DON CONCENTRATION PREDICTION REPORT

Data Exploration and Preprocessing

- **Data Cleaning:** The dataset is loaded in vs code and the first five rows are printed to visualize the data after that it is checked if data has zero values if found the rows with zero values are removed and also it is checked whether there are missing values in data it is checked and removed. Statistical summary of bands is checked to understand the statistical distribution of data. Boxplots and histogram and QQ plots are plotted to visualize the quantity of outliers or the distribution of outliers in data. Outliers are removed using Interquartile Range Method. The Interquartile Range (IQR) Method is a statistical technique used to identify outliers in a dataset. The IQR represents the range between the first and third quartiles (Q1 and Q3), which correspond to the 25th and 75th percentiles of the data, respectively. The IQR is used to determine the spread of the middle 50% of the data and helps in detecting values that are unusually high or low compared to the rest of the data. After removing the outliers boxplots are again plotted to visualize whether the outliers are perfectly removed or not. Noise Analysis of the outlier removed data is done to check if smoothing is required or not and it is found out that the data is appearing clean and smoothing is not required.

- **Normalization of Data:** At first it is checked if normalization or standardization is required for data it is found out that normalization is required as because the data showed high variability in terms of numeric features. Normalization is done using Min-Max Scaling Method and the Pre-processed data is saved in csv file.

- **Data Visualization:** To visualize the trend of average reflectance over wavelengths a line plot is drawn. In the given case wavelengths are plotted on x axis and the corresponding average reflectance on the y axis for trend identification and wavelength analysis. To understand correlations between data heatmaps are drawn and pairplots are drawn to understand pairwise relations between several variables.

Model Selection, Training and Evaluation

- The pre-processed data is loaded and methods like SelectKBest and Random Forest Feature Importance are used to select the best features and vif is used to address multicollinearity between selected features.

- After selecting best features data is loaded and feature scaling is done to standardize the data to bring in a range where the features have a mean of 0 and a standard deviation of 1. After that that data is splitted into training and test set and put into model training. Six models are taken that is RandomForestRegressor, GradientBoostingRegressor, Ridge, Lasso, SVR (Support Vector Regression),Neural Network Model after training the model in data and evaluating it, it is found out that the model Lasso gave good R2 score and low MAE and RMSE value which makes Lasso the best model and the model is saved in pkl format for further use. Scatter plot is plotted to compare actual values with predicted values and residual analysis is performed to assess the model and SHAP is used to explain the model's predictions. Further accuracy of model can be improved by using K fold cross validation technique and hyperparameter tuning methods.

Streamlit App Creation

- A streamlit app is created which will be serving as an interactive user interface to predict DON concentration that is the features which is selected for the trained model will be taken and the value will be put by the user and the vomitoxin concentration will be predicted based on the input features. This app helps in quick decision making based on model's predictions.

Unit Tests Functionalities

- Logging setup is created to track the execution flow of program especially when model is deployed in production environments. It records model loading success or failure, prediction success, warnings and errors. It helps you to monitor the system For potential issues such as version issues or loading failures. The model is loaded and prediction is done to check if prediction has been done correctly or not.
- Unit tests are done to ensure that predict function behaves correctly under different conditions. Automated validation of prediction logic helps ensure that the model and its integration with code continue to work as expected as changes are made over time. It helps in maintaining reliability and Maintainability and easy debugging and confidence in Model Predictions.

FastAPI Application Creation:

- FastAPI is used to create the web service which allows users to send HTTP requests to make predictions about vomitoxin concentration in corn samples. The API expects to receive an input of 23 features which it scales and uses to predict the vomitoxin

level by using a pre-trained model saved in pkl format. It helps in ensuring proper data scaling, input validation, and error handling and uses logging for monitoring and debugging.