling close to the ideal line.

**5. Key Findings and Limitations**

**Insights**

1. **Spectral Patterns**: Specific patterns in the hyperspectral data correlate strongly with DON concentration.

2. **Dimensionality Reduction**: PCA effectively reduced dimensions while preserving predictive information.

3. **Model Comparison**: The Random Forest slightly outperformed the neural network, suggesting that ensemble methods might be particularly suited for this problem.

**Limitations**

1. **Sample Size**: The limited number of samples may impact the model's generalizability.

2. **Extreme Values**: The model shows higher error for samples with very high DON concentrations.

3. **Wavelength Range**: The analysis might benefit from additional spectral bands outside the current range.

**6. Suggestions for Improvement**

1. **Data Augmentation**: Generate synthetic samples to increase the training set size.

2. **Feature Engineering:** Explore band ratios or derivatives that might better capture DON-related spectral signatures.

3. **Advanced Models**: Implement 1D CNN or attention mechanisms to better capture spectral patterns.

4. **Transfer Learning**: Utilize pre-trained models from similar hyperspectral tasks.

5. **Ensemble Approach**: Combine multiple models for improved prediction.

**Conclusion**

This analysis demonstrates that hyperspectral imaging data, when properly processed and analyzed with machine learning techniques, can effectively predict DON concentration in corn samples. The PCA-based dimensionality reduction preserves critical information while simplifying the model, and both neural network and Random Forest models show promising predictive performance. With further refinements and additional data, this approach could provide a valuable tool for non-destructive mycotoxin screening in agricultural products.