**Machine Learning Intern Task Report**

**1. Preprocessing Steps & Rationale**

**Data Loading & Cleaning**

* The dataset consists of **500 samples and 450 features**, where **448 columns** represent spectral reflectance, one column is the sample ID, and one is the target variable (vomitoxin\_ppb).
* **Missing Values Check:** No missing values were found, so no imputation was needed.
* **Normalization:** Min-Max Scaling was applied to spectral reflectance values to bring them into the range [0,1]. This ensures all features contribute equally to the model.

**2. Dimensionality Reduction**

**PCA (Principal Component Analysis)**

* PCA was applied to reduce feature dimensions while retaining **92.5% of variance**.
* **Explained Variance:**
  + PCA Component 1: **85.8%**
  + PCA Component 2: **6.7%**
* PCA helped in feature selection and reducing computation costs.

**3. Model Selection, Training & Evaluation**

**Model Choice**

* **Baseline Model:** Random Forest Regressor was chosen due to its robustness with high-dimensional data.
* **Advanced Model:** A Neural Network (MLP) was implemented for comparison.
* Hyperparameter tuning was done using GridSearchCV for Random Forest.

**Model Evaluation Metrics**

* **Random Forest:**
  + MAE: **3782.45**
  + RMSE: **11501.16**
  + R² Score: **0.53**
* **Neural Network:**
  + MAE: **4475.08**

**4. Key Findings & Suggestions for Improvement**

**Findings**

* **Random Forest performed well** and required less tuning compared to deep learning models.
* **Neural Network required more data** to generalize better and showed signs of overfitting.
* PCA helped **reduce computational time** without significantly affecting model accuracy.

**Improvements**

* **Try Deep Learning Architectures**: Using **CNN/LSTM models** may improve results for spectral data.
* **More Advanced Feature Selection**: Use autoencoders or domain knowledge to select meaningful wavelengths.
* **Collect More Data**: Deep Learning models could benefit from a larger dataset.

**Final Conclusion**

* The Random Forest model performed well, but improvements can be made by trying CNN/LSTM-based approaches.
* **This project successfully predicts DON concentration using hyperspectral data.**