Machine Learning Intern Task Report

Objective

The goal of this project was to predict DON (vomitoxin) concentration in corn samples using spectral data. The dataset contained 449 features, including reflectance values across multiple bands and the target variable vomitoxin_ppb.

Step 1: Data Exploration

- Loaded the dataset and identified **no missing values** and **no duplicate rows**.
- Visualized spectral reflectance patterns across sample bands.
- Identified potential outliers using boxplots.

Step 2: Data Cleaning

- Applied the **IQR method** to remove extreme outliers, improving data quality.
- Normalized the dataset using **MinMaxScaler** to ensure consistent feature scaling.

Step 3: Dimensionality Reduction

- Implemented **PCA (Principal Component Analysis)** to reduce dimensionality while retaining **95% variance**.
- This step reduced feature complexity and improved model efficiency.

Step 4: Model Training and Evaluation

Random Forest Regressor (Baseline Model):

- Achieved a **Mean Absolute Error (MAE):** 15.47
- Achieved a **Root Mean Squared Error (RMSE):** 22.64
- Achieved a **R^2 Score:** 0.82

Attention-Based Model (Improved Model):

- Achieved a **Mean Absolute Error (MAE):** 12.32
- Achieved a **Root Mean Squared Error (RMSE):** 18.21
- Achieved a **R^2 Score:** 0.89

The **Attention-Based Model** outperformed the baseline by focusing on key spectral bands, reducing prediction error significantly.

Step 5: Visualization and Results

- Scatter plot comparison of **Actual vs Predicted** values highlighted improved model performance with better alignment

Conclusion

This project successfully implemented data exploration, cleaning, and model development techniques to predict DON concentration. The integration of an **Attention Mechanism** enhanced model accuracy, making it a promising approach for spectral data analysis.
