In this video, we are going to discuss about a new clustering algorithm, which is called as DBscan clustering. Now, already in our previous video we have discussed about K-means clustering. We have discussed about hierarchical mean clustering. Now DBscan clustering is also a clustering technique. We will try to understand how DBscan clustering will actually make clusters. And along with that, in the upcoming videos we will also be discussing about the advantages and the disadvantages. Okay. So let's go ahead and let's understand how does DBscan clustering actually work. Now this is one of the diagram that I've actually taken. And the source is Wikipedia. Definitely I would like to give a huge amount of credit to Wikipedia. So please make sure that you support Wikipedia by doing some donation because it is a non-profit organization. Right. So let's go ahead and let's understand okay. Over here in this diagram you will be finding three main points okay. So let's consider one is the red point. And this I will be considering as a core point. I'm just mentioning it as a core point okay. Let's go ahead with the next point. Over here you'll be also seeing Yellow point. And uh this yellow point is nothing, but it is called as a border points okay. And finally here you will be also seeing one blue color. So let me just quickly take one blue color over here and let me draw it like this. And this particular point is specifically called as outlier okay. Now let's understand which all points becomes a core point in this. And obviously this point are just like my data points that are distributed in this 2D plane. Right? So which point actually becomes a core point? Which point actually becomes a border point and which point actually becomes an outlier point. We'll try to discuss okay. So first of all let's go ahead and discuss about this specific core points. And along with this we will also consider two more parameters. One is minimum points that we have to probably consider. In this particular example we have considered four minimum points. And the second one is something called as epsilon which is nothing but the radius that we are going to use. Uh, to draw a circle around the any specific points. Okay. So this is the two hyper parameters we will basically be using for DBscan clustering. And how do we select this. Again some kind of hyper parameter tuning is also done with respect to this. We play with different different values. But at the end of the day when we will be continuing the series, we will be understanding a topic which is called as silhouette scoring, which will actually help us to select this both the values. Okay, now let's consider that for this particular example I have taken minimum points is equal to four. And this epsilon is nothing, but it is called a. It is a radius some value. Okay. Now for the first one. Right. I'm just going to consider for the red point okay. So let's go ahead and discuss. When does a point specifically point call be called as a core point okay. This is the first scenario. Let's say I have a point I have a data point over here. And we take some epsilon value and we draw a circle around it okay. Now why I have marked this as a red point or a core point over here. There are some conditions. First is that there should be at least. Some minimum number of points that should exist within this circle. So let's say if my minimum point is four. That basically means if I have four points within this and I can also have greater than four. Right. But if I have at least that many number of minimum points within this, then I will consider this as a core point. So here my first condition is that number of points. Within the radius right or epsilon should be. Greater than. Or equal to four. Okay, that is nothing. But that is my specific minimum points. Okay. So this is the most important point over here. So if this particular point satisfies this condition and these points are nothing but the other data points that will be existing in this space. Right. So if this many number of points have been there within the epsilon, then I can definitely consider this as a core point. Okay. Super important. Now let's discuss about the next point which is called as a border point. Now what does a border point actually specify? Okay. Let's say if this is a border point. And I create an circle considering one epsilon value okay. Epsilon value here also epsilon is there. Let's say if there are number of data points within this epsilon value. But. The number of data points will be less than the number of data points. Will be. Within this radius. Okay. Within this radius. Will be. Less than. Or let me write it properly. More. Will be. Less than minimum points. So here what we have done is that here clearly we have seen that when does a point is basically called as a border point? If we have, let's say if you have created the circle with some radius epsilon, if the number of points within this particular circle is less than the minimum number of points, that is, it is less than four. In this particular case I have three. So that point of time I will basically call it as a border point. Okay. It's super important to understand because then only we'll be able to understand this particular diagram, why this is specifically a red point, why this is specifically a border point. Okay. Now coming to the next point, which is also called as a outlier point. Outliers. Now, in this particular case, whenever I take any point and create a circle within the epsilon. No other point will be existing, then I can definitely say this as outlier. Okay, this says outlier. These are the basic differences between all this particular points. Okay. Now in DBscan clustering what happens is that based on density, you know, based on density in various regions, you know we can actually group based on this core points border points and outliers also. Now let's let's consider this specific example here why this red point is marked as a core point. Because if I draw a circle with some epsilon value, let's say this is my epsilon value that I'm going to consider. Or I can also use some other red point over here. Let's say if this is my epsilon. Now within this epsilon you will be able to see there are at least four number of minimum points one, two, three, four. So if this satisfies this condition um so sorry it is 1234 okay. We can also combine this specific point and we can also count it okay. So if that many number of minimum points are basically present I will mention this as a core point okay. Let's say why this is also marked as a red point. Because if I probably draw a circle right over here also you can see 1234. So four minimum points is basically present. So that is the reason why we are marking this as the red point. Similarly let's say whether this this is also a red point or not. If I probably draw a circle with the same epsilon one two, three, four is basically existing. Right? So that that is the reason why we basically mentioned this as a, uh, core point right now. Let's go ahead and see with respect to this specific border point now here, this is nothing, but this is my border point. This is my border point. Right. And if I probably consider this this is my core point. Right. And finally here you'll be able to see. This is my. Outlier point, right? Outlier. Okay. So or I can also say this as noise. Noise basically means outlier only okay. So this is completely noise. So this outlier I can also mention it as noise. Okay. Super important. Okay. So here why we are calling this as a border point. See when we draw this particular circle you know that minimum number of points are four. But here within this circle I just have two points one and two. Right. So here this particular condition is also not getting satisfied. So I'm marking this as a border point. Similarly I'm marking this as a border point okay. Now based on this what will happen different uh similar clustering will basically happen. Border point will be clustered in a separate way. This core points will be separate, clustered in a separate way. And now is an outlier is going to skip from that specific data. So one important thing with respect to DBscan clustering is that you know it. Even though your data will have noise, your DBscan clustering will be able to handle it in amazing way. Right. So this is the major, major advantage with respect to DBscan clustering okay. Now let me go ahead and show you some of the examples like how the clustering will basically happen. And here you'll be able to see this this this technique that we are specifically using in DBscan clustering. This will also able to handle non-linear clustering also okay non-linear clustering also. So let me just show you some examples and see some kind of outputs directly and how it looks like okay.

So guys now let's go ahead and try to see some of the examples. You know after we apply DBscan clustering on a specific data, what kind of groups we will be able to get. You know, and uh, both these images have been taken again from Wikipedia as the source. Again, a huge shout out to Wikipedia for this amazing resource that they have actually provided to the entire world right now. Over here, you'll be able to see that, uh, you know, in the left hand side here is one specific image. And this is basically saying DBscan can find non-linear separable clusters. This data set cannot be adequately clustered with k means or Gaussian mixture or Em clustering okay. These are some of the other clustering techniques. Now here you can see that all these points that are actually present over here these are nothing. But these are my noise okay. So with the help of DBscan clustering we will not be taking up any noise data. Right. So it is being able to separate completely the noise data. Right. And this all data points that you will be able to see that based on that particular core points and border points, it is being able to group it right. Similarly here also a non-linear structure approach is also there not linear non-linear because it is completely in a different shape. Right. So this is the output that we usually get with DBscan. The second example is quite interesting over here. Now see in the left hand side when we have this specific data, if we try with K-means clustering or if we try with hierarchical mean clustering over here with respect to this specific data, here you will be able to see that it is not able to group together. You know, it is just considering all these things within a single cluster. And even it is taking all the outliers. But after we apply DBscan clustering here, you'll be able to see that we are getting specific different kind of groups, right. So as said for non-linear data it is being able to work well. So here you can see one one group is basically getting categorized over here. This is my another group. This is my third group fourth group fifth group six group. And there are also some points that has been left out over here which is basically an outliers. Right. So at the end of the day when we are applying techniques which are related to this core point, border points and outliers, we will definitely be able to get this amazing clustered groups. Uh, but always understand that at the end of the day we have two important parameters. One is the minimum points and one is the radius. So in the practical example I'll try to show you how to do with uh, how to do with an sklearn library. And we'll try to perform this DBscan clustering by initializing some minimum points and epsilon values as radius. And we'll try to then see in a visualized manner that whether we are able to get some good, uh, number of points or not. Okay. Uh, and we'll also be discussing the advantages and disadvantages with respect to DBscan clustering. Right. So yes I will see you all in the next video. But I hope you have understood how does our output looks looks after we apply a DBscan clustering. So yes I will see you all in the next video.

In this video, we are going to discuss about the advantages and disadvantages of DB scan algorithm. Okay. Now here I have completely displayed the Wikipedia page. The reason is that because it has covered almost all the points, I don't have to write this particular point. Again, I'm just trying to save my some amount of efforts. But just by reading this advantages and disadvantages, you will be able to understand and each and every point. I will also try to make you understand if there is any kind of confusion. Okay, so first point uh, to go ahead with is that DB scan does not require one to specify the number of clusters. That basically means like how in K-means we specify the number of clusters like that in db scan. You don't have to do it because in DB scan we have core points. We have uh, border points, we have outliers or noise. Uh, and based on this, all the points, it will be able to do the clusters properly, which we have already seen in the diagram. Okay. Here also in this particular Wikipedia, here you will be able to see this particular part. See all this clusters is made together right. This entire red points are clustered together because these all are core points. Right. And these are my border points right. So similarly you don't have to any specify any, any, any uh you don't have to specify the number of clusters. In short, coming to the second point DB scan can find arbitrary space clusters. It can even find a cluster that is completely surrounded by a different cluster due to the minimum power points parameter. Now what does this basically mean? First of all, here in the right hand side you can see that we have got a DB scan clustering and it is completely in a non-linear separable clusters. You can also get a linear separable clusters. And you can also get uh clusters which may look something like this. Let me just show you. Suppose I have some data points which may be looking like this. And let's say these are, uh, these are one clusters like this. Okay. And within this you may have another point, which may be having another clusters which may look like this. Right. So it will also be able to create this kind of clusters also. Right. So that is what it is basically saying. It can even find a cluster completely surrounded. Why. Because we use something called as minimum points parameter. So in this case we will be using minimum points parameter and epsilon. Because of this two and obviously because of the functionalities of core point, border point and noise point, we will be able to get this specific kind of clusters. Also the coming to the third point, it is super, super important. Uh, DB scan has a notion of noise and is robust to outliers since it is being able to detect all the outliers over here, and it is not taking it as a part of the specific cluster, it is basically robust to outliers. Super important point with respect to anything. If you really want to understand okay. Uh, and that is one of the most important properties. Like if you have an outlier, I would suggest always go ahead with DB scan. Okay? DB scan requires just two parameters and mostly insensitive to the ordering of the points in the database. Okay. This is just just a simple definition. It is mostly insensitive to the ordering of the points. You know coming to the list. Last one uh sorry. Fifth one DB scan is designed for use with database that can accelerate region queries using R-tree. Okay. So probably there is a, there is a, uh, new amazing technology that is probably coming up or new amazing subject which is called as graph knowledge. There probably DB scan will definitely be getting used. Okay. So it can um, it is used for designed for the use with the databases that can escalate region queries because common queries, that is being basically pulled up. We can directly pick it up from there, and we can retrieve it to the users so that it will be very much faster. And obviously one of the important thing is graph knowledge. If you know about graph knowledge, if you know of some of the concepts with respect to graph knowledge, DB scan will definitely be getting used. The parameters minimum points and epsilon can be set by a domain expert. As I said that these are kind of hyper parameters. So with the help of a domain expert we can basically set it up if you have understood the data properly. Okay. Now these are some of the advantages. Now let's go with respect to disadvantages DB scan is not entirely deterministic. A border point that are reachable from one cluster can be a part of the other cluster. Okay, so what does this basically mean? Let's say that I have a border point over here. Okay. Let's say, uh, let me go over here. Let's say I have a border point over here. This border point can be the border for this cluster also. It can also be the border for this cluster also. Right. So this kind of scenarios will definitely be there. So DB scan that is the reason it says that it is not entirely deterministic okay. This is one of the disadvantage that we can basically see okay. Now uh let me drop this so that you'll be able to understand more points. The quality of the DB scan depends on the distance measure okay. Distance measure used. Right now most of the distance measure that we use is something called as Euclidean distance. But as we keep on changing the distance measure we use Manhattan distance. And there is also another distance measure uh, which is specifically called as region query okay okay. Region query is a function which takes the distance. So that is also different. Different distance measure like Manhattan distance. Um Euclidean distance. If we if we change that distance measuring parameter then we may have a different quality of clusters. We may have a different different kind of clusters. And because of that the quality of the DB scan may vary. Okay. DB scan cannot cluster data set with larger differences in densities. Now what does this basically mean. Let's say I have some data points and uh let's say over here I'm just going to draw it. Let's say uh, some of the data points over here are densely available. Let's say these are densely available over here. And some of the data sets may be. The density may be a little bit more. Right. So like this. Or let me take another example, uh, with respect to this. And some will be like may have they may have higher densities like this. So if we try to apply DB scan for this kind of datas, uh, it will not be able to cluster the data because why. Because we use a minimum threshold that is radius right. Epsilon. Right. So it will not be able to cluster the data which has varying densities okay. Varying densities. In this particular thing you will be able to see that there is a varying density. And this all the other points will be within epsilon. But in this some of the points may become um, you know, outliers. And some of the points may even become a border point. Okay. So when you have this kind of scenarios, DB scan will not be able to perform well if the data and scale are not well understood. Choosing a minimum distance threshold epsilon can be difficult. So if your data points, let's say if I have f1, f2, f3 feature, right? And we don't know the scale that is basically used to measure all these things at that point of time, your epsilon value will be a difficult task to select. So usually what we do in this case we standardize our data set right, even though we don't know the scale just to overcome this particular issue. Okay. So I hope I have covered most of the points with respect to this, and I hope you have got an idea about what are the major advantages and disadvantages with respect to, uh, DB scan. Yes. In the upcoming videos we are going to see some practical implementation. I'll see you all in the next video. Thank you. Bye.