REPORT

Preprocessing Steps and Rationale

The dataset (TASK-ML-INTERN.csv) contains 500 samples with 448 hyperspectral features (0-447) and vomitoxin_ppb (0-131,000 ppb). Preprocessing includes loading via pd.read_csv(), scaling with StandardScaler (assumed from imports), and splitting via train_test_split. Scaling ensures equal feature contribution for the neural network, while splitting evaluates generalization. The skewed target range suggests a need for log transformation, not yet applied.

Insights from Dimensionality Reduction

PCA and TSNE imports indicate dimensionality reduction. PCA likely reduces 448 features to fewer components capturing most variance, addressing redundancy in hyperspectral data. t-SNE may visualize sample clustering by DON levels. This suggests a compact feature set (e.g., 10-20 components) could suffice, simplifying modeling.

Model Selection, Training, and Evaluation

A Sequential neural network (MLP) from tensorflow.keras was selected for its ability to model non-linear relationships. Training used a train-test split, targeting vomitoxin_ppb with probable MSE/MAE loss. Evaluation includes a scatter plot (y_test vs. y_pred) and metrics (MAE, MSE, R^2), assessing prediction accuracy, though results aren't shown.

Conclusion

The approach leverages HSI and neural networks for DON prediction, with preprocessing and reduction tackling complexity. Refining the target, model, and validation could boost performance for practical use.