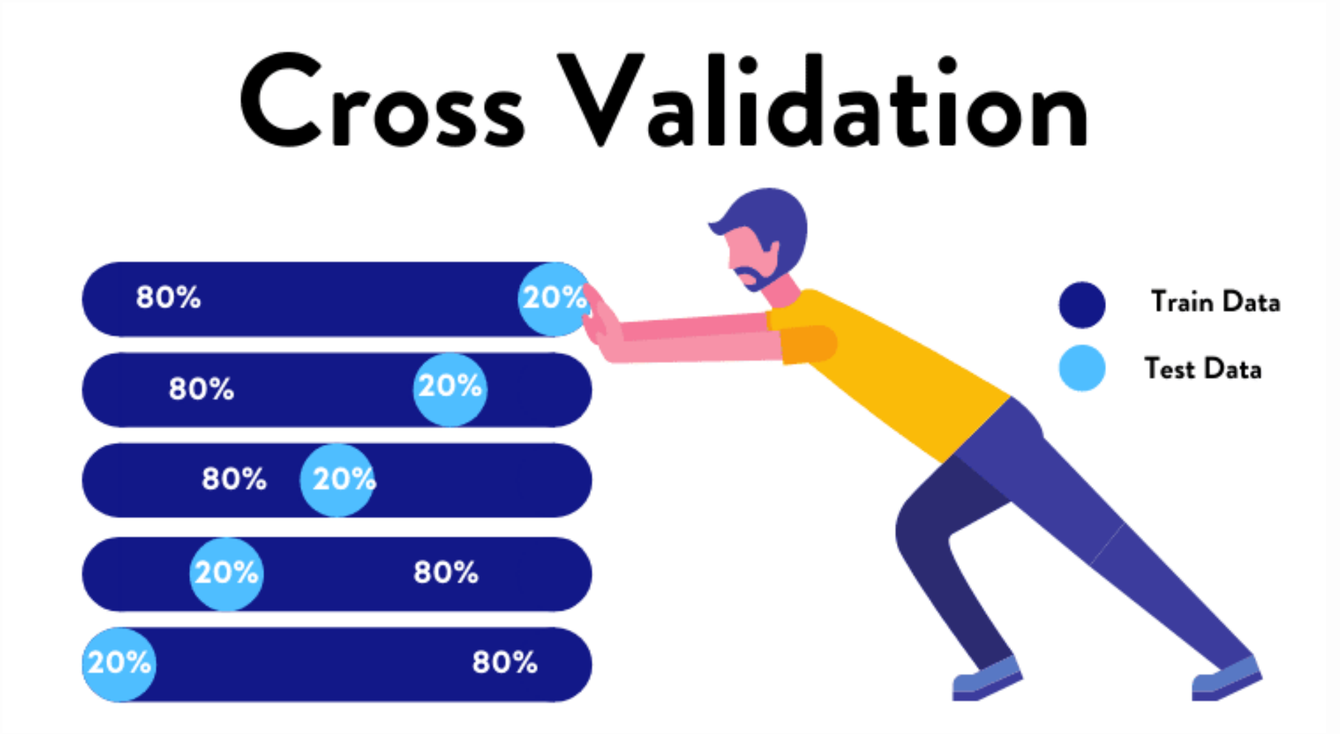


CROSS-VALIDATION

GROUP 3

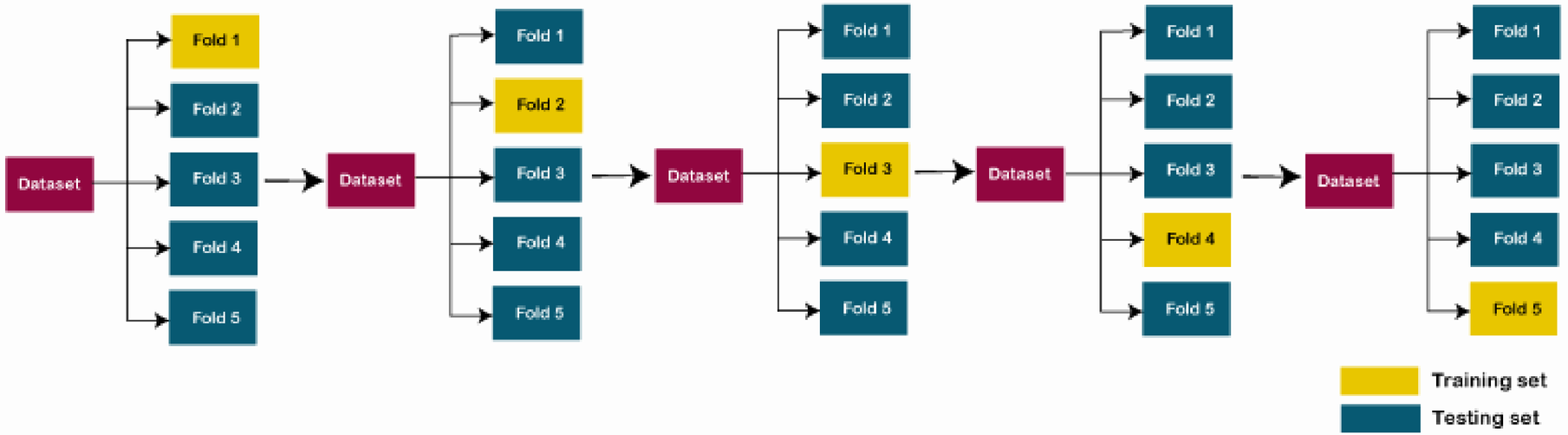
TRAIN TEST SPLIT

Train-test split is a technique used in machine learning and statistical modeling to evaluate the performance of a predictive model. It involves splitting the available dataset into two disjoint subsets: the training set and the testing set.

Limitation: It's highly dependent on a single split which can lead to errors in the evaluation of model performance (overly optimistic or overly pessimistic)

CROSS VALIDATION

Cross-validation is a technique that allows us to get a more reliable estimate of our model's performance by using multiple test sets. The idea is to divide our data into several subsets, or "folds" and iteratively train and evaluate the model on different combinations of these folds.



METHODS OF CROSS VALIDATION

K-Fold Cross-Validation

- K-fold cross-validation is a technique used to evaluate the performance of a machine learning model by partitioning the dataset into K equally sized subsets, or folds

Stratified K-Fold Cross-Validation

- a variation of K-fold cross-validation that ensures each fold preserves the class distribution of the target variable.

Leave-One-Out Cross-Validation (LOOCV)

- is a special case of K-fold cross-validation where K is equal to the number of instances in the dataset

Leave-P-Out Cross-Validation (LPOCV)

- LPOCV is a generalized version of LOOCV. Instead of leaving out only one sample as in LOOCV, LPOCV involves leaving out P samples as the test set and using the remaining samples for training.

Time Series Cross-Validation

- takes into account the temporal order of the data. It involves splitting the data into training and test sets based on time, ensuring that the test set comes after the training set in terms of time.

Nested Cross-Validation

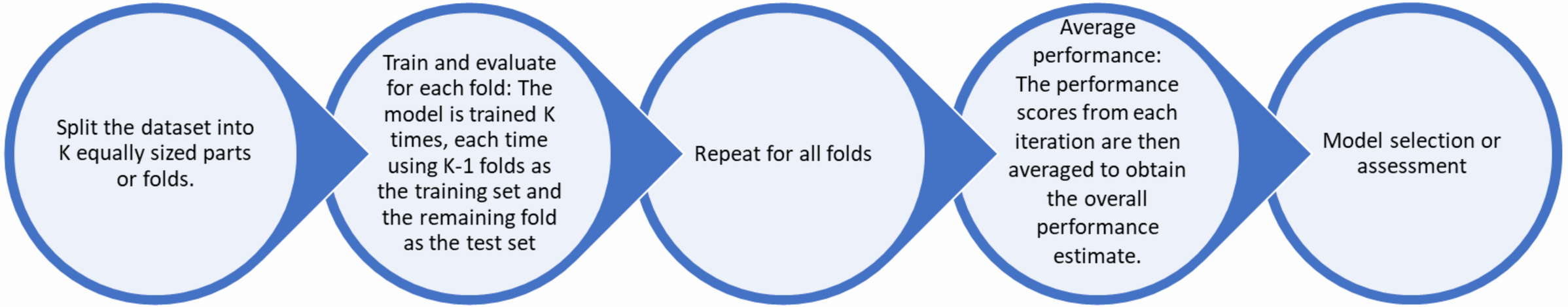
- Nested cross-validation is used when there is a need to tune hyperparameters for the model

METHODS OF CROSS VALIDATION

K-FOLD CROSS-VALIDATION

K-Fold cross-validation is the most common and basic type of cross-validation.

K-FOLD CROSS-VALIDATION STEP BY STEP



LIMITATIONS OF CROSS-VALIDATION

Data Leakage: Data leakage occurs when information from the test set leaks into the training process, leading to overly optimistic performance estimates.

Cross-validation **does not shuffle** the data before splitting it into folds

Computationally Expensive: Cross-validation can be computationally expensive, especially for large datasets or complex models

Sample Size Dependency: The effectiveness of cross-validation depends on having a sufficient amount of data. If the dataset is small, cross-validation may not provide reliable performance estimates

THE END!

VIVIAN
WILFRED
ANTONY
SHEILA
EUNITA
BRYAN
CYNTHIA

ANY QUESTIONS