Decision Tree

Introduction

- Non-parametric → Non-linear data
- white box model → Inference
- Mother of all tree based algo
- Work with both classification and Regression.

Intuition behind DT

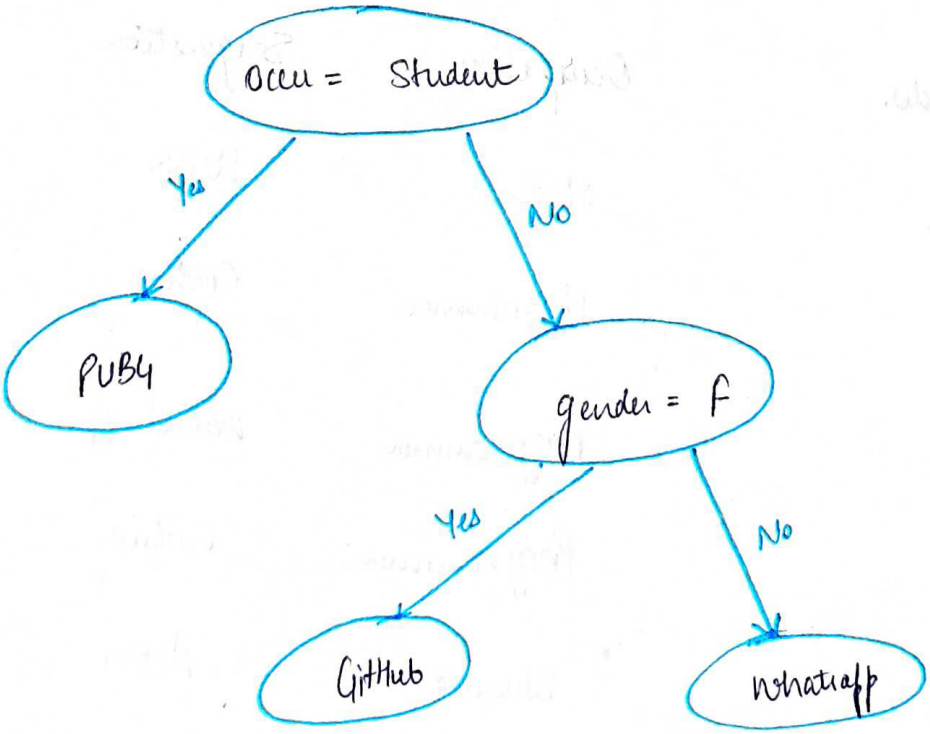
Gender	Occupation	Suggestion
F	student	PUBG
F	Programmer	Github
M	Programmer	Whatsapp
F	Programmer	Github
M	Student	PUBG
M	Student	PUBG

Giant Nested if - else structure

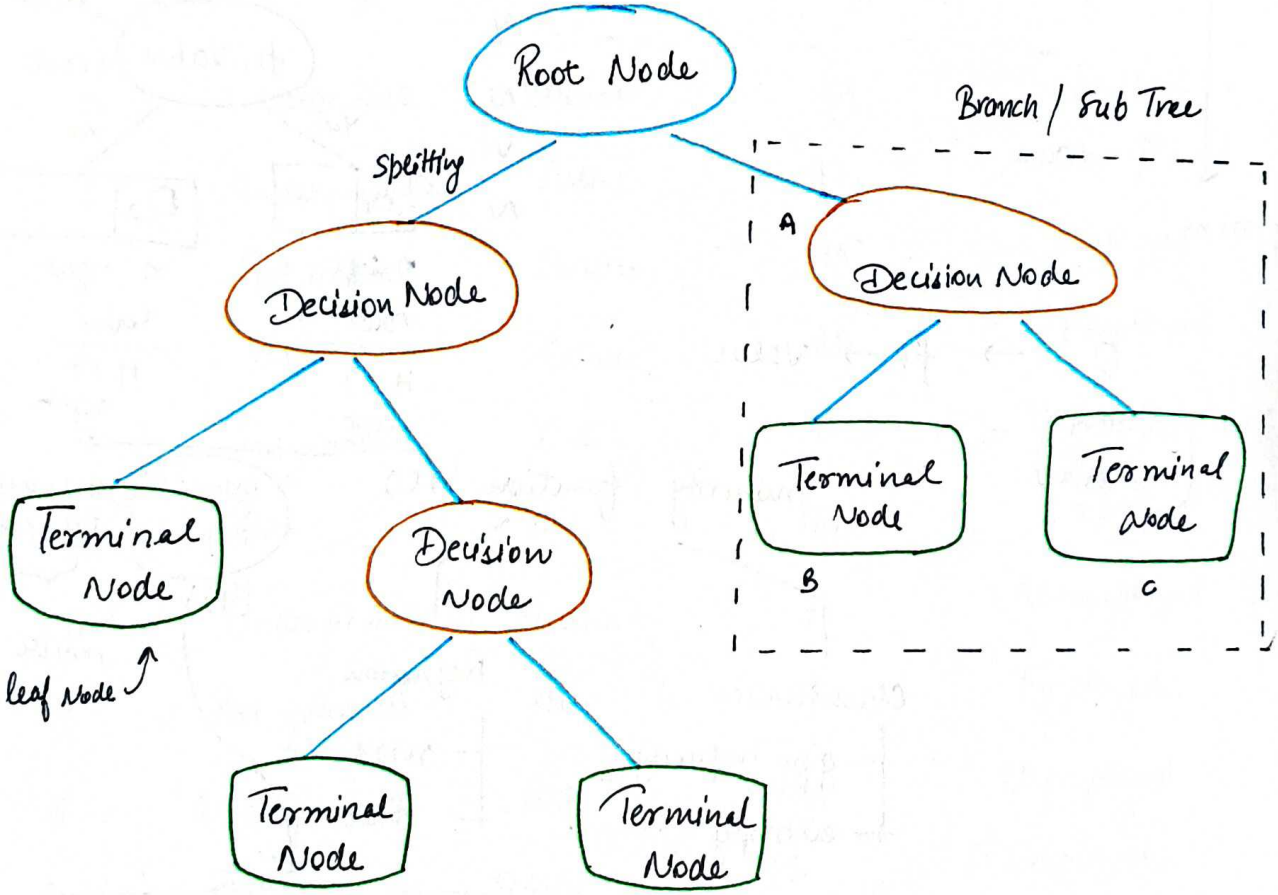
ex:-

```
if occupation = student
    print ("PUBG")
else:
    if gender = f
        print ("Github")
    else:
        print ("whatsapp")
```

basic decision Tree



Vocab



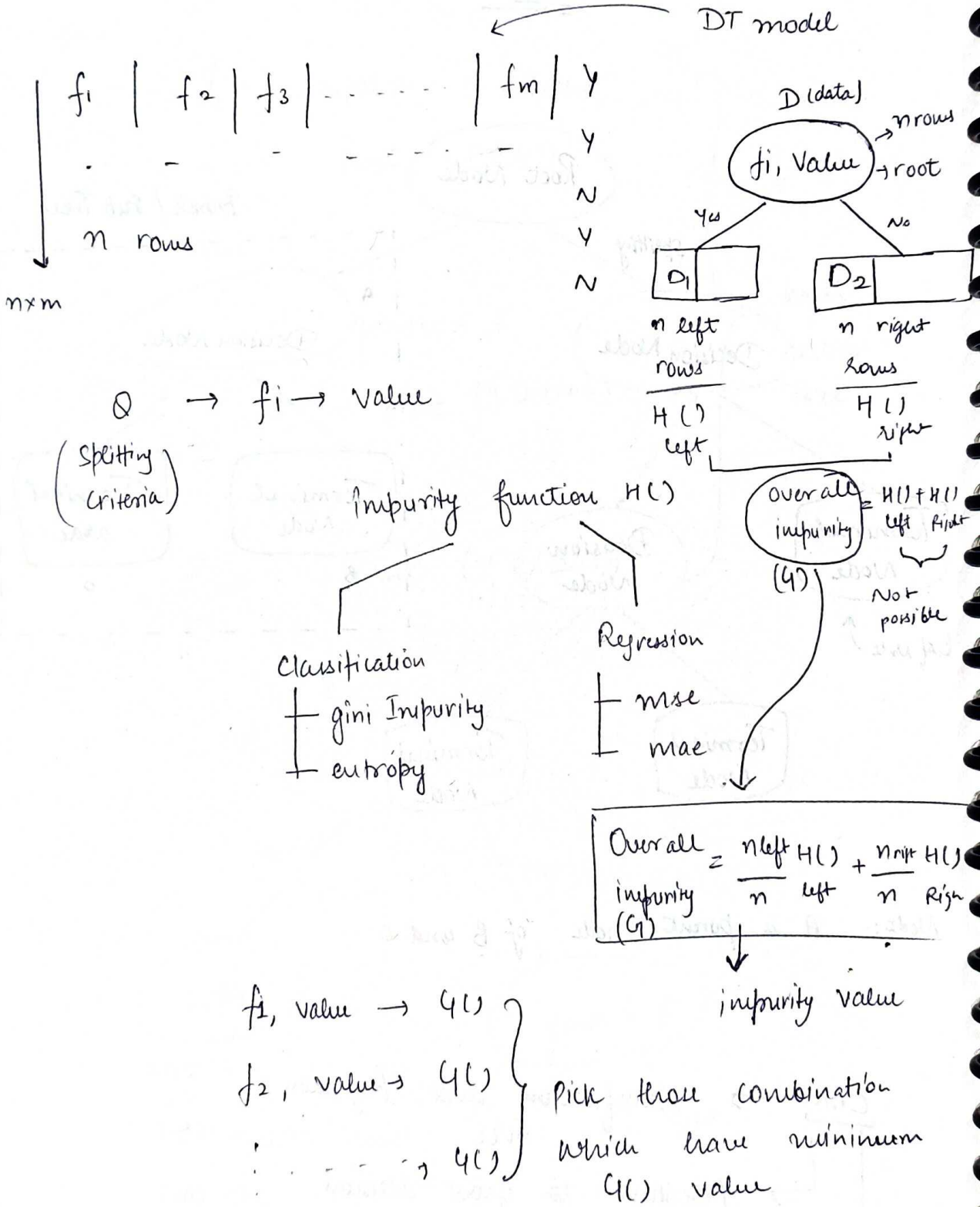
Node :- A is parent node of B and C

CART → Classification and Regression Tree
- ID3
- C4.5
- CART

→ Algorithm to grow decision Tree

→ Sklearn use internally

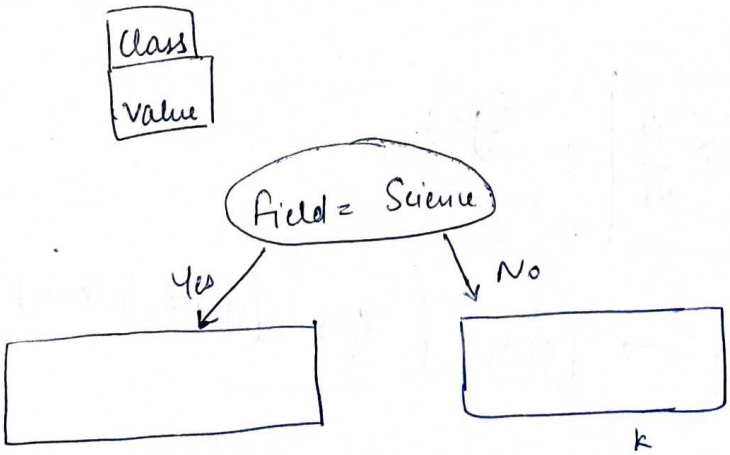
Classification



Splitting Categorical Features

	f_1 ↓ Degree_type	f_2 ↓ Field	f_3 ↓ Average_Grade	Job_Outcome
0	Undergraduate	Science	89	Employed
1	Undergraduate	Arts	92	Unemployed
2	Post graduate	Science	95	Employed
3	PhD	Science	85	Employed
4	Postgraduate	Arts	98	Unemployed
5	PhD	Arts	90	Employed
6	Undergraduate	Science	88	Unemployed
7	Post graduate	Arts	93	Employed
8	Undergraduate	Arts	94	Unemployed
9	PhD	Science	86	Employed

multi class ↑ binary class ↑ Numerical ↑

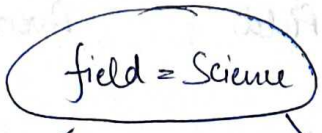


$$\text{gini impurity} = 1 - \sum_{i=1}^{k} P_k^2$$

P_k = Problem of each classes

$P_{Science} = \left(\frac{5}{10}\right)$ $P_{Arts} = \left(\frac{5}{10}\right)$

$$= 1 - \left(\frac{5}{10}\right)^2 - \left(\frac{5}{10}\right)^2$$



Yes

No

emp	unemp
4	1

emp	Unemp
3	2

$P_{emp} = \frac{4}{5}$ $P_{unemp} = \frac{1}{5}$

$P_{emp} = \frac{3}{5}$ $P_{unemp} = \frac{2}{5}$

↓ gini

$$1 - \left(\frac{4}{5}\right)^2 - \left(\frac{1}{5}\right)^2$$

$$1 - \left(\frac{16}{25}\right) - \left(\frac{1}{25}\right) = \frac{8}{25}$$

↓ gini

$$1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2$$

$$1 - \frac{9}{25} - \frac{4}{25} = \frac{12}{25}$$

$$\frac{5}{10} \times \frac{8}{25} + \frac{5}{10} \times \frac{12}{25} \quad g()$$

$0.16 + 0.24 \rightarrow \boxed{0.40}$ $q \rightarrow$ (field, Science)

multiclass Value → UG / PG / PhD

degree, UG	degree, PG	degree, PhD
GI	GI	GI
	min	

degree = UG

Yes → | emp | Unemp |
|---|---|
| 1 | 3 |

$1 - \left(\frac{1}{4}\right)^2 - \left(\frac{3}{4}\right)^2$

0.3

No → | emp | Unemp |
|---|---|
| 5 | 1 |

$1 - \left(\frac{5}{6}\right)^2 - \left(\frac{1}{6}\right)^2$

0.4

$\frac{4}{10} \times 0.3 + \frac{6}{10} \times 0.4 = 0.5$

degree = PG

Yes → | emp | Unemp |
|---|---|
| 2 | 1 |

$1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2 = 0.25$

No → | emp | Unemp |
|---|---|
| 4 | 3 |

$1 - \left(\frac{4}{7}\right)^2 - \left(\frac{3}{7}\right)^2 = 0.35$

$\frac{3}{10} \times 0.25 + \frac{7}{10} \times 0.35 = 0.65$

degree = UG < degree = PG

↓

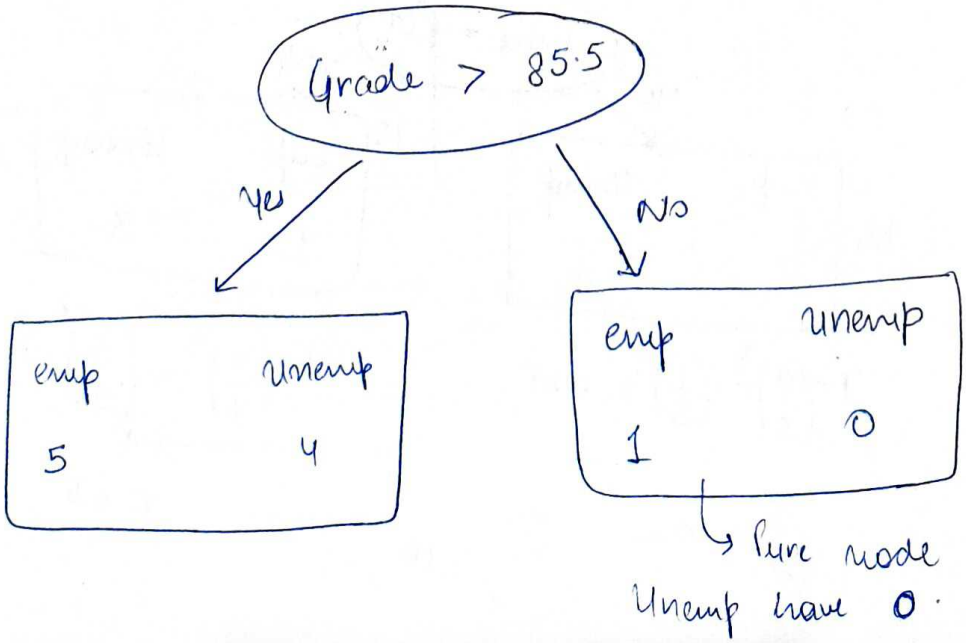
reject Criteria

because greater value

Numerical column

↓ sort

	Avg-Grade	job_outcome
	85	emp
85.5	86	emp
87	88	unemp
88.5	89	emp
89.5	90	emp
91	92	Unemp
92.5	93	emp
93.5	94	Unemp
94.5	95	emp
	98	Unemp



$$1-\left(\frac{5}{9}\right)^2-\left(\frac{4}{9}\right)^2 \qquad 1-\left(\frac{1}{1}\right)^2-\left(\frac{0}{1}\right)^2$$

$$1-\frac{25}{81}-\frac{16}{81} \qquad 1-1=0$$

$$1-\frac{41}{81}=\frac{40}{81}$$

$$G(1) = \frac{9}{10}\times\frac{40}{81}+\frac{1}{10}\times 0 = 0.25$$

Let assume,

0.4 → 85.5

0.3 → 87

0.2 → 88.5

0.1 → 89.5

0.8 → 91

0.7 → 93.5

94.5

grade > 89.5 → 0.1

this value compare with other value

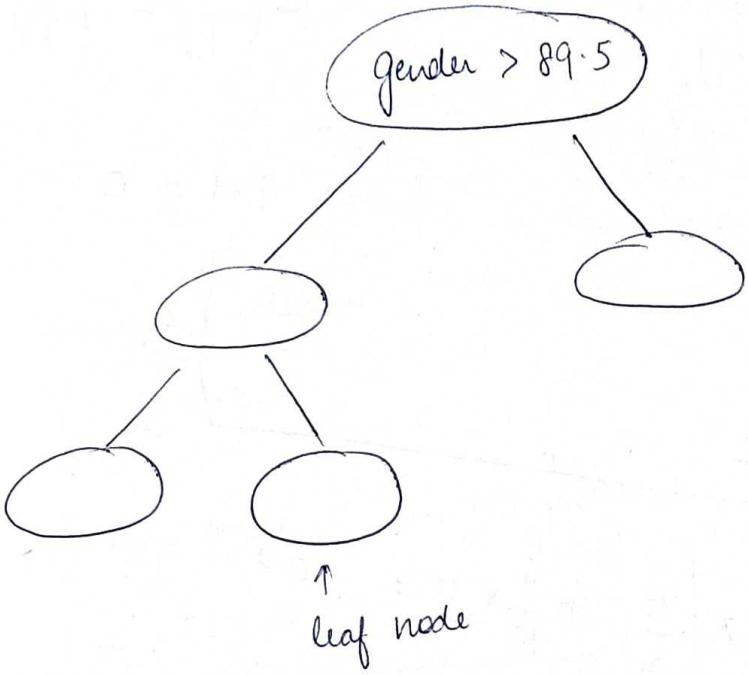
0.1 → Small

0.4 ↑ Field

↑ Average_Grade

Degree_Type

Job-occupation

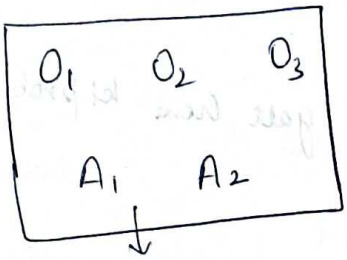


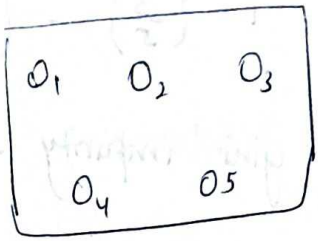
Understanding Gini Impurity?

The Gini impurity is a measure of how often a randomly chosen element from the set would be incorrectly labelled if it was randomly labelled according to the distribution of labels in the subset.

$$H() = 1 - \sum_{i=1}^{k} P_k^2$$

example:





3 Oranges

2 Apples

	O_1	O_2	O_3	A_1	A_2
O_1	✓				
O_2		✓			
O_3			✓		
A_1				✓	
A_2					✓

	O_1	O_2	O_3	A	A
O	✓	✓	✓	X	X
O	✓	✓	✓	X	X
O	✓	✓	✓	X	X
A	X	X	X	✓	✓
A	X	X	X	✓	✓

$$1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2$$

gini impurity → guess galat hone ki prob.

Geometric Intuition of Decision Tree

cgpa	iq	placement
-	-	Yes
-	-	No
-	-	Yes