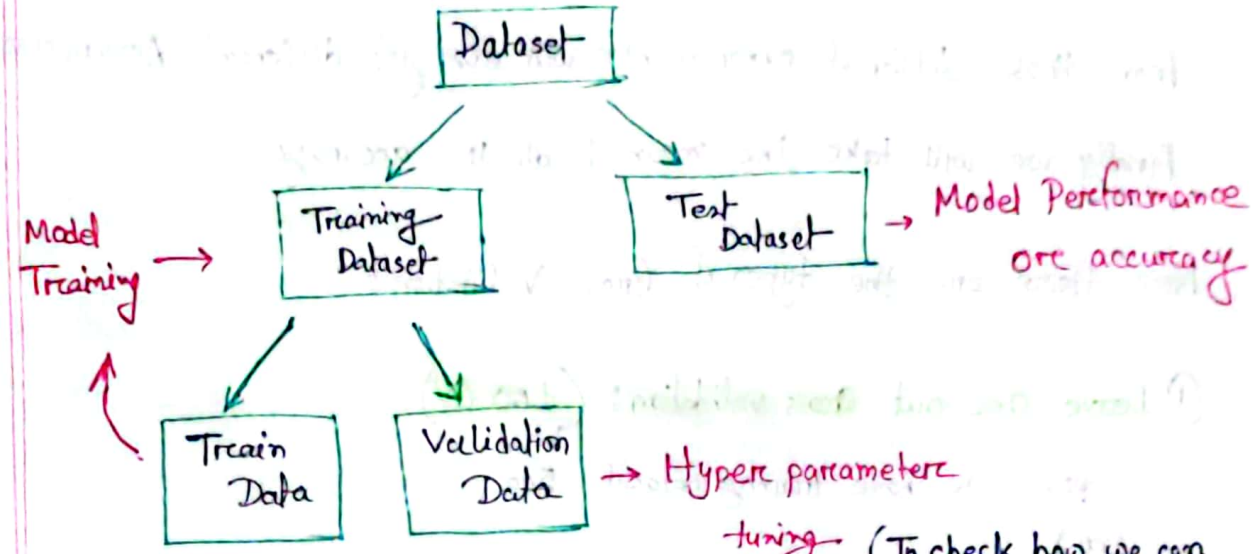
Cross Validation and it's type:



In scikit-learn we use the model ~~test~~ train, test, split, in that module we use a "random state" parameter.

What this parameter does is, it converts the training dataset into train data and validation data according to the given value.

Accuracy of the model depends on this ~~xxx~~ parameter.

Every time you change the parameter value, it will produce different number of Train and Validation data and for that reason each time model accuracy will also change.

That's ~~my~~why to bring out the actual accuracy, we need cross validation.

CamScanner

In cross validation, we will have 5 ~~validations~~ experiments.

Every Experiment there will be different train and validation data.

From those different experiments will also get different Accuracies.

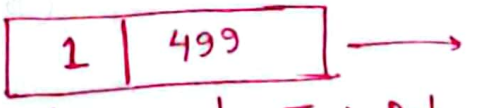
Finally we will take the mean of all the accuracy

Now, Here are the types of Cross Validation:

① Leave One out Cross validation: (LOO CV)

Suppose we have training dataset = 500

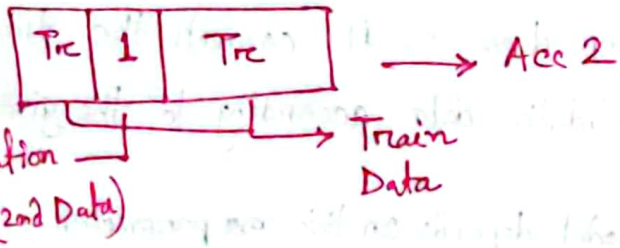
Exp 1:



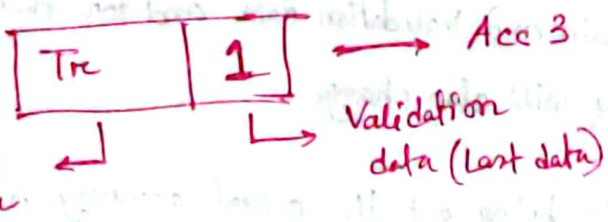
Train data → Model training

Validation data → Model testing to predict Acc.

Exp 2:



Exp 500:



So, we have conducted 500 Exp for 500 dataset and got 500 Accuracy results. Now, we will take the "mean" of 500 accuracy.

That is the leave one out Cross Validation.

The issue with this Cross validation technique:

① Time Complexity is huge for training dataset. (Suppose for 1M dataset)

② Usually Model overfit ⟶ Training Acc ↑↑
Validation Acc ↓↓

2nd type of Cross Validation Technique:

② Leave P out of Cross Validation:

Here will place P records in the validation data instead of 1.

P can be = 10, 20, 30, 40 ⟶ Hyper parameter

Here the number of Experiments will be less as P values are high.

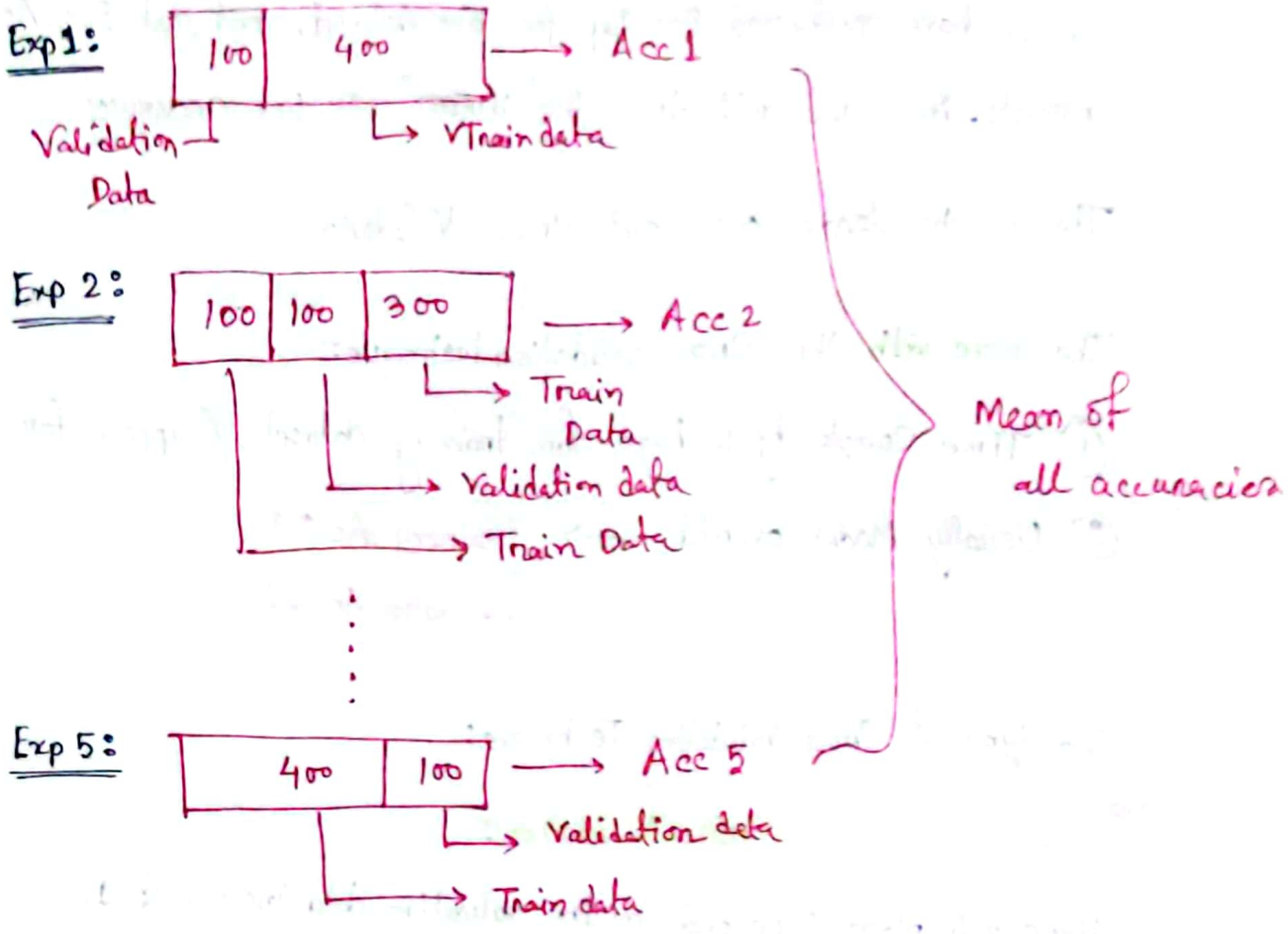
K Fold Cross Validation Technique: (The most useful)

We have to specify a K value (say 5) and Total Training dataset say: 500

∴ Total experiments = $\frac{500}{5} = 100$

This 100 is the validation size.

So our experiment would be ⟶



4th type of Cross Validation Technique:

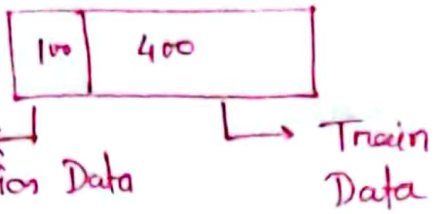
Stratified K Fold Technique: (used For imbalanced Dataset)

Suppose our dataset is Imbalanced: $\{ 350 \rightarrow 1,\ 150 \rightarrow 0 \}$

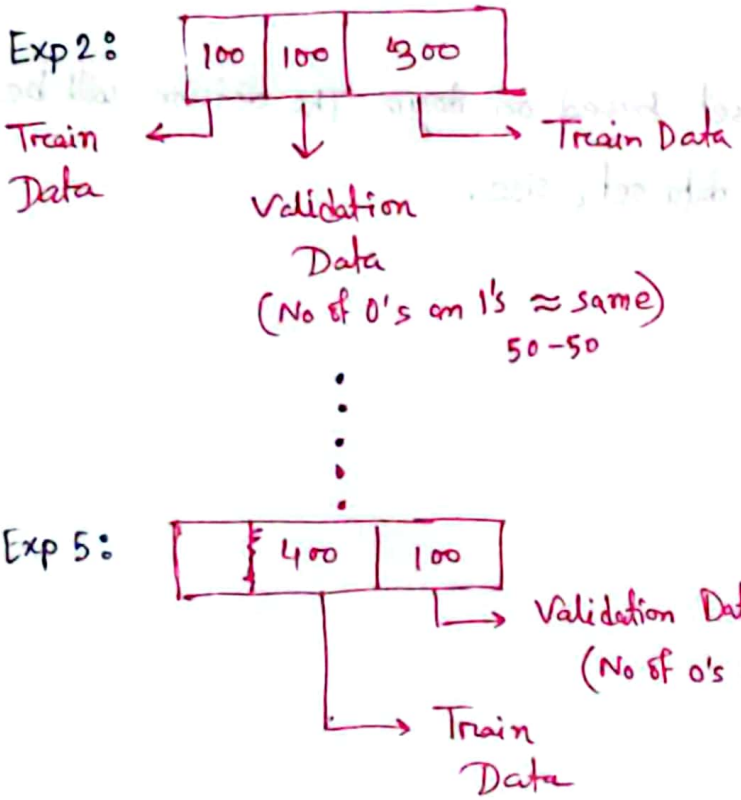
$K=5, \quad n=500$ (dataset) $\therefore$ validation size $=100$

But here what the difference is, this tecnique kept the approximately same number of 0 and 1 dataset in the validation part in every experiment

Exp 1:



Exp 2:



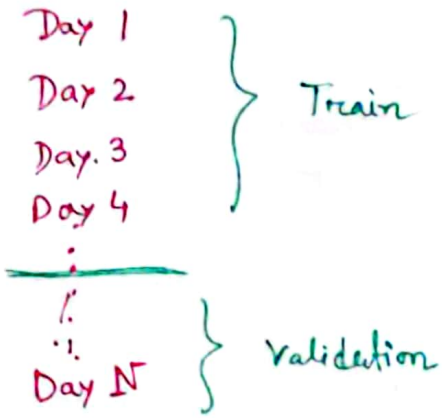
Exp 5:

5th type of Cross Validation:

Time Series Cross Validation:

Suppose we have to make a sentiment analysis for Amazon product review.

We have reviews from Jan-Dec.



Here, we split the dataset based on days. The division will be dependent on based on data set, size.