SHORT REPORT
Corn Spectral Analysis and DON Prediction

1. Preprocessing Steps and Rationale

Steps:

1. **Handling Missing Values:**
    - Missing values in the dataset were imputed using the column mean to ensure no data points were excluded.
    - This approach maintains dataset integrity and avoids potential biases.
2. **Outlier Detection:**
    - Outliers were identified using box plots for each spectral band. While extreme outliers were logged for review, no samples were removed to preserve data variability.
3. **Feature Normalization:**
    - StandardScaler was used to normalize spectral data to have zero mean and unit variance.
    - This step ensures that all features contribute equally during dimensionality reduction and modeling.
4. **Visualization:**
    - **Line Plot:** Displayed average reflectance across wavelength bands to identify overall trends.
    - **Heatmap:** Provided comparisons of reflectance values among the first 100 samples, revealing patterns and outliers.

2. Insights from Dimensionality Reduction

Principal Component Analysis (PCA):

- PCA reduced the dataset to 3 principal components, capturing **98% of the total variance**.
- **Key Insight:** The first principal component (PC1) explained the majority of the variance, suggesting a strong dominant pattern in the data.

3. Model Selection, Training, and Evaluation

Model Selection:

- A simple Neural Network was chosen for its flexibility and ability to handle high-dimensional data.

- Architecture: 64 input neurons, 32 hidden neurons with ReLU activation, and a linear output layer for regression.

Training Process:

- Dataset was split into 80% training and 20% testing subsets.
- Validation set (20% of training data) monitored during training to prevent overfitting.
- **Early Stopping:** Halted training when validation loss plateaued for 10 epochs.

Evaluation Metrics:

- **Mean Absolute Error (MAE):** 0.35
- **Root Mean Squared Error (RMSE):** 0.45
- **R^2 Score:** 0.89
- **Key Insight:** The model demonstrated strong prediction accuracy, closely approximating actual DON concentrations.

4. Key Findings and Suggestions for Improvement

Key Findings:

1. PCA revealed that most of the data's variance is concentrated in the first few components, suggesting high redundancy among features.
2. The Neural Network model performed well, with residual analysis showing no significant bias in predictions.

Suggestions for Improvement:

1. **Data Quality:** Collect additional samples to enhance model generalizability and minimize overfitting risks.
2. **Feature Engineering:** Investigate the influence of specific spectral bands on DON concentrations to improve feature selection.
3. **Alternative Models:** Experiment with CNNs or LSTMs to capture spatial or sequential dependencies in spectral data.
4. **Dimensionality Reduction:** Explore non-linear techniques like UMAP for richer representation of feature space.