Data Preparation Documentation

1. Raw Data Description

The raw dataset used in this project is **Dataset_ATS_v2.csv**, stored inside the `Data_Preparation/Raw_Data` folder.

This dataset contains customer-level information from a telecom company. The main objective of using this data is to understand customer behaviour and prepare clean, structured data for churn analysis and clustering.

Each row represents one customer, and each column represents a customer attribute or service detail.

---

2. Raw Data Variables Description

Variable Name	Description
gender	Gender of the customer (Male or Female)
SeniorCitizen	Indicates whether the customer is a senior citizen (1 = Yes, 0 = No)
Dependents	Whether the customer has dependents (Yes or No)
tenure	Number of months the customer has stayed with the company
PhoneService	Whether the customer has phone service (Yes or No)
MultipleLines	Whether the customer has multiple phone lines (Yes or No)
InternetService	Type of internet service used by the customer (DSL or Fiber optic)
Contract	Type of customer contract (Month-to-month, One year, Two year)
MonthlyCharges	Monthly charges billed to the customer
Churn	Whether the customer has left the company (Yes or No)

---

3. Prepared Data Variables Description

After data preparation, several changes were applied to the original variables to make the data suitable for machine learning and clustering.

Variable Name	Description (After Preparation)
gender	Converted from Male/Female to numeric values using label encoding

Variable Name	Description (After Preparation)
SeniorCitizen	Already numeric (0 or 1), no change required
Dependents	Converted from Yes/No to numeric values
tenure	Scaled using StandardScaler to normalise customer duration
PhoneService	Converted from Yes/No to numeric values
MultipleLines	Converted from Yes/No to numeric values
InternetService	Converted from service type to numeric values
Contract	Converted from contract type to numeric values
MonthlyCharges	Scaled using StandardScaler to balance feature impact
Churn	Converted from Yes/No to numeric target variable

---

4. Explanation of Train and Test Data

To evaluate the model correctly, the dataset was divided into training and testing data.

- **X (Features):** All independent variables used to learn patterns (customer details and services)
- **y (Target):** The dependent variable (Churn)

Split explanation:

- **X_train:** 80% of feature data used to train the model
- **X_test:** 20% of feature data used to test model performance
- **y_train:** Churn values corresponding to X_train
- **y_test:** Churn values corresponding to X_test

This separation ensures the model is tested on unseen data, which reflects real-world performance.

---

5. Data Preparation Scripts Overview

Data preparation was carried out using structured Python scripts to ensure clarity, reusability, and teamwork.

Script 01: Data Loading and Validation

File: `01_data_loading_and_validation.py`

Purpose of this script:

- Load the raw dataset from the Raw_Data folder
- Validate dataset structure and quality

Key checks performed:

- Dataset shape (rows and columns)
- Column names verification
- Data types of each column
- Missing values check
- Unique values for each column

Outcome:

- Dataset contains 7043 rows and 10 columns
- No missing values found
- Data types and categories are valid for further processing

Script 02: Encoding and Scaling

File: `02_encoding_and_scaling.py`

Purpose of this script:

- Convert categorical variables into numeric format using label encoding
- Apply feature scaling to numerical variables
- Prepare clean input data for modeling

Categorical Encoding Details

The following categorical variables were converted into numeric values using **Label Encoding**. The mappings below were generated from the dataset:

- **gender**: Female = 0, Male = 1
- **Dependents**: No = 0, Yes = 1
- **PhoneService**: No = 0, Yes = 1
- **MultipleLines**: No = 0, Yes = 1
- **InternetService**: DSL = 0, Fiber optic = 1
- **Contract**: Month-to-month = 0, One year = 1, Two year = 2
- **Churn**: No = 0, Yes = 1

This conversion allows machine learning algorithms to process categorical information correctly.

Numerical Scaling

- **tenure** and **MonthlyCharges** were scaled using **StandardScaler**

- This transforms values to have mean = 0 and standard deviation = 1
- Scaling ensures fair contribution of features during model training

Train-Test Split

- Training data: 80% (5634 records)
- Testing data: 20% (1409 records)
- Stratified split based on the **Churn** variable to preserve class balance

---

Script 03: Save Processed Data

File: `03_save_processed_data.py`

Purpose of this script:

- Save cleaned and processed datasets for reuse

Outputs saved in `Data_Preparation/processed_data` folder:

- X_train.csv
- X_test.csv
- y_train.csv
- y_test.csv

This ensures consistent data usage across notebooks and scripts.

---

6. Final Prepared Data Summary

After data preparation:

- All categorical variables are converted to numeric format
- Numerical variables are standardized
- Data is split into training and testing sets
- Processed datasets are saved and version-controlled in GitHub

This prepared data is now ready to be used for:

- Clustering analysis
- Churn prediction modeling
- Further exploratory analysis

---

5. Notes for Team Members

- Do not modify raw data files

- Always use processed data for modeling
- Follow script execution order for consistency
- Update documentation if any new preprocessing step is added