Summary steps and insights  on project(Hyperspectral imaging data- mycotoxin prediction)

## Preprocessing steps:

1. **Missing Values** – Data doesn't have any missing values
2. **Outlier**– Data contains outliers in range of 0(2) to 15, removing directly results in losing about 20% data. So Capping outliers to extreme values.
3. **Normalization** – Standardizing or scaling wavelength reflectance values.
4. **Exploratory Data Analysis (EDA):**
    - **Average Reflectance Curve** – Examining spectral patterns.
    - **Correlation Analysis** – Identifying highly correlated bands with DON concentration(target), highly corelated bands can be selected as important feature

### Insights from PCA:

1. **Explained Variance** – PCA components capture significant variance.
2. **Cumulative Variance curve** – Helps determine the number of principal components needed, kept 97% variance with 24 principal components
3. **Scatter Plot of Principal Components** – shows how principal components covers variation in data by reducing dimension

## Model Selection, Training, and Evaluation Summary:

As data is small (500 examples) started with **linear model**,

- Not using Neural nets cause data is small that might lead to high overfitting (high variance)
- **Random-Forest** increases performance but overfits
- **XGBoost :** Captures complexity well but initially overfits.
- **Hyperparameter Tuning:** Used Random Search to optimize XGBoost, reducing variance while maintaining performance.

## Key Findings and Suggestions for Improvement:

- Data does have outlier in reflectance values that might be due to measurement error or natural thing so keeping outlier is good but instead of caping maybe log transformation would have kept variation.
- Adding more data help in reducing variance/ generalizing to unseen data
- Instead of applying PCA on all wavelengths, we can remove highly correlated wavelengths and added as independent feature after applying pca to rest of wavelengths