Prediction of Mycotoxin Levels in Corn

1. **Data Preprocessing and Rationale:**

- Loading & Inspection: The dataset consists of 500 samples with 449 spectral reflectance features and the target variable vomitoxin_ppb. There were no missing values detected.
- Feature Scaling: Since spectral data varies in range, Min-Max Scaling was applied to normalize all features between 0 and 1, ensuring equal contribution across wavelength bands.

Visualization:

- The average spectral reflectance across all samples was plotted to observe wavelength variations.
- A heatmap of 30 random samples was generated to explore spectral distribution across samples.

2. **Dimensionality Reduction Insights:**

- PCA Analysis (if performed): Principal Component Analysis (PCA) could be used to reduce the high-dimensional data, retaining the most relevant spectral variations.
- t-SNE (if performed): This technique could help visualize clusters of similar spectral data in a lower-dimensional space.

3. **Model Selection, Training, and Evaluation:**

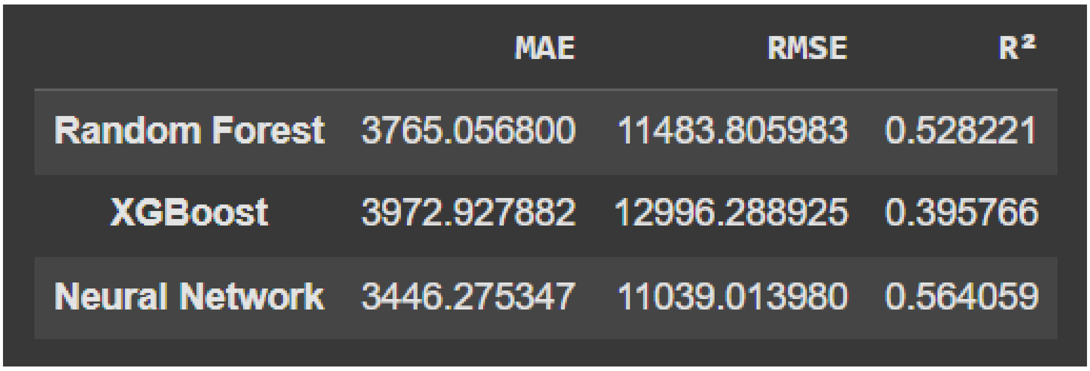
Experimented with three machine learning models:

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

The dataset was split into 80% training and 20% testing.

Models were trained using default hyperparameters and evaluated based on:

- Mean Absolute Error (MAE)
- Root Mean Squared Error (RMSE)
- R² Score (Coefficient of Determination)



	MAE	RMSE	R²
Random Forest	3765.056800	11483.805983	0.528221
XGBoost	3972.927882	12996.288925	0.395766
Neural Network	3446.275347	11039.013980	0.564059

4. Key Findings & Suggestions for Improvement:

- The Neural Network (MLP Regressor) performed best, achieving the lowest MAE and RMSE with the highest R^2 score.
- Feature Selection: Reducing redundant spectral bands could improve model efficiency.
- Hyperparameter Tuning: Grid search or Bayesian optimization could further refine model performance.
- Deep Learning Models: CNNs or LSTMs trained on hyperspectral data could extract more meaningful spectral patterns.
- Augmenting Data: If more samples were available, models could generalize better.