So guys, in this video we are going to discuss about select clustering. Now already in our clustering algorithms we have covered something like DB scan. We have covered K-means clustering hierarchical clustering. Let's say with the help of K-means clustering right. Let's say for a specific problem statement. In the K-means clustering we have selected the k value is equal to four with the help of elbow method. Right. Now how do I validate this specific model or this k value is super suitable for this particular problem statement. Right. How do we validate that k is equal to four will actually give us the best results. You know, similarly let's say if you have considered any supervised machine learning algorithm like classification problem statement okay. In that we will be using some performance metrics. Right. What kind of performance metrics will be using will be we may be using things like accuracy. We may be using precision recall and why we are using those performance metrics to basically validate this classification model. Right. To see whether the my model is performing well or not. So similarly, slight scoring is one amazing technique which will actually help you to validate the unsupervised machine learning algorithm like k means or hierarchical mean clustering. Okay, so we'll try to understand what exactly it's like clustering. How do we basically apply it. What are the steps over here in in understanding or in getting the centroid score okay. So here is all my steps. This is my first step. I will discuss about it with and provide you all the examples. Uh second step. And finally this is my third step right. So let's go ahead and let's understand over here. Now the first step says that for every data point that belongs to C of I. Now let's consider that this is my cluster okay. And in this specific cluster I have all this data points okay. These are all my data points that are present inside this cluster okay. Now the first point in silhouette scoring or silhouette clustering, it basically says that for every data point I belongs to C of I, let's say this is my C one cluster. And within this particular cluster I'm just going to consider a point that is this specific point. Okay. Let's consider this is I okay. We are going to compute the distance. And this distance will be the distance. And we can also say this distance can be the average distance from this particular point to all the other points that are present in this cluster. Okay. So that is what this first statement says. For data point I belongs to C of I data point I in the cluster C of I let a of I. So here we are basically computing a of I and a of I is nothing, but it is the mean distance between I and all the other data points in the same cluster. Okay, so that is the reason why we are dividing by one by c of I minus one. Why minus one. Because I'm not computing the distance from this particular point to itself. Right. So that is how we basically compute a of I by using this specific formula. Okay. Please remember this formula. In short what we are doing we have we have just considered one cluster. We have taken one point. And from that particular point we have computed the distance with average distance with all the other points. Right. So we have done the summation of all the distances and then divided by total number of points minus one okay. So this is the first step. And with by this way we will be able to compute a of I okay. Now let's go to the second step. In the second step it says that we then define the mean dissimilarity of I to some cluster C of j. Let's say that this was my cluster over there and this was my point okay. Let's say this is my C one cluster. Now the first step what we did is that we from this particular point we computed all the average distance. Right. Like this. And we computed a of I. This is perfectly fine. Now we'll go and compute B of I. Now what exactly is B of I? What it will do is that we'll take the nearest cluster from this particular cluster. Let's say this is my nearest cluster C of two which is very near to this specific cluster. There may be other clusters also. There may be other clusters that may be available over here. But the nearest cluster is this specific cluster. Let's say this is C three. Now let's say here we also have some data points. Now what we are going to do in the second step is that from this particular data point, right from the same data point that we used in the cluster one, we will try to compute the distance. And this distance will again be the average distance average summation of the

distance. Right. So here what we are doing. We are just computing it. And by this what we will be getting we'll get B of I. Right. Minimum J equals to I. That basically means which is the nearest cluster over here. And we are trying to find out the average distance from this particular point to all the other points that are present in this specific cluster. And by that specific way, we are we are we will be able to get B of I. Okay. Again, let me repeat all the points. We then define the mean dissimilarity of point I to some other cluster C of j. So this will be the nearest cluster as the mean of the distance from point I to all the other points in C of J. So for each data point I belongs to C of I, we now define uh b of I I. This minimum is basically used because we are going to consider the minimum, uh uh, the nearest cluster which has the minimum distance. And from this specific point we can go ahead and compute the average distance of all the other points. That is the reason why we are dividing by C of J, because that many number of points are present inside this particular cluster. Right. So in this particular case this will be my C of I and this will be my C of J. Okay. So in this particular case I will be getting B of I. Now obviously from this particular scenario what observation we can definitely take out that a of I is definitely going to be B of less than B of I if my clustering is done well, right? Obviously, if my clustering is done well, whatever a of I am getting from this, the distance over here, it will obviously be less than b of I. Right. If B of I uh if a of I is greater that basically means the clustering will not be done. Well. Right. Obviously the clustering, if it is not done well at that point of time, we will be getting a of I. That will be more than B of I right now. Let's go to the third point. Third point which is super important point. Then we calculate the C Lloyd score right. So then we calculate the C Lloyd score. And the formula is very simple B of I minus a of I maximum by a of I comma b of I right. And whenever this particular uh formula we basically apply our value of uh Lloyd score will be ranging between one minus one to plus one. If the more near to plus one. Again I'm going to write more near to plus one better. Clustering model we have created. Better clustering model. We have created. Okay. So here you can basically see B of I minus a of I. This one will be getting the value between minus one to plus one. We can also write this equation like this one minus a of I divided by b of I. Here if a of I is less than b of, I will be getting some values. It will be zero. When a of a and b of I are equal, and if a of I is greater than b of I, then obviously we are going to get this specific value which will be towards minus one, right? So from the definition it is clear that we will be getting this specific value right now. In this particular case, if A of I is less than b of I, then we know that our value will be nearer to one approximately equal to one or nearer to one. If a of I is greater than b of I, then this will be moving towards minus one, right? So that is what you will be able to see. You will be getting the negative values over here and here. We'll be getting the positive values right. So this is how we use for validating the model. So it is a super important concept in silhouette scoring. Again let me repeat in the first step I take a specific cluster. Let's say these are my points. And then with respect to this particular points what we do we calculate the average distance. We will consider this. We'll calculate the average distance. Once we calculate the average distance we are basically going to find out a of I. Then in the next step we we take up another cluster. And from that same point we will try to calculate the average distance to the other points in this specific cluster. And we compute b of I write the formula is basically given up. And then here you'll be able to see once we apply the silhouette score formula, we will be getting the value between minus one to plus one. The more the value towards plus one the more the better model, the more the better the clustering model has been created. Uh, if you are getting the value near to minus one, that basically means B of I is a of I is greater than B of I. So definitely there is something wrong with respect to your clustering model. Right? So this was an idea about the entire silhouette clustering. Uh, in the next video we are going to see some practical implementation. And we'll try to see that by applying K-means

clustering. You know, are we able to get uh, are we able to validate it with the silhouette score or not? Okay. So that I'll be covering in my next video. Thank you.