Assignment Report: Predicting DON (Vomitoxin) Concentration from Hyperspectral Data

1. Preprocessing: Data Cleaning and Normalization

- **Preprocessing:** We start by reading the raw hyperspectral data (spectral features for corn samples) and the corresponding DON concentration values.
- **Normalization:** Spectral data typically have varied scales across wavelengths, so standardization (using methods like StandardScaler) is applied to ensure that each feature contributes equally to the model. This reduces the risk of features with larger scales dominating the learning process.
- **Feature Filtering:** A correlation analysis is performed to assess the relationship between each spectral feature and the target variable. Based on a chosen threshold (e.g., features with an absolute correlation above 0.1), only the most relevant features (265 out of the 448 in the original set) are retained for modeling.
  Rationale: Reducing the number of features not only decreases computational complexity but also mitigates multicollinearity, which is common in hyperspectral datasets. But since this step reduced the model performance, contrary to the expectations, we retained all the features for final model training and testing.

2. Insights from Dimensionality Reduction

Principal Component Analysis (PCA):

- PCA was applied to visualize the variance captured by the spectral data. A plot of cumulative explained variance revealed how many principal components are needed to retain most of the information.
- A 2D scatter plot of the first two principal components (color-coded by DON concentration) provided insights into the data structure. However, the lack of clear clustering suggests that the signal related to DON might be distributed across many wavelengths.

t-SNE Visualization:

- t-SNE was used as an alternative non-linear dimensionality reduction technique. The 2D visualization of features did not show distinct clusters corresponding to different levels of DON concentration.

Interpretation:

- Both PCA and t-SNE indicate that while the spectral features do not form perfect clusters by DON concentration, the variance is spread across multiple dimensions. This reinforces the need for models that can capture subtle, distributed signals rather than relying on a few dominant features.

3. Model Selection, Training, and Evaluation Details

Model Selection:

- **Random Forest Regression:**
  Selected due to its robustness in handling high-dimensional, multicollinear data and its ability to capture non-linear relationships. GridSearchCV was used to tune hyperparameters (number of trees, max depth, etc.).
- **XGBoost Regression:**
  Also applied with hyperparameter tuning. However, initial experiments indicated that XGBoost did not perform as well as Random Forest in this context—possibly due to overfitting or sensitivity to redundant features.
- **Partial Least Squares (PLS) Regression:**
  Chosen because it is well-suited for data with high collinearity. Cross-validation was used to determine the optimal number of latent components that best predicted DON concentration.
- **Neural Network (PyTorch):**
  A fully connected feed-forward network was built with dropout layers and L1 regularization to mitigate overfitting. Early stopping was implemented to halt training when validation loss plateaued. This model gave the best performance among all the others.

Training and Evaluation:

- **Cross-Validation:**
  For each regression method, cross-validation was applied to reliably assess model performance.
- **Evaluation Metrics:**
  Performance was primarily measured using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the R^2 score.
- **Visualization of Results:**
  For all models, scatter plots comparing actual vs. predicted DON concentrations were generated to visually assess prediction accuracy. Loss curves (for the neural network) helped in monitoring training progress and ensuring early stopping worked as intended.

4. Key Findings and Suggestions for Improvement

Key Findings:

- **Preprocessing Impact:** The feature filtering based on correlation helped reduce the dimensionality from 448 wavelength bands to 205, but failed in focusing the model on the most relevant features and improving overall model performance. This can be because of the lack of clearly dominating features that govern the predictions.
- **Dimensionality Reduction Insights:** Both PCA and t-SNE did not reveal perfectly distinct clusters by DON concentration, suggesting that the signal is subtle and spread across many features. This justifies the use of models capable of capturing complex, distributed patterns.
- **Model Performance:**
  - Random Forest performed robustly given its ensemble nature and ability to manage redundancy in the features. But still needs a lot more improvement.

    - XGBoost required more careful hyperparameter tuning and might have suffered from overfitting, as its performance was not as good as Random Forest.
    - PLS regression proved to be useful for datasets with high collinearity. It gave a little better results as compared to RF Regression on this task.
    - The neural network, with regularization techniques and early stopping, provided the best results but might benefit from further tuning of its architecture and hyperparameters.

Suggestions for Improvement:

- **Hyperparameter Optimization:**
    - Exploring more extensive hyperparameter tuning for XGBoost and the neural network.
- **Ensemble Methods:**
    - Stacking or blending models (e.g., combining Random Forest, PLS, and Neural Network predictions) to leverage the strengths of each approach.
- **Data Augmentation:**
    - If possible, increasing the sample size through data augmentation or acquiring additional data could help improve model robustness, especially for deep learning approaches.