UEM
श्रद्धावान लभते ज्ञानम्
UNIVERSITY OF ENGINEERING & MANAGEMENT
Good Education, Good Jobs

# UNIVERSITY OF ENGINEERING & MANAGEMENT, KOLKATA

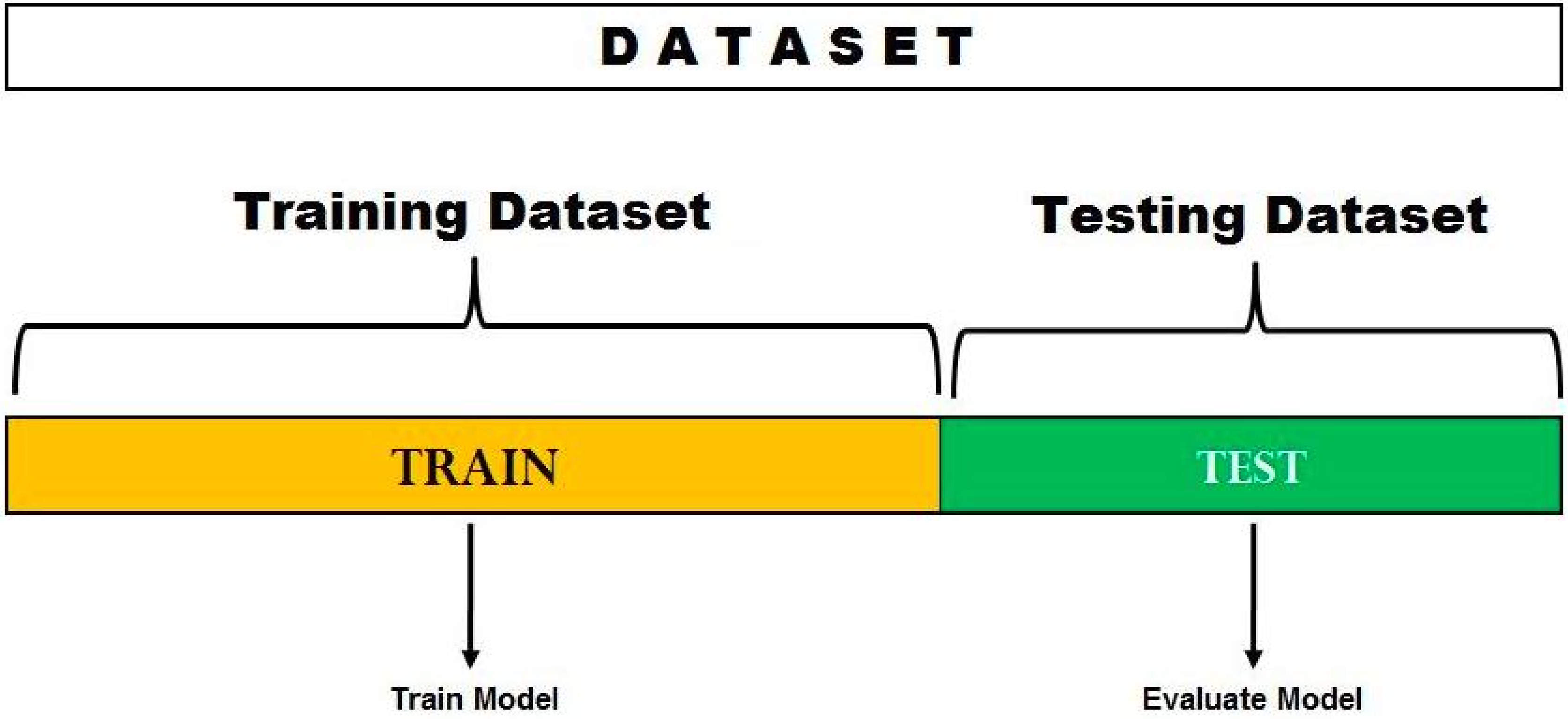
## Course Name : AI & ML

## Training a Model (For Supervised Learning)

- Holdout Method
- The hold-out method for training a machine learning model is the process of splitting the data in different splits and using one split for training the model and other splits for validating and testing the models. The hold-out method is used for both **model evaluation** and **model selection.**

Process of using the hold-out method for model evaluation:

- Split the dataset into two parts (preferably based on 70-30% split; However, the percentage split will vary)
- Train the model on the training dataset; While training the model, some fixed set of hyper parameters is selected.
- Test or evaluate the model on the held-out test dataset
- Train the final model on the entire dataset to get a model which can generalize better on the unseen or future dataset.



DATASET

Training Dataset

Testing Dataset

TRAIN

TEST

Train Model

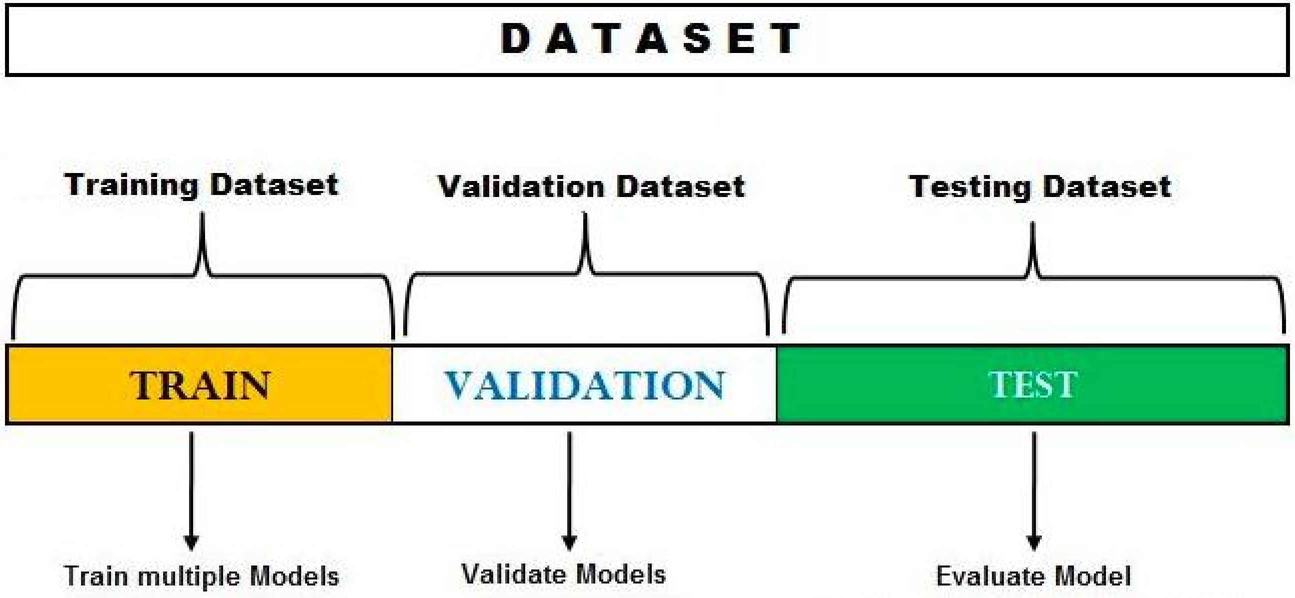
Evaluate Model

# Hold-out method for Model Selection

- The hold-out method can also be used for model selection or hyper-parameters tuning. As a matter of fact, at times, the model selection process is referred to as hyper-parameters tuning. In the hold-out method for model selection, the dataset is split into three different sets – training, validation, and test dataset.

Process represents the hold-out method for model selection:

- Split the dataset in three parts – Training dataset, validation dataset and test dataset.
- Train different models using different machine learning algorithms. For example, train the classification model using logistic regression, random forest, XGBoost.
- For the models trained with different algorithms, tune the hyper-parameters and come up with different models. For each of the algorithms mentioned in step 2, change hyper parameters settings and come with multiple models.
- Test the performance of each of these models (belonging to each of the algorithms) on the validation dataset.
- Select the most optimal model out of models tested on the validation dataset. The most optimal model will have the most optimal hyper parameters settings for specific algorithm. Going by the above example, lets say the model trained with XGBoost with most optimal hyper parameters gets selected.
- Test the performance of the most optimal model on the test dataset.



DATASET

Training Dataset

Validation Dataset

Testing Dataset

TRAIN

VALIDATION

TEST

Train multiple Models

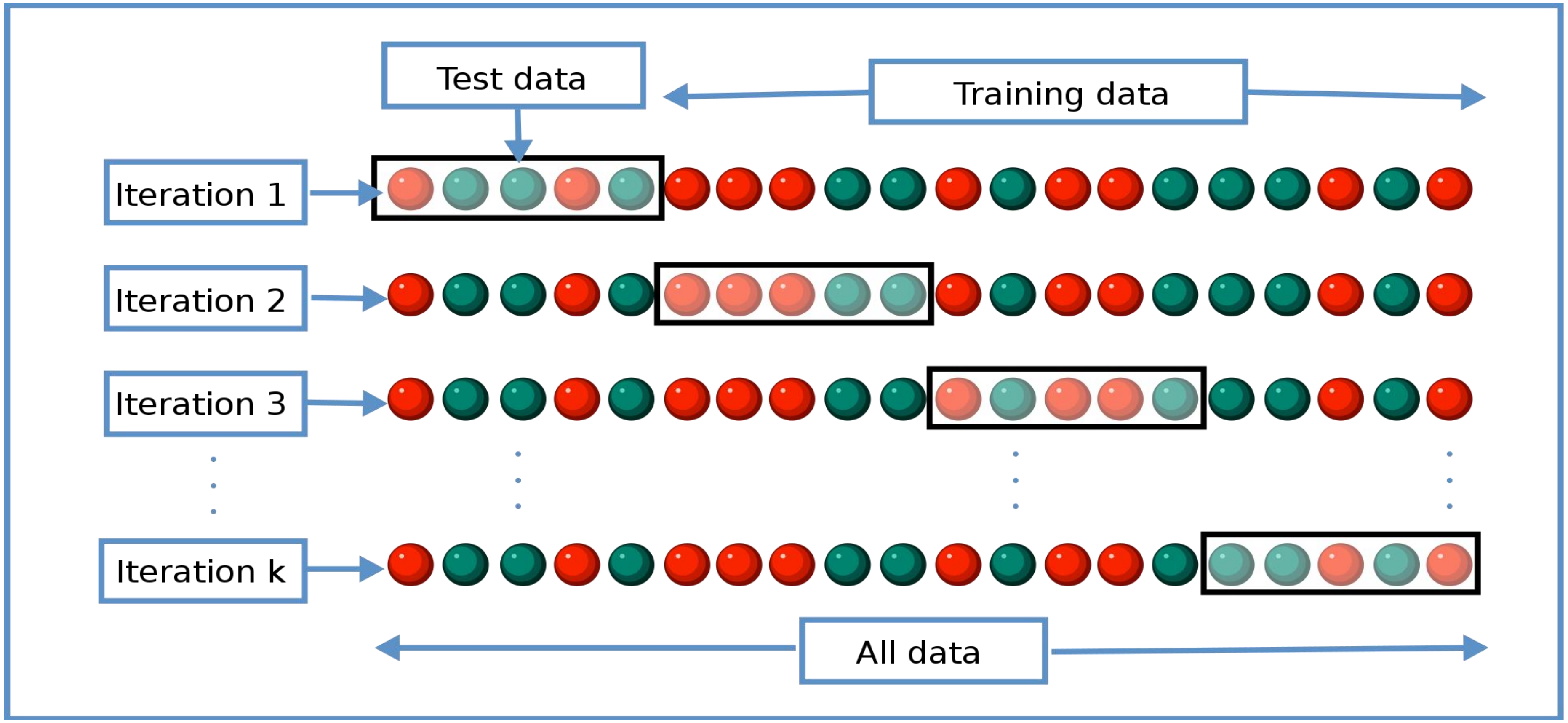
Validate Models

Evaluate Model

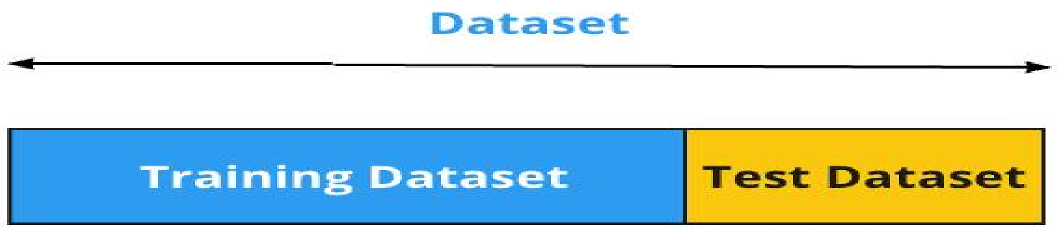
# k-Fold Cross-Validation

- Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample.
- The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. As such, the procedure is often called k-fold cross-validation. When a specific value for k is chosen, it may be used in place of k in the reference to the model, such as k=10 becoming 10-fold cross-validation.

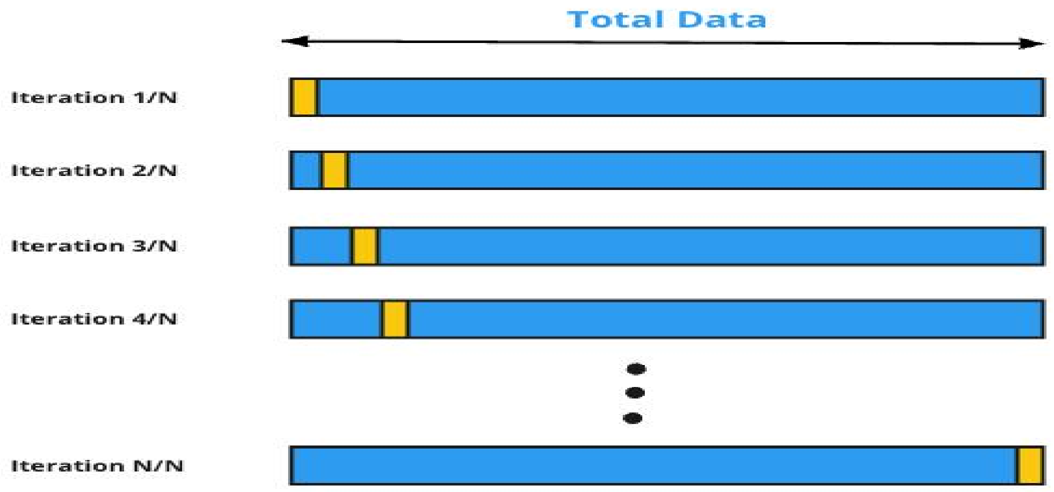
- The general procedure is as follows:
- Shuffle the dataset randomly.
- Split the dataset into k groups
- For each unique group:
  - Take the group as a hold out or test data set
  - Take the remaining groups as a training data set
  - Fit a model on the training set and evaluate it on the test set
  - Retain the evaluation score and discard the model
- Summarize the skill of the model using the sample of model evaluation scores
- Importantly, each observation in the data sample is assigned to an individual group and stays in that group for the duration of the procedure. This means that each sample is given the opportunity to be used in the hold out set 1 time and used to train the model k-1 times.


Test data
Training data
Iteration 1
Iteration 2
Iteration 3
Iteration k
All data

- LOOCV : The **L**eave-**o**ne-**o**ut **C**ross **V**alidation or LOOCV is a type of **cross-validation method** that involves leaving out one sample from the training set and using the remaining samples to train the model.
- Leave-one-out cross validation (LOOCV) is a type of cross-validation method in which a single data point is removed from the dataset, and the model is trained on the remaining data points.
- The removed data point is then used as a test case to evaluate the model's performance. This process is repeated for each data point in the dataset, and the results are averaged to obtain an estimate of the model's performance.
- LOOCV is an effective technique for **model evaluation**, as it uses all available data points for training and testing. It can be especially useful when there is limited labeled data available, as it provides a way to estimate the model's performance without requiring a separate validation set.



## LOOCV: Leave One Out Cross Validation



dataaspirant.com