Report 2: Mycotoxin Prediction in Corn Using Machine Learning

- Clipping Method is used for handling outlliers in target variable

1. Preprocessing Steps & Rationale

1.1 Data Cleaning & Normalization

- **Outlier Removal:** Applied the **Interquartile Range (IQR) method** to remove extreme values, ensuring model stability.
- **Feature Scaling:** Used **MinMaxScaler** to normalize spectral reflectance data, ensuring all features are within the same range for better model convergence.
- **Clipping Extreme Values:** Instead of log transformation, we applied **IQR-based value clipping** to the target variable.
    - **Rationale:** This prevents extreme target values from distorting predictions while maintaining meaningful variation.

1.2 Dimensionality Reduction (PCA)

- **Applied Principal Component Analysis (PCA)** to retain **95% variance**, reducing 448 spectral bands to the top 5 most important components.
- **Rationale:** PCA helped remove noise and redundant information, improving model interpretability.
- **Insights: PC2 and PC3 were consistently the most important features across all models.**

2. Model Selection, Training & Evaluation

2.1 Models Tested

Model	MAE (Lower is Better)	RMSE (Lower is Better)	R^2 Score (Higher is Better)

Model	MAE (Lower is Better)	RMSE (Lower is Better)	R^2 Score (Higher is Better)
Random Forest	664.61	892.87	**0.5275**
XGBoost	650.61	934.76	0.4822
Tuned XGBoost	695.77	921.78	0.4964
MLP (500 iterations)	768.95	1059.00	0.335
MLP (2500 iterations)	702.12	943.41	0.4725
MLP (Tuned)	698.26	1020.17	0.3832

2.2 Best Model: Random Forest

- **Performance:**
  - **MAE: 664.61**
  - **RMSE: 892.87**
  - **R^2 Score: 0.5275** (Best among all models tested)
- **Feature Importance Insights:**
  - **PC2 (43.6%) & PC3 (23.9%) were the most important features**.

---

3. Key Findings & Suggestions for Improvement

3.1 Key Findings

Random Forest performed best among all models tested based on R^2 score.

XGBoost showed competitive results, but tuning did not significantly improve performance.

MLP required more iterations to improve performance but still lagged behind tree-based models.

PCA significantly improved model accuracy by removing redundant spectral bands.

IQR-based value clipping effectively handled extreme values, but further analysis is needed to confirm its effectiveness compared to log transformation.

3.2 Observations

- **Random Forest performed best overall, suggesting that simpler models work well for this dataset.**
- **MLP models required significantly more iterations to approach tree-based models' performance.**
- **XGBoost tuning did not lead to meaningful improvements, possibly due to data characteristics.**

- **PC2 and PC3 were consistently the most important principal components.**

3.3 Areas for Improvement

Explore Hybrid Models → Combine RF + XGB for ensemble learning.
Try Log Transformation Again → Compare its effectiveness against clipping for handling extreme values.
Increase Dataset Size → Deep learning models (MLP) may improve with more training data.
Optimize Feature Engineering → Instead of PCA, explore domain-specific spectral feature extraction.
