Classifiers: toy example of decision tree

Data mining lab 3

Input data

	skin	color	size	flesh	class
1	hairy	brown	large	hard	safe
2	hairy	green	large	hard	safe
3	smooth	red	large	soft	dangerous
4	hairy	green	large	soft	safe
5	hairy	red	small	hard	safe
6	smooth	red	small	hard	safe
7	smooth	brown	small	hard	safe
8	hairy	green	small	soft	dangerous
9	smooth	green	small	hard	dangerous
10	hairy	red	large	hard	safe
11	smooth	brown	large	soft	safe
12	smooth	green	small	soft	dangerous
13	hairy	red	small	soft	safe
14	smooth	red	large	hard	dangerous
15	smooth	red	small	hard	safe
16	hairy	green	small	hard	dangerous

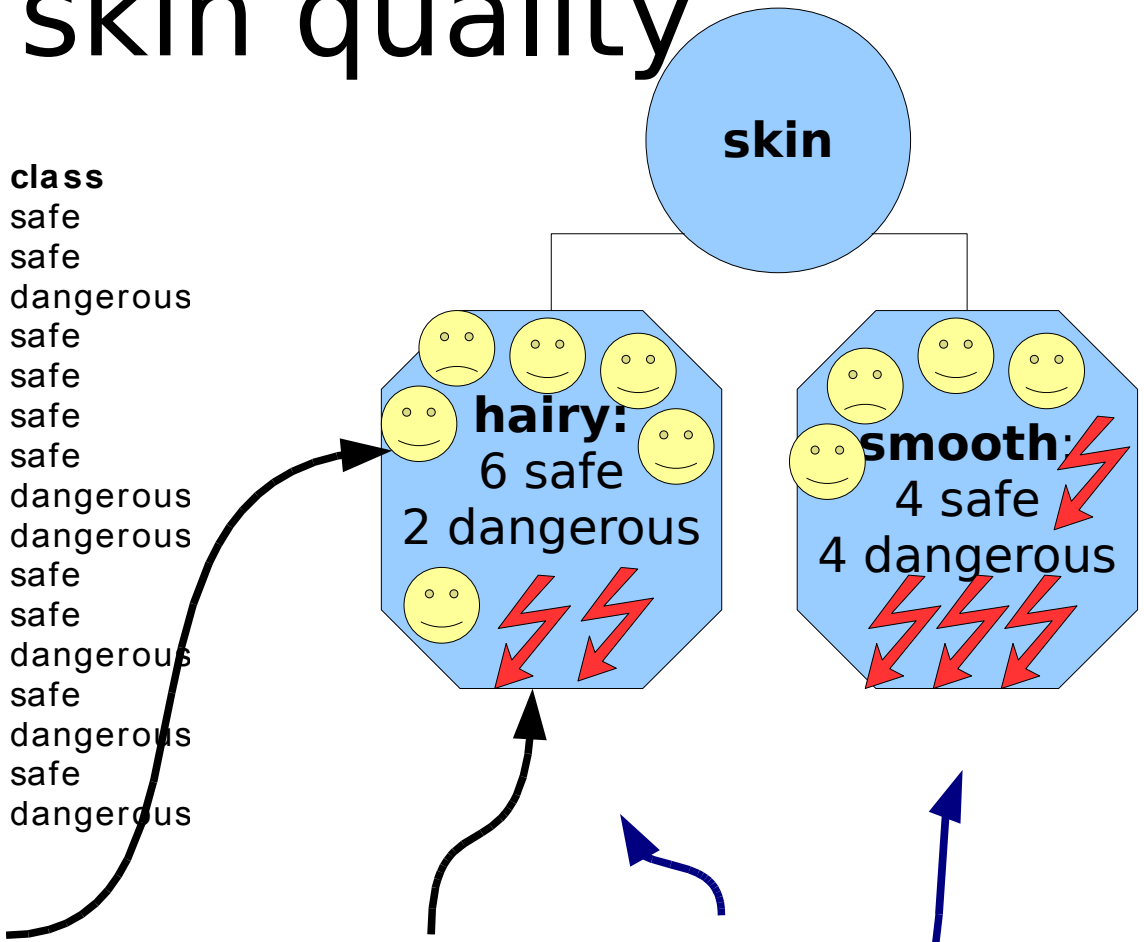
Query: hairy skin, red color, large with soft flesh.
The question is: is it safe to eat this unknown animal?

Decision tree classifier

- We can split the data based on any given nominal (non-numeric) attribute
- We compute the entropy (purity) of the tree nodes after the split and we choose the split wich gives the smallest average entropy for all the tree nodes created after the split

Split on skin quality

	skin	color	size	flesh	class
1	hairy	brown	large	hard	safe
2	hairy	green	large	hard	safe
3	smooth	red	large	soft	dangerous
4	hairy	green	large	soft	safe
5	hairy	red	small	hard	safe
6	smooth	red	small	hard	safe
7	smooth	brown	small	hard	safe
8	hairy	green	small	soft	dangerous
9	smooth	green	small	hard	dangerous
10	hairy	red	large	hard	safe
11	smooth	brown	large	soft	safe
12	smooth	green	small	soft	dangerous
13	hairy	red	small	soft	safe
14	smooth	red	large	hard	dangerous
15	smooth	red	small	hard	safe
16	hairy	green	small	hard	dangerous



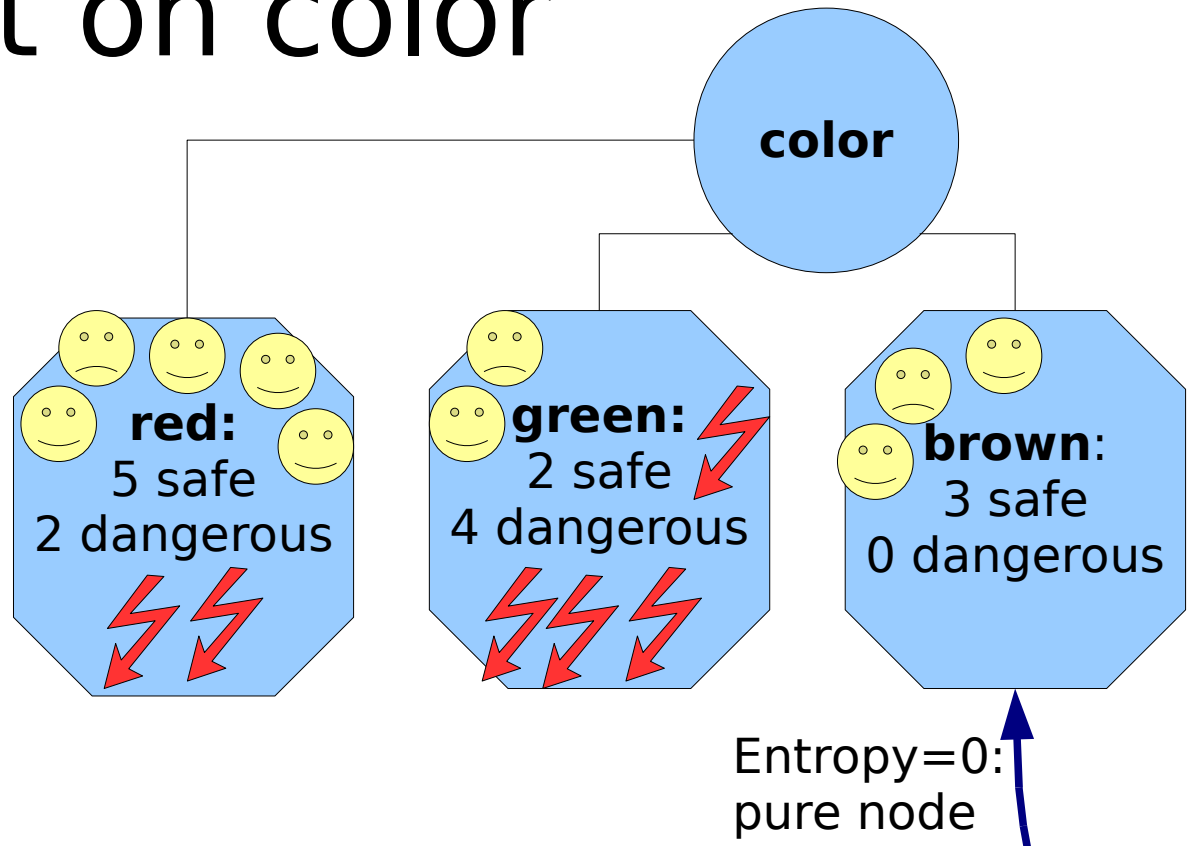
Entropy for the node [skin:hairy]=-6/8 log (6/8) -2/8 log(2/8)=0.24

Entropy for the node [skin:smooth]=-4/8 log (4/8)-4/8 log(4/8)=0.30

Averaged entropy of the split on skin attribute=8/16 * 0.24 +8/16* 0.30 = **0.27**

Split on color

color	size	flesh	class
brown	large	hard	safe
green	large	hard	safe
red	large	soft	dangerous
green	large	soft	safe
red	small	hard	safe
red	small	hard	safe
brown	small	hard	safe
green	small	soft	dangerous
green	small	hard	dangerous
red	large	hard	safe
brown	large	soft	safe
green	small	soft	dangerous
red	small	soft	safe
red	large	hard	dangerous
red	small	hard	safe
green	small	hard	dangerous



Entropy for the node [color:red]=-5/7 log (5/7) -2/7 log(2/7)=0.28

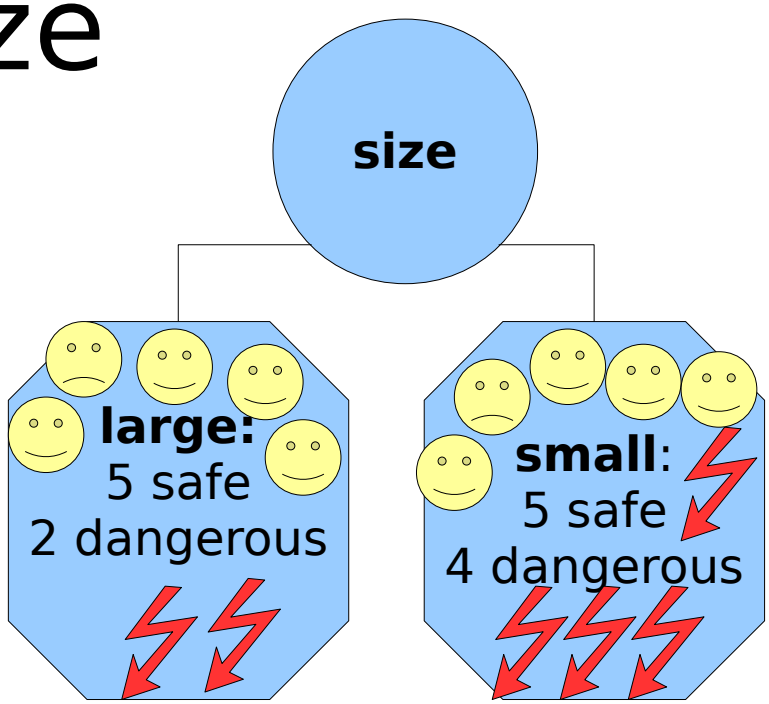
Entropy for the node [color:green]=-2/6 log (2/6)-4/6 log(4/6)=0.15

Entropy for the node [color:brown]=-3/3 log (3/3)=0

Averaged entropy of the split on color attribute=7/16 * 0.28 +6/16* 0.15 = **0.17**

Split on size

size	flesh	class
large	hard	safe
large	hard	safe
large	soft	dangerous
large	soft	safe
small	hard	safe
small	hard	safe
small	hard	safe
small	soft	dangerous
small	hard	dangerous
large	hard	safe
large	soft	safe
small	soft	dangerous
small	soft	safe
large	hard	dangerous
small	hard	safe
small	hard	dangerous



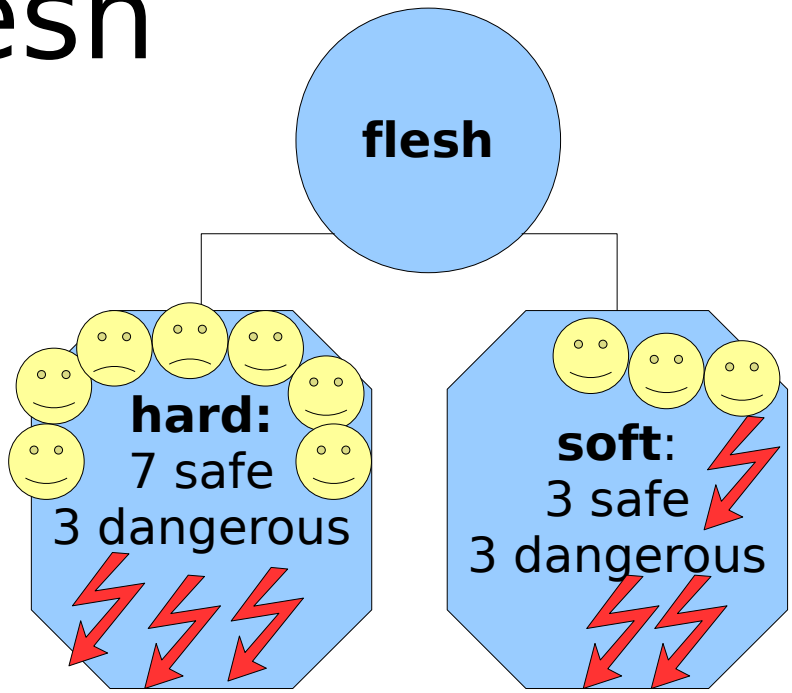
Entropy for the node [size:small]=-5/7 log (5/7) -2/7 log(2/7)=0.26

Entropy for the node [size:large]=-5/9 log (5/9)-4/9 log(4/9)=0.30

Averaged entropy of the split on size=7/16 * 0.26 +9/16* 0.30 = **0.28**

Split on flesh

flesh	class
hard	safe
hard	safe
soft	dangerous
soft	safe
hard	safe
hard	safe
hard	safe
soft	dangerous
hard	dangerous
hard	safe
soft	safe
soft	dangerous
soft	safe
hard	dangerous
hard	safe

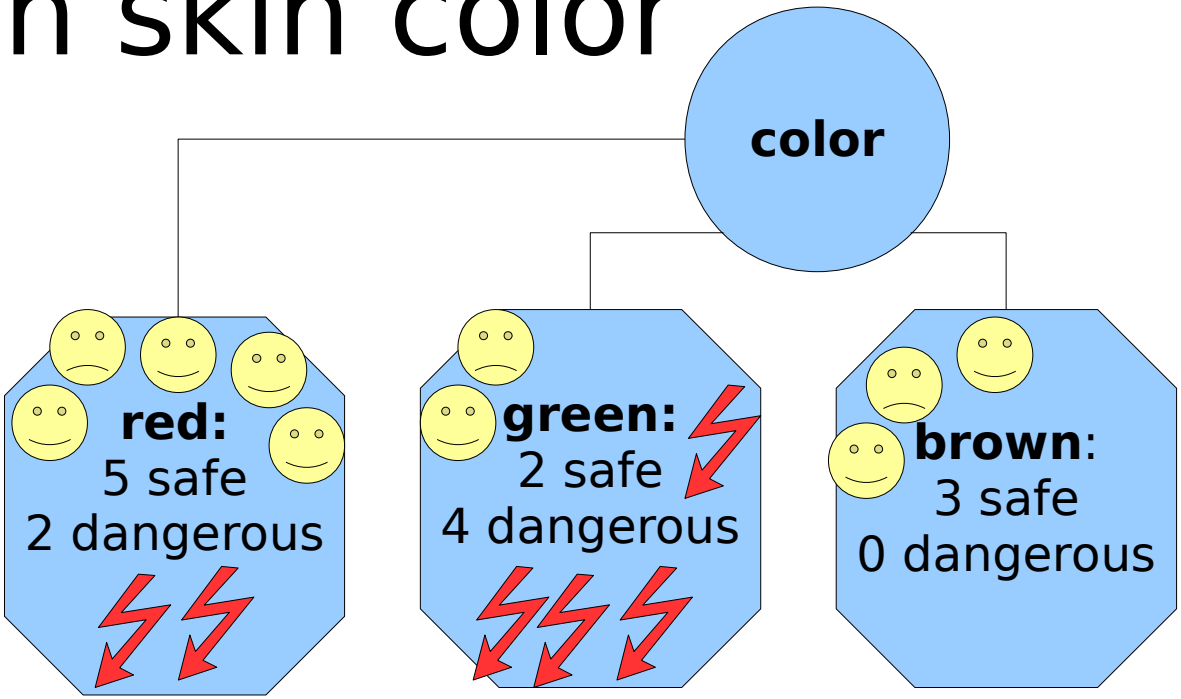


Entropy for the node [flesh:hard]=-7/10 log (7/10) -3/10 log(3/10)=0.27

Entropy for the node [flesh:soft]=-3/6 log (3/6)-3/6 log(3/6)=0.33

Averaged entropy of the split on flesh=10/16 * 0.27 +6/16* 0.33 = **0.29**

Split on skin color



The best split: the smallest entropy of the nodes after the split.
We choose this split to be a root node of the decision tree

Splitting node: color=red

color=red

	skin	size	flesh	class
3	smooth	large	soft	dangerous
5	hairy	small	hard	safe
6	smooth	small	hard	safe
10	hairy	large	hard	safe
13	hairy	small	soft	safe
14	smooth	large	hard	dangerous
15	smooth	small	hard	safe

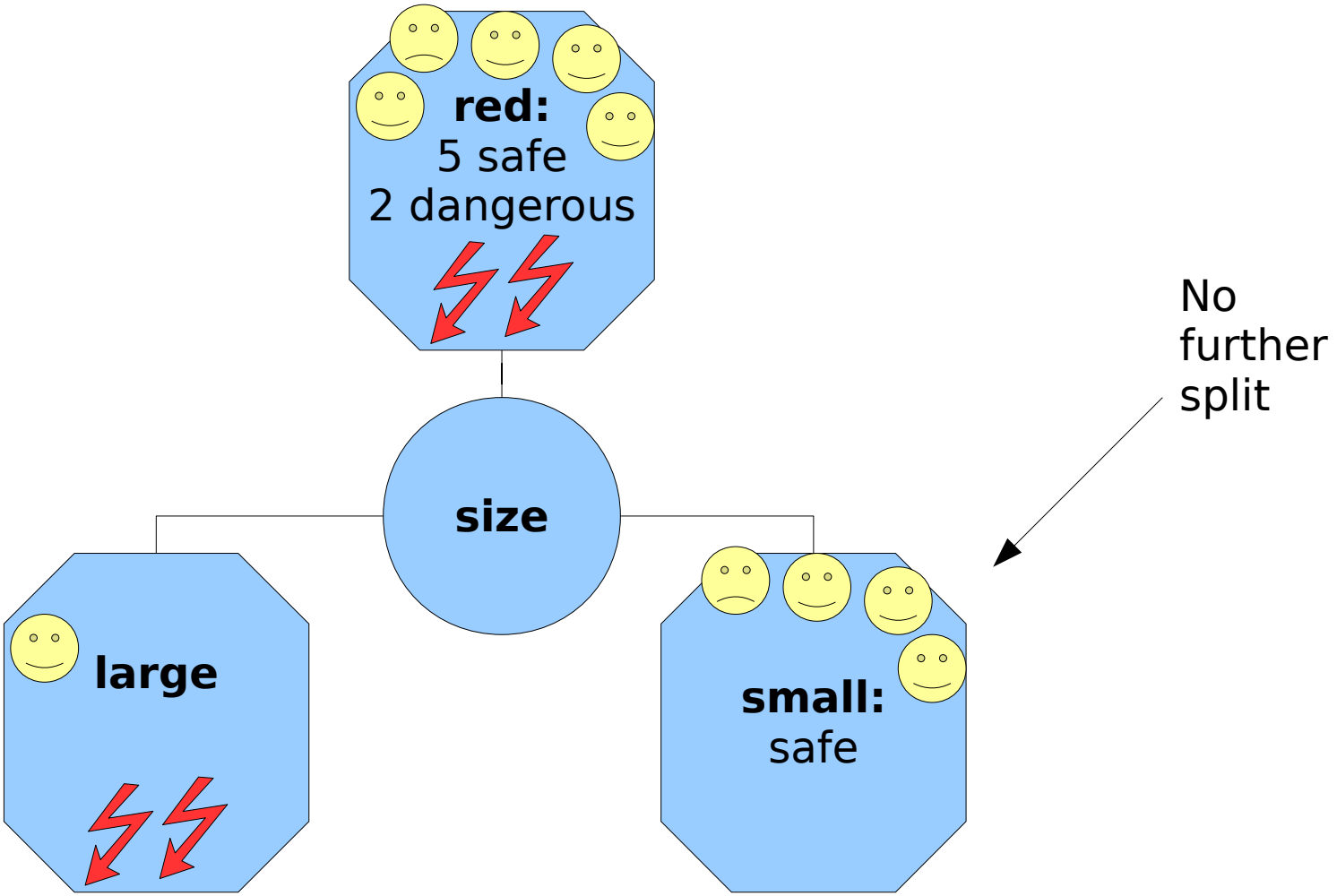
Entropy of the node [skin=smooth]=-2/4 log 2/4 – 2/4 log 2/4=0.30
Entropy of the node [skin=hairy]=-3/3 log 3/3 =0
Averaged entropy of the split on skin=4/7 *0.30 = **0.17**

Entropy of the node [size=large]=-1/3 log 1/3 – 2/3 log 2/3=0.28
Entropy of the node [size=small]=-4/4 log 4/4 =0
Averaged entropy of the split on size=3/7 * 0.28 = **0.12**

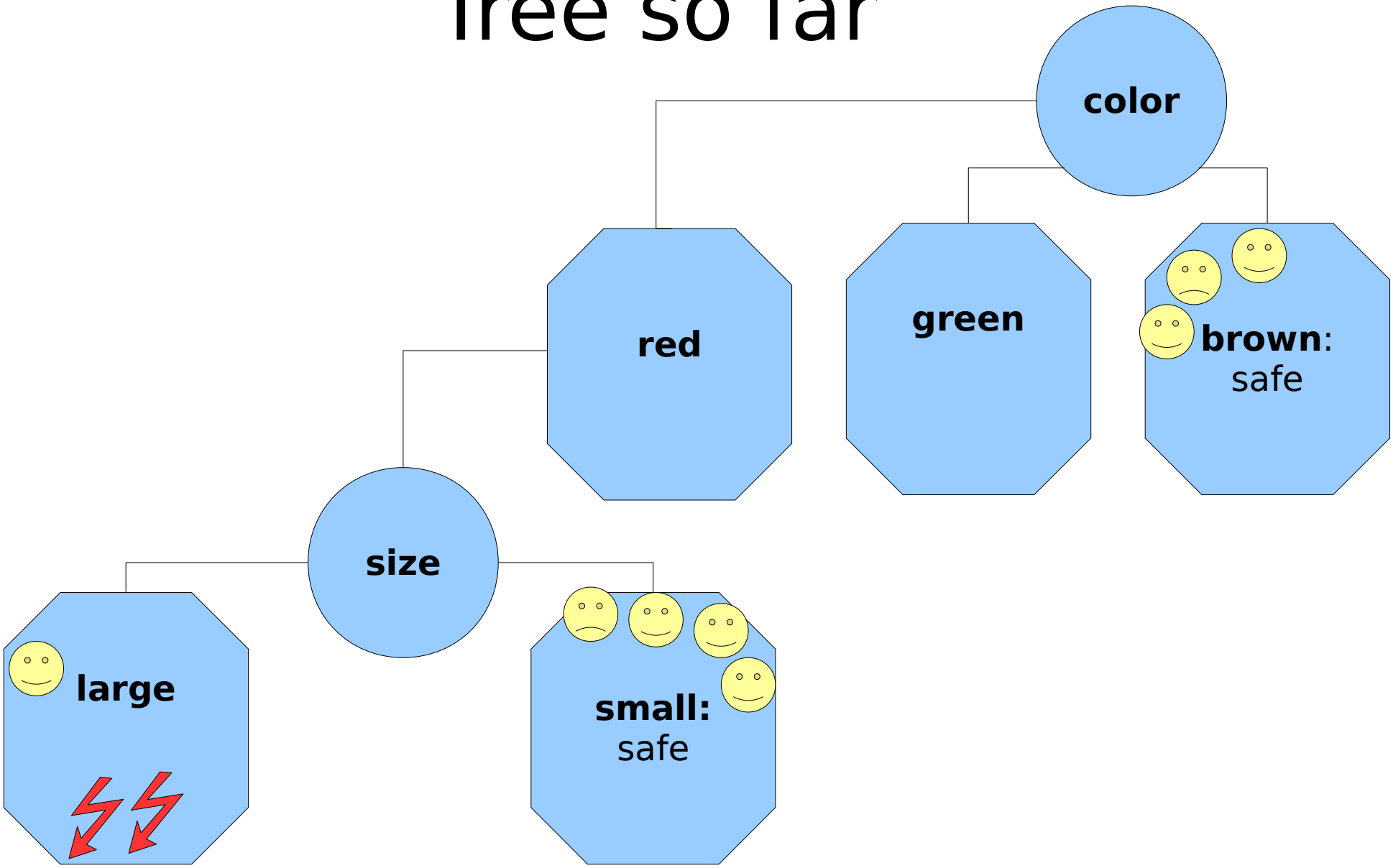
The best

Entropy of the node [flesh=soft]=-1/2 log 1/2 – 1/2 log 1/2=0.30
Entropy of the node [flesh=hard]=-4/5 log 4/5 – 1/5 log 1/5=0.22
Averaged entropy of the split on flesh=2/7 *0.30 +5/7*0.22 = **0.24**

Results of the split:



Tree so far



Splitting node: color=green

color=green

	skin	size	flesh	class
2	hairy	large	hard	safe
4	hairy	large	soft	safe
8	hairy	small	soft	dangerous
9	smooth	small	hard	dangerous
12	smooth	small	soft	dangerous
16	hairy	small	hard	dangerous

Entropy of the node [skin=smooth]=-2/2 log 2/2 =0

Entropy of the node [skin=hairy]=-2/4 log 2/4-2/4 log 2/4 =0.30

Averaged entropy of the split on skin=4/6 *0.30 = **0.20**

The best

Entropy of the node [size=large]=-2/2 log 2/2 =0

Entropy of the node [size=small]=-4/4 log 4/4 =0

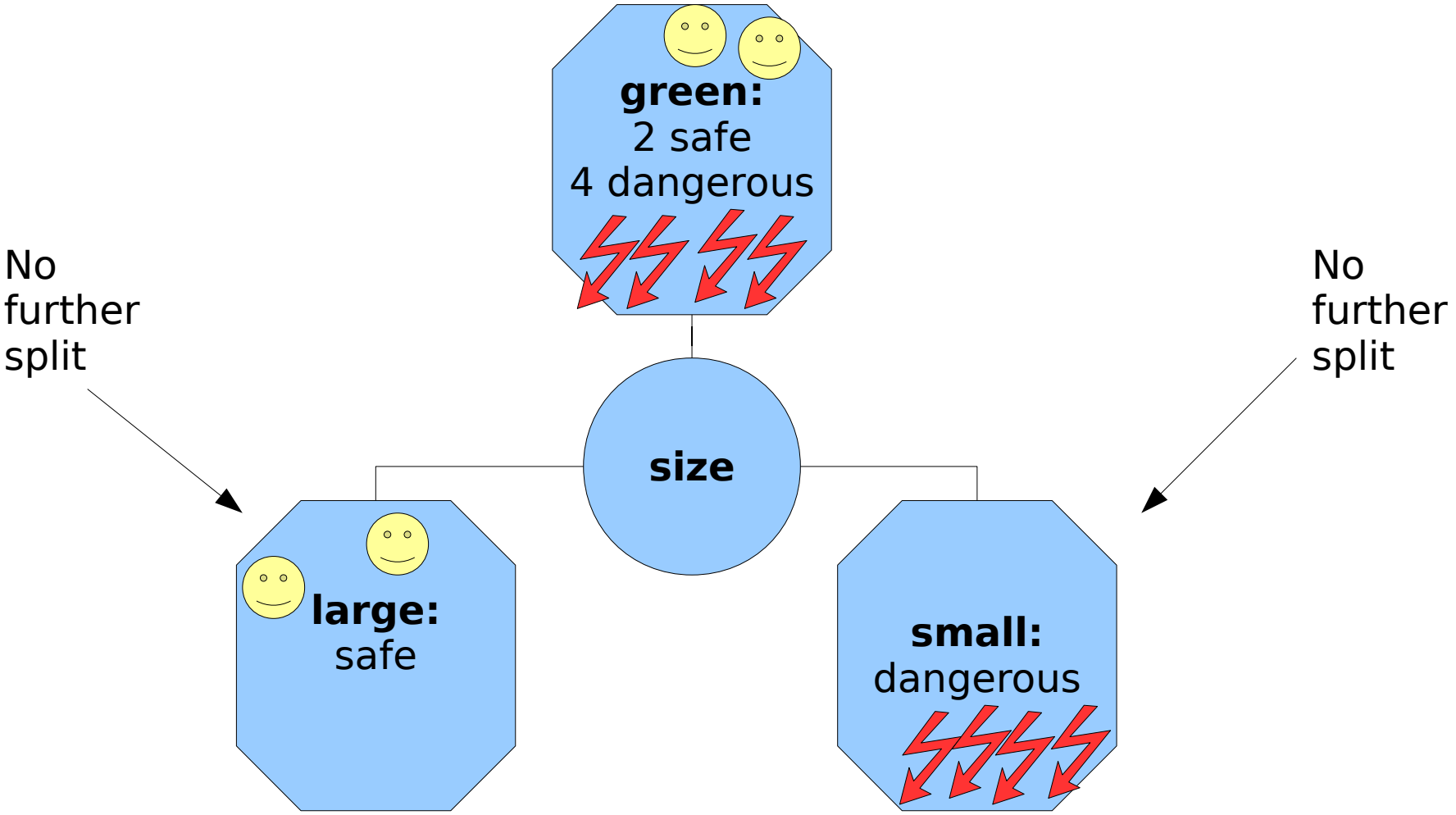
Averaged entropy of the split on size=**0.00**

Entropy of the node [flesh=soft]=-1/3 log 1/3 – 2/3 log 2/3=0.28

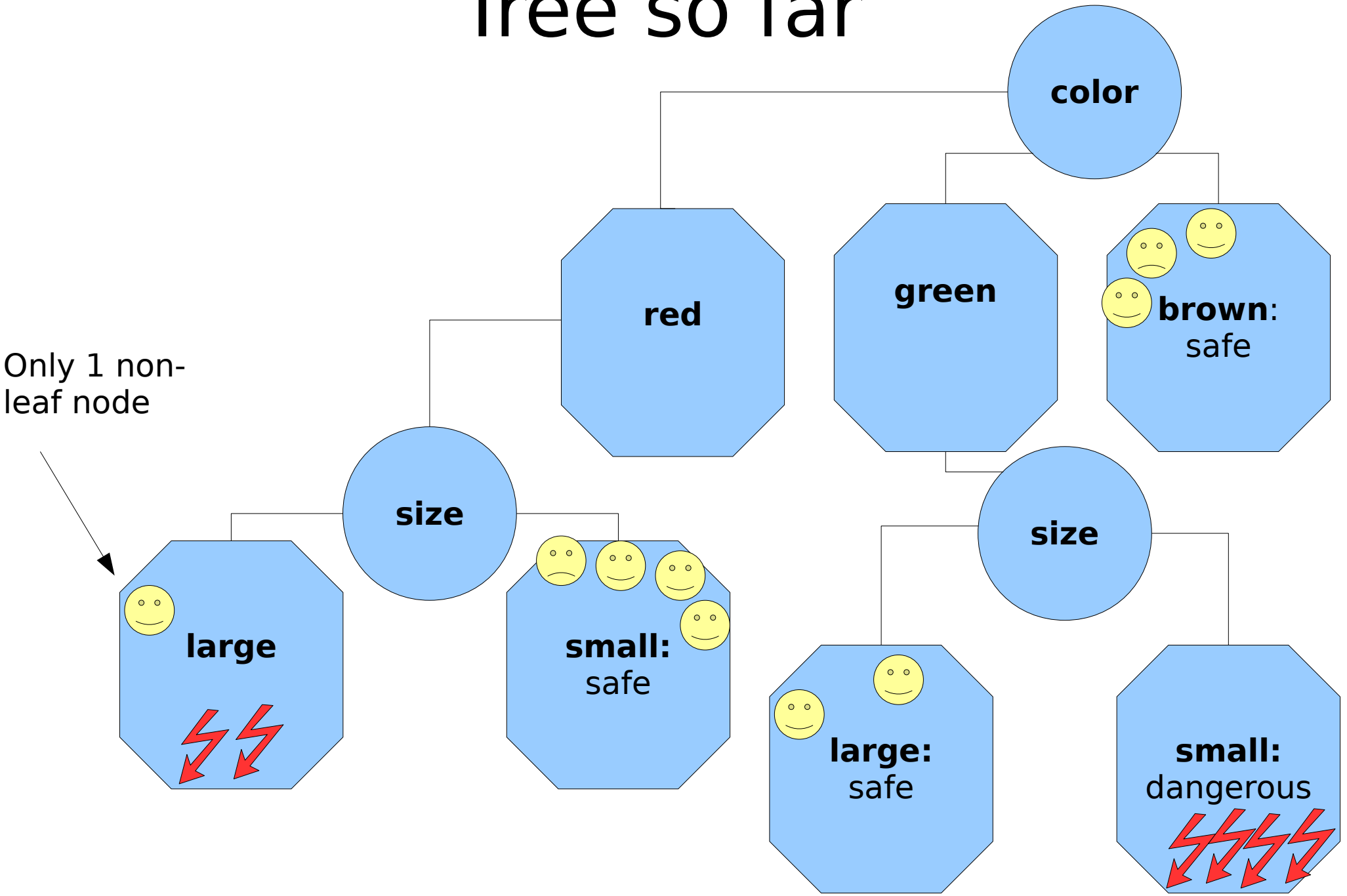
Entropy of the node [flesh=hard]=-1/3 log 1/3 – 2/3 log 2/3=0.28

Averaged entropy of the split on flesh=3/6 *0.28 +3/6*0.28 = **0.28**

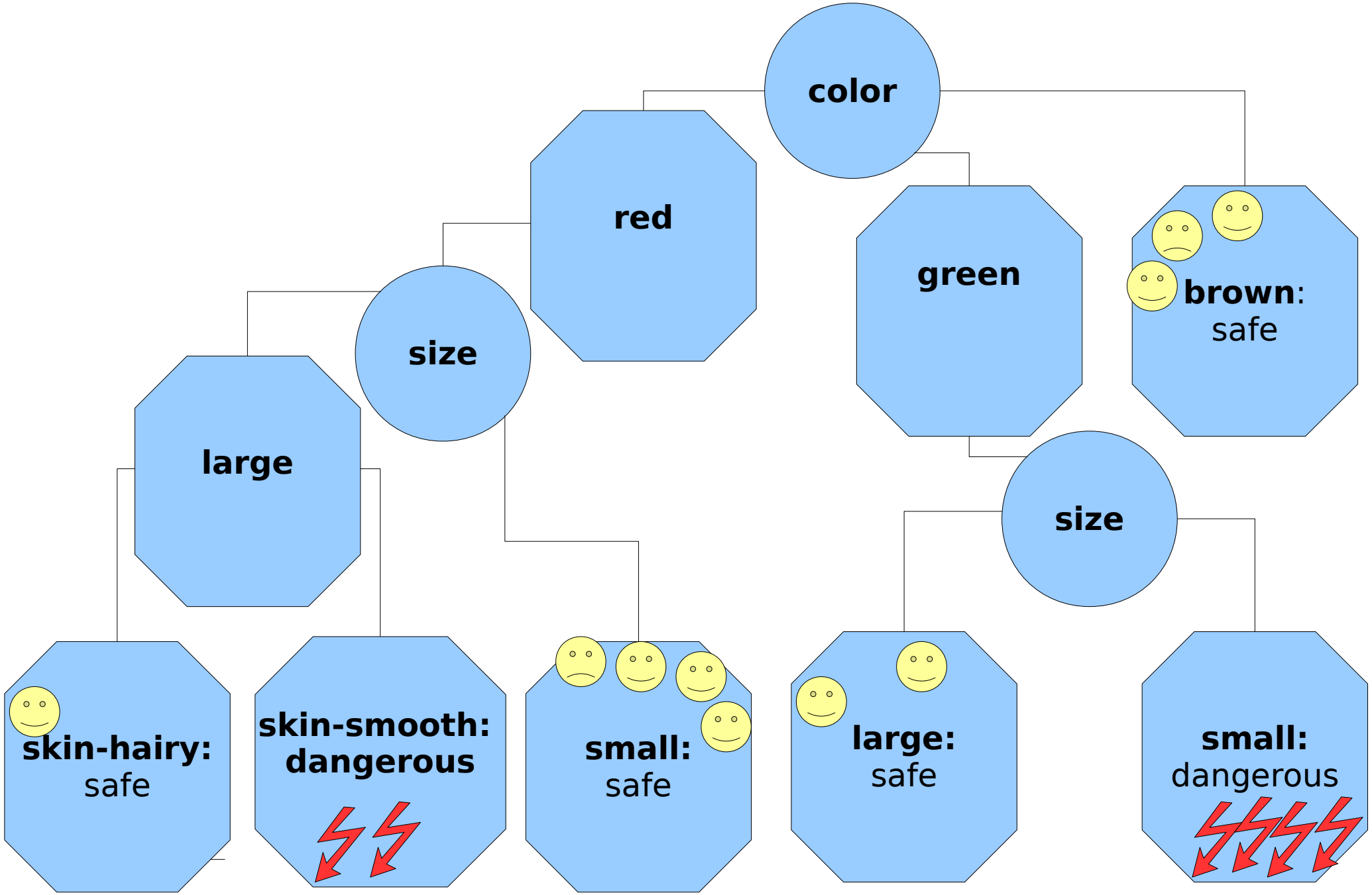
Results of the split:



Tree so far


color
red
green
brown: safe
Only 1 non-leaf node
size
size
large
small: safe
large: safe
small: dangerous

The final tree



Classify:

hairy skin, red color, large with soft flesh: safe