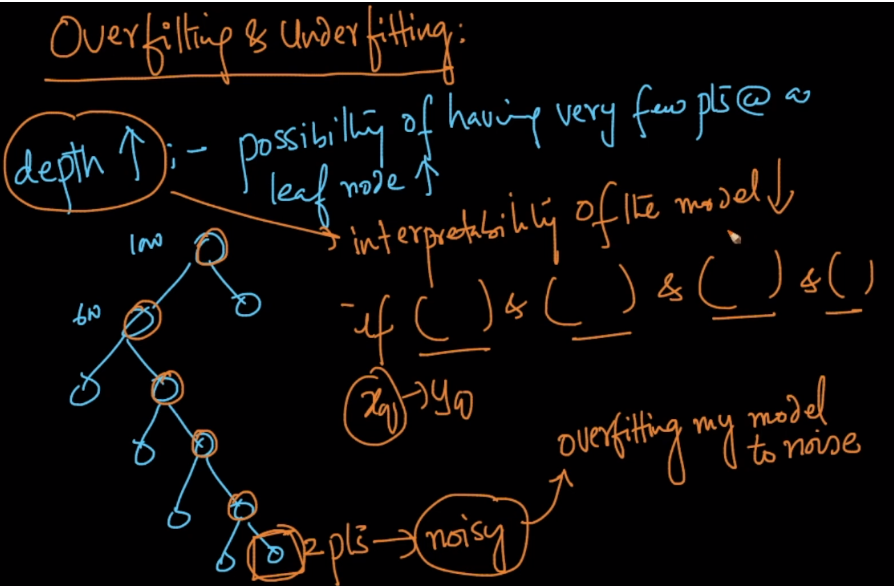
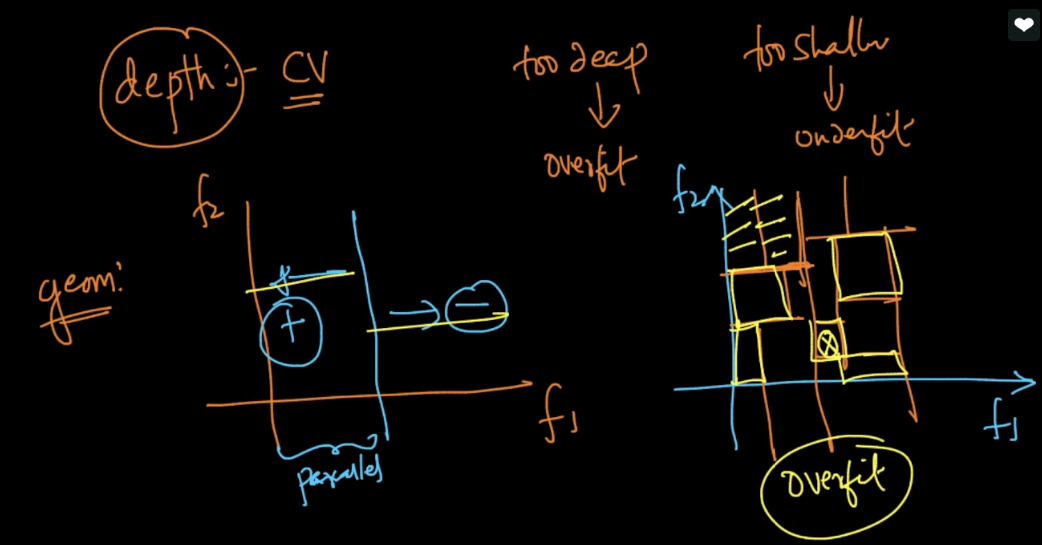
If depth of tree is more, then there will be very low no. of points remaining on the lower depth nodes which will be used for classification, and there may be case where that points are outliers or noisy, so with more depth overfitting happens.



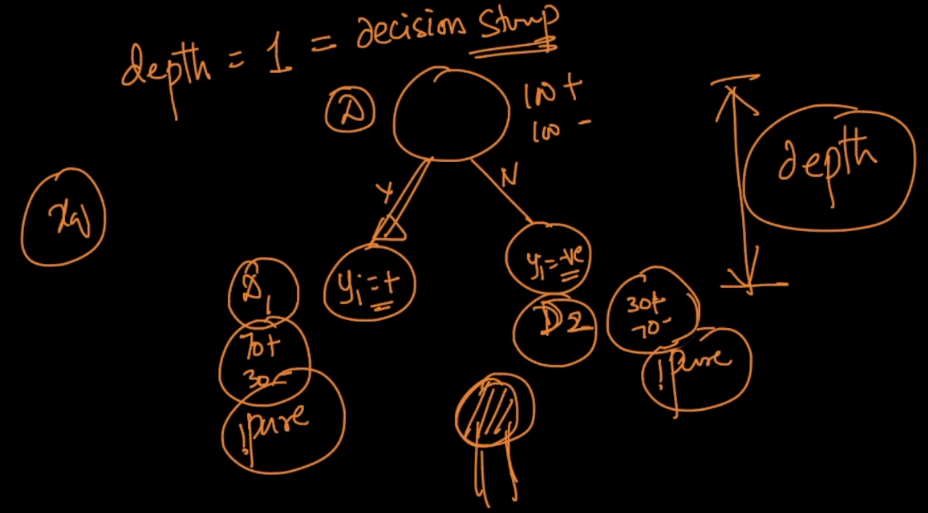
See if there are more depth then there will be too many small blocks, and some blocks may have 1 or 2 points which decides the output of new data point if fall in that region, this is like knn with k = 1, now if those points are outliers then our DT also trying to fit into it, and classifying according to outliers class which is overfitting.



But how do we know the output we keep depth less and we won’t get pure nodes.

Let’s take example where d = 1.

So by doing this at left node we have 70 +Ve and 30 –ve and at right node we have 30 +Ve and 70 –ve, Now at whatever node we reach, we’ll do majority voting, like if we reach at left node, then we’ll do majority voting and since here +ve > -ve, we tells that it has +Ve class.



But d = 1 will cause underfit, so what we do is we’ll use cross validation to find the appropriate value of d, such that our model is neither overfit nor under fit.

