Spectral Data Analysis and Prediction – Report

Author: Mohd Fahad Chougle

Date: 14/3/2025

1. Introduction:

This report outlines the steps taken to preprocess spectral data, apply dimensionality reduction, train predictive models, and evaluate their performance. The goal is to develop an accurate model for predicting vomitoxin levels based on spectral reflectance data.

2. Data Preprocessing:

2.1 Data Cleaning:

- Loaded the dataset and inspected it for missing values or inconsistencies.
- Removed non-feature columns (e.g., hsi_id) and ensured the target variable (vomitoxin_ppb) was properly separated.

2.2 Feature Scaling & Normalization:

- Applied **StandardScaler** to standardize spectral features, ensuring a uniform distribution for PCA and model training.

3. Dimensionality Reduction:

3.1 Principal Component Analysis (PCA):

- **Applied PCA** to reduce dimensionality while preserving **97% variance** in the dataset.
- The number of retained components was determined automatically based on a variance threshold.
- PCA-transformed data was used for model training to improve efficiency and mitigate redundancy.

4. Model Selection & Training:

4.1 Models Considered:

Three models were tested and compared:

- **Linear Regression**
- **Random Forest**
- **XGBoost**

4.2 Model Training:

- **Data Split:** 80% training, 20% testing.
- **Hyperparameter Tuning:** Used **RandomizedSearchCV** to optimize **Random Forest** and **XGBoost** models.
- The best hyper-tuned **XGBoost** model achieved **94% accuracy**.

5. Model Evaluation:

Model	RMSE	R^2 Score	MAE
Linear Regression	10376.65	0.5111	4618.28
Random Forest	6693.15	0.7966	2635.68
XGBoost	2674.22	**0.9675**	**1523.81**

- **XGBoost performed the best**, achieving the **lowest RMSE (2674.22)** and **highest R^2 score (0.9675)**.
- **Linear Regression performed poorly** due to the non-linearity in data.
- **Random Forest showed decent results** but was outperformed by XGBoost.

6. Key Findings & Future Improvements:

6.1 Observations:

- **Dimensionality Reduction:** PCA effectively reduced feature count while maintaining high variance.
- **Model Performance:** XGBoost emerged as the best-performing model.

6.2 Potential Enhancements:

- **Further Hyperparameter Optimization**: Additional fine-tuning may improve XGBoost performance.
- **Feature Engineering**: Incorporating domain-specific features might enhance predictions.
- **Deep Learning Approach**: Exploring CNNs or LSTMs for improved predictions.
- **Transformers & Attention Mechanisms**: Could enhance predictive capabilities.

7. Deployment: Streamlit App:

A **Streamlit application** was developed for interactive model usage:

- **Users can upload spectral data** (CSV/XLSX) for real-time predictions.
- The app **applies scaling, PCA transformation, and XGBoost inference** to generate predictions.
- Results are **displayed in a table** with a **downloadable CSV output**.

8. Conclusion:

This project successfully developed a robust **vomitoxin prediction model** using **XGBoost** and **PCA-transformed spectral data**. The results highlight the importance of **feature reduction and model selection** in optimizing performance. Future enhancements may include **deep learning models** and **further fine-tuning** for improved accuracy.