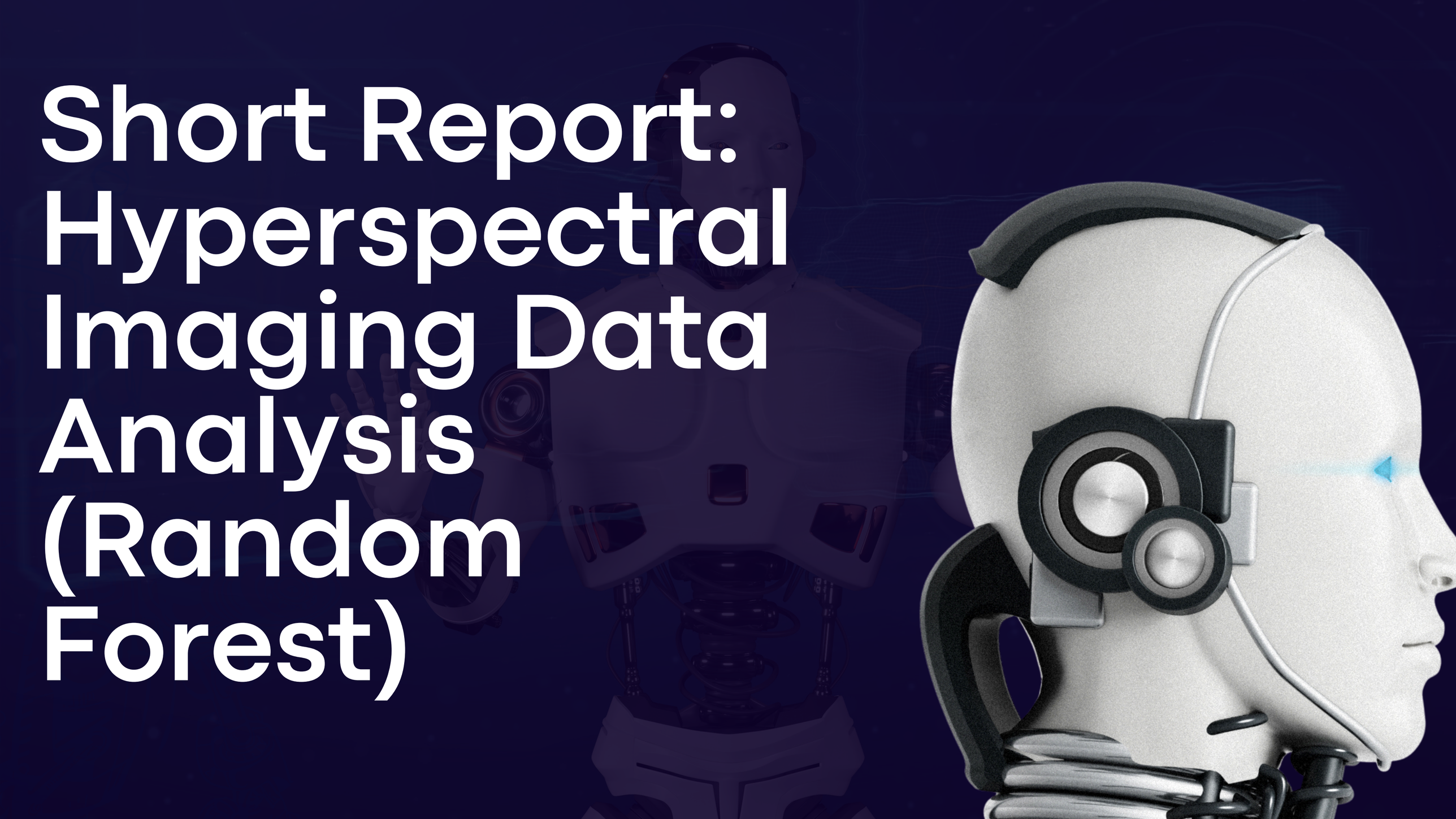
MACHINE LEARNING

This presentation will introduce our machine learning project, highlighting its objectives, methodologies, and potential impact.

www.reallygreatsite.com

Short Report: Hyperspectral Imaging Data Analysis (Random Forest)



1. Preprocessing Steps and Rationale

Data Cleaning:

- Missing Values: Imputed using column-wise mean to maintain data consistency.

- Outlier Detection & Removal: Applied the Interquartile Range (IQR) method to detect and remove extreme values, ensuring a cleaner dataset for training.

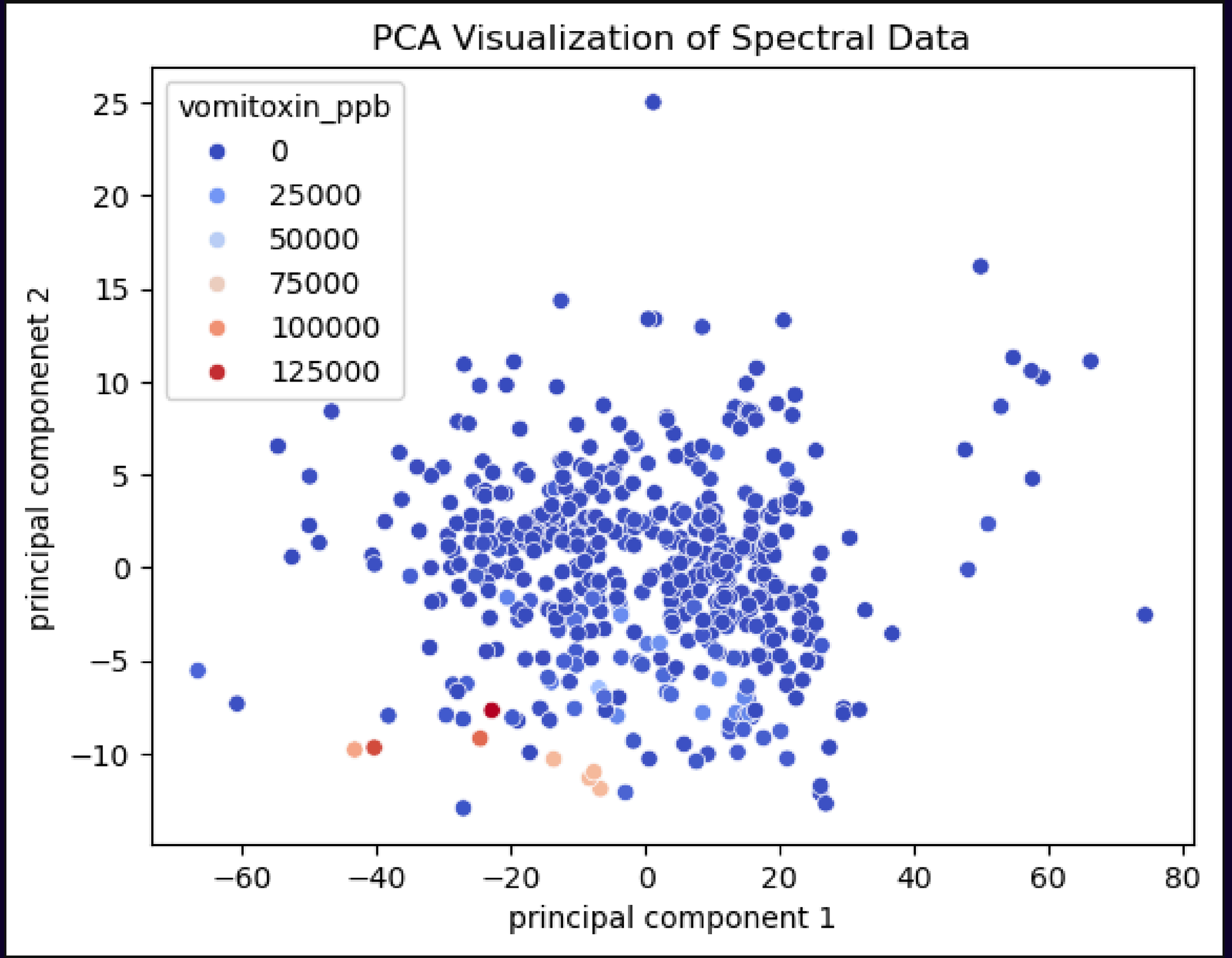
Data Transformation:

- **Feature Scaling**: Used StandardScaler to normalize spectral bands, ensuring all features have a mean of 0 and a standard deviation of 1, improving model performance.
-
- **Feature Selection**: The first column (categorical identifier) was removed before training.

2. Insights from Dimensionality Reduction

05

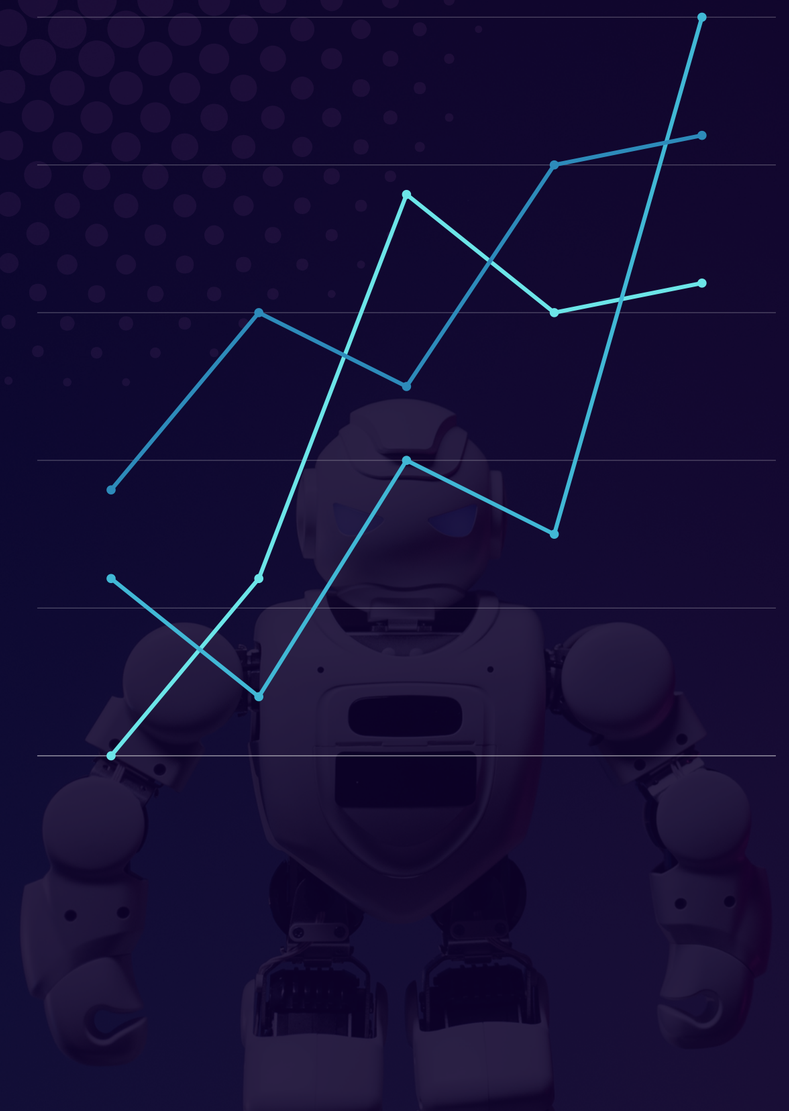
- Principal Component Analysis (PCA) was explored to reduce feature dimensions and visualize data in a lower-dimensional space.
- Most of the variance was retained, suggesting that a lower-dimensional representation could still preserve critical information.


PCA Visualization of Spectral Data
vomitoxin_ppb
0
25000
50000
75000
100000
125000
principal componenet 2
principal component 1
25
20
15
10
5
0
−5
−10
−60
−40
−20
0
20
40
60
80

3. Model Selection, Training, and Hyperparameter Tuning

Model Choice:

- Random Forest Regressor was selected due to its ability to handle high-dimensional spectral data, non-linearity, and interpretability via feature importance



Training Details:

08

- 80%-20% Train-Test Split to evaluate performance.
- **Hyperparameter Optimization using RandomizedSearchCV:**
  - Number of estimators: 100, 200, 300
  - Maximum depth: None, 10, 20
  - Minimum samples split: 2, 5, 10
  - Minimum samples leaf: 1, 2, 4
- Optimal parameters were selected to improve generalization and reduce overfitting.

Evaluation Metrics:

- **Mean Absolute Error (MAE)**: Measures absolute differences between actual and predicted values.
- **Root Mean Squared Error (RMSE)**: Penalizes large deviations more heavily than MAE.
- **R^2 Score**: Explains the proportion of variance captured by the model.

Performance Results:

- **MAE: 4594.57**
- **RMSE: 11981.91**
- **R^2 Score: 0.185**

4. Key Findings and Suggestions for Improvement

Findings:

- Feature importance analysis showed that some spectral bands contribute significantly more to prediction than others.

- The model's R^2 score indicates that a significant portion of variance is unexplained, suggesting potential improvements.

Suggested Improvements:

- Try Boosting Algorithms: Using **XGBoost** or **LightGBM** for better predictive performance.
- **Feature Engineering**: Deriving new spectral indices to enhance the model's capability.
- **Hyperparameter Fine-Tuning**: Expanding the search space for better optimization.

Conclusion :

- This study successfully applied Random Forest Regression with hyperparameter tuning to predict mycotoxin levels in corn samples using hyperspectral data. However, the current model's performance suggests room for improvement through alternative ensemble techniques, feature engineering, and advanced hyperparameter tuning.