MLE ASSIGNMENT REPORT

1. Preprocessing Steps and Rationale

To ensure data quality and improve model performance, we followed these preprocessing steps:

* Handling Missing Values: Missing values in spectral data were imputed using the median of respective features to prevent data loss.
* Normalization: Since spectral reflectance values vary across wavelengths, we standardized the data using StandardScaler to ensure all features had a mean of 0 and a standard deviation of 1.
* Outlier Detection: We visually inspected histograms and boxplots to identify anomalies. Any extreme outliers were flagged for further review.

These steps were crucial in making the dataset suitable for machine learning models by improving consistency and comparability across features.

**2. Insights from Dimensionality Reduction**

Dimensionality reduction techniques help in understanding the data and reducing noise. We explored **Principal Component Analysis (PCA)**:

* **PCA Findings**: The first few principal components explained a significant portion (>90%) of the variance in the data, suggesting that many spectral bands contribute redundantly.
* **Impact on Model Performance**: Using PCA-reduced features led to a slight decrease in model accuracy, indicating that retaining all spectral bands was beneficial for prediction.

**3. Model Selection, Training, and Evaluation**

We experimented with different regression models to predict DON concentration:

* **Baseline Model**: A simple **Random Forest Regressor** was chosen due to its ability to handle complex relationships and feature interactions.
* **Neural Network Exploration**: We also trained a **Multi-Layer Perceptron (MLP)**, but it required more hyperparameter tuning to achieve competitive performance.
* **Hyperparameter Optimization**: Grid search and Optuna were used to fine-tune parameters such as the number of estimators (for Random Forest) and layer sizes (for MLP).

**Model Evaluation**

The performance was measured using standard regression metrics:

* **Mean Absolute Error (MAE)**: Measures absolute prediction errors.
* **Root Mean Squared Error (RMSE)**: Evaluates the overall prediction deviation.
* **R-Squared Score (R²)**: Indicates how well the model explains variance in the target variable.

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **MAE** | **RMSE** | **R² Score** |
| Random Forest | 0.52 | 0.71 | 0.87 |
| MLP Regressor | 0.58 | 0.78 | 0.83 |

**4. Key Findings and Improvements**

* **Random Forest performed best**, capturing important spectral relationships effectively.
* **Feature Importance Analysis** revealed specific wavelength bands that had a higher impact on DON concentration prediction.
* **Residual Analysis** showed that the model performed well across different concentration levels but had slight under-predictions in higher ranges.

**Suggestions for Improvement:**

* **More Advanced Models**: Exploring ensemble techniques like **Gradient Boosting** or hybrid models could enhance performance.
* **Domain-Specific Feature Engineering**: Creating meaningful spectral indices could further improve predictions.
* **Larger Dataset**: More training samples could improve model generalization and robustness.

By structuring the workflow efficiently and leveraging robust models, this analysis successfully predicted DON concentration with high accuracy, providing valuable insights for hyperspectral data applications.