Now, here we come to the most important part of this project.

General procedure we use in machine learning is to simply divide the data into training set and testing set and then create model on training set and test it on testing set but sometimes this is not a good thing to do because when we play with the things called hyperparameter tuning then these things may cause a problem in a way that we may be become more dependent on the testing dataset. The result comes out with the test dataset and we start looking for accuracy on training dataset this creates the dependency on the test dataset.

In order to handle that we come up with the validation dataset, here we divide the dataset into 3 parts rather than two parts.

Here, those three parts are:

1. Training set
2. Cross-validation set
3. Testing set

In training set, we are going to build our model.

In cross-validation, we treat it as a test set where we test our hyperparameter tuning and all the other stuff. We going to know that what the best parameter we get out of it.

In test set, we do not touch it yet, when we completely build our model then we use this test dataset for predictions and testing stuff of our dataset.

Here, we can adopt any method like 70-20-10 distribution etc.

I have adopted first split 80-20 in training and test respectively. Then again splitting the training into 80-20 ratio in training and cross-validation respectively.

Now, after splitting it is very important that we do notice that distribution of the data happened properly because if the distribution does not happen properly many kinds of problem can occur like:

* Here we have 9 classes, suppose class 1 components are stayed in the training dataset and there is a chance that they haven’t gotten into the cross-validation and testing set. If this happens then there is no way to bringing in and verify that whether my model is to bringing the class 1 is good or not.
* It can also happen that suppose class 7 components we have here is not much present in the training set, so if that happens then we do not have much information in training set to train our model and get accurate predictions because not much patterns available to study or to learn from the training set.

So, classes distribution from the raw data is very difficult to compare. Hence, as a data scientist I try to visualize this dataset distribution.