DON Concentration Prediction on Corn Samples

The project aims to predict the concentration of vomitoxin (DON) in corn samples using hyperspectral imaging data. The dataset comprises spectral reflectance values measured at hundreds of wavelengths for each corn sample, along with the corresponding DON concentration. It involves data preprocessing, regression modeling using a neural network and model evaluation.

Preprocessing Steps:

1. Data Loading and Inspection - The dataset was loaded from a CSV file. Spectrals columns (all columns except hsi_id and vomitoxin_ppb) were identified and converted to numeric values for consistency in data types
2. Handling Missing Values - Missing values in numeric columns were imputed using mean imputation to maintain dataset size while minimizing bias
3. Data Standardization - The spectral features were standardized using StandardScaler to center the data at zero with unit variance for faster and more stable convergence

Model Selection, Training and Evaluation:

1. Model Selection: A custom neural network regression model was implemented in PyTorch. The model comprises an input layer (matching the number of features), two hidden layers (with 128 and 64 neurons respectively), dropout for regularization, and an output layer with one neuron.
2. Training: The model was trained using the Adam optimizer with a learning rate set to 0.002. A training loop with early stopping was implemented based on the validation loss to prevent overfitting. Experiments included adjusting the learning rate and number of epochs (with early stopping typically triggering at 163 epochs out of a maximum of 350).
3. Evaluation:

- Metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), R^2 Score, and Median Absolute Error were computed.
- Compared to a baseline model (which simply predicts the mean DON value), our model achieved significant improvements (e.g., MAE reduced from 6432.880 to 3815.786, R^2 improved to 0.609)
- Visualizations (scatter plots and residual plots) confirmed that the model performed well for most samples though some high-outlier predictions remain a challenge

Key Findings:

1. The model explains roughly 61% of the variance in DON concentration
2. The majority of predictions are accurate (low median error), but a few extreme high value samples still result in large errors as reflected in the RMSE