ARTIFICIAL INTELLIGENCE

DECISION TREES

CS-632

Dr. Muhammad Aqib

University Institute of Information Technology

PMAS-Arid Agriculture University Rawalpindi

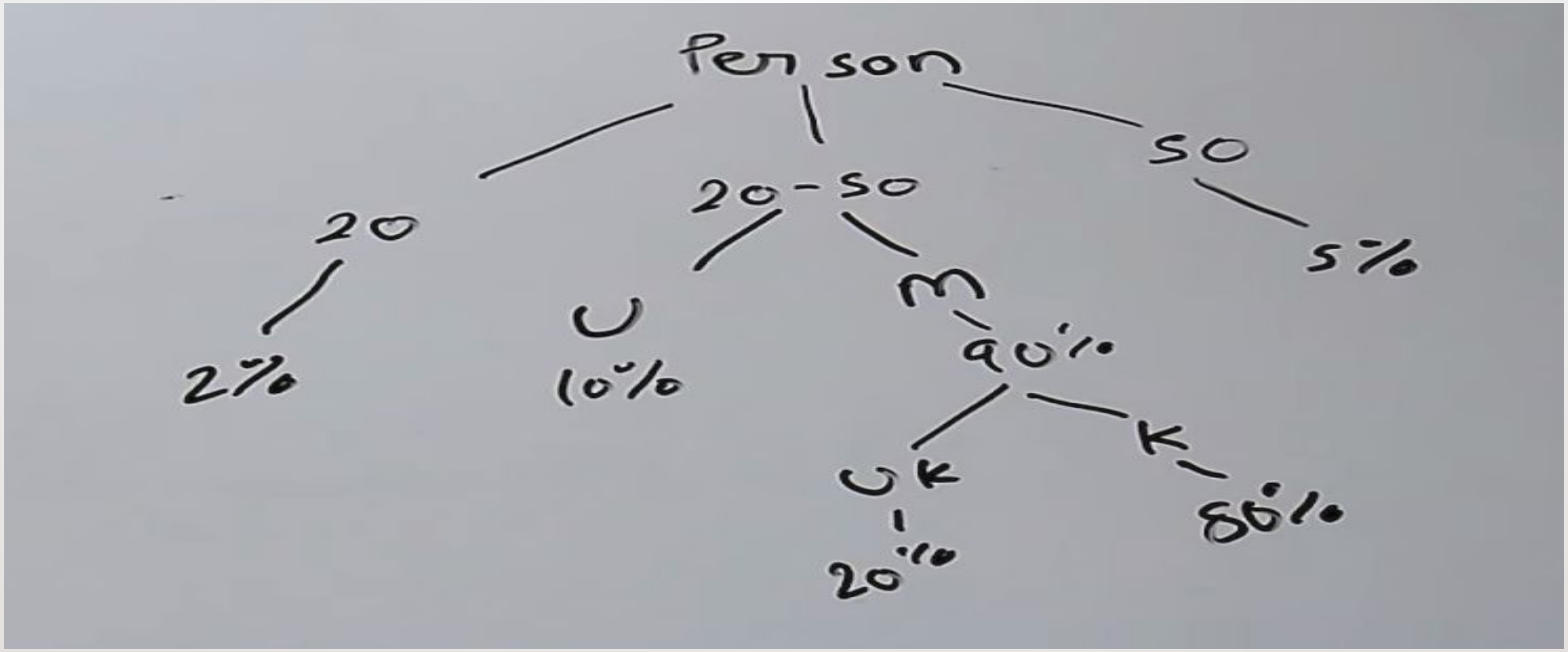
DECISION TREES

- Falls in the category of supervised learning.
- Can be used for both, regression and classification problems.
- Uses the tree representation to solve the problem.
- Each leaf node corresponds to a class label.
- Attributes are represented on the internal node of the tree.
- Boolean function on discrete attributes can be represented using the decision tree.

DECISION TREE

- Used for decisions
- Understand by example
    - A loan company want to know the persons who will be interested in loans
    - Categories on basis of Age

DECISION TREE



- Lets suppose you have been given a problem to find whether your company will be in profit or loss?

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

- First of all, find root node
  - Some formulas to identify
  - Last attribute is class attribute and will be leaf of tree
  - Calculate Entropy of class attribute

ENTROPY CALCULATION

- Lets have P and N
  - P is positive signs
  - N for Negative signs

  - P = 5 → Up

  - N = 5 → Down

Entrophy (class)

$$= \frac{-P}{P+N} \log_2\left(\frac{P}{P+N}\right) - \frac{N}{P+N} \log_2\left(\frac{N}{P+N}\right)$$

$P = 5$ $\quad N = 5$

$$= \frac{-5}{5+5} \log_2\left(\frac{5}{10}\right) - \frac{5}{5+5} \log_2\left(\frac{5}{5+5}\right)$$

$$= \frac{-5}{10} \log_2\left(\frac{5}{10}\right) - \frac{5}{10} \log_2\left(\frac{5}{10}\right)$$

$= \quad 1$

- Now we calculate 3 things
- First of all Information gain

For each attribute

$$I(P_i,N_i) = \frac{-P}{P+N}\log_2\left(\frac{P}{P+N}\right) - \frac{N}{P+N}\log_2\left(\frac{N}{P+N}\right)$$

  - When we use a node in a decision tree to partition the training instances into smaller subsets the entropy changes. Information gain is a measure of this change in entropy.

Entrophy (attribute)

$$= \frac{\sum P_i + N_i}{P+N}\left(I(P_i,N_i)\right)$$

- Second Calculate Entropy
  - Entropy is the measure of uncertainty of a random variable, it characterizes the impurity of an arbitrary collection of examples. The higher the entropy more the information content.
- Third

Gain

$$= \underset{\text{Class}}{\text{Entrophy}} - \text{Entrophy (attribute)}$$

- Now will compare gain
    - Maximum gain attribute will be root

Age	P_i	N_i	$I(P_i , N_i)$
Old	0	3	0
Mid	2	2	1
New	3	0	0

How many Up in Class attribute

How many Down in Class attribute

- If one of the number is zero then information gain is zero
- If Both are same then Information gain is 1

ENTROPY OF COMPLETE AGE ATTRIBUTE

- Now will calculate Entropy of Age attribute

$$\text{Entrophy (attribute)} = \frac{\sum P_i + N_i}{P+N} \left(I(P_i, N_i) \right)$$

$$= \cancel{\frac{0+3}{5+5}(0)} + \frac{2+2}{5+5}(1) + \cancel{\frac{3+0}{5+5}(0)}$$

$= 0.4$

NOW GAIN

Gain
= Entrophy Class − Entrophy (attribute)

- Gain = Entropy of Class - Entropy (Age)

  = 1 - 0.4

  = 0.6

- Now will calculate for rest of the attributes (Competition)

$$I(P_i, N_i)_{Yes} = \frac{-1}{1+3}\log_2\left(\frac{1}{1+3}\right) - \frac{3}{1+3}\log_2\left(\frac{3}{1+3}\right)$$

$$= \frac{-1}{4}\log_2\left(\frac{1}{4}\right) - \frac{3}{4}\log_2\left(\frac{3}{4}\right)$$

$$= 0.81127$$

$$I(P_i, N_i) = \frac{-4}{6}\log_2\left(\frac{4}{6}\right) - \frac{2}{6}\log_2\left(\frac{2}{6}\right)$$

$$= 0.918295$$

Competition

	Pi	Ni	I(Pi, Ni)
Yes	1	3	
No	4	2	

$I(P_i, N_i)_{Yes} = 0.81127$

$I(P_i, N_i)_{No} = 0.918295$

Entrophy (Competition)

$$= \frac{(1+3)}{5+5}(0.81127) + \frac{(4+2)}{5+5}(0.918295)$$

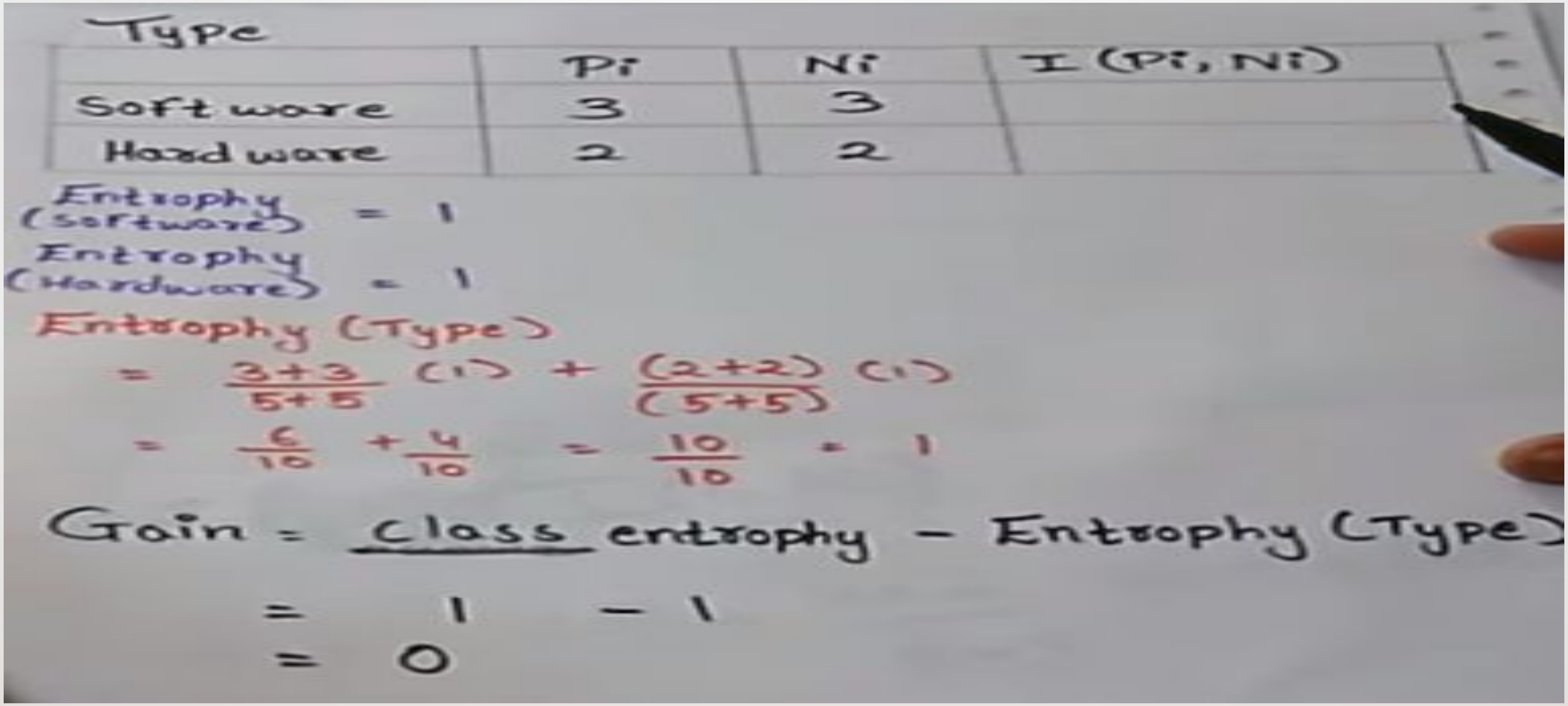
$$= 0.8754$$

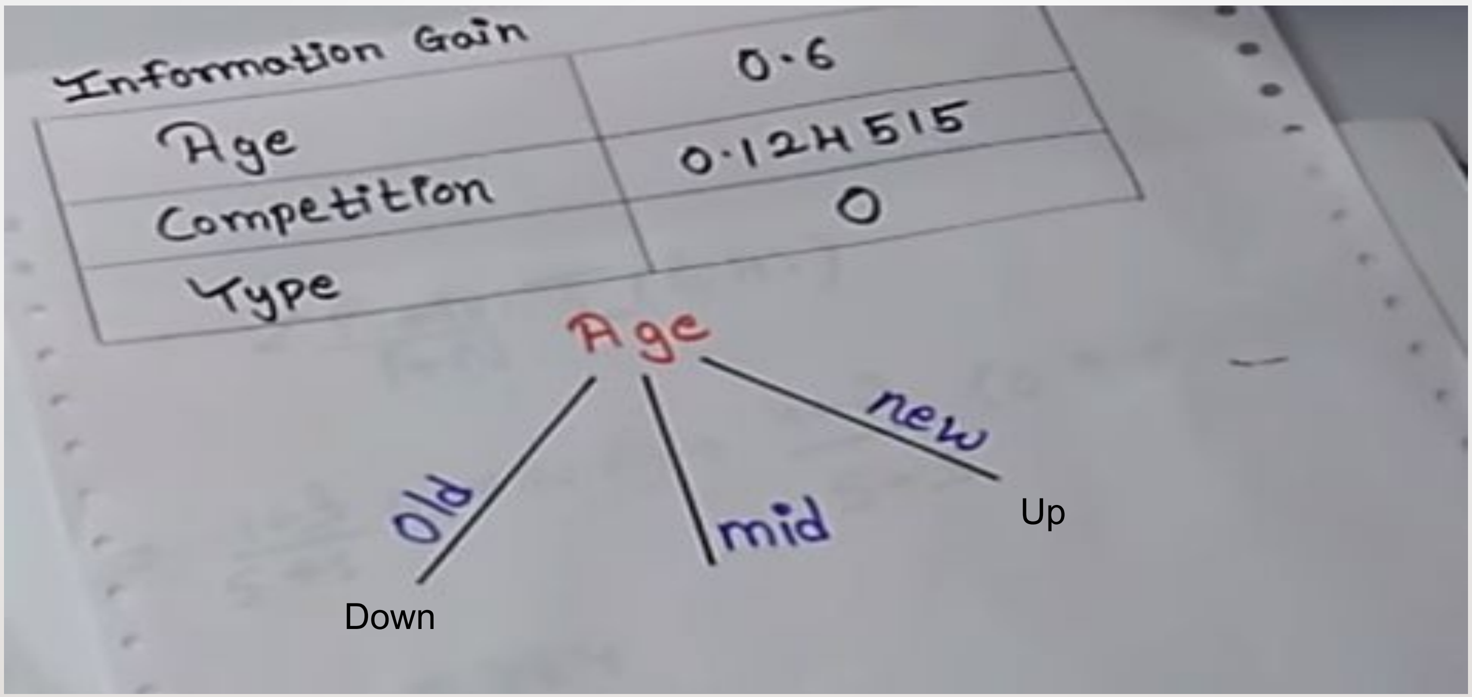
Gain = Class Entrophy − Entrophy (competition)

$$= 1 - 0.8754$$

= 0.124515

TYPE


Type
Pi
Ni
I (Pi, Ni)
Software
3
3
Hardware
2
2
Entrophy (software) = 1
Entrophy (Hardware) = 1
Entrophy (Type)
= 3+3 / 5+5 (1) + (2+2) / (5+5) (1)
= 6/10 + 4/10 = 10/10 = 1
Gain = Class entrophy − Entrophy (Type)
= 1 − 1
= 0



?

It can be either Type or Competition

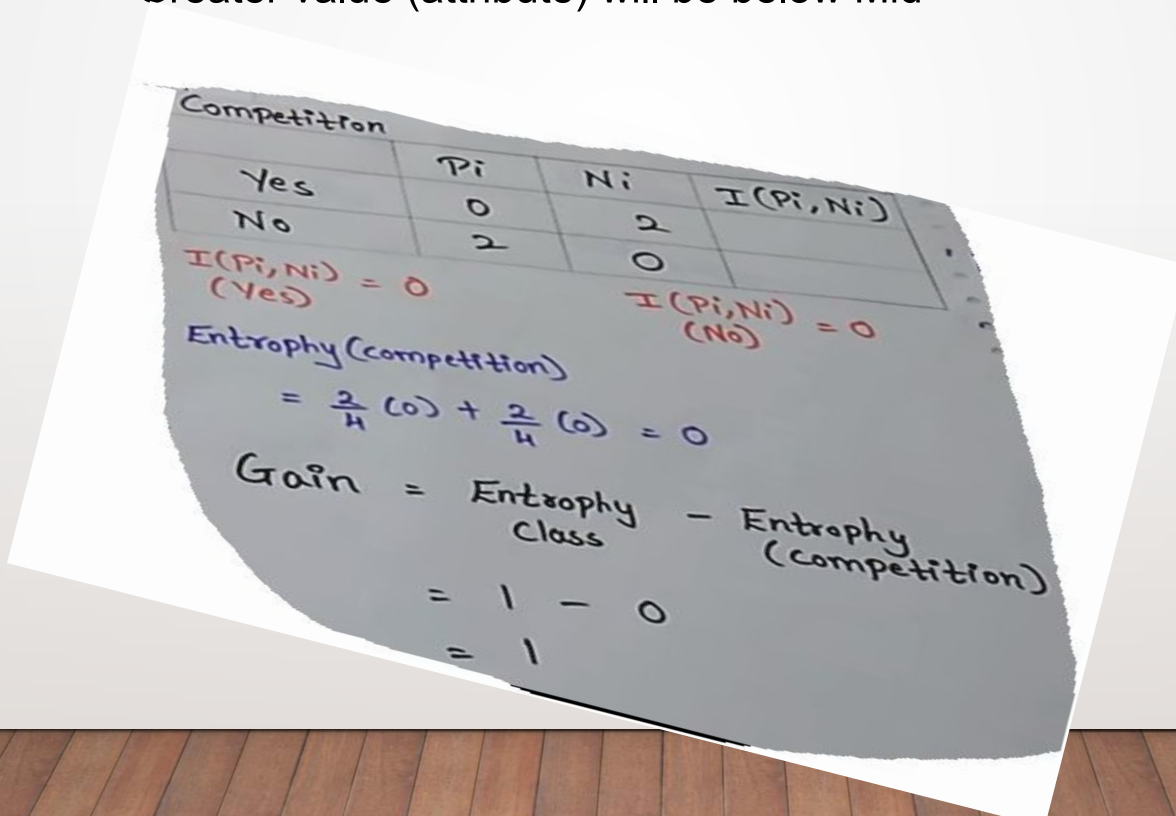
- Now create table of Mid part only

Age	Competition	Type	Profit
mid	Yes	Software	Down
mid	Yes	Hardware	Down
mid	No	Hardware	Up
mid	No	Software	Up

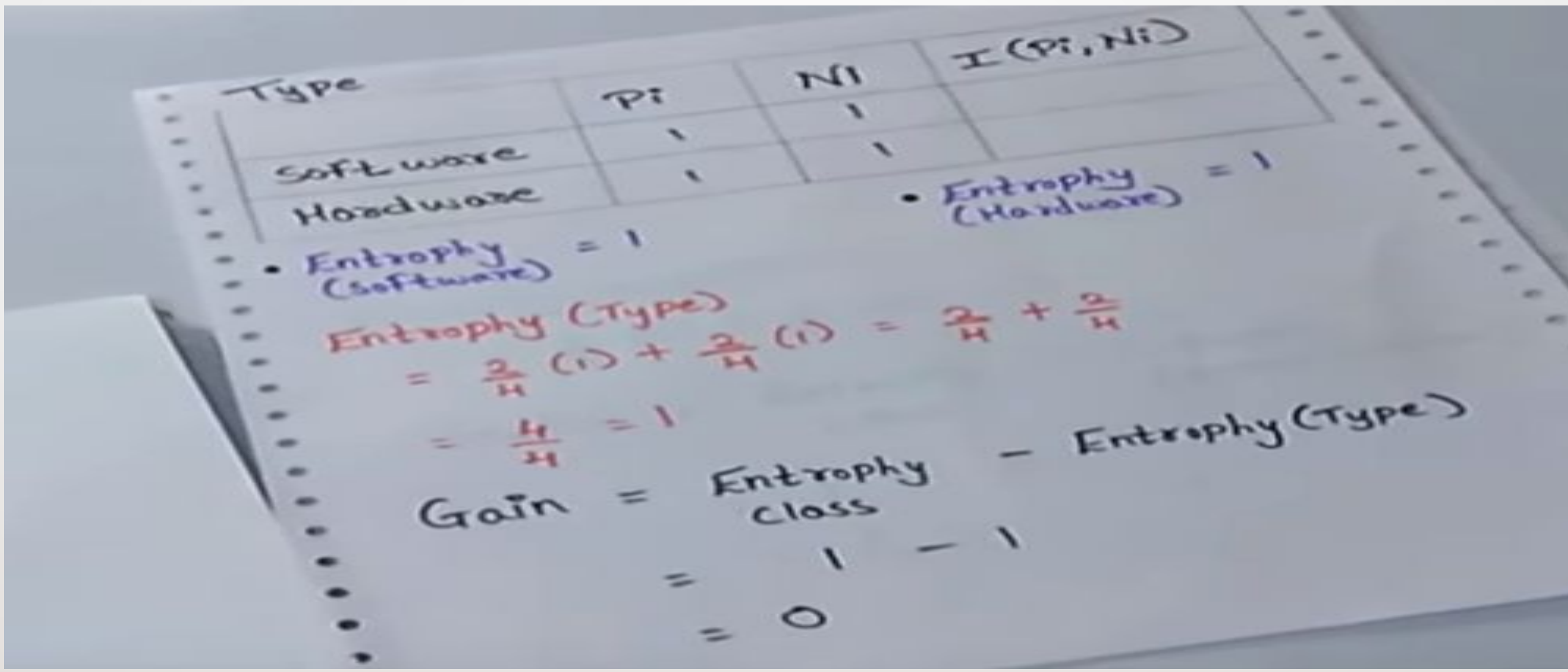
P = 2 , N = 2

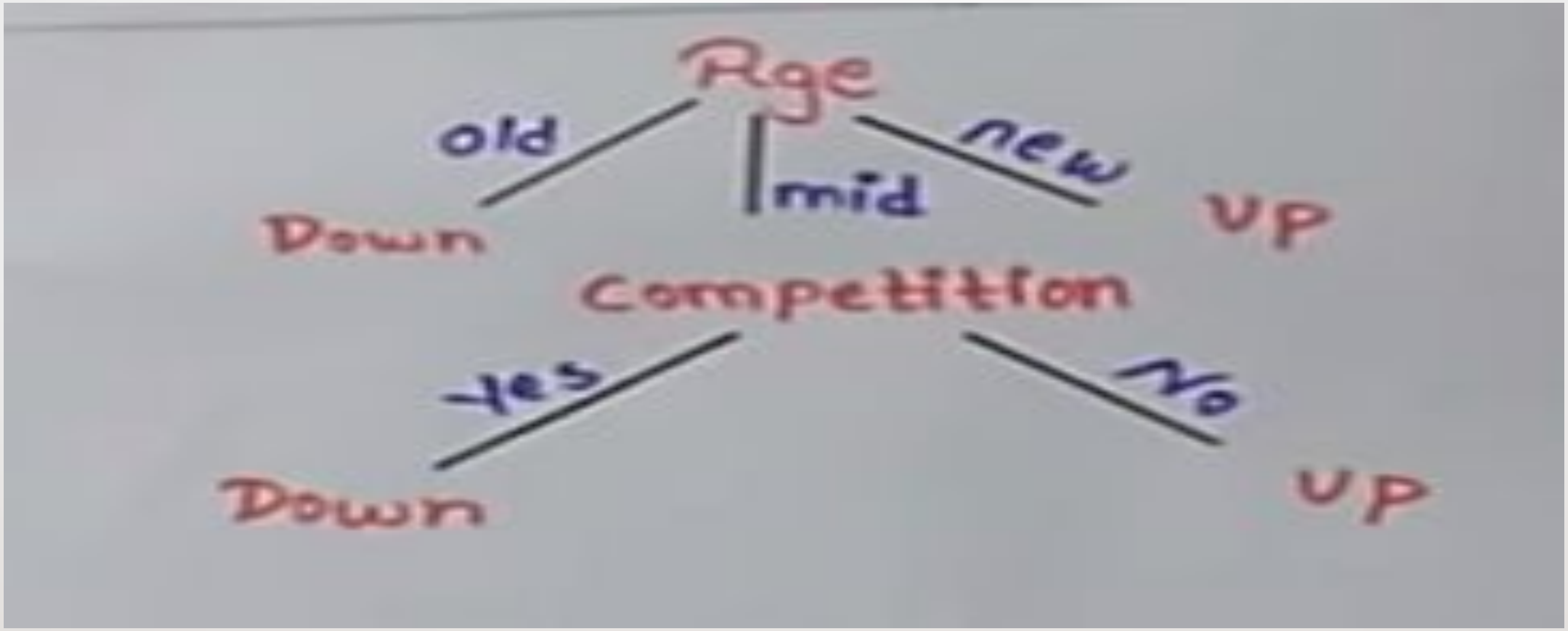
Entrophy (Profit) = 1

- Now find information gain of Competition and type
  - Greater value (attribute) will be below Mid



GAIN OF TYPE


Type | Pi | Ni | I(Pi, Ni)
Software | 1 | 1
Hardware | 1 | 1
Entrophy (Software) = 1
Entrophy (Hardware) = 1
Entrophy (Type) = 2/4 (1) + 2/4 (1) = 2/4 + 2/4
= 4/4 = 1
Gain = Entrophy class − Entrophy (Type)
= 1 − 1
= 0


Rge
old
mid
new
Down
UP
Competition
Yes
No
Down
UP

Thank You