Training and Testing Sets - Size & Composition Documentation

1. Overview

As part of Stage 2 – Data Preparation, the original preprocessed customer churn dataset was split into **training** and **testing** datasets. This split allows the predictive model to be trained on one subset of data and evaluated on unseen data to ensure generalisability and avoid overfitting.

2. Dataset Source

- **Original Dataset:** Dataset_ATS_v2.csv
- **Purpose:** Customer churn prediction for a telecommunications company
- **Target Variable:** Churn (Binary: 0 = No churn, 1 = Churn)

All missing values were handled, categorical variables were encoded, and numerical features were scaled prior to splitting.

3. Train–Test Split Strategy

- **Split Ratio:**

  Training Set: 80% **Testing Set:** 20%

- **Method:** Random stratified split
- **Reason:** Ensures the churn vs non-churn distribution remains consistent across both datasets

This approach supports reliable model training and fair performance evaluation.

4. Training Dataset Details

- **File Name:** training_set.csv
- **Purpose:** Used to train the predictive (ANN) model
- **Approximate Size:**

  ~80% of total records

- **Composition Includes:**

  - Customer demographic attributes
  - Service-related features (e.g. contract type, tenure, charges)
  - Encoded categorical variables
  - Scaled numerical features
  - Target variable (Churn)

The training dataset contains sufficient data diversity to allow the model to learn complex patterns associated with customer churn.

5. Testing Dataset Details

- **File Name:** testing_set.csv
- **Purpose:** Used exclusively for evaluating model performance
- **Approximate Size:**

  ~20% of total records

- **Composition Includes:**

  - Same feature structure as the training dataset
  - Identical preprocessing and scaling
  - Target variable retained for evaluation metrics

The testing dataset was not exposed to the model during training and represents unseen customer data.

6. Feature Consistency

Both datasets:

- Contain identical columns
- Use the same encoding and scaling logic
- Maintain consistent data types and formats

This ensures compatibility with machine learning models and accurate performance comparison.

7. Files Included in Submission

Data_Preparation/
– training_set.csv
– testing_set.csv
– Training_Testing_Documentation.pdf

8. Conclusion

The prepared training and testing datasets meet all **Stage 2 Data Preparation deliverable requirements**. They are clean, well-structured, scaled, and suitable for predictive modelling and validation tasks in later stages of the project.