MYCOTOXIN LEVEL PREDICTION IN CORN SAMPLES:

A MACHINE LEARNING APPROACH

1. Introduction

Mycotoxins, specifically **deoxynivalenol (DON)**, are toxic compounds produced by fungi in agricultural products like corn. Detecting DON concentration is crucial for food safety. This project leverages **machine learning (ML)** to predict DON concentration in corn samples using **hyperspectral imaging data**.

The goal is to develop an effective predictive model that can help in early detection and analysis of DON concentration, improving food quality monitoring.

---

2. Data Preprocessing and Rationale

The dataset consists of hyperspectral imaging features used to predict DON concentration. Proper preprocessing is crucial to ensure model accuracy and robustness.

2.1 Handling Missing Values

- The dataset was checked for missing values.
- Rows containing missing values were dropped to ensure clean data.

2.2 Feature Scaling

- Since machine learning models perform better with normalized data, all numerical features were standardized using **StandardScaler** from sklearn.preprocessing.
- This ensures that all features contribute equally to model training without being biased by scale differences.

2.3 Train-Test Split

- The dataset was split into:
    - **80% Training data**
    - **20% Testing data**
- This split ensures the model can learn patterns from the training data and generalize to unseen data.

---

3. Model Selection, Training, and Evaluation

3.1 Choice of Model

A **Multi-Layer Perceptron (MLP) Regressor** was chosen due to its ability to capture **nonlinear relationships** in complex datasets like hyperspectral imaging. The MLP model consists of:

- **Two hidden layers** with **64 and 32 neurons** respectively.
- **ReLU activation function**, which helps in learning complex features.
- **Adam optimizer**, which adapts learning rates dynamically for efficient training.
- **500 maximum iterations** to ensure sufficient training time.

3.2 Model Training Process

The MLPRegressor was trained using the standardized feature set. The model attempts to learn patterns that link hyperspectral imaging features to DON concentration levels.

3.3 Model Evaluation Metrics

To assess model performance, the following evaluation metrics were used:

- **Mean Squared Error (MSE)**: Measures the average squared difference between actual and predicted values.
- **R^2 Score**: Determines how well the model explains the variance in the target variable.

3.4 Model Performance Visualization

A **scatter plot of actual vs. predicted values** was created to visually assess model performance. The ideal prediction follows a **45-degree diagonal line**, but deviations indicate model error.

---

4. Model Interpretability Using LIME

Machine learning models, especially deep models like MLP, can be **black boxes**, making it difficult to understand their decisions. To address this, **LIME (Local Interpretable Model-agnostic Explanations)** was used.

4.1 How LIME Works

- LIME explains model predictions by creating **local approximations** of the complex model using simpler interpretable models (like linear regression).
- It perturbs the input features and observes changes in the model's predictions, helping to determine the most influential features.

4.2 Implementing LIME

- The **LIME explainer** was initialized using the standardized training dataset.
- A **random test sample** was selected for explanation.
- The **top 10 most influential features** were identified, revealing which factors contributed most to the DON prediction.

4.3 Key Findings from LIME

- Some spectral features had **significant influence** on the predictions.
- The model's response was **nonlinear**, validating the need for deep learning techniques like MLP.
- LIME provided an **intuitive breakdown of feature importance**, making it easier to understand the decision-making process of the model.

---

5. Key Findings

5.1 Key Insights

- The MLP model successfully captured relationships in hyperspectral imaging data.
- The scatter plot showed **a good correlation between actual and predicted values**, but some deviations indicate room for improvement.
- **LIME analysis** highlighted the **most important features**, offering transparency into the model's decision-making process.

---

6. Conclusion and Future Work

This study demonstrates the **effectiveness of ML models** in predicting mycotoxin levels using hyperspectral imaging data. The MLP model provided reasonable predictions, and **LIME explainability techniques** helped interpret the results.

Future Work

- **Exploring Advanced Models**: Investigate **XGBoost, CNNs, or hybrid deep learning models**.
- **Feature Engineering**: Identify new **hyperspectral bands** that might correlate better with DON levels.
- **Data Augmentation**: Use synthetic data generation techniques to improve model robustness.

By refining preprocessing, model selection, and interpretability techniques, the accuracy of DON prediction can be significantly improved, enhancing agricultural safety and quality control.

---

7. How to Run the Project

7.1 Requirements

Ensure you have the following Python packages installed:

pip install numpy pandas matplotlib seaborn scikit-learn lime

7.2 Running the Code

1. Place the dataset (MLE-Assignment.csv) in the working directory.
2. Run the Python script containing the ML pipeline.
3. The script will preprocess data, train the MLP model, evaluate its performance, and generate LIME explanations.