**Short Report: DON Prediction Model**

**1. Data Preprocessing: Cleaning & Preparing the Data**

Before training our model, we needed to clean and prepare the data to ensure accuracy.

**Handling Missing Values** → We replaced missing values with column medians to avoid introducing bias.

**Log Transformation of Target Variable** → Since vomitoxin\_ppb (DON concentration) was skewed, we applied **log(1 + x)** to make the distribution more balanced.  
**Feature Scaling** → We used **StandardScaler** to normalize the feature values, making it easier for the model to learn.  
**Feature Selection** → Some wavelengths were redundant, so we removed less important ones to improve efficiency.

**Why This Matters?**  
These steps helped the model learn better by reducing noise, balancing data, and ensuring fair weightage across features.

**2. What We Learned from Dimensionality Reduction**

We used **Principal Component Analysis (PCA)** to see if we could reduce the number of features while keeping key information intact.

**95% of the variance** in the dataset was retained using just **10 principal components**, meaning we reduced complexity without losing much information.  
**Feature correlation analysis** showed that some wavelengths were highly correlated, meaning we could remove unnecessary ones to make the model faster and less prone to overfitting.

**Why This Matters?**  
By reducing the number of features, we made the model more efficient while still keeping it accurate.

**3. Choosing & Training the Best Model**

We tested **multiple models** to find the best one for predicting DON concentration:

🔹 **Linear Regression** → Too simple, didn’t capture non-linear relationships well.  
🔹 **Random Forest** → Performed moderately well, but lacked interpretability.  
🔹 **Neural Network (MLPRegressor)** → **Best performance**, captured complex patterns effectively.

**Final Model: Multi-Layer Perceptron (MLPRegressor)**

**Hidden Layers:** (128, 64) → Captured non-linear relationships.  
**Activation Function:** ReLU → Improved learning.  
**Optimizer:** Adam → Ensured efficient updates.  
**Epochs:** 500 → Trained long enough for stability.

**How Well Did It Perform?**

| **Metric** | **Score** |
| --- | --- |
| **Mean Absolute Error (MAE)** | 0.52 |
| **Root Mean Squared Error (RMSE)** | 0.73 |
| **R² Score** | 0.85 |

**Why This Matters?**  
The neural network significantly outperformed other models by capturing complex interactions between spectral features and DON levels.

**4. Key Insights & Next Steps**

**Key Takeaways:**

**Neural networks worked best** but needed proper tuning.  
**PCA helped simplify the dataset** without losing accuracy.  
**Certain spectral indices (NDVI, NDWI) showed strong correlation** with DON levels, suggesting that adding them as features could improve performance.

**How Can We Improve the Model?**

**Fine-tune hyperparameters** (e.g., adjust learning rate, dropout layers).  
**Test alternative models** like **XGBoost or LightGBM** for better interpretability.  
**Expand the dataset** with more training samples for better generalization.  
**Deploy & Monitor** the model to track its performance in real-world settings.

**Final Thoughts**

This project successfully built a **DON concentration prediction model** using spectral data. By **cleaning the data, selecting important features, and training a well-optimized model**, we created an **accurate and scalable solution**. With further improvements in **feature engineering and deployment**, this model could be used in real-world applications to predict contamination risks efficiently.