02-11-2025

Agenda:

→ Code Implementation :—

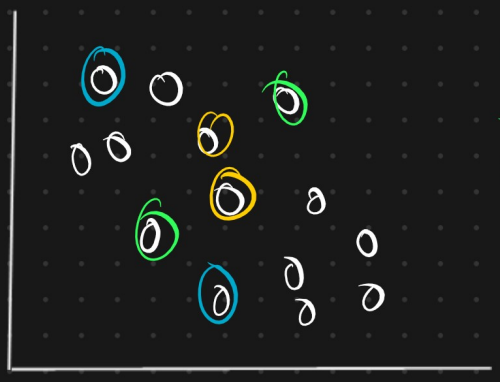
→ Naive Bayes

↳ Categorical NB

↳ Gaussian NB

→ K-means clustering

→ Decision Trees (ML Algo)



↓ init

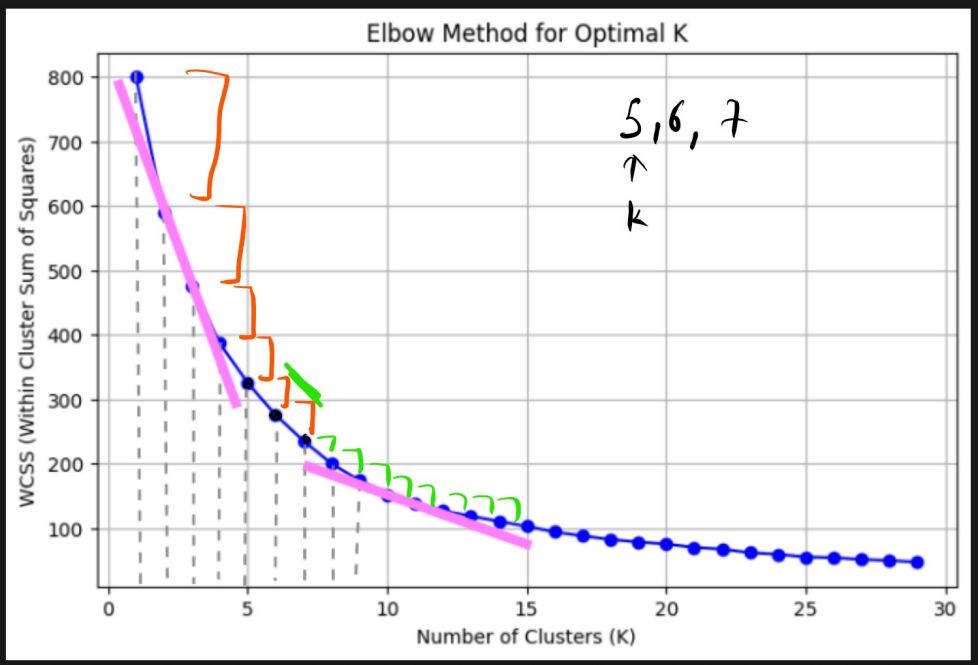
random → model → wcss

random → model → wcss → } 3 = n_init

random → model → wcss

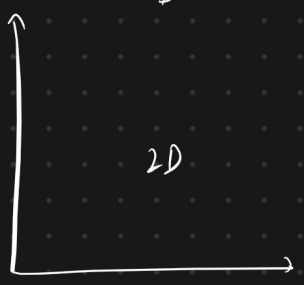
→ divide the dataset optimal

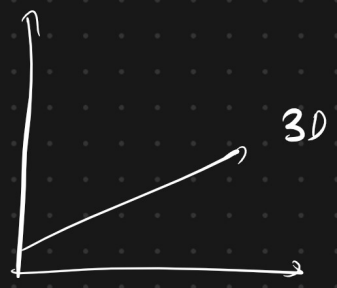
→ i want to divide the dataset into 2 categories



x_1 x_2 x_3 x_4 → Red

Blu



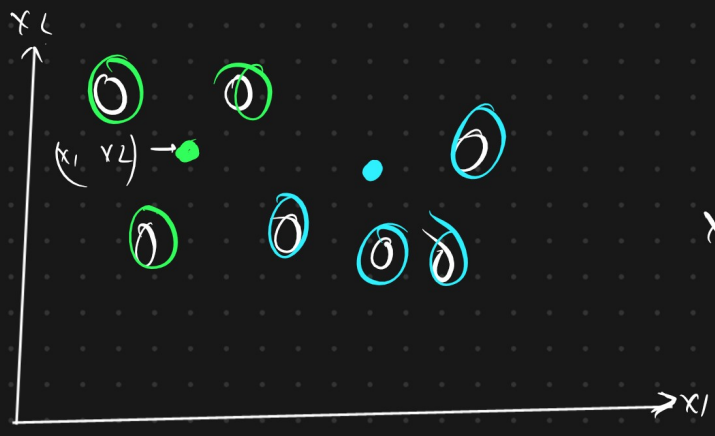


PCA → algo → High dimension data → low dimensional data

principal component Analysis

x_1 x_2 x_3 x_4 → x_1, x_2 (2D)

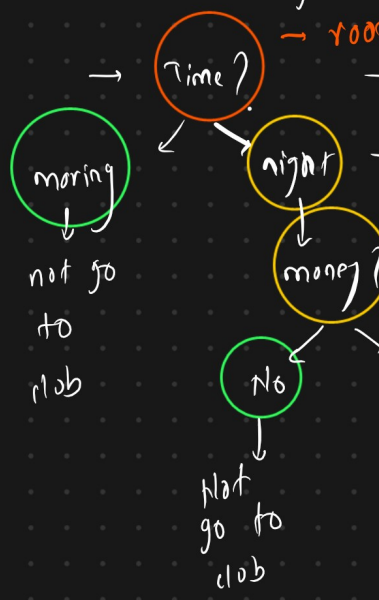
→ loss information



→ Decision Trees:

- → supervised
- → ease of interpretation
- → both regression & classification problems.

Task: Go to night club

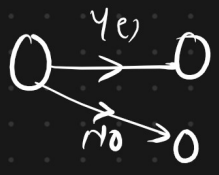


→ Series of sequential decision

→ not a node

→ decision nodes

→ leaf nodes → also called as terminal nodes, they represent predicted output for the model



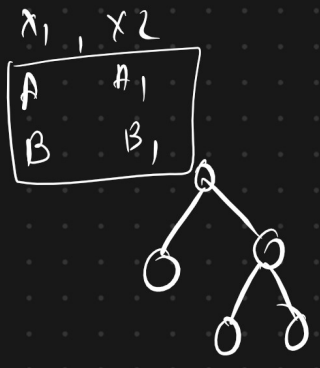
For decision tree, the resulting flow-like structure is navigated via conditional control statements, or if-then rules.

---

if-else statement → why it is a ML algo?

x_1	x_2	x_3	x_4	x_5	x_6

→ millian of rows



Decision Tree classifier:

ID	Weather	Temp	play
1	Sunny	HOT	NO
2	sunny	Mild	NO
3	cloudy	HOT	Yes
4	Rain	Mild	Yes
5	Rain	cool	Yes
6	cloudy	cool	Yes
7	Rain	HOT	NO
8	Sunny	cool	Yes

impurity = how good a split is

→ Entropy: measure of uncertainty: $-\sum p_i \log_2(p_i)$

→ Gini impurity: impurity: $1 - \sum p_i^2$

information gain:

: How much uncertainty (entropy) is reduce after splitting on a feature.

$$\text{Information Gain} = \text{Entropy}_{parent} - \sum \frac{n_{child}}{n_{total}} \times \text{Entropy}_{child}$$

Gender: [M, F, F, F, M, F, F, M, F] → impurity → high / low

(1) (2) 3 4 5 6

① ② ⑤

M - 3
F - 6

→ Entropy ← Impurity

Gini - impurity

Gender → G1 [M, F, F, F, F] → Entropy
→ G2 → [M, M, F] → Entropy

High Information gain

low IG

→ Better impurity G1 compared to previous split
→ Not the best IG, still higher side.

~~ID~~ Weather Temp Play

Step 1: Calculate parent Entropy (x_1)

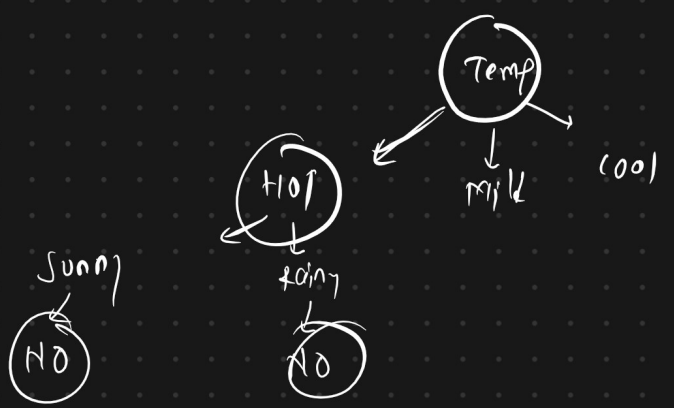
Step 2: Try each feature

Weather
↓
Entropy (x_2)
↓
IG (parent - child)
↓
($x_1 - x_2$)
→ A1 (1.23)

Temp
↓
Entropy (x_3)

IG (parent - child)
→ ($x_1 - x_3$)
→ A2 (5.6)

→ whichever is having the higher IG becomes the root node
→ A2 → Temperature



ID	Weather	Temp	Play
1	Sunny	HOT	NO
2	sunny	MILD	NO
3	Cloudy	HOT	Yes
4	Rain	MILD	Yes
5	Rain	cool	Yes
6	Cloudy	cool	Yes
7	Rain	HOT	NO
8	Sunny	cool	Yes

Step 1: Calculate parent entropy:

	prob
Yes → 5	5/8 = 0.625
NO → 3	3/8 = 0.375

Entropy: $-(P_i \log_2 P_i) \rightarrow -(0.625 \log_2 0.625 + 0.375 \log_2 0.375)$

$\rightarrow$ 0.954

Step 2: Try each feature

Feature 1: weather

weather	Count	Yes	NO	Entropy
Sunny	3	1	(2)	$-(\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log_2\frac{2}{3}) = 0.918$
Cloudy	2	2	0	0
Rain	3	2	1	0.918

weighted average entropy after split:

$$\frac{3}{8} \times 0.918 + \frac{2}{8} \times 0 + \frac{3}{8} \times 0.918 \rightarrow 0.688$$

Information Gain (weather) = 0.954 − 0.688

= 0.266

Feature 2: Temp:

Temp	Count	Yes	No	Entropy
HOT	3	1	2	0.918
Mild	2	1	1	1
Cool	3	3	0	0

weighted Entropy:

$$\frac{3}{8} \times 0.918 + \frac{2}{8} \times 1 + \frac{3}{8} \times 0 = 0.594$$

Information Gain (Temp) = 0.954 − 0.594

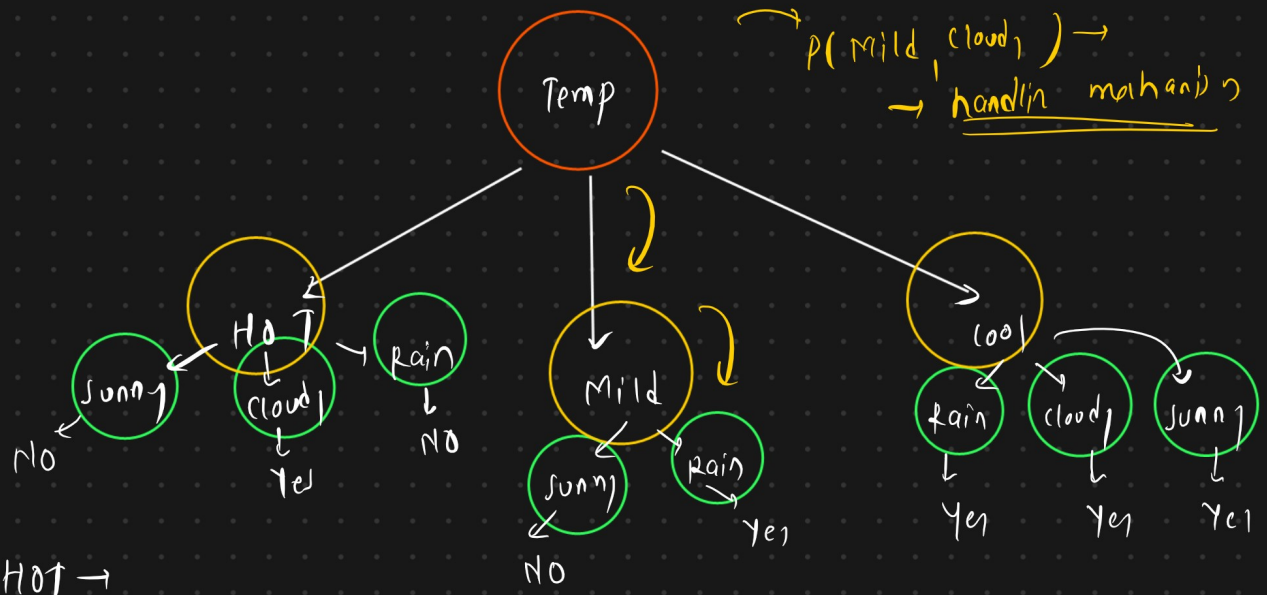
= 0.360

IG (weather) < IG (Temp)

0.266 < 0.360

↓

Temp becomes our root node.



P(Mild, Cloudy) → → handling mahanbo

Hot →

weather	play
Sunny	NO
Cloudy	Yes
Rain	NO

Yes → 1
NO → 2

Hot	Count	Yes	NO	Entropy
	3	1	2	0.918 (same as before)

↓
we must split again to make it pure

Mild →

	play
Sunny	NO
Rain	Yes

Entropy : 1.0 (mixed) (same as before) → split