Mycotoxin Prediction Report

KATTA BHAVANA

24.03.2025

B.TECH 3RD YEAR

INTRODUCTION

Problem Statement:
The objective of this project is to predict the concentration of vomitoxin (**DON**) in corn samples using hyperspectral imaging data.

Goal:
Develop a robust machine learning model to predict DON levels based on spectral reflectance values.

Approach:
The pipeline involves data preprocessing, exploratory data analysis (**EDA**), model training, hyperparameter tuning, evaluation, and interpretation.

DATA PREPROCESSING

Dataset Overview:
448 spectral reflectance values representing different wavelengths
Target variable: vomitoxin_ppb (DON concentration).

Data Cleaning:
Missing values were filled using median imputation.

Feature Selection:
Only numeric columns were used.
Non-relevant columns like hsi_id were removed.

Normalization:
Features were normalized using StandardScaler for better model performance.

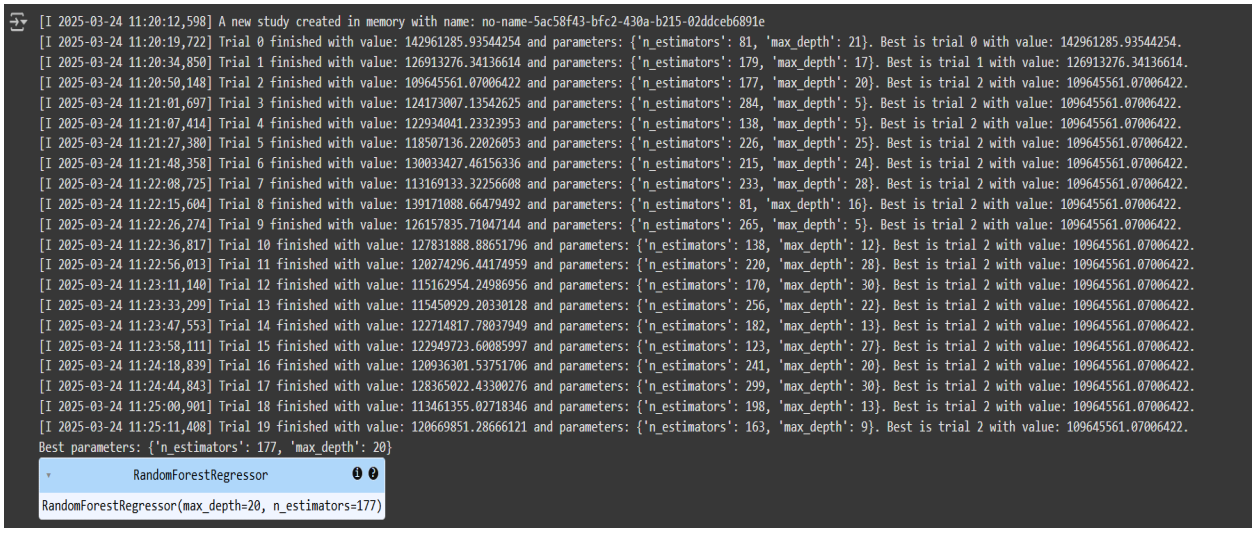
MODEL TRAINING

Model Selection:
Random Forest Regressor was chosen for its interpretability and robustness with high-dimensional data.

Hyperparameter Tuning:
Optuna was used for automated hyperparameter tuning.

Best Parameters:
`n_estimators`: 177
`max_depth`: 20

Data Splitting:
80% Training, 20% Testing using `train_test_split`.



MODEL EVALUATION

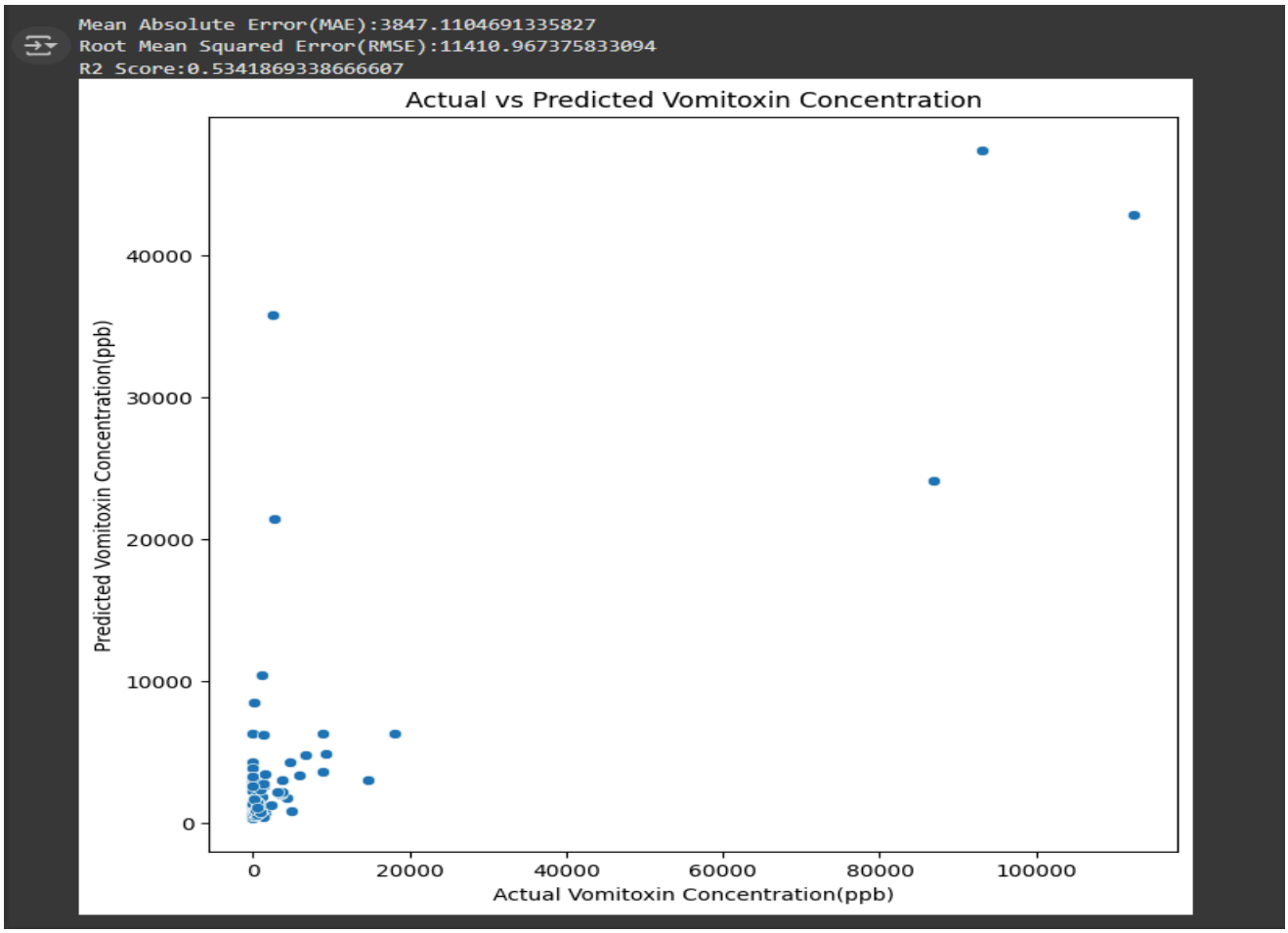
Metrics Used:

Mean Absolute Error (MAE): Measures average prediction error.
Root Mean Squared Error (RMSE): Evaluates overall prediction accuracy.
R^2 Score: Measures how well the model explains variability in the data.

Results:
After training with the best hyperparameters:
MAE: 3847.11 ppb
RMSE: 11,410.97 ppb
R^2 Score: 0.534

These results indicate the model has moderate predictive capability. While the model captures some patterns in the data, further improvements are possible.

Visualization:
A scatter plot comparing actual vs. predicted values showed a moderate correlation, indicating the model's ability to capture general trends.

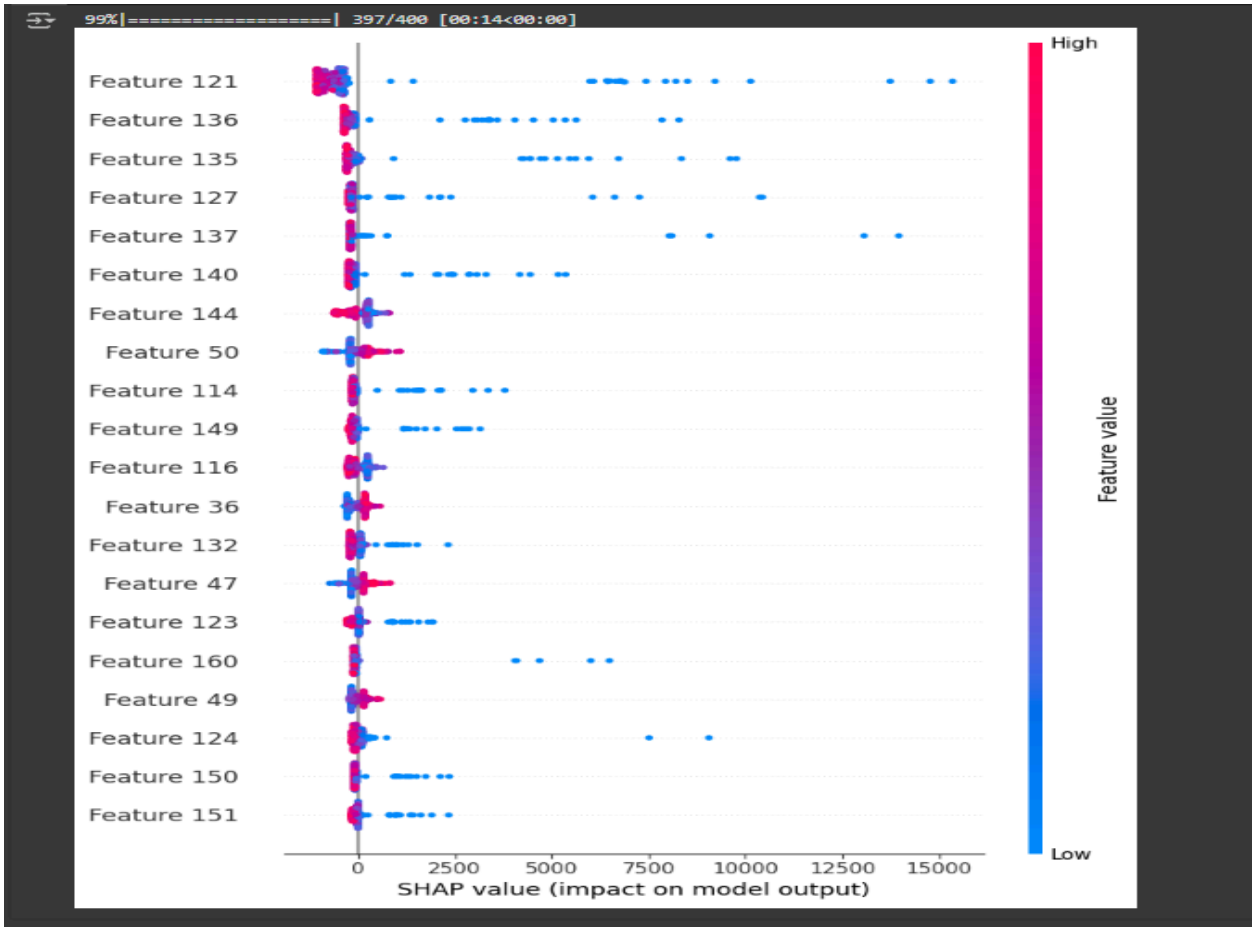


MODEL INTERPRETABILITY

SHAP (SHapley Additive exPlanations) was used to interpret feature importance.

Key Insights from SHAP Plots:
Certain spectral bands contributed significantly to the prediction.
The model relied heavily on specific wavelengths that may indicate early signs of contamination.

Some bands had lower importance, suggesting potential redundancy in the data. Further analysis of these wavelengths could enhance domain knowledge and aid in refining the model.



CONCLUSION

The Random Forest Regressor achieved a reasonable performance with an R^2 score of **0.534.**
The model shows promise in predicting vomitoxin concentration but has room for improvement.
Future improvements can include:

- Trying ensemble models like XGBoost or LightGBM.
- Performing further feature engineering to create meaningful spectral indices.
- Collecting more data or reducing noise in the reflectance measurements.

REFERENCES

1. **Scikit-Learn:** For model training and evaluation.
2. **Optuna:** For hyperparameter tuning.
3. **SHAP:** For interpretability.
4. **Matplotlib & Seaborn:** For data visualization.