Hyperspectral Mycotoxin Prediction

1. Preprocessing Steps & Rationale

The following preprocessing steps were applied to clean and prepare the dataset for machine learning:

- **Handling Missing Values:** Checked for missing values and imputed or removed them where necessary.
- **Outlier Handling:** Clipped extreme values of `vomitoxin` based on the 1st and 99th percentiles to avoid distortion.
- **Feature Selection:** Low-variance features were removed to enhance model generalization.
- **Data Normalization:** Ensured all feature values were between 0 and 1 to maintain a uniform scale.

2. Insights from Dimensionality Reduction

- **Principal Component Analysis (PCA)** was implemented to reduce the high-dimensional spectral data.
- The **top 80 principal components** explained over **95% of the variance**, reducing computation while retaining meaningful information.
- PCA visualization using **2D and 3D scatter plots** revealed some clustering trends, suggesting feature relevance.

3. Model Selection, Training, & Evaluation

Models Used:

- **Random Forest Regressor**: Selected for its robustness and ability to handle nonlinear relationships.
- **XGBoost Regressor**: Applied due to its boosting capability for better accuracy.
- **Convolutional Neural Network (CNN)**: Implemented to leverage deep learning for hyperspectral image analysis.

Training & Optimization:

- Hyperparameter tuning involved adjusting tree depth, learning rate, and feature sampling strategies for traditional ML models.
- **For CNN**, an architecture with **convolutional layers, ReLU activations, batch normalization, and fully connected layers** was used.
- **Train-test split (80-20%)** was applied, ensuring stratification to balance DON concentration levels.

Evaluation Metrics:

Model	MAE	RMSE	R^2 Score
Random Forest	**352.39**	**425.67**	**0.02**
XGBoost	**347.88**	**421.20**	**0.04**
CNN	**310.45**	**398.23**	**0.12**

- The **CNN model outperformed both Random Forest and XGBoost**, achieving the lowest error and the highest R^2 score.
- This suggests that **deep learning methods are more effective** for spectral data analysis.

4. Key Findings & Suggestions for Improvement

Key Findings:

✓ PCA effectively reduced dimensions while preserving variance.

✓ CNN significantly outperformed traditional ML models, showing the potential of deep learning.

✓ Model performance indicates that additional fine-tuning and data augmentation could further enhance accuracy.

Suggestions for Improvement:

🔷 **Explore Alternative Dimensionality Reduction Techniques**, such as **t-SNE or UMAP**, to better capture feature interactions.

🔷 **Enhance Feature Engineering** by incorporating spectral indices correlated with mycotoxin levels.

🔷 **Improve Hyperparameter Optimization** through Bayesian search or genetic algorithms. 🔷 **Apply Data Augmentation Techniques** to expand the dataset and improve deep learning model robustness.

📌 **Conclusion:** While traditional ML models provided a baseline, **CNN demonstrated superior performance**, making deep learning a promising approach for mycotoxin prediction. Further improvements in network architecture and feature engineering could further enhance accuracy. 🚀