Short Report: Mycotoxin Prediction in Corn Using Machine Learning

1. Preprocessing Steps & Rationale

1.1 Data Cleaning & Normalization

- **Outlier Removal**: Applied the **Interquartile Range (IQR) method** to remove extreme values, ensuring model stability.
- **Log Transformation**: Transformed the target variable (DON concentration) using `log1p()` to reduce skewness.
- **Feature Scaling**: Used **MinMaxScaler** to normalize spectral reflectance data, ensuring all features are within the same range for better model convergence.

1.2 Dimensionality Reduction (PCA)

- **Applied Principal Component Analysis (PCA)** to retain **95% variance**, reducing 448 spectral bands to a smaller set of meaningful components.
- **Rationale**: PCA helped remove noise and redundant information, improving model performance.
- **Insights**: **PC2, PC3, and PC1 were the most influential components**, indicating that not all spectral bands contribute equally to prediction.

2. Model Selection, Training & Evaluation

2.1 Models Tested

Model	PCA Used?	MAE (Lower is Better)	RMSE (Lower is Better)	R² Score (Higher is Better)
Random Forest	□ Yes	□ 1.81	□ 2.46	□ 0.24
Random Forest (Tuned)	□ Yes	□ 1.76	□ 2.37	□ 0.30
XGBoost	□ Yes	□ **1.70**	□ **2.35**	□ **0.31** □
XGBoost (Tuned)	□ Yes	□ 1.82	□ 2.48	□ 0.23
MLP (Neural Network)	□ Yes	□ 1.85	□ 2.44	□ 0.26
MLP (Tuned)	□ Yes	□ 3.00	□ 4.05	□ -1.05

2.2 Best Model: XGBoost (Untuned)

- **Performance**:
  - **MAE**: **1.70**
  - **RMSE**: **2.35**
  - **R² Score**: **0.31** (Best among all models tested)
- **Feature Importance Insights**:
  - **PC3 (29.5%) & PC2 (27.8%) were the most important features**.

3. Key Findings & Suggestions for Improvement

3.1 Key Findings

□ **XGBoost (Untuned) performed best** among all models tested based on raw performance.
□ **PCA significantly improved model accuracy** by removing noise from spectral bands.
□ **Random Forest was a strong alternative due to stability & interpretability.**
□ **Deep Learning model (MLP) improved slightly with tuning but was still outperformed by tree-based models.**

3.2 Observations

- **XGBoost (Untuned)** had the best overall accuracy, but tuning worsened its performance, likely due to overfitting.
- **Random Forest (Tuned)** provided the second-best results, showing that tuning helped this model.
- **MLP** had inconsistent results, indicating that deep learning models need more data to perform well.
- **PCA** was essential in improving model accuracy by removing redundant spectral bands.

3.3 Areas for Improvement

□ **Increase Dataset Size** → Deep learning models (MLP) may improve with more training data.
□ **Try Advanced Hyperparameter Tuning** → Bayesian Optimization instead of GridSearchCV.

Test Different Feature Engineering Approaches → Instead of PCA, extract domain-specific spectral features.
Explore Hybrid Models with Different Base Models → Combine RF + MLP instead of RF + XGB.
