ICT Academy of Kerala

Building the Nation's Future

Feature Engineering and Tuning

A **GOVT. OF INDIA** SUPPORTED, **GOVT. OF KERALA** PARTNERED SOCIAL ENTERPRISE.

Model Selection

Underfitting and Overfitting

Underfitting vs Overfitting

High Bias, Low Variance	Low Bias, High Variance
Performs poorly on training data, also on unseen data	Performs well on training data, poorly on unseen data
Training Accuracy and Validation accuracy are poor	Training Accuracy is very good but Validation Accuracy is poor
Happens when we have very less amount of data	Happens when we train our model a lot over noisy datasets

Bias:

Algorithms tendency to consistently learn the wrong thing by not taking into account of all the information in the data

High Bias:

It is a result of the algorithm missing the relevant relationship between features and target outputs

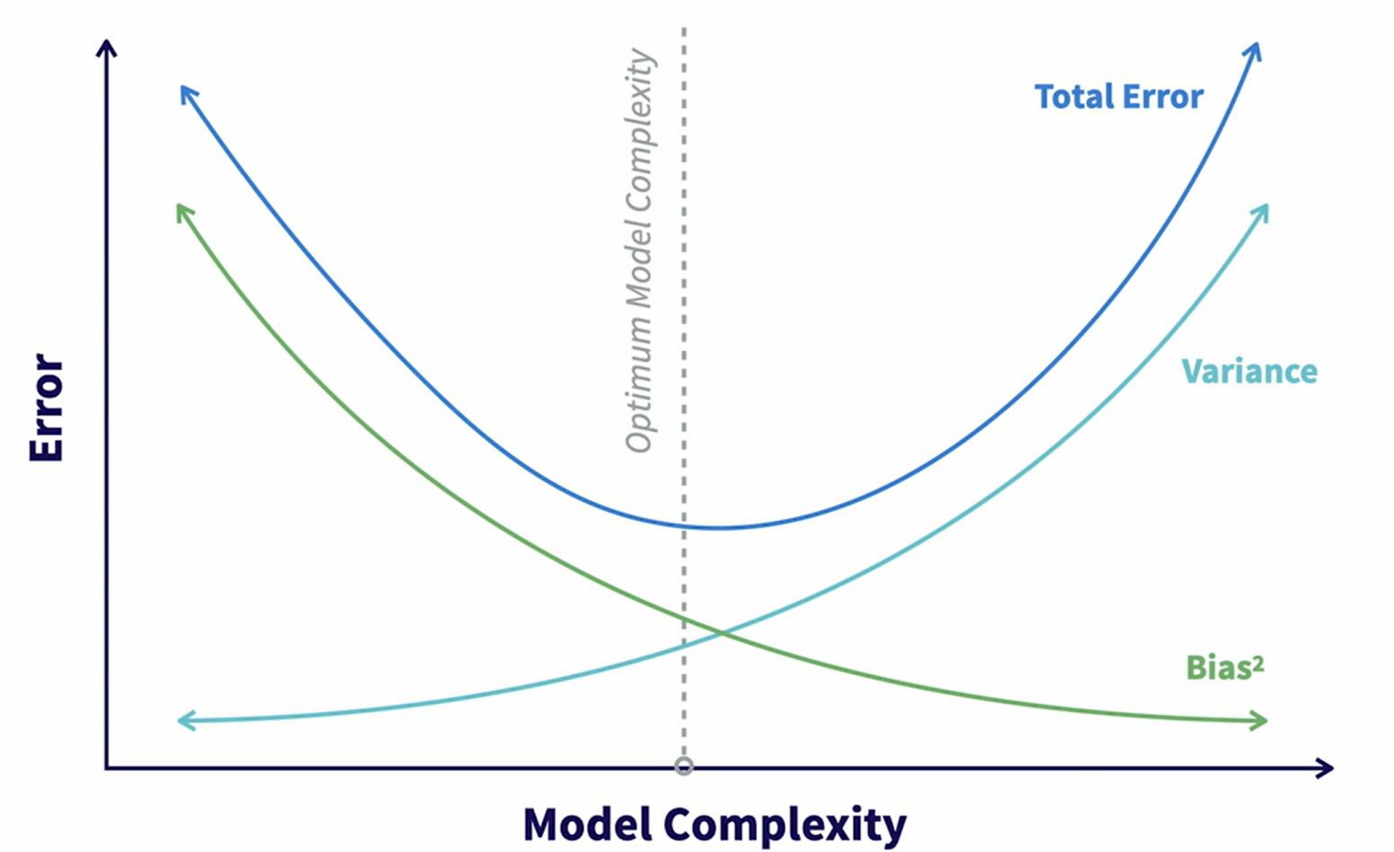
Variance:

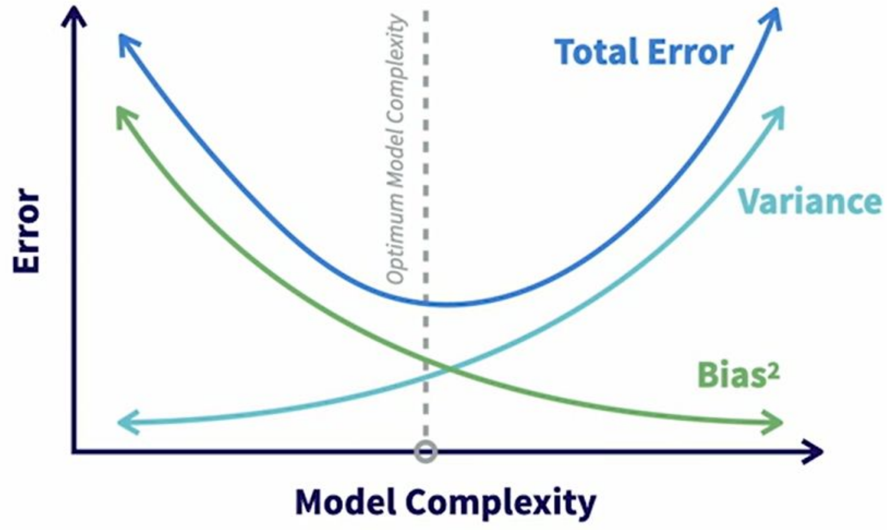
It refers to an algorithm's sensitivity to small fluctuations in the training set

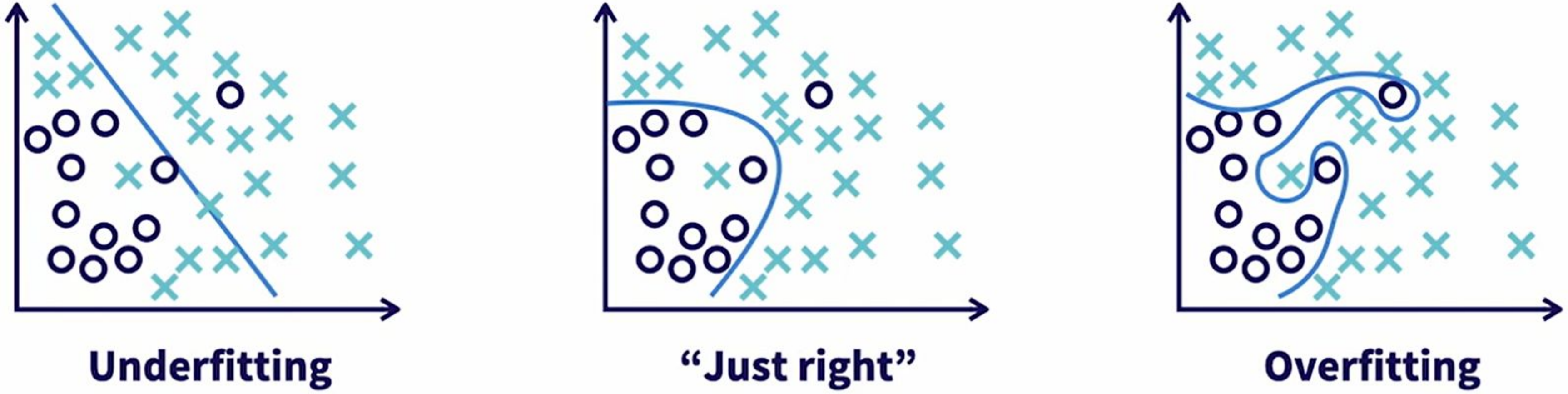
High variance:

It is a result of the algorithm fitting to random noise in the training data

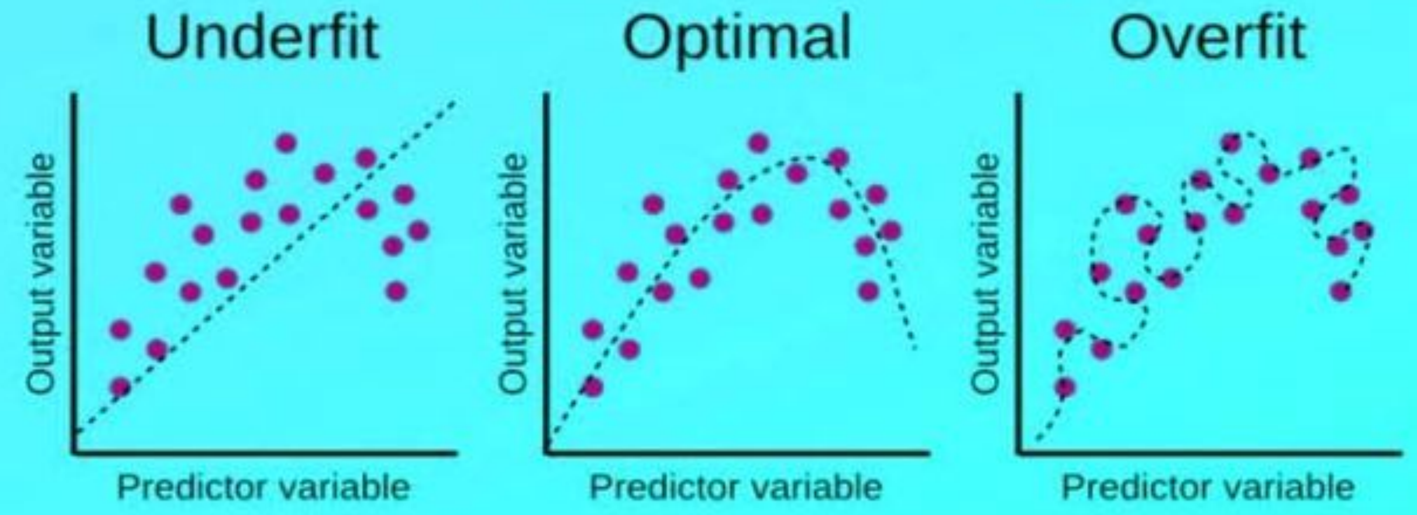
Bias - Variance Tradeoff




Error
Optimum Model Complexity
Total Error
Variance
Bias²
Model Complexity


Underfitting
“Just right”
Overfitting

What is overfitting and underfitting



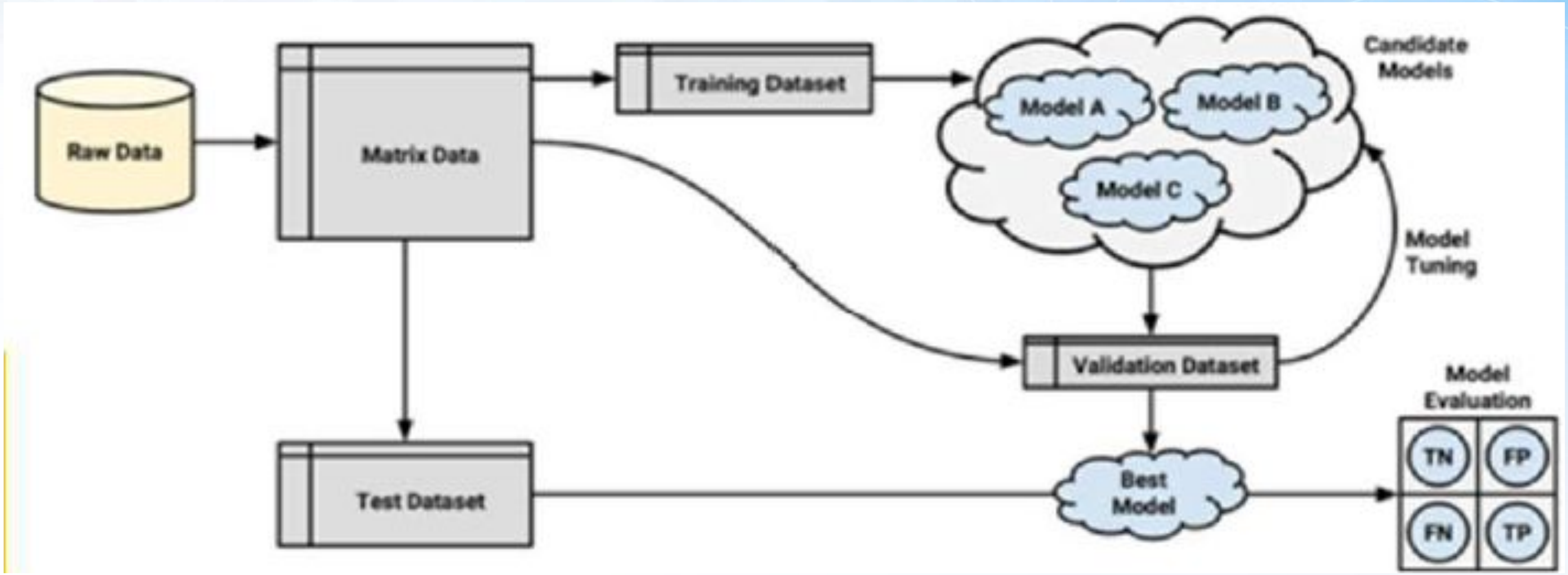
Cross Validation

- Divide training set into 2 parts
  - Training
  - Validation
- Training part is used to find the hypothesis
- Validation set is used to test the generalization ability
- If training and validation sets are large enough, the hypothesis that is the most accurate on the validation set is the best one
- This process is called cross-validation

Cross Validation

- The validation dataset would be used for iterating and refining the model or models chosen, leaving the test dataset to be used only once as a final step to report an estimated error rate for future predictions
- Typical split – Training: validation: test = 50:25:25

Cross Validation



Cross Validation Steps

1. Randomly split training set into several subsets (n) of the same size. Each subset is called a fold.
2. Train models with (n-1) folds and validate with last fold.
3. Repeat n times varying the folds used for training and validating
4. Calculate average of metric to get final value.
5. Evaluate model using test set

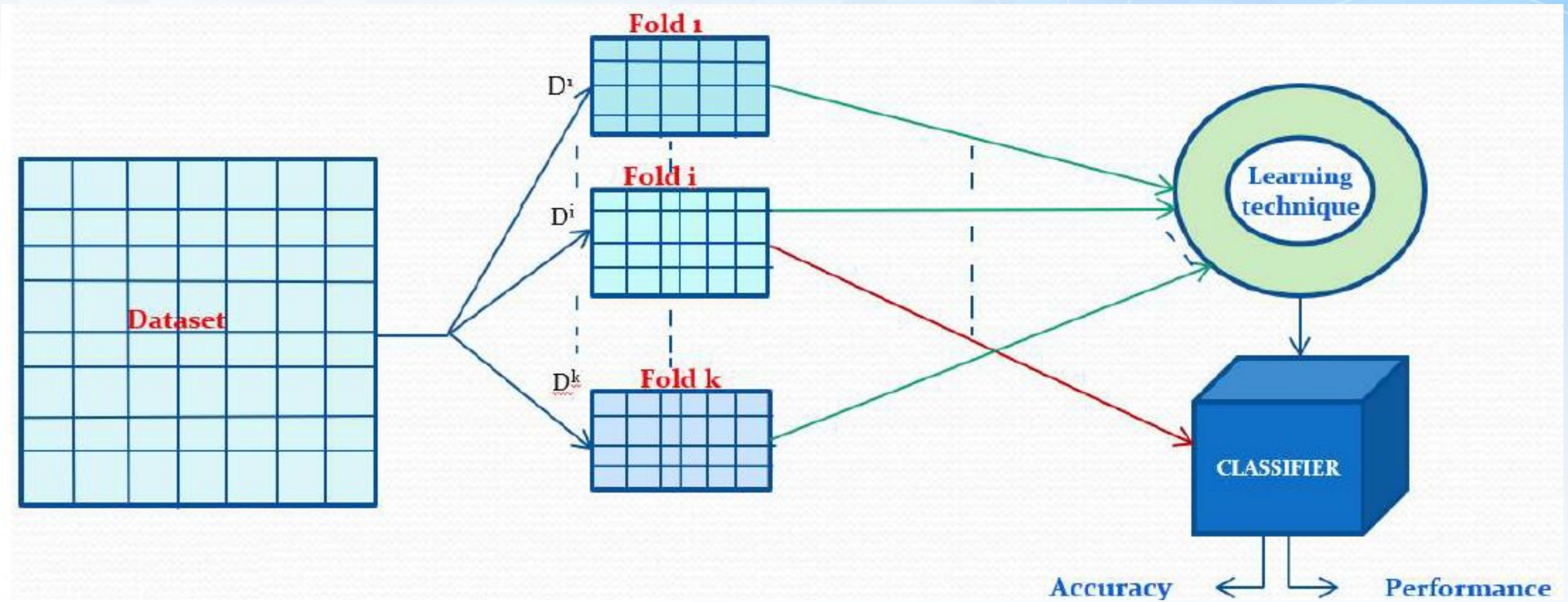
Cross Validation Types

1. K-fold cross validation
2. Leave-One-Out Cross Validation (LOOCV)
3. Stratified Cross Validation

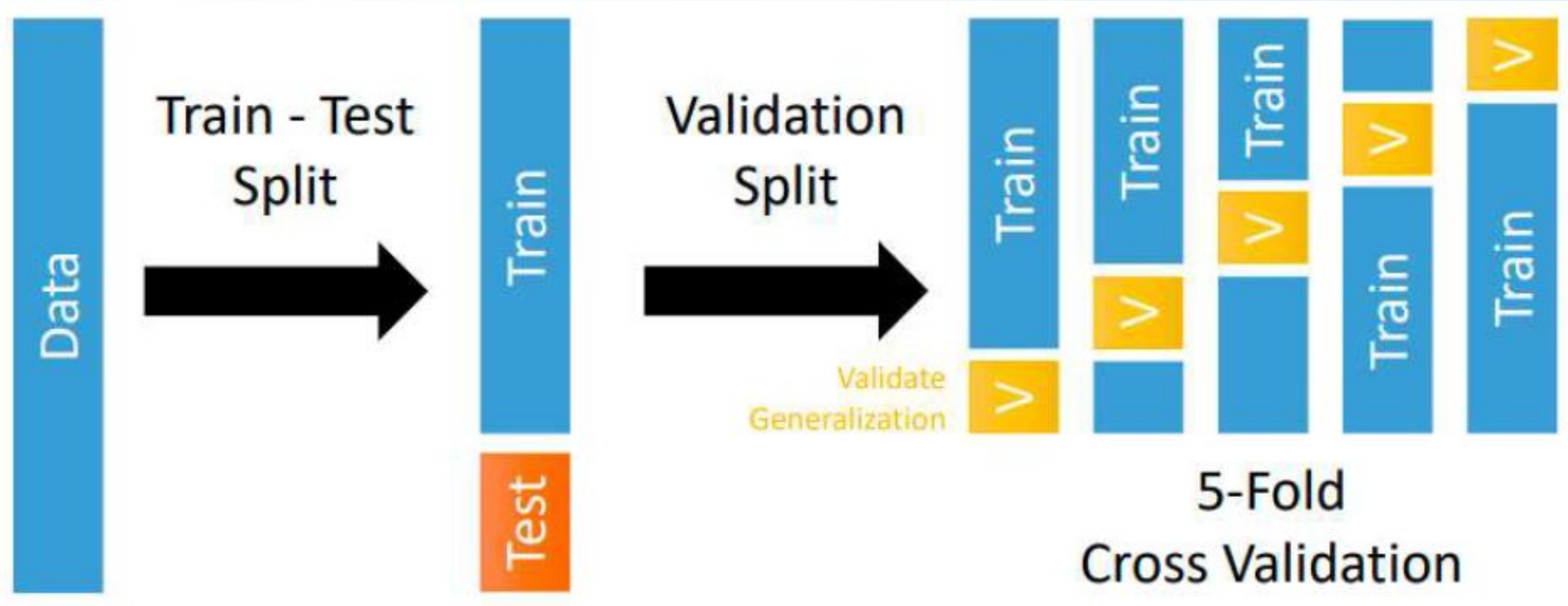
K-fold Cross Validation

- Partition the data into k non-overlapping subsets
- All available data is partitioned into k groups (folds)
- Each group has size (N/k)
- k-1 groups are used to train and validated on remaining group
- Repeat for all k choices of held out group
- Performance scores from k runs are averaged
- If k = N, this is the leave-one-out method

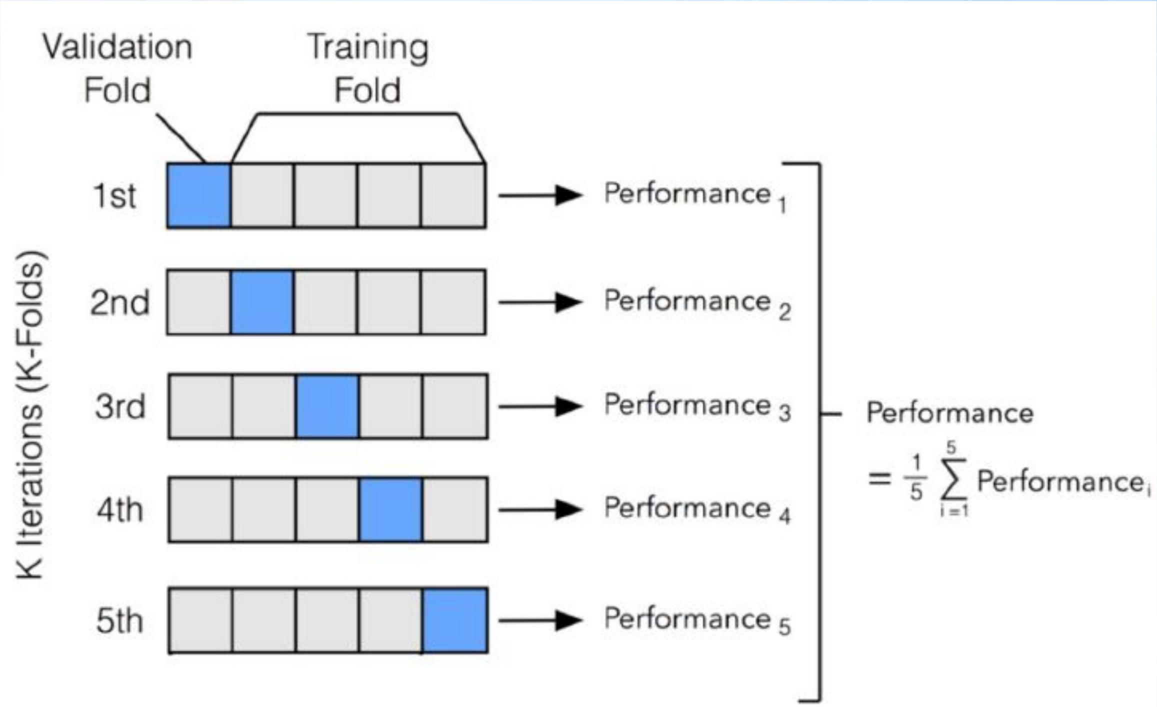
K-fold Cross Validation



K-fold Cross Validation



K-fold Cross Validation



Grid Search Cross-Validation (CV)

- Grid Search CV is a technique to tune hyperparameters by exhaustively searching through a specified subset of hyperparameter combinations
- **Working:**
  - Define a grid of hyperparameter values to explore.
  - Perform cross-validation for each combination.
  - Select the combination that yields the best cross-validation performance
- **Advantages:**
  - Exhaustive search ensures thorough exploration of hyperparameter space.
  - Guarantees finding the optimal combination within the specified grid.
- **Limitations:**
  - Computationally expensive, especially with large hyperparameter grids.
  - Prone to overfitting if the grid is not well-defined

Randomized Search Cross-Validation (CV)

Randomized Search CV is a technique to tune hyperparameters by sampling a specified number of combinations randomly from the hyperparameter space

Workflow:

- Define a probability distribution for each hyperparameter.
- Randomly sample combinations from these distributions.
- Perform cross-validation for each sampled combination.
- Select the combination that yields the best cross-validation performance.

Advantages:

- More efficient than Grid Search CV, especially with large hyperparameter spaces.
- Allows exploration of a broader range of hyperparameter values.
- Less susceptible to overfitting compared to Grid Search CV.

Limitations:

- May miss optimal combinations present in the unexplored regions of the hyperparameter space.
- Less deterministic compared to Grid Search CV.

How to select?

Grid Search CV exhaustively searches through a predefined grid of hyperparameters, while Randomized Search CV samples combinations randomly from the hyperparameter space.

Recommendation:

- Use Grid Search CV for smaller hyperparameter spaces or when computational resources allow.
- Use Randomized Search CV for larger hyperparameter spaces or when computational resources are limited.