ImagoAI ML TASK

● Preprocessing steps and Rationale

- At first check the dataset if it contains null values, in this case no null values so use df.drop to drop rows with null values but normally use sklearn Imputer method by using different strategy like mean,median or mode.

- Next step to standardize the value (z-score normalization) was applied to ensure features had zero mean and unit variance, improving model convergence.

- Spectral reflectance profiles were plotted to observe trends across wavelengths.
- Heatmaps were generated to compare spectral variations and DON concentrations.

● Insights from dimensionality reduction

- PCA was applied to reduce feature dimensionality while retaining significant variance. As from the PCA distribution graph the majority of data points are concentrated near the center, suggesting that most samples have similar spectral characteristics.

- The spread of data along the PC1 is wider than along PC2, suggesting that PC1 captures more variance in the spectral data.

● Model Selection,training and evaluation details

- Used CatboostRegressor model which is a gradient boosting algorithm optimized for categorical data and fast training, which also achieved better training and validation R-squared error than XGBoost,Random Forest, SVM or Neural Network.

- All the model are hyperparameter tuned using GridSearchCV and KFold cross validation
- The Metrics from the CatboostRegressor is as follows:
Mean Absolute Error of Catboost Model: 3184.939832486601
Root Mean Square Error of Catboost model: 6482.071702321353
R2 Score of Catboost model: 0.8496877414518142
- Which is the best score among all the models listed above.

● Key Findings and suggestions for improvement

- Key Findings

- PCA provided valuable insights into data structure.
- CatBoost outperformed XGBoost and Neural Networks in predictive accuracy.
- Data scaling and hyperparameter tuning significantly improved model performance.

- Suggestions

- Experiment with feature engineering techniques (e.g., polynomial features, domain-specific transformations).
- Utilize advanced ensembling techniques (stacking multiple models) for further improvement.