Machine Learning Task Report

1. Data Preprocessing

The dataset consists of 500 samples and 450 columns, including 448 spectral features, an hsi_id column, and the target variable (vomitoxin_ppb).

Preprocessing Steps:

- Checked for Missing Values: No missing values were found.
- Feature Scaling: Applied MinMax Scaling to normalize spectral values between 0 and 1.
- Train-Test Split: The dataset was split into 80% training and 20% testing.

2. Data Visualization & Insights

To understand the dataset better, two key visualizations were created:

- Average Spectral Reflectance: Shows spectral response patterns.
- Spectral Reflectance Heatmap: Displays a subset of 30 samples.

3. Model Selection, Training & Evaluation

Three regression models were used to predict vomitoxin levels:

- Random Forest Regressor
- XGBoost Regressor
- Neural Network (MLP Regressor)

Performance Metrics (Lower MAE and RMSE, Higher R^2 are better):

Model	MAE	RMSE	R^2 Score
Random Forest	3765.06	11483.80	0.528
XGBoost	3972.92	12996.28	0.395
Neural Network	3446.27	11039.01	0.564

4. Key Findings & Suggestions for Improvement

Key Findings:

- The Neural Network performed best, achieving the lowest error and highest R^2 score.
- XGBoost performed the worst, likely due to high feature correlation.
- The dataset is high-dimensional, making feature selection important.

Suggestions for Improvement:

- Apply PCA or t-SNE for dimensionality reduction.
- Optimize hyperparameters using Grid Search.
- Experiment with Convolutional Neural Networks (CNNs).
- Develop an interactive UI for model predictions using Streamlit.

5. README: Running the Project

Installation & Dependencies:

Ensure you have Python and required libraries installed:

```
pip install pandas numpy matplotlib seaborn scikit-learn xgboost
```

How to Run:

1. Clone the repository or download the files.
2. Place TASK-ML-INTERN.csv in the working directory.
3. Run the main script:

```
python main.py
```