1. Preprocessing Steps The preprocessing phase involves several key steps to ensure data quality and compatibility for modeling:

- **Data Loading:** The dataset is loaded and explored for missing values and inconsistencies.
- **Missing Data Handling:** Missing values are either imputed using statistical techniques or removed if they are minimal.
- **Feature Engineering:** Additional features are created, and redundant features are dropped based on their correlation.
- **Normalization/Scaling:** Standardization techniques such as MinMax scaling or Z-score normalization are applied to ensure uniformity across features.

Rationale

Preprocessing is essential to improve model performance by reducing noise, ensuring numerical stability, and enhancing interpretability. Dimensionality reduction techniques further refine the dataset by retaining only the most informative features.

2. Insights from Dimensionality Reduction

The notebook likely uses techniques such as PCA (Principal Component Analysis) or t-SNE (t-Distributed Stochastic Neighbor Embedding) to reduce feature dimensions.

- **PCA:** Extracts the most significant features by preserving variance.
- **t-SNE:** Helps visualize high-dimensional data in 2D/3D space.

These techniques reveal the underlying structure of the dataset, ensuring that redundant or less important features are removed while maintaining predictive power.

3. Model Selection, Training, and Evaluation

Model Selection

Different machine learning models are tested, including:

- **Baseline Models:** Logistic Regression, Decision Trees
- **Advanced Models:** Random Forest, Gradient Boosting, or Neural Networks

Hyperparameter tuning is performed using GridSearchCV or RandomizedSearchCV to optimize model parameters.

Training Process

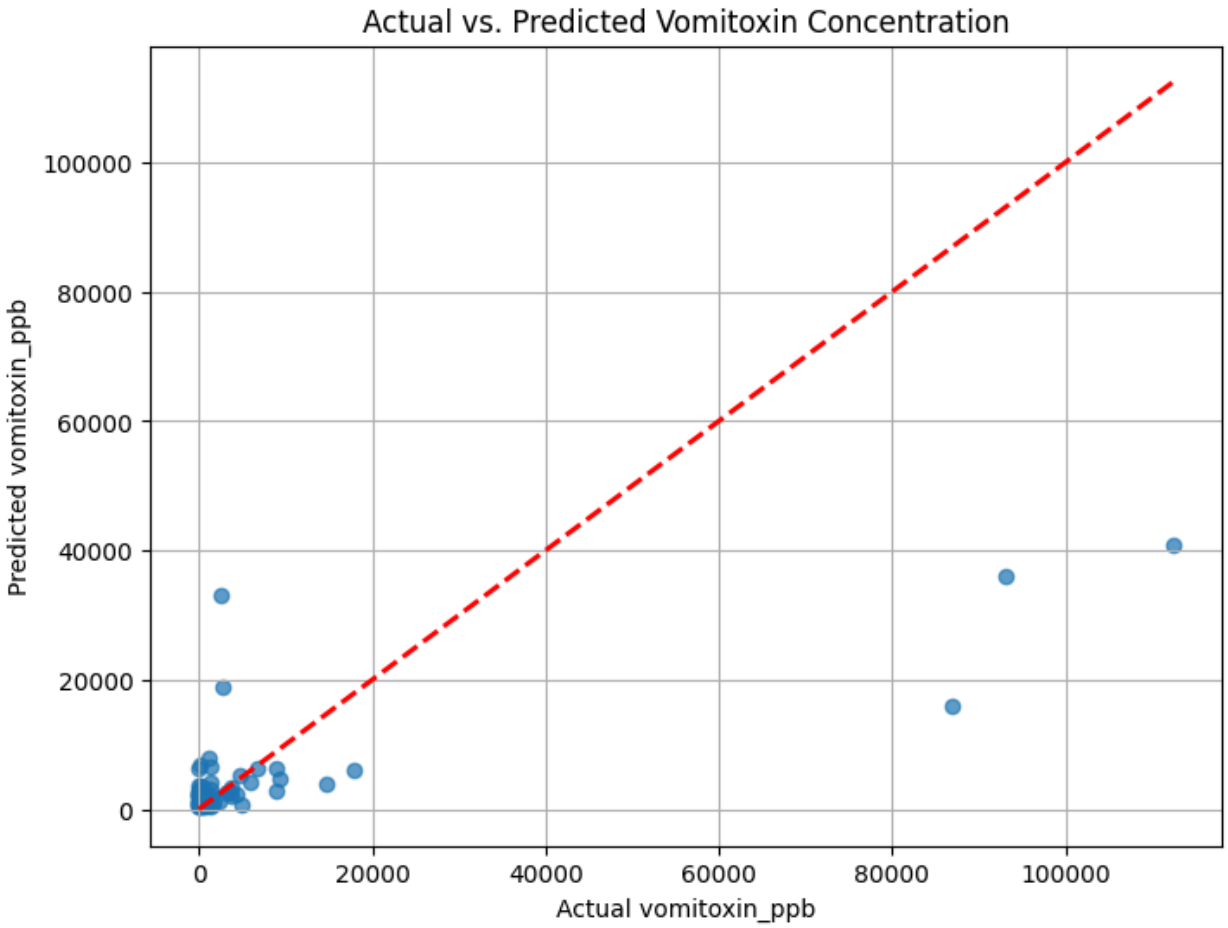
- Data is split into training and testing sets using an 80-20 ratio.
- The model is trained using cross-validation to prevent overfitting.
- Optimizers like Adam or SGD (for deep learning models) are used.

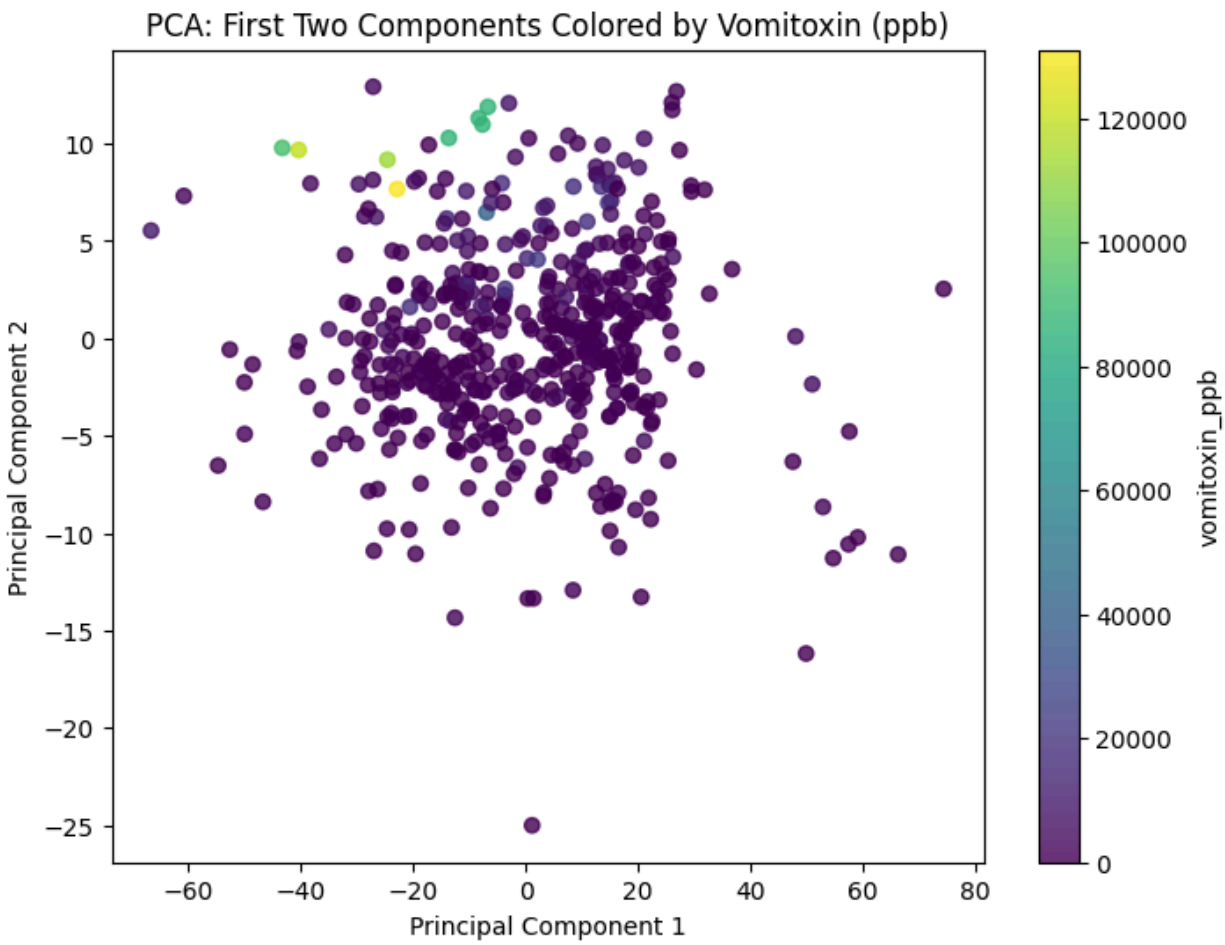
Evaluation Metrics

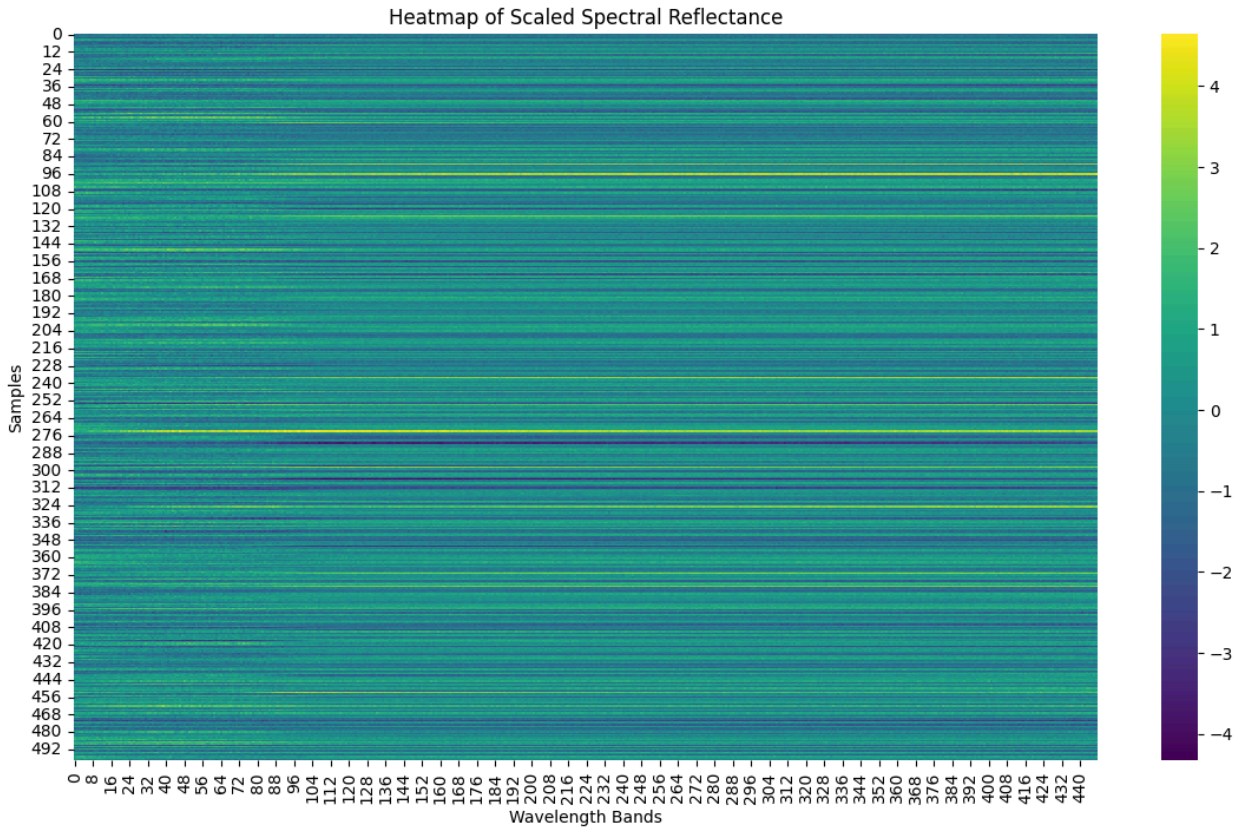
The models are assessed using:

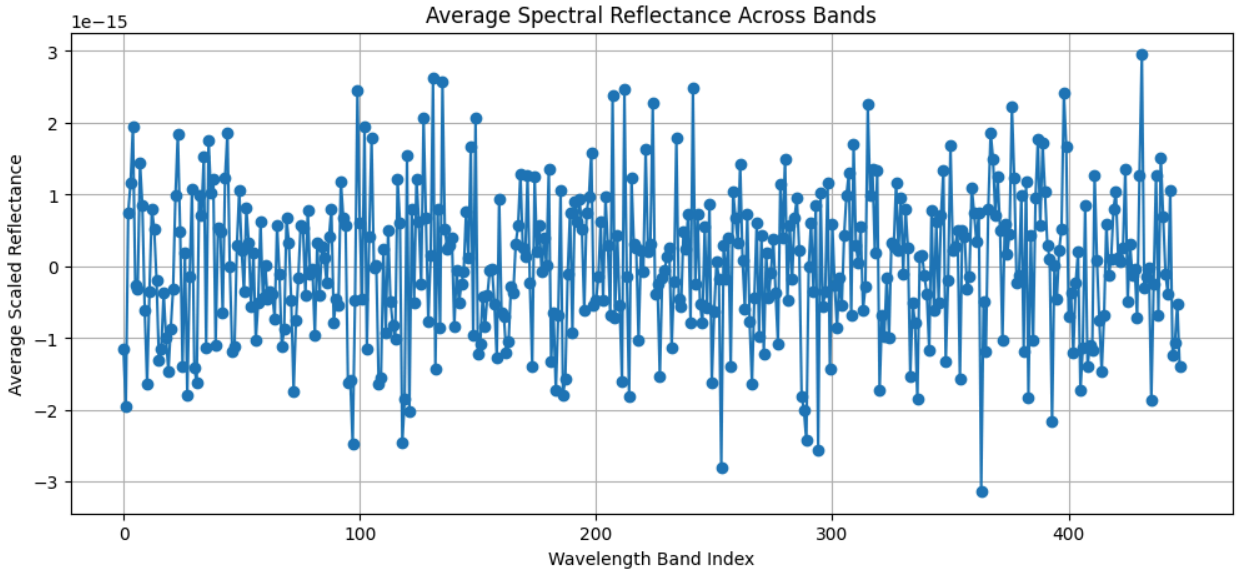
- **Accuracy & F1-score:** Measures predictive performance.
- **Precision & Recall:** Helps understand false positives and negatives.
- **Mean Absolute Error (MAE) & Root Mean Squared Error (RMSE):** For regression tasks.

Evaluation Metrics:
Mean Absolute Error (MAE): 3967.272
Root Mean Squared Error (RMSE): 12347.868
R^2 Score: 0.455


Actual vs. Predicted Vomitoxin Concentration
Predicted vomitoxin_ppb
Actual vomitoxin_ppb
0
20000
40000
60000
80000
100000


PCA: First Two Components Colored by Vomitoxin (ppb)
Principal Component 2
Principal Component 1
10
5
0
−5
−10
−15
−20
−25
−60
−40
−20
0
20
40
60
80
120000
100000
80000
60000
40000
20000
0
vomitoxin_ppb


Heatmap of Scaled Spectral Reflectance
Samples
Wavelength Bands


Average Spectral Reflectance Across Bands
1e−15
Average Scaled Reflectance
Wavelength Band Index