Brief Report: Methodology Used in Data Processing, Feature Engineering, and ML Model

1. Data Processing:

The data processing steps involve reading the labeled dataset and splitting it into features (x) and labels (y). Additionally, we also read the unlabeled dataset for making predictions later.

Preparing Features and Labels: Next, we separate the target variable (label) from the features. The first column of the DataFrame is considered as the target variable (y), and all the remaining columns are used as features (x).

Flattening Labels: To ensure proper compatibility with the scikit-learn model, we flatten the target variable using the `flatten` function from NumPy.

Train-Test Split: We split the labeled dataset into training and testing sets using the `train_test_split` function from scikit-learn. The training set will be used to train the logistic regression model, and the testing set will be used to evaluate the model's performance.

2. Feature Engineering:

There is no explicit feature engineering used in our code.

3. Machine Learning Model:

The machine learning model used in this code is logistic regression, which is a popular binary classification algorithm.

Saving Predictions: Finally, the predicted labels for the unlabeled data are added as a new column to the original unlabeled dataset. The complete dataset (unlabeled data with predicted labels) is saved to a new CSV file called "output.csv" .

Conclusion:

The methodology followed involves basic data processing, where the labeled dataset is read, features and labels are extracted, and the data is split into training and testing sets. No specific feature engineering is used. The logistic regression model is then trained on the labeled data and used to predict labels for the unlabeled data, with the final output saved in a new CSV file.