**Train-Test Split Diagram**

Original Dataset (X, y)

+----+----+----+----+----+----+----+----+----+----+

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |

+----+----+----+----+----+----+----+----+----+----+

↓ Split (test\_size=0.2, 20% test, 80% train)

Training Set (80%)

+----+----+----+----+----+----+----+----+

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |

+----+----+----+----+----+----+----+----+

X\_train → Features for training

y\_train → Corresponding target values

Testing Set (20%)

+----+----+

| 9 | 10 |

+----+----+

X\_test → Features for testing

y\_test → Corresponding target values

**Explanation**

1. **Training Set**
   * Used to **teach the model** patterns in the data.
   * Usually **larger portion** of the dataset (e.g., 70%-80%).
2. **Testing Set**
   * Used to **evaluate the model** on unseen data.
   * Helps measure **real-world performance**.
3. **Random Split**
   * Data is usually shuffled before splitting so that the **train and test sets are representative**.
   * random\_state=42 ensures the split is **reproducible** every time.

**Quick Tip**

* **Too small test size** → model evaluation may not be accurate.
* **Too large test size** → model may not learn enough from training.
* **Common split** → 80% training, 20% testing (or 70%-30%).