

Decision Tree

attributes

→ Let's take Profit as Target Variable

Page No.

Date:

$$P: up = 5$$

$$P: down = 5$$

$$\text{1) } \rightarrow I(P, N) = -\frac{P}{P+N} \log_2 \left(\frac{P}{P+N} \right) - \frac{N}{P+N} \log_2 \left(\frac{N}{P+N} \right)$$

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

$$= 1$$

Information gain of overall dataset is equal to 1.

Now, calculating the entropy of each attribute

2) Entropy (Age)

age	P_i	n_i	$I(P_i, n_i)$
old	0	3	$I(0, 3) = 0$
mid	2	2	$I(2, 2) = 1$
new	3	0	$I(3, 0) = 0$

$$E(A) = \sum_{i=1}^v \frac{P_i + n_i}{P+N} (I(P_i, n_i))$$

$$\begin{aligned} E(A) &= \frac{0+3(0)}{5+5} + \frac{2+2(1)}{5+5} + \frac{3+0(0)}{5+5} \\ &= 0 + 0.4 + 0 \\ &= 0.4 \end{aligned}$$

Entropy of age is 0.4

$$\text{Gain (Age)} = Ig - \text{Entropy(Age)}$$

$$= 1 - 0.4$$

$$= 0.6$$

3) Entropy (Competition)

competition	P_i	n_i	$I(P_i, n_i)$
-------------	-------	-------	---------------

$$\text{Yes} \quad 1 \quad 3 \quad 0.81$$

$$\text{No} \quad 4 \quad 2 \quad 0.91$$

$$E(c) = \sum_{i=1}^v \frac{P_i + n_i}{P+N} (I(P_i, n_i))$$

$$= \frac{1+3}{5+5} (0.81) + \frac{4+2}{5+5} (0.91) = 0.8754$$

$$G(\text{competition}) = 14 - \text{Entropy}(C)$$

$$= 1 - 0.8754$$

$$= 0.125$$

Page No.

Date :

3) Entropy (Type)

type	p_i	n_i	$I(p_i, n_i)$	
software	3	3	1	new
hardware	2	2	1	new

$$E(T) = \sum_{i=1}^V \frac{p_i + n_i}{P+N} (I(p_i, n_i))$$

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

$$\text{Gain}(T) = 14 - \text{Entropy}(T)$$

$$= 1 - 1$$

$$= 0$$

Gain of age = 0.6

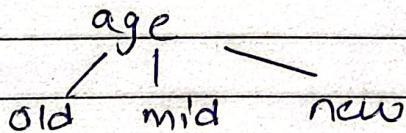
Gain of competition = 0.125

Gain of type = 0

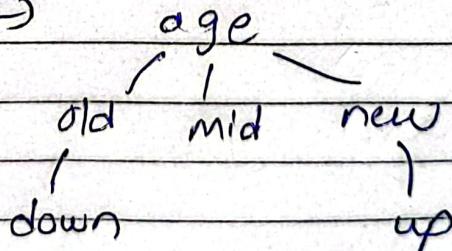
gain of age is maximum, therefore it will be our root node.

age

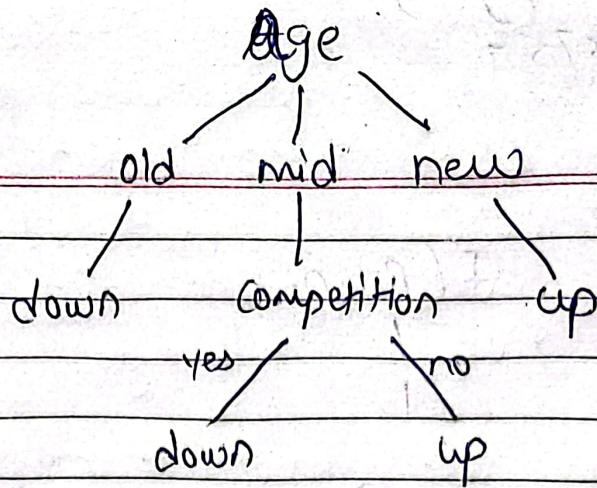
age has 3 attributes old, mid, new



as our data for old all is down & for new all is up. therefore,



mid has both up & down, so next highest gain is competition



Q1 Star

Date
Day
Month
quarter
Year

Customer
Customer ID
Customer Name
Gender
Age
Membership Level

Employee
Emp ID
Emp Name
Title
Department
Region

Sales
Sale ID
Product ID
Store ID
Customer ID

Store
Store ID
Store Name
Grade Level

2) Slow flake

Time
Order ID
Year
Quarter
Month

Sales
Sale ID
Product ID
Store ID
Customer ID

Product
Product ID
Product Name
Product category ID

Product Category
ID name
Description
Unit price

City
City ID
City Name
State
Country

Customer
Customer ID
Customer Name
Address
City ID

Employee
Emp ID
Emp Name
Region
Department ID

Department
Department ID
Name
Location

3) fact constellation

Location

Street

City

Country

Page No.

Date:

Time	Sales	Shipping	Shippers
Order ID	Sale ID	item key	Shipper-key
Year	Product ID	location	location
Quarter	Shipper ID	units-shipped	Shipper-type
Month	Customer ID	Shipper-key	

Customer	Employee	Department
Customer ID	Emp ID	DepartmentID
Customer Name	Emp Name	Name
Address	Region	Location
City ID	DepartmentID	DepartmentID

Q2) OLAP operation

DM Table

Time (Year)	Location (Country, State)	Product (Category, Product Name)	Sales Amount
2024	USA, California	Electronics, TV	5000
2024	USA, California	Electronics, Laptop	7000
2024	USA, Texas	Furniture, Sofa	3000
2025	USA, California	Electronics, TV	6000
2025	India, Maharashtra	Electronics, Mobile	8000
2025	India, Maharashtra	Furniture, Chair	2000
2025	India, Gujarat	Furniture, Table	2500

1) Slice

Time = 2025

Time	Location	Product	Sales Amount
2025	USA, California	Electronics, TV	6000
2025	India, Maharashtra	Electronics, Mobile	8000
2025	India, Maharashtra	Furniture, Chair	2000
2025	India, Gujarat	Furniture, Table	2500

2) Dice - selects two or more dimension & creates a smaller sub-cube.

Page No. _____

Country = India And Product Category = Furniture

Date: _____

Time	Location	Product	Amount
2025	Maharashtra	Chair	2000
2025	Gujarat	Table	2500

3) Drill Down - going from higher-level summary to detailed down data.

Drill down from Year → Quarter → Month
assuming

Time	Location	Product	Sales Amount
Q1 - 2025	USA, California	Electronics, TV	6000
Q2 - 2025	India, Maharashtra	Electronics, Mobile	8000
Q3 - 2025	India, Maharashtra	Furniture, Chair	2000
Q4 - 2025	India, Gujarat	Furniture, Table	2500

4) Roll-up - summarizing or grouping data

Roll-up from City → State → Country

Country	Amount
USA	14000
India	12500

5) Pivot - means rotating the view to see data from a diff perspective

Product Name	USA Sales	India Sales
TV	11000	20000
Laptop	7000	15000
Sofa	3000	10000
Mobile	-	8000
Chair	-	2000
Table	-	2500