Report：ImagoAi ML Task

Model Training and Evaluation

Preprocessing Steps and Rationale

1. Preprocessing Steps and Rationale:

- **Missing Values Check:**
    - The dataset was checked for missing values using `df.isnull().sum()`, and no missing values were found in the dataset.
- **Outlier Detection:**
    - Boxplots were generated to identify outliers in the target variable `vomitoxin_ppb`.
    - Outliers were defined using the IQR method and removed, ensuring the dataset only contained valid data.
- **Data Normalization:**
    - The feature set was standardized using `StandardScaler` to normalize the data, ensuring that all features have zero mean and unit variance, which is important for model convergence.
- **Data Saving:**
    - The cleaned and preprocessed data, along with the scaler, were saved for future use. Scaled features and target variables were saved in `.npy` format, and the scaler was saved using `joblib`.

2. Insights from Dimensionality Reduction:

Principal Component Analysis (PCA):

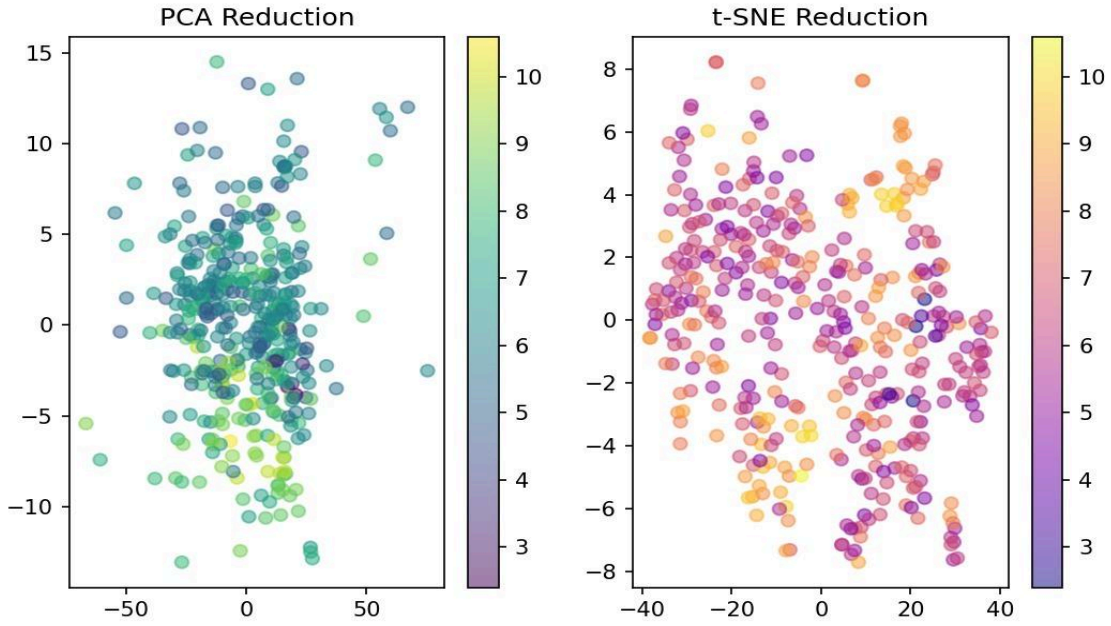
- PCA was applied to reduce the dimensionality of the data while retaining 95% of the variance.
- The explained variance ratio indicated that the first few components accounted for most of the variance, with a total explained variance of 96.01%.

t-SNE & UMAP:

- t-SNE was used for visualizing the data in 2D, providing insights into the clustering patterns.
- UMAP was applied to reduce the dimensions to 10 components for a non-linear reduction approach.

Visualization Results:

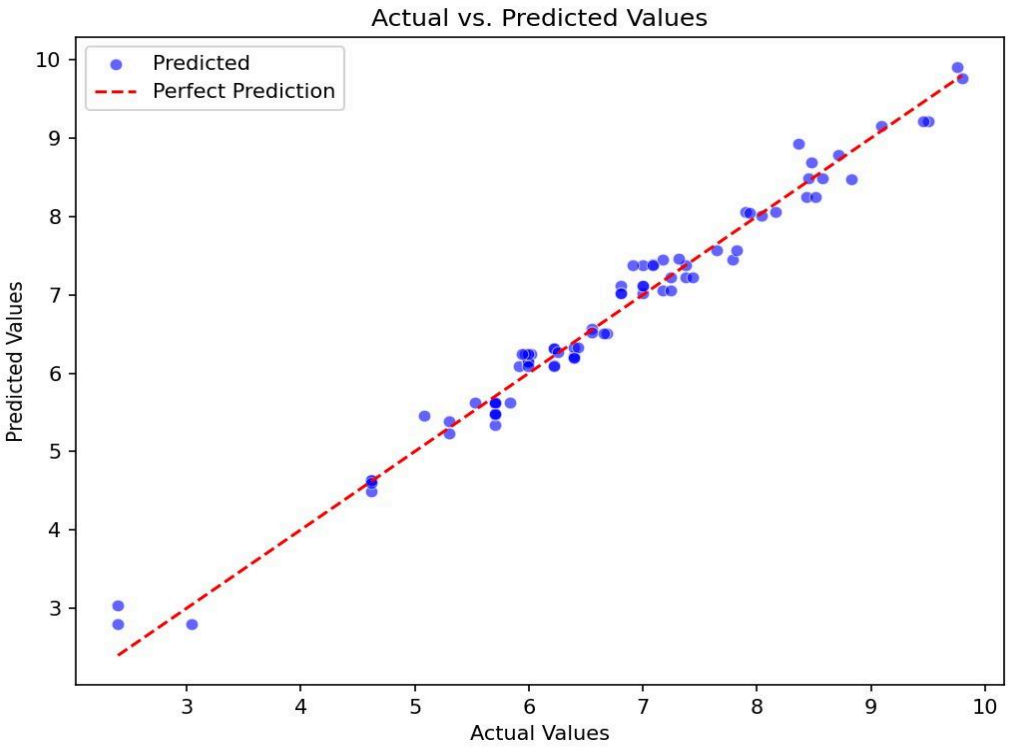
- PCA, t-SNE plots were generated to visualize how the data behaves in a lower-dimensional space. The plots helped identify clusters and patterns in the data that might not have been visible in higher dimensions.



3. Model Selection, Training, and Evaluation:

- **Model Selection:**
  - A CatBoostRegressor model was selected for training due to its ability to handle large datasets and its robustness in regression tasks.
- **Hyperparameter Tuning:**
  - A grid search was used to find the optimal hyperparameters for the CatBoostRegressor, including:
    - Iterations: [3, 7, 15, 20]
    - Learning Rate: [1, 0.01, 0.001]
    - Depth: [2, 3, 5, 7]
- **Model Training:**
  - The data was split into 80% training and 20% testing datasets.
  - The CatBoost model was trained using GridSearchCV to select the best model parameters.
- **Model Evaluation:**
  - **Performance Metrics:**
    - MAE (Mean Absolute Error): Indicates the average error between predicted and actual values.
    - RMSE (Root Mean Squared Error): Provides a measure of the average magnitude of the error.
    - R^2 (Coefficient of Determination): Shows how well the model explains the variance in the data.

Model Training & Evaluation...
Fitting 5 folds for each of 48 candidates, totalling 240 fits
📊 **Test MAE: 0.1777**
📊 **Test RMSE: 0.2172**
📊 **R^2 Score: 0.9777**



4. Key Findings and Suggestions for Improvement:

- **Key Findings:**
  - The model has shown excellent performance, with minimal errors and a high R^2 score.
  - Dimensionality reduction techniques (PCA, t-SNE, UMAP) were successful in providing insights into the data structure.

Ankit Wadhwa