**Assessment | Short Report**

**Preprocessing Steps and Rationale**

To begin with, I studied mycotoxins in food to understand their impact and the significance of spectral reflectance data. I also researched how spectral reflectance data is collected to better interpret the dataset. Basic data analysis was performed, including handling missing values, plotting histograms to examine data distribution, and using box plots to detect outliers.

The histogram analysis revealed that the data was not normally distributed. Since reflectance values range from 0 to 1, I initially applied MinMaxScaler for normalization. However, this approach did not effectively address outliers, leading me to use StandardScaler instead, which resolved the issue.

For outlier removal, I applied the Z-score method with a threshold of 3, eliminating extreme values while retaining meaningful variations in the data. Given that the dataset contained 448 wavelengths with high correlations among them, I used Principal Component Analysis (PCA) to reduce dimensionality. I retained components that explained 95% of the variance, reducing the data to six principal components.

**Insights from Dimensionality Reduction**

PCA effectively reduced the number of features from 448 to 6 while preserving most of the variance in the dataset. The correlation matrix before PCA indicated strong multicollinearity among wavelengths, validating the need for dimensionality reduction. This transformation simplified model training and reduced computational complexity without significant loss of information.

**Model Selection, Training, and Evaluation**

I experimented with multiple models, including Linear Regression, Neural Networks (MLP), and Decision Trees. Both Linear Regression and MLP showed poor performance, indicating non-linearity in the data. To address this, I transformed the data using second-degree polynomial features, leading to improved model performance.

After fine-tuning the models, I explored boosting techniques to enhance predictive accuracy. Gradient Boosting yielded an R² score of 0.94, outperforming other individual models. To further improve robustness, I implemented an ensemble technique using Decision Tree and Gradient Boosting as base models, with Linear Regression as the final predictor.

**Model Interpretability Using SHAP**

To better interpret the model’s predictions, we used SHAP (SHapley Additive exPlanations). The SHAP summary plot illustrates the impact of each feature on the model's output.

From the visualization:

* Feature 5 and Feature 1 have the highest impact on the model predictions, with high feature values generally pushing predictions positively.
* Feature 2 and Feature 4 also contribute significantly, but their SHAP values vary widely.
* Some extreme SHAP values indicate that certain observations have a disproportionately high effect on the model, reinforcing the importance of feature selection and regularization.
* Features 0 and 3 have lower impacts on model predictions, suggesting they could be less critical.

These insights validate the feature selection process and confirm that reducing dimensionality did not compromise model performance.

**Key Findings and Possible Improvements**

* PCA significantly reduced dimensionality while retaining critical information, simplifying model complexity.
* Initial models performed poorly due to the non-linearity in data, which was addressed using polynomial feature transformation.
* Gradient Boosting achieved the best performance with an R² score of 0.94.
* An ensemble approach further improved model robustness by leveraging the strengths of multiple models.
* SHAP analysis provided valuable interpretability, highlighting the most influential features in model predictions.

Future improvements could include:

* Exploring other non-linear transformation techniques such as kernel PCA.
* Tuning hyperparameters further using automated optimization techniques.
* Incorporating domain-specific feature engineering to enhance model interpretability and performance.

Additionally, deployment strategies have not been explored in this study, as I have not worked with or been trained in deployment techniques. Future work could involve learning and implementing model deployment strategies to make the model accessible for practical applications.

These steps and refinements would contribute to a more accurate and reliable predictive model for mycotoxin analysis in food.