Machine Learning Pipeline for Mycotoxin Prediction in Corn Samples

1. Introduction

This documentation outlines the solution implemented for predicting DON concentration in corn samples using hyperspectral imaging data. The pipeline follows a structured approach to data preprocessing, model training, evaluation, and deployment, ensuring accuracy and production-readiness.

2. Problem Statement and Solution Approach

The objective is to predict DON concentration using spectral reflectance data. The pipeline addresses the problem through:

- **Preprocessing:** Cleaning, normalizing, and enhancing the data.
- **Visualization:** Understanding spectral characteristics through plots.
- **Model Training:** Implementing a neural network and advanced models.
- **Evaluation:** Using robust metrics and interpretability tools.
- **Deployment:** Creating a modular, production-ready pipeline.

3. Dataset Description

- **Features:** Spectral reflectance values at multiple wavelengths.
- **Samples:** Each row represents a corn sample.
- **Target Variable:** DON concentration (continuous numerical value).

4. Data Exploration and Preprocessing

4.1 Data Loading & Inspection

- Loaded dataset using Pandas and inspected for missing values and outliers.
- Used `.info()`, `.describe()`, histograms, and boxplots for statistical insights.

4.2 Preprocessing Steps

- **Handling Missing Data:** Imputed missing values using the median.
- **Normalization:** Applied Min-Max scaling for feature consistency.
- **Outlier Detection:** Used Z-score and IQR to remove anomalies.
- **Feature Engineering:** Created additional spectral indices to enhance predictive power.
- **Visualization:**
    - Line plots for reflectance trends.
    - Heatmaps for feature correlation.
    - Pairplots to analyze relationships between wavelengths.

4.3 Advanced Data Quality Checks

- Automated checks for sensor drift and inconsistencies.
- Validated data integrity using statistical tests.

5. Model Training

5.1 Model Selection

- **Baseline Model:** Feed-forward neural network with optimized hyperparameters.
- **Advanced Models:** Compared XGBoost, Gradient HistBoosting, and Transformer-based models with attention mechanisms for performance improvements.

5.2 Data Splitting

- Train-test split (80%-20%) for model validation.
- k-Fold cross-validation for stability.

5.3 Hyperparameter Optimization

- Used an **iterative approach** to refine hyperparameters dynamically based on model performance.
- Experimented with custom loss functions to enhance prediction accuracy.

6. Model Evaluation

6.1 Performance Metrics

- Evaluated using:
    - **Mean Absolute Error (MAE)**
    - **Root Mean Squared Error (RMSE)**
    - **R^2 Score**

6.2 Visualization of Results

- Scatter plots for actual vs. predicted values.
- Residual analysis to detect systematic errors.

6.3 Model Interpretability

- Applied SHAP for feature importance analysis.
- Documented model insights and limitations.

7. Pipeline Integration and Deployment

7.1 Code Quality

- Modularized code into independent scripts for:
    - Data preprocessing
    - Model training and evaluation
    - Inference and deployment
- Implemented unit tests to validate pipeline components.
- Integrated logging for debugging and runtime tracking.

7.2 Deployment

- Packaged as a **Flask/FastAPI service** for real-time predictions.
- Containerized using **Docker** for seamless deployment.
- Designed API endpoints to accept new spectral data dynamically.

8. Deliverables

- **GitHub Repository** with:
    - Well-documented Jupyter Notebook and Python scripts.
    - Report summarizing preprocessing, model selection, and findings.
    - README with setup instructions.
    - Deployment scripts for API integration.

9. Conclusion

The pipeline successfully predicts DON concentration using hyperspectral imaging. Future enhancements include:

- **Transformer-based models** for improved feature extraction.
- **Interactive Streamlit app** for real-time user interaction.
- **Ensemble learning** for improved generalization.