Preprocessing Steps and Rationale

1. **Handling Missing and Infinite Values**
    - Replaced inf values with NaN to ensure data consistency.
2. **Log Transformation on Target Variable**
    - Applied log transformation to vomitoxin_ppb to address sparsity and handle zeros in the target variable, normalizing its distribution.
3. **Standardization of Spectral Features**
    - Standardized spectral features using StandardScaler to achieve zero mean and unit variance, ensuring equal contribution of features during modeling.
4. **PCA for Dimensionality Reduction**
    - Applied PCA to reduce dimensionality, with the first 4 principal components capturing most of the variance (as shown in the cumulative variance plot). This reduced complexity while retaining essential information.
5. **Outlier Detection and Management**
    - Identified outliers using the IQR method and addressed them to prevent skewing model performance.

Rationale:

- Log transformation was used to normalize the skewed target variable and handle zeros.
- PCA was applied to reduce dimensionality, with the first 4 components capturing significant variance, ensuring efficient feature representation.

Model Selection, Training, and Evaluation

- **Model**: KNN Regression with Multiplicative Signal Correction (MSC) for noise reduction.
- **Hyperparameters**: n_neighbors=25, metric=euclidean, weights=distance.
- **Metrics**:
    - MSE: 6.1307, RMSE: 2.4760, MAE: 1.8357, SMAPE: 49.14%, R^2: 0.2448.

Key Findings & Suggestions

- **Findings**: MSC improved noise handling; performance peaked at n_neighbors=25. Moderate R^2 indicates potential for improvement.
- **Suggestions**: Try advanced models (e.g., Random Forest, Neural Networks), additional preprocessing (e.g., Savitzky-Golay), and hyperparameter tuning.

Please refer the notebook for the details about the DL models, and other ML models for their performances ,and the conclusions made from it, could’nt fit them here