**Project Overview: Hyperspectral Data Analysis for Corn Quality Prediction**

* Loaded the dataset using **pandas** and inspected for any missing values.
* Conducted **Exploratory Data Analysis (EDA)** and visualized the spectral reflectance for each corn sample.
* The dataset contains **wavelengths** as independent variables and **vomitoxin (DON) ppb** as the target variable.
* **Vomitoxin (DON) Interpretation:**
  + **0 ppb** is safe for human consumption (no toxin).
  + Humans can withstand up to **2000 ppb** of DON.
  + Crops with **reflectance values between 0.6 - 0.8** are considered good.
  + Some crops exceed **0.9 reflectance**, indicating toxic contamination.
* Identified **outliers (above 0.9 reflectance)** which are likely toxic corn samples.
* **Statistical Analysis:**
  + The target variable does not follow a normal distribution.
  + Data preprocessing is needed, including skewness correction and outlier removal.
* **Wavelength Region Analysis:**
  + **0-100 nm (Blue Region)**: Low reflectance due to absorption by chlorophyll.
  + **100-200 nm (Green Region)**: Higher reflectance, contributing to the green color of corn.
  + **200-300 nm (Red Region)**: Strong absorption by chlorophyll, leading to lower reflectance.
* Differentiated **high and low ppb crops** using a line plot.
  + **Bad corn** exhibits extreme reflectance values.
  + **Good corn** has a gradual increase in reflectance across wavelengths.
  + High reflectance is due to **dehydration, disease, or extreme contamination.**
* **NIR Reflectance vs Vomitoxin Analysis:**
  + Identified **sensor calibration issues** as a possible cause of extreme outliers.
  + Data points showed **poor structure**, making regression modeling challenging.
* **Statistical Summary of Target Variable:**
  + Mean: **3410 ppb**, Min: **0 ppb**, Max: **131,000 ppb**
  + Standard Deviation: **13,095.80 ppb** (high variance, indicating outliers).
  + **Skewness is positive**, meaning most crops have low DON levels, but extreme outliers exist.
  + Peaks observed at **0, 100, 200, and 400 ppb**.
* Analyzed extreme **high-ppb corn samples**, confirming that **higher reflectance values** indicate **toxic corn**.
* **Correlation Analysis:**
  + Wavelengths showed **strong correlation** at **low wavelengths**, but a sudden drop at **100 nm**, which is a **notable feature of corn crops**.
* **Good Corn Analysis:**
  + **88 samples were identified** as good-quality corn.
  + **Wavelength distribution follows a near-normal distribution** with low skewness.
  + **Standardization is not required** before PCA since it will be handled during PCA processing.

**Dimensionality Reduction Using PCA**

* Performed **Principal Component Analysis (PCA)** to reduce dimensionality.
  + **447 independent features** were reduced to **3 PCA components**.
  + **Variance explained by PCA components:**
    - **PC1:** 87.2%
    - **PC2:** 5.7%
    - **PC3:** 2.2%
    - **Total Variance Explained:** **95.1%** (sufficient for analysis).
* **PCA Interpretation:**
  + PCA helps **capture key patterns** in high-dimensional spectral data.
  + The first **few components retain the most valuable information**.
  + The red-dashed line at **95% variance retention** ensures minimal data loss.
* **PCA Scatter Plot Analysis:**
  + No clear **clusters or patterns** were observed, indicating complex data.
  + t-SNE clustering also showed **no strong groupings**, meaning DON levels do not form tight clusters.
* **Conclusion from PCA and Clustering:**
  + Data **exhibits non-linearity**, suggesting the need for a **non-linear model** like **XGBoost**.
  + Regression models are preferred over classification since DON levels are continuous.

**Target Variable Transformation (Vomitoxin ppb)**

* **Skewness Reduction Techniques Applied:**
  + **Outlier removal** using **IQR method** reduced skewness to **1.98**.
  + **Normalization** after outlier removal **did not improve** R² or MAE.
  + **Yeo-Johnson transformation** reduced skewness to **-0.03**, significantly improving model performance.
  + **Log transformation** and **Huber loss regression** did not yield better results.
* **Final Decision:**
  + **Yeo-Johnson transformation** was selected as it produced the best **R² and MAE values**.

**Model Selection & Performance Evaluation**

* **Linear Regression Model:**
  + **MAE:** 0.62
  + **R² Score:** 0.33
  + **Findings:**
    - It **generalizes well** but is **not the best performer** for this dataset.
* **Random Forest & XGBoost:**
  + Performed **hyperparameter tuning using GridSearchCV**.
  + Achieved **R² scores of 0.27 and 0.26**, respectively.
  + Results indicate **no significant improvement** over linear regression.
* **Attention Model:**
  + Provided similar generalization capabilities as XGBoost.

**Final Conclusion**

* **Linear Regression outperformed all other models**, achieving an **MAE of 0.6**.
* Since the data is complex, **MAE is the primary evaluation metric**, rather than R².
* Future work includes **deploying the model using Streamlit and MLflow**, but due to time constraints, this was not implemented in the current iteration.

Since the data was skewed, I learned various data preprocessing techniques like **Yeo-Johnson** and **Huber loss**. The data was complex and non-linear, so understanding it took **two days**.

Using **447 samples** as independent variables, I achieved a **93% R-squared value**, but the **MAE was very high**. This taught me that **R-squared is not reliable for highly skewed data**, which helped me avoid misclassification.

For skewed data, the focus should be on **reducing MAE** rather than relying on R-squared. I also learned how to use **MLflow** (watched tutorials on YouTube), but due to time constraints, I couldn’t implement it.

EDA was crucial in understanding the data. From **reflectance values**, I figured out how to classify whether the corn is toxic or not.