# Verification of Data Handling in Model Development

To ensure that at no stage the model in the provided notebook was tested on the training data, confirming adherence to best practices in machine learning. After analysis the whole code and the documentation I would present the:

**Summary of Findings**

1. **Data Splitting**:
   * The dataset was properly split into training and testing sets using an 80-20 ratio.
   * The split ensured that:
     + **Training Data** (X\_train, y\_train) was used for model training and hyperparameter tuning.
     + **Testing Data** (X\_test, y\_test) was reserved exclusively for evaluating the model's performance on unseen data.



1. **Model Training**:
   * A machine learning pipeline was created to include preprocessing steps (e.g., feature scaling) and logistic regression for prediction.
   * Hyperparameter tuning was conducted using **GridSearchCV** with cross-validation (cv=5), ensuring robust evaluation during training.
   * All training steps and evaluations during tuning were confined to the training set.

A computer screen shot of a black rectangular with white text

Description automatically generated

1. **Model Evaluation**:
   * The testing set (X\_test, y\_test) was used for final evaluation after training.
   * No evidence was found of the model being tested on the training data (X\_train, y\_train).