Short Report: ML Intern Task

1. Preprocessing Steps and Rationale

Data Loading and Cleaning

- The dataset was loaded using `pandas`.
- Missing values in numeric columns were filled with their respective column means to maintain data integrity.
- The target variable **vomitoxin_ppb** was separated from features.
- Only numeric features were retained for analysis.

Feature Scaling

- Standardization was applied using `StandardScaler` from `sklearn.preprocessing` to normalize spectral reflectance values.

2. Dimensionality Reduction Insights

Principal Component Analysis (PCA)

- PCA was implemented to reduce feature dimensions while preserving variance.
- The top principal components explaining the highest variance were identified.
- A 2D scatter plot visualized the transformed data.

3. Model Selection, Training, and Evaluation

Model Selection

- A **Random Forest Regressor** was chosen due to its robustness in handling high-dimensional data and non-linearity.
- Data was split into **80% training** and **20% testing**.

Training Process

- The model was trained using the training set.
- Hyperparameter tuning (e.g., number of estimators, max depth) was explored.

Model Evaluation

- The model was evaluated using:
  - **Mean Absolute Error (MAE**): Measures average absolute errors.
  - **Root Mean Squared Error (RMSE):**Penalizes larger errors more heavily.
  - **R^2 Score**: Represents the proportion of variance explained by the model.
- A scatter plot of actual vs. predicted values illustrated performance.

4. Key Findings and Suggestions for Improvement

Findings

- PCA reduced dimensionality effectively while preserving information.
- Random Forest performed well, achieving a reasonable R^2 score.
- Feature scaling improved model performance.

Suggestions for Improvement

- Experiment with other models like **XGBoost or Neural Networks.**
- Perform feature engineering to extract meaningful spectral patterns.
- Test additional dimensionality reduction techniques like **t-SNE**.
- Implement automated hyperparameter tuning for optimization.