به نام خدا

کارگاه علم داده با پایتون پیشرفته

جلسه ششم: درخت تصمیم

مدرس :

مهرناز جلیلی

دانشجو کارشناسی ارشد علم داده ها

دانشگاه شهید بهشتی

Classification

Decision Trees / Intro

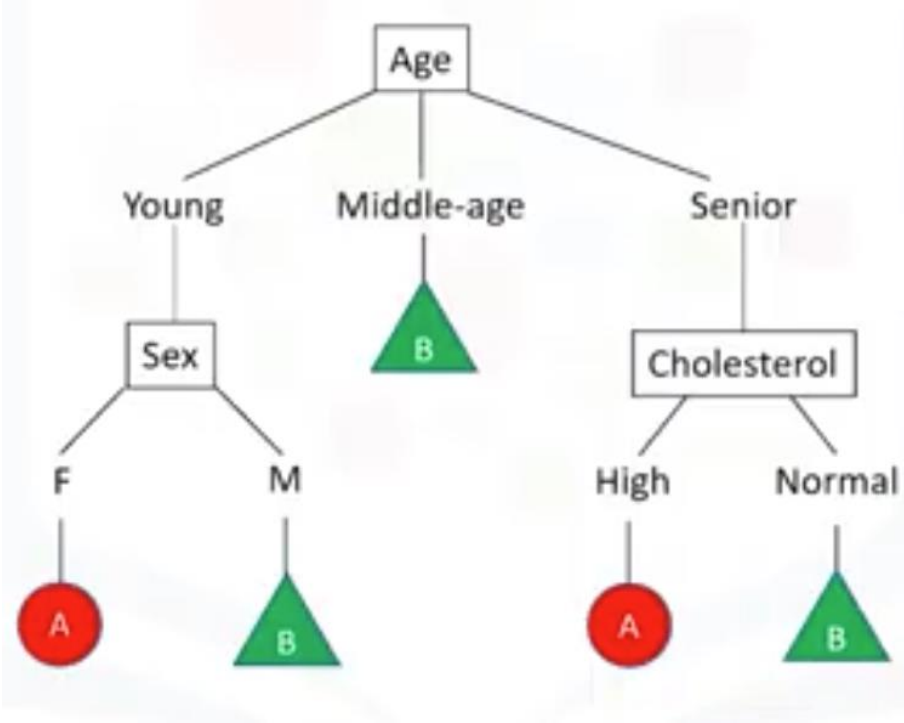
Patient ID	Age	Sex	BP	Cholesterol	Drug
p1	Young	F	High	Normal	Drug A
p2	Young	F	High	High	Drug A
p3	Middle-age	F	Hiigh	Normal	Drug B
p4	Senior	F	Normal	Normal	Drug B
p5	Senior	M	Low	Normal	Drug B
p6	Senior	M	Low	High	Drug A
p7	Middle-age	M	Low	High	Drug B
p8	Young	F	Normal	Normal	Drug A
p9	Young	M	Low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

Classification

Decision Trees / Intro

Internal Node (test), branch (result of test) & leaf (class)

1. Choose attribute from dataset
2. Calculate the significance of the attribute in the splitting of data
3. split data based on value of the best attribute
4. replete !

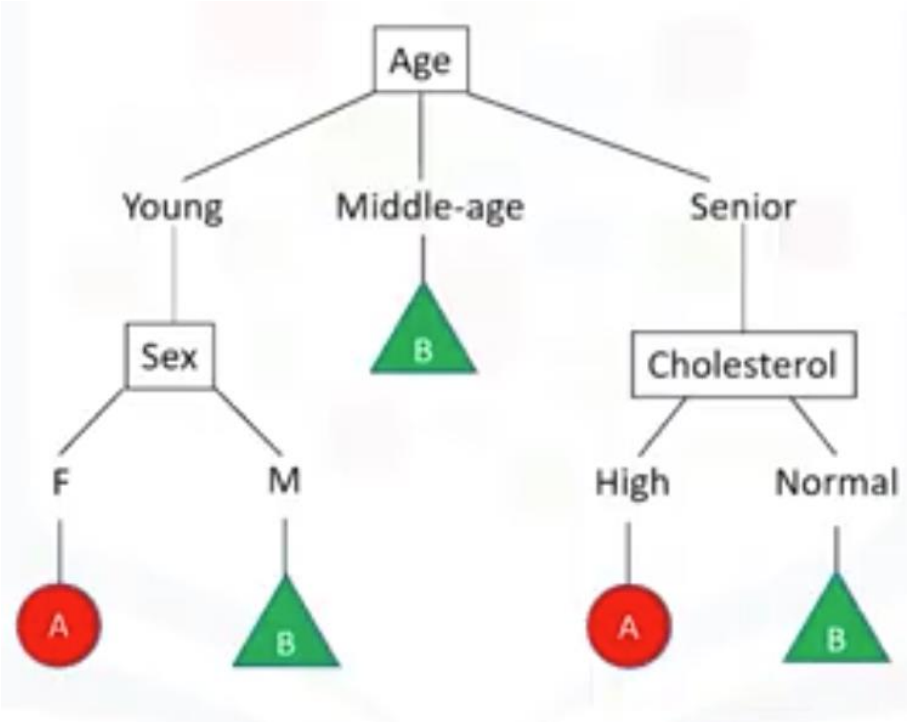


Classification

Decision Trees / Building

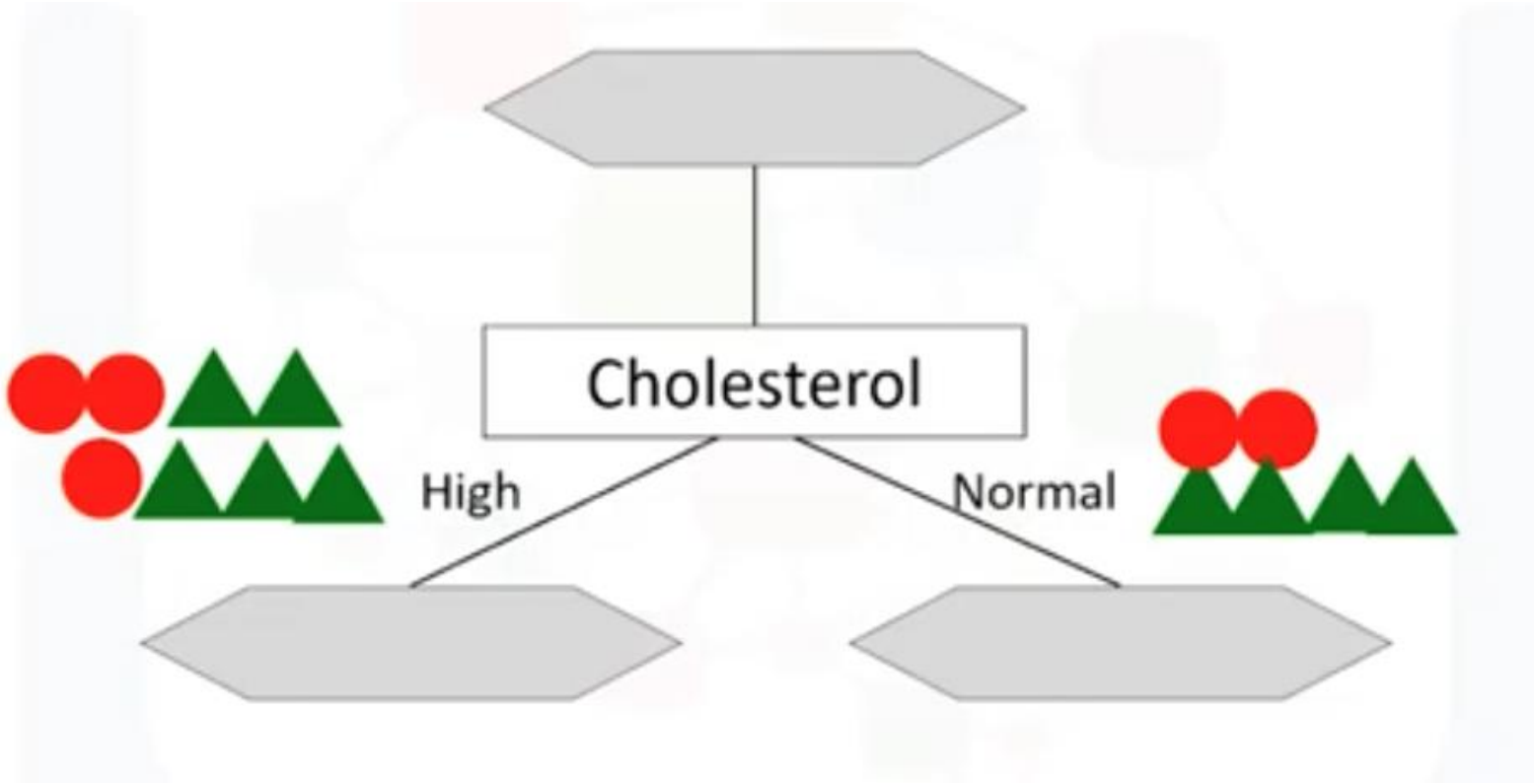
Decision trees are built using recursive partitioning to classify the data.
Cholesterol? Sex? ...

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	?



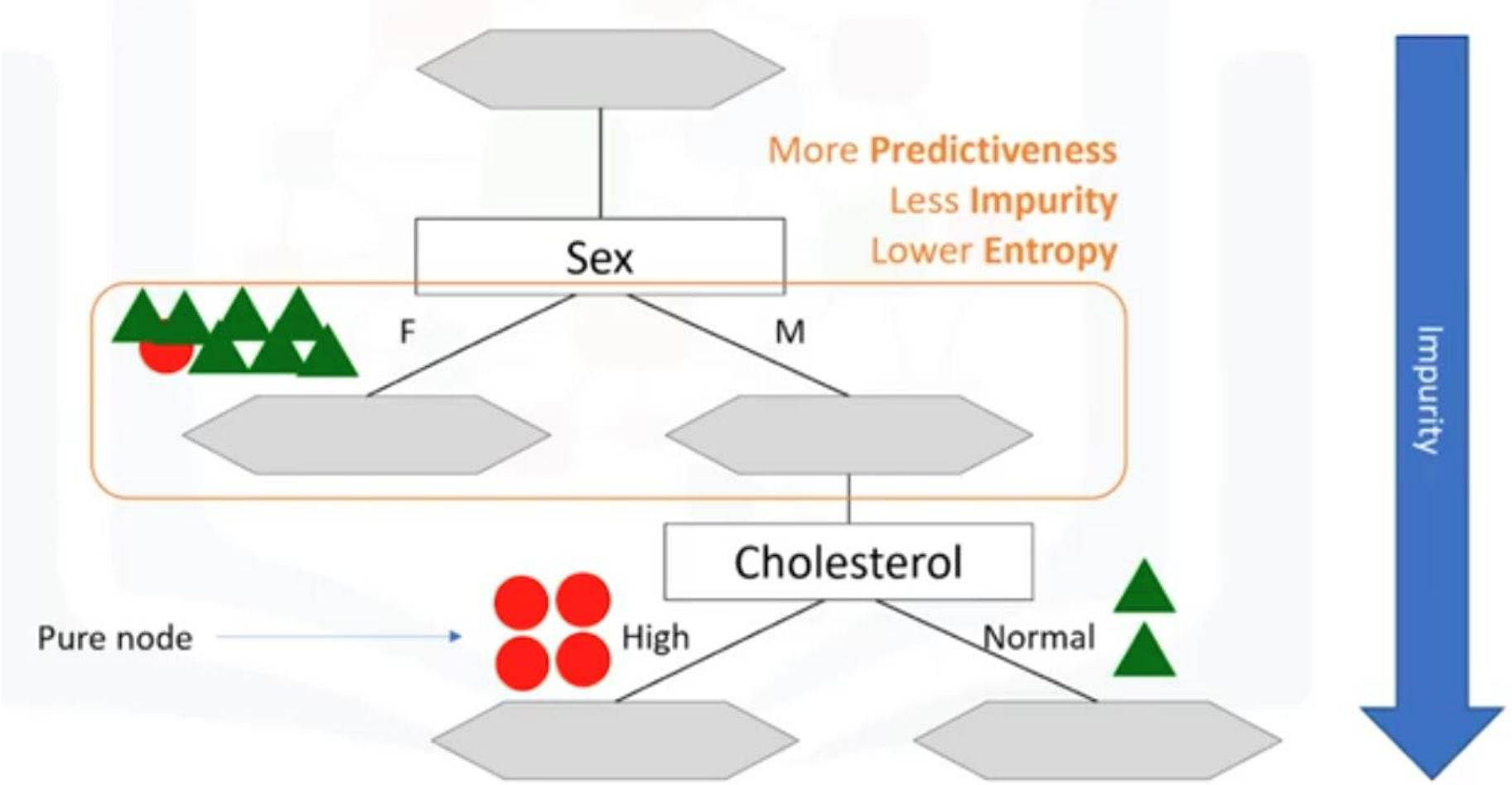
Classification

Decision Trees / Building



Classification

Decision Trees / Building



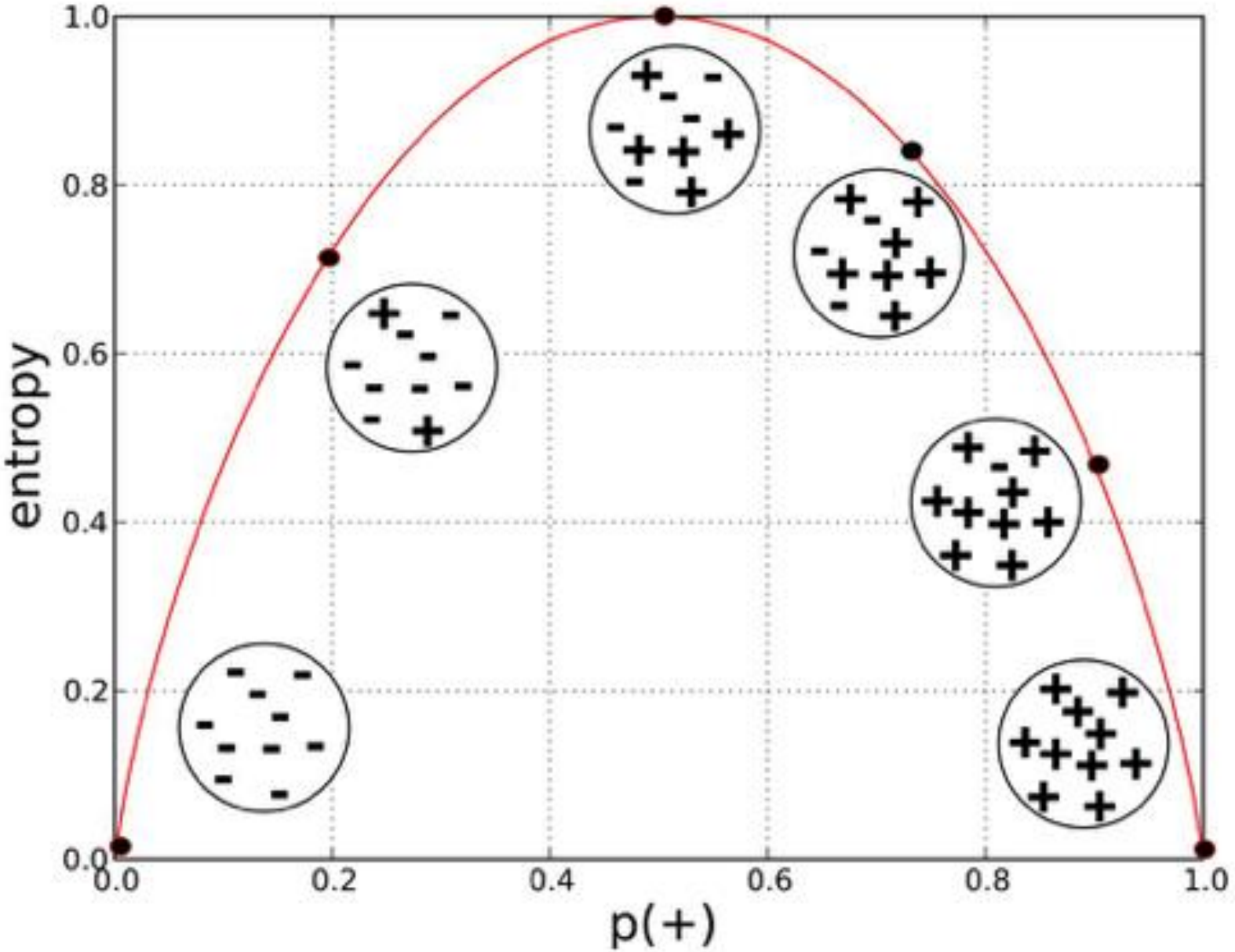
Entropy

$$Entropy(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$$

Bucket 1
Entropy: 0

Bucket 2
Entropy: 0.81125

Bucket 3
Entropy: 1


entropy
p(+)
0.0
0.2
0.4
0.6
0.8
1.0

Gini

Impurity Criterion

Gini Index

$$I_G = 1 - \sum_{j=1}^{c} p_j^2$$

p_j: proportion of the samples that belongs to class c for a particular node

Entropy

$$I_H = -\sum_{j=1}^{c} p_j log_2(p_j)$$

p_j: proportion of the samples that belongs to class c for a particular node.

*This is the the definition of entropy for all non-empty classes ($p \neq 0$). The entropy is 0 if all samples at a node belong to the same class.

The Gini coefficient is calculated for all possible splits and then the splits with the lowest value of the Gini coefficient is selected.

Gain

- Select splits maximizing **information gain**, i.e. **entropy reduction**:

$$IG = E(R) - [f_1 E(R_1) + \ldots + f_\ell E(R_\ell)]$$

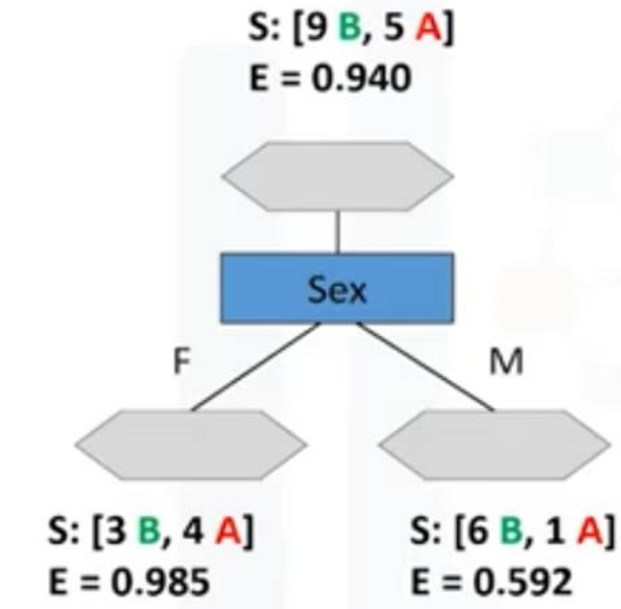
- R: parent node.
- $R_1, \ldots, R_\ell$: children nodes.
- f_k: fraction (proportion) of instances in child node R_k over all instances.

Classification

Decision Trees / Building

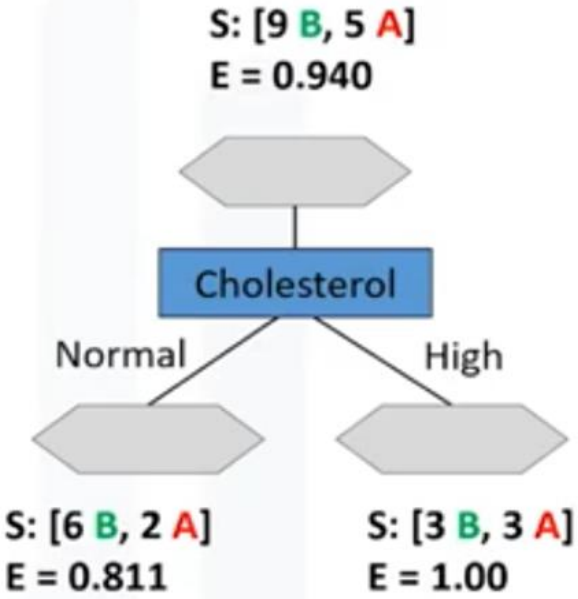
The best? the one with the most information gain
Information gain is the information that can increase the level of certainty after splitting.

- IG = Entropy before split - Weighted entropy after split.



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	High	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

?



Gain (s, Sex)
= 0.940 – [(7/14)0.985 + (7/14)0.592]
=0.151

Gain (s, Cholesterol)
= 0.940 – [(8/14).811 + (6/14)1.0]
=0.048

Lab: Decision Trees

- L06-MehrnazJalili