So in this video we are going to discuss about a new machine learning algorithm, which is called as logistic regression. In our previous video already we have completed linear regression algorithm. Now let's go ahead and understand the in-depth maths intuition behind logistic regression. But remember this logistic regression is actually used for solving a binary classification problem. Uh, we'll try to understand the entire maths how logistic regression is basically works. And uh, you know we'll try to discuss it some use cases and we'll try to understand what is the cost function that is used and other things. So without wasting any time let's go ahead and let's try to solve this. Now first of all whenever I say binary classification problem, what does binary classification problem actually mean? Let's say I have a specific data set. Now in this particular data set, as I said already right. My output feature will be a categorical feature right over here. Specifically, we'll be having binary categories okay. The output feature or the dependent feature. Let's say I'm just going to consider with one feature as an example, so that, uh, I will make you understand with this, and then we'll try to understand the other examples also. So let's say in my feature one I have something called as study hours okay. And the output feature is that whether the person is going to pass or fail okay. So this two specific feature I have and let's say I have a data points like two hours. If the person studies for two hours, probably he's going to fail. Okay. Uh, if the person probably studies for three hours, he is going to fail. Let's say if the person is going to study for four hours, uh, you know, I'm just saying that, okay, the person may fail. Okay? You may be thinking, Krish, why you're writing, all the persons are actually failing. I'm just taking it as an example. Okay, so let's say if the person is studying for five hours, uh, for a particular exam, at that point of time, the person is actually passing similarly 67I have lot of data points like this. Okay. And this data points also will consider it as pass and pass. Okay. Now, now how can we basically create a machine learning algorithm which will first of all you know whenever they take this study of hours. So suppose if my machine learning algorithm takes study of hours as an input, it should be able to predict the output whether the person is actually passing or failing, and why we say this as a classification problem, because in our output we have two categories or binary categories. So that is the reason why we use logistic regression to solve a binary classification problem. Now let's let's think over this and probably let's plot this in one kind of diagram. Okay. Now uh, the first question that may arise, you know, Krish, why the name is logistic regression, even though it is a classification problem. Right. Why do we say regression? Okay. And if I consider this data set, can we solve this particular problem statement with the help of linear regression that we have learnt in the past? Okay. So let me do one thing. Let me take just just this as an example. So suppose if I start plotting okay, so let's say in my x axis I actually have study hours okay. And uh let's say with respect to my y axis I will be having two categories. Either this will be pass or this will be fail. Right. So this specific point will be fail I will indicate pass by one okay. I'm just assigning a label. Right. So let's say pass is one and fail is zero okay. And I have study hours. So let's say this is two for. 2345678. Right now, if I try to train this model, then let's say we are trying to solve this particular problem statement with the help of linear regression. So over here you can see for two hours the person is going to fail. So over here let me make a point over here. So for that if the person studies for two hours, the person is going to fail. If the person is going to study for three hours, four hours and five hours, the person is going to fail. And that is what my specific data set says. Let's say this five hours is fine because five hours, if the person studies for five five hours, the person is going to pass. So I'm just plotting this particular data points. When it is two hours the person is going to fail. That is zero. Uh, if the person is studying for three hours is going to fail, he or she is going to fail. And if the person is studying for four hours, he or she is going to fail. Similarly, when the person is actually studying for five hours, the person is going to pass. So let me just, uh, make some points like this so that, uh, this will indicate my one

line. So I'm just going to make this specific line, or let me put some dotted points so that it will indicate you. Okay. So when the person is actually studying for five hours I'm basically going to make a tick over here. Okay. That basically means the person is going to pass. Similarly, when the person is basically going to study for six hours, he's going to pass, uh, he or she is going to pass. And if the person is basically studying for seven hours, he or she is going to pass. Okay. Now, whenever I have this kind of data points, if I really want to solve this with linear regression, if I want to solve this with linear regression, then how can I solve understand the main aim of linear regression? We try to find out a best fit line right now with respect to this, I will try to create a best fit line. Let's say this is my best fit line okay, this is my best fit line. Now this specific best fit line, what does it indicate. Because obviously right. Whichever level whichever will be the minimum error. If we combine it will try to create the best fit line that we have already seen in linear regression. This midpoint is basically point five. Now with respect to this particular best fit line, we are trying to just understand can we solve this problem statement with the help of linear regression. First of all okay. So with the help of linear regression definitely we created this. Now what I'm saying is that anything anything right. Any with respect to any value. Uh, suppose uh with respect to this particular experience, what should be your output. So I will try to predict it like this. Right. So from this particular point, let's say this is my new test data. I will try to predict it over here. And this will in turn get predicted over here. And if I say that okay, if my value is less than or equal to point five, I'm going to say my output is zero. And with respect to any study hours, suppose that this particular study hours, if I try to plot it, I'm getting somewhere here. And this is above four uh, point five right. This is above point five. So anything that is greater than point five, we are going to consider it as one. So this is how I'm trying to solve this binary classification problem okay. So this is perfectly fine right. So definitely you can see over here with the help of linear regression. Also we are able to solve by putting this particular condition that with respect to any new data points with respect to any new data points, if it is coming less than point five, then we are going to basically consider it as zero. And with respect to any new data point, suppose if it is coming greater than point five, this is point five, right? This is point five, right. If it is coming greater than point five then I'm going to basically consider this as I'm going to basically consider this as uh, the person is going to pass. Okay. So this is what we have basically considered. Now what is the problem with this? Okay. Now you may be thinking this fine. This is fine. Right? We are able to solve this with the linear regression. Now why we have to actually learn logistic regression. Now let me say that I'm introducing one outlier. Let's say if the person is studying for somewhere around uh, 12 hours. So this will be the new point over here. So obviously when the person is studying for 12 hours, obviously he has to pass. But now let's say in my training data set I have added an outlier. So one outlier I've added if the person is studying for 12 hours. So he's also going to pass now because of this outlier. If I now again retrain my linear regression model, now, you will be seeing that my line will basically get slanted towards this. Right. So my best fit line will completely change because of the outlier right now. Just see this scenario. Now you can consider that if I try to plot five okay. And here you can see that I'm plotting from here. And it is coming less than point five. Right. So this basically means even though the person is studying for five hours, he is he or she is actually failing. But in our training data set, it shows clearly that if the person is studying for five hours, he or she is actually passing. So this is the biggest blunder, right? With the help of linear regression, because as soon as a new data point came because of the outlier, now you can see that I'm getting a new best fit line. So this is my new best fit line, right? Because of the new point best fit line right now here you can clearly understand what is actually happening. Is that because as an outlier, my entire line got changed. And now because of this, most of my data points, you know, that were initially, uh, with. Back to that many number of hours, which is giving one now it is

basically giving zero as an output. So this is the biggest blunder. So this is one of the issue with linear regression. The second issue is that now see let's consider this um this white line okay. Now with respect to the eight hours. Right. Obviously if I try to plot this over here, if I try to plot this over here, and with respect to this, if I try to plot it now, in this scenario, you can see that my output is coming greater than one, right? Greater than one. And similarly, let's consider one more point, which is somewhere here. Right. Let's consider this. And now if I try to plot it over here, and if I try to plot it over here, I'm getting a negative value. Right. So I'm also getting a value that is greater than one or and I'm also getting a value that is less than one. Sorry less than zero right. So this two scenarios is also there. But with respect to binary classification we should always think of getting the value between 0 to 1, right. This condition should always keep on uh matching right. This condition should be fulfilled if it if it is less than uh, less than or equal to 0.5, it should be zero. If it is greater than 0.5, it should be one. But here you can see that it is either going greater than one. Also it is going less than zero also. So we are also getting a negative. So this is again a problem with linear regression. So we should try to find out a way in wherein we should try to not extend this line from here. Instead we should try to squash this line from here. So if it squash this line from here, that basically means I'm not extending the best fit line. I'm also squashing it from here. And I'm also squashing it over here. So if I squash it from here and here, you will be able to see I will always get my answer between 0 to 1. And this squashing is only possible with the help of logistic regression. Okay. So I'm just going to make two important points. Why we cannot why we cannot use. Why we cannot use linear regression for classification okay. Linear regression for classification. So just pause the video and think over it okay. First is first is if I have an outlier then there is a biggest issue, right. Uh, entirely my best fit line changes, right? The best fit line completely changes. Best fit line changes. This is one of the issue, right. Let's talk about the second one. Apart from the outlier, what is the other issue that you saw? I'm actually getting greater than one. My output is also coming as greater than one. And it can also come less than zero. So what we should do, we should basically squash that line. And squashing that line is not possible by linear regression. Right. How squashing. You're saying like I'm restricted restricting the output that it should be between 0 to 1. So this is only possible through this squashing. So how the squashing will happen. We'll try to understand in logistic regression. But I think through this video you have understood why we cannot use linear regression to solve a classification problem. Right. So I hope you have understood this in the next video will try to understand okay, about logistic regression and will try to see how the squashing will also happen. So yes, I'll see you all in the next video. Thank you.

Now we'll try to understand how logistic regression solves a classification problem. Already from our previous video, we understood why linear regression cannot solve a classification problem over here because we have two issues. Because of the outliers. This line, the best fit line, may move abruptly here and there. Okay. And the second thing is that we were trying to get an output that is greater than one or less than zero. Okay. So this two conditions were there. Now let's go ahead and let's try to understand now as I told you, that, uh, if we really want to prevent from getting an output that is greater than one, we have to basically squash this result. Similarly, we need to squash the result over here. Right. So this two squashing needs to be done already. We know that this best fit line can be created with the equation f of x is equal to theta zero plus theta one multiplied by x one. Right. So this we have already learned in the linear

regression. Right. But and this this best this is nothing but it actually gives us the best fit line. So this yellow line that you basically see right. This is basically given for this particular equation. But we don't want this specific equation. Right. Because from here we really need to squash it and how the squashing is basically possible. That is what we are going to see. So guys today I'm going to introduce you to a new activation function. We also say this as an activation function which is called as sigmoid activation okay. Sigmoid activation or sigmoid activation. What is the so important thing about sigmoid activation is that this sigmoid activation is basically given by the equation one plus one, one divided by one plus e to the power of minus z. Okay, now whatever the z value is, you know, whatever the z values. And if I try to plot this sigmoid activation function it looks something like this. Okay. So always remember the sigmoid activation function. If you probably try to play with any any z value okay. The sigmoid activation function that you are probably going to get, and I will probably give this the notation as sigmoid of z. Right. So with respect to any z value, always remember your sigmoid outputs between 0 to 1 okay 0 to 1. So this is one condition. And the sigmoid activation will have a curve which will look something like this okay. It will look something like this. Let me just again draw it a little bit properly. So this is how a sigmoid activation function will look like. That basically means your value will always be between 0 to 1. And this mid one is basically your point five. That basically means whenever your z value is greater than zero. Understand from this, whenever your z value is greater than zero, then your sigmoid of z right is always greater than point five. Okay, so greater than or equal to point five. If I say greater than or equal to zero okay. So at any instance, whatever z value you select and you try it out, you probably go and Google over there and you just say that, draw me the curve of this specific equation which is called a sigmoid curve. Right. So this sigmoid curve will always give this and whatever z value, irrespective of any z value, will let it be a negative or positive z value. Your values will always be ranging between 0 to 1. Now because of this equation, what we are going to do on this best fit line, on this best fit line, we are just going to apply the sigmoid activation function. Okay, so that is what we are actually going to do in order to achieve this. Okay. So in order to achieve this now what will happen is that both the problems will actually get solved since we are squashing it over there. Right. So any outliers that also comes in the future over here will also be giving the output as one. Okay. So this is how we are going to do it. Now let's go ahead and write out our uh all the equations that is required okay. Now first step is that in logistic regression we also create a best fit line. And then on top of it we apply a sigmoid activation function. So let me go ahead and write our new function right. So I will write h theta of x. Now always remember when I am writing h theta of x I'm basically going to use a notation sigmoid because I'm going to apply a sigmoid activation function on this specific equation if I have just one independent feature right. If I have multiple independent features, then it becomes theta 1X1 plus theta 2X2 plus theta 3X3. But let me just show you with respect to this. And you will be able to understand that okay. Now what we are going to do over here is that we are just trying to apply this, uh, activation function, which is basically given in the sigmoid. And here this sigma, I'm going to denote it by one plus e to the power of minus z. Okay. And let's consider that this specific value is z. Now theta zero plus theta one into x one okay. So let's consider this value as z. So in short I can basically write sigma by z okay. Which in short I can basically write as t of x which is the uh usually when whenever we even in linear regression right. We took theta zero plus theta one multiplied by x. That is the equation. Now instead of x, the hypothesis that we are going to apply will be basically one to. The power of e to the power of minus z, where z is nothing but this specific value, which is nothing but theta zero plus theta one into x one, okay, x one or x anything. Right. So what we are doing, first of all a best fit line is getting created. On top of it we are applying a sigmoid activation function. So always our output will be between 0 to 1 right. So this is the entire hypothesis that we are going to use. Uh let me

uh again highlight it. So this is the entire hypothesis that we are going to use. So. I'll just mark it like this. So this is my logistic regression hypothesis. Okay. So this is my logistic regression hypothesis. Logistic regression hypothesis. Okay. And here you know that Z is nothing, but Z is theta zero plus theta one into x one. Okay. So guys, I hope you have understood, uh, the logistic regression hypothesis, uh, as I discussed. So this is basically the equation of the hypothesis. Now what I am actually going to do is that let's go ahead and consider the linear regression cost function. And then we'll try to modify the logistic regression cost function okay. So I am just going to write over here linear regression cost function I hope everybody remembers what linear regression cost function is. Okay, because we have spent a lot amount of time in understanding the theory of that. So this is basically J theta zero comma theta one is nothing but one by two m summation of I is equal to one to m okay. And this will be h theta of x of I okay. Minus y of I whole square. Right. So obviously you know what is theta of x is in the case of linear regression it is nothing but theta zero plus theta one into x. If you just have one single independent feature. Now similarly what I'm actually going to do, I'm going to write it for the logistic regression also. So logistic regression cost function since we are using regression somewhere. Right. And later on we are applying an activation function which is called as sigmoid. So logistic regression if I really want to write I can basically give the same notation over here. Also something like this. So this will be j theta zero comma theta one which is nothing but one by two m summation of I is equal to one to m. Here. Also, we'll try to use the same notation just a second. But what is the thing that is basically changing. It is nothing but your h of x. So here I can definitely write my h. Theta of x is nothing but one plus e to the power of minus z, and z is basically given by theta zero plus theta one into x one. Okay, so this is how we have probably learned it from here. Right. Because z is nothing but theta zero plus theta one into x one. So now if I use this notation understand what will happen if I directly use this cost function. And inside this cost function what is x of x. It is nothing but this sigmoid activation function, right? In short this is basically sigmoid activation function. Now can we get the gradient descent curve with the help of this particular cost function? The answer is no guys. Why? Because this cost function. If I try to plot it with respect to theta and cost, this is going to give me a non-convex function okay. Non convex function. And this equation that we probably saw this will give me a convex function. Now the gradient descent. Now for this what what is the gradient descent that is usually created when you try to create a plot with respect to theta and j of theta, you will be seeing you will be getting this kind of gradient descent right. And over here this will basically be your global minima. Right. Global minima. So this is specifically your convex function. So I really want to call this as my convex function. So gradient descent is actually a convex function right. Gradient descent is actually a convex function. Now if I try to use this equation that is there and try to plot it, then what kind of curve will I get. So if I try to plot it over here with respect to theta and j of theta here, you'll be seeing that we will be getting some this kind of curve, something like this. Okay. And what is the problem with respect to this particular curve. It is first of all it is non convex okay. So it is non convex. Whenever it is non convex. That basically means obviously this is my global minima okay. But here you have some more points. And this points will basically create something called as local minima. Now what is the importance. What is what is the disadvantage with respect to local minima. Because whenever I come at this particular point and if I try to find the slope, the slope will always be zero. And if the slope is always zero, you know the theta will be remaining stagnant at this place. It won't be able to increase or reduce right to come to the global minima. So it will be very, very difficult to come over here. So this is basically my global minima. So if I probably use this equation right. This equation uh, and if I try to create my um relation between theta and j theta I'm going to get this kind of non convex function. So it is and it is proven also guys uh I don't want to do the proof entirely and probably

just try to, you know plot with respect to each and every point. But if we try to use this we are going to get a non convex function. Now in order to prevent this what we can do is that, uh, let me again write the cost function of logistic regression. So let's say this is my cost function here I'm basically writing one by two m summation of I is equal to one to m. And here I am writing h of x of I minus y of I, minus y of I whole square. Right. And uh, obviously everybody knows what is uh, h of x. It is nothing. But uh, h of x of I is nothing. But one plus e to the power of minus z, where z is nothing but theta zero plus theta one into x one. Right. So this we are basically discussing from that much time. Right. This is very very much simple. Now let's consider this is my cost function. Now instead of using this as my cost function let's consider this I'm going to indicate this as cost state of x of I comma y of I. Let's say I'm denoting it. Let's denote let's denote. This value as this. Okay, so what we can do in order to get a convex function, we can basically write cost of a state of x of I comma y of I. I can use this equation. And yes, we can also prove this. But I'm just going to write down this equation over here because the proof is not that important. If I say minus log of x theta of x okay. And if my y value is one okay. So if my y value in this particular case, if my y is one, I'm going to use this as my cost function which is basically going to get replaced over here okay. And one more condition is that if my y is equal to zero, I may probably use something called as minus log one minus h theta of x and y. You are using this because this two specific thing, this two log, uh we basically say this as log loss. Okay log loss. And if you are using this log loss, this will be able to create a convex function that is basically required in terms to get the global minima. Okay. So we are going to use this two conditions over here. Right over here when my y value is one I'm going to use this equation. When my y value is equal to I'm going to use this equation uh and this two equations when we are using with respect to different different y value we basically get a convex function okay. So that is what uh we get it mathematically. Again I don't want to prove this as a formula. You just need to remember it. So in short if I try to combine this my cost this this specific value can be replaced by something like this. So I will go and replace it something like this where I'll write cost. Sorry cost not cost. It is cost. Did I write cost. Uh cost. Cost. Okay, fine. Uh, if I try to write cost, it'll be cosine. Right. So here I'm going to write cos h theta of x of I comma y of I. If I try to use this log loss we can definitely create a equation which is like y log of h theta of x of I uh. Y log of h theta of x. Just a second. Okay. So I'm just going to write minus y. Log off. H theta of x minus one, minus y of one minus h theta of x. Now here you can definitely see that if my y value is one, if I replace y value as one, I'm going to get this specific equation okay. If I replace y value with zero then what will happen. This will entirely become zero. This into this will become zero. So I am actually going to uh, get something called as C one -0. This will be one. So I'm actually going to get uh okay. I missed one thing over here. Just a second. This equation will be changing a little bit to um, after this minus I'll be using one minus y log of. One minus h theta of x. Okay. So in short, if I replace y with one, if I replace y with one, then I'm going to remain with only this specific equation because one minus one will be entirely zero. This entire term will be zero. If I replace y with zero at that point of time, this will become entirely zero, and I'll be, uh, you know, remaining with this specific equation. So with the help of this, what we will be able to do is that we'll be able to get a cool convex function as a gradient descent. Okay. So that is the reason why we specifically use it. Now what I'm actually going to do this entire cost function that I've actually written over here, I'm going to replace it with this specific value. So finally my final conclusion with respect to J theta theta zero comma theta one is nothing but minus one by two m okay. Summation of I is equal to one to m. Why I'm taking minus because here both are minus right. So I'm going to take out this minus outside. So finally I'll be able to write y of I okay log. H theta of x of I minus one. Minus y of I. Log. Log. One minus h theta of x of I. So this will be my entire cost function. In short okay so this will be my cost function. Now what what do we do with

the help of cost function. And always remember this cost function will actually give us a convex function okay. This cost function will be actually giving us a convex function. Or I can also probably call it as gradient descent. Okay. So this is my cost function. Again the lines are not uh getting created properly. Let me try it out. Once again sorry for the inconvenience. So here is my cost function. Now our major aim is always to minimize. Cost function by changing value. So here my cost function j theta zero comma theta one by changing. By changing theta zero and theta one, right? So in short, if you remember again we have to write the convergence algorithm. And then this convergence algorithm we repeat. So here I'm just going to say convergence algorithm. We are going to just repeat. One important step. That is nothing. But we are just going to change theta j value, which is nothing but theta j minus learning rate. Derivative of derivative of theta j with respect to theta with respect to j theta zero comma theta one. So in short we are going to find out the slope to find out the best fit line. And here your j value is zero and one. And obviously that derivative part, whatever we have written in the simple linear regression that same thing will be writing it over here. Right. So this is what we have done. You need to really understand why we did not use this specific cost function, because this was giving us a non convex function. In order to replace this non convex function we use something called as log loss. And this is basically my log loss wherein I'm using two specific thing with respect to this cost theta x of x I minus y of I. Whenever y is equal to one I'm going to use this. If y is equal to zero I'm going to use this okay. So instead of x is nothing. But this is where I'll do the predictions right. Predictions with respect to any value. So finally I write my cost theta of x as this specific equation. This will definitely create a convex function. Convex function or gradient descent. So finally we will be able to see that I will be writing my cost function as j theta zero comma one. And this is the entire equation. And our main aim is to minimize the cost function by changing theta zero and theta one, and will try to repeat the convergence algorithm where will keep on updating theta j value, and where j is equal to zero and one. And similarly, if I have multiple features, uh, the j will keep on increasing like 0123456. So I hope you have got an idea about the entire mathematical intuition and how we are using logistic regression to solve the classification problem, right? So yes, in the upcoming videos we are also going to discuss about the performance metrics. Uh, this was it from my side. I'll see you in the next video. Bye bye.

So guys, in this video we are going to discuss about performance metrics which are specifically used in binary classification, multi-class classification. We are going to cover topics like confusion, matrix accuracy, precision recall and F beta score. Now guys, let's say I have created a model using logistic regression. So I'm just going to run this. So here let's say uh I'm using a logistic regression. And in this logistic regression you know we are able to classify what what what is the main aim in logistic regression. This is basically a classification problem. And we try to split this by creating a linear line okay. And already we have discussed about the cost function and all. Now suppose if I probably get a new point over here, this basically means it belongs to this category. Or suppose if I get a new point over here, it basically means it belongs to this

particular category because it is below the line. Okay, now if I really want to understand how my model is actually performing. So in linear regression we discussed about something called as r squared right or r squared and adjusted r square. Now this were the performance metrics that were used in the case of regression problem statement. Right now in the case of classification what all things we can actually use. So we use specifically called as confusion matrix accuracy precision recall and f beta score. And in this video we'll understand how through this particular topics we can understand whether our model is performing well or not. Okay. Now let's consider this simple data set I had. Okay. Now in this particular data set I have all this data points X1X2. And this is my output feature. This is my output feature. This is my input feature right. And this is my predicted feature. Predicted feature basically means uh this is basically predicted by model okay. This y hat okay. So this basically shows that when my true output was zero, my predicted model predicted it as one. And similarly over here, when my true output was one my model predicted as one, here it was 0011110110. Okay, so this is a simple data set. And I probably trained my model. And I'm getting this kind of predictions. Now the first thing that you really need to understand is something called as confusion matrix okay. Whenever you get this kind of outputs okay. Basically once you find out a prediction the first thing is that you create a confusion matrix. Now what is a confusion matrix? Confusion matrix in the case of binary classification is a two cross two matrix okay. So I'm just going to draw this two cross two matrix on the top. I'm basically going to draw uh I'm just going to write 1010 basically means uh this is with respect to the output. Uh, you know ones and zeros are my output. Similarly on the left hand side we are going to write one zero. Now you really need to understand what the top one zero basically indicate. This indicates my actual values okay. Everything will make sense guys. Just focus on this actual values and what this one zero indicates. It indicates something called as predicted values. Okay. Now now let's try to. Now suppose if I have this real output and the predicted output, how can I feel the confusion matrix with this data? Okay, now let's say my real output was zero but my model predicted it as one. So over here how do I map it over here. So my real output is zero. My model predicted it as one because I told you that this is the predicted part, right? So if it is zero and my model is predicted one, that basically means I've got this particular scenario. So I'm going to increase the count by one. So here let me write it down as one okay. So I've got this particular count. And this belongs to this particular box okay. Similarly in the next field you can see when my model uh when my real output is one, my model predicted it as one. So this is fine, right. This is incorrect. This is fine. Right. Our prediction should also happen like this. So where does this one one actually fall. So one one will actually fall over here. Right. So we have something over here and I'm basically writing it okay. Perfect. Then I have zero zero. So I have zero zero. It will be coming over here. So I'm going to increase the count by one over here. Then I have one one again one one is there. So what I'm actually going to do I'm going to rub this and I'm going to make this as two I'm going to increase the count. Then again I have one one. Then I'm going to increase the count by one more then uh 123. Then again I have 0101 basically means this is zero and this is one. So I'm going to increase the count to two okay. So let's say before it was one. Now I'm going to make it as two perfect. And then I have one zero. One zero is nothing but this specific field. So I'm going to write over here. Now over here. You know that if I consider one one value, this basically means my model has predicted correctly. Right. Because this scenario, this scenario is that my model have done the prediction correctly because my real output was one and my prediction was also one. And in this particular scenario, when my actual value was zero, my model is predicting zero. So this is also correct right. So out of this all results this diagonal elements are the actual correct. Results. Right. And this diagonal elements are actually the wrong results because when it was zero my model is predicting one and when it was one my model is predicting zero. So these are my wrong results.

Now this confusion matrix I'm again going to draw over here. And I'm going to give you some notations over here okay. So let me draw it over here and let me provide you some notations. Now notation one is over here I'm using 1010 whenever I have in the actual as one and predicted as one. So we usually say this is a correct scenario. So I'll be writing it as true positive when the actual value is zero. And probably my model has predicted zero. I'm going to basically write this as true negative, and when my actual output is zero, but my predicted is one that basically means this is false positive okay. And similarly if I if my actual output is one and if my predicted is zero, I'm going to basically call it as false negative. Now this is super, super important in respect to deriving all this formula of accuracy, precision recall and f beta score. Okay. How we're going to discuss about it. Now out of this, if I talk about true positive and true negative, these are my correct results. And remaining F false positive and false negative. These are my wrong results. So if I really want to calculate my accuracy. So this formula will be true positive plus true negative divided by true positive plus false positive plus false negative plus true negative. Right. So this is the specific formula with respect to this. Now what is true positive over here. From this we can actually see this particular field. We'll just try to populate it. So it is nothing but three plus one because one is over here, three is over here. And then I'm going to add with three plus what is false positive. Obviously it is two. Then false negative is one and this is one. So finally I'll be able to see that I'm getting four by seven. So this whatever value I'm getting is basically my accuracy for this particular model output with respect to this. Right. For with respect to this particular confusion matrix. So guys now we have understood about accuracy okay. Now let's go ahead and understand about the next performance metric which is called as precision and recall. And we'll try to understand when should we use precision and when should we use recall. Now let's consider that I have a data set okay. So I'm just going to consider let's say I have a data set. Now in this particular data set let's say that I have 1000 data points okay. So my entire training data set has basically 1000 data points. That basically means 1000 rows. And from this I have 900 output categories as one. So let's say this is a binary classification. Binary classification. Okay, now, one important thing that I really want to talk about is that suppose if we have a multi classification, then the confusion matrix will not become two cross two. Let's say if I have the number of output categories is three, then this will basically become three cross three matrix. If the number of if the output categories the number of output categories are four, then this becomes a four cross four confusion matrix. Okay. So let's consider that in my binary classification I have 900 as one and I have 100 as let's say 100 number of categories are basically saying it as zero okay. So in this particular scenario you can see the ratio of the output categories with respect to ones and zeros are nine is to one right. So this basically becomes an imbalanced data set. Right. Now in the case of imbalanced data set I cannot directly use accuracy. See I cannot directly use accuracy accuracy. If I directly try to use accuracy. Let's say if I just create a blunt model, which is just saying that for all the input data points, I'm going to get one. Then if you try to calculate right will be getting 90% accuracy. But in the real it is not true, right? In the imbalanced data set, it can give you a completely biased result. So in order to fix this, we will try to use something called as precision and recall. Now we'll try to understand what exactly is precision recall and what is the exact formula with respect to precision recall. Okay. So first of all let's go ahead and let's talk about the formula with respect to precision okay. So first topic that I am going to cover over here is precision. Now I hope you have understood when we should not use accuracy. Let's say if you have a imbalanced data set at that point of time, you know, I cannot use it because I'm getting this high accuracy for that, even though my model is bluntly saying one to every data points as an output. So precision is basically given by the formula true positive divided by true positive plus false positive. Now over here the most priority is basically given on the false positive. Now let me make you

understand what this formula basically says. So I'm just going to draw the confusion matrix again okay. And let's say this is my ones and zeros of the actual point. And this is ones and zeros of the predicted point. So this is my true positive. This is my false positive. This is my false negative. This is my true negative. Now this precision basically says that out of all the see when I say uh, with respect to true positive and false positive. Right. So we are basically focusing on this two. Right. So what what what does this particular formula say. Right. And we know that this is basically predicted and this is my actual result. Right. So with respect to this I can definitely say one very important statement. Okay. From this particular formula, out of all the predicted result, out of all the predicted result, how many of them are correctly predicted? Okay, this formula basically says this, right? Again, let me write it down. Out of all the all the predicted results. Because if we are referring to this, right, or uh, if I am just referring to this right, I can definitely say, okay, let me just make a little bit change over here again. Okay. Let me make a simple change because again, I miss, uh, you know, it always there is always a confusion with respect to precision recall. So let me just make this statement again, okay? Out of all the actual values, okay. Because when I'm referring to this two. Right. So in short I'm referring to both the actual values ones and zeros. Okay. Out of all the actual values, how many are correctly predicted? That is what we are basically going to find out with the help of prediction precision. Okay, so out of all the actual values, how many are correctly predicted. That is what this precision basically talks about. Okay. So again I'm going to repeat it out of all the correct results or actual values. Okay. Uh, out of all the actual values or true values, we are going to find out that how many of them are correctly predicted. That is the reason why we are dividing true positive by true positive plus false positive. Okay. So this is the example. And again I'll be talking about more examples as we go ahead. When should we use precision and when should we use recall. Now the second thing if I talk about the recall formula okay, my recall formula is something like true positive divided by true positive plus false negative. So now in the case of recall I am considering this specific part. That is true positive and false negative. Now you you can basically guess and I can basically write out of all the predicted out of all the predicted values, how many are correctly predicted, right, correctly predicted. This is what we are trying to. Find out with the help of recall. And always remember with respect to precision and recall, we should try to reduce this false positive and we should try to reduce the false negative. Now you may be thinking, okay Chris, this is the formula. Fine. But tell me, when should we use precision and when should we use recall? Okay, now let me give you some of the use case okay. Let's say my first use case is I want to create a model which will basically do spam classification. Okay. Spam classification. Now, in spam classification, let's say my model. If my truth value is right, if I get a mail and this is spam, right? This is spam. And if my model predicts this is a spam, then this is quite amazing, right? So this is the true positive result. So in short my model is absolutely working fine. So this is good right? This is how my model should perform. But let's say if my mail is not spam. And if my model predicts it's at spam, then what will happen? This is a huge blunder. Okay, blunder. Because even though my mail was not a spam, my model is predicting it to the spam and it is sending to the spam folder. So by that way, I may miss my more precious email, right? So in this particular case, whether we should reduce false positive or false negative. Now you need to understand over here obviously and obviously this all knowledge comes from domain experts also. But you should start thinking whether I should use precision or recall or something else. Okay. So let's say this is one zero. This is one zero. Now in this particular case if I say one one basically means spam, zero basically means not spam, right. So let's say this is spam and this is not spam. This is the predicted output. Now in this particular scenario when my mail actual value is here, right? When my mail is not spam, not spam basically means here and it is model is predicting as spam. So this is the error. It is

basically gone, right? It is basically coming up with false positive error. Right. This is obviously true negative. This is false negative. But in this particular case this is a blunder right. So in this use case I should try to reduce false positive because we understand that this kind of situation is a major blunder okay. Major major blunder. Because when the mail is not a spam, my model is predicting it as a spam. So when my mail is not a spam zero, my model is predicting as one. So this is basically false positive. So we should try to reduce false positive over here. And for this particular case here I will be using precision because precision formula is nothing but true positive divided by true positive plus false positive. Right. So this is the formula of precision. Now you may be thinking okay this is my use case one. You may be thinking Chris, then where can you tell us some use case which actually deals with false negative? Okay. Let's go with the use case two okay. Use case two. Let's say over here my problem statement is to detect. To detect or to predict. To predict. Whether. Person. Has cancer. I'll not say cancer. That is a very big disease. So whether the person has diabetes or not okay. Let's consider this particular problem statement. Now in this scenario, let's say if the in truth right, let's say in uh, if I probably take as an example, let's say. In the truth. The person has a diabetes. Okay. And my model predicts the person does not have a disability. And my model predicts the person doesn't have a diabetes. Then this is also a very big blunder, right? This is also a very big blunder because if the person is having diabetes, my model should be able to predict that the person should have diabetes. Let's say in truth, if the person is having diabetes and my model predicts it has diabetes, then this is good, right? My model is performing absolutely fine. Similarly, if I say in truth, let's say my model is not having diabetes, sorry. In truth, my the person is not suffering from diabetes, but my model predicts that the person is having diabetes. So this is also wrongly predicted. But can I say this is a blunder scenario, right? Even though the person is not having diabetes, but the model predicts as diabetes. But it's okay, right? The person will go and probably again recheck or take second opinion in the hospital. Right. So this this kind of error is fine. It can happen. Not a worries with that. But we need to not focus much on this because the person can go with respect to second opinion and they can actually check okay. They can actually check in the hospital by doing various tests. Right. So in this particular scenario we will not focus on this, but this scenario will try to focus. Now in this scenario should we focus more on false positive and false negative. Let's say I am just again creating my uh confusion matrix. And here I'll say person has diabetes and person has no diabetes. Okay. And this is with respect to the actual value. And here also person has diabetes and person has no diabetes. Now in this particular scenario the truth is that the person has diabetes okay. But the model is predicting the person does not have diabetes. So what in this particular case what is happening. We are getting false negative as a blunder. Right. So this is the most important error in this use case. In the use case of medical right. The use case of diseases. And this gets applied to every disease case. Right. So we should try to reduce false negative because if the person has diabetes and if my model is saying it does not have diabetes, the person does not have diabetes, then the person may also not go even for checkup. And he may miss that. And because of that, you know, the diabetes may be, uh, dangerous disease going ahead. Right. And if we try to treat it in the early stages, some type of fixtures, some, uh, will be able to prevent it, right? In the initial days if you are able to capture that. Right. So in this particular case I should try to reduce false negative. So in this case I'll be using a recall formula. Because in my recall the formula is true positive divided by true positive plus false negative. So in this case we will focus more on false positive to reduce it. Right. So in this use case our main focus should be that we should try to reduce false negative. Right. So now I have told you with with respect to different different use cases when you should use precision and when you should use recall. Okay. Uh, let me tell you one more, uh, simple problem statement. Okay. Uh, and then you try to do it as an assignment,

okay. And probably write down in the comment section, what if I really want to like my model wants to predict tomorrow the stock market will crash on it. Stock market will crash or not. Crash or not. Now suppose if this is my use case, what do you think we should focus on? Reducing false positive or reducing false negative? Now this should be the answer coming from you. Just try to plot that again confusion matrix and try to understand and get it okay. So you can definitely try to do this and uh let me know the answer. This will be quite amazing if you are able to talk about it. Okay. Now let's go. With respect to the one more amazing performance metrics, which we basically say as f beta score. Nine F beta score. The formula that is basically given is one plus beta square precision. Multiplied by recall, divided by precision. Plus the call. Okay. So this is the formula that we basically use for this okay. Now let's say if. Because what will be this beta value? We will be selecting it if false positive and false negative. Because in false some of the use case, both false positive and false negative can also be important. If false positive and false negative are both important and we really need to reduce both of them, then my beta value is basically selected as one. So here now if I talk about this this will basically become F1 score. F1 score is equal to one plus one square and this will be precision multiplied by recall. Divided by precision. Plus recall. So, in short, if I probably consider this, this value will become two multiplied by this one. And this term is basically called as harmonic mean. Okay. Harmonic mean because this is nothing but the harmonic mean. You I think you have heard of geometric mean harmonic mean something like that okay. So this formula basically says about F1 score. Now this is my first condition. My second condition is that let's say if. FP is more important. That is false. Positive is more important than false negative. Then false negative. Then what we do in this particular scenario is that we try to reduce our beta value. So now we will be selecting our beta value is 0.5. So this f 0.5 score will be in this particular scenario what FP is more important than f n. So here I will be getting one. Plus the formula is beta square which is nothing but 0.25 okay. And this will be precision multiplied by recall divided by precision plus recall. So this is the formula with respect to this. The final condition is that if. False negative is more important than false positive. At that point of time, my beta value will become two. So f two score is nothing but one plus four precision multiplied by recall precision plus recall. So this is my. Final output with respect to all the things. So if you consider a false positive is negative. Important if false negative is important. If both are important, you can also use f beta score. Otherwise you can also go with precision and recall. So I hope you have understood about all the performance metrics that we have discussed. This is just the part one. In the part two we are going to discuss about ROC curve. So yes, this was it. And in the next video we'll be discussing about that specific thing. Thank you I'll see you all in the next video.

In this video, we are going to discuss about logistic regression for multi-class classification. And uh, in our previous video we already seen for binary classification. Right. And in binary classification, what we have done is then what was the main aim using logistic regression was that let's say if I have some data points like this of binary classes, let's say I may consider two classes like this okay. Our main aim in logistic regression was to basically create a best fit line

such that it should be able to divide the specific points. But what if in our data set, I have three output categories or more than two output categories? Now in this particular case this is my class A and this is my class B right. This is my class B. Now let's say that in one of the use case I have more than two output categories. Let's say in this case let's let me draw it again. Suppose over here, let's say this is my independent feature x and y and I have three specific categories. My three output categories which looks like this. Or more than two. It can be more than two okay. But we will take an example with respect to three okay. Just to understand multi-class classification. So let's say if we have this three categories as my output then how we can solve this multi-class classification problem. Multi-Class classification problem. Multi-Class classification problem with the help of logistic regression. Okay. So that is what we are going to discuss in this particular video. With the help of. Logistic regression. So one important parameter in logistic regression is something called as one versus rest. Okay. And in order to solve. This classification problem for multiclass, we can use this version of logistic regression and we'll try to understand how this. Technique will be working if we use one versus rest. Now the main aim of one versus rest and let let let's consider some examples. Let's say if I have some f1, f2, f3 feature and this is my output feature, let's say I have some data points like this. And this is my output. One more data points. Let's say this is my output. Two more data points like this. Let's say this is my output three. Again I can have output one getting repeated. Let's say this is output three. Let's say this is output two. Okay. So let's say these are my input features. And here you can see in the output feature I have three output categories. That is 010203. So that basically means I have three categories right. Output categories. And let's say these are my data points. And it looks basically like this now with the help of one versus Rest. What we do is that we try to create, you know internally multiple models. And each and every model will act as a binary classification. It will try to solve a binary classification problem. So how does this basically work? Let's say I'm trying to create a model M one okay. Okay. And this in model M1, you know, just imagine that this is some internal model. And this what it will do is that it will try to create a best fit line in such a way that it will make sure that all these points that are below this best fit line will be one category, and all the other points that are above this best fit line. This will be another category, right? So here both these categories can be combined into one category. So this model will be able to differentiate this class and this specific both the classes okay. So here what we are doing is that internally we created one M1 model. And this will basically perform binary classification. And the binary classification will be performed in such a way that we will be having two set of categories. This is not going to get considered as two separate categories, but this both categories will be combined and this will be one data set and this will be the other data set. Okay. So what about the next model. Like since we have three categories three models will get created. So right now we have will get again another M two model will get created. Now this model will make sure that it will try to create this as a best fit line, where it is going to consider this as one category. And combined both this as another category. So this will basically be my M2 model and it will solve this binary classification. Okay. And similarly, internally we'll have one more model which is like M3. And here what it is going to do in this M3 model, it will try to create another best fit line which will look like this, which will consider this as one category and this both combined as the another category. So here again a binary classification will get created. Okay, so at the end of the day, what we are going to do is that we are going to combine both all of this outputs, and then we are going to decide in which block it basically belongs. Okay. So that is the entire agenda of this one versus rest. Okay. So to start with what it will do is that, uh, all this 010203 you know, it is just going to do something called as one hot encoding. So here I will have 010203 and for every model how it is going to take the input and the output, we are going to discuss. So whenever oh one is present this will

basically become 100. So let me just write this as wherever one is present, this will become one remaining. All will become zero. Wherever oh two is present, this will become one and remaining all will be zero. Wherever oh three is present, the last one will become one, and remaining all will be zero. Similarly, over here you can see that we'll be having again 100, uh, 001 and 010. Now, at the first instance, if I probably create my M one. Right. So let's say our main model is this. Internally we'll be having m1, m2, m3, y three models because I specifically have three outputs. And finally, uh, the averaging of all these things will give me the model output model of this classification. Now in the M one, what will happen is that we will take all this as my independent features, and the output feature will be basically zero one. Right. So what will happen in M one? We will be taking input features which will be f1, f2, f3 and the output feature will be. Or one right or one basically means this, this entire column. Okay. Similarly, in M2 my input feature will be f1, f2, f3, but the output feature will be oh two. Okay. So the O2 will be the output feature for model two. And similarly here we'll be taking the input as f1 f2 f3. And finally you'll be able to see that my output will basically be my O3 model. Okay. Now let's understand how the prediction for a new test data is basically done by using this specific concept. Now as soon as I pass my new test data, you know. Now first of all, this gets passed to M1 model. Let's say M1 model is giving you some probability okay. So obviously M1 model over here is the output is oh one. Right. So based on this let's say it is giving some probability. Let's say over here I'm getting 0.25. Now. Similarly when this test data is again passed to M2 model, the same M2 model let's say over here I'm going to get 0.20 as the output okay. So this is the first output. This is the second output. And then finally when this test data is again given back to M3 model and let's say the remaining um thing that we are going to get over here is 0.55. So this is the probability that I'm going to get. Now here you can see the combination is 0.250.20.55. Now what does this basically indicate here. This is the output for one uh model. This is my or I can basically write m1 m2 like this. Right. So this is my M1 model. This is my M2 model. This is my M3 model. Now you know that the probability of M3 that is going to come up is somewhere around 0.55. Right. So whenever this value is more, this value is more right from and when compared to M1 and M2, obviously .55 is more. So this M3 model, you have to see for what output it has got trained, it has got trained for oh three. Then definitely we can say that since the probability is high, this 0.55 will basically give us the output of 0303 basically means output three, which is nothing, but this is nothing. But this is basically my category three. Okay, so this is the entire understanding. Whenever we get a new test data, all the three models will give you different different probability, whichever will be the highest probability. Then we can understand that for which model is basically giving that specific probability and what this model has taken as the output feature while training. Right. So here you can definitely see oh three has been taken by the M3. So that is the reason you can basically say that for this new test data the output is category three. Now what we are going to do is that we are going to see some practical application with both one versus one and one versus rest. And here you'll be able to understand most of the things. So I will see you all in the next video. Thank.