# Short Report: Machine Learning Pipeline for Vomitoxin Prediction

**1. Data Preprocessing Steps and Rationale**

* **Handling Target Variable:** The target variable (vomitoxin\_ppb) exhibited a highly skewed, long-tailed distribution with extreme outliers. Multiple transformations (log, square root) were tested, but none significantly improved performance. Standard scaling was applied to both features and target for better model performance.
* **Feature Engineering:** The dataset comprised 449 spectral reflectance features. Exploratory analysis revealed a Gaussian-like distribution with some outliers. A variance-based approach was considered to eliminate redundant features, but dimensionality reduction techniques were further explored.
* **Train-Test Split & Scaling:** An 80:20 train-test split was performed, followed by standard scaling on both X and Y. The transformation was fitted on Y\_train and applied to Y\_test to maintain consistency.

**2. Dimensionality Reduction Insights**

* While feature values showed consistent variance, principal component analysis (PCA) and other dimensionality reduction techniques were tested but did not yield significant improvement.
* SHAP analysis revealed that certain features had near-zero impact on predictions, indicating possible candidates for removal.
* Variance in reflectance values remained consistent across contaminated and uncontaminated samples, suggesting no extreme feature disparities.

**3. Model Selection, Training, and Evaluation**

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **MAE** | **RMSE** | **R² Score** |
| Linear Regression | 2.2117 | 2.7485 | -2.8846 |
| Random Forest | 0.2863 | 0.8248 | 0.6502 |
| XGBoost | 0.3038 | 0.7942 | 0.6756 |

* **XGBoost (Tuned):** Best-performing model with hyperparameters optimized via grid search.
* **Residual Analysis:** Residuals followed a normal distribution, indicating unbiased predictions with consistent variance.

**4. Key Findings and Future Improvements**

* **Model Selection:** XGBoost outperformed Random Forest and Linear Regression, demonstrating better generalization.
* **Feature Importance:** SHAP analysis provided insights into influential and redundant features, suggesting scope for feature selection refinement.
* **Error Analysis:** While the model captured general trends, certain extreme values deviated from expected predictions, indicating room for improvement in handling outliers.
* **Next Steps:**
  + Experiment with ensemble learning (stacking models) for enhanced performance.
  + Apply advanced outlier detection techniques for robust training.
  + Explore domain-specific spectral transformations for improved feature representation.

The trained XGBoost model has been saved and deployed via Streamlit for practical evaluation.