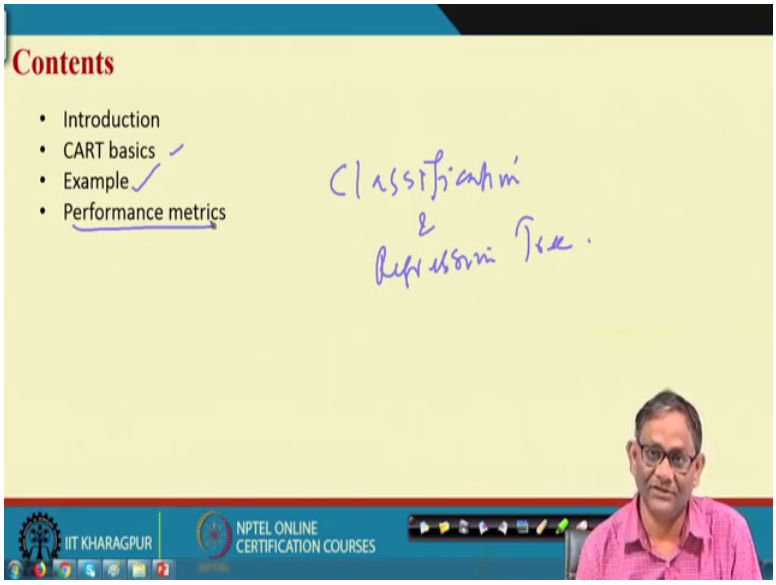
Industrial Safety Engineering
Prof. Jhareswar Maiti
Department of Industrial and Systems Engineering
Indian Institute of Technology, Kharagpur

Lecture – 50
Accident Data Analysis: Classification Tree

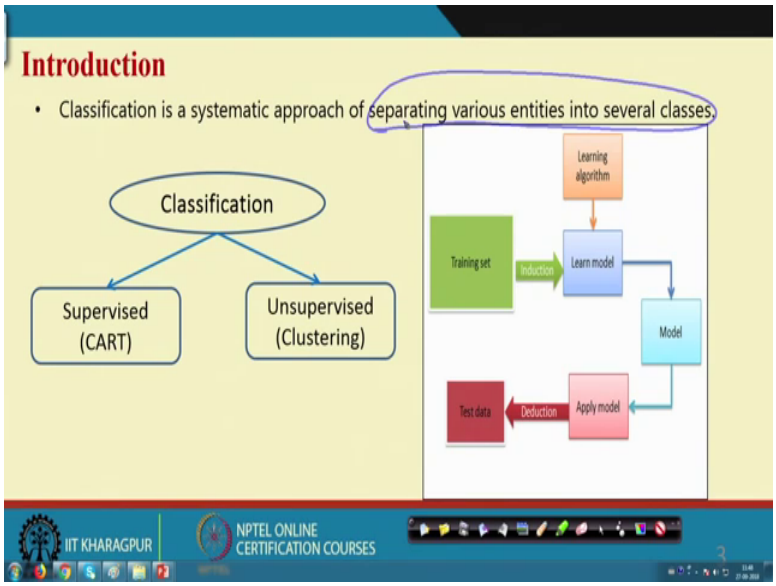
Hello everybody. We will discuss Classification Tree as a part of Accident Data Analysis. You have seen in last class we have discussed regression, we have discussed only the concepts and the basics and some results. For classification tree also I will do the same thing will not be dealing the little mathematical issues here because our purpose is not to treat the mathematics our purpose is to see how classification tree will be useful in accident data analysis.

(Refer Slide Time: 00:55)



So, the concept will be discussed that CART concept, where CART stands for Classification And Regression Tree classification and regression tree and then some one example we will see and then performance metrics of classification tree will be discussed. In fact, this performance metrics will or our classification regression tree or in case of prediction also it is used so, general one.

(Refer Slide Time: 01:33)



So, what is classification? Classification is a systematic approach by separating various entities into several classes separating various entities into several classes. For example, if I say that living animals. So, if you say that the living bodies, then you can find out animal and non animal like human. Again, animal you can find out different kind of some animals are wild animals some animals and domestic animals. So, suppose given a animal to you, you are asked the animal from which class if it is cow you will say it is domestic animal, but it is tiger you say it is a wild animal. So, that is basically the concept of classification.

So, another concept tree classification problem will be suppose, you are the bank manager and then you know you your disbursing loan to people. So, there can be a group of people who are always the filter they will not repay the loan. So, giving the person individual value individual attributes, the persons income and social economic status and you can vary you can find a more build a classification model in which will ultimately tells you later giving these information whether the particular person will be repaying the loan or if repay what is the probability of repaying something like this.

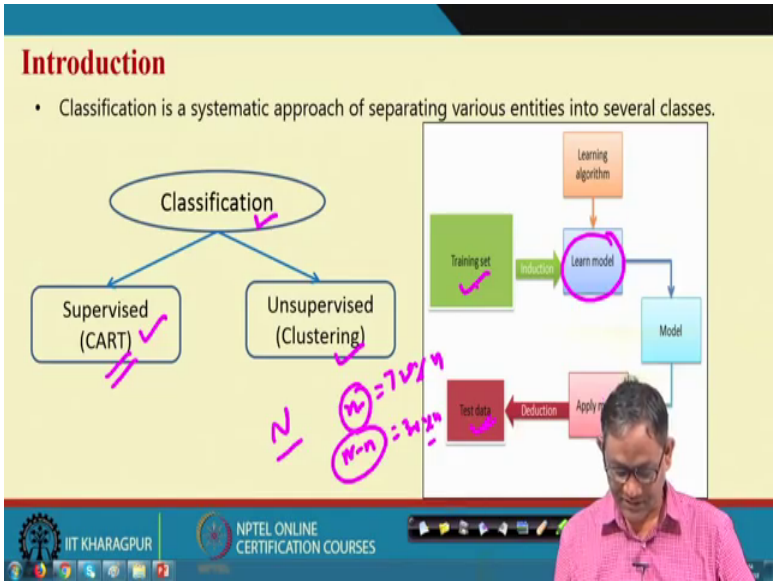
So, you may and same thing can be thought of in accident scenarios also because you have seen that accidents are result of deterioration or deviation from normal condition in from the process point of view, procedure point of view, from the technology point of view, from the environment point of view, from organization point of view many things.

So, in the in the lower level that they are basically unsafe apps and unsafe condition, incompetency's, inspection, lack of inspection lack of removal many other things are there.

So, now, what may it may so happen that situation the situation can be of different severity level a situation can be of satellite problem situation another can be that high severity another situation can be of less severe or maybe severe point severity point of view. So, the if the ~~if the~~ workplace conditions are given to you and then you are asked that you tell that and if today from one wants to work there so, what is the exposure to what severity level of exposure he or she will be having, then also you can go for classification model because here the classes are class is the severity and there will be different severity class~~,..~~

There could be many other examples in the safety domain that can be that can be talked about. So, you just think and you find out the way, but here we are basically interested to tell you that classification is a good to when you have large number in enough data and then you have and the attributes of the variables of interested also and logically and correctly quantified and data are collected accordingly, then you can use classification model to predict accident situation or accident severity or incident categories or something like this.

(Refer Slide Time: 05:31)



Now, classification is having of two different types one is supervised, another one is unsupervised when you know the class. For example, for bank it is defaulter or not defaulter, for accident the situation is grave or situation is normal, for people for people maybe he given this job he will be committing some mistakes or not. So, in that case now in the classes are known it will be supervised, but in some cases classes are not known what actually will happen, but there are so much of features or the characteristics which are interest.

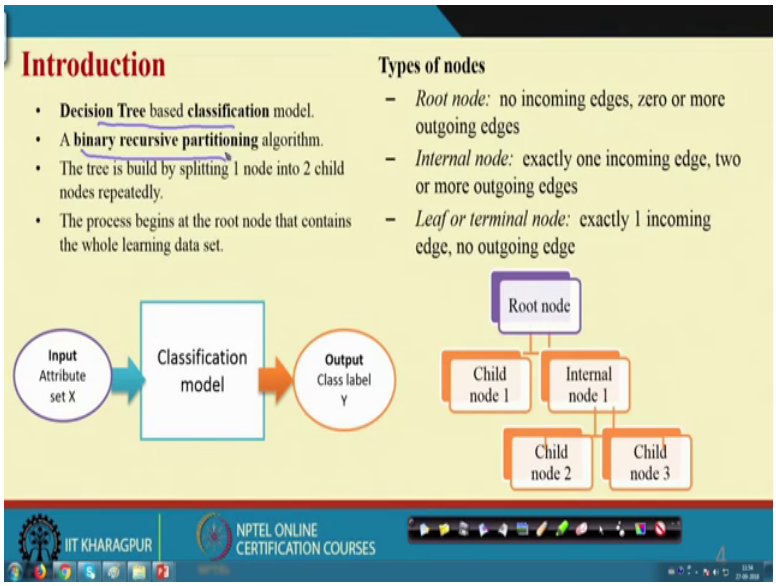
For example, we have different sections with the all the work system component values characteristics values. Suppose, you want that you that cluster the groups or the work places into different zone of zone of safety level point of view and then ultimately you have to go for unsupervised method like clustering and then after that seeing the features you will name the clusters,.

So, one of the supervised classification method is classification and regression tree. This is the determining techniques and it is basically used in also coming under the machine learning concept. So, what is the concept here basically this model this model first learn using the training data and then we apply the model using test data. So, even that is true for the regression also suppose, you have N number of capital N number of data points you take small n data for training then N minus n whatever another small amount small is there like 70 percent of this is equal to 70 percent of n then this maybe 30 percent of n.

So, you use this to learn the model and you and the small data for apply the model to test with the model your working or not. Here what happened basically you will create a learning algorithm and that algorithm will learn by considering the instances one after another and then one model is model is adequate enough. So, you would have to test whether what is learn adequately you apply the model with another set of data and you see that whether actually model and adequately or not or you required to refine the model,.

So, this is true for all prediction model including CART regressions and other things.

(Refer Slide Time: 08:08)



So, now let us see the basics of CART. Actually it is a decision tree based classification model. It basically split the entire data into different parts. The splitting is binary in nature means two like, I have the entire data set like this first I split like this is one this another, then again I split this making it to another one this making it to like this binary splitting continuous. So, the tree build by splitting one node into two child nodes. So, as a result we will start when you start with all the data all data are in the root node then you considering certain attributes the split will take place and it will go to child node 1 and then internal child node 1.

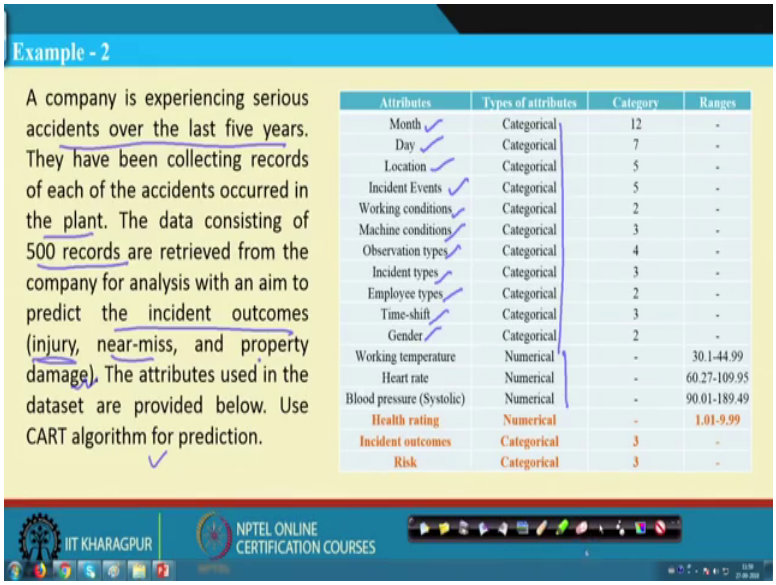
So, this is basically node 1 and node 2. This two are child children to this root node. Now, when we are saying child node 1 here means this is not further splitted, this is the end here. Now, at the internal node it can become parent to another child nodes, ok. So, ultimately at the bottom ~~a child nodes~~child nodes will be there. What will be the child node? Child nodes is when one where the data is homogenous in nature or further splitting will not increase the homogeneity of the data~~,~~.

So, that is what is the way classification works. So, root node: root node, no incoming edges 0 or more outgoing edges internal node exactly one incoming edge one incoming edge, two or more outgoing edge and leaf or terminal node these are child node exactly one incoming edge, but no outgoing edge. So, how this cart model will use they basically consider different sets of attributes these are X or other way in the regression

independent variables we say same manner then the model learns and output will be the class level which class of in stand it is. If it is individual related what type of individual you have see is, if it is work place regretted what type of work place it is it is severely accident prone or it is a safe or it or it is it is basically moderately dangerous something like this.

So, what I said that you have data, huge data for different attributes. So, this data will be splitted. The split take place in binary mode first route data set will be split into two child children nodes and then the again they will be split and ultimately at the bottom there will be only the leaf node or terminal node. So, what is the basic of when do you stop? We stop when the homogeneity of the data is not further improved in the child nodes.

(Refer Slide Time: 11:29)



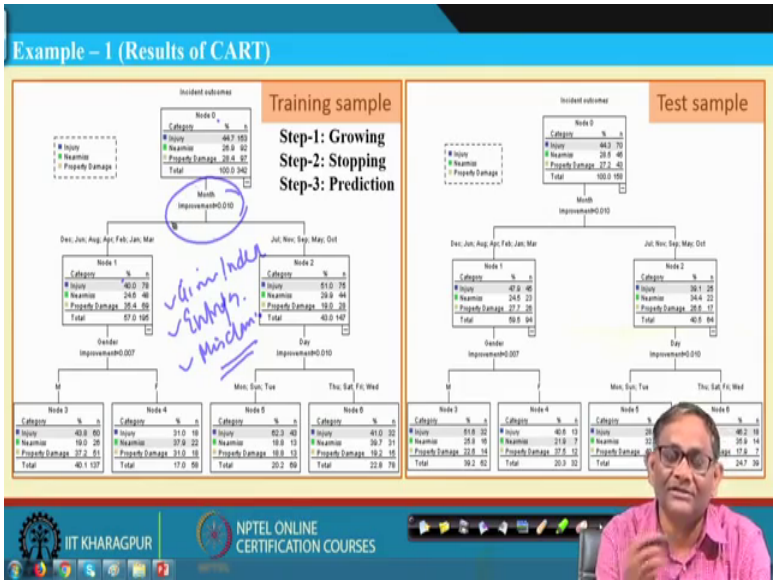
So, let us see one example suppose a company is experiencing serious accident over the last five years, they have been collecting records of each of the accident occurred in the plant. The data consisting 500 records are retrieved from the company for analysis with an aim to predict the incident outcomes like whether injury near miss or property damage what kind of accident will take place. The attributes used in the dataset provided below use cart algorithm for prediction.

So, what are the attributes? Month, which day, which location, Incident, event, what type of and what is working condition machine condition observation types incident types employee time generation. So, many. Soso, mainly these are all location

timestamp, process related information, procedure relating information, then your inquiry level information. ~~and~~ And finally, obviously, what you want to predict so that categorical variables most of them are categorical in nature some are numerical ok. So, there are different categories this is what the dataset.

Now, you want use this dataset and you want show that how CART we will help us. Please keep in mind that, this data set is not real data this is a hypothetical data again. Although there is touch of reality from the attributes point of view, but data are not data not real, it is there data we have it.

(Refer Slide Time: 13:09)



So, we have use some software ok. So, you use software I already given in last class different software. So, use any of the software and then what happened you will basically find out interestingly that how many classes are there? One is injury class, near_miss class, property damage three ~~class~~classes that is what you want to predict. Given the workplace, attributes, individual people working attributes all those things. So, you want to know that who is a given that situation which one is most likely to occur injury damage or property damage something like this.

So, here at the root node you see that only these three classes are there injury near miss and property with this number of observation there are total number of observation 342~~,~~. Then it is questions comes that what are the, which attribute you will choose basically you tried to use one of the attribute at a time and then split do the split here. The split is

split takes place depending based on certain mathematics. So, that that is known as that is known and basically, growing.

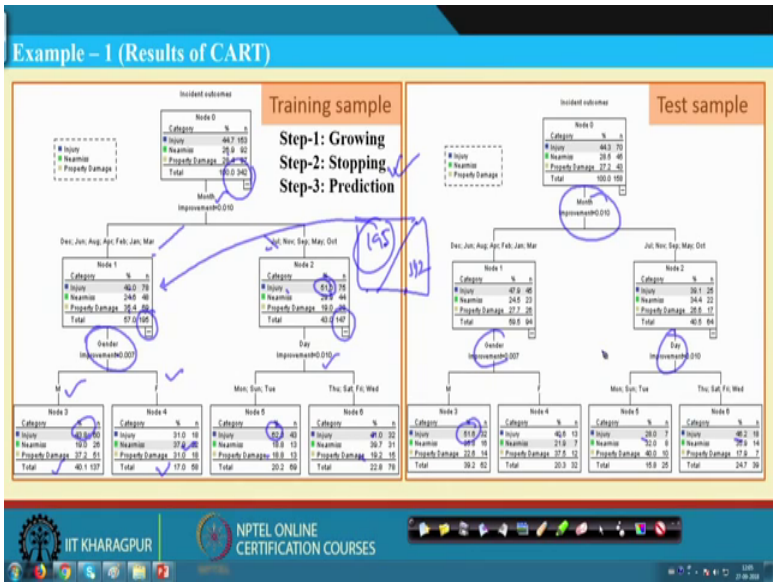
So, what you will do at this point we you want to you have to find out consider each of the variables and then or attributes and then find out that which attributes is giving you the maximum improvement in terms of homogeneity. So, here the class homogeneity these are this, the class homogeneity here, if the child node they would be different and here the homogeneity will be less compare to this. So, that is the issue or other way we can see in improve it.

So, so you will choose that variable who is in the split which gives you the maximum improvement from homogeneity point of view, maximum gain from homogeneity point of view that mean the impurity will be reduced maximum. So, as there are many variables to attributes together you will you will take that one which gives you the maximum at the beginning that is the starting point. So, we are saying basically that improvement 0.0 this.

So, please understand that there are different index like Gini index Gini Gini index then entropy index then miss classification ready their misclassification probability. So, there are many such measures in index are their which used to find out the find out the impurity the impurity here when they are compared here impurity will be less. So, that is the improvement or impurity in other sense you can say homogeneity and homogeneity is here or heterogeneity. So,

So, using those three, one of the three indices like a Gini index entropy another criteria you find out which variable is giving you the maximum benefit, ok.

(Refer Slide Time: 16:24)



So; that means, the entire dataset you see 342 was the initial, now it is split to two parts, in this case where (Refer Time: 16:29) it is 195. So, one ninety 342 minus 193 will be coming here. The, what is the benefit you found out here injury 44, 26, 28 here 40, 24, 35, but here in this class 51 percent injury side. So, it is basically improved one it is again it is not a very good splitting in the sense that the homogeneity is not that much improved, but it is improved one.

Then, now you are dealing with this set. So, originally 342;- ~~So~~so, then 195, 195 here then 147 here now this one this is here. Now, this one again this data you are your finding out with the all the variable which one is giving you the maximum benefit here we found the gender is giving maximum input. So, gender male and female. So, then again the split has taken place similarly this one also like this. So, here what happened this become 43 percent here this is a maximum, here the 62 percent, here it is again 41 percent.

So, in this manner you continue and finally, you will you will finish or you will stop you will stop when there is no further improvement possible given the attributes and using any of the impurity calculation criteria calculation methods. So, here what happened 51 percent maximum it is 40, 28, 48 ok. So, the data is not that good here because or other way I can say the variables the for the purpose it is used they are not able to explain that

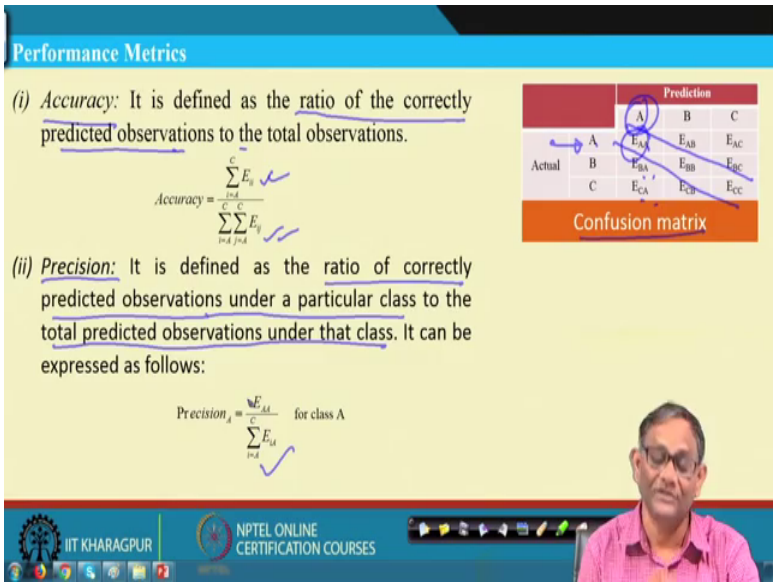
much, but nevertheless it has given you some information that what which of the variables contributing you started with month, then gender here, day again month comes.

So, then again what happen gender comes to month and gender and they these are coming in this data because hypothetical data we have constitute again. We might have common mistake in putting the values, in reality it may not be this much it may be much better even or much was also absolutely. But, this is what is the method of cart then what happened given the given attributes you will be able to predict which class it is there. So, that is the last data will the prediction,,.

So, so, that means, you please understand that this is basically tree is cart tree is generated. So, what ultimately we have done I just missed forgot to tell you. So, this side we are talking about the training side and this is a test. I am sorry that I have started from here and finally, landed here. But, it is not that using the 70 percent of the training data you got this is the end and here test data and almost there training and test giving you the similar kind of results and that is that is that is to be,,.

So, let me repeat growing stopping prediction this is important. So, there are many criteria in this that to be used. Now, you start with the root node in you ultimately go for the children that leaf node final one. The leaf node is one where you cannot further split because that split will not give you any improvement. So, there will be two sets of data training sample and test sample and in the training sample like this and the test sample like this is given here. It is not that from here that again day actually here month-month, gender-gender, day-day this is coming. It is it can be it can be further split. But, we have not shown because the further split is not giving you better result,,. So, correct it.

(Refer Slide Time: 20:41)



Now, what will ~~happen~~ happen? So, using the model the CART so, you will be able give in the given the attributes you will be able to predict under which class. So, here the actual class these are this predicted class. These are the predicted class, these are the actual class. When you observe that data you can find out a classification table like this, where actual class and predicted class will be will be represented in this manner.

Then, this one E AA or E BB or E CC these are the frequencies which talks about that actual class and predicted class matches and raised that are of diagonal element raised of diagonal element frequencies are mismatch or misclassification. This metrics is known as confusion metrics or misclassification metrics. So, in the confusion metrics or misclassification metrics the diagonal elements are the correct classification off diagonal element are incorrect classification.

So, the classifier accuracy or classifier performance will be tested using this confusion metrics. There are several parameters or several criteria have been have been developed one of them is or performance metrics have been developed one of them is accuracy another one precision and some more are there.

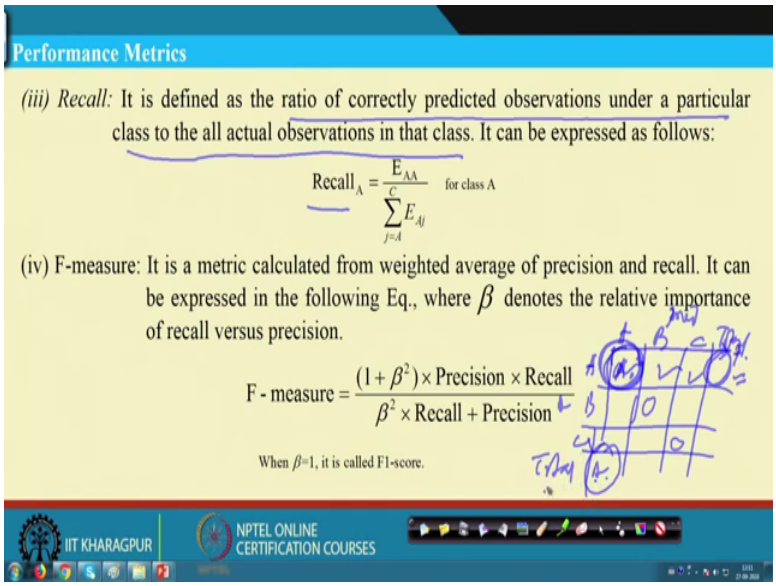
So, what is accuracy? ~~it~~ It is defined as the ratio of the correctly predicted observations ratio of correctly predicted observation to the total observations. I told you the diagonal frequencies are correctly predicted. So, I equal to 1 to A to C E II AA BB CC divided by

total observation all. So, that mean actual one from A to C and predicted one from A to C that I A to C j get soon every observations you have consider that is the accuracy.

Then another one is the precision. It is defined as the ratio of correctly predicted observations ratio of correctly predicted observations under a particular class to the total predicted observation under that class. So, you actual class you consider class A. So, what is the ratio of correctly predicted observations of that class that is nothing, but A then what is the basically total predicted observation under this class? So, under this class total predicted observation is AA E BA E CA A E AA E BA E CA i equal to A to C, now that is your precision.

So, what do you want? You want accuracy should be high precision should be high that is the, that is the issue.

(Refer Slide Time: 23:51)



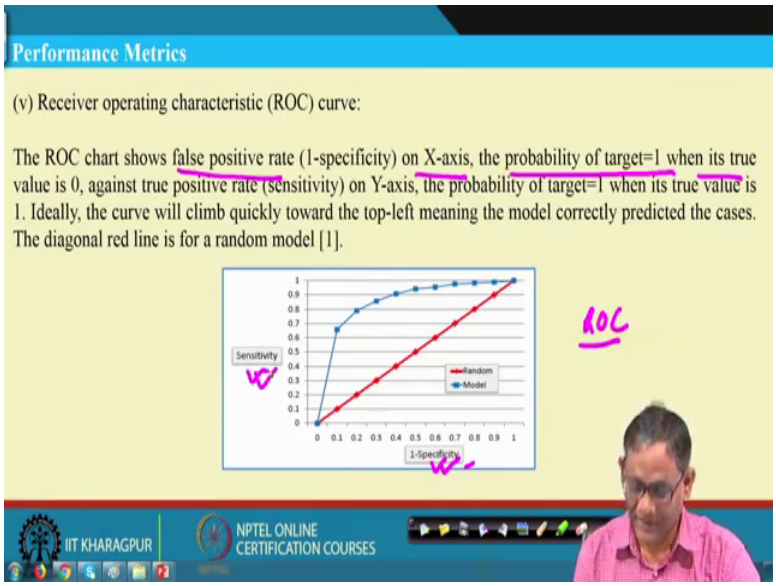
So, now the third one is recall.. What is recall? Recall is E AA by this. So, so, you see it is the ratio of correctly predicted observations under a particular class to actual observation in this class you have this is the predicted part and this is the actual part. The correct one is in our case correct one is this one is correct, this is correct, this is correct. So, this is actual B and C predicted A, B and C. What is recall? Ratio of correctly predicted observations, so, this one divided by the total number of observations that is basically that is actual observations in this class;. ~~So~~so, this one.

This plus this plus this total and if you take divide by this total then this will be precision. If you divide the correctly classified observation by the column the root total it is giving you recall, if you write like this is the predicted and this is actual. If you if you do the same thing that may divide the correctly classified one by the column total in this question then that is will the recall.

So, now what happened they this precision and recall are very important one. So, these two are combined in a matrix which is known as F-measure. F-measure. What is F-measure? It is a metric calculated from the weighted average of precision and recall. It can be expressed in the following equation where beta denotes the relative importance of equal versus precision. So, what is F-measure? F measure 1 plus beta square precision times recall by beta square into recall plus precision when beta equal to 1, it is F 1 score, ok. So, that is basically weighted combination.

Now, what will be the beta value ? So, beta 1 we can be other values also. So, so, what we say that F 1 score is used or other measure F values you can take, but F 1 score is mostly used. So, this basically gives you a metric which is basically combining the two important performance metrics called precision and recall.

(Refer Slide Time: 26:29)

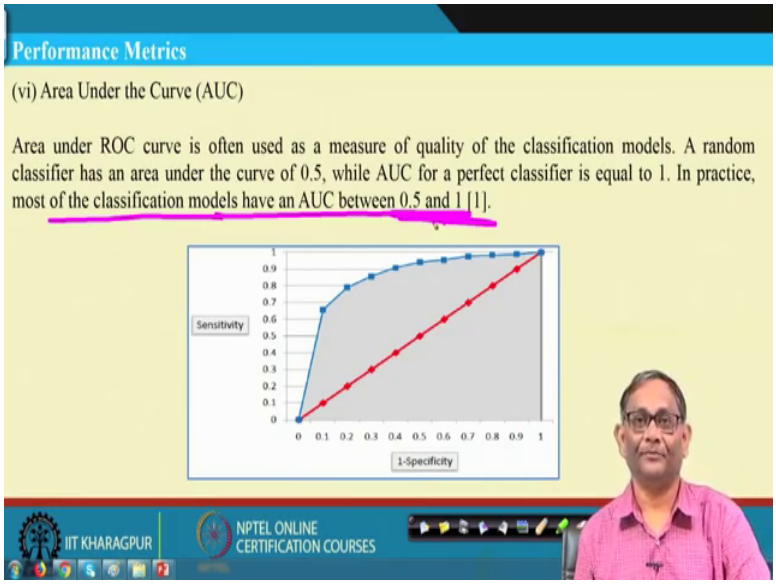


Then another one; that is all important one, which is known as receiver operating characteristic curve. So, this is basically the sensitivity and specificity is the, they are the ROC chart shows that false positive on X-axis and probability of target when it is true

values something like this. So, the area under this curve is very important and that will basically talked about the performance of that is.

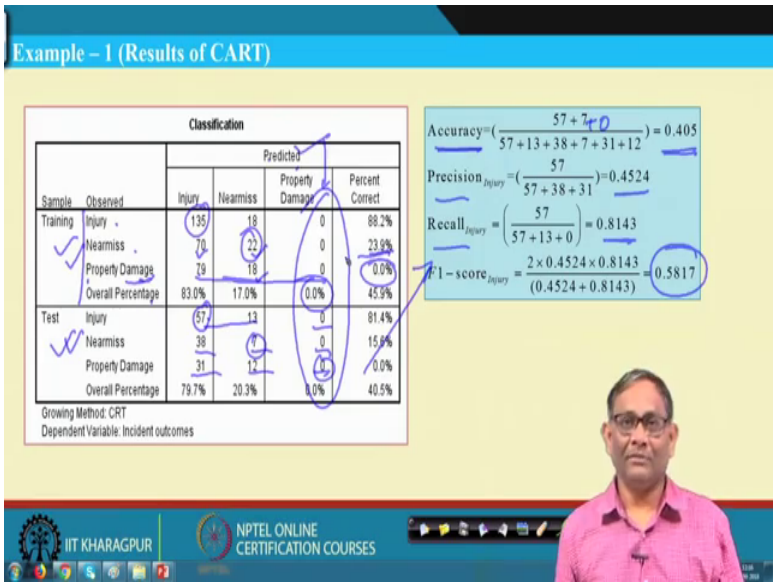
So, true positive true negative and other things. So, I will not discuss here. So, what I request all of you just please go through this is a ROC curve little more detailed and then the true positive, true negative, then sensitivity and all those the specificity all those majors also please go through and do your home assignment otherwise you if you are not able to understand put in the discussion forum because these things are important point not only for understanding ROC curve or also for the examination purpose.

(Refer Slide Time: 27:35)



So, the area under the curve is very important, that area under the curve we basically talk about what is the what is the performance of your that classifier. So, I under the curve the more that is a better one. In practice for most in practice for most practical purposes this will be 0.521.

(Refer Slide Time: 28:06)



Now, for that particular data what we have developed the classification tree. So, we are basically seeing that what is the measure. So, in the training data and test data two kind of data set we have used. In the training data that injury, nearmiss, property damage and overall;- ~~So~~so, the predicted value and this is the actual value. So, 135, 22 and here it is 0. So, so, that means, you see that so much of misclassification is taken place. So, the accuracy is basically total accuracy which side the same thing from the test data also you can do and the, this one given from the test data point of view.

So, correctly classified 57 and plus 7 plus 0 57 this plus 0 by total observation 57 plus 13 plus 0 38 plus 7 plus 0 and your other one is 31 plus 12 plus 0, ok. ~~So,~~ -So, actually if I see the actual property damage case this if I see the predicted property damage case in both case it is 0. So, it is very bad model I can tell you and accordingly what happen accuracy is 0.405 precision is 0.4524 recall is 0.81 ok. Recall is better and F 1 score is like this, but particularly I am surprised to see this that these side it is completely 0 ok.

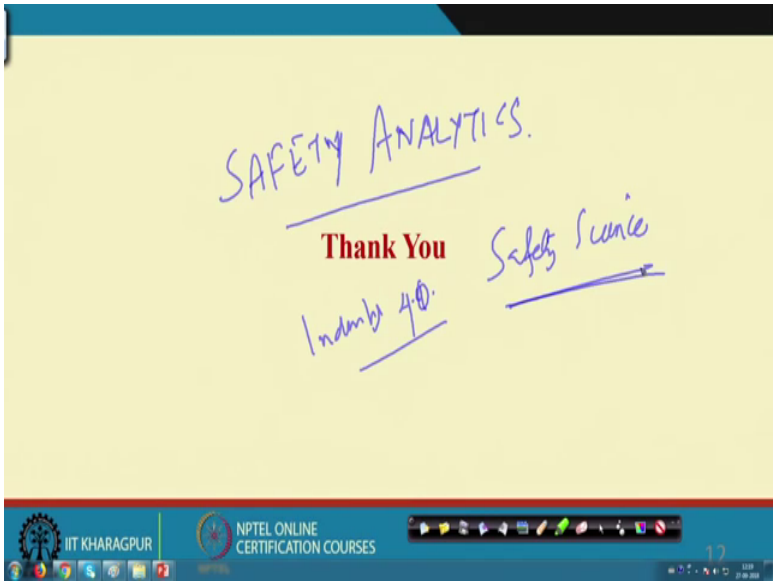
But, if I if I see that these are the predicted than the actual one it is a huge number is there 79 plus 28 plus 18 and here 70 plus 22 that 23 percent [FL] property damage case, ok. This is basically version correct predicted one is 0. So, this is not a big issue for all of us what is the important thing is that, so, CART can be used for production purposes and once you use any software you will be getting such confusion or classification table.

You will be getting it for the both the training and test data then you can find out all the performance measures with the reference to the reference to the cart model and then accuracy precision recall all should be value should be high value this will be good value. So, then ultimately there will be there is no such hundred percent absolute threshold value for all those things. But, it should be as good as possible.

So, if you go by that is then literature so, you may find out some of the some of the useful values for accuracy precision recall and F score. So, these are the issues what is to be dealt with before using this model or prediction.

So, I hope that this lecture make some sense to you also. This is a classification model CART basically we are shown in the classification tree not the regression tree, but the classification and regression tree it is commonly used in a. So, so, it is also similar to your regression, but in classification tree you will see the attributes are categorical in nature and integration tree attributes will be continuous in nature.

(Refer Slide Time: 32:26)



Thank you, very much from the accident investigation and accident analysis data analysis point of view this is what we have coved. This is a huge topic, in fact, there is one concept called Safety Analytics, this in today's very hot topic is safety analytics. Safety analytics, in fact, under industry 4., industry 4.0.

So, you will see that are now this analytics AI machine learning, then the sensor data all those things means safety, maintenance operation related all data together you will use and ultimately you find out that it will give you a better prediction model for a safety analytics is a is an area to work on particularly if you if you are interested in safety analytics. So, there is there is a journal called Safety Science. Safety Science, there is a special issue. Safety Science special issue on safety analytics where you can if you are researcher, if you are practitioner who want to disseminate your knowledge you can we can send this paper this is basically we are organizing this special issue ok.

Thank you very much. Have a good day.