Data Splitting Strategies

🎲 Random Split

Simplest approach for large, i.i.d. datasets

✓ Best for: Large, independent datasets

⚖️ Stratified Split

Maintains class distribution in each subset

✓ Best for: Imbalanced classification

📅 Time-based Split

Essential for temporal or sequential data

✓ Best for: Time series, forecasting

👥 Group-based Split

Prevents data leakage from related samples

✓ Best for: Patient/user-level data

⚠ Dataset Size Consideration

Larger datasets allow smaller validation/test percentages

✓ Key Principle

All splits must be representative of population