Prediction of Mycotoxin Concentration in Corn via Hyperspectral Imaging

Introduction

This document outlines the creation of a machine learning pipeline to predict deoxynivalenol (DON) concentration in corn samples from hyperspectral imaging data. The work entails preprocessing intricate spectral data, using dimensionality reduction methods, choosing and tuning regression models, and making the pipeline modular and production ready.

Data Preprocessing

The preprocessing step starts with data loading of hyperspectral images and retrieving useful features. The response variable, DON concentration (vomitoxin_ppb), is separated to be predicted. Owing to the high-dimensional nature of hyperspectral data, preprocessing is essential to improve model performance and interpretability.

Missing values in spectral features are treated through median imputation to preserve the dataset intact without any biases arising from mean-based imputation. Feature scaling is done through standardization (z-score normalization) to scale the spectral reflectance values to the same magnitude for different wavelengths. For model training data preparation, the data is divided into a training set and a test set, 80-20 ratio, to check the generalization ability of the predictive model.

Dimensionality Reduction

Hyperspectral data contains hundreds of wavelength-specific reflectance values, which may introduce redundancy and collinearity. Dimensionality reduction methods like Principal Component Analysis (PCA) are used to keep the most informative components and remove noise. Dimensionality reduction optimizes the model training process, resulting in improved generalization and lower computational expenses.

Exploratory data analysis encompasses spectral signature visualization to recognize wavelength bands where substantial reflectance value variations are present among samples. These findings are useful for grasping the spectral feature and DON concentration relationship and informing feature selection.

Model Selection, Training, and Evaluation

A regression model based on deep learning is used to forecast DON concentration. The structure has an input layer as per the number of spectral features, several hidden layers with rectified linear unit (ReLU) activation, and a single-node output layer for predicting continuous

values. The network uses dropout regularization (30%) to prevent overfitting, making it resilient to unseen data.

The training procedure uses the Adam optimizer with Mean Squared Error (MSE) as the loss function. To optimize model efficiency, early stopping is used to avoid unnecessary iterations when the validation loss no longer improves. Model performance is evaluated using important regression metrics like Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the R-squared (R^2) score, which together give an idea of prediction accuracy and model reliability.

For explanation of the contribution of various spectral features, SHAP (SHapley Additive exPlanations) analysis is conducted. This method points out which bands of wavelengths have the strongest contribution to DON concentration prediction, and it supports feature importance analysis.

Key Findings and Suggested Improvements

The resultant machine learning algorithm effectively identifies spectral patterns to foresee DON concentration with acceptable accuracy. Standardization of the spectral features is helpful in maintaining stability across wavelength reflectance values. It is found through analysis that some wavelengths are more significantly contributing to predictions, implying scope for feature selection to enhance efficiency.

Additional improvements may be pursued by including spectral indices, e.g., NDVI-like transforms, which may improve feature representation. Other machine learning models, e.g., Random Forest and XGBoost, could be compared with the deep learning-based methods to determine the best model for this application. Hyperparameter tuning, e.g., the number of hidden layers, learning rate tuning, and dropout rates, may also result in performance improvements.

Lastly, adding data augmentation methods to mimic spectral changes would enhance the robustness of the model such that the pipeline is more capable of handling real-world situations.

Conclusion

The project confirms the viability of mycotoxin content prediction in corn via hyperspectral imaging and machine learning. With successful preprocessing, feature reduction, and model tuning, a stable regression model is established. Further improvements in feature engineering, model choice, and interpretability methods can make predictive performance even better, leading to more accurate and scalable Agri-monitoring solutions.