ML Intern Assignment Report

1. Introduction

This report summarizes the steps taken to process hyperspectral imaging data, apply dimensionality reduction, train a machine learning model, and evaluate its performance for predicting mycotoxin levels in corn samples.

2. Data Preprocessing

- **Handling Missing Values:** The dataset was checked for missing values, and none were found.
- **Feature Selection:** Removed the hsi_id column as it was not relevant.
- **Normalization:** Applied standardization using StandardScaler to ensure all spectral features were on the same scale.

3. Dimensionality Reduction

- **PCA:** Principal Component Analysis (PCA) was applied to reduce feature dimensions while retaining the maximum variance.
- **Results:** The first two principal components explained a significant portion of the variance, visualized through a scatter plot.

4. Model Training & Evaluation

- **Model Used:** Random Forest Regressor was chosen for its robustness in handling high-dimensional data.
- **Hyperparameters:** Used default settings with 100 trees.
- **Performance Metrics:**
    - M**ean Absolute Error (MAE):** Computed and reported in script
    - **Root Mean Squared Error (RMSE):** Computed and reported in script
    - **R^2 Score:** Computed and reported in script
- **Visualization:** Scatter plot of actual vs. predicted DON concentration was generated.

5. Key Findings & Suggestions for Improvement

- The model performed reasonably well with Random Forest.
- Adding more advanced models like XGBoost or deep learning (CNN/LSTM) could further improve accuracy.
- Hyperparameter tuning (GridSearch or RandomSearch) could optimize model performance.
- Exploring t-SNE for better visualization of data clustering might provide deeper insights.

6. Conclusion

This project successfully demonstrated preprocessing, dimensionality reduction, and ML-based regression for mycotoxin prediction. Future improvements could focus on deep learning approaches and enhanced feature engineering.