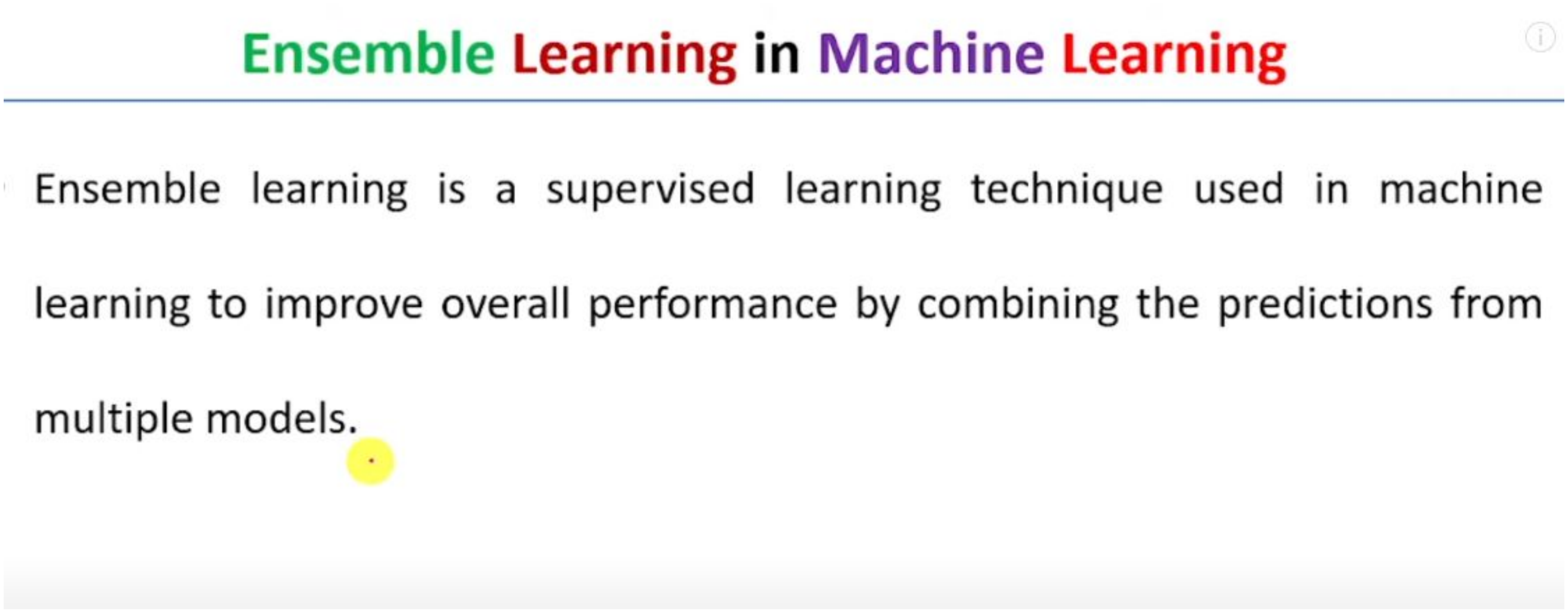


Ensemble Learning in Machine Learning

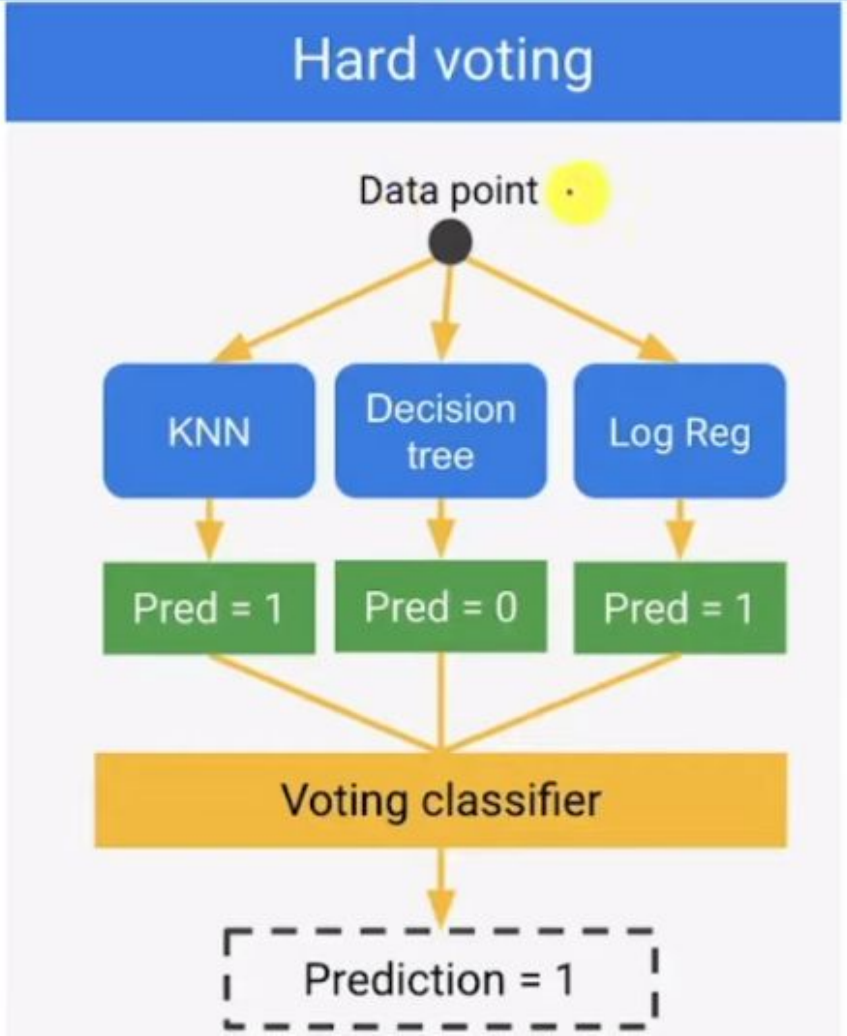
Ensemble learning is a supervised learning technique used in machine learning to improve overall performance by combining the predictions from multiple models.

Ensemble Learning - Types of Ensemble Methods

- Voting (Averaging)
- Bootstrap aggregation (bagging)
- Random Forests
- Boosting
- Stacked Generalization (Blending)

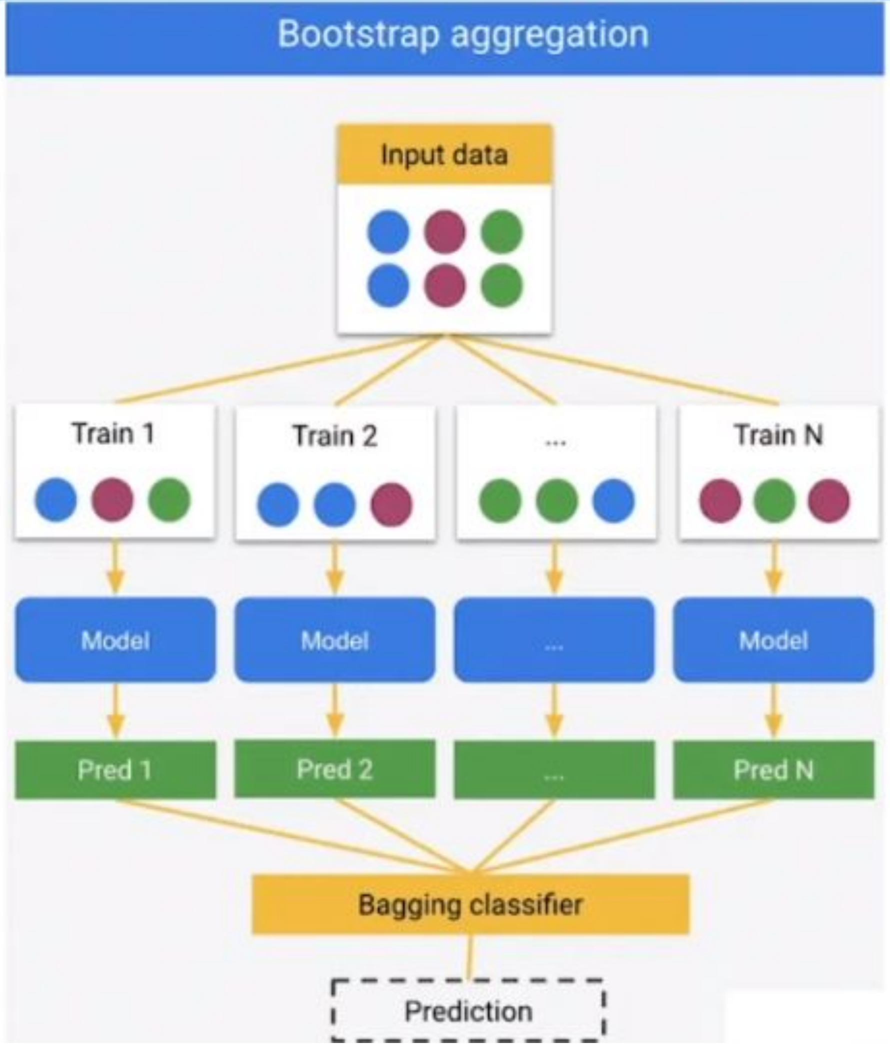
Ensemble Learning – Voting (Averaging)

- Voting is an ensemble machine learning algorithm that involves making a prediction that is the average (regression) or the sum (classification) of multiple machine learning models.



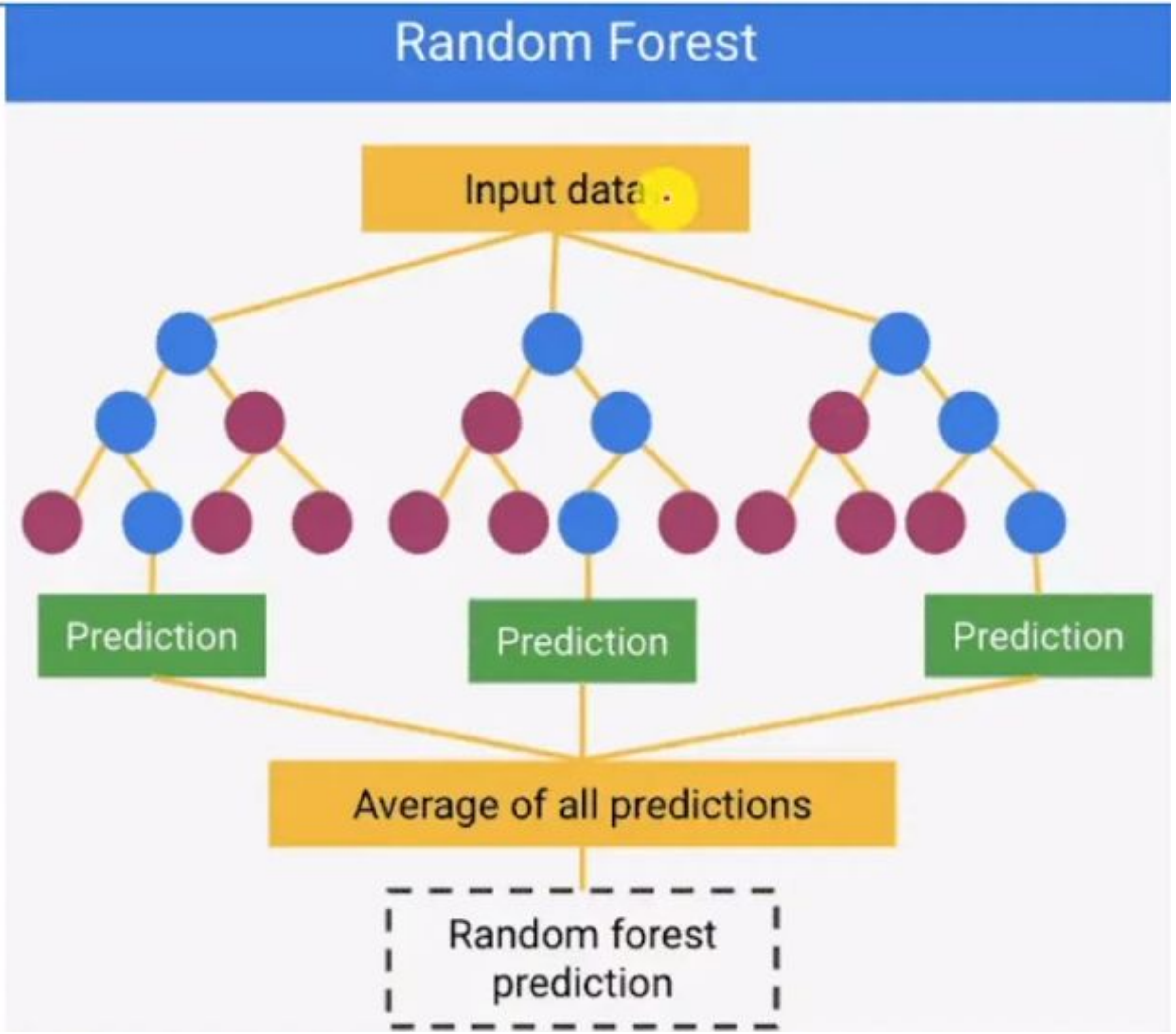
Ensemble Learning – Bootstrap aggregation (bagging)

- Bootstrap Aggregating, also known as bagging, is a machine learning ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms like classification and regression.
- It decreases the variance and helps to avoid overfitting.
- It is usually applied to decision tree methods.
- Bagging is a special case of the model averaging approach.



Ensemble Learning – Random Forest

- Random forest is a commonly-used machine learning algorithm.
- A random forest is an ensemble learning method where multiple decision trees are constructed and then they are merged to get a more accurate prediction.



Ensemble Learning – Boosting

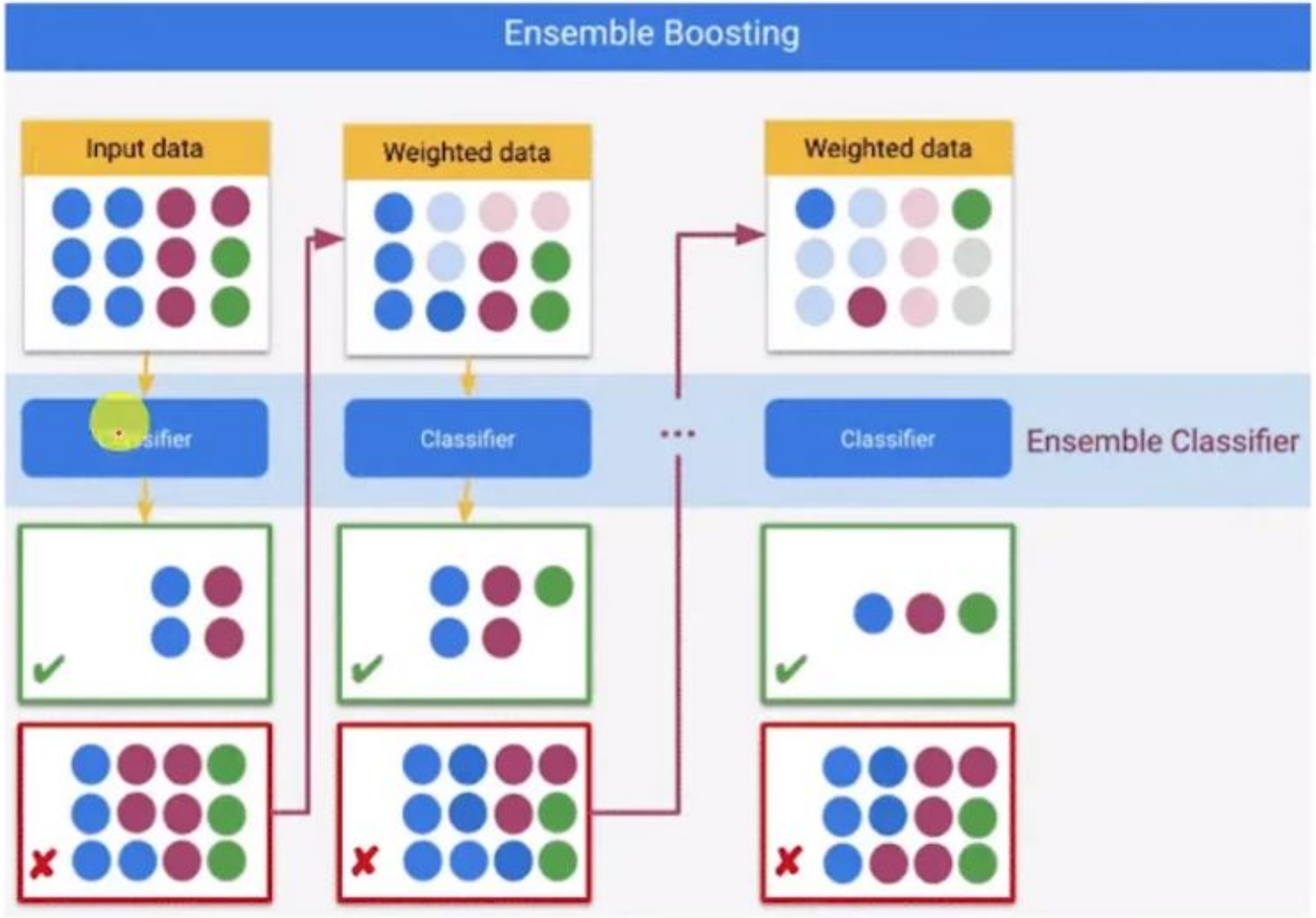
- Boosting is an ensemble modeling technique that attempts to build a strong classifier from the number of weak classifiers.
- It is done by building a model by using weak models in series.
- Firstly, a model is built from the training data.

C

I

- Then the second model is built which tries to correct the errors present in the first model.
- This procedure is continued and models are added until either the complete training data set is predicted correctly or the maximum number of models is added.

Ensemble Learning – Boosting

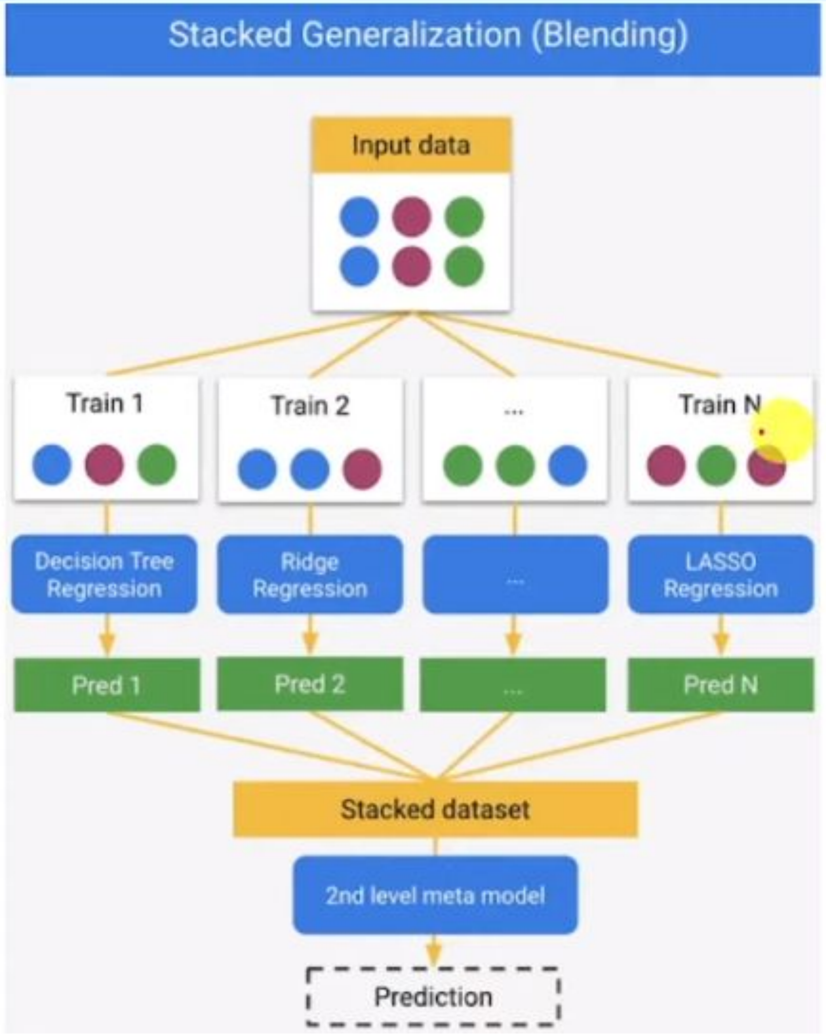


Ensemble Learning – Stacked Generalization (Blending)

- Stacking, Blending and and Stacked Generalization are all the same thing with different names. It is a kind of ensemble learning.
- In traditional ensemble learning, we have multiple classifiers trying to fit to a training set to approximate the target function.
- Since each classifier will have its own output, we will need to find a combining mechanism to combine the results.
- This can be through voting (majority wins), weighted voting (some classifier has more authority than the others), averaging the results, etc.

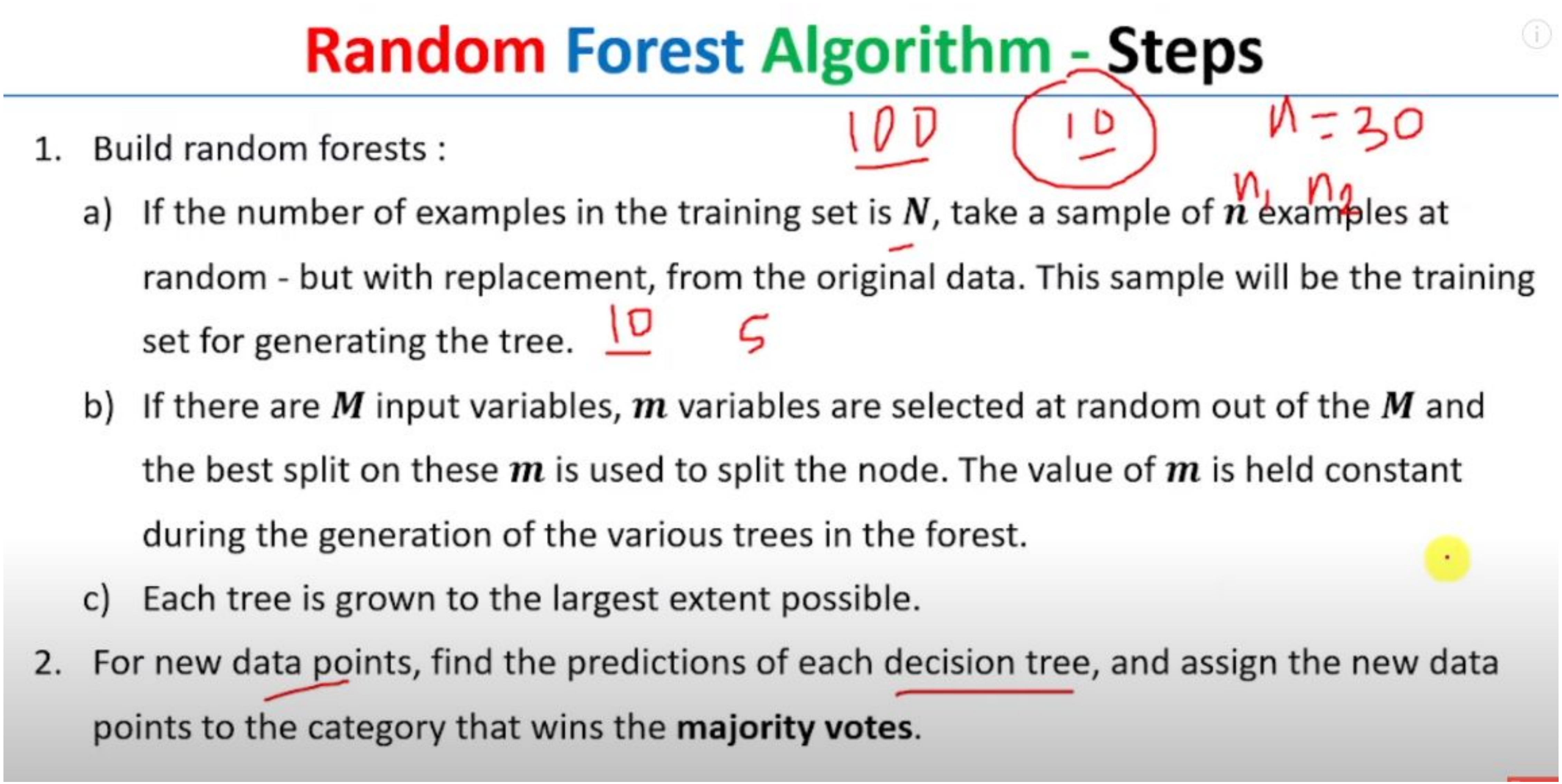
Ensemble Learning – Stacked Generalization (Blending)

- In stacking, the combining mechanism is that the output of the classifiers (Level 0 classifiers) will be used as training data for another classifier (Level 1 classifier) to approximate the same target function.
- Basically, you let the Level 1 classifier to figure out the combining mechanism.



Random Forest Algorithm

- Random forest is a commonly-used machine learning algorithm.
- A random forest is an ensemble learning method where multiple decision trees are constructed and then they are merged to get a more accurate prediction.
- Random forest became popular because of its ease of use and flexibility in handling both classification and regression problems.



Random Forest Algorithm - Steps

100 10 $n=30$

1. Build random forests :

   n_1 n_2

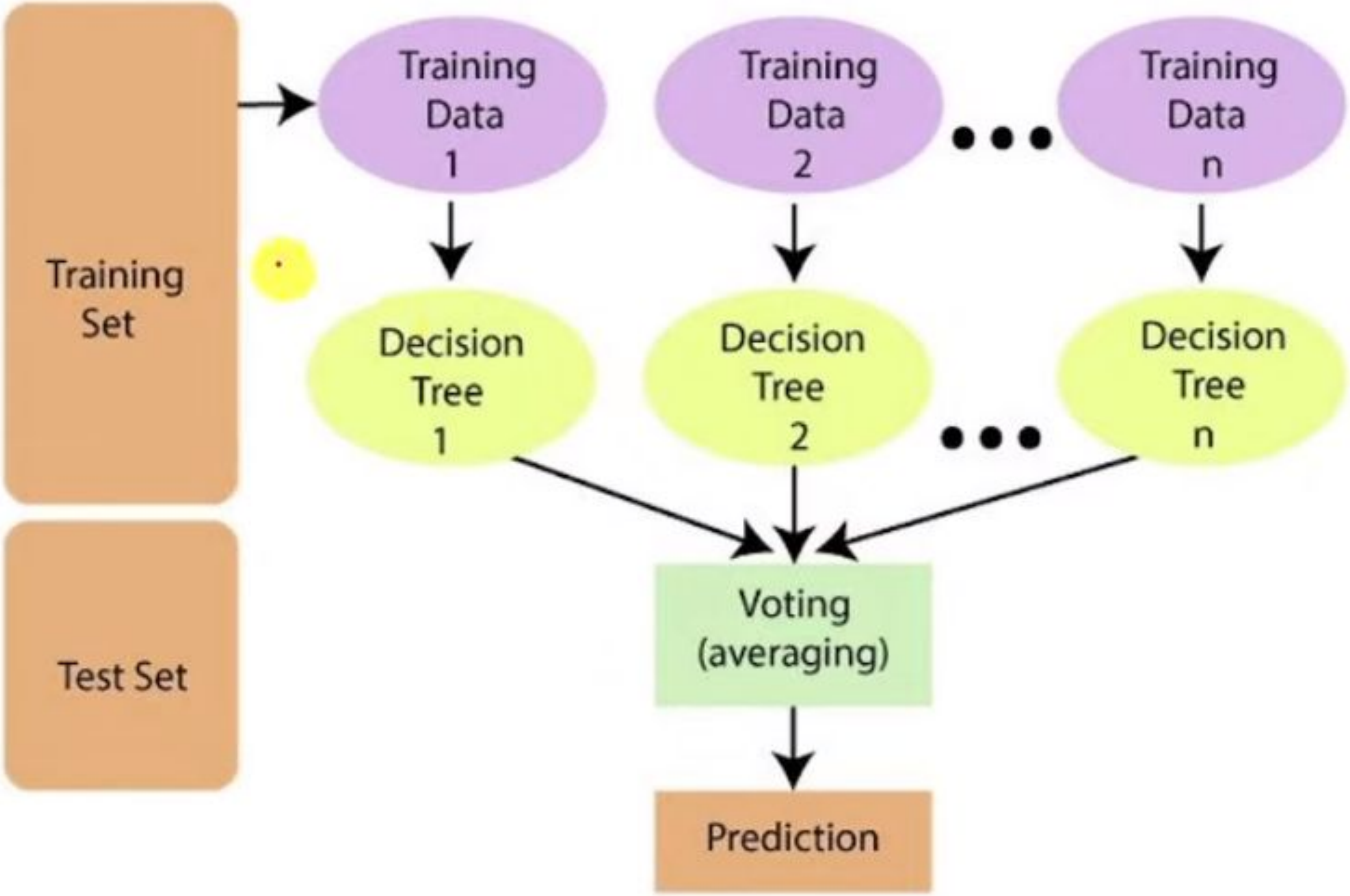
   a) If the number of examples in the training set is N, take a sample of n examples at random - but with replacement, from the original data. This sample will be the training set for generating the tree. 10 5

   b) If there are M input variables, m variables are selected at random out of the M and the best split on these m is used to split the node. The value of m is held constant during the generation of the various trees in the forest.

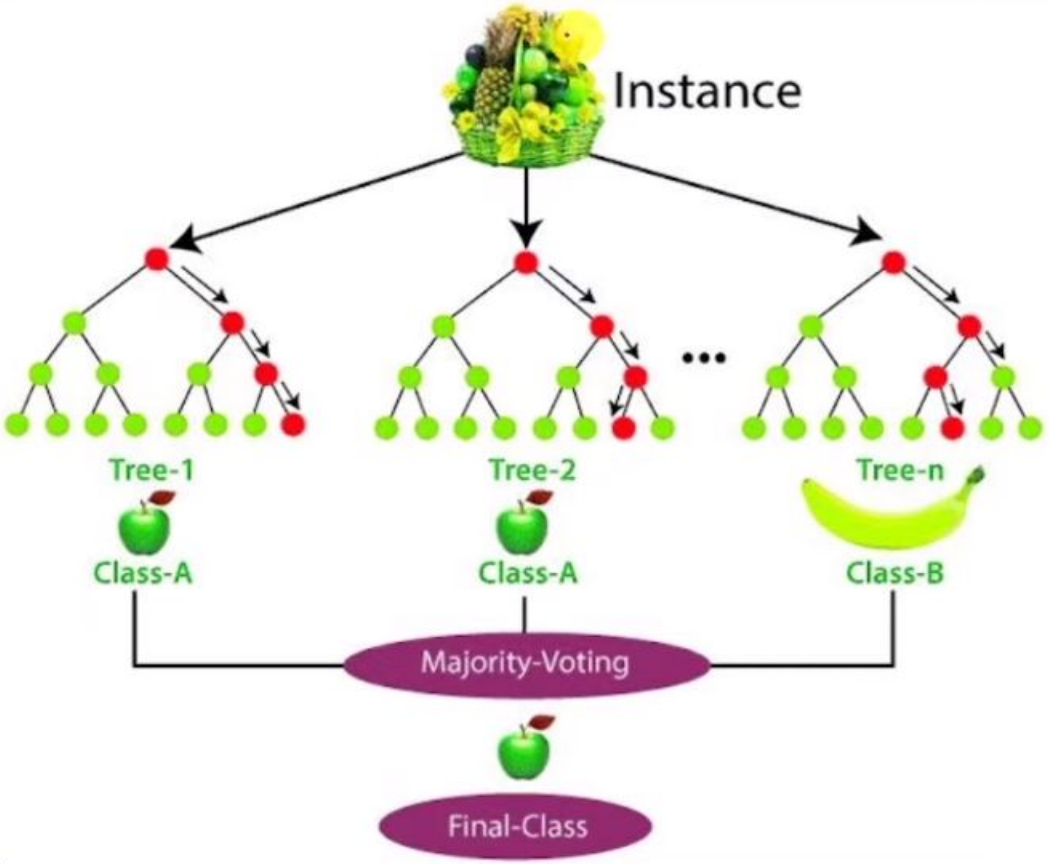
   c) Each tree is grown to the largest extent possible.

2. For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the **majority votes**.

Random Forest Algorithm - Steps



Random Forest Algorithm - Steps



Random Forest Algorithm - Strengths

1. It takes less training time as compared to other algorithms.
2. It predicts output with high accuracy, even for the large dataset it runs efficiently.
3. It can also maintain accuracy when a large proportion of data is missing.

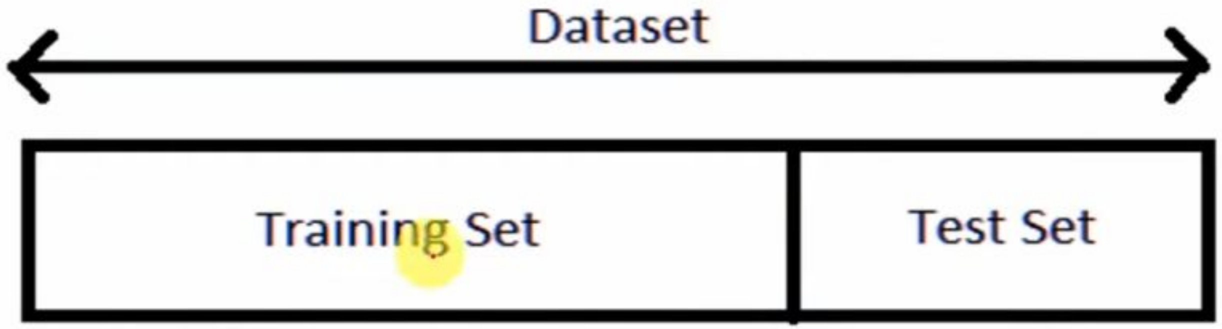
Random Forest Algorithm - Weaknesses

1. A weakness of random forest algorithms is that when used for regression they cannot predict beyond the range in the training data, and that they may over-fit data sets that are particularly noisy.

2. The sizes of the models created by random forests may be very large. It may take hundreds of megabytes of memory and may be slow to evaluate.

3. Random forest models are black boxes that are very hard to interpret.

Train Test datasets in Machine Learning

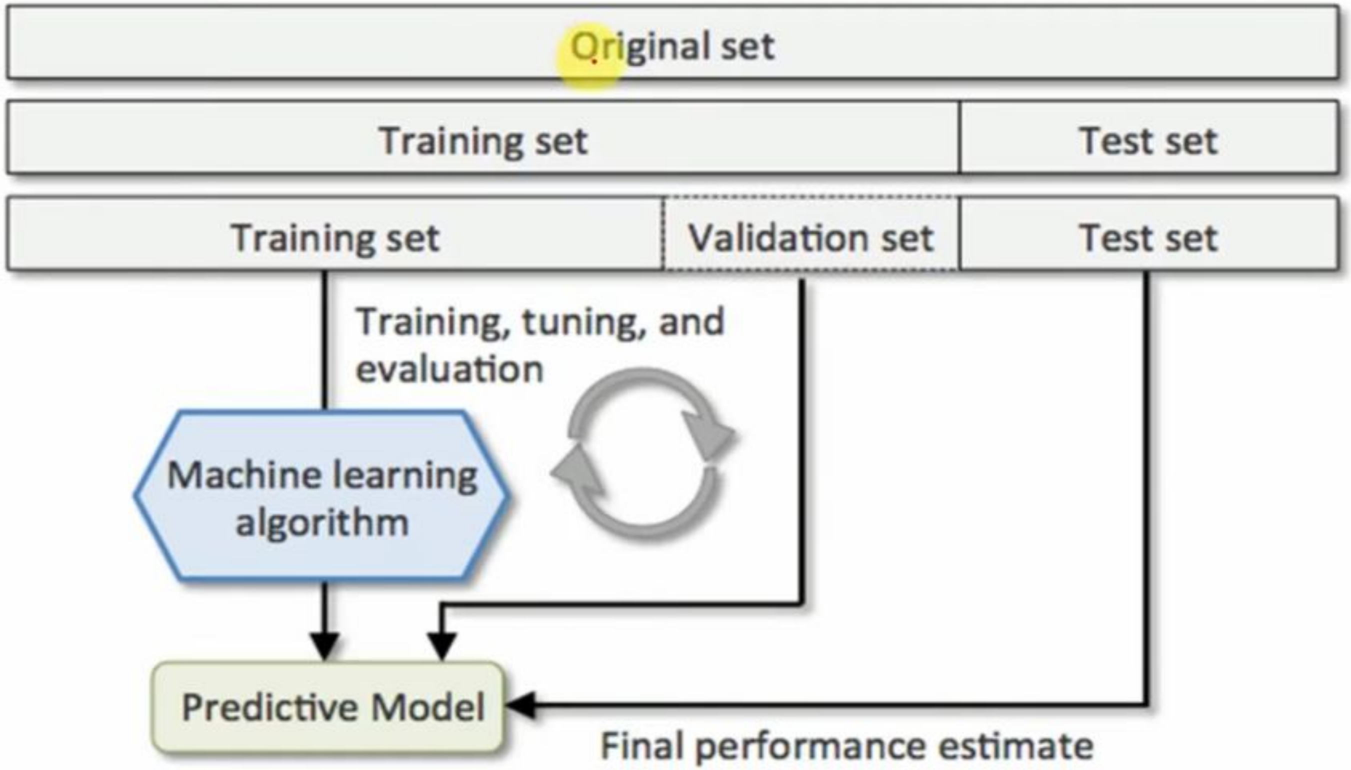


Dataset

Training Set

Test Set

Train Test Validation datasets in Machine Learning



Original set
Training set
Test set
Training set
Validation set
Test set
Training, tuning, and evaluation
Machine learning algorithm
Predictive Model
Final performance estimate

Cross Validation in Machine Learning

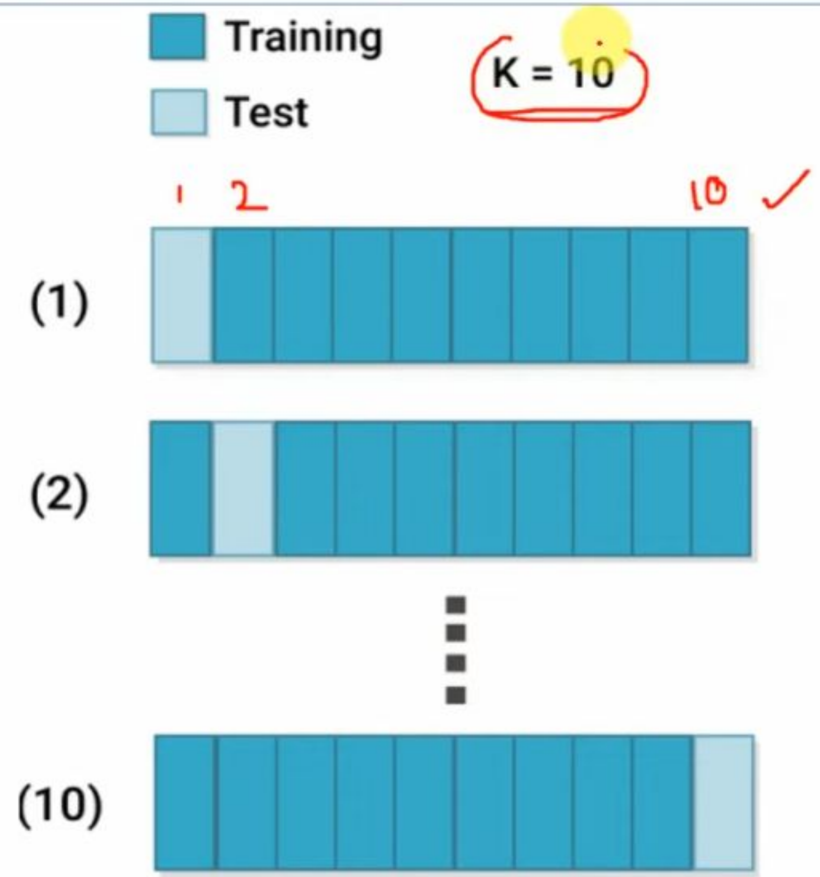
- The idea of cross-validation arises because of the *problems* with train test model or train test validation model.
- It basically wants to guarantee that the score of our model does not depend on the way we picked the train and test set.

Types Cross Validation in Machine Learning

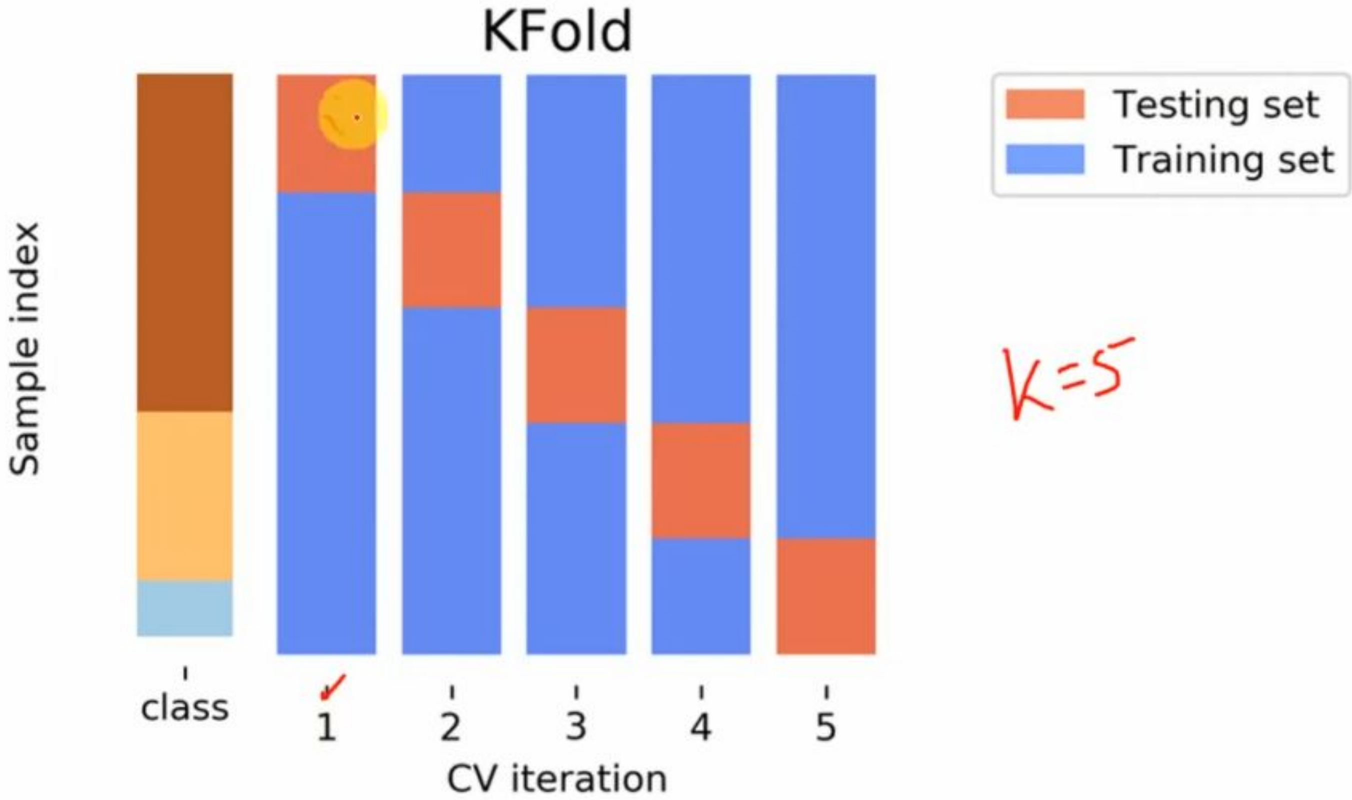
Following are the type of Cross Validation Techniques

- K-folds ✓
- Stratified K-folds
- Leave-one-out
- Leave-p-out

K-Fold Cross Validation in Machine Learning

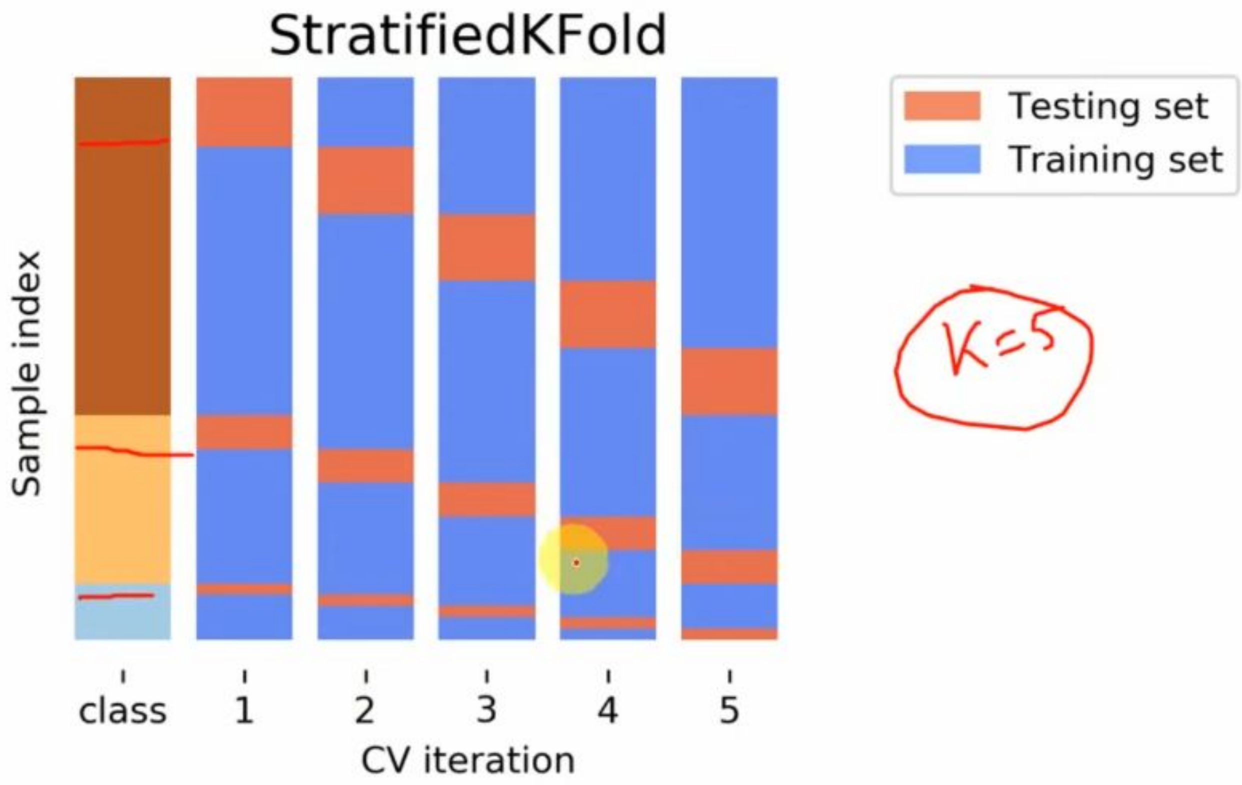


K-folds Cross Validation in Machine Learning



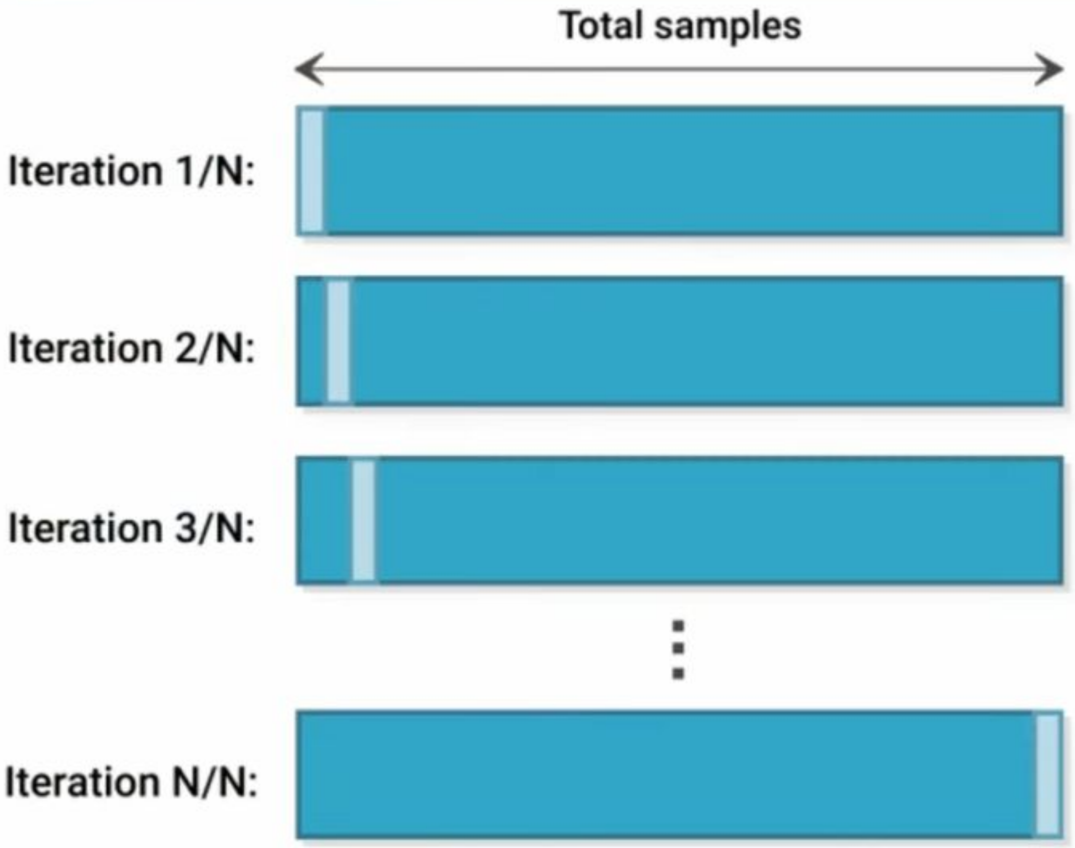
K=5

Stratified K-folds Cross Validation in Machine Learning

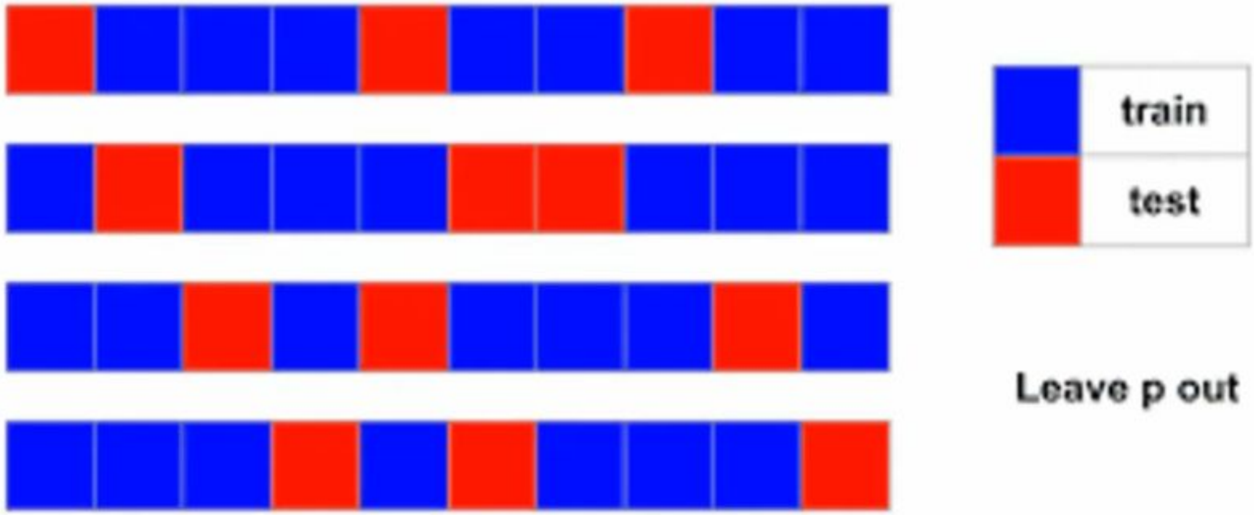


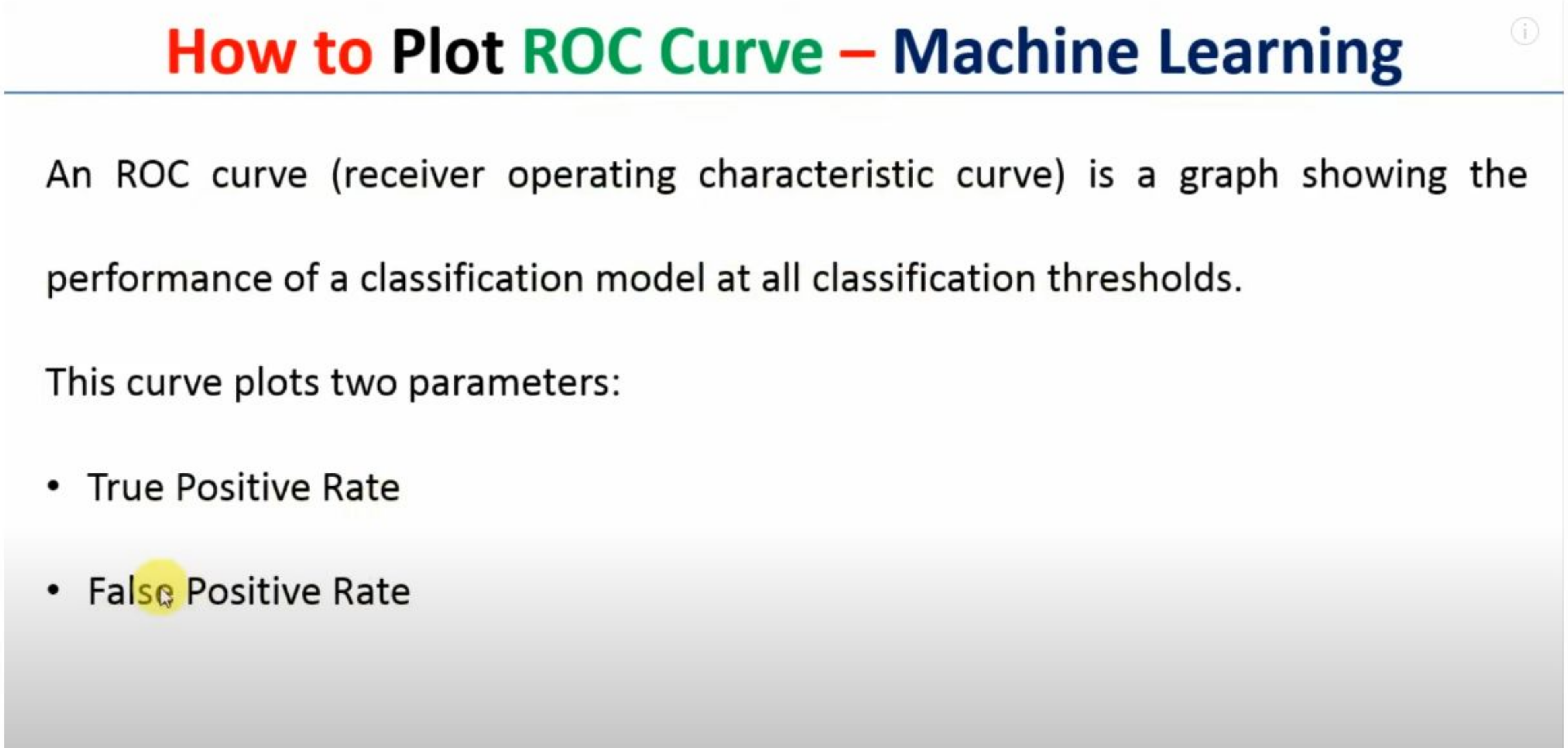
K=5

Leave-one-out Cross Validation in Machine Learning



Leave-P-out Cross Validation in Machine Learning





How to Plot ROC Curve – Machine Learning

An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds.

This curve plots two parameters:

- True Positive Rate
- False Positive Rate

How to Plot ROC Curve – Machine Learning

Tuple	Class	Prob	TP	FP	TPR	FPR
1	P	0.90				
2	P	0.80				
3	N	0.70				
4	P	0.60				
5	P	0.55				
6	N	0.54				
7	N	0.53				
8	N	0.51				
9	P	0.50				
10	N	0.40				

P → P

N → P

✓ TP = True Positive

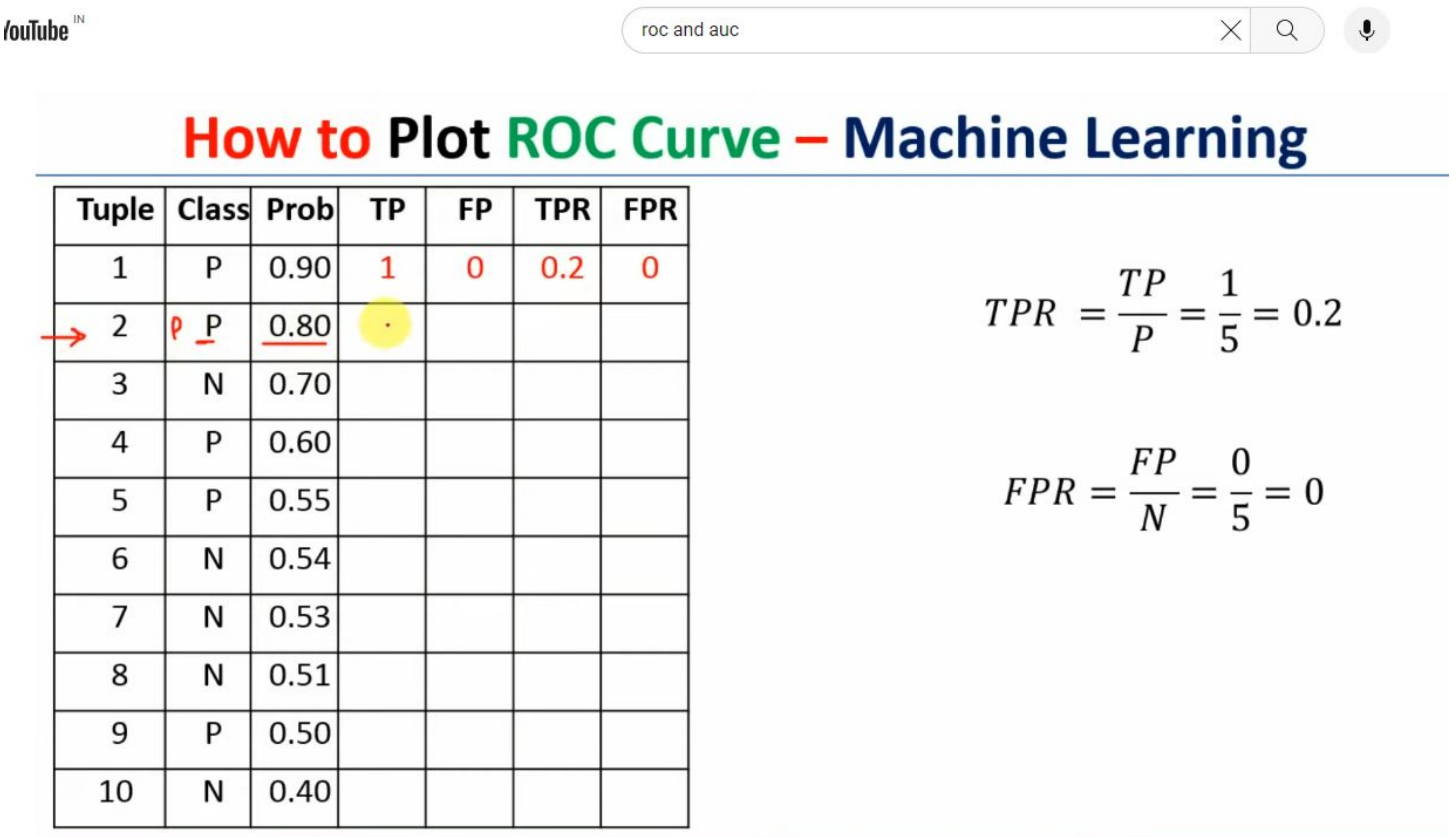
✓ FP = False Positive

TPR = True Positive Rate

$$\underline{TPR} = \frac{TP\,✓}{P\,✓}$$

FPR = False Positive Rate

$$FPR = \frac{FP}{N}$$



How to Plot ROC Curve – Machine Learning

Tuple	Class	Prob	TP	FP	TPR	FPR
1	P	0.90	1	0	0.2	0
→ 2	P P	0.80				
3	N	0.70				
4	P	0.60				
5	P	0.55				
6	N	0.54				
7	N	0.53				
8	N	0.51				
9	P	0.50				
10	N	0.40				

$$TPR = \frac{TP}{P} = \frac{1}{5} = 0.2$$

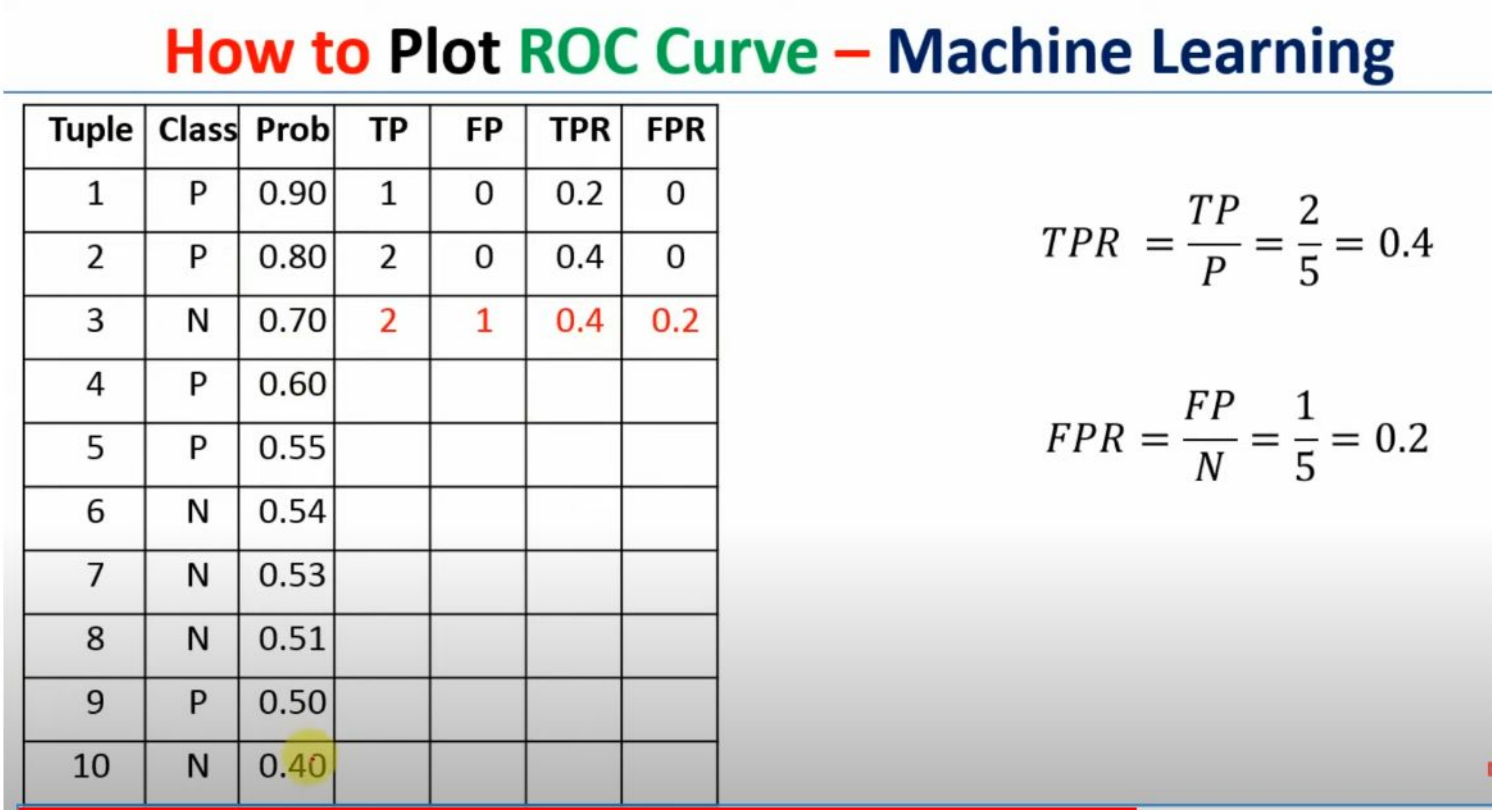
$$FPR = \frac{FP}{N} = \frac{0}{5} = 0$$

How to Plot Roc Curve – Machine Learning

Tuple	Class	Prob	TP	FP	TPR	FPR
1	P	0.90	1	0	0.2	0
2	P	0.80	2	0	0.4	0
3	N	0.70				
4	P	0.60				
5	P	0.55				
6	N	0.54				
7	N	0.53				
8	N	0.51				
9	P	0.50				
10	N	0.40				

$$TPR = \frac{TP}{P} = \frac{2}{5} = 0.4$$

$$FPR = \frac{FP}{N} = \frac{0}{5} = 0$$



How to Plot ROC Curve – Machine Learning

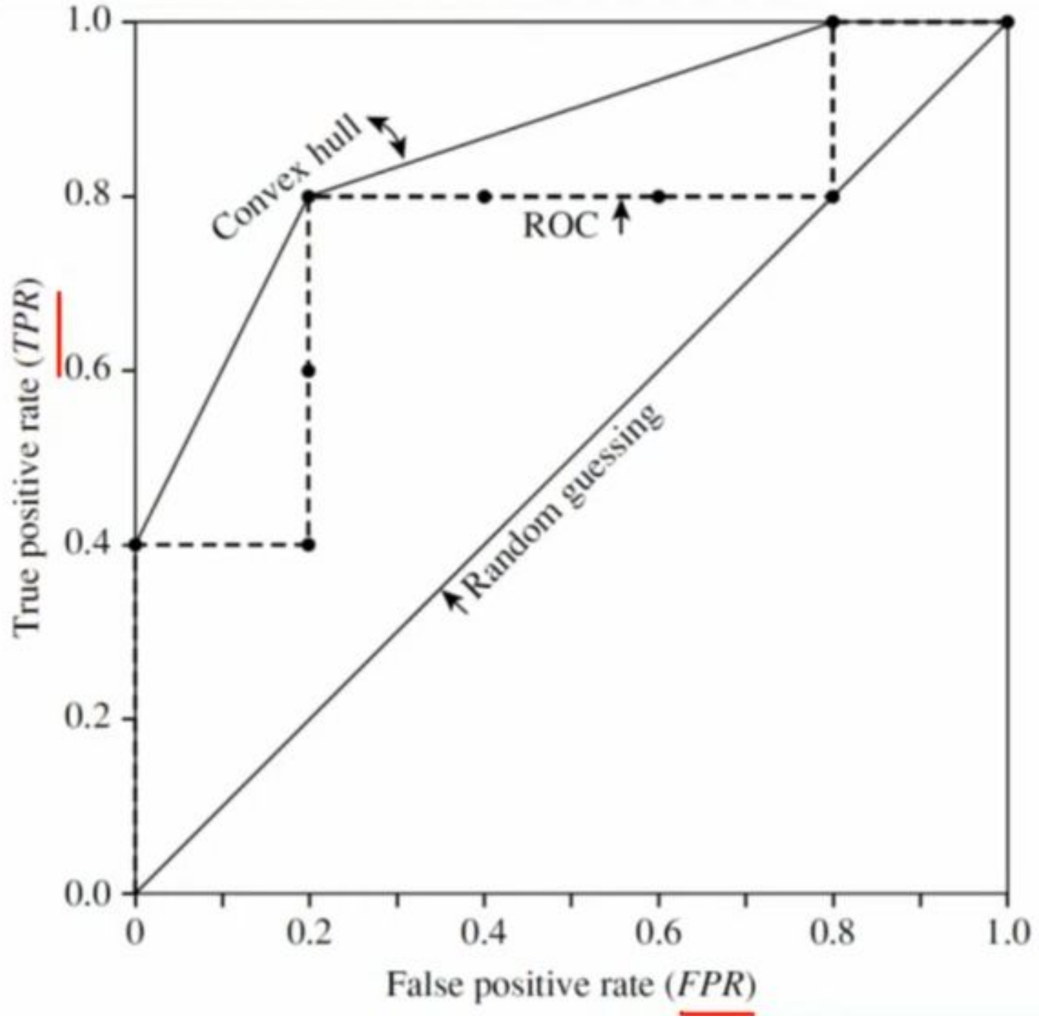
Tuple	Class	Prob	TP	FP	TPR	FPR
1	P	0.90	1	0	0.2	0
2	P	0.80	2	0	0.4	0
3	N	0.70	2	1	0.4	0.2
4	P	0.60				
5	P	0.55				
6	N	0.54				
7	N	0.53				
8	N	0.51				
9	P	0.50				
10	N	0.40				

$$TPR = \frac{TP}{P} = \frac{2}{5} = 0.4$$

$$FPR = \frac{FP}{N} = \frac{1}{5} = 0.2$$

How to Plot Roc Curve – Machine Learning

Tuple	Class	Prob	TP	FP	TPR	FPR
1	P	0.90	1	0	0.2	0
2	P	0.80	2	0	0.4	0
3	N	0.70	2	1	0.4	0.2
4	P	0.60	3	1	0.6	0.2
5	P	0.55	4	1	0.8	0.2
6	N	0.54	4	2	0.8	0.4
7	N	0.53	4	3	0.8	0.6
8	N	0.51	4	4	0.8	0.8
9	P	0.50	5	4	1.0	0.8
10	N	0.40	5	5	1.0	1.0


True positive rate (TPR)
Convex hull
ROC
Random guessing
False positive rate (FPR)