Khoi Duong

Prof. Yang

CS483

8/20/2022

HW#4

1.

Based on the document about entropy calculation, we have the new formula for Entropy:

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

We can calculate the entropy of root for 3 classes (apple, grape, and lemon):

$$\text{E (root)} = -\frac{2}{5}\log_2(\frac{2}{5}) - \frac{2}{5}\log_2(\frac{2}{5}) - \frac{1}{5}\log_2(\frac{1}{5}) = 1.5219$$

We have a great value of entropy, which means a high level of disorder.

Is color green?

For level 2, we have the entropy of the left subtree is:

$$\text{E (! = green)} = -\frac{1}{4}\log_2(\frac{1}{4}) - \frac{1}{4}\log_2(\frac{1}{4}) - \frac{2}{4}\log_2(\frac{2}{4}) = 1.5$$

The entropy of the right subtree is 0 (since we know the impurity = 0, which means there is no disorder on the right subtree)

Weighted average of entropy:

$$E\ (\text{green ?})\ =\ \frac{4}{5} * 1.5 = 1.2$$

Information gain: IG (root, green?) = E(root) - E(green?) = 1.5219 - 1.2 = ***0.3219***

Is diameter >= 3?

For level 2, we have the entropy of the left subtree is 0 (since there is only one class: Grape)

The entropy of the right subtree is:

$$E(\text{diameter}\ >=\ 3) = -\frac{2}{3}\log_2(\frac{2}{3}) - \frac{1}{3}\log_2(\frac{1}{3}) = 0.9183$$

Weighted average of entropy:

$E\ (\text{diameter}\ >= 3\ ?)\ =\ \frac{3}{5} * 0.9183 = 0.5510$

Information gain: IG (root, diameter >= 3?) = E(root) - E(diameter >= 3?)

= 1.5219 - 0.5510 = ***0.9709***

Is color red?

E(== red) = 0 (since there is only one class: Grape)

$$E(! = \text{red}) = -\frac{2}{3}\log_2(\frac{2}{3}) - \frac{1}{3}\log_2(\frac{1}{3}) = 0.9183$$

Weighted average of entropy:

$E\ (\text{red?})\ =\ \frac{3}{5} * 0.9183 = 0.5510$

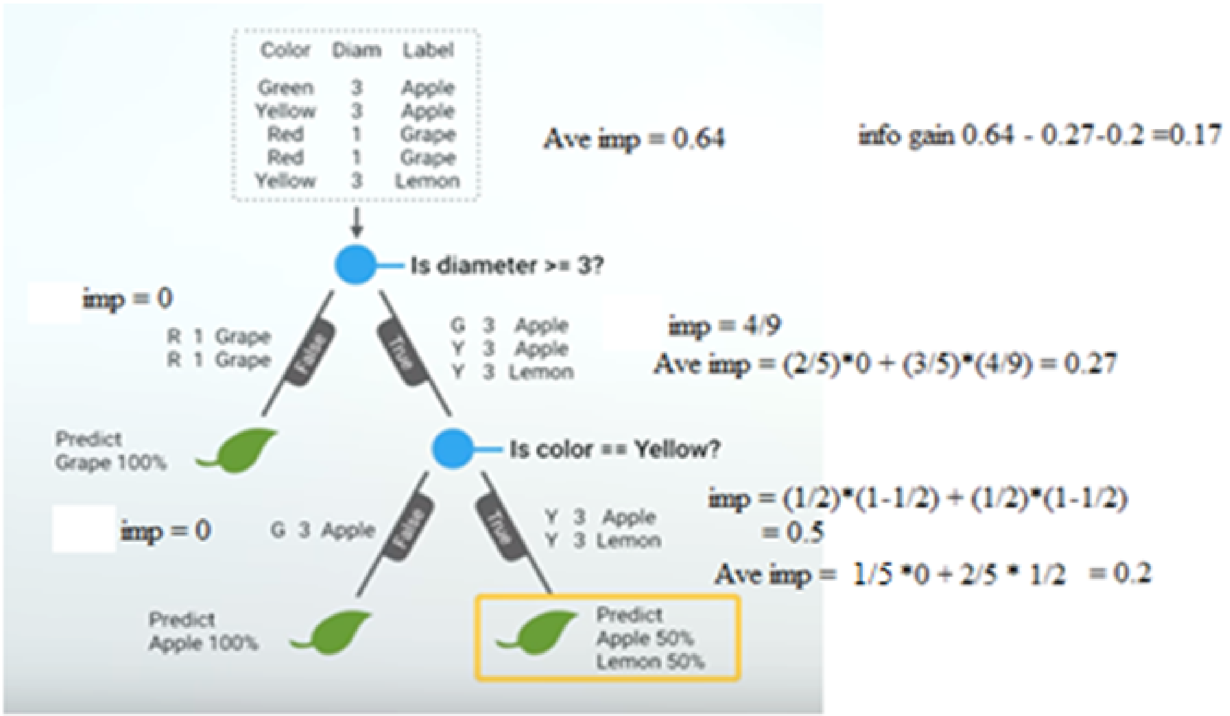
Information gain: IG (root, red?) = E(root) - E(red?) = 1.5219 - 0.5510 = ***0.9709***

Is diameter >= 1?

Since none of the classes has a diameter < 1, therefore the branch will not separate anything from the root branch. Thus, information gain IG = 0

We will take "Is diameter >= 3?" as the next level branch.

Go to the next level by different branches and compare info gains.



Color	Diam	Label
Green	3	Apple
Yellow	3	Apple
Red	1	Grape
Red	1	Grape
Yellow	3	Lemon

Ave imp = 0.64

info gain 0.64 - 0.27-0.2 =0.17

imp = 0

imp = 4/9

Ave imp = (2/5)*0 + (3/5)*(4/9) = 0.27

imp = (1/2)*(1-1/2) + (1/2)*(1-1/2)
= 0.5

imp = 0

Ave imp = 1/5 *0 + 2/5 * 1/2 = 0.2

<u>Is color == Yellow?</u>

E(! = Yellow) = 0 (since there is only one class: Apple)

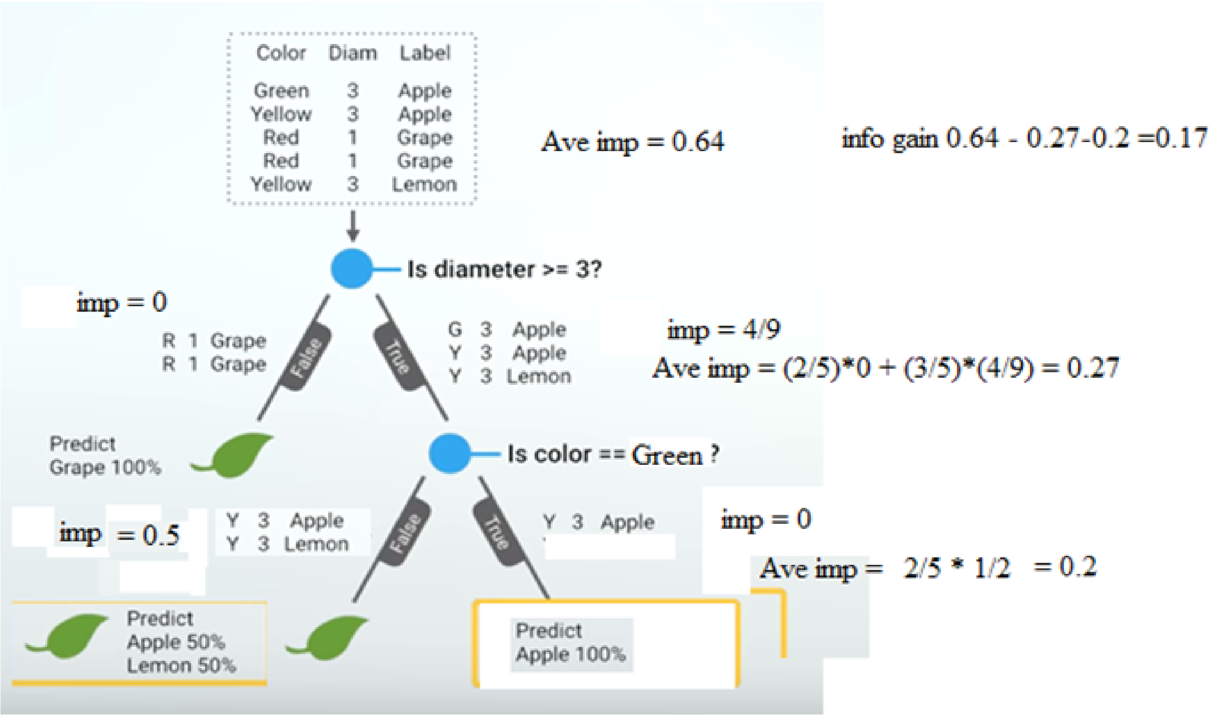
$$E(== \text{Yellow}) = -\frac{1}{2}\log_2(\frac{1}{2}) - \frac{1}{2}\log_2(\frac{1}{2}) = 1$$

Weighted average of entropy:

$E\ (\text{Yellow?}) = \frac{2}{5} * 1 = 0.4$

Information gain: IG (diameter >= 3? , yellow?) = E(diameter >= 3) - E(yellow?)

= 0.5510 - 0.4 = ***0.1510***



Color	Diam	Label
Green	3	Apple
Yellow	3	Apple
Red	1	Grape
Red	1	Grape
Yellow	3	Lemon

Ave imp = 0.64

info gain 0.64 - 0.27-0.2 =0.17

imp = 0

imp = 4/9

Ave imp = (2/5)*0 + (3/5)*(4/9) = 0.27

imp = 0.5

imp = 0

Ave imp = 2/5 * 1/2 = 0.2

<u>Is color == Green?</u>

E(== Green) = 0 (since there is only one class: Apple)

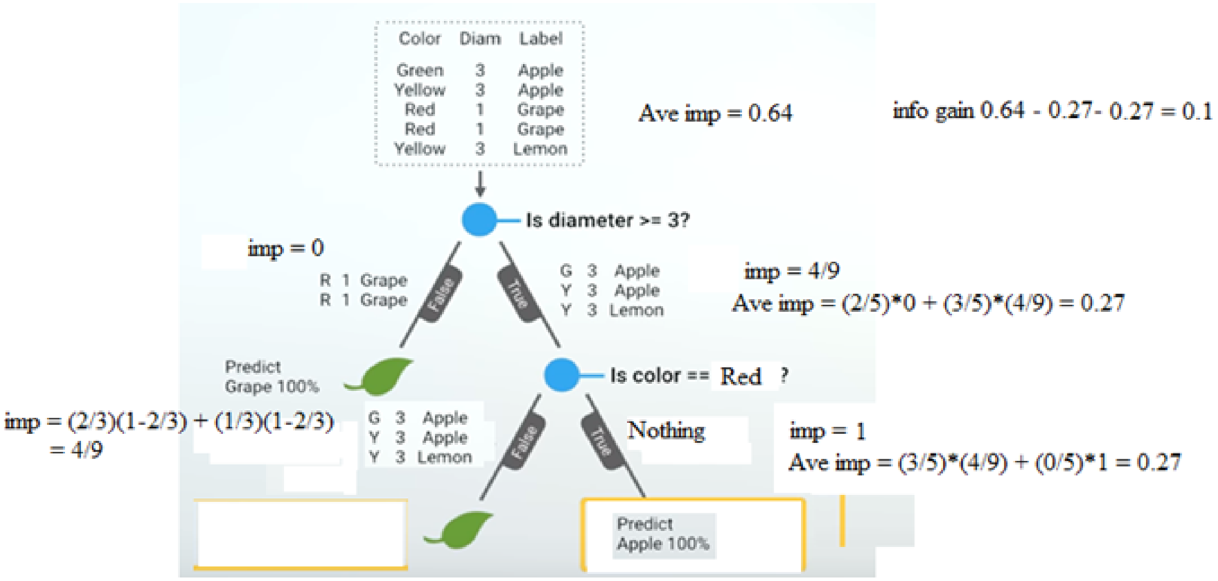
$$E(! = \text{Green}) = -\frac{1}{2}\log_2(\frac{1}{2}) - \frac{1}{2}\log_2(\frac{1}{2}) = 1$$

Weighted average of entropy:

$E\ (\text{Green?}) = \frac{2}{5} * 1 = 0.4$

Information gain: IG (diameter >= 3? , green?) = E(diameter >= 3) - E(green?)

= 0.5510 - 0.4 = ***0.1510***



Color	Diam	Label
Green	3	Apple
Yellow	3	Apple
Red	1	Grape
Red	1	Grape
Yellow	3	Lemon

<u>Is color == Red?</u>

E(== Red) = 1 (since this side of the branch cannot specify any member)

$$E(! = Green) = -\frac{2}{3}\log_2(\frac{2}{3}) - \frac{1}{3}\log_2(\frac{1}{3}) = 0.9183$$

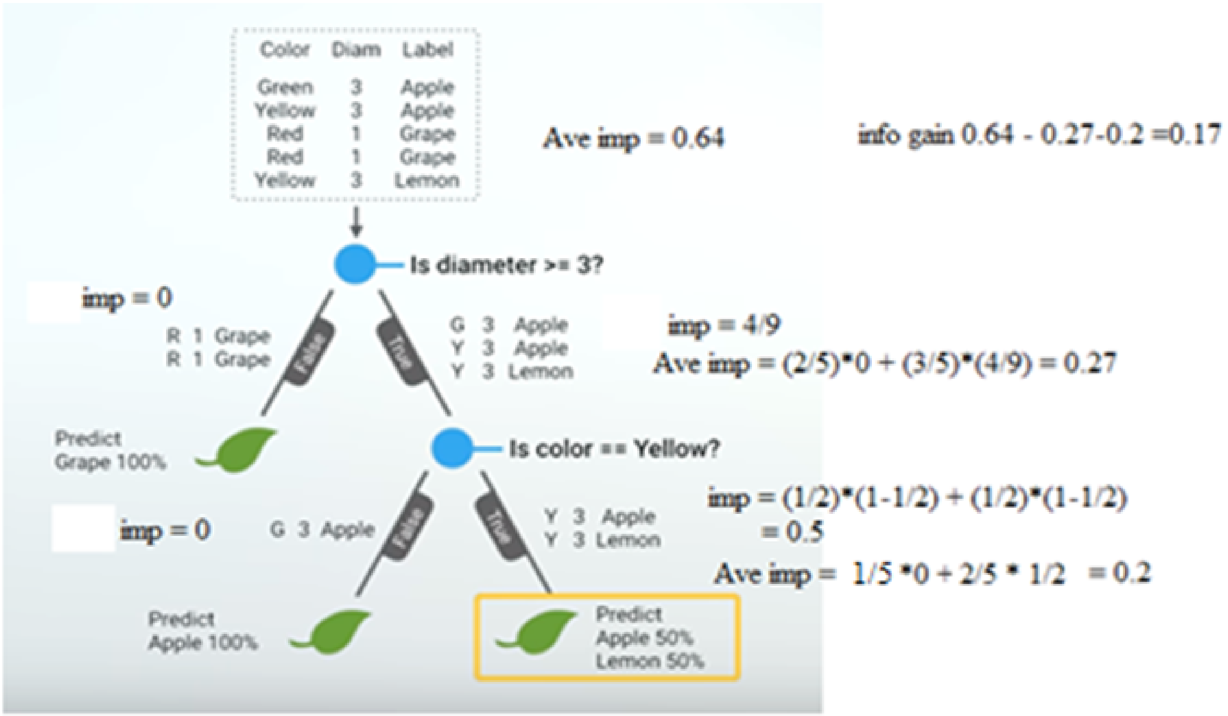
Weighted average of entropy:

$E\ (Green?)\ =\ \frac{3}{5} * 0.9183 = 0.5510$

Information gain: IG (diameter >= 3? , green?) = E(diameter >= 3) - E(green?)

= 0.5510 - 0.5510 = ***0***

So, the Gini impurity method and the entropy method will yield out the same result for the best info gain on each branch level. In this case, we can choose one of the best decision trees as follow:

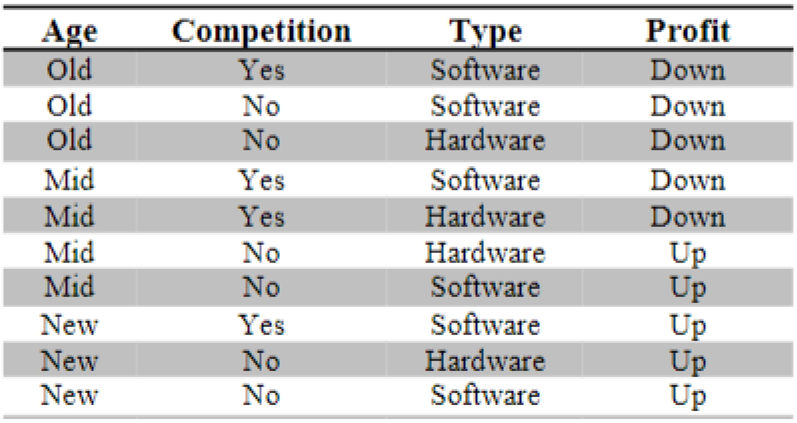


Color	Diam	Label
Green	3	Apple
Yellow	3	Apple
Red	1	Grape
Red	1	Grape
Yellow	3	Lemon

For calculation, the index of the entropy is larger than that of the Gini impurity. Therefore, when calculating the info gain with the entropy method, the result will be higher than the info gain calculated in the Gini impurity method.

2.

We have the dataset below:



Age	Competition	Type	Profit
Old	Yes	Software	Down
Old	No	Software	Down
Old	No	Hardware	Down
Mid	Yes	Software	Down
Mid	Yes	Hardware	Down
Mid	No	Hardware	Up
Mid	No	Software	Up
New	Yes	Software	Up
New	No	Hardware	Up
New	No	Software	Up

(Assume that the ID for the dataset is from 0 to 9)

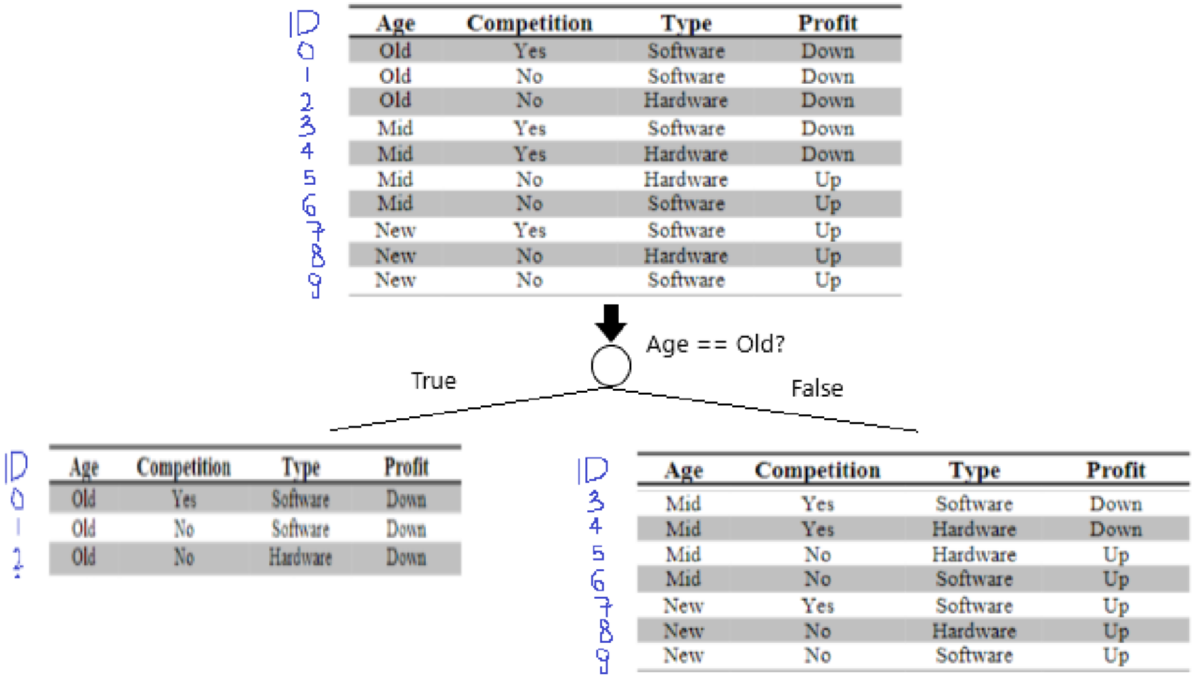
We have the imp. of root = 1 - (5/10)^2 - (5/10)^2 = ½ = 0.5

We have the condition list below:

ID	Condition list
0	Age == Old?
1	Age == Mid?
2	Age == New?
3	Competition?
4	Type?

ID 0: Age == Old?

We have the system below:



ID	Age	Competition	Type	Profit
0	Old	Yes	Software	Down
1	Old	No	Software	Down
2	Old	No	Hardware	Down
3	Mid	Yes	Software	Down
4	Mid	Yes	Hardware	Down
5	Mid	No	Hardware	Up
6	Mid	No	Software	Up
7	New	Yes	Software	Up
8	New	No	Hardware	Up
9	New	No	Software	Up

ID	Age	Competition	Type	Profit
0	Old	Yes	Software	Down
1	Old	No	Software	Down
2	Old	No	Hardware	Down

ID	Age	Competition	Type	Profit
3	Mid	Yes	Software	Down
4	Mid	Yes	Hardware	Down
5	Mid	No	Hardware	Up
6	Mid	No	Software	Up
7	New	Yes	Software	Up
8	New	No	Hardware	Up
9	New	No	Software	Up

LHS imp. = 0

RHS imp. = 1 - (2/7)^2 - (5/7)^2 = 20/49

LHS ave. imp. = 0

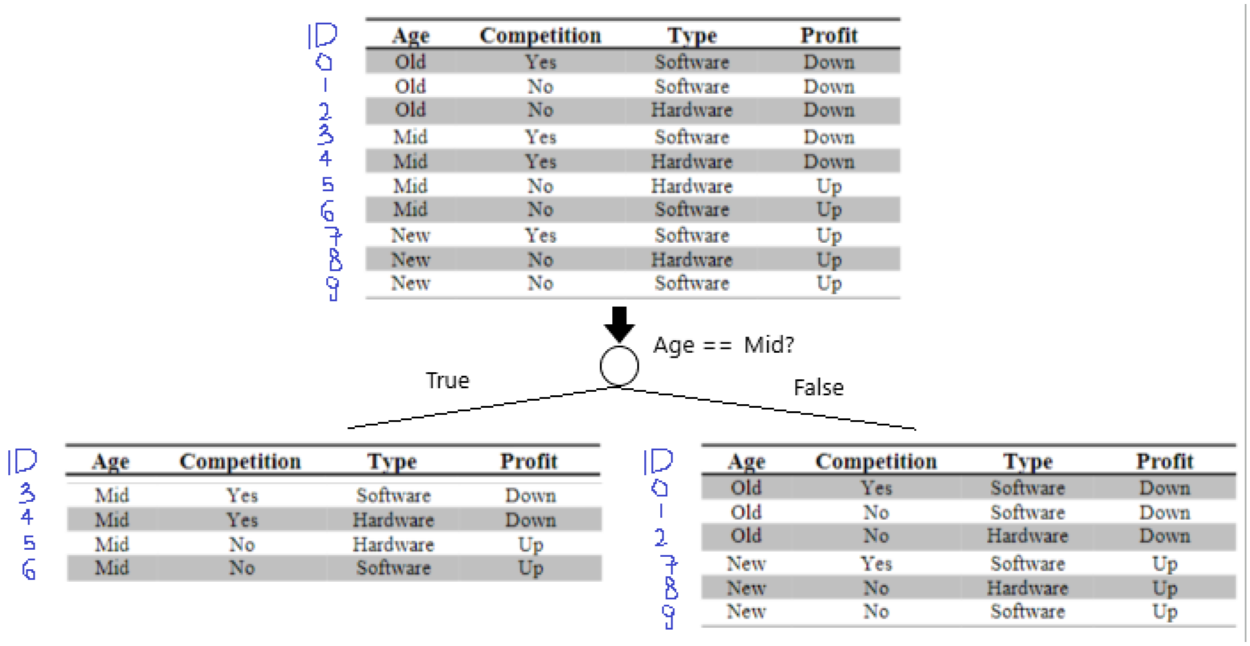
RHS ave. imp. = (7/10) * (20/49) = 2/7

Total ave. imp = 2/7

Info gain = 0.5 - 2/7 = 3/14 = 0.2143

<u>ID 1: Age == Mid?</u>

We have the system below:



ID	Age	Competition	Type	Profit
0	Old	Yes	Software	Down
1	Old	No	Software	Down
2	Old	No	Hardware	Down
3	Mid	Yes	Software	Down
4	Mid	Yes	Hardware	Down
5	Mid	No	Hardware	Up
6	Mid	No	Software	Up
7	New	Yes	Software	Up
8	New	No	Hardware	Up
9	New	No	Software	Up

ID	Age	Competition	Type	Profit
3	Mid	Yes	Software	Down
4	Mid	Yes	Hardware	Down
5	Mid	No	Hardware	Up
6	Mid	No	Software	Up

ID	Age	Competition	Type	Profit
0	Old	Yes	Software	Down
1	Old	No	Software	Down
2	Old	No	Hardware	Down
7	New	Yes	Software	Up
8	New	No	Hardware	Up
9	New	No	Software	Up

LHS imp. = 1 - (½)^2 - (½)^2 = ½

LHS ave. imp. = 0.4 * ½ = 0.2

RHS imp. = 1 - (½)^2 - (½)^2 = ½

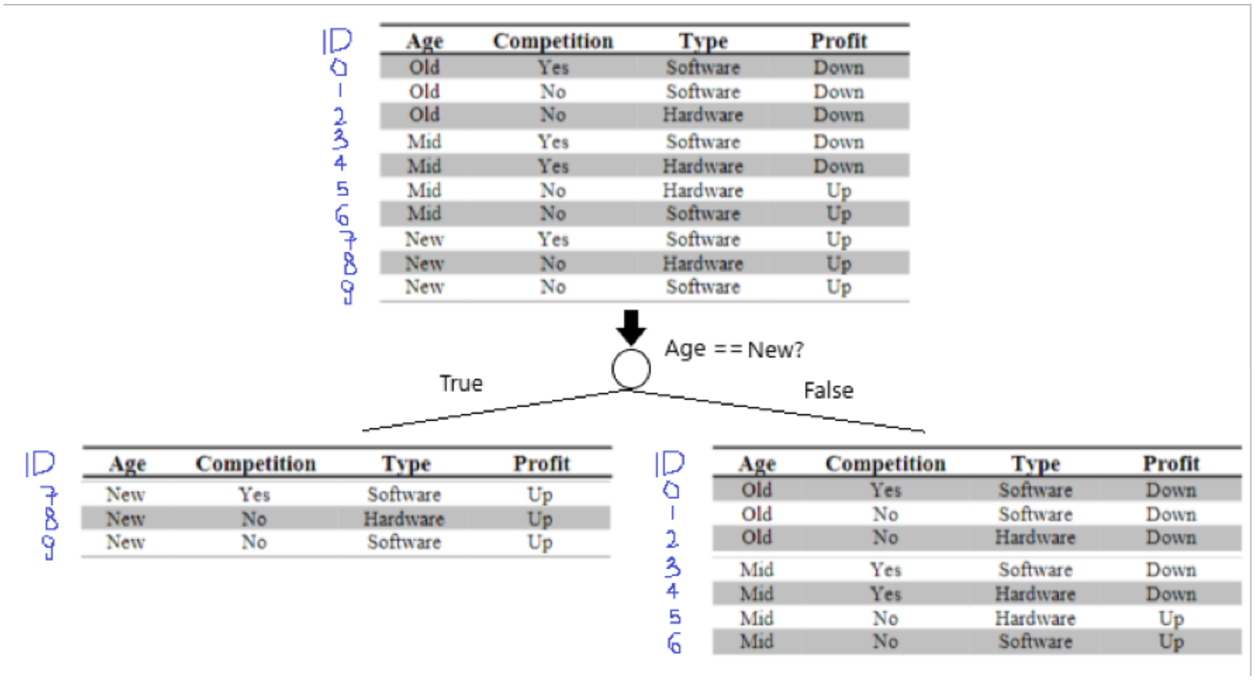
RHS ave. imp. = 0.6 * ½ = 0.3

Total ave. imp. = 0.2 + 0.3 = 0.5

Info gain = 0.5 - 0.5 = 0

<u>ID 2: Age == New?</u>

We have the system below:



ID	Age	Competition	Type	Profit
0	Old	Yes	Software	Down
1	Old	No	Software	Down
2	Old	No	Hardware	Down
3	Mid	Yes	Software	Down
4	Mid	Yes	Hardware	Down
5	Mid	No	Hardware	Up
6	Mid	No	Software	Up
7	New	Yes	Software	Up
8	New	No	Hardware	Up
9	New	No	Software	Up

ID	Age	Competition	Type	Profit
7	New	Yes	Software	Up
8	New	No	Hardware	Up
9	New	No	Software	Up

ID	Age	Competition	Type	Profit
0	Old	Yes	Software	Down
1	Old	No	Software	Down
2	Old	No	Hardware	Down
3	Mid	Yes	Software	Down
4	Mid	Yes	Hardware	Down
5	Mid	No	Hardware	Up
6	Mid	No	Software	Up

LHS imp. = 0

RHS imp. = 1 - (2/7)^2 - (5/7)^2 = 20/49

LHS ave. imp. = 0

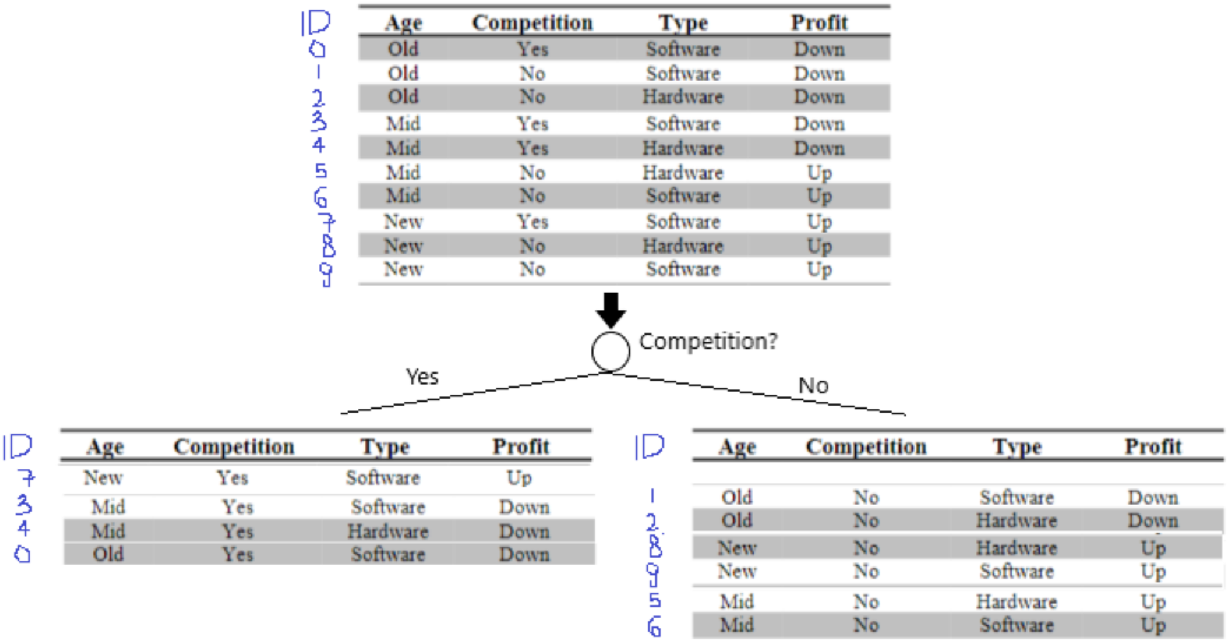
RHS ave. imp. = (7/10) * (20/49) = 2/7

Total ave. imp = 2/7

Info gain = 0.5 - 2/7 = 3/14 = 0.2143

<u>ID 3: Competition?</u>

We have the system below:



ID	Age	Competition	Type	Profit
0	Old	Yes	Software	Down
1	Old	No	Software	Down
2	Old	No	Hardware	Down
3	Mid	Yes	Software	Down
4	Mid	Yes	Hardware	Down
5	Mid	No	Hardware	Up
6	Mid	No	Software	Up
7	New	Yes	Software	Up
8	New	No	Hardware	Up
9	New	No	Software	Up

ID	Age	Competition	Type	Profit
7	New	Yes	Software	Up
3	Mid	Yes	Software	Down
4	Mid	Yes	Hardware	Down
0	Old	Yes	Software	Down

ID	Age	Competition	Type	Profit
1	Old	No	Software	Down
2	Old	No	Hardware	Down
8	New	No	Hardware	Up
9	New	No	Software	Up
5	Mid	No	Hardware	Up
6	Mid	No	Software	Up

LHS imp. = 1 - (¼)^2 - (¾)^2 = ⅜

RHS imp. = 1 - (⅓)^2 - (⅔)^2 = 4/9

LHS ave. imp. = 0.4 * ⅜ = 3/20

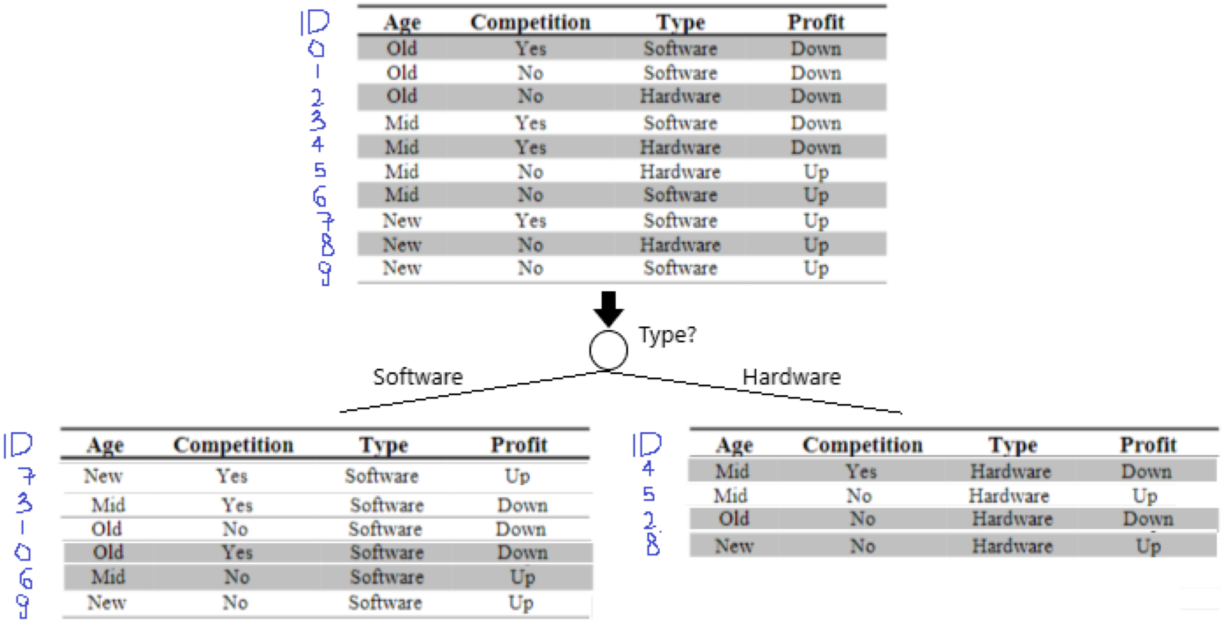
RHS ave. imp. = 0.6 * 4/9 = 4/15

Total ave. imp. = 3/20 + 4/15 = 5/12

Info gain = 0.5 - 5/12 = 1/12 = 0.0833

<u>ID 4: Type?</u>

We have the system below:



ID	Age	Competition	Type	Profit
0	Old	Yes	Software	Down
1	Old	No	Software	Down
2	Old	No	Hardware	Down
3	Mid	Yes	Software	Down
4	Mid	Yes	Hardware	Down
5	Mid	No	Hardware	Up
6	Mid	No	Software	Up
7	New	Yes	Software	Up
8	New	No	Hardware	Up
9	New	No	Software	Up

ID	Age	Competition	Type	Profit
7	New	Yes	Software	Up
3	Mid	Yes	Software	Down
1	Old	No	Software	Down
0	Old	Yes	Software	Down
6	Mid	No	Software	Up
9	New	No	Software	Up

ID	Age	Competition	Type	Profit
4	Mid	Yes	Hardware	Down
5	Mid	No	Hardware	Up
2	Old	No	Hardware	Down
8	New	No	Hardware	Up

LHS imp. = 0.5

RHS imp. = 0.5

LHS ave. imp. = 0.6 * 0.5 = 0.3

RHS imp. = 0.4 * 0.5 = 0.2

Total ave. imp. = 0.3 + 0.2 = 0.5

Info gain = 0.5 - 0.5 = 0

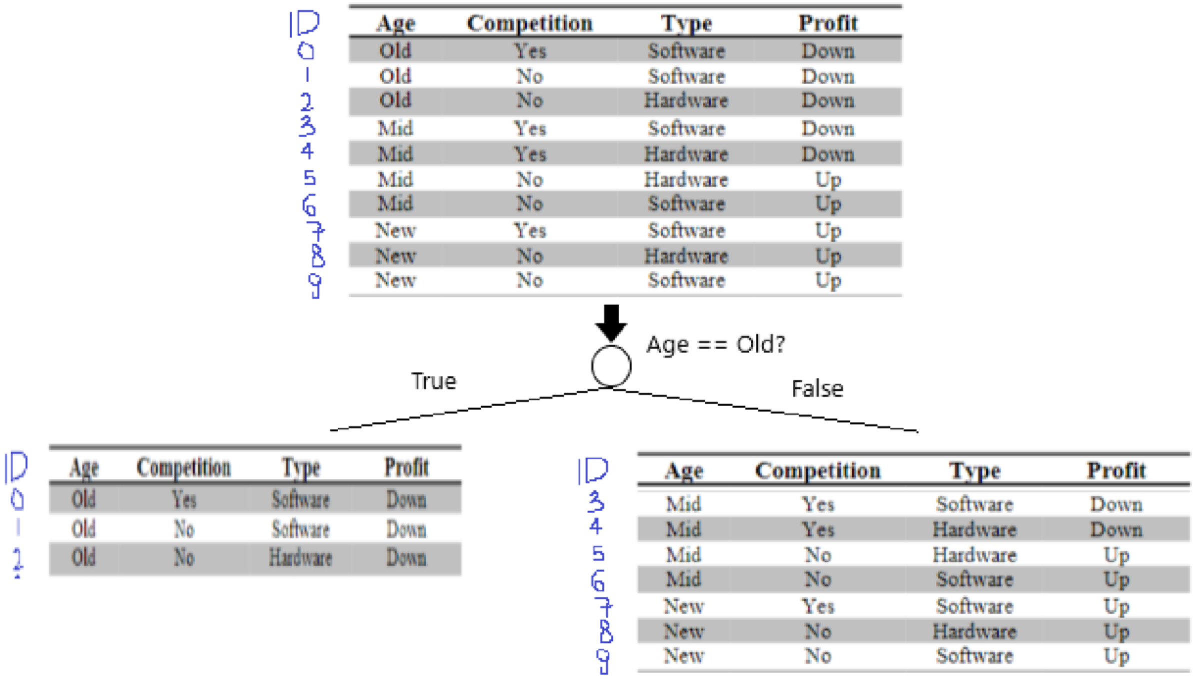
INFO GAIN COMPARISON

Age == Old?	Age == Mid?	Age == New?	Competition?	Type?
0.2143	0	0.2143	0.0833	0

Both “Age == Old?” and “Age == New?” will give the best info gain.

We choose “Age == Old?” as the next level of the decision tree.

We have the schema below:



ID	Age	Competition	Type	Profit
0	Old	Yes	Software	Down
1	Old	No	Software	Down
2	Old	No	Hardware	Down
3	Mid	Yes	Software	Down
4	Mid	Yes	Hardware	Down
5	Mid	No	Hardware	Up
6	Mid	No	Software	Up
7	New	Yes	Software	Up
8	New	No	Hardware	Up
9	New	No	Software	Up

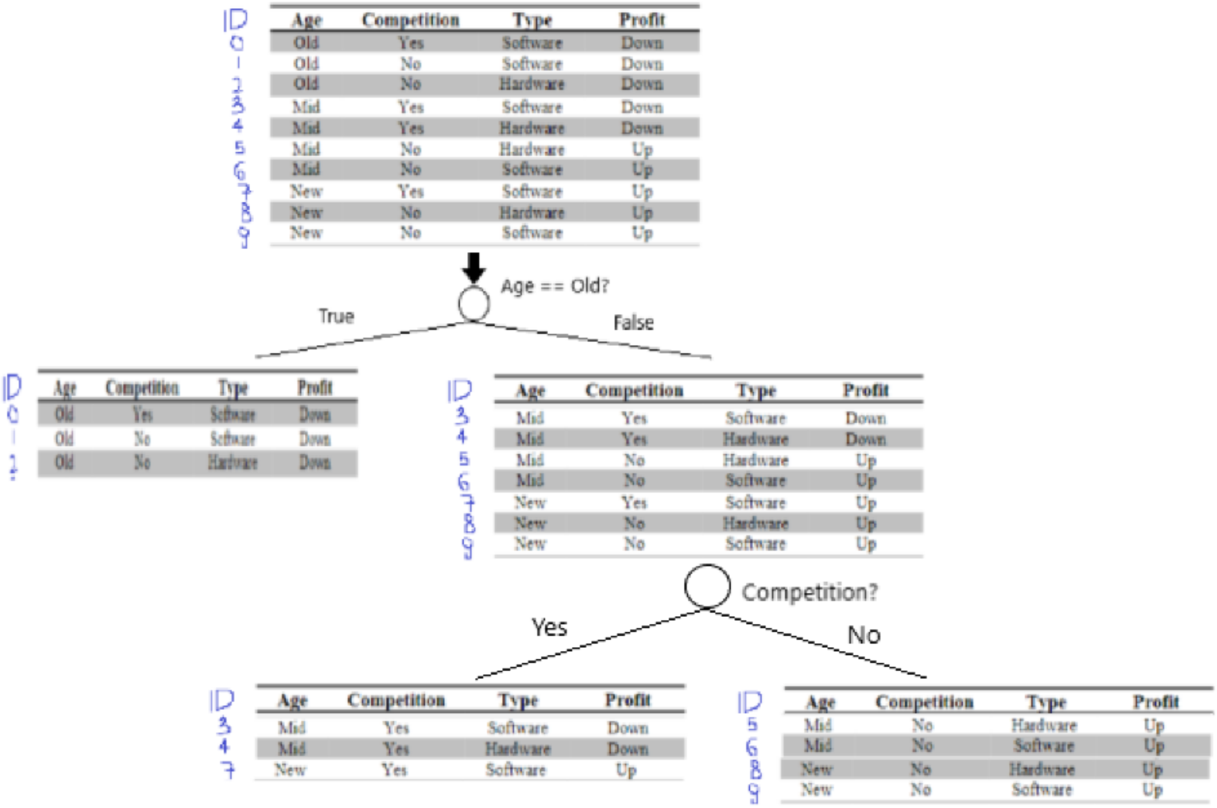
ID	Age	Competition	Type	Profit
0	Old	Yes	Software	Down
1	Old	No	Software	Down
2	Old	No	Hardware	Down

ID	Age	Competition	Type	Profit
3	Mid	Yes	Software	Down
4	Mid	Yes	Hardware	Down
5	Mid	No	Hardware	Up
6	Mid	No	Software	Up
7	New	Yes	Software	Up
8	New	No	Hardware	Up
9	New	No	Software	Up

Go to the next level by different branches and compare info gains.

<u>Competition?</u>

We have the schema below:



ID	Age	Competition	Type	Profit
0	Old	Yes	Software	Down
1	Old	No	Software	Down
2	Old	No	Hardware	Down
3	Mid	Yes	Software	Down
4	Mid	Yes	Hardware	Down
5	Mid	No	Hardware	Up
6	Mid	No	Software	Up
7	New	Yes	Software	Up
8	New	No	Hardware	Up
9	New	No	Software	Up

ID	Age	Competition	Type	Profit
0	Old	Yes	Software	Down
1	Old	No	Software	Down
2	Old	No	Hardware	Down

ID	Age	Competition	Type	Profit
3	Mid	Yes	Software	Down
4	Mid	Yes	Hardware	Down
5	Mid	No	Hardware	Up
6	Mid	No	Software	Up
7	New	Yes	Software	Up
8	New	No	Hardware	Up
9	New	No	Software	Up

ID	Age	Competition	Type	Profit
3	Mid	Yes	Software	Down
4	Mid	Yes	Hardware	Down
7	New	Yes	Software	Up

ID	Age	Competition	Type	Profit
5	Mid	No	Hardware	Up
6	Mid	No	Software	Up
8	New	No	Hardware	Up
9	New	No	Software	Up

LHS imp. = 1 - (⅓)^2 - (⅔)^2 = 4/9

RHS imp. = 0

LHS ave. imp. = 0.3 * 4/9 = 2/15

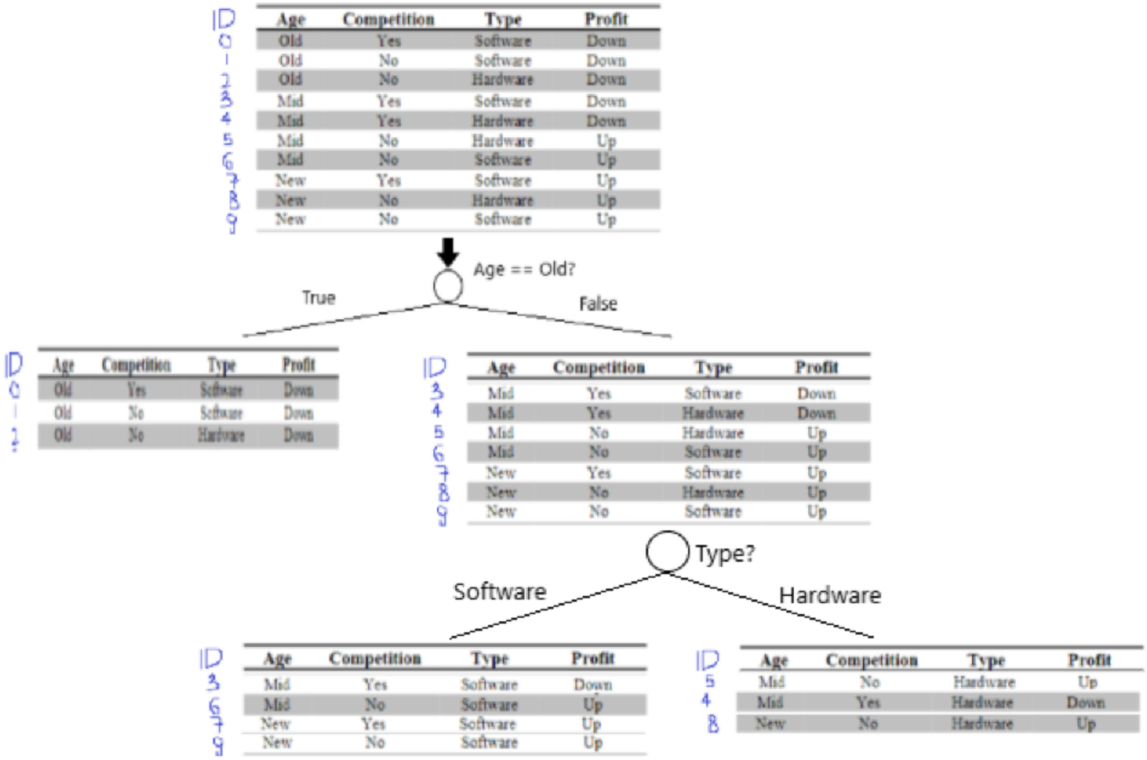
RHS ave. imp. = 0

Total ave. imp. = 2/15

Info gain = 2/7 - 2/15 = 0.1524

<u>Type?</u>

We have the schema below:



ID	Age	Competition	Type	Profit
0	Old	Yes	Software	Down
1	Old	No	Software	Down
2	Old	No	Hardware	Down
3	Mid	Yes	Software	Down
4	Mid	Yes	Hardware	Down
5	Mid	No	Hardware	Up
6	Mid	No	Software	Up
7	New	Yes	Software	Up
8	New	No	Hardware	Up
9	New	No	Software	Up

ID	Age	Competition	Type	Profit
0	Old	Yes	Software	Down
1	Old	No	Software	Down
2	Old	No	Hardware	Down

ID	Age	Competition	Type	Profit
3	Mid	Yes	Software	Down
4	Mid	Yes	Hardware	Down
5	Mid	No	Hardware	Up
6	Mid	No	Software	Up
7	New	Yes	Software	Up
8	New	No	Hardware	Up
9	New	No	Software	Up

ID	Age	Competition	Type	Profit
3	Mid	Yes	Software	Down
6	Mid	No	Software	Up
7	New	Yes	Software	Up
9	New	No	Software	Up

ID	Age	Competition	Type	Profit
5	Mid	No	Hardware	Up
4	Mid	Yes	Hardware	Down
8	New	No	Hardware	Up

LHS imp. = 1 - (¼)^2 - (¾)^2 = ⅜

RHS imp. = 1 - (⅓)^2 - (⅔)^2 = 4/9

LHS ave. imp. = 0.4 * ⅜ = 3/20

RHS ave. imp. = 0.3 * 4/9 = 2/15

Total ave. imp. = 3/20 + 2/15 = 17/60 = 0.2833

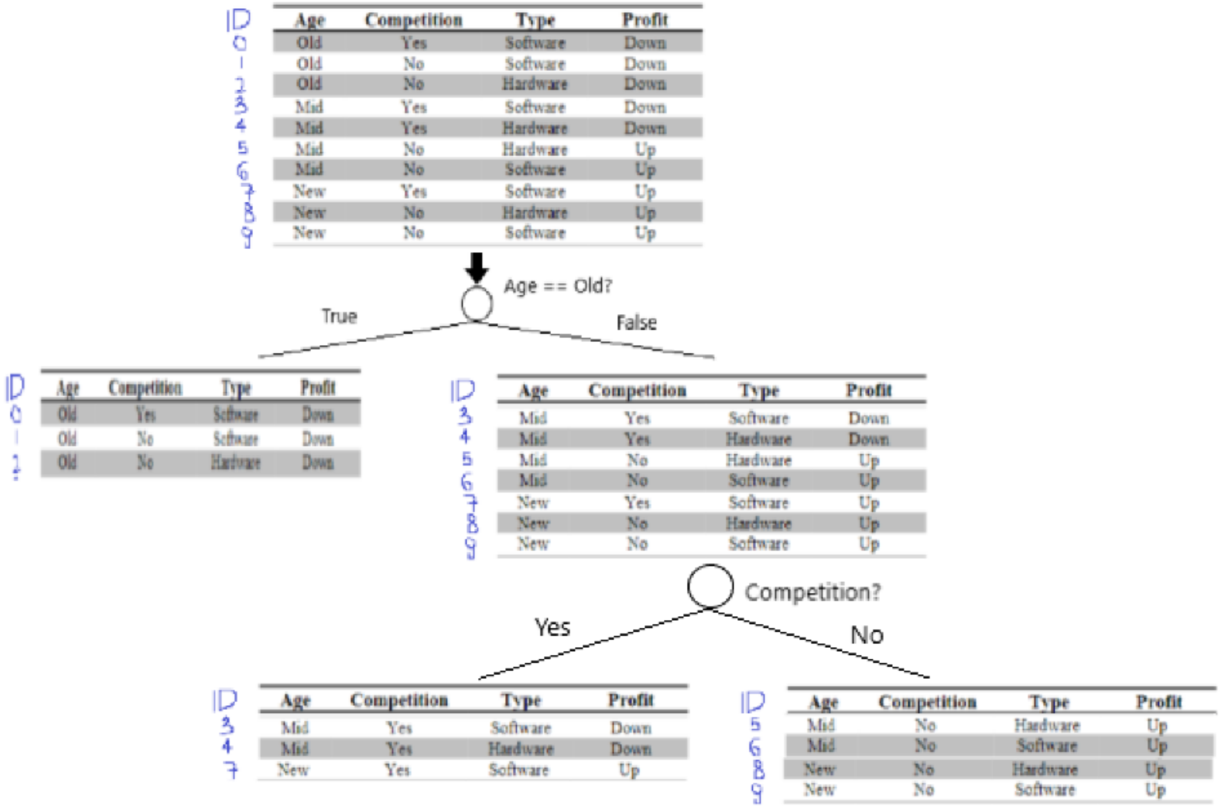
Info gain = 2/7 - 17/60 = 1/420 = 0.0024

INFO GAIN COMPARISON

Competition?	Type?
0.1524	0.0024

“Competition?” will give the best info gain. Thus, we choose “Competition?” as the next level.

We have the schema below:



ID	Age	Competition	Type	Profit
0	Old	Yes	Software	Down
1	Old	No	Software	Down
2	Old	No	Hardware	Down
3	Mid	Yes	Software	Down
4	Mid	Yes	Hardware	Down
5	Mid	No	Hardware	Up
6	Mid	No	Software	Up
7	New	Yes	Software	Up
8	New	No	Hardware	Up
9	New	No	Software	Up

ID	Age	Competition	Type	Profit
0	Old	Yes	Software	Down
1	Old	No	Software	Down
2	Old	No	Hardware	Down

ID	Age	Competition	Type	Profit
3	Mid	Yes	Software	Down
4	Mid	Yes	Hardware	Down
5	Mid	No	Hardware	Up
6	Mid	No	Software	Up
7	New	Yes	Software	Up
8	New	No	Hardware	Up
9	New	No	Software	Up

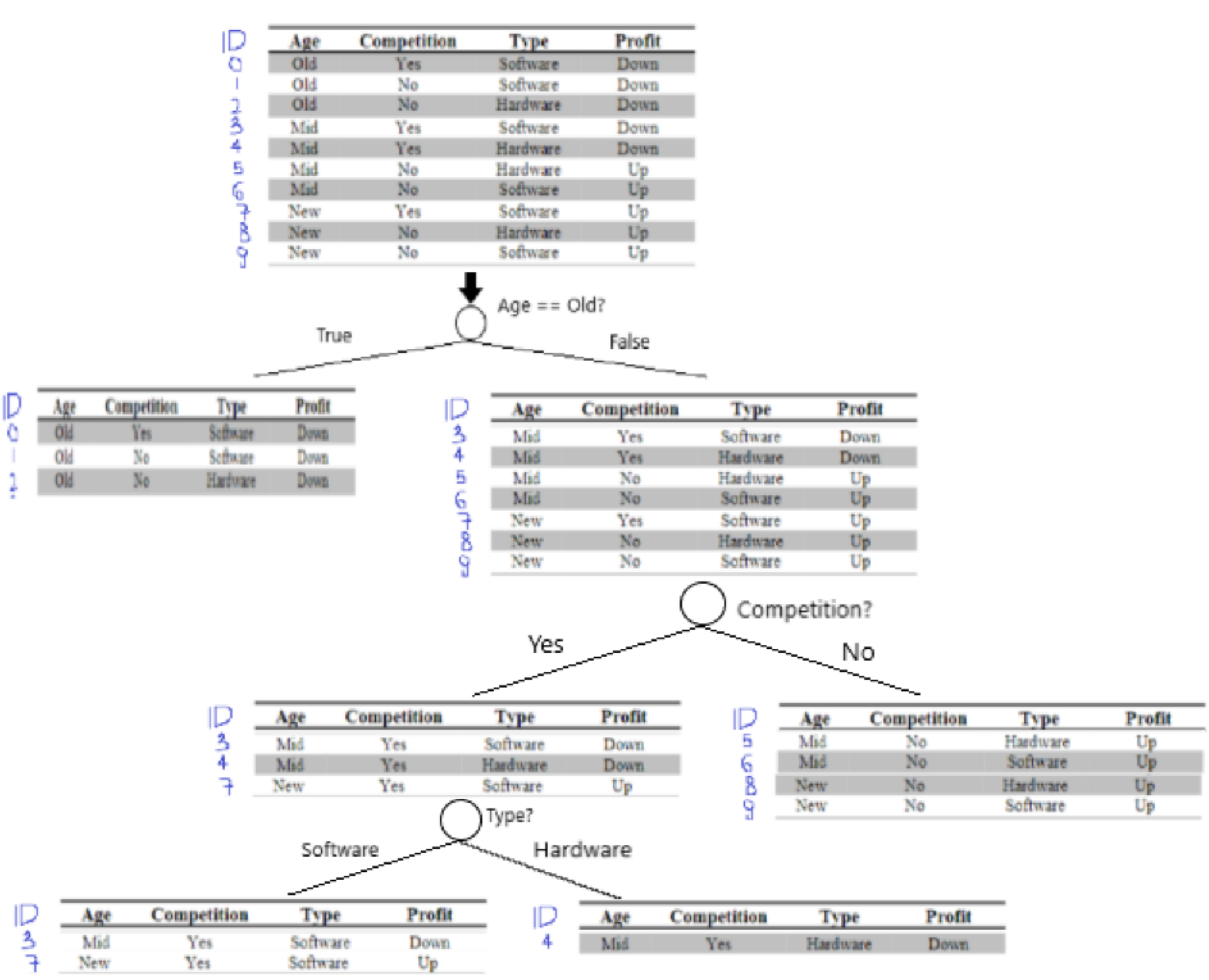
ID	Age	Competition	Type	Profit
3	Mid	Yes	Software	Down
4	Mid	Yes	Hardware	Down
7	New	Yes	Software	Up

ID	Age	Competition	Type	Profit
5	Mid	No	Hardware	Up
6	Mid	No	Software	Up
8	New	No	Hardware	Up
9	New	No	Software	Up

Try the same method and we will have “Age == Mid?” giving the best info gain.

(This can be concluded by looking at the similarity of the two classes in “Profit” - “Down” & “Up” and the class in the “Age” column. Both members of class “Down” has the same “Age == Mid”. Thus, choosing “Age == Mid?” will separate the rest of the disorder and make the impurity equal to 0)

We will have the final decision tree as below:



ID	Age	Competition	Type	Profit
0	Old	Yes	Software	Down
1	Old	No	Software	Down
2	Old	No	Hardware	Down
3	Mid	Yes	Software	Down
4	Mid	Yes	Hardware	Down
5	Mid	No	Hardware	Up
6	Mid	No	Software	Up
7	New	Yes	Software	Up
8	New	No	Hardware	Up
9	New	No	Software	Up

ID	Age	Competition	Type	Profit
0	Old	Yes	Software	Down
1	Old	No	Software	Down
2	Old	No	Hardware	Down

ID	Age	Competition	Type	Profit
3	Mid	Yes	Software	Down
4	Mid	Yes	Hardware	Down
5	Mid	No	Hardware	Up
6	Mid	No	Software	Up
7	New	Yes	Software	Up
8	New	No	Hardware	Up
9	New	No	Software	Up

ID	Age	Competition	Type	Profit
3	Mid	Yes	Software	Down
4	Mid	Yes	Hardware	Down
7	New	Yes	Software	Up

ID	Age	Competition	Type	Profit
5	Mid	No	Hardware	Up
6	Mid	No	Software	Up
8	New	No	Hardware	Up
9	New	No	Software	Up

ID	Age	Competition	Type	Profit
3	Mid	Yes	Software	Down
7	New	Yes	Software	Up

ID	Age	Competition	Type	Profit
4	Mid	Yes	Hardware	Down

For the new data, we have:

Age: Mid Competition: No Type: Hardware Profit: ?

Level 1: Age == Old? False

Level 2: Competition? No

=> Profit: Up

Python program to verify the design of the decision tree

Source code:

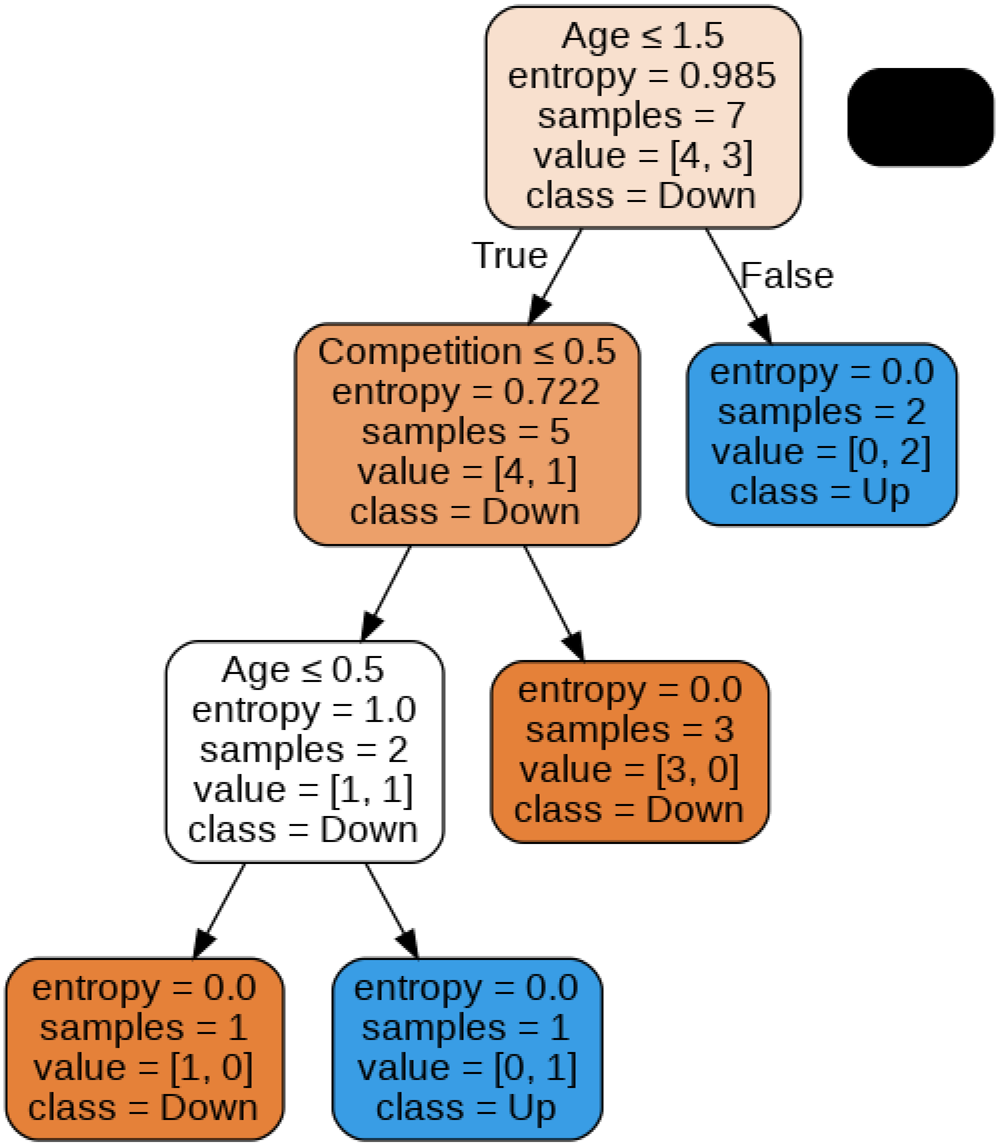
```
from google.colab import drive
drive.mount('/content/drive')
data_path = "/content/drive/My Drive/Colab Notebooks/hw4_ex2.csv"
```

```
import pandas as pd
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import train_test_split
from sklearn import metrics
col_names = ['Age', 'Competition', 'Type', 'Profit']
df = pd.read_csv(data_path, header = None, names = col_names)

# Replace the class in 'Age', 'Competition', and 'Type' columns into integers
ex2 = df.replace(regex={'Old': 0, 'Mid': 1, 'New': 2, 'Yes': 1, 'No': 0,
'Software': 0, 'Hardware':1})


feature_cols = ['Age', 'Competition', 'Type']
a = ex2[feature_cols]
b = ex2.Profit
a_train, a_test, b_train, b_test = train_test_split(a,b, test_size=0.3,
random_state=1)
clf = DecisionTreeClassifier(criterion="entropy", max_depth=3)
clf = clf.fit(a_train,b_train)
!pip install graphviz
!pip install pydotplus
from sklearn.tree import export_graphviz
from six import StringIO
from IPython.display import Image
import pydotplus
dot_data = StringIO()
export_graphviz(clf, out_file=dot_data, filled=True, rounded=True,
special_characters=True, feature_names =
feature_cols,class_names=['Down','Up'])
graph = pydotplus.graph_from_dot_data(dot_data.getvalue())
graph.write_png('hw4_ex2.png')
Image(graph.create_png())
```

Run program & result:



Thus, it has the same structure to the decision tree calculated by hand.