DON Concentration Prediction Using Machine Learning

1. Project Overview

This project focuses on predicting DON concentration in corn samples using machine learning. DON contamination is a major concern in food safety and agriculture, and this model aims to provide accurate predictions to support quality control measures.

2. Dataset Description

- **Features:** Spectral reflectance values at different wavelengths.
- **Target Variable:** DON concentration (continuous numerical value).
- **Samples:** Each row represents a unique corn sample.

3. Data Preprocessing Steps & Rationale

- **Handling Missing Values:** Checked for null values and handled them.
- **Duplicate Removal:** Identified and removed any duplicate rows.
- **Feature Selection:** Dropped irrelevant columns like `hsi_id` to avoid redundancy.
- **Outlier Detection & Treatment:** Used histograms and boxplots to visualize distributions and identify extreme values.
- **Feature Correlation Analysis:** Selected top 10 features with the highest correlation to DON concentration.

4. Dimensionality Reduction Insights

- **PCA (Principal Component Analysis)** was applied to reduce feature dimensions while preserving variance.
- **Results:** The first few principal components explained most of the variance, helping in reducing noise and improving model efficiency.

5. Model Selection, Training & Evaluation

- **Algorithm Used:** MLPRegressor (Multi-layer Perceptron for regression tasks).
- **Hyperparameter Tuning:** Experimented with different hidden layers and activation functions.
- **Evaluation Metrics:**
  - **R^2 Score:** Measures how well the model explains variance in the data.
  - **Mean Squared Error (MSE):** Evaluates prediction accuracy.
- **Findings:** Model performed well on training data but showed some variance in predictions on test data.

6. Key Findings & Possible Improvements

- **Findings:**
  - Some extreme values caused a few prediction deviations.
  - Model captured general trends but struggled with high DON concentrations.