Task for ML Intern

Objective

This assignment assesses your ability to process hyperspectral imaging data, perform dimensionality reduction, and develop a machine learning model to predict mycotoxin levels (e.g., DON concentration) in corn samples.

Problem Statement

You are provided with a compact hyperspectral dataset containing spectral reflectance data from corn samples across multiple wavelength bands. Your task is to:

- Preprocess the data (e.g., handle missing values, normalize features).
- Visualize spectral bands to explore data characteristics.
- Reduce dimensionality using PCA or t-SNE and interpret the results.
- Train a machine learning model (e.g., Random Forest, XGBoost, or a simple neural network) for regression (or classification, if specified).
- Evaluate the model and present actionable insights.

Dataset Description

- Features: Spectral reflectance values across multiple wavelength bands (columns).
- Rows: Individual corn samples.
- Target Variable: DON concentration (continuous, for regression).

Tasks

1. Data Exploration and Preprocessing

- Load the dataset and inspect for missing values, outliers, or inconsistencies.
- Apply normalization or standardization to the spectral data as needed.
- Visualize spectral bands (e.g., line plots for average reflectance, heatmaps for sample comparisons).

2. Dimensionality Reduction

- Implement Principal Component Analysis (PCA) or t-SNE to reduce feature dimensions.
- Report the variance explained by the top principal components (for PCA) or clustering patterns (for t-SNE).
- Visualize the reduced data (e.g., 2D/3D scatter plots).

3. Model Training

- Select a model: Deep Learning, CNN, GNN, or LSTM.
- Split the dataset into training (e.g., 80%) and testing (e.g., 20%) sets.
- Train the model and optimize hyperparameters (e.g., using grid search or random search).

4. Model Evaluation

- Evaluate using regression metrics:
    - Mean Absolute Error (MAE)
    - Root Mean Squared Error (RMSE)
    - R^2 Score
- *(If adapted to classification: Accuracy, Precision, Recall, F1-Score)*
- Visualize results:
    - Scatter plot of actual vs. predicted values (regression).
    - *(Optional) Confusion matrix (classification).*
- Summarize model performance and limitations.

Deliverables

Submit a GitHub repository containing:

1. Jupyter Notebook or Python Script:
    - Clean, modular, and well-commented code covering all tasks.
2. Short Report (1-2 pages, PDF or Markdown):
    - Preprocessing steps and rationale.
    - Insights from dimensionality reduction.
    - Model selection, training, and evaluation details.
    - Key findings and suggestions for improvement.
3. README File:
    - Instructions to install dependencies and run the code.
    - Brief overview of the repository structure.

Evaluation Criteria

- Code Quality (30%): Clean, organized, and documented code.
- EDA & Visualization (25%): Effective data exploration and clear visualizations.
- Model Performance (25%): Appropriate model choice, training, and evaluation.
- Interpretability (20%): Insightful explanations and improvement ideas.

Bonus (Optional)

- Implement a attention mechanism, or transformer and compare performance.
- Create a Streamlit app for interactive predictions from user-uploaded spectral data.

Submission Guidelines

- Deadline: March 14, 2025
- Submission: Email the GitHub repository link to satyam.kumar@imagoai.com.

Tips for Success

- Focus on clarity and simplicity in your approach.
- Justify your choices (e.g., preprocessing techniques, model selection).
- Highlight trade-offs or challenges encountered.

Good luck, and we look forward to your submission! 🚀