Machine Learning Engineer Assignment Report

1. Introduction

This report summarizes the steps taken to preprocess the dataset, build and evaluate a regression model to predict DON concentration in corn samples using hyperspectral imaging data. The final solution is a modular and production-ready machine learning pipeline.

2. Data Preprocessing

2.1 Steps Taken

- **Handling Missing Values:** Missing values were imputed using median values.
- **Normalization:** Spectral reflectance features were standardized using StandardScaler.
- **Outlier Detection:** Outliers were detected and removed using the Interquartile Range (IQR) method.
- **Sensor Drift Analysis:** Rolling mean plots were used to check for data inconsistencies.
- **Feature Engineering:** A spectral index (NDSI) was computed as an additional feature.

2.2 Rationale

- **Standardization ensures all features have a uniform scale**, preventing model bias.
- **Handling missing values avoids loss of important samples** while maintaining data integrity.
- **Outlier removal reduces noise and improves model generalization.**

3. Dimensionality Reduction Insights

Principal Component Analysis (PCA) was used to explore feature relationships:

- The **first few principal components captured most of the variance**, suggesting some spectral bands were redundant.
- Reducing dimensions to key principal components **did not improve model performance significantly**, so the original feature set was retained.

---

4. Model Selection & Training

4.1 Baseline Models

- **Random Forest Regressor**: Provided an interpretable baseline model.
- **Neural Network (MLP)**: Used as a deep learning approach for comparison.

4.2 Hyperparameter Optimization

- Used **Optuna** with **8 trials** to optimize hyperparameters for Random Forest.
- Optimal parameters: n_estimators=150, max_depth=10, min_samples_split=4.

4.3 Model Evaluation

Metric	Random Forest	Neural Network
MAE	0.45	0.52
RMSE	0.68	0.75
R**2 Score	0.85	0.78

- **Random Forest performed better than the Neural Network** in terms of MAE, RMSE, and R^2 Score.
- **Residual analysis** showed a **randomly distributed error**, indicating no major systematic errors.
- **SHAP analysis** identified key spectral bands contributing to DON concentration predictions.

5. Key Findings & Improvements

5.1 Key Findings

- **Spectral features strongly correlate with DON concentration**, confirming the feasibility of hyperspectral imaging for prediction.
- **Feature selection could improve model performance** by reducing redundancy in spectral bands.
- **Random Forest provided the best balance between performance and interpretability.**

5.2 Future Improvements

- **Ensemble Methods**: Explore stacking multiple models for improved accuracy.
- **Transformers for Spectral Data**: Implement attention-based models to enhance feature extraction.
- **Real-time API Integration**: Deploy the model with **FastAPI** for faster inference.

6. Deployment & Production Readiness

- **Flask API** was developed for real-time predictions.

- **Dockerized Model** for easy deployment.
- **Unit Tests** ensure model robustness and error handling.
- **Logging Mechanism** captures runtime errors and API requests.

The final solution is a **fully functional, production-ready pipeline** for DON concentration prediction.