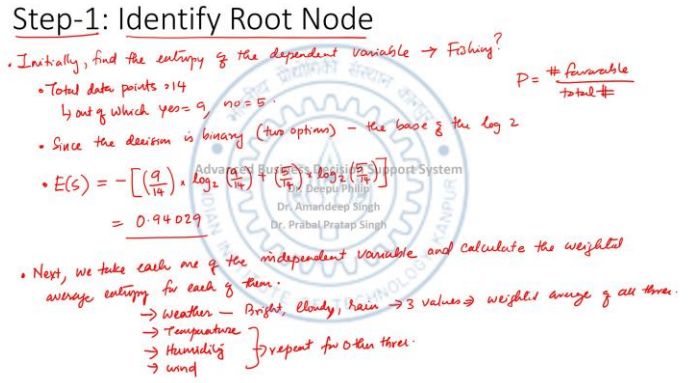
Advanced Business Decision Support Systems

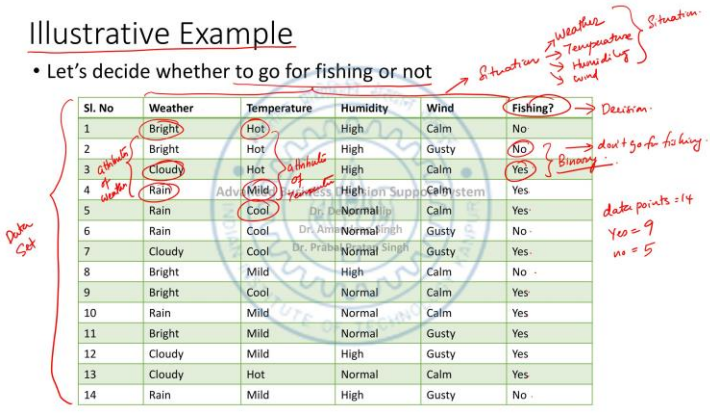
Professor Deepu Philip
Department of Industrial Engineering and Management Engineering
Indian Institute of Technology, Kanpur
Professor Amandeep Singh
Imagineering Laboratory
Dr. Prabal Pratap Singh
Indian Institute of Technology, Kanpur

Lecture 14

Decision Tree Algorithm for Business Decision (Part 2 of 3)

Good afternoon, everyone. Welcome to the fourth week lecture of the Business Decision Support System, the advanced course of the Web-Based Decision Support System under the NPTEL MOOC's program from IIT Kanpur. I am Dr. Deepu Philip and along with me Dr. Amandeep Singh Oberoi and Dr. Prabal Pratap Singh are teaching this course. So, now let us get back to this example and let us try to see whether you know we can start building the decision tree.





So, the first step is to identify the root node and we know root node is, where the tree starts. So, the first step is you initially find the entropy of the dependent variable . What is a dependent variable? Is the fishing.

Whether to go for fishing or not, that is the dependent variable. So, if you look into it, there are 14 data points all together and of which, you have, let us count data points equals 14 . How many yeses? How many no's? So, there are 5 numbers, so that should be 9 yes. So, let us write that down here total data points is equal to 14 out of which yes is equal to 9 and no is equal to 5. So that is one part. So, we have already calculated that right.

Then, since the decision is binary, the reason we call it as binary is there are only 2 options. The base of the log is 2 where is the log? Log I told you earlier, there is this log, we are talking about this log. Log is 2 right.

So, hence, plugging in this value,

$$E(s) = -\left[\left(\frac{9}{14}\right) * log2\ \left(\frac{9}{14}\right) + \left(\frac{5}{14}\right) * log2\ \left(\frac{5}{14}\right)\right]$$

How do we calculate this? What is the one way to calculate this? So, let me quickly show you, how do we calculate that. We use an excel sheet. So, it will make your life easy. So, if you look into this, so this one is, if you see it is written as $\frac{9}{14}$. So, I calculated first $\frac{9}{14}$, this is $\frac{5}{14}$.

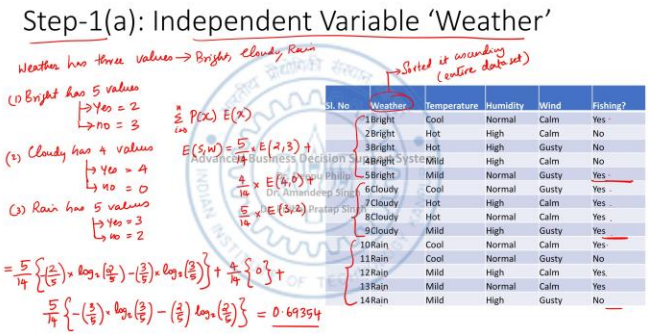
So, both of them are calculated $\frac{9}{14}$ and $\frac{5}{14}$ are calculated here . So, then the next thing to do is, I can say equals there is a function called a log in excel log and you put the number here and my comma and you give the base is 2, put that and that is this number minus 6.473. I have already calculated that, but I am just showing you how it is done. And, same way, if you say, you calculate log using this number n, 2. So, these are the 2 logs.

Now, what you need to do is, you need to multiply this fraction with that of the log. So, that is equal to this fraction, this log value multiplied by this fashion, that is minus, I have already done these numbers here and you can pull it down. So, instead of pulling it down, I will just show you so that, you will not forget multiply this log with this number and this. So, now you have 2 of these numbers.

So, now to calculate your $E(x_i)$ you just need to do is, you need to sum them equals this plus this . So, you will get minus 94021, that is the so I am just going to delete these so because it just would confuse you. So, the 94029 minus and there is a minus sign outside, so you get a positive value. So, going back to this, we can see that, the summation inside was minus. So, we will get it as 0.94029.

If you want to round it off, you can round it out to 0.94 but let us use the whole number what it is . So, first we calculated the entropy of the dependent variable. So, you have now probably clear with how we did that. Then what do we do? First, we take each one of the independent variable and calculate the weighted average entropy for each of them.

So, you have 4 independent variable. You will take it as the weather is the first one then, you do, I think it was temperature, I believe humidity and wind temperature. So, in weather, we have already seen from the previous data, there is bright cloudy and rain. So, it has bright, cloudy and rain . So, since it has 3 values or 3 attributes, weighted average of all 3 is what we will end up calculating. Then, we repeat it for other 3 . So, that is what we will do as part of this.



So, now let us do the first one, the step 1. I am just calling it as 1(a) the independent variable is weather. So, if you look at the data set, now we are focusing on weather.

So, what we do here is, you look at weather first thing I did is, I take the data set and sorted it ascending entire data set, just do not sort the weather a lot sort the entire data set. So, I have now what you called as I said earlier, weather has 3 values. So, 3 values are bright cloudy and rain, here is bright, here is cloudy and here is rain. So, bright has how many values? It has 5 values. So, bright has 5 values until here, that is bright of which yes is equal to and no is equal to .

So, yes is 2 and 3 no and same way, so next is cloudy. Cloudy has 4 values and how many yes and no. So, cloudy has 4 yes and 0 no, there is no noes as part of it. Then, the third one is rain. Rain has 5 values yes equal to and no is equal to. So, how many yes are there 3 yes and 2 no.

So, this is the split of cloudy and this is the this one. So, you have we have counted that. So, now then, based on our equation, look into this, this is our equation the weighted

average. So, we have to do the probability of X and then, you multiply, so the weighted average is this one multiplied by the E (x).

So, the Information Gain. So, the equation here is,

$$\sum_{i=0}^{n} P(x)\, E(x)$$

So, if I write that down then the expected value or the weighted average entropy of the weather from the with respect to the decision, will be the first one is the bright .

$$E(S,W) = \frac{5}{14} * E(2,3) + \frac{4}{14} * E(4,0) + \frac{5}{14} * E(3,2)$$

So, this will be your weighted average entropy of 2, 3. So, now how do you calculate this one? This is an interesting calculation and we will use excel to do this, but I will expand this equation .

So, the first one is 5 by 14, then within braces, now we need to expand the E(2, 3) correct. So, E(2, 3) is 2 out of 5, so that will be the first one will be 2 out of 5, there is a minus sign 2 by 5 and why is there a minus sign because this is the equation.

$$\frac{5}{14}\left\{-\left(\frac{2}{5}\right) * log2\ \left(\frac{2}{5}\right) - \left(\frac{3}{5}\right) * log2\ \left(\frac{3}{5}\right)\right\} + \frac{4}{14}\{0\} + \frac{5}{4}\left\{-\left(\frac{3}{5}\right) * log2\ \left(\frac{3}{5}\right) log2\ \left(\frac{2}{5}\right)\right\}$$

So, how do you calculate this now? So, again you would need excel because this log calculation and if you have a scientific calculator, you can do this also, but the easiest way to do it is excel. So, I have done that, here so temperature the first one is kind of did it in a weird fashion.

I have written it some place here; we can freshly calculate it. The easiest way to do that would be the first one is equals $\frac{5}{14}$ that is the first one, then the second one will be $\frac{4}{14}$ then the third one is equals $\frac{5}{14}$. So, these are the 3 ones and then the next thing you can do is, you have to calculate 3 by 5 and the log of that value. So, this is the initial. So, I am giving it a space that is equals $\frac{2}{5}$ and the second option is this is $\frac{2}{5}$ and the next is $\frac{3}{5}$

Then, you have is the log of calculate, the log of this one for the base 2 and same is with the log of the second number. You got 2 logs here correct. Now, what you do is, you multiply these 2 logs equals this multiplied by this and you get both of those logs calculated out. So, that is the initial value of this.

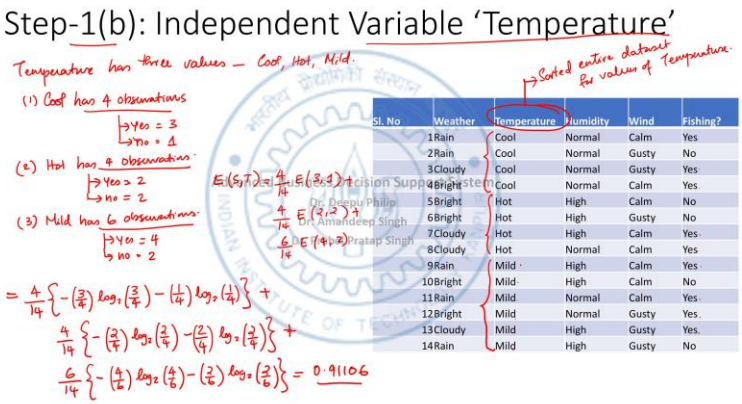
Then, what you do is, you sum them here. So, I am just putting the equals this plus this, that is the sum . So, this is the thing that is underneath this 2 by log and 3 by 5 . Now, that need to be multiplied by 3 by 5. So, the product of that will be equals this number multiplied by the first one is 5 by 14 .

So, that will be your first value . Now, the second one is 0. So, we already done the 0 part, so second is equal to 0, so we just multiply this with this should give you 0 . And, then the third set up for this one is, what I am meaning is, this 5 by 14 part. So, that is 3 by 5 and 2 by 5.

So, that is equals 3 by 5 equals 2 by 5 . So, that is the first 2 values then, we need to find the log of them equals log. This number comma 2 got both logs . So, the fraction and the log now need to be multiplied equals this multiplied by this and same thing both of them are there. So, now you do is you do plus, this the sum both the minuses will get summed up. So, then the third one, that is inside the parenthesis is fraction.

So, this need to be multiplied with this weight, weight is this one. So, you have both of those. Now, what happens is, you have now calculated this value then, the 0, 4 by 14 and the 5 by 14, all these 3 are, so you sum them equals this plus this plus, all of that is minus 0.69354 and there is a negative sign outside, we all knew that because if you look at the equation, there is a negative outside. So, using that, this value will become the positive value.

So, the value will then become equal, so if you look at the excel sheet, you have what you call as 0.69354 . So, that is the value. So, it is 0.69354, this is the value of the weighted entropy of the independent variable weather. I hope that you guys could understand this calculation clearly.



Now, let us do the next one which is the independent variable step 1(b). Now, we do that for the temperature. So, now what I have done is, if you look into this, I sorted the entire

data set for values of temperature. So, the temperature has also same way that is cool, hot and mild.

So, temperature has 3 values that are the cool, hot and mild . So, how many observations you have cool? This is cool, this much is hot, rest of it is mild. So, cool has 4 observations. In this 4 observations, S is equal to and no is equal to, how many S? So, in the cool, it is 3 S and 1 no, then same way. We have is hot having 4 observations.

Of these 4 observations, we have S and no. So, S will be 2 and 2 no. The third one is mild; can I do mild? Mild has how many observations 6 observations. Of this there is S and no, so S is 4 that means, there are 2 no. So, now how do we calculate this as we did in the previous one, we do is the temperature. Same way, we did for weather, we do it for temperature, that is the entropy.

$$E(S,T) = \frac{4}{14}E(3,1) + \frac{4}{14}E(2,2) + \frac{5}{14}E(4,2)$$

How do you calculate?

$$\frac{4}{14}\left\{-\left(\frac{3}{4}\right)log2\left(\frac{3}{4}\right)-\left(\frac{1}{4}\right)log2\left(\frac{1}{4}\right)\right\}+\frac{4}{14}\left\{-\left(\frac{2}{4}\right)log2\left(\frac{2}{4}\right)-\left(\frac{2}{4}\right)log2\left(\frac{2}{4}\right)\right\}$$
$$+\frac{6}{14}\left\{-\left(\frac{4}{6}\right)log2\left(\frac{4}{6}\right)-\left(\frac{2}{6}\right)log2\left(\frac{2}{6}\right)\right\}$$

So, again we go back to excel. We keep this notes here and the way to do it is, we can read and the calculations here I think, I have done it earlier. Here also, so this is 4 by 14, 6 by 14, so I done the calculation here.

So just for you, make your life easy, so we will run through the calculation. So, first one is, we need to do these all weights, the 3 weights get calculated out, that is 4 by 14, 4 by 14, 6 by 14, all the 3 weights get calculated out then, what we calculate is the fraction inside the first one, the 3 by 4 this one so that is 3 by 4 and 1 by 4. So, that is the first value is 3 by 4. The next value is 1 by 4. Why did I do 2 by 4? I did it in the other fashion, the raw format, so you may get confused.

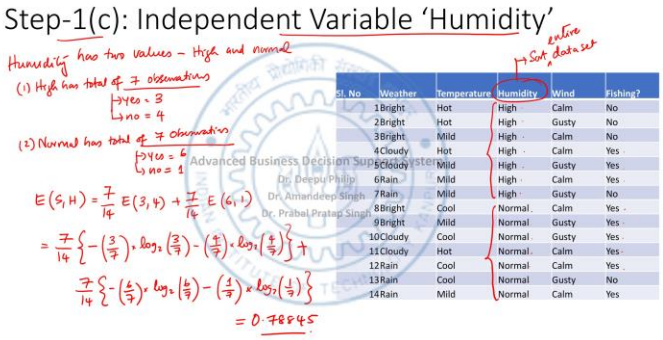
So, let us repeat the same process so that, you do not get confused. So, we will do it here forget about that other thing that is equals, the first one is let us first calculate the weights equals 4 by 14, equals 6 by 14, so 3 weights get calculated out. Then, first what we do is, we calculate the inside, the first part 3 by 4 and 1 by 4. So, the first one is equals 3 by 4 and equals 1 by 4. So, both of them are there then, what you need to do is, you need to take the log of this equals log of number comma 2 and same thing gets repeated to next of them.

Now, what we do is, we multiply them, this multiplied by that value. The probability and the value get multiplied. You get two of those values, both of them are there. So, now what you do is, you sum them equals this plus this, you get that. So, this sum value is the thing within the curly bracket now, what you need to do is, you need to multiply that with the help of equals this multiplied by the weight.

So, that gives you the first coefficient the second one is, we look into this that is 2 by 4 and 2 by 4. So, what we do is, equals 2 by 4 Now, we do is log of both of them is taken for the base 2, take that then what we do is, we take the product of them, this multiplied by product. So, you get the two numbers.

What you do is, you sum them this plus this. You sum that values. So, you get that, so this one what you do is now, this weight, which is the value inside this, we did that you multiply that with the weight. That weight is, so that will be this multiplied by the second weight. You get that value here and the third one is, what you need to calculate is, the stuff inside 4 by 6 and 2 by 6, so that is equals 4 by 6 equals 2 by 6. So, now you need to find the log of these numbers comma 2 and both the logs gets calculated. Next one is you take the product of these numbers multiplied by the product.

So, you get both these products then what you do is, you sum them equals this plus this, that is your product so now, the weight you multiplied by the weight equals, this multiplied by the weight. So, all the three values we have calculated. So, then that is equal to sum them, this plus all the three values 0.91106, that is what you get the value here and since, there is a minus sign outside, we all remember that minus makes this plus so, 0.91106, so we can say that as 0.91106. So, that is the weighted average entropy of the independent variable temperature.



Now, continuing the same fashion, we move to what we call as the Independent Variable 'Humidity'. So, now humidity, the things are slightly different because when you look into it, now again, what I did is sort entire data set and for this, you get the high to be this

many and normal to be this many. So, the humidity has values high and normal by one high has total of 7 observations. There are 7 observations, let us take yes how many no? how many the yes? 3 yes and that means, there will be 4 no same way, normal has total of 7 observations how many yes? how many no? Yes is 6 yes and 1 no.

$$E(S,H) = \frac{7}{14}E(3,4) + \frac{7}{14}E(6,1)$$

So, now how do we expand this?

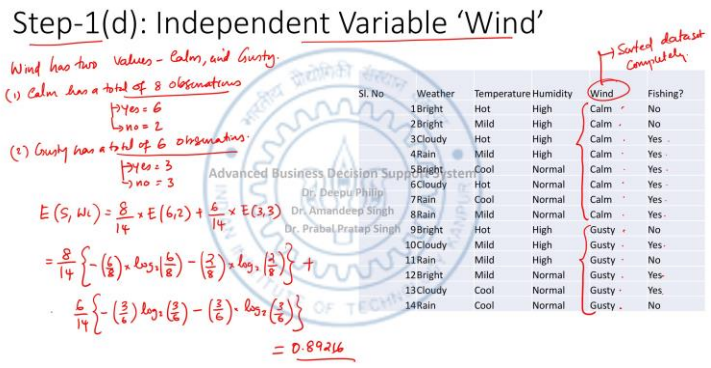
$$\frac{7}{14}\left\{-\left(\frac{3}{7}\right) * log2\left(\frac{3}{7}\right) - \left(\frac{4}{7}\right) * log2\left(\frac{4}{7}\right)\right\} + \frac{7}{14}\left\{-\left(\frac{6}{7}\right) * log2\left(\frac{6}{7}\right) - \left(\frac{1}{7}\right) * log2\left(\frac{1}{7}\right)\right\}$$

So, how do we do that? We can calculate that again just I will refresh the calculation so that, you have your idea here. So, the coefficients we need to calculate are 7 by 14, so that 2, 7 by 14. So, that is equals 7 by 14 equals 7 by 14. Two of them and then the interior values are 3 by 7 and 4 by 7 for the first one.

So, equals 3 by 7 equals 4 by 7, now the first term is, if you remember 3 by 7 and the log of that and then product of that, so that equals the log of this number to the base of two and then, this we push it this way and we do is, take the product of these both products are done then, what we do is, we sum them, this plus this and that value gets weighted. So, that weight part is equals this number multiplied by the weight that we already calculated out that is the first part. Then, the second part is, we know that the fraction is 6 by 7 and 1 by 7, the second 7 by 14 that is 6 by 7 and 1 by 7. So, we do is, 6 by 7 equals 1 by 7, two numbers we calculate the log of them to the base 2, 2, then both of them need to be multiplied. So, we take this product of the log and the probability in both of these values, now what we do is, we sum them equals this plus this.

Now, what happens is, equals this times, The second weight these two so what we do is, we take this value plus this value gives you 0.78845, then minus sign outside, so we will have that value.

So, 0.78845, so that will be 0.78845. That will be the entropy of the humidity, so I think now, you have seen, how we calculate the weather which is 3 output temperature, which is 3 and humidity, which is 2.



We will do the last one which is a step d or 1(d) which is the Independent Variable 'Wind' and if you see again, I have taken the wind and sorted data set completely and you can see that, there is two values of it, one is calm and other is gusty. So, we can say that, wind has two values, Calm and Gusty. So, number one Calm has a total of how many? 8 observations of this, yes is equal to no is equal to, S is equal to 6. So, there are 6 S and 2 no. Number two, Gusty has a total of 6 observations of which yes no, 3 S and 3 no.

So, now what we need to do is, we need to do the same thing which entropy of wind.

$$E(S, Wi) = \frac{8}{14} * E(6,2) + \frac{6}{14} * E(3,3)$$

So, if we expand this,

$$\frac{8}{14}\left\{-\left(\frac{6}{8}\right) * log2\left(\frac{6}{8}\right) - \left(\frac{2}{8}\right) * log2\left(\frac{2}{8}\right)\right\} + \frac{6}{14}\left\{-\left(\frac{3}{6}\right) * log2\left(\frac{3}{6}\right) - \left(\frac{3}{6}\right) * log2\left(\frac{3}{6}\right)\right\}$$

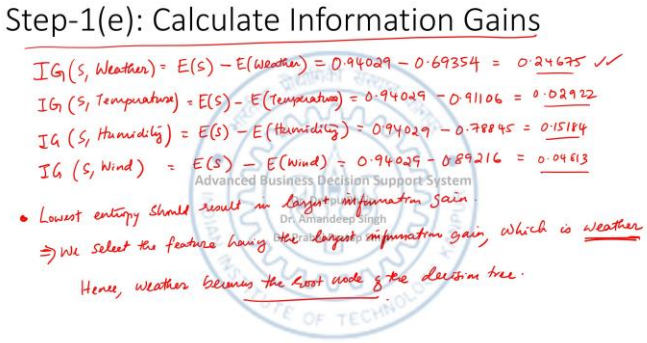
So, again, we do the calculation quickly. So, that you have an idea that in the next steps, we can expedite the calculation. So, first we calculate the weights are 8 by 14 and 6 by 14. So, that is equals 8 by 14, you have 6 by 14, two of those. Now, the first parenthesis you are looking at 6 by 8 and 2 by 8, so we calculate that equals 6 by 8 equals 2 by 8.

Now, as we know, we calculate the log of those because that is the probability and the log of the probability, both of those and then, we multiply the probability and the log of its probability to the base to get these two numbers, then what we do is, we sum them, this plus this, then what you do is, you take the fraction which is multiplied by the weight. So, this number multiplied by the weight, we get the first part, which is the first 8 by 14. Now, we do the 6 by 14 part and the fraction inside is 3 by 6. So, we have 2, 3

by 6 S, so that is we can do the same exact fashion equals 3 by 6. So, its life is, so that we have been following a process, so let us not confuse ourselves N31, 2, both of them we take the product of them by the probability.

Both the values now, you do is, you sum them. We know it is going to be 1 but that is, sum them up then, this value gets multiplied by its weight, this multiplied by the second weight, we get these two numbers. So, the sum is this plus this gives you 0.89216 minus 0.89216, the minus signs out, we will get the positive one, that is zero point.

So, that is the third one. So, we have now calculated the entropy, the average weighted entropy for all the four variables and we have come calculate the entropy of the Independent Variable.



Now, if that is the case, then the last part of the step is calculate the Information Gains. So, how do we calculate the Information Gain? So, the Information Gain (IG) the equation for this is remember,

$$IG(S, A) = E(S) = -\sum_{i=0}^{n} P(x)\, E(x)$$

So, to using that equation,

$$IG(S, Weather) = E(S) - E(Weather) = 0.94029 - 0.69354 =$$

$$IG(S, Temperature) = E(S) - E(Temperature) = 0.94029 - 0.91106 =$$

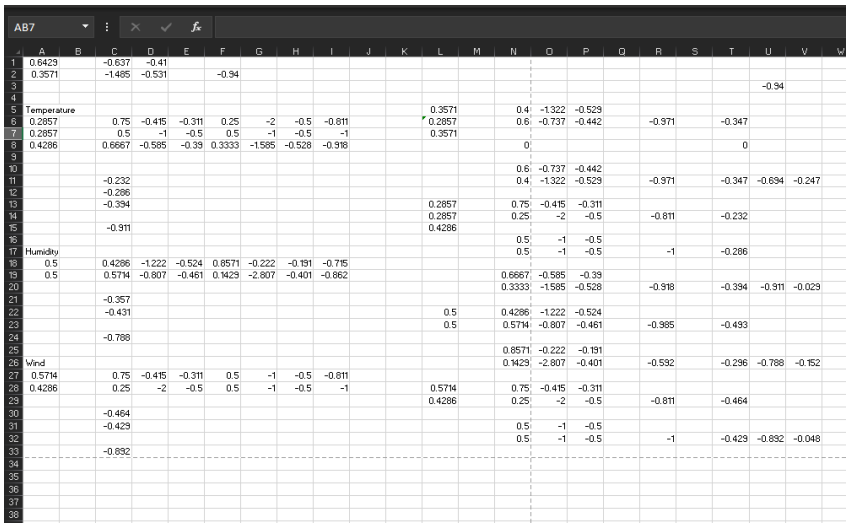
$$IG(S, Humidity) = E(S) - E(Humidity) = 0.94029 - 0.78845 =$$

$$IG(S, Wind) = E(S) - E(Wind) = 0.94029 - 0.89216 =$$

So, how do we calculate all these values? We can just again use Excel. So, we have the base what we calculated as 0.924025. I am just going to put it here equals this so the thing is, we have to first calculate the 0.69345. So, that is the difference is what we need to calculate.

So, what we need to do is, this value equals this. So, we should get a value of 0.24675, that is the first one. Then, we will calculate the same way, equals this number plus this one. So, if you remember, this is the so first is, we are doing this calculation the 0.91106 is the second calculation 0.78845 is the third one, what did I do wrong here, I made a mistake or not plus, it is minus because I needed to be equals because the signs are there, this one minus this that is 0.154 and then, the last one is the main one, this one minus this, last value we calculated it out.

So, now 0.24675 is the entropy against weather, so 0.24675 is the first one that we calculated out go to the next one, that is 991 that is minus point so 0.02922 and then, the fourth one means, go to the excel sheet, this is minus 0.15184 and then, the last one goes to excel again and we calculate dot minus so it is 0.04813.



-MS. Excel Demonstration

So, we have all the values we have calculated out now, the idea in this is the thing is remembered, the lowest value of entropy is what is best, so that means, lowest value of entropy will result in the largest Information Gain. Lowest entropy should result in largest information gain. So, what we do, which implies, we select the feature having the largest information gain among this which is largest value is this, which is weather. Hence, weather becomes the root node of the Decision Tree. So, the root node of the

Decision Tree is weather at this point, so we will do one thing and now we will start building the decision Tree after this.

So, I know there is a long process, so far, I have to explain to you the algorithm and the calculations behind this. So, we will speed up in the next one. Now, I hope that you understood the calculations and we will complete the tree in the next session. Thank you for your patient here.