Machine Learning Pipeline Report

1. Data Preprocessing
- We started by loading the dataset and removing any non-numeric columns.
- We chose `vomitoxin_ppb` as our target variable.
- We normalized the features using `StandardScaler` to ensure they were on a similar scale.
- The data was then split into training (80%) and testing (20%) sets.

2. Dimensionality Reduction Insights
- We didn't use any specific dimensionality reduction techniques.
- Feature scaling helped improve the convergence of our models.

3. Model Selection and Training

Random Forest Regressor
- We picked this model because it can handle nonlinear relationships well.
- We used 100 estimators for training.
- The model was trained on the preprocessed data without any reshaping.

LSTM Neural Network
- This model was chosen for its ability to detect sequential patterns.
- It had two LSTM layers with dropout to prevent overfitting.
- We used the Adam optimizer and Mean Squared Error as the loss function.
- The data was reshaped to fit the LSTM's input requirements.

4. Model Evaluation
- We evaluated the models using Mean Absolute Error (MAE), Mean Squared Error (MSE), and R^2 Score.
- The Random Forest model performed well with a lower MAE.
- The LSTM model needed more training epochs to achieve better performance.

5.Key Findings & Suggestions
- The Random Forest model provided faster and more interpretable results.
- The LSTM model could benefit from further hyperparameter tuning and more training epochs.
- Consider exploring feature selection to optimize the models.
- Collecting more data could help improve the accuracy of both models.