Splitting the Dataset into Training and Testing Sets

The dataset is split into training and testing sets with an 80-20 split, where 80% of the data is used for training and 20% for testing. This split ensures that the model has sufficient data to learn from while keeping a portion of the data for unbiased evaluation.

Size and Composition

Training Set

- **Size:** The training set consists of 80% of the total data.
- **Composition:** It includes the scaled features and corresponding labels for training the machine learning model.

Testing Set

- **Size:** The testing set consists of 20% of the total data.
- **Composition:** It includes the scaled features and corresponding labels for evaluating the machine learning model.

Procedure:

- To evaluate the model's performance, we split the dataset into two parts: a training set and a testing set. The training set (80% of the data) is used to train the model, while the testing set (20%) is used to validate its performance on unseen data. This split is achieved using the train_test_split function.

```
from sklearn.model_selection import train_test_split

# Define features (X) and target (y)

X = encoded_data.drop('Churn_Yes', axis=1)

y = encoded_data['Churn_Yes']

# Split the dataset into training and testing sets (80% train, 20% test)

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Display the shapes of the training and testing sets

X_train.shape, X_test.shape, y_train.shape, y_test.shape
```