

Course Name:Business Intelligence and Analytics
Professor Name:Prof. Saji.K.Mathew
Department Name:Department of Management Studies
Institute Name:Indian Institute of Technology Madras
Week:07
Lecture:25

DECISION TREES | Business Intelligence & Analytics

Hello, and welcome to this session on Decision Trees . This is a very important topic as far as exploring algorithms for business analytics is concerned. We have just finished a discussion on classification, as a broad area or a broad category of problems. And I also indicated that there are several techniques available to address classification problems and one of them is decision trees. Decision trees, you know, it is a very interesting combination of two words. Decision of course, you can connect with real life, you can connect with business and management, but trees belongs to the biological domain or it belongs to the nature.

So there is a metaphor in the title, the tree is a metaphor. So let us see what is the tree structure. It is a structural metaphor, I would say. It is sort of the tree gives us a sense of what is the structure of a solution that we are trying to build to classify.

So decision tree is an important class of problems, class of techniques or algorithms that is widely used for classification. And decision tree is a technique that belongs to the supervised learning. Supervised meaning there is a target variable and in terms of our different categories of techniques which we discussed earlier, this is supervised and this also is algorithmic. This cannot be called a statistical technique, it is an algorithmic technique which originated in the CS, from the CS community. So we will first try to understand what is the structure called decision tree or what is the thinking behind such a algorithm or method and what is being done and how is it being done, what are the sources of data and also other inputs that are required for the algorithm to function.

So we will get into details slowly as we go and also try understand different types of decision tree algorithms towards the end. And I must also caution you that this is an introductory session. I am introducing you to the topic of decision trees. Decision trees is an area of research, it is an area of attention both for researchers as well as for teachers in itself, meaning this is a topic in itself for a course. You can study this in much more detail.

Sunburn at the beach ?

NAME	HAIR	HEIGHT	WEIGHT	LOTION	RESULT
Sarah	Blonde	Average	Light	No	sunburn
Dana	Blonde	Tall	Average	Yes	None
Alex	Brown	Short	Average	Yes	None
Annie	Blonde	Short	Average	No	sunburn
Emily	Red	Average	Heavy	No	sunburn
Pete	Brown	Tall	Heavy	No	None
John	Brown	Average	Heavy	No	None
Katie	Blonde	Short	Light	Yes	None

Example from Winston (1996)

BUSINESS INTELLIGENCE & ANALYTICS

So therefore you can imagine that when decision tree is covered as one topic within a course on analytics, you get an overview and that should be the expectation from the session. Let me start with a very basic example, an example with which many of you could connect. This is about a group of foreign visitors going to Chennai's Marina Beach. We get a lot of visitors from other countries, visiting scholars and visiting students from other countries particularly the Nordic countries or European countries where they do not have a sea close by. And so is the case with students who come from northern India.

Many of them have not seen a sea or a seashore or a beach, you know, rarely some of them have not even seen a beach. So they rush to the Marina Beach immediately after they join IIT Madras. And I am just suggesting that this example is something of that kind where a group of eight students, eight visitors go to a beach and that was a great experience. You can see their names Sarah, Dana, Alex, Annie, Emily, Pete, John and Katie, both boys and girls. And they do go and they had a nice time.

We can spend a whole day or a half day or an evening and then of course, the next day they were having their breakfast and while having breakfast they looked at each other

and then they found that some people have got sunburn. They have got sunburn in the beach and while some others did not get sunburn in the beach. You see some got sunburn and others did not get. You see a binary problem here sunburn or no sunburn, that was the outcome, that is the result. I am conveniently quoting an example from the book of Winston ordered in 1996 and I am just sort of contextualizing it.

But this is the case, you know, group of eight people or eight objects here in our machine learning language, eight objects and the outcomes are binary, either they got a sunburn or they did not get sunburn. And someone in the team was very curious, what explains? You know, we all have a need to explain things, what explains this different outcomes? When if everyone got it then the curiosity would have been less but there would still be a curiosity but if none got it nobody cares, it is an indifferent case. There are several such cases where we have, we are indifferent to it. If nobody has a sickness or nobody has a situation or an outcome, no objects in our case then it is not an interesting case. But this is interesting because the result is mixed, both sunburn and no sunburn.

Indifferent cases, as something that I will point out when we discuss various measures for distance in cluster analysis, that is a subsequent session. Alright, so coming back here, somebody was curious and she wanted to find out what explains this differential outcome. And she decided to collect data about the different individuals and that is what is listed here. She imagined or reasonably imagined, this it is a recent imagination that it could be hair color, it could be height, it could be weight, it could be lotion used. There is, there are lotions that you can use against sunburn when you go to the beach.

Very popular in western countries but not so much in our country because we do not care. But in any case, that was another variable. You can see her sharp sense of variable collection. These are variables which she thought is relevant and is also, or these variables are also variables for which data is easily available, data availability. So you know very interesting small project.

So here is the data, 8 records and the corresponding data. And you know, if you look at this data, can you form some rules which leads to a particular outcome or in other words, what are the conditions for getting a sunburn based on the 4 attributes? We will think that name is an attribute of person but it has no influence on the outcome. And therefore, there are 4 variables or 4 attributes or your feature set consists of 4 features, hair, height, weight, lotion used. Now if you manually look at this and you can sort of, go by each record and see what are some patterns or common rules that leads to a particular result. So a condition is written using 'if then' clause, you know that if then you write the condition.

If let us look at the last variable which we all will think will be the most important variable like lotion used or lotion not used. So if lotion used is yes, let us look at that condition. If lotion used is yes here, result no sunburn, yes here, no sunburn, yes here, no sunburn. So we have a condition here, if that is fulfilled, then the outcome is always one value, outcome is always one value, there is no mixed result or you know in other words the probability is 1. So if lotion used is equal to yes, then outcome is sunburn or in our case, result is the name of the variable.

So let me write result is equal to none. Lotion used is not there, so let me actually only put that much. If lotion is equal to yes, then result is equal to none. This is a valid rule that you can extract from this small set of data. But if lotion used is no, do people get sunburn all the time? Let us check that.

One case yes, they got sunburn, second case yes, third case yes, but here it is no, but no sunburn. John also did not use sunburn, but still he did not get a sunburn. And therefore, this is not a rule for which you can have 100 percent confidence. So therefore, the probability is not 1, there is some probability that one would get a sunburn or one would not get a sunburn, but it is not 100 percent. So this can also be a rule since there is some confidence that you can see, but maybe it is at this point you try to combine this condition with some other condition or you try to put a 'and clause'.

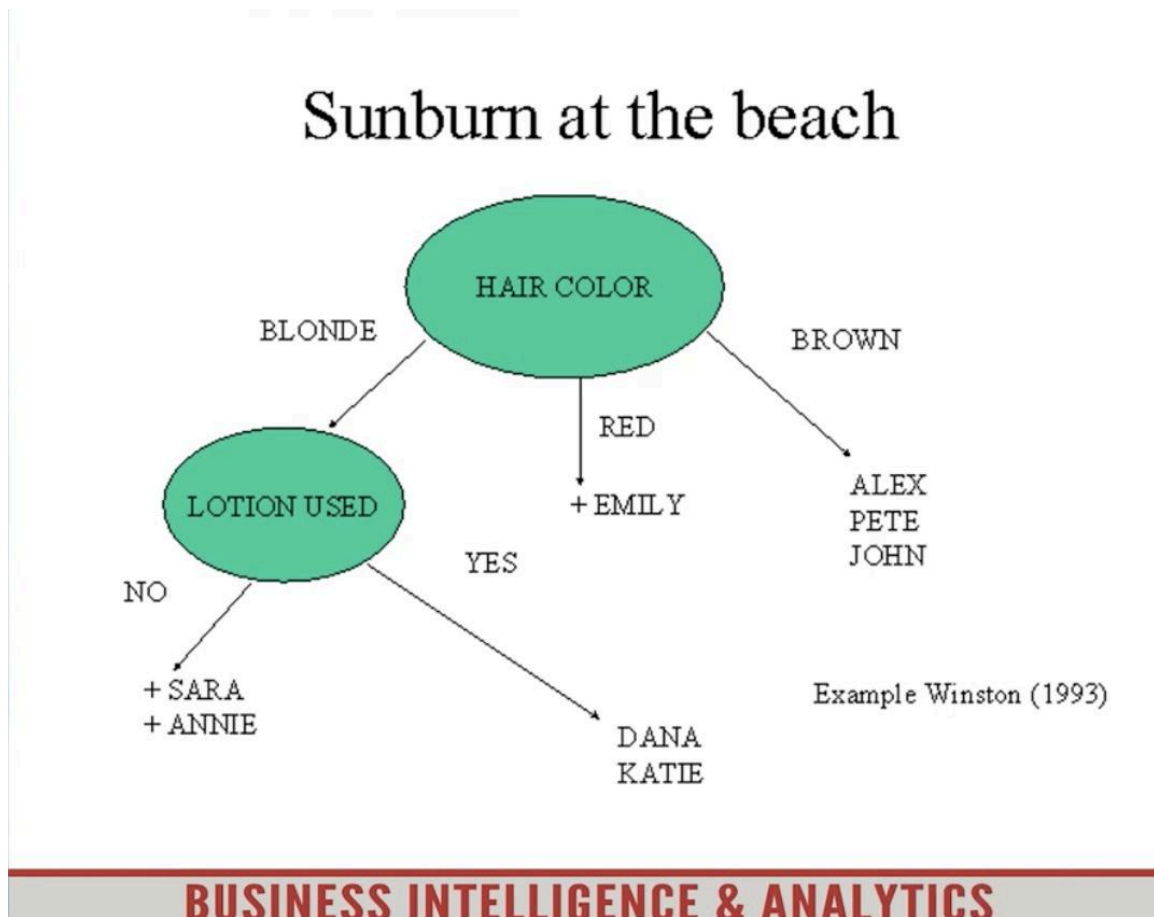
What is that 'and clause'? Quickly looking at this, you can see, look at the hair colour. If hair colour is blonde and lotion used is no, then you get sunburn. Yes, this is the case, this is the case, you get sunburn. This is the case of yes. So another rule that you can easily extract is, if hair is equal to blonde and lotion is equal to no, then result is equal to sunburn.

Result is equal to sunburn. This is also, we have 100 percent confidence. Probability is 1 because we see two such records which fulfill these conditions. So what are we trying to do here? What are we trying to do here? We have a set of data or a set of 8 records. There is an outcome variable which is a result, where the result has two values- sunburn and none, binary outcome.

And we think that this binary outcome is determined or influenced by certain attributes of the objects of the individuals who get sunburn and who does not get sunburn, depends on certain variable. So this is a dependence technique. We know as in regression, there is a y variable which is a function of x variables. But the only difference is that we are not having a mathematical formula here to connect y and x but instead of that we are actually expressing the mapping between y and x in the form of rules or algorithms in

the form of rules. So essentially we are extracting rules from a set of data, the rules that determines a particular outcome and in this case a binary outcome. This is what we try to do through this small example. Let us move on.

Now whatever we attempted to do, if we put this or if you try to visualize it, if you try to visualize it, this can be visualized something like this as shown in this slide. As shown in this slide again, I am referring to Winston's book. So where do we start? We start with the whole population, whole data.



So when we begin our analysis, we start with the whole data set. In this case a data set of 8 records. Now we first split that data set into 3 subgroups. This is 1, this is 2 and this is 3. 3 subgroups you can see. We are splitting the data set based on a rule or a condition. So here we are saying, if hair color is a variable as we have seen which has 3 values. So it is a categorical variable with 3 values, blonde, red and brown. So we use hair color as a subsetting criteria. We use hair color as a attribute to split a data set.

You know, the other term instead of subsetting in decision trees the more widely used term is split. So we are splitting a data set into multiple subgroups based on the attribute

values of one variable, one attribute. So we are choosing hair color. Why are we choosing hair color and not some other variable? Well, that is a question we need to answer as we go. But here we choose hair color and if hair color is blonde, then you know that you know a subset is formed with all the records having hair color as blonde.

If hair color is red, a subset is formed which has certain members and if hair color is brown, you have 3 people with hair color brown who are Alex, Pete and John. Now you can see as we did previously, you put one more condition if lotion used is no. Then there are two, there is another subgroup or another child that is formed from the parent. One way also to look at this kind of a structure is, you know look at in a hierarchical format. Just like we build the family hierarchies like, this is like the grandfather, grandparents, parents and children, something like that.

So in which case, the starting is the origin, that is the genesis, you know the genesis begin there. So this is called the root node. This is the root node in which case since we use this term node, each of the subgroup is a node. Each subgroup is a node and when a node is split based on certain condition related to an attribute, then subgroups are formed or children are formed. So 'lotion used' node is a comparison here but before that there is a group here that group is formed from the parent.

So this group is again split into two based on lotion used is yes or no, two groups formed. So each of them, each of the green circle that you see here, they are all nodes, this is a node. But is this a node, Sara and Annie? This is a node too. These nodes are called terminal nodes. This is a terminal node. This is a terminal node. A node that does not split any further is a terminal node. So therefore this is a terminal node. This is a terminal node. So how many terminal nodes? One, two, three, four. There are four terminal nodes. Terminal nodes are also called leaf nodes. So here we are bringing back the metaphor of the tree. So essentially you can see, that a tree has a root. Tree has branches.

Branching is based on subsetting criteria. And when the branches are formed, there are certain node from which branches split. So it is a tree structure and finally what you see in a tree is a branch keeps splitting, splitting. And finally there are leaves. These are the leaves. The final elements of a tree. But there is a difference here in the tree metaphor as you must have already imagined, the tree is inverted. The tree is upside down. Root is at the top and leaves are at the bottom. So it is a, to be accurate, this is an inverted tree.

It is not other way around. And you can draw it the other way around also but it is more intuitive or it is more easy to grasp the idea if you look at it in a inverted format. So all I was trying to do here was to explain to you how a tree structure is formed from a set of

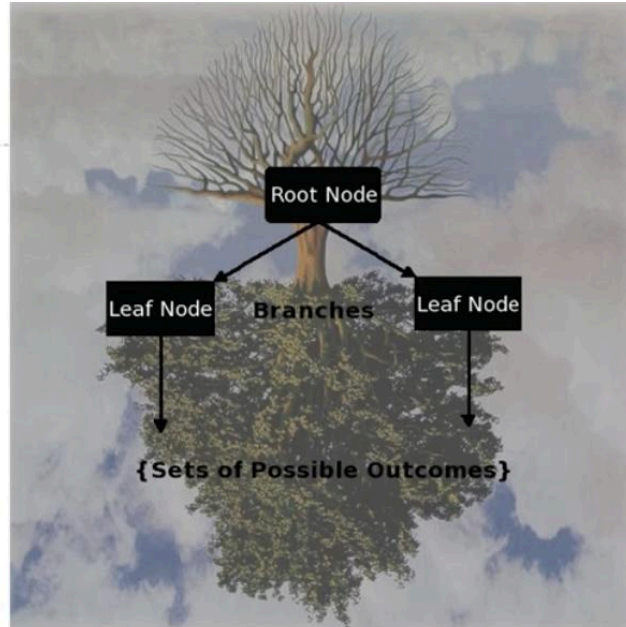
data with an outcome variable and other attributes. And I want to stress on one more aspect of the decision tree splitting.

This is a decision tree. This is a decision tree. And what is that? We must at this point ask this question. All good. You have chosen certain attributes to split a node into subgroups. But what is the basis? What are you trying to achieve? What is the objective of splitting a tree? There has to be some objective function or there has to be some ultimate goal. The goal in decision tree split, splitting is to create subgroups or children, child nodes such that a child node is more uniform or homogeneous, homogeneous subgroups.

A child node should be more homogeneous than a parent node. But whenever you talk about homogeneity, it has to be with respect to something. We say it is a homogeneous population. It does not mean anything, unless you say homogeneous with respect to what? That is a logical question. Homogeneity could be with respect to your culture, your language, it could be gender, it could be economic status.

There could be so many different variables with which homogeneity can be defined. And therefore, here homogeneity means homogeneous, homogeneity of the subgroups with respect to, is already there in WRT with respect to the outcome variable with respect to target variable. What is the target variable here? The result, which is sunburned, none, sunburned or no sunburn. So you can see that when you formed the subgroups and look at the leaf nodes or the terminal nodes, you can see this particular terminal node has only 2 members and both are homogeneous with respect to sunburn, both got sunburned. Look at this particular node, there is only 1 member, but of course, there is no point in defining homogeneity there.

It is 100 percent homogeneous one. Look at this node, Dana and Katie, both did not get sunburned. Again, they are homogeneous. Look at this sub node, Alex, Pete and John, they did not get sunburned and therefore, it is homogeneous with respect to the outcome variable. Now, we can actually try to define what is a decision tree. A decision tree is a set of rules that split a heterogeneous population or data set into homogeneous subgroups.



▶ Decision tree induction

▶ Classification

A decision tree is a set of rules that split a heterogeneous population into homogeneous subgroups. And that is example that we have just seen. So with this fundamentals, let us go forward as I already indicated to you, a decision tree is a inverted tree with root node at the top and leaf nodes at the bottom and branches in between. And what we have done in the previous exercise is something known as decision tree induction. Decision tree induction is an activity where we extract rules or we, in other words, in our modeling terms, it is the training, training space, training activity.

We actually take a training data set and then identify rules that lead to certain homogeneous outcome. And the decision tree training or decision tree building is known as decision tree induction. You are inducing from data. There are no prior rules. And therefore, it is not deduction. This is induction. This is bottom up. We are identifying rules from a set of observations and therefore, it is induction, not deduction. So, the training phase can also be called decision tree induction. And once a decision tree is induced or built, we are ready to use it for classification.

A tree, in other words, is a classifier. A tree is a classifier because we know from the tree, what are the rules that leads to a particular outcome. In our case, we have outcome which are either 0 or 1. A group has either, either sunburn or no sunburn. But in practical

settings, when we apply decision tree algorithm, we will see that there will be result nodes or terminal nodes, which will not be 100 percent pure or homogeneous. And therefore, what you get is a probability score based on the distribution of data, distribution of the classes within that subgroup.

So, and since we know the purity of each node, based on that, we can actually set a cutoff. You know, cutoff is a term that we already discussed in classification and classifier performance and keep that in mind, decision tree subgroups or decision tree terminal nodes will have a probability value. And say 0.8 and it is up to the analyst to decide whether a node with 0.8 should be used or not used. That is a cutoff, that will have an implication on, you know, sensitivity and specificity, etc, which we already discussed. So essentially, we are saying that decision tree induction is a model building activity, and then you test it and see how well the model performs during classification. That is the prediction phase.

Decision Tree

- ▶ A decision tree model consists of a set of rules for dividing a large heterogeneous population into smaller, more homogeneous groups with respect to a target variable
- ▶ This structure divides up a large collection of records into successively smaller sets of records with simple decision rules- resulting sets become more homogeneous
- ▶ Target variable is generally categorical, input variables could be any combination of categorical or metric

BUSINESS INTELLIGENCE & ANALYTICS

So, let me use text to explain to you what is a decision tree. A decision tree model consists of a set of rules, as I said, for dividing a large heterogeneous population into

smaller, more homogeneous group with respect to a target variable, with respect to an outcome variable.

This structure divides up a large collection of records into successively smaller set of records. So, decision tree building is a process, it splits and splits and splits, just like a tree grows and grows and grows, finally resulting in terminal node. And target variable is generally categorical. And if not categorical, you have to actually discretize the data, if it is continuous valued variable, discretize it and then you can use it as categorical variables. And the most common case is that of the binary outcome variable or binary target variable. And also the most common case is that of a binary tree, meaning every node splits into two and not more than two, that is called a binary split.

Rules (If... then) in Fraud detection at HSBC

- ▶ 3 claims in last 2 years
- ▶ Credit card used in different locations
- ▶ Credit card used at petrol station and then in high-value store!

BUSINESS INTELLIGENCE & ANALYTICS

Well, this is the case presented by a student when that student did a project at a organization and they extracted, the student and the team extracted certain rules based on the data set. The outcome variable was fraud, fraudulent credit card transaction. What

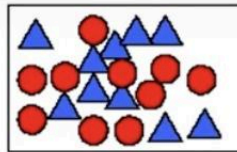
are the rules they identified? Of course, this is a case of imbalance data set and all that, but they identified a few rules, three claims in last two years. And I must put, these are rules and credit card used in different locations and credit card used at petrol station and then in a high value store. So, this is a condition that go together. Then probability of fraud is high. And of course, it leads to a more highly likely fraudulent transaction. That is a set of rules they identified from the data set. So, this is the practical example and also informs us that decision trees have practical use in fraud detection.

A case of internal fraud

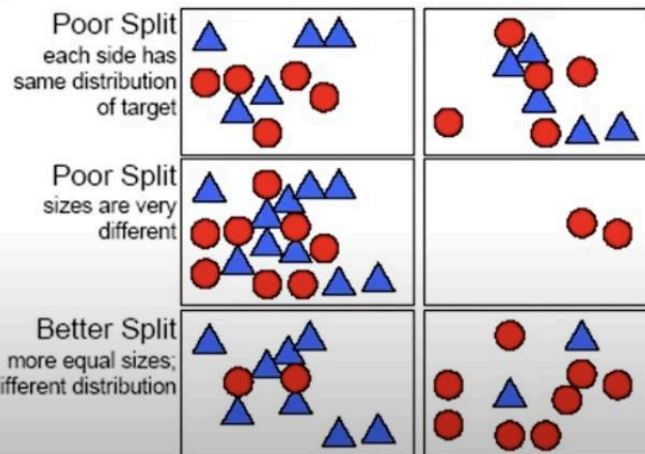
- A bank auditor found that the credit card balances written off as uncollectible had an excessive level of numbers with first two digits 24
 - The investigation found that \$2,500 was an internal write-off limit
 - One employee was responsible for most of the 24s by working with friends and having them apply for a card and then running up a balance to just below \$2,500. The employee then write the debt off.
 - The systematic nature of the fraud was evident from the first two digits
-

BUSINESS INTELLIGENCE & ANALYTICS

And this is another example about a bank, how they used decision tree to identify another fraudulent activity in the bank. And I share this with you. This is just an example that is mentioned in a book. But let us move on, growing decision trees. So, let us get into the algorithmic aspect of a decision tree, as to how does a decision tree grow. So as we have seen, a decision tree starts with a heterogeneous data set. And in this case, you know again, what is depicted in the data set is the value of the outcome variable, they are either blue triangles or red circles. Blue triangles or red circles.



Original Data



Now that is the original data set and you can see that it is mixed up, you know, both the records, both categories are mixed up in the original data set. Our aim is to split it and arrive at leaf terminals, which will be more homogeneous. So, some decision tree algorithm is applied and then you see it resulted in a split.

2, it is a binary split. So, this is split into 2. And these are the results, A and B are the results, A and B. Let me call this as A, this as B. In the first plate, you can see when you formed 2 A and B, 2 subsets, it is not a good split. Reason is that it also looks visually, it is very clear to us that it looks as heterogeneous as the original data, there is no improvement in homogeneity and therefore, it is a poor split. And the next split, if you look at one of the subgroups at least, this is very homogeneous, only red circles there.

Good, but not really good. What about the size? The resultant subgroup or the terminal node is very small. So, if you apply this rule, you will get a very homogeneous outcome or a subgroup, but that will be very few in numbers. And if you are actually applying decision tree algorithm for market segmentation or customer segmentation products, any activity, a segment which is very small in size is not attractive, it is not useful for action. And therefore, this is not practically useful and therefore, it is a poor split. And what is a better split? That is what is shown at the end, you can see that visually examining this, you know the split has reasonable size and also more homogeneous, it looks more

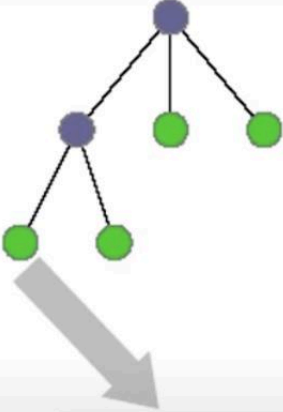
homogeneous.

The left node A has more blue triangles and the right node has more red circles and therefore, it is more homogeneous and also reasonable in size. So, this also gives us some basics as to how a tree should be induced or created. Let us move on to the next item. So as I said, a tree algorithm will split a root node and finally, the green leaves will be formed, leaves are the final results. And if you use a good software, which implements a decision tree algorithm, it will report the results, the final results.

NPTEL

DECISION TREES | BI&A | Prof. Saji K Mathew

The Leaves Contain the Scores



After the tree has been built, each leaf has a proportion of classes. The densest class is chosen as the classification for that node. Its probability is the density at the node. In this case, the leaf would give a score of 96.5% to NO.

YES	3.5%
NO	96.5%
Size	11,112

And it will qualify the leaf nodes with certain useful information. And what is that useful information? Number one, it gives the size of the node. And that is a very important information. And of course, size is something again, you can control. We will see what are the hyper parameters while building a decision tree, but size is reported and then what? The distribution of the two classes in the group. Please keep in mind that a subgroup or a node, particularly a leaf node will have a particular distribution of the different classes.

Here there are only two classes, yes and no. So yes is 3.5 percent in the subgroup and no is 96.5 percent or in terms of probability 0.035 is yes and 0.965 is no. And therefore, this is a no group. And if you want to target the no group, then this is a, this rule or this particular node is useful. One must actually focus on a group like this. That is the idea you get. So when the leaf nodes are reported with its underlying profile, like the size and the distribution and probabilities etc, then you get actionable information.