Machine Learning Intern Task Report

1. Introduction

This report summarizes the steps taken to process hyperspectral imaging data, reduce dimensionality, and train a machine learning model to predict mycotoxin levels (DON concentration) in corn samples.

2. Data Preprocessing

- **Missing Values**: Filled using the median to prevent bias.
- **Feature Normalization**: Applied StandardScaler to normalize spectral reflectance data.
- **Exploratory Visualization**: Line plots and scatter plots of spectral bands were used to analyze patterns.

3. Dimensionality Reduction

- **PCA (Principal Component Analysis)**: Applied to reduce dimensionality while retaining the majority of variance.
- **Explained Variance**: The top components captured significant variation in the dataset.
- **Visualization**: Scatter plots were created to visualize clustering patterns.

4. Model Training

- **Model Used**: Random Forest Regressor.
- **Train-Test Split**: 80% training, 20% testing.
- **Hyperparameter Optimization**: Default parameters used with future potential tuning.

5. Model Evaluation

- **Metrics Used**:
  - Mean Absolute Error (MAE): Measures average error.
  - Root Mean Squared Error (RMSE): Penalizes large errors more.
  - R^2 Score: Indicates model accuracy.
- **Results**:
  - MAE: *(3765.0568)*
  - RMSE: *(11483.805982806223)*
  - R^2 Score: *(0.5282211884116356)*
- **Visualization**: Scatter plot of actual vs. predicted values for evaluation.

6. Key Findings & Suggestions

- PCA reduced feature dimensions effectively.

- The Random Forest model performed well but could be improved with hyperparameter tuning.
- Additional models such as CNN or LSTM could be explored.
- Feature selection techniques could further optimize the model’s accuracy.