Name: Manav Bakliwal

Email: manavbakliwal792@gmail.com

Mobile No.: 9673130424

Prediction of Mycotoxin Levels in Corn Samples

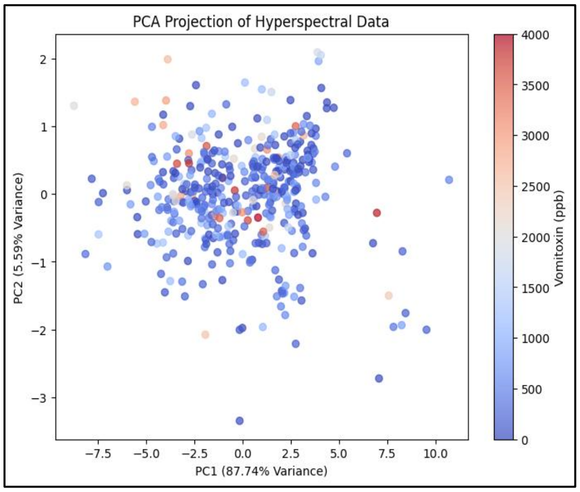
1. Data Preprocessing

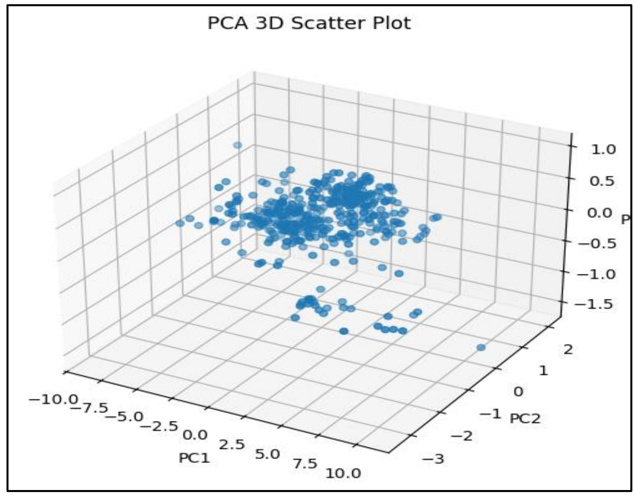
- Steps Taken:
  - Data Loading: The dataset (TASK-ML-INTERN.csv) was loaded using pandas
  - Handling Missing Values: Checked for missing values, though imputation techniques were not explicitly mentioned
  - Outlier Removal: Used Interquartile Range (IQR) method to remove extreme values in the target variable.
  - Feature Scaling: Normalization (MinMaxScaler) scales features between 0 and 1.
  - Target Scaling: Standardization (StandardScaler) ensures zero mean and unit variance.
- Rationale:
  - Removing outliers reduces noise and prevents extreme values from skewing training.
  - Scaling ensures that models sensitive to feature scale (e.g., SVM, neural networks) perform optimally

2. Dimensionality Reduction

- Principal Component Analysis (PCA):
  - PCA was applied to reduce feature dimensions while retaining maximum variance
  - Visualized transformed data to understand clustering patterns.
- Insights:
  - PCA helped determine if a lower-dimensional representation could be used for modeling.
  - A small number of principal components explaining most variance can reduce computation while maintaining accuracy.

- PCA Projection of Hyperspectral Data





3. Model Selection, Training, and Evaluation

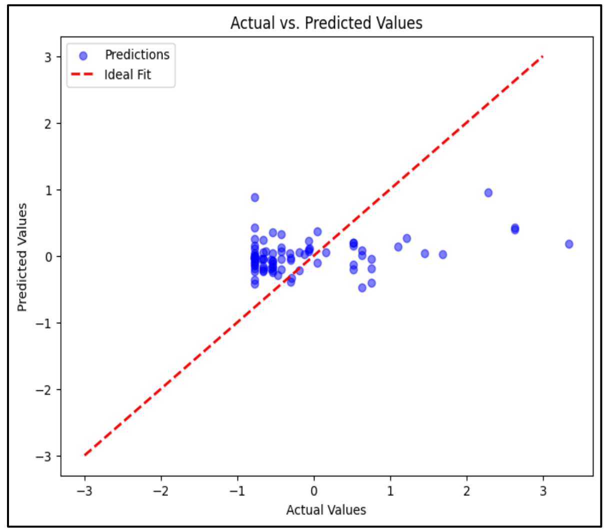
- Traditional Machine Learning Models:
  - Multiple Linear Regression: Assumed a linear relationship.
  - Support Vector Regressor (SVR): Used kernel functions for non-linearity
  - Random Forest Regressor: Handled complex feature interactions well.
  - XGBoost Regressor: Optimized gradient boosting model.

- Deep Learning Models:
  - CNN: Extracted features from spectral data.
  - CNN with Attention: Focused on important spectral bands.
  - LSTM: Captured sequential dependencies.
  - LSTM with Attention: Improved LSTM focus.
  - Multi-layer Perceptron (MLP): Used fully connected layers.

4. Results Evaluation and Conclusion
   - Evaluation Metrics:
     - Mean Absolute Error (MAE), Mean Squared Error (MSE), R-squared Score (R^2).

- Evaluation Metric Results and Scatter Plot

	Model	MAE	RMSE	R² Score
0	Multiple Regression	0.663344	0.830716	0.087034
1	Support Vector Regression	0.608705	0.888708	-0.044883
2	XGBoost	0.678535	0.815421	0.120342
3	Random Forest Classifier	0.714779	0.900628	-0.073102
4	CNN Without Attention	0.653290	0.870440	-0.002369
5	CNN With Attention	0.725516	0.891965	-0.052556
6	LSTM Without Attention	0.725999	0.892172	-0.053044
7	LSTM With Attention	0.754225	0.905789	-0.085434
8	Neural Network	0.558795	0.932261	-0.149807



- Findings:
  - Random Forest and XGBoost performed best among traditional models.
  - CNN and LSTM with Attention Mechanisms performed well in deep learning models.
  - Linear Regression and SVR had weaker performance, indicating non-linearity in data.