Learning Analytics Tool

Professor. Ramkumar Rajendran

Department of Educational Technology

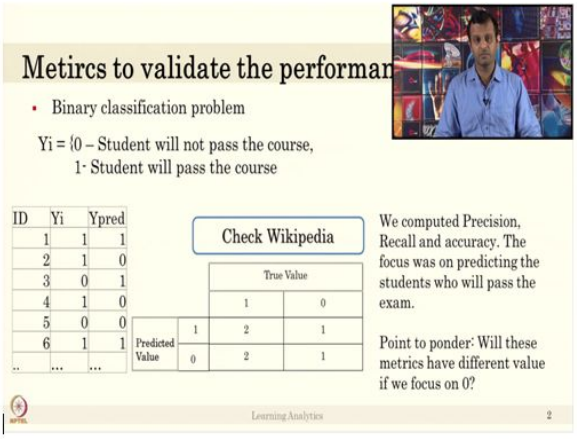
Indian Institute of Technology, Bombay

Lecture No. 3.5

Performance Metrics - II

In this course, we will talk about Performance Metrics, detail second part of Performance Metrics. So, we saw Performance will take like Recall, Precision and Accuracy in our last video. Let us look at more detailed Performance Metrics which compares different classifiers.

(Refer Slide Time: 00:34)

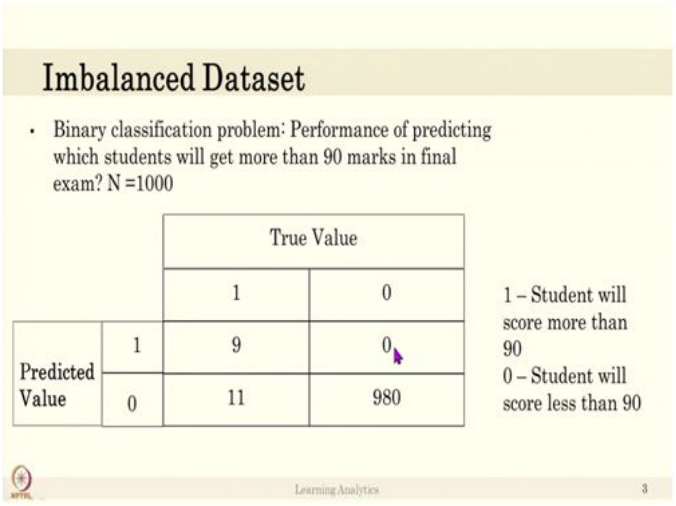


Consider there is a binary classification problem and the classes are 0 if the student did not pass the course, 1 if the student will pass the course. There are only two classes. So, it is a binary classification problem. And you have a table like this. y_i is the actual value and $y_{predict}$ is the output of the classifier. And we computed the Confusion Matrix. We have computed/calculated precision, recall, and accuracy. The focus was on predicting students who will pass the exam. So, I want to tell you that so the focus was on the student's who will pass the exam. If I have a focus

on 0, that is who will not pass the exam. What will be the change? I am not going to discuss what will happen if you focus on 0 in this video, but this is for you to think. Take it as extra work.

So, please check Wikipedia. There are very good resources. How to check Wikipedia find this particular page, just type Precision, Recall, Accuracy, Wikipedia and Google you will get the page.

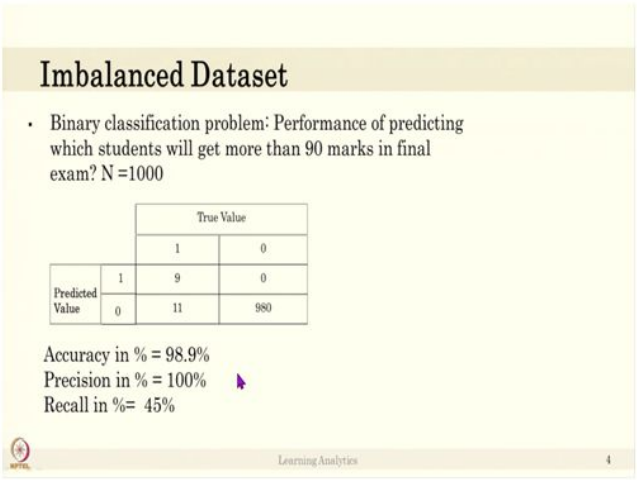
(Refer Slide Time: 01:49)



Let us check the next problem and classification. Consider your binary classification problem. That is two classes, the performance of predicting which students will get more than 90 marks final exam? And which students will not get more than 90 marks in the final exam? And you have N equal to 1000 samples that are thousand students data from historical semesters and courses.

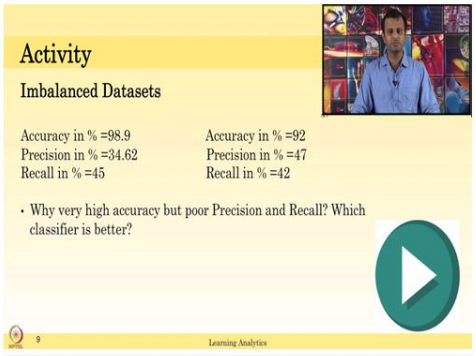
As you know, very few number of students will get more than 90 marks, they will get University ranks or something like that. So, the number will be very less so consider there are only 20 students who got more than 90 and the other 980 students got less than 90 marks. This data set is imbalanced. For example, in a true value, there are only 20 positive classes and 980 negative class that is scoring much more than 90 marks. So if we have this kind of imbalanced data set, what will happen?

(Refer Slide Time: 02:49)



So Let us compute Precision, Accuracy and Recall. Accuracy is 980 correct plus 9 correct, 989 by thousand 98.9 percentage is a very high accuracy you can get and Precision is it predicted all the 9, 9 divided by 9 plus 0 is 100 percentage precisely predicting, recall is 9 by 11. So, what do you think about this? Is it good? This value is interesting?

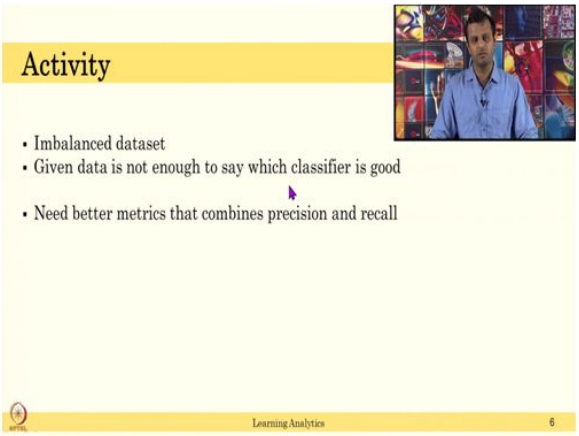
(Refer Slide Time: 03:23)



Let us see. Consider you have two classifiers, this also is not same as what we discussed in the last slide. This is a bit different consider the two classifiers which use the same 1000 data set and which gives the results like this Accuracy in percentage that 98.9 percentage and Precession 34 and Recall 45 percentage. Classifier here you might have used Decision tree or here you would have used some Naive Bayes classifier. Classifier 1, Classifier 2, results on the same data set is given.

Why it has very high accuracy but very poor Recall infection rate? And which classifier is better? Please pause the video Think about it. Take a minute Think about it, why these two classifiers gave very good accuracy and very less poor operation and Recall. Think about it and which classifier is better after you this done your answer you can resume continuing.

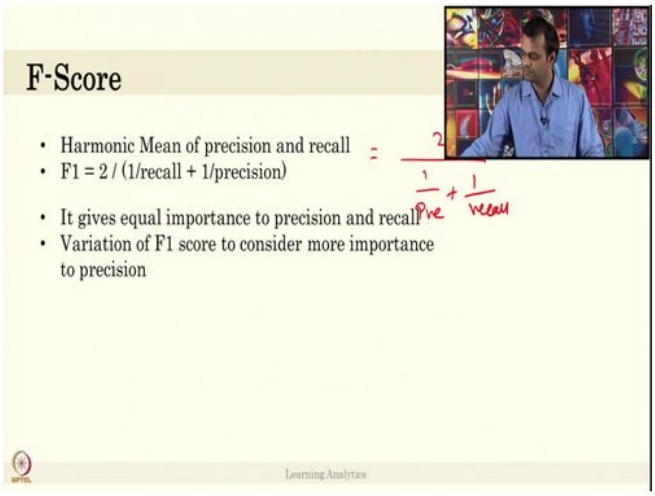
(Refer Slide Time: 03:23)



So, the reason for very high accuracy is the imbalanced data set. Given that the data is 980 is positive or negative. So, if a system which classifies everything as negative still will get 98 percentage Accuracy. No need to even try to create a new logic, the simple rule can classify all the classes into majority class. So, you get higher accuracy if it is imbalanced data set.

And which classifier is better? Given the data set, it is not enough to say which classifier is good because it depends on the research goal. So in order to make a decision, which classifier is better or which performance is doing good, we need a score, we need a metric which combines Precision and Recall or some other kind of metrics. Let us look at those metrics in this video.

(Refer Slide Time: 05:21)



One of those metrics is F score or F1 score, it actually the Harmonic Mean of precision and recall. what is Harmonic Mean? Harmonic Mean is simple. It is one kind of averaging technique. So Harmonic mean is simply it is 2 divided by 1 by precision plus 1 by the recall

$$H = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}}$$

So, it gives importance to both position and recall. So, is it good? Should we give importance to both precision and recall? In the last slide, I mentioned there are some research questions which will need better precision compared to recall, some research questions which need better recall than precision. Can you guess, Can you think of one such problem? This is not an activity but you can pause it and think about that. So, we will talk about that such questions later, but please think about it.

In order to avoid this challenging situation, the F score is giving equal importance to both precision and recall. What we can do is we can have a variation of F1 score computation methods that is it gives more importance to the precision or recall by adding weight to it. So, there is a variation of F1 score that can be considered if you give weights to precision but we are not discussing that in this video, you can check the Wikipedia page on what is formulations to do that.

(Refer Slide Time: 07:07)

So there is another metric, which is developed by Jacob Cohen's is called Kappa. Kappa is developed to measure the Inter-rater agreement of two raters. What is Inter-rater agreement of two raters? Let us take an example.

So, there are two raters, two researchers all watching the students facial expressions. There are say 10 students in the class or taking a class/attending a class. Raters are looking at the student's facial expressions and body gesture, the tone everything classifying them as one of the affective states bored, confused or engaged something like that.

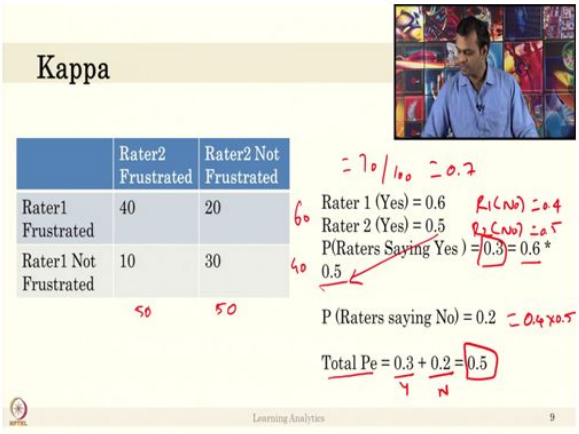
So that two raters, we can not have two raters for our complete research we want to use one rater for 5 sets of students now they differ by students. But how do we avoid the bias between these raters? So that is called Inter-rater agreement. Initially, we have to ask two raters to observe the same set of students and check whether the two raters agree on their classification. That is if you have items to classify, say boredom, confusion, engagement, and you want to classify into two or three categories, and now these two categories are accepted at both raters. So, in order to measure whether there is an agreement between these two raters, Cohen's Kappa is used. So, Cohen's Kappa the formulation is

$$k = \frac{P_0 - P_e}{1 - P_e}$$

Let us see what is P_0 *and* P_e ?

P_0 is accuracy. From the confusion table, we can compute the accuracy which we computed in our last class. And P_e is the hypothetical probability of chance agreement. What is the hypothetical probability that both raters will agree, is like what is the minimum value they can agree. So, how to compute Pe sum of the estimated probability that both raters agree for K number of items. We will see an example of how to compute Kappa score.

(Refer Slide Time: 09:14)



Let us take this table. Let us understand the table first. There are two raters looking at students facial expressions they looked at So, they looked at around 100 instances of facial expression. I am not telling 100 students that might be 2 students, there might be like 50 students, but they have 100 instances of facial observations. Both observed all the 100 instances.

Rater1 said at 40 times frustrated the same as the Rater2, but Rater1 said that 20 more times students are frustrated but Rater2 did not say that. He might have said not frustrated. So, this is a simple confusion table similar to what we saw in the classification problem. So, Rater1 and Rater2 agree they are frustrated Rater1 and Rater2 not agreeing on student frustration, this is the cross-validation studies Rater1 not agreeing it is frustration but Rater2 marks as a frustration. This is the wrongly classified problems there is no agreement between these two Raters. What is the accuracy simple to compute this 40 plus 30 divided by the total number of observations that is 70 by 100. Very simple to compute. What is Rater1 Yes agreement? Let us compute Pe now. What is Rater1's agreement? Rater1 says Yes for 60 per cent of the time compared to all 100

samples. 60 per cent of the time he says yes that is 40+ 20= 60. Rater2 says frustration 50 per cent of the times, this is 50.

If you just add these two values, so Rater1 says 60 per cent of the time Yes and Rater2 says Rater 2 says Yes for 50 per cent of the time that is Rater1 saying No is 0.4. What it says that Rater1 has a bias of telling Yes more compared to No, which means when you see a small slight expression in the face, you might mark it as a frustrated that is a Rater1 bias.

In a Rater2 it is equally saying yes and no, for example, he says yes 0.5 times. So, Rater2 saying yes is 0.5 and Rater2 saying No is also 0.5. This is from this value $\frac{50}{100}$. So, what is the probability that both Raters will say Yes

$$P(\textit{both Rater will say yes}) = P(\textit{Rater}1\ \textit{will say Yes}) \times P(\textit{Rater}2\ \textit{will say Yes})$$

$$P(\textit{both Rater will say yes}) = 0.6 \times 0.5$$

$$P(\textit{both Rater will say yes}) = 0.3$$

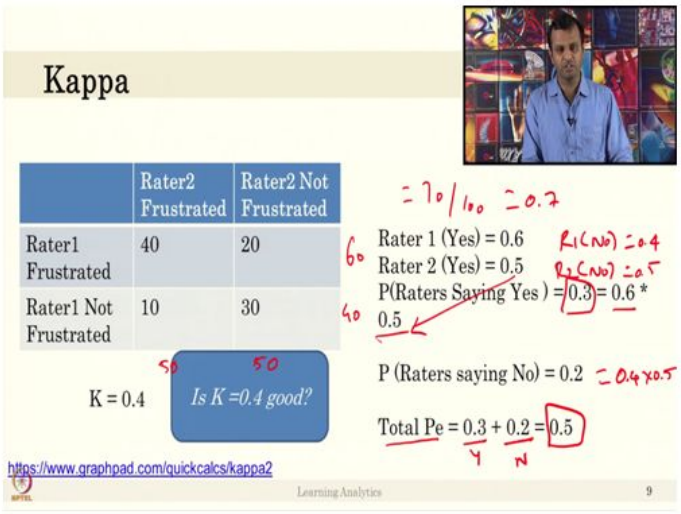
Similarly, for Raters saying No will be

$$P(\textit{both Rater will say NO}) = P(\textit{Rater}1\ \textit{will say NO}) \times P(\textit{Rater}2\ \textit{will say NO})$$

$$P(\textit{both Rater will say NO}) = 0.4 \times 0.5$$

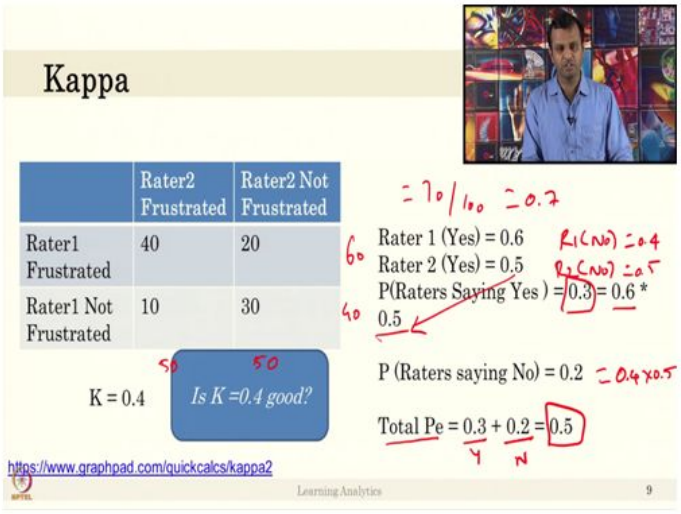
$$P(\textit{both Rater will say NO}) = 0.2$$

(Refer Slide Time: 12:51)



This we saw that some of the hypothetical property of chance agreement that is P_e as an observation accuracy we say.

(Refer Slide Time: 12:59)



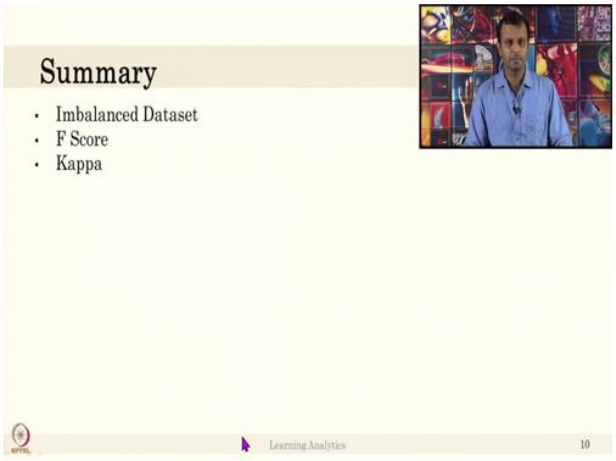
So if you use the 0.7 value in the formula, you get the Kappa score equal to 0.4 you can compute that. Please apply that in the formula we gave it in the previous slide and do that. Is K equal to 0.4? Good? Is the question. Think about it. If you want, can pause and go searching the internet

and see his Kappa score 0.4 is good. There is no answer, no exact answer to this, but it depends on domain the K value good or bad can be inferred.

So in this scenario, that is both raters will agree will be 0.4. That is not so good score for which is not so good score for Inter-rater agreement reliability. So you compute Kappa, do you need to do it every time like this says simple website which uses like a two cross two table confusion table.

Just enter the values in a table and say calculate it will calculate and give the Kappa Score. This is a website and a lot of tools like, lot, a lot have tools or script language, the software library and the machine learning tools have the library to compute Kappa Score easily.

(Refer Slide Time: 14:23)



So in this video, we saw what is imbalanced data set that is in a data set there are too many positive cases, too many negative cases. In order to pick up the better performing classifier, we have to comes up with a new score which combines accuracy, precession and recall.

One can be an F score. A simple one to start with is Cohen's Kappa. Cohen's Kappa is used widely to pick the right classifier. We will look at better classifiers or better metrics. In the next video, we will check more metrics on picking the right classifier. Thank you