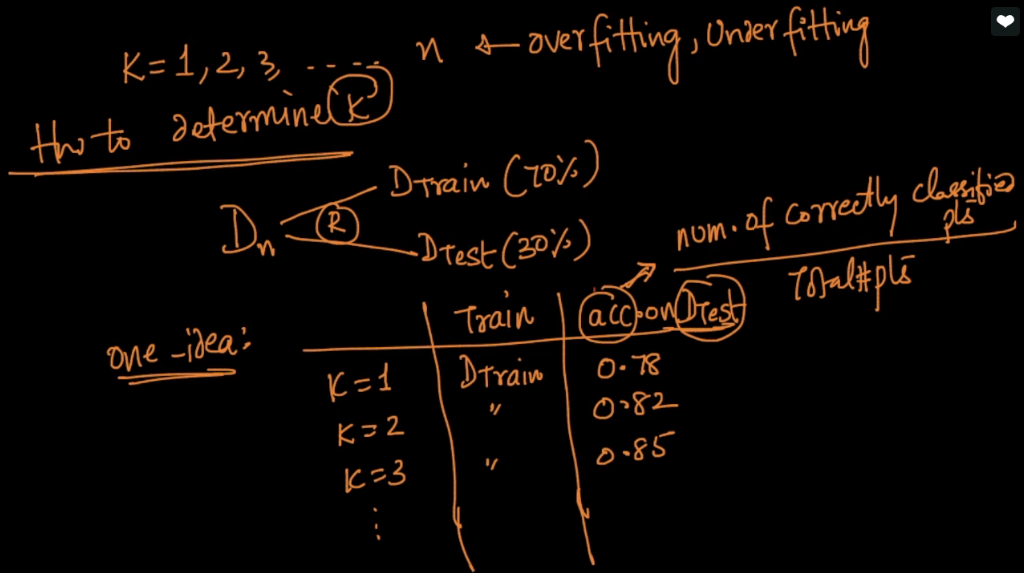
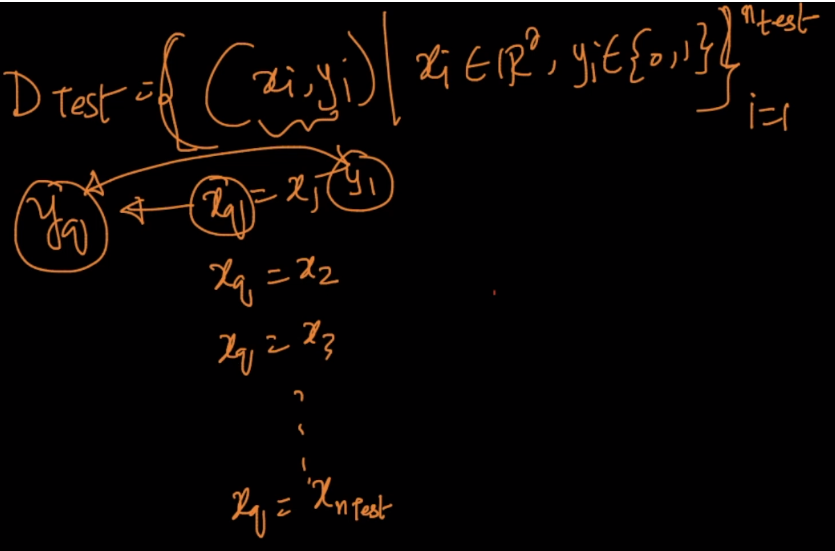
**How to determine K, such that model don’t get overfit and underfit:**

We’ll split dataset into training and test datset, and we train k-NN on training dataset and check for accuracy on test dataset for different k values.

Accuracy = no. of correctly classification points / total no. of points.

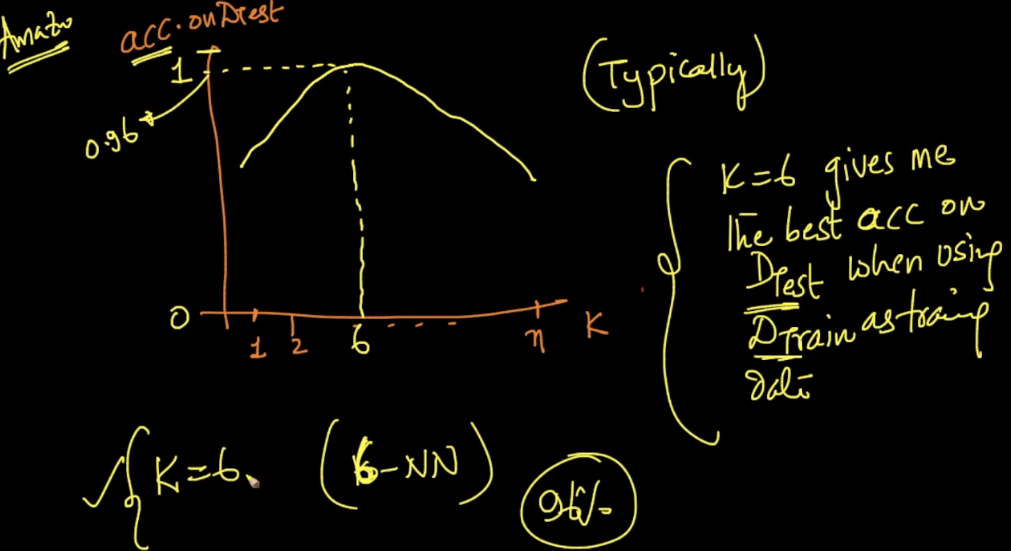
We can check for correctly classification as for each data point in test dataset we have it’s original output associated with it.

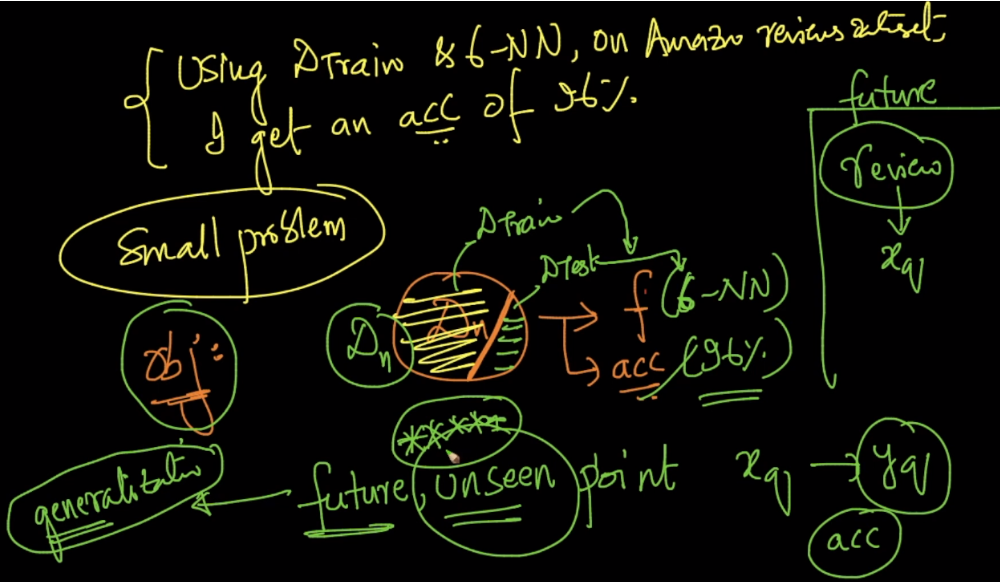




Then we plot a graph for this where K lies on X-axis and accuracy on test data lies on Y-axis.

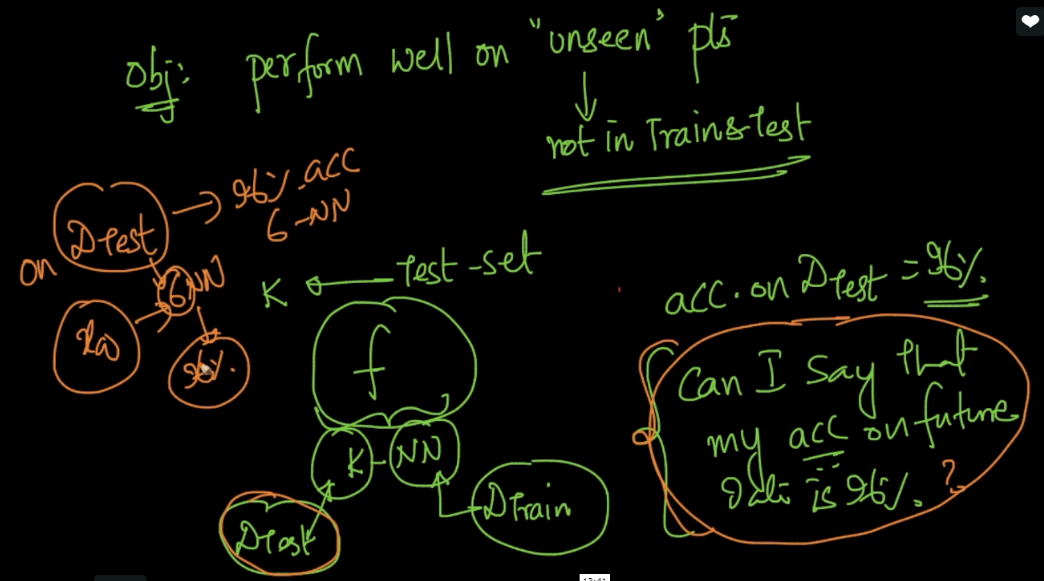
The value of K where we get highest accuracy, we’ll pick that value, as for current ex we are getting highest accuracy of 0.96 at 6 as k value.





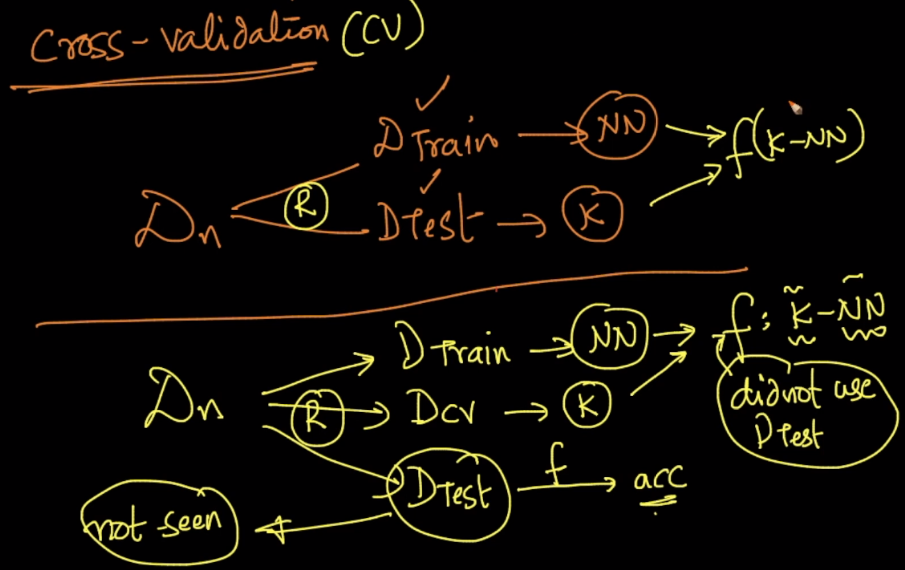
Now Since we get 96% of accuracy of our k-NN for k=6, so can we say that we’ll get accuracy of 96% on future unseen datapoint.

Answer is no, because we have trained model using training dataset, and for picking up correct K value we’ve used test data. So both are used, and it’s not evaluated for unseen data, therefore answer is no.

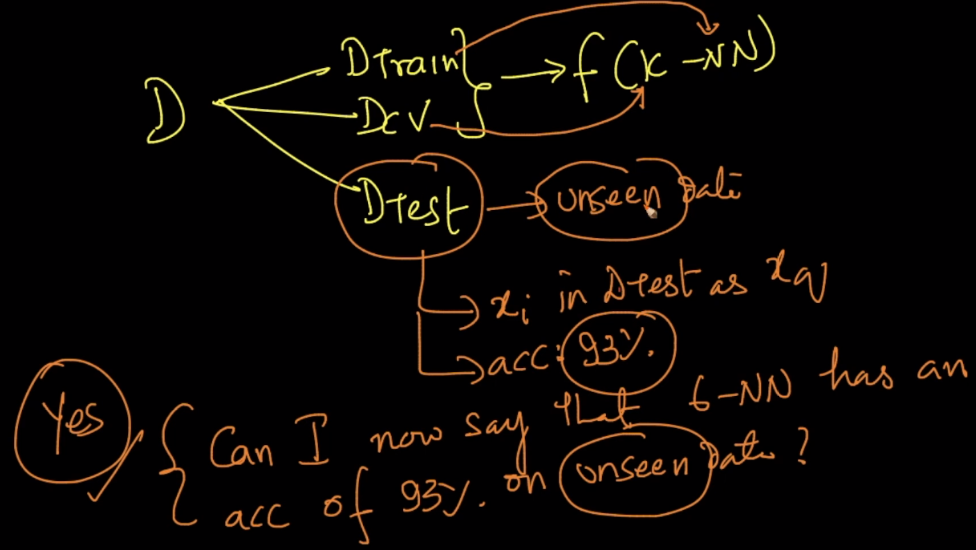


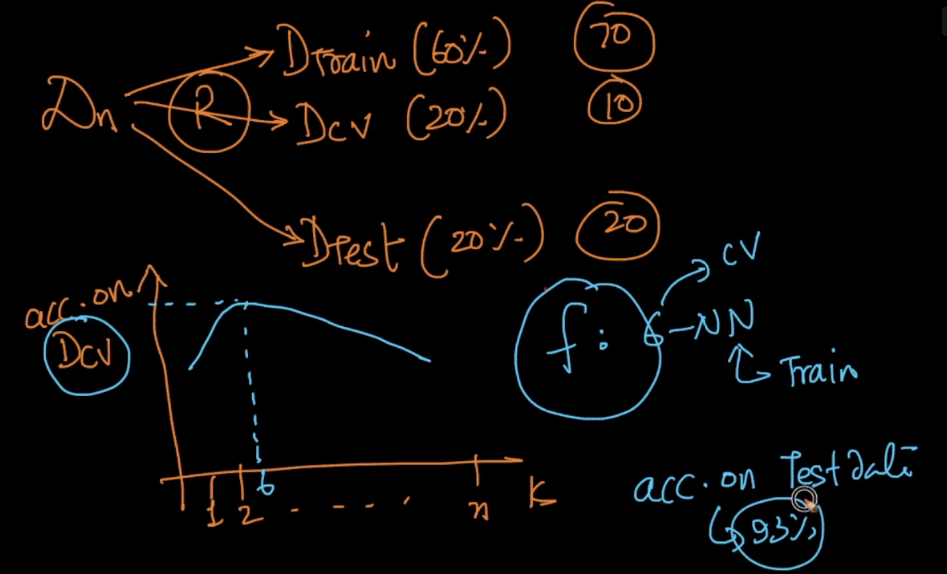
Now to determine accuracy for future unseen data we now divide dataset into 3 datasets.

* Training dataset: used to train k-NN
* Cross Validation dataset: used to identify correct K value by cross validataion.
* Test dataset: used to calculate accuracy for unseen data.



Now after splitting dataset into 3 datasets, the accuracy calculated using test dataset(let say we get 93%) can be said as **that for future unseen data we’ll get 93% accuracy.**





The final accuracy we get on test dataset is called generalization accuracy and 100-accuracy is called generalization error.

