**Mycotoxin Prediction using Hyperspectral Imaging Data**

**Objective**

The goal of this project is to develop a machine learning pipeline to predict vomitoxin (DON) concentration in corn samples based on spectral reflectance data. We preprocess, visualize, and train various regression models, including a Random Forest Regressor, a Neural Network, an Attention-Based Transformer, and an Ensemble Model (Stacking Regressor), to achieve optimal performance.

**Step 1: Exploration and Data Loading Activities:**

* loaded a dataset with 450 columns and 500 samples, where the first column is an identifier (hsi\_id).
* Spectral reflectance values at various wavelengths are represented by 448 columns.
* The target variable (vomitoxin\_ppb) is the last column.
* Missing values were checked, but none were discovered.
* No duplicate rows were found after summary statistics were calculated.
* Spectral reflectance values were visualised to comprehend the distribution of the data.

**Step 2: Preprocessing Data Measures Done:**

* missing values that were imputed (none were found, so no imputation was required).
* StandardScaler was used to standardise all wavelengths in the spectrum data.
* Finding Outliers:
* To find extreme outliers, the Z-score method (threshold = 3) was used.
* The dataset was reduced from 500 to 480 samples after 20 samples were eliminated.

**Step 3: Model Training**

**3.1 Baseline Model**

Random Forest Model

Used: RandomForestRegressor (n\_estimators=100)

Performance:

MAE: 3765.06

RMSE: 11483.81

R2 Score: 0.5282

Interpretation: Only half of the variance in the target variable is explained by the model, with a high RMSE (~11,483 ppb) suggesting large prediction errors. Try feature engineering, hyperparameter tuning, or more sophisticated models for improvement.

**3.2 Model of Neural Networks Employed:**

* Three-layer Neural Network:
* 64 neurones in
* Layer 1, ReLU activation
* Layer 2: ReLU activation, 32 neurones
* One neurone is the output (for regression).
* Adam is the optimiser.
* The loss function MSE
* Periods: 50
* Performance:
* 3493.94 MAE
* RMSE: 11,521.87
* superior to Random Forest in terms of MAE (the lower the better).
* RMSE is a little worse, suggesting that significant errors are still present.
* The model suffers from complexity and a lack of data.

**3.3 Attention-Based Transformer Model**

Model Employed:

* Transformer Block with: 64 embedding dim, 2 heads, multi-head attention
* Network feedforward (128 hidden units)
* Adam Optimiser, Performance of MSE Loss:
* RMSE: 12,527.19; MAE: 3830.54 👉 Meaning:
* marginally worse than neural networks and random forests.
* For transformers to learn efficiently, larger datasets are needed.
* Pretraining on related spectral datasets might be helpful.

**3.4 Stacking Regressor Ensemble Model Model Employed:**

* Random Forest (n\_estimators=100) is the base model.
* RBF kernel, or SVR
* Ridge Regression Final Estimator Performance: MAE: 4503.74
* Analysis: Worse than every earlier model.
* implies that SVR and Random Forest are not working well together.
* Next Steps: Use XGBoost or Gradient Boosting in place of SVR.

**Model Comparison & Best Model Selection**

| **Model** | **MAE (Lower is Better)** | **RMSE** | **R² Score** |
| --- | --- | --- | --- |
| Random Forest | **3765.06** | 11483.81 | **0.5282** |
| Neural Network | **3493.94** | 11,521.87 | **-** |
| Transformer | 3830.54 | 12,527.19 | **-** |
| Stacking (RF + SVR) | 4503.74 | **Worst** | **-** |

* **Best Performing Model**: **Neural Network (Lower MAE, Stable Performance)**
* **Worst Performing Model**: **Stacking (SVR did not contribute positively)**

**Important Results & Recommendations for Enhancements**

**1. Model Interpretation & Performance**

The best-performing models were Neural Networks and Random Forest.  
Ensemble's Stacking Regressor had the worst performance.  
Large datasets are necessary for transformers to function effectively.

**2. Upcoming Enhancements**

**Tuning hyperparameters:**  
Random Forest grid search (n\_estimators, max\_depth)  
Try varying the batch size and learning rate in neural networks.

**Engineering Features:**  
Extract spectral indices (e.g., NDVI, water absorption bands).  
To reduce dimensionality, use Principal Component Analysis (PCA).

More Data: To properly train deep learning models, gather more maize samples.  
Use hyperspectral imaging datasets to investigate transfer learning.

**Final Conclusion**

**Best Model**: **Neural Network (Lowest MAE: 3493.94)**