Week 3 (index 1) Introduction to Decision Trees

In this video, we're going to introduce an examine decision trees. So, let's get started.

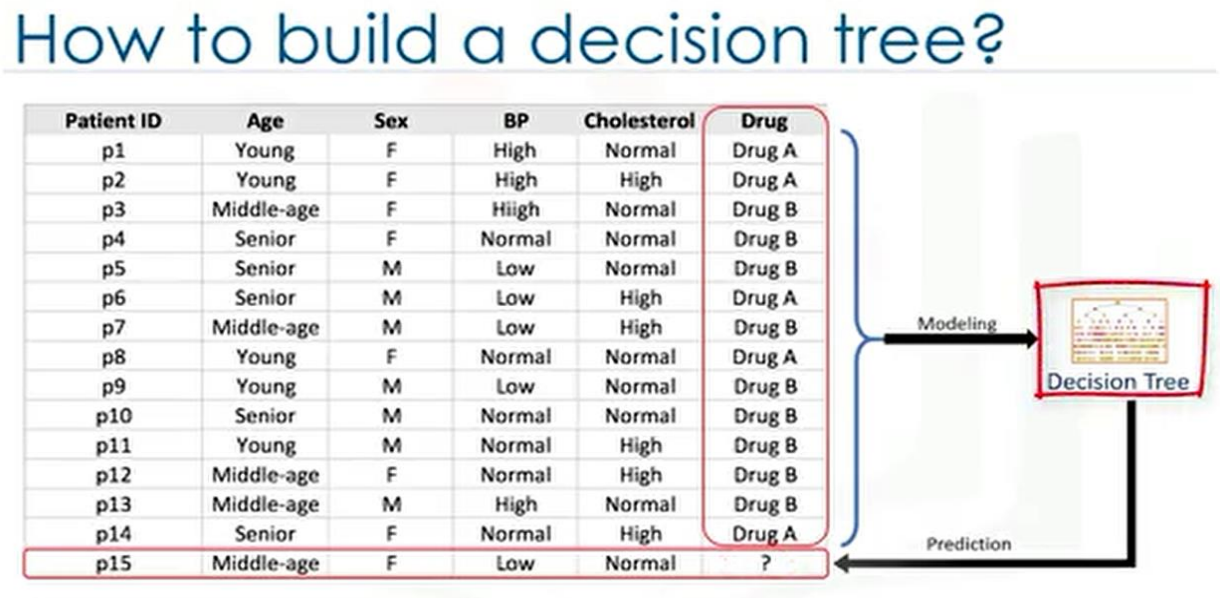
- ✓ What exactly is a decision tree?
- ✓ How do we use them to help us classify?
- ✓ How can I grow my own decision tree?

These may be some of the questions that you have in mind from hearing the term decision tree. Hopefully, you'll soon be able to answer these questions and many more by watching this video.

Imagine that you're a medical researcher compiling data for a study. You've already collected data about a set of patients all of whom suffered from the same illness. During their course of treatment, each patient responded to one of two medications. We call them drug A and drug B.

Part of your job is to build a model to find out which drug might be appropriate for a future patient with the same illness. The feature sets of this dataset are age, gender, blood pressure, and cholesterol of our group of patients and the target is the drug that each patient responded to.

It is a sample of binary classifiers, and you can use the training part of the data set to build a decision tree and then use it to predict the class of an unknown patient.



Patient ID	Age	Sex	BP	Cholesterol	Drug
p1	Young	F	High	Normal	Drug A
p2	Young	F	High	High	Drug A
p3	Middle-age	F	Hiigh	Normal	Drug B
p4	Senior	F	Normal	Normal	Drug B
p5	Senior	M	Low	Normal	Drug B
p6	Senior	M	Low	High	Drug A
p7	Middle-age	M	Low	High	Drug B
p8	Young	F	Normal	Normal	Drug A
p9	Young	M	Low	Normal	Drug B
p10	Senior	M	Normal	Normal	Drug B
p11	Young	M	Normal	High	Drug B
p12	Middle-age	F	Normal	High	Drug B
p13	Middle-age	M	High	Normal	Drug B
p14	Senior	F	Normal	High	Drug A
p15	Middle-age	F	Low	Normal	?

In essence, to come up with a decision on which drug to prescribe to a new patient. Let's see how a decision tree is built for this dataset.

Decision trees are built by splitting the training set into distinct nodes, where one node contains all of or most of one category of the data. If we look at the diagram here, we can see that it's a patient's classifier.

Note: So as mentioned, we want to prescribe a drug to a new patient, but the decision to choose drug A or B will be influenced by the patient's situation.

We start with age, which can be young, middle aged or senior. If the patient is middle aged, then we'll definitely go for drug B. On the other hand, if he has a young or a senior patient, will need more details to help us determine which drug to prescribe.

The additional decision variables can be things such as cholesterol levels, gender or blood pressure. **For example**,

- ✓ if the patient is female, then we will recommend drug A, but
- ✓ if the patient is male, then will go for drug B.

As you can see, decision trees are about testing an attribute and branching the cases based on the result of the test.

Each internal node corresponds to a test, and each branch corresponds to a result of the test, and each leaf node assigns a patient to a class.

Q: Now the question is, how can we build such a decision tree?

Here is the way that a decision tree is build. A decision tree can be constructed by considering the attributes one by one.

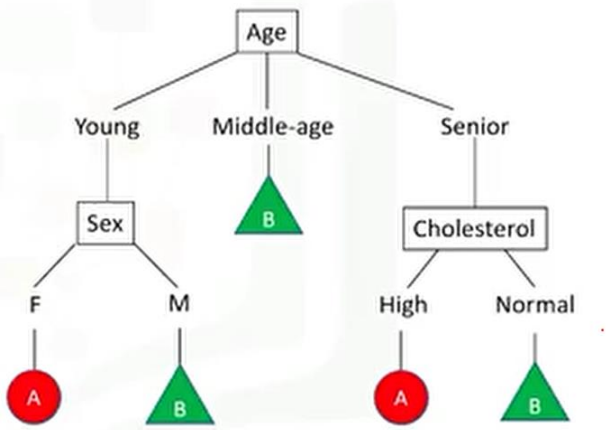
- First, choose an attribute from our dataset.
- Calculate the significance of the attribute in the splitting of the data.
- Next, split the data based on the value of the best attribute,

Then go to each branch and repeat it for the rest of the attributes.

After building this tree, you can use it to predict the class of unknown cases; or in our case, the proper drug for a new patient based on his or her characteristics.

Decision tree learning algorithm

1. Choose an attribute from your dataset.
2. Calculate the significance of attribute in splitting of data.
3. Split data based on the value of the best attribute.
4. Go to step 1.



Week 3 (index 2): Building Decision Trees

In this video, we'll be covering the process of building decision trees. So, let's get started. Consider the drug data set again.

Q: The question is, how do we build a decision tree based on that data set? Decision trees are built using recursive partitioning to classify the data.

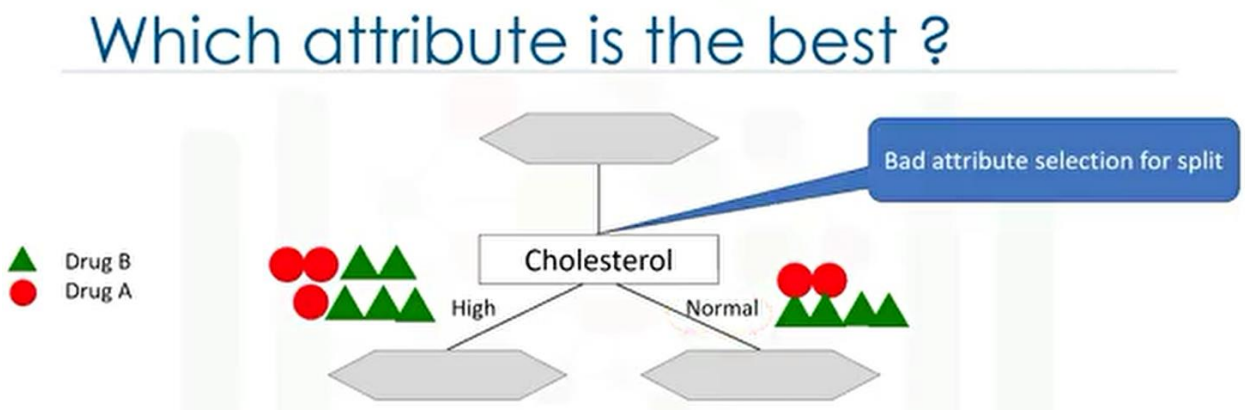
Let's say we have 14 patients in our data set, the algorithm chooses the most predictive feature to split the data on.

What is important in making a decision tree, is to determine which attribute is the best or more predictive to split data based on the feature.

Cholesterol attribute

Let's say we pick cholesterol as the first attribute to split data, it will split our data into two branches.

- ✓ As you can see, if the patient has high cholesterol, we cannot say with high confidence that drug B might be suitable for him.
- ✓ Also, if the patient's cholesterol is normal, we still don't have sufficient evidence or information to determine if either drug A or drug B is in fact suitable.



It is a sample of bad attributes selection for splitting data.

Sex attribute

- ✓ So, let's try another attribute. Again, we have our 14 cases, this time we picked the sex attribute of patients. It will split our data into two branches,

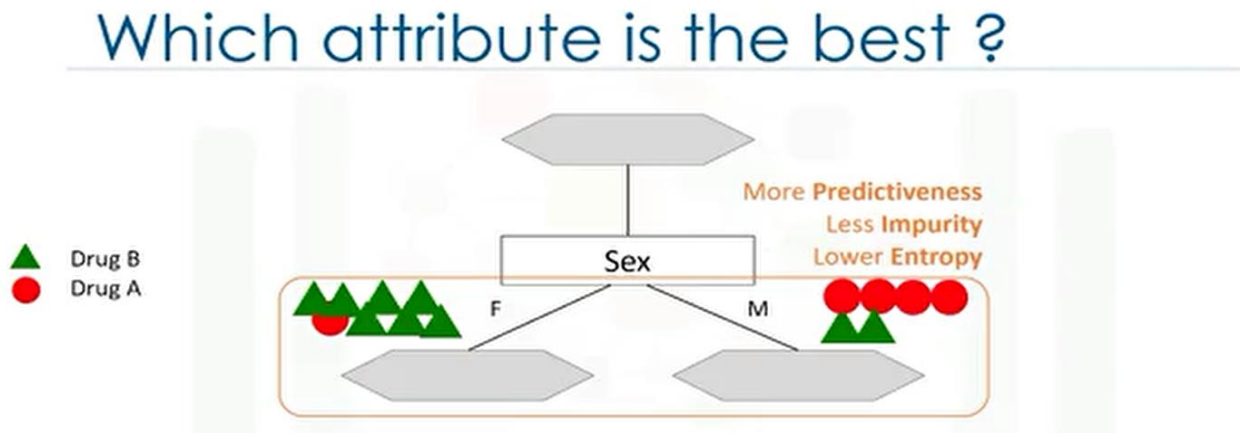
male and female. As you can see, if the patient is female, we can say drug B might be suitable for her with high certainty.

- ✓ But if the patient is male, we don't have sufficient evidence or information to determine if drug A or drug B is suitable.

It means nodes that are either mostly drug A or drug B. So, we can say the sex attribute is more significant than cholesterol, or in other words it's more predictive than the other attributes.

Note: Indeed, predictiveness is based on decrease in impurity of nodes. We're looking for the best feature to decrease the impurity of patients in the leaves, after splitting them up based on that feature.

So, the sex feature is a good candidate in the following case because it almost found the pure patients. Let's go one step further.



Let's go one step further. For the male patient branch, we again test other attributes to split the sub-tree. We test cholesterol again here, as you can see it results in even more pure leaves. So, we can easily make a decision here. **For example**,

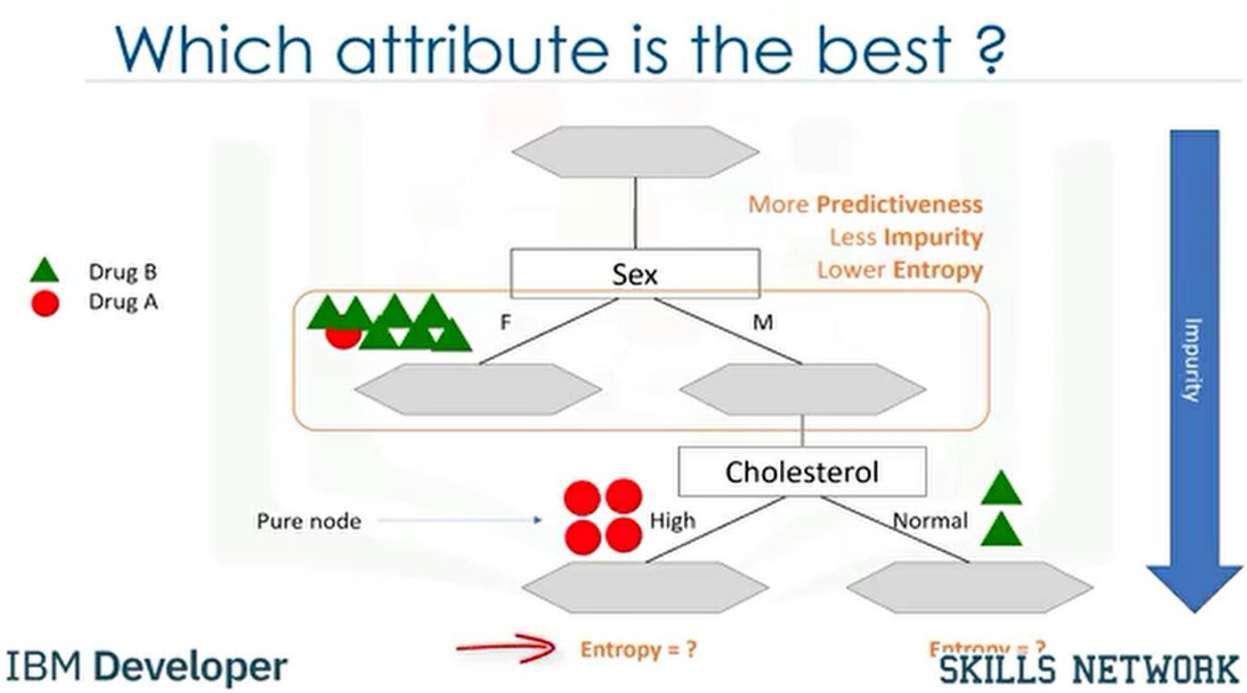
- ✓ if a patient is male and his cholesterol is high, we can certainly prescribe drug A,
- ✓ but if it is normal, we can prescribe drug B with high confidence.

Note: As you might notice, the choice of attribute to split data is very important and it is all about purity of the leaves after the split.

Note: A node in the tree is considered pure if in 100 percent of the cases, the nodes fall into a specific category of the target field.

Note: In fact, the method uses recursive partitioning to split the training records into segments by minimizing the impurity at each step.

Impurity of nodes is calculated by entropy of data in the node.



Q: So, what is entropy?

Entropy: Entropy is the amount of information disorder or the amount of randomness in the data.

The entropy in the node depends on how much random data is in that node and is calculated for each node.

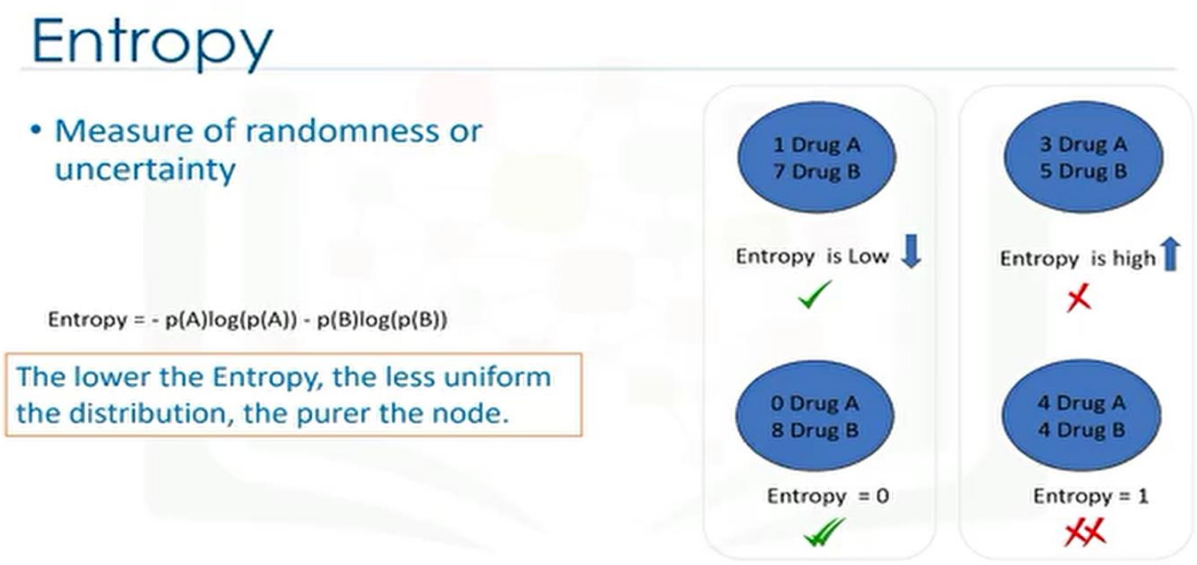
In decision trees, we're looking for trees that have the smallest entropy in their nodes.

The entropy is used to calculate the homogeneity of the samples in that node.

- ✓ If the samples are completely homogeneous, the entropy is zero and
- ✓ if the samples are equally divided it has an entropy of one.

This means

- if all the data in a node are either drug A or drug B, then the entropy is zero, but
- if half of the data are drug A and other half are B then the entropy is one.

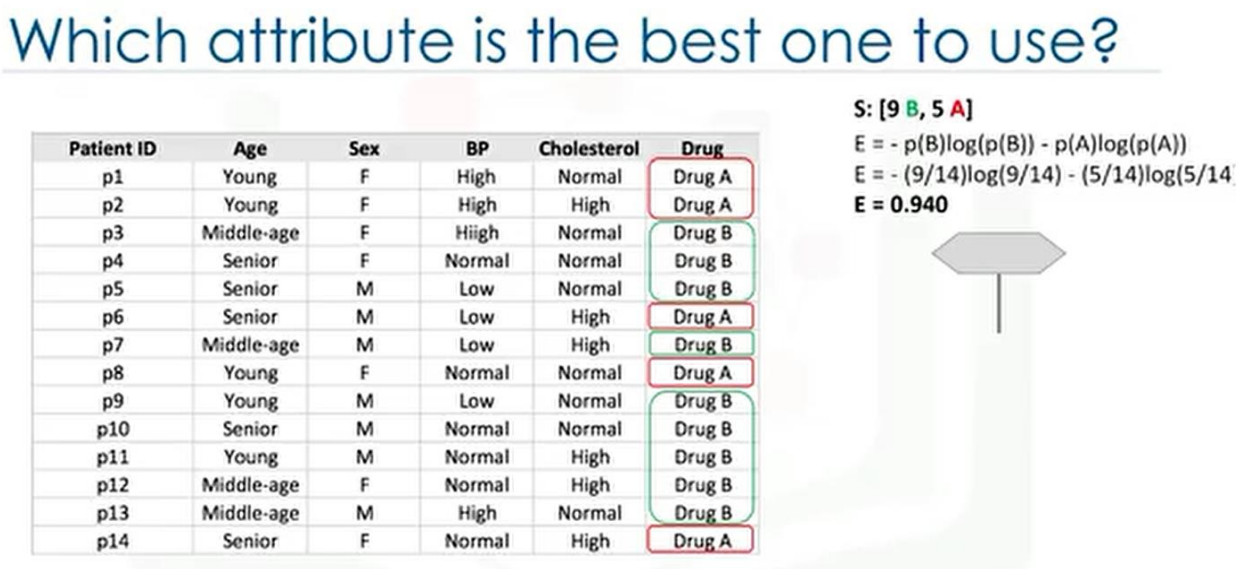


You can easily calculate the entropy of a node using the frequency table of the attribute through the entropy formula where P is for the proportion or ratio of a category, such as drug A or B. Please remember though that you don't have to calculate these as it's easily calculated by the libraries or packages that you use.

Entropy of the data set before splitting

As an example, let's calculate the entropy of the data set before splitting it.

We have nine occurrences of drug B and five of drug A. You can embed these numbers into the entropy formula to calculate the impurity of the target attribute before splitting it. In this case, it is E=0.94.



Patient ID	Age	Sex	BP	Cholesterol	Drug
p1	Young	F	High	Normal	Drug A
p2	Young	F	High	High	Drug A
p3	Middle-age	F	Hiigh	Normal	Drug B
p4	Senior	F	Normal	Normal	Drug B
p5	Senior	M	Low	Normal	Drug B
p6	Senior	M	Low	High	Drug A
p7	Middle-age	M	Low	High	Drug B
p8	Young	F	Normal	Normal	Drug A
p9	Young	M	Low	Normal	Drug B
p10	Senior	M	Normal	Normal	Drug B
p11	Young	M	Normal	High	Drug B
p12	Middle-age	F	Normal	High	Drug B
p13	Middle-age	M	High	Normal	Drug B
p14	Senior	F	Normal	High	Drug A

Entropy after splitting

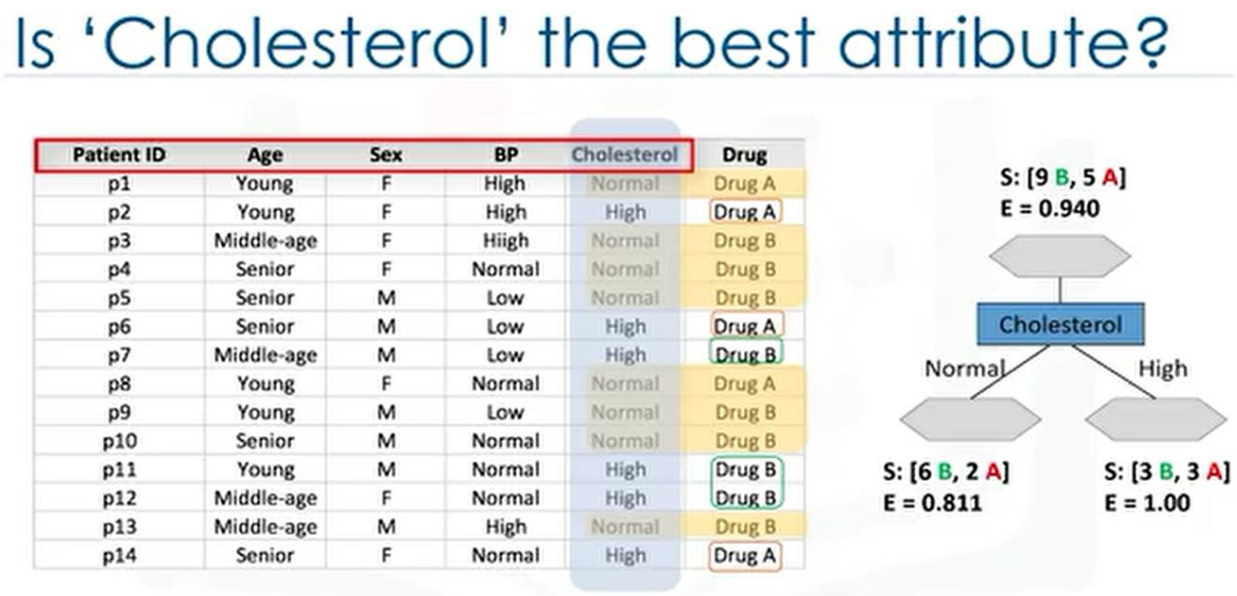
So, what is entropy after splitting?

Q: Now, we can test different attributes to find the one with the most predictiveness, which results in two more pure branches.

Cholesterol attribute

Let's first select the cholesterol of the patient and see how the data gets split based on its values. **For example**,

- when cholesterol is normal, we have six for drug B, and two for drug A. We can calculate the entropy of this node based on the distribution of drug A and B which is E = 0.8 in this case.
- But, when cholesterol is high, the data is split into three for drug B and three for drug A. Calculating it's entropy, we can see it would be E=1.0.



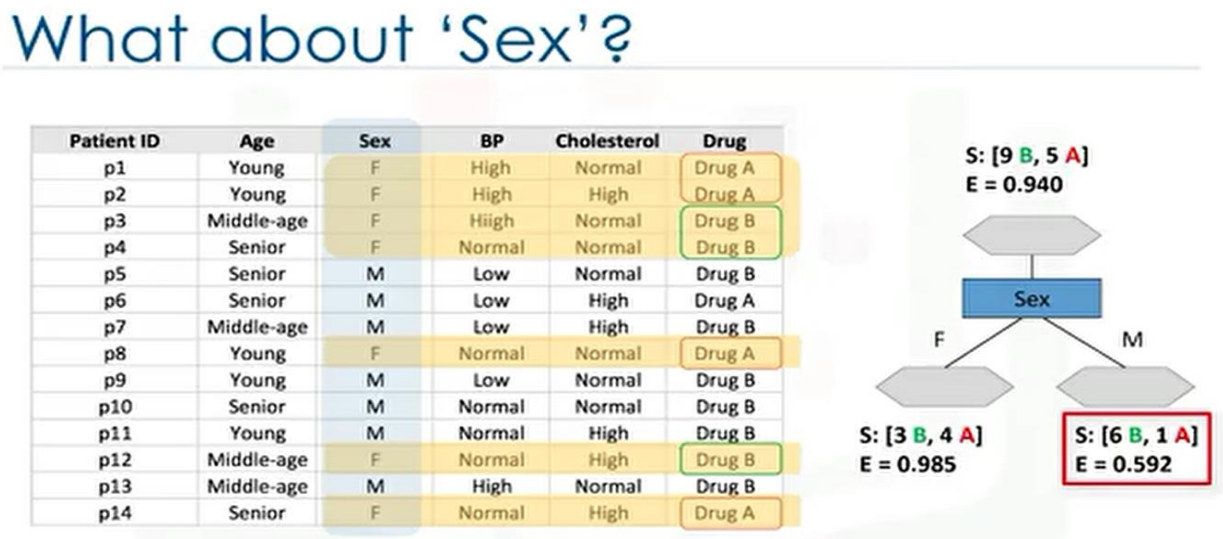
We should go through all the attributes and calculate the entropy after the split and then choose the best attribute.

Sex attribute

Let's try another field. Let's choose the sex attribute for the next check. As you can see, when we use the sex attribute to split the data,

when its value is **female**, we have three patients that responded to drug B and four patients that responded to drug A. The entropy for this node is E=0.98 which is not very promising. However, on the other side of the branch,

when the value of the sex attribute is **male**, the result is purer with sex for six drug B and only one for drug A. The entropy for this group is E=0.59.

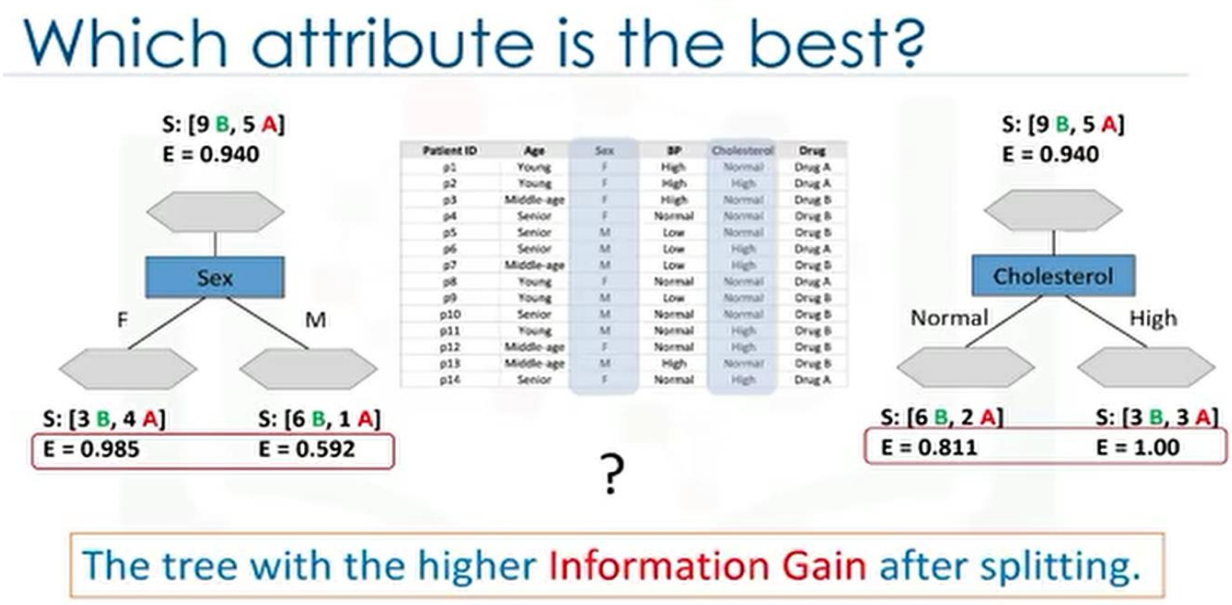


Comparison between the cholesterol and sex attributes

Now, the question is between the cholesterol and sex attributes

- ✓ which one is a better choice?
- ✓ Which one is better at the first attribute to divide the data-set into two branches?
- ✓ Or in other words, which attribute results in more pure nodes for our drugs? Or in which tree do we have less entropy after splitting rather than before splitting?

The sex attribute with entropy of 0.98 and 0.59 or the cholesterol attribute with entropy of 0.81 and 1.0 in it's branches. The answer is the tree with the higher information gain after splitting.

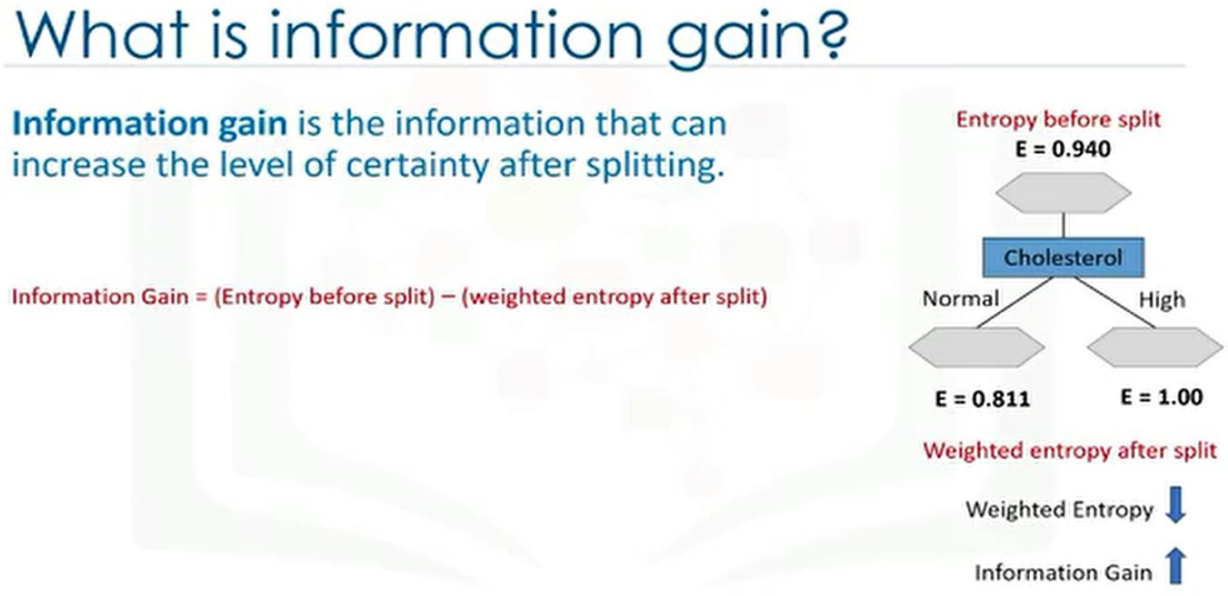


Q: So, what is information gain?

Information gain: Information gain is the information that can increase the level of certainty after splitting.

It is the entropy of a tree before the split minus the weighted entropy after the split by an attribute.

Note: We can think of information gain and entropy as opposites. As entropy or the amount of randomness decreases, the information gain or amount of certainty increases and vice versa.



Note: So, constructing a decision tree is all about finding attributes that return the highest information gain.

Let's see how information gain is calculated for the sex attribute. As mentioned, the information gained is the entropy of the tree before the split minus the weighted entropy after the split.

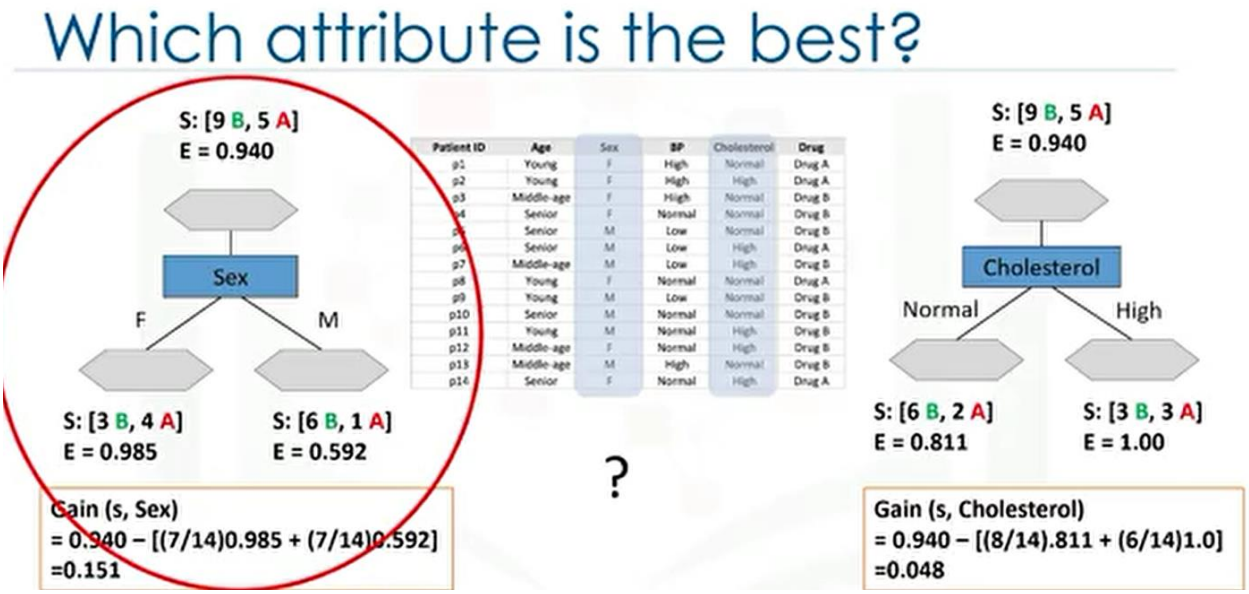
The entropy of the tree before the split is 0.94, the portion of female patients is seven out of 14 and its entropy is 0.985. Also, the portion of men is seven out of 14 and the entropy of the male node is 0.592. The result of a square bracket here is the weighted entropy after the split. So, the information gain of the tree if we use the sex attribute to split the data set is 0.151.

As you could see, we will consider the entropy over the distribution of samples falling under each leaf node and we'll take a weighted average of that entropy weighted by the proportion of samples falling under that leave.

We can calculate the information gain of the tree if we use cholesterol as well. It is 0.48.

Q: Now, the question is, which attribute is more suitable?

Well, as mentioned, the tree with the higher information gained after splitting, this means the sex attribute. So, we select the sex attribute as the first splitter.



Q: Now, what is the next attribute after branching by the sex attribute?

Well, as you can guess, we should repeat the process for each branch and test each of the other attributes to continue to reach the purest leaves. This is the way you build a decision tree.