**NPTEL**

**NPTEL ONLINE CERTIFICATION COURSE**

**Introduction to Machine Learning**

**Decision Trees Tutorial**

Hello and welcome to this tutorial on decision trees in the preceding lectures we have looked at some of the theory behind decision trees, in this tutorial we will get some hands-on experience actually building trees using some of the concepts we have learned for building decision tree models with real data we will of course resort to packages such as Becca however in this tutorial we will be building trees for a very small data set in order to understand the process involved in building decision trees.

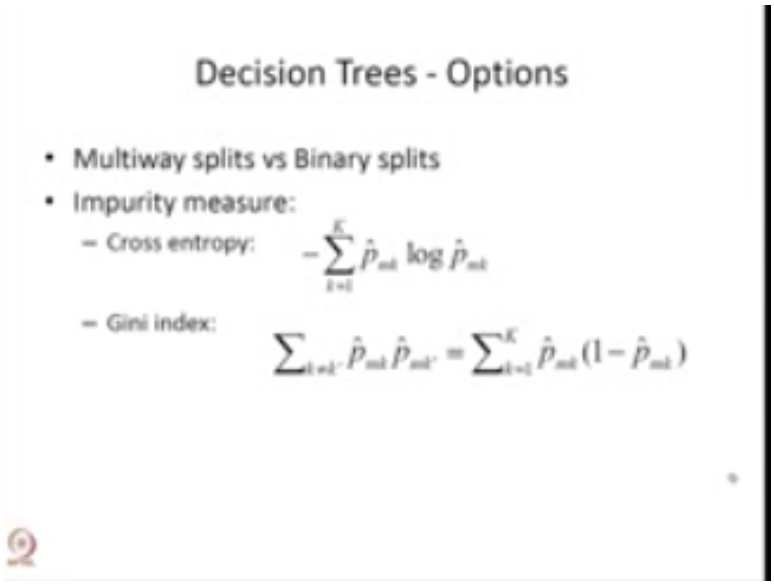
(Refer Slide Time: 00:51)

Data Set

| age | income | student | credit_rating | buys_computer (target) |
|---|---|---|---|---|
| youth | high | no | fair | no |
| youth | high | no | excellent | no |
| middle_aged | high | no | fair | yes |
| senior | medium | no | fair | yes |
| senior | low | yes | fair | yes |
| senior | low | yes | excellent | no |
| middle_aged | low | yes | excellent | yes |
| youth | medium | no | fair | no |
| youth | low | yes | fair | yes |
| senior | medium | yes | fair | yes |
| youth | medium | yes | excellent | yes |
| middle_aged | medium | no | excellent | yes |
| middle_aged | high | yes | fair | yes |
| senior | medium | no | excellent | no |

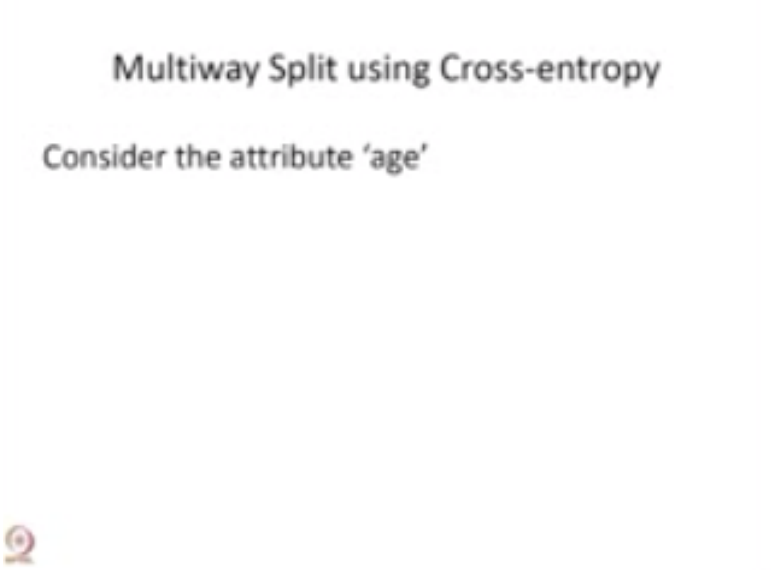
This is the dataset we will use for this exercise as you can see there are four different attributes with a binary valued target hat is bias computer note that the attributes age and income can take three different values whereas student and credit rating are binary valued.

(Refer Slide Time: 01:14)




In the theory lectures we have already seen the different options available to us when building binary trees for example the first thing we have to consider is the type of tree which we want to build we can either have binary trees or multi way trees depending upon the branching praetor at each node another option available to us is the impurity measure used in this tutorial we will be looking at two different impurity measures which are cross and trophy and the Gini index. Another option which we do not consider here is the pruning technique used.

(Refer Slide Time: 02:01)




To start with let us try and build a tree using multi-way splits and cross entropy as the impurity measure, the first thing that we have to do is to identify the root node this is done by considering each attribute in turn calculating the cross entropy value for that attribute and identifying the attribute which uses the lowest value let us start by considering the attribute age.

(Refer Slide Time: 02:32)

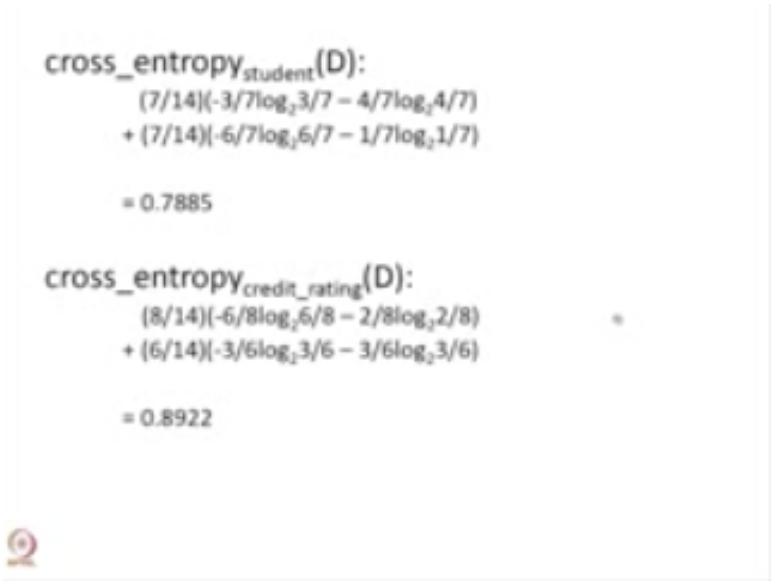
| age | income | student | credit_rating | buys_computer (target) |
|---|---|---|---|---|
| youth | high | no | fair | no |
| youth | high | no | excellent | no |
| middle_aged | high | no | fair | yes |
| senior | medium | no | fair | yes |
| senior | low | yes | fair | yes |
| senior | low | yes | excellent | no |
| middle_aged | low | yes | excellent | yes |
| youth | medium | no | fair | no |
| youth | low | yes | fair | yes |
| senior | medium | yes | fair | yes |
| youth | medium | yes | excellent | yes |
| middle_aged | medium | no | excellent | yes |
| middle_aged | high | yes | fair | yes |
| senior | medium | no | excellent | no |

5 points with age = youth; 2 with buys_computer = yes & 3 with buys_computer = no

From the table we observe that the attribute age can take on three distinct values which are youth middle-aged and senior, going back to the formula for cross and drop e we see that for each node we need the proportion of class k observations in that node in case of a two class problem such as the one we are considering we can use the simpler expression - p log p – 1- p log 1 - p where p is the proportion of observations for the positive clause since we need to calculate the cross entropy for an attribute with three distinct values we will have three components.

Let us first consider the value youth this is highlighted in the table we observe that out of the 14 different data points five observations have aged equals to youth among them two are belonging to the positive class and three belong to the negative class.

(Refer Slide Time: 03:51)

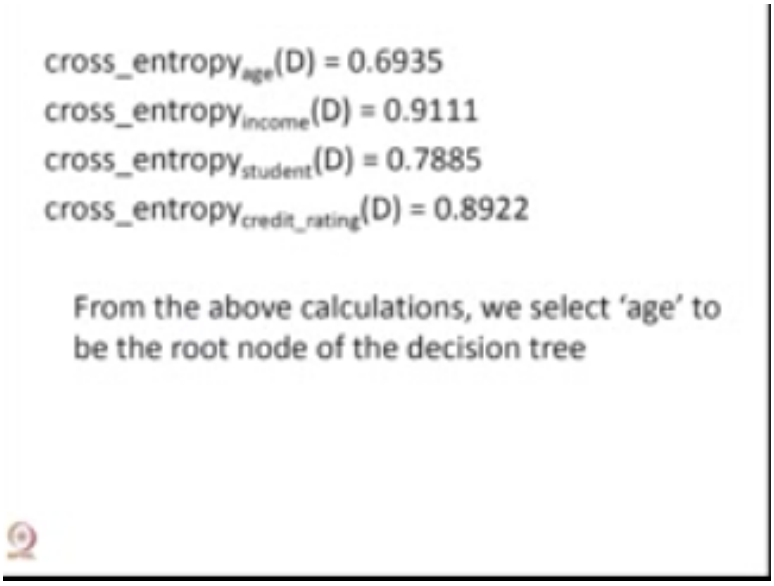



Using this information we have -2 / 5  log 2 by/ 5 that is the proportion of observation belonging to the positive class and - 3 / 5 log 3 / 5 for the negative class this expression is multiplied by the ratio 5 :14 which indicates which is a weight on the which is a normalizing factor since five out of the14 data points had aged equals to youth continuing with this manner we take up the next value that is age equals two middle-aged and observe that among the14 there are four points where age equal to middle-aged and for all of them buys computer equals two years that is they all belong to the positive class.

This gives us the second component as you can see we do not necessarily need to calculate this but we have put it there just for your reference the final component comes when we consider age is equals to senior again there are five points with age is equals to senior of them we observe that three belong to the positive class and to belong to the negative class putting it all together we get a value of cross and drop e note that all logarithms used here are using the base too.

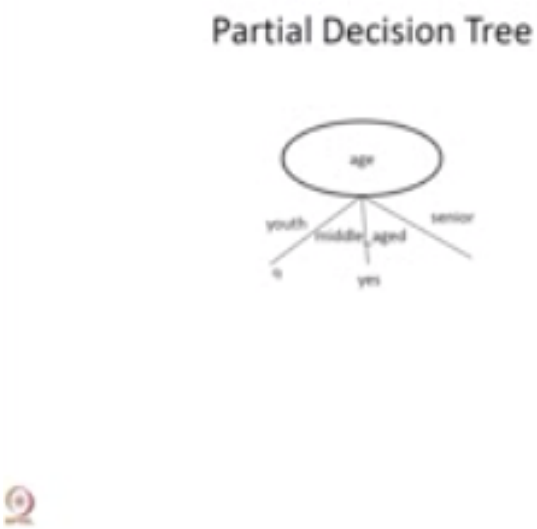
Make sure that you are able to follow the calculations especially how we were able to write down each of the components of the cross entropy expression using the same process we now consider the attribute income and find the cross entropy width next we find the cross entropy for the attribute student and a cross and trouble for the credit rating note that here we have only two components because both of these are binary valued.

(Refer Slide Time: 06:07)




Finally we compare each of the course entropy values and in this case observed that the attribute age gives the lowest cross entropy value and hence is the optimal AB tribute to use as the root of the our decision tree.

(Refer Slide Time: 06:24)

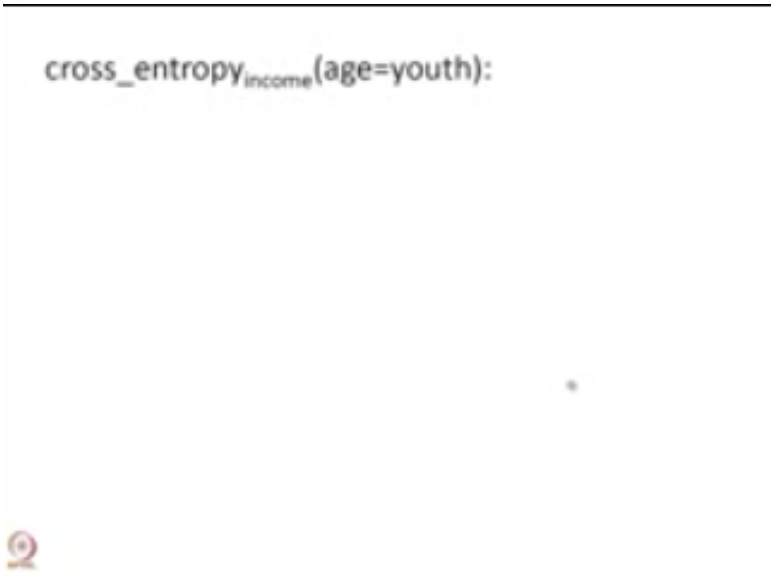



Thus we obtain the partial decision tree with age as the root attribute and three branches corresponding to the three distinct values that the attribute age can take note that the middle-aged that is the branch where a age equals two middle-aged has been labeled with yes indicating that this is a leaf node where any observation following along this branch will be labeled yes this is because if we go back to the table we observe that when age equals two middle-aged buys computer equals two yes.

Thus along this branch of the tree there is no need to further grow the tree since from the training data given to us we can directly conclude that if we observe age to be middle-aged then we can label the class and the observation as positive that is the person will buy a computer. Now we have created this partial decision tree so how do we proceed? Essentially it is a recursive process we started at the root node we were able to find the root node to be the attribute age.

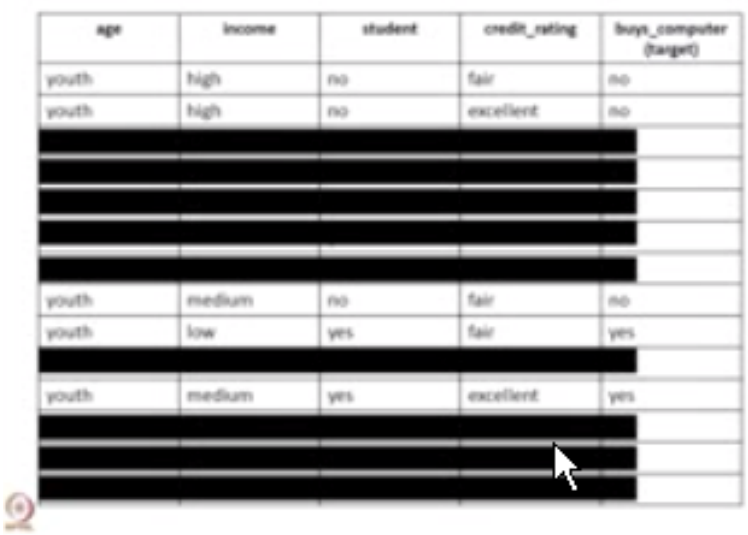
Now along each of the remaining branches where we have not found the note to be a leaf node we have to repeat the same process. So let us first look at the branch is equals to youth we have already considered the attribute age, so there are three attributes left to us using a process similar to what we have just seen we try to identify the best attribute to use at this position.

(Refer Slide Time: 08:29)




So we consider the cross entropy of income where age equals to youth, now we are not now we will not be considering the entire data set but will consider the restricted data set where age equals to youth.

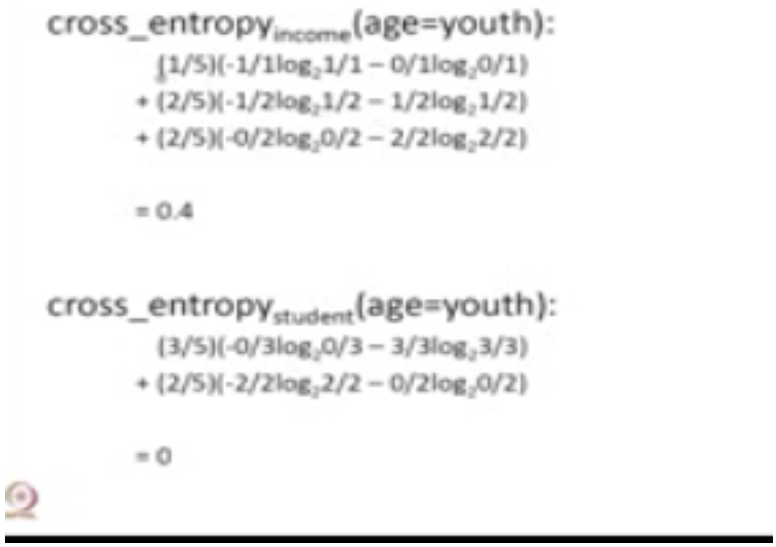
(Refer Slide Time: 08:42)



| age | income | student | credit_rating | buys_computer (target) |
|---|---|---|---|---|
| youth | high | no | fair | no |
| youth | high | no | excellent | no |
| youth | medium | no | fair | no |
| youth | low | yes | fair | yes |
| youth | medium | yes | excellent | yes |

This is illustrated in this table where we have crossed out all observations where age is not youth so essentially we repeat the same entire process with this restricted data set note that the attribute age has already been considered so we are left with the remain three attributes and these are the values that are to be considered.

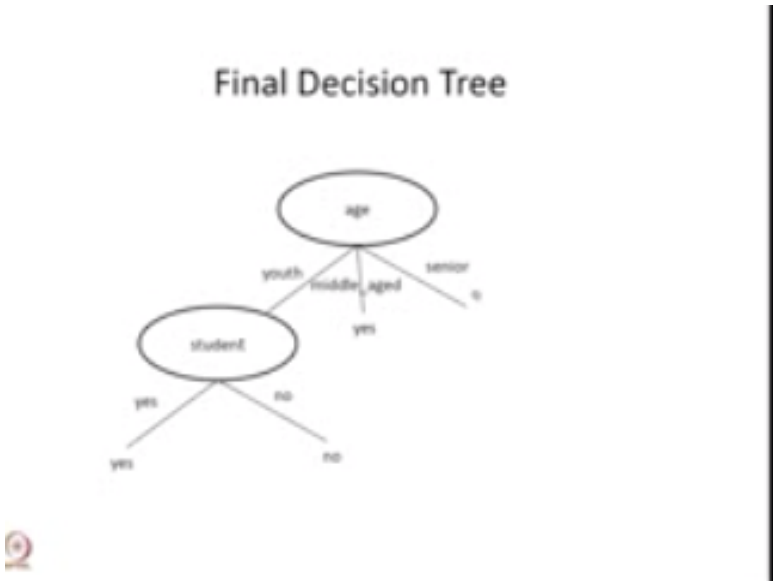
(Refer Slide Time: 09:11)




Thus we have cross entropy of income when age equals to youth you can go back and verify that these are the values you will obtain next we have cross entropy of student when they is equal to youth here we observe that the cross entropy is actually 0 going back to the table we see that in a equals to youth and student is no bias computer is no and when student is yes buy computers yes so this leads us to a pure leaf we can go ahead and calculate the cross entropy for credit rating as well when age is equals to youth but since we will not get a value less than 0.

We can stop the process here and get the partial tree where we have selected the attribute student with the least value of cross entropy.

(Refer Slide Time: 10:07)




As you can see we have labeled the branches yes and no because these are leaf nodes which are pure there is no mixture, now as you can see we have this branch this branch in this branch are all leaf node so last the remaining at branch to consider is when age is equals to senior again we look at the table where we disregard all observations where it is not senior and follow the same calculations.

(Refer Slide Time: 10:45)

$$\text{cross\_entropy}_{\text{income}}(\text{age=senior}):$$
$$(2/5)(-1/2\log_2 1/2 - 1/2\log_2 1/2)$$
$$+ (3/5)(-2/3\log_2 2/3 - 1/3\log_2 1/3)$$
$$= 0.9510$$

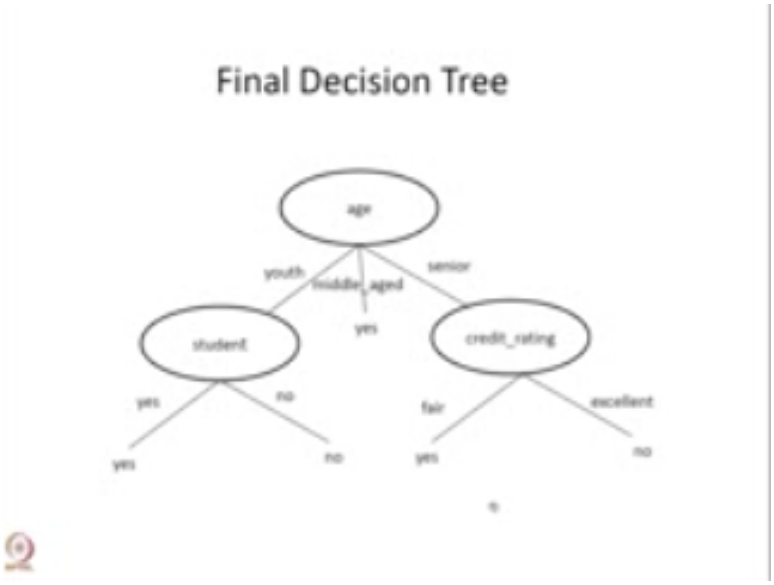
$$\text{cross\_entropy}_{\text{student}}(\text{age=senior}):$$
$$(2/5)(-1/2\log_2 1/2 - 1/2\log_2 1/2)$$
$$+ (3/5)(-2/3\log_2 2/3 - 1/3\log_2 1/3)$$
$$= 0.9510$$

We get cross entropy and we look at income when h is equal to senior and cross entropy of student when age equals to senior.

(Refer Slide Time: 10:55)


And cross interview of credit rating when is equal to senior, here again we find a cross entropic value of 0 which is the minimum and if we go back to the table wherever credit rating is fair we have bias computer equals two years here as well and whenever credit rating is excellent you have by some high school to no. So this allows us to create a decision tree.

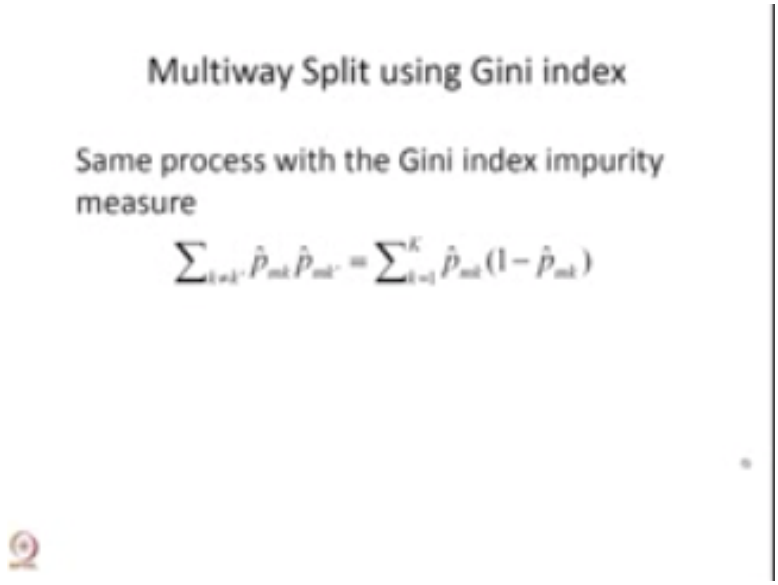
(Refer Slide Time: 11:21)




Where each of the leaf node is a pure nod in case we did not get a cross entropy value of zero let us say we have a different value cross entropy here we would again continue the process and the last situation is when we have exhausted all attributes available to us and we still do not have a pure leaf what do we do then essentially let us say when we follow this branch there were five points of which three were positive and two were negative then this would have been labeled yes.

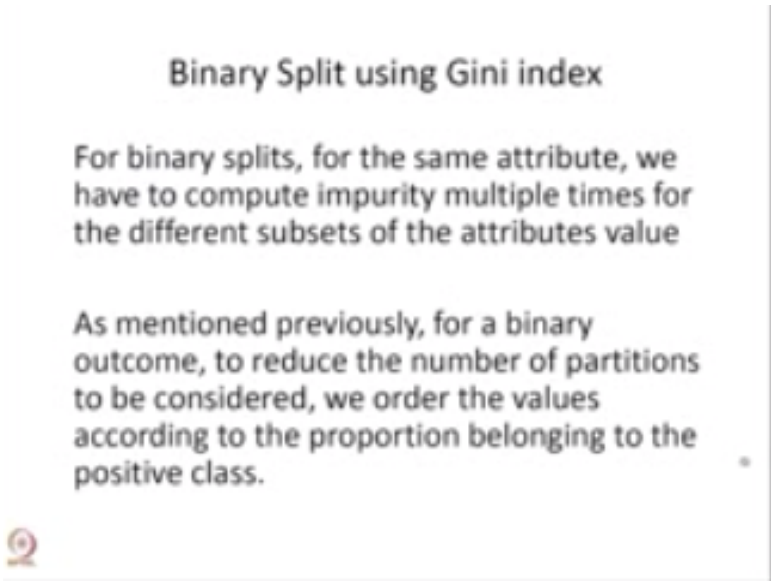
Because the majority of the data points have a positive Cubs have a positive belong to the positive class fortunately for us in this example we have obtained all leaf nodes as pure but this will always this will not always be the case.

(Refer Slide Time: 12:24)




Next we will look at decision trees using multivariate and the Gini index, essentially the process is the same except that we replace across entropy measure with the Gini index impurity measure this will be left as an exercise.

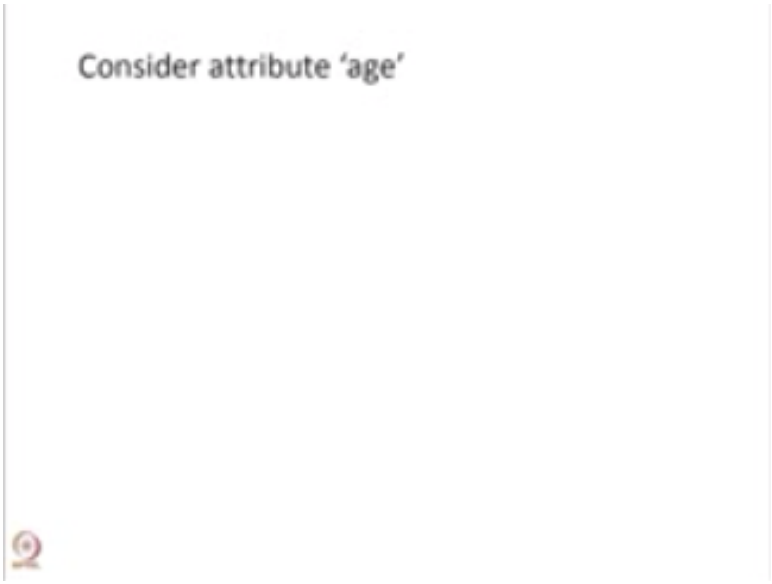
(Refer Slide Time: 12:41)




Now the other type of tree that can be built is a binary tree for this exercise we will look at using the Gini index impurity measure recall that when we were creating multi-way trees at and to select an attribute for a node we have to consider each attribute only once, however in the case of binary trees since for each attribute there may be different subsets to consider that is where to split.

We may have to look at attributes multiple times also as was mentioned in the theory lectures in case of a binary outcome we can reduce the number of partitions that have to be considered by ordering the values according the proportion belonging to the positive class since our data set has binary valued outcome we will see how this process works.

(Refer Slide Time: 13:41)




Again we start by considering the attribute age.

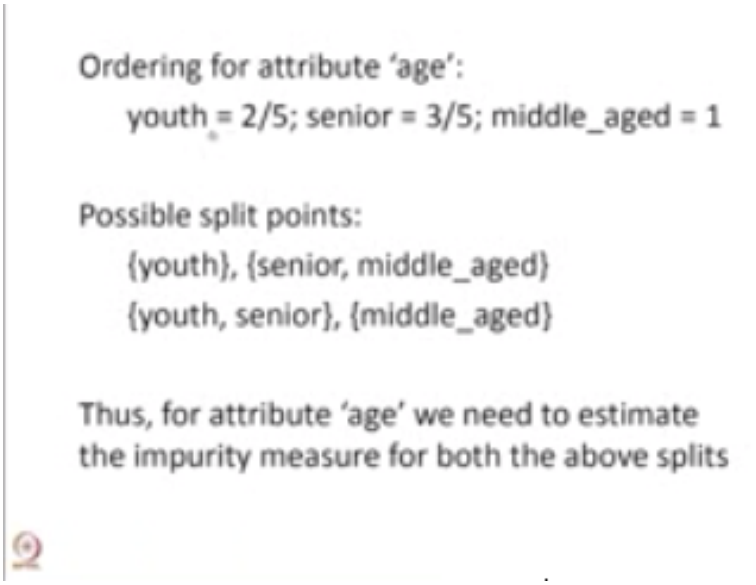
(Refer Slide Time: 13:44)

| age | income | student | credit_rating | buys_computer (target) |
|---|---|---|---|---|
| youth | high | no | fair | no |
| youth | high | no | excellent | no |
| middle_aged | high | no | fair | yes |
| senior | medium | no | fair | yes |
| senior | low | yes | fair | yes |
| senior | low | yes | excellent | no |
| middle_aged | low | yes | excellent | yes |
| youth | medium | no | fair | no |
| youth | low | yes | fair | yes |
| senior | medium | yes | fair | yes |
| youth | medium | yes | excellent | yes |
| middle_aged | medium | no | excellent | yes |
| middle_aged | high | yes | fair | yes |
| senior | medium | no | excellent | no |

Positive class proportions: youth: 2/5; middle_aged: 1; senior: 3/5

As we can see from the table H can take three distinct values, now we look at the positive class proportions for each of the values we see that when age is equals to youth two out of five times two out of five observations belong to the positive class age is equal to middle-aged each of the four observations belong to the positive class and when a is equals to senior three out of five observations belong to the positive class.

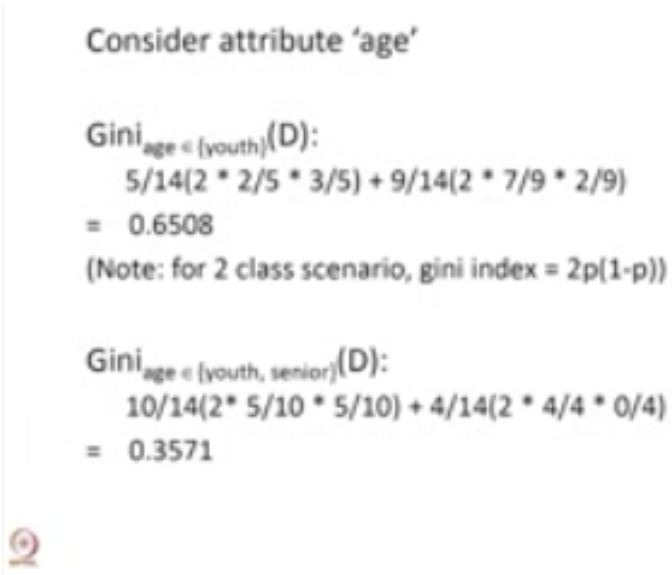
(Refer Slide Time: 14:21)




Thus we have an out ordering for the attribute H youth senior and middle aged, what this essentially means is that we have two possible spit points youth and youth along one branch and senior and middle age on the other or youth and senior along one branch and middle-aged along the other note that the attribute age actually has a notion of order that is youth middle-aged and senior.

If we want to retain that notion of order then we would only consider the split points where youth is along one branch and the rest the other two are along the other branch or youth and middle-aged is along the one branch and the seniors along the other we have considered the attitude age unordered here to illustrate how you would go about ordering values for the rest of this exercise we will use the specific ordering for the attribute age, now that we have identified the possible split points.

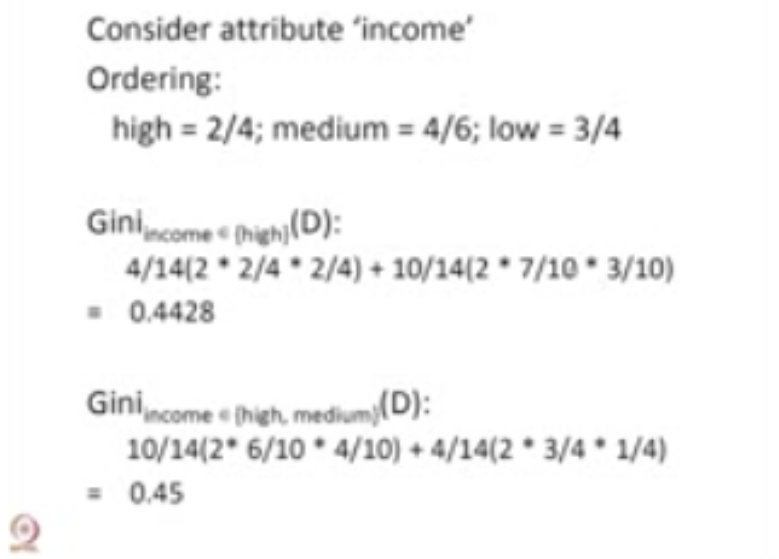
(Refer Slide Time: 15:39)




We estimate the impurity measure here using Gini index for both possibilities note that since we are in a two class scenario we use the simplified formula of the Gini index equals to 2 P x 1 - P where P is the proportion of observations belonging to the positive class. Go back to the table and verify that these are the values obtained we see that when we calculate Gini index where we are considering the split where youth is in one branch and the remaining two are on the other branch we get a value of 0.6508.

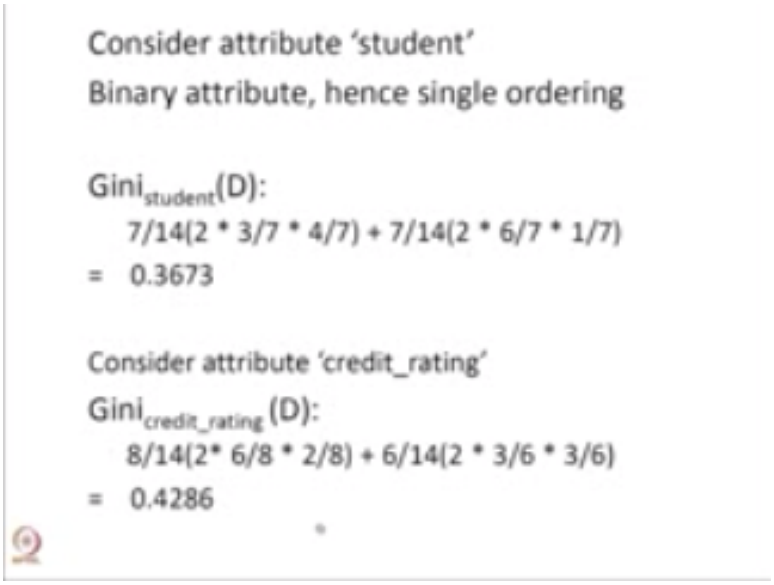
And where we consider the alternate split where youth and senior belong in one branch and middle-aged it belongs to the other we get a lower value of 0.3571 thus among the among these two this is the split that will be preferred. Now this calculation is just for the attribute age we need to repeat the same process for the remaining three attributes.

(Refer Slide Time: 16:48)



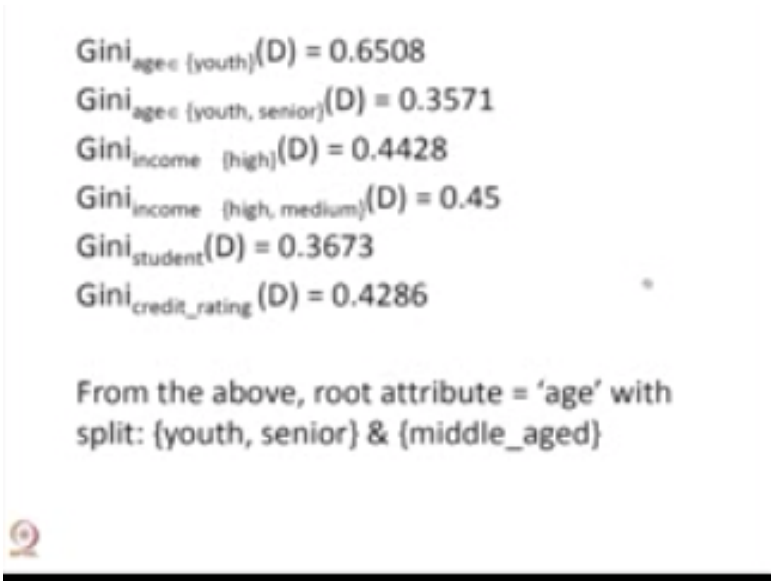

Thus we have the attribute income, going back to the table we will see that this is the ordering for the three values that the attitude income can take and the corresponding Gini index values.

(Refer Slide Time: 17:04)




And next we have the translations for the attribute student and credit rating here there is only a single possible ordering because both of these are binary valued attributes.

(Refer Slide Time: 17:17)




Finally we compare all the results and we observe that the attribute age where the split is with youth and senior along one side and middle later on the other is the optimal gives the optimal value and thus we select this particular attribute with this particular split point as the root.

(Refer Slide Time: 17:43)


Thus we create this partial T three entry again from our previous exercise we know that when age is equals to middle-aged all our observations are positive, so we do not need to grow the tree beyond this thus our not now we focus on the branch where equals to youth or senior is.

(Refer Slide Time: 18:05)

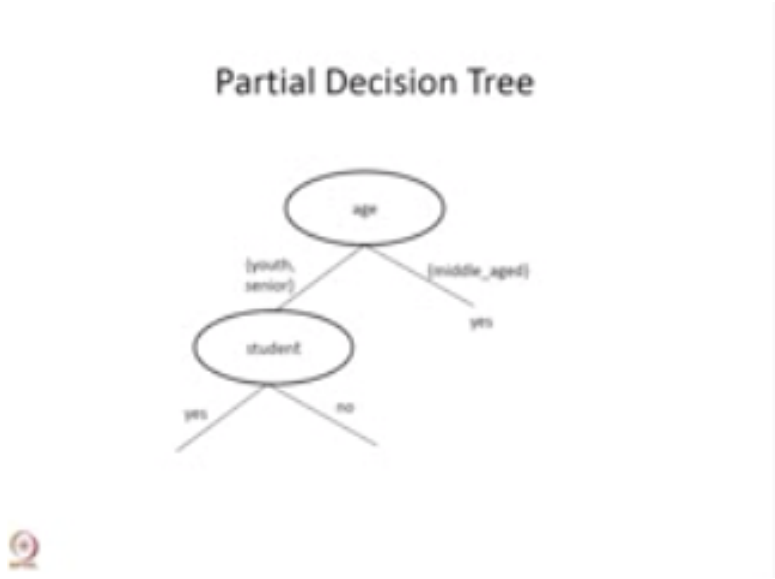
| age | income | student | credit_rating | buys_computer (target) |
|---|---|---|---|---|
| youth | high | no | fair | no |
| youth | high | no | excellent | no |
| [redacted] | | | | |
| senior | medium | no | fair | yes |
| senior | low | yes | fair | yes |
| senior | low | yes | excellent | no |
| [redacted] | | | | |
| youth | medium | no | fair | no |
| youth | low | yes | fair | yes |
| senior | medium | yes | fair | yes |
| youth | medium | yes | excellent | yes |
| [redacted] | | | | |
| [redacted] | | | | |
| senior | medium | no | excellent | no |

Ordering (income): high = 0/2; medium = 3/5; low = 2/3

So again from previous experience we know that to create and to grow the tree along this branch we need to consider only observations where ages youth or senior thus we have disregarded all observations where you age equals to middle-aged. Now we repeat the same process we have already consumed the attribute age we have three remaining attributes of these income has three distinct values so we need to identify the optimal split point that is we need to first consider the ordering shown here for student and credit rating both are binary values.

So there is only going to be in one straight point so if you repeat the calculations for this subset of the data set.

(Refer Slide Time: 19:02)




We will identify that the next node should use the attribute student however this is not the end since we do not obtain pure nodes here and this process has to be continued again we will leave this as an exercise, hopefully this tutorial would have clarified some of the concepts that we came across in the theory lectures and helped you in understanding how decision trees are created of course for real-world data as well as the programming assignments that will be released we will be using a tool such as vacca where you have lot more options for example pruning which we were unable to cover in this short tutorial for any doubts regarding any of the concepts covered here please use the forums.

**IIT Madras production**

Funded by
Department of Higher Education
Ministry of Human Resource Development
Government of India

www.nptel.ac.in