<u>DON Concentration Prediction Report</u>

1. Preprocessing Steps & Rationale

The dataset underwent several preprocessing steps to ensure data quality and consistency:

- **Handling Missing Values:** No missing values were found; hence, no imputation was necessary.
- **Outliers Retention:** Instead of removing outliers, they were retained as they may carry valuable domain-specific information, helping preserve real-world variability and prevent information loss.
- **Feature Scaling:** Spectral features were normalized using MinMaxScaler to improve model performance by ensuring uniform input distributions.
- **Data Splitting:** The dataset was divided into 80% training and 20% testing to ensure robust model evaluation.

2. Insights from Dimensionality Reduction & Scaling

- **Feature Normalization:** Standardizing spectral reflectance values helped stabilize training and improved convergence in neural network models.
- **Feature Importance:** Analysis using LIME revealed that certain spectral bands (e.g., wavelengths corresponding to indices 121, 175, and 135) had the highest impact on predictions.
- **No PCA Applied:** Given that spectral reflectance data often retains essential patterns, dimensionality reduction was not performed to avoid loss of crucial information.

3. Model Selection, Training & Evaluation

Three regression models were trained and evaluated:

- **Feedforward Neural Network (FNN):** Trained using TensorFlow/Keras with ReLU activation and Adam optimizer. The model was optimized using Optuna for hyperparameter tuning. **It achieved the best performance with MAE = 2611.85, RMSE = 6971.05, and R^2 = 0.8262.**
- **XGBoost Regressor:** Tuned for max depth, learning rate, and boosting iterations. However, it performed poorly, with **MAE = 4231.75, RMSE = 13368.41, and R^2 = 0.3607.**
- **Ensemble Model (FNN + XGBoost):** A weighted averaging approach was used to combine predictions from FNN and XGBoost, but it did not improve performance over FNN alone, yielding **MAE = 3399.82, RMSE = 9666.29, and R^2 = 0.6657.**

4. Key Findings & Areas for Improvement

Model Performance & Residual Analysis

- **FNN outperformed both XGBoost and the Ensemble Model, achieving the best overall accuracy and lowest error.**
- **XGBoost struggled significantly, with a high RMSE and a poor R^2 score, indicating it failed to capture relationships in the data effectively.**
- **The Ensemble Model did not improve over FNN alone, suggesting that XGBoost predictions were too weak to enhance overall performance.**
- Residual analysis showed that both FNN and XGBoost had some systematic errors in high-concentration cases, while FNN alone minimized these errors more effectively.
- LIME explanations indicated that a few spectral bands dominated predictions, highlighting the need for further feature engineering.

Possible Improvements

- **Drop XGBoost and focus on improving FNN further.**
- **Additional Feature Engineering:** Creating spectral indices (e.g., NDVI-like indices for hyperspectral data) could enhance feature representation.
- **Advanced Model Architectures:** Trying CNNs for spatial feature extraction from spectral bands.
- **Hyperparameter Fine-tuning:** Exploring Bayesian optimization or Genetic Algorithms for further performance gains.
- **Data Augmentation:** Generating synthetic spectral samples using techniques like SMOTE to improve model generalization.
- **Implement MLflow:** Using MLflow for model tracking, versioning, and experiment logging would streamline model comparison and reproducibility, making it easier to test various improvements systematically.

Final Recommendation

Given the results, the **FNN model should be used for final deployment**, as it achieved the best performance in predicting DON concentration. The ensemble model and XGBoost should not be prioritized unless further improvements are made to enhance their predictive capabilities.