Analysis of Training and Test Data

**1. Training Data Overview**

* The training dataset consists of **76,518 rows** and **38 columns**.
* The columns include various features such as:

- **Demographics**: Marital status, Nationality, Age at enrollment.

- **Academic Details**: Previous qualification and grade, Admission grade, Curricular unit performance (1st and 2nd semester).

- **External Factors**: Unemployment rate, Inflation rate, GDP.

- **Target Variable**: ‘Target’, which includes three classes: ‘Graduate’, ‘Dropout’, and ‘Enrolled’.

* All columns are either ‘int64’ or ‘float64’ types, indicating no categorical or string columns.

**2. Target Variable Distribution**

* The ‘Target’ column is **imbalanced**, with the following distribution:

- ‘Graduate’: **36,282 instances (47.4%)**.

- ‘Dropout’: **25,296 instances (33%)**.

- ‘Enrolled’: **14,940 instances (19.6%)**.

**Observation**:

The imbalance in the `Target` variable suggests that modeling may require techniques to handle this imbalance, such as using balanced class weights or resampling methods.

**3. Missing Values in Training Data**

* The training dataset contains **no missing values** across all 38 columns. This ensures that the dataset is ready for preprocessing and modeling without requiring imputation.

**4. Test Data Overview**

* The test dataset consists of **51,012 rows** and **37 columns**, similar in structure to the training data but without the ‘Target’ column.
* Key features include:

- Academic and demographic data (e.g., ‘Previous qualification’, ‘Admission grade’).

- Performance metrics for curricular units in the 1st and 2nd semesters.

- External factors (‘Unemployment rate’, ‘Inflation rate’, ‘GDP’).

**5. Missing Values in Test Data**

* Similar to the training data, the test dataset contains **no missing values** in any column, ensuring consistency and compatibility with the training dataset.

**6. Insights from the Data**

* **No Missing Data**: Both the training and test datasets are complete, with no null values in any columns.
* **Target Imbalance**: The `Target` variable in the training dataset is imbalanced, which could influence model performance if not addressed.
* **Consistency**: The test dataset mirrors the structure and formatting of the training dataset (excluding the `Target` column), ensuring compatibility for predictive modeling.
* **Rich Features**: The data includes a mix of academic, demographic, and external socioeconomic factors, providing a comprehensive basis for building predictive models.

This analysis confirms that the dataset is clean, consistent, and ready for further preprocessing and model development.