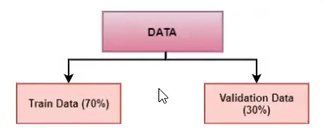
**Types of Cross Validation Final**

**1) Holdout cross-validation:-**

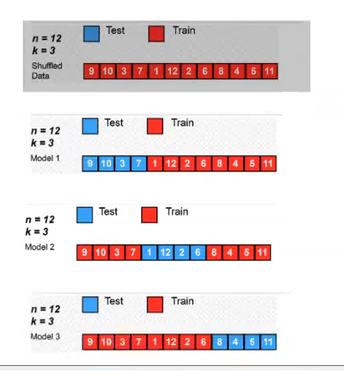
The holdout technique is an exhaustive cross-validation method, that randomly splits the dataset into train and test data depending on data analysis.



In the case of holdout cross-validation, the dataset is randomly split into training and validation data. Generally, the split of training data is more than test data. The training data is used to induce the model and validation data is evaluates the performance of the model.

**2) K fold Cross Validation:-**

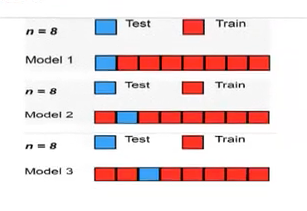
In k-fold cross-validation, the original dataset is equally partitioned into k subparts or folds. Out of the k-folds or groups, for each iteration, one group is selected as validation data, and the remaining (k-1) groups are selected as training data. The process is repeated for k times until each group is treated as validation and remaining as training data.



The final accuracy of the model is computed by taking the mean accuracy of the k-models validation data.

**3) Leave-one-out cross-validation:-**

Leave-one-out cross-validation (LOOCV) is an exhaustive cross-validation technique. For a dataset having n rows, 1st row is selected for validation, and the rest (n-1) rows are used to train the model. For the next iteration, the 2nd row is selected for validation and rest to train the model. Similarly, the process is repeated until n steps or the desired number of operations.



**Balanced and Unbalanced dataset:**

# Balanced dataset:

Suppose you have a dataset where your y column data is like 0,1,1,1,0,0,1,1,0,1,0,0.So here you can see your y column or output column has six 1 and six 0's that means your dataset is balanced.

# Unbalanced dataset:

Suppose you have a dataset where your y column data is like 0,1,1,1,0,1,1,1,1,1,0,0.Here you can see your y column or output column has eight 1's and four 0's.That means here 1's are in majority and 0's are in minority and this is bad for machine learning. This is the example of Unbalanced dataset.

So to balancing the data set we have to use SMOTE technique to balance an unbalance dataset.

SMOTE works like in above example you see there is eight 1's and four 0's in y column so it will increase the 0's data to eight also and its adding newly four rows for newly added four 0's output. So now your dataset have 16 rows where in y column there is eight 1's and eight 0's are present.

In another method we can describe Unbalanced dataset as:-

Imagine, you have two categories in your dataset to predict - Category-A and Category-B. When Category-A is higher than Category-B or vice versa, you have a problem of imbalanced dataset.

**SMOTE (Synthetic Minority Over-sampling Technique):-**

SMOTE is an over-sampling method. It creates synthetic samples of the minority class. Hence making the minority class equal to the majority class. SMOTE does this by selecting similar records and altering that record one column at a time by a random amount within the difference to the neighbouring records.

**Curse of dimensionality reduction:-**

When your model has too much column like 60,70,200 so model did not learn well so in this situation we have to use a technique named PCA()- Principle Component Analysis. It will reduce your dataset column to your desired number by merging highly correlated column.

Eg-

Suppose your dataset has 100 of columns so you can write like :

PCA[n\_components=15]

So your 100 columns reduced to 15.