DECISION TREES (CART)

Cleanliness of information. A dataset is "clean" or statistically clean if the Gini-Index of the dataset is zero, and is "dirty" if the Gini-Index is one.

Dataset: S {Apple, Orange, Strawberry, Apple, Lemon}

$$Gini = 1 - \sum_{i=1}^{n} p_i^2 \qquad p_i \equiv \text{probability of element } i$$

$$Gini = 1 - \left(\frac{2}{5}\right)^2 - \left(\frac{1}{5}\right)^2 - \left(\frac{1}{5}\right)^2 - \left(\frac{1}{5}\right)^2 = 0'72$$

prob. Apple Orange Strw.

We use this concept to make decisions: based on previous decisions, we can make a decision algorythm: DECISION TREE

Case of Mary & John. Decision tree "When to have sex".

	Cleanliness	Emotional Talking	Fitness	Moon	Sex
1.	dirty	often	regular	full	Yes
2.	dirty	seldom	intense	decreasing	No
3.	clean	seldom	none	new	Yes
4.	mid-clean	often	regular	increasing	Yes
5.	dirty	seldom	intense	full	No
6.	dirty	often	none	decreasing	Yes
7.	clean	seldom	regular	new	Yes
8.	dirty	often	none	increasing	No
9.	clean	seldom	regular	full	Yes
10.	mid-clean	often	intense	decreasing	No

We go through each category and find the smallest Gini-Index. This category has the highest cleanliness and helps us make a better decision.

Cleanliness.	Yes	No	#
dirty	2	3	5
mid-clean	1	1	2
clean	3	0	3

$$\text{Gini}(\text{Cleanliness dirty}) = 1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2 = 0'48$$

$$\text{Gini}(\text{Cleanliness mid-clean}) = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = 0'5$$

$$Gini(\text{Cleanliness clean}) = 1-\left(\frac{3}{3}\right)^2-\left(\frac{0}{3}\right)^2 = 0$$

$$Gini(\text{Cleanliness}) = \frac{5}{10}\cdot 0'48 + \frac{2}{10}\cdot 0'5 + \frac{3}{10}\cdot 0 = 0'34$$

Emotional talking (ET)	Yes	No	#
often	3	2	5
seldom	3	2	5

$$Gini(\text{ET often}) = 1-\left(\frac{3}{5}\right)^2-\left(\frac{2}{5}\right)^2 = 0'48$$

$$Gini(\text{ET seldom}) = 1-\left(\frac{3}{5}\right)^2-\left(\frac{2}{5}\right)^2 = 0'48$$

$$Gini(\text{ET}) = \frac{5}{10}\cdot 0'48 + \frac{5}{10}\cdot 0'48 = 0'48$$

Fitness	Yes	No	#
regular	4	0	4
intense	0	3	3
none	2	1	3

$$Gini(\text{F regular}) = 1-\left(\frac{4}{4}\right)^2 = 0$$

$$Gini(\text{F intense}) = 1-\left(\frac{3}{3}\right)^2 = 0$$

$$Gini(\text{F none}) = 1-\left(\frac{2}{3}\right)^2-\left(\frac{1}{3}\right)^2 = 0'44$$

$$Gini = \frac{4}{10}\cdot 0 + \frac{3}{10}\cdot 0 + \frac{3}{10}\cdot 0'44 = 0'133$$

Moon	Yes	No	#
full	2	1	3
decreasing	1	2	3
new	2	0	2
increasing	1	1	2

$$Gini(M\ full) = 1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2 = 0'44$$

$$Gini(M\ decr.) = 1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2 = 0'44$$

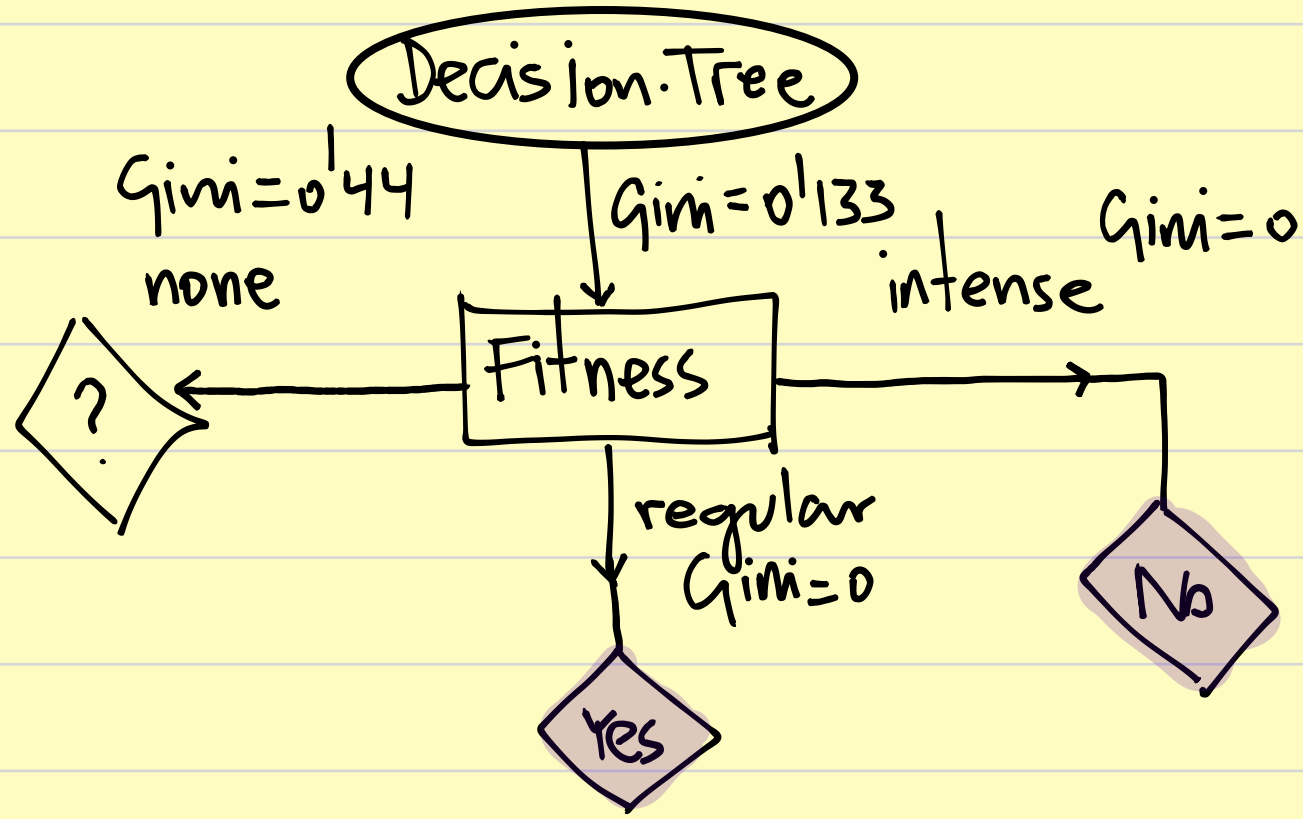
$$Gini(M\ new) = 0$$

$$Gini(M\ increasing) = 0'5$$

$$Gini(Moon) = \frac{3}{10} \cdot 0'44 + \frac{3}{10} \cdot 0'44 + 0 + \frac{2}{10} \cdot 0'5 = 0'364$$

Cleanliness	Gini = 0'34
ET	Gini = 0'48
F	Gini = 0'133
M	Gini = 0'364

The first decision criteria is fitness. Smallest Gini.



⊙ Fitness None + { C, ET, M }

Fitness None + Cleanliness	Yes	No	#
dirty	1	1	2
clean	1	0	1

$$Gini(\text{Fitness None + C dirty}) = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = 0'5$$

$$Gini(\text{Fitness None + C clean}) = 1 - \left(\frac{1}{1}\right)^2 = 0$$

$$Gini(\text{Fitness None + C}) = \frac{2}{3} \cdot 0'5 + \frac{1}{3} \cdot 0 = 0'33$$

Fitness None + ET	Yes	No	#
often	1	1	2
seldom	1	0	1

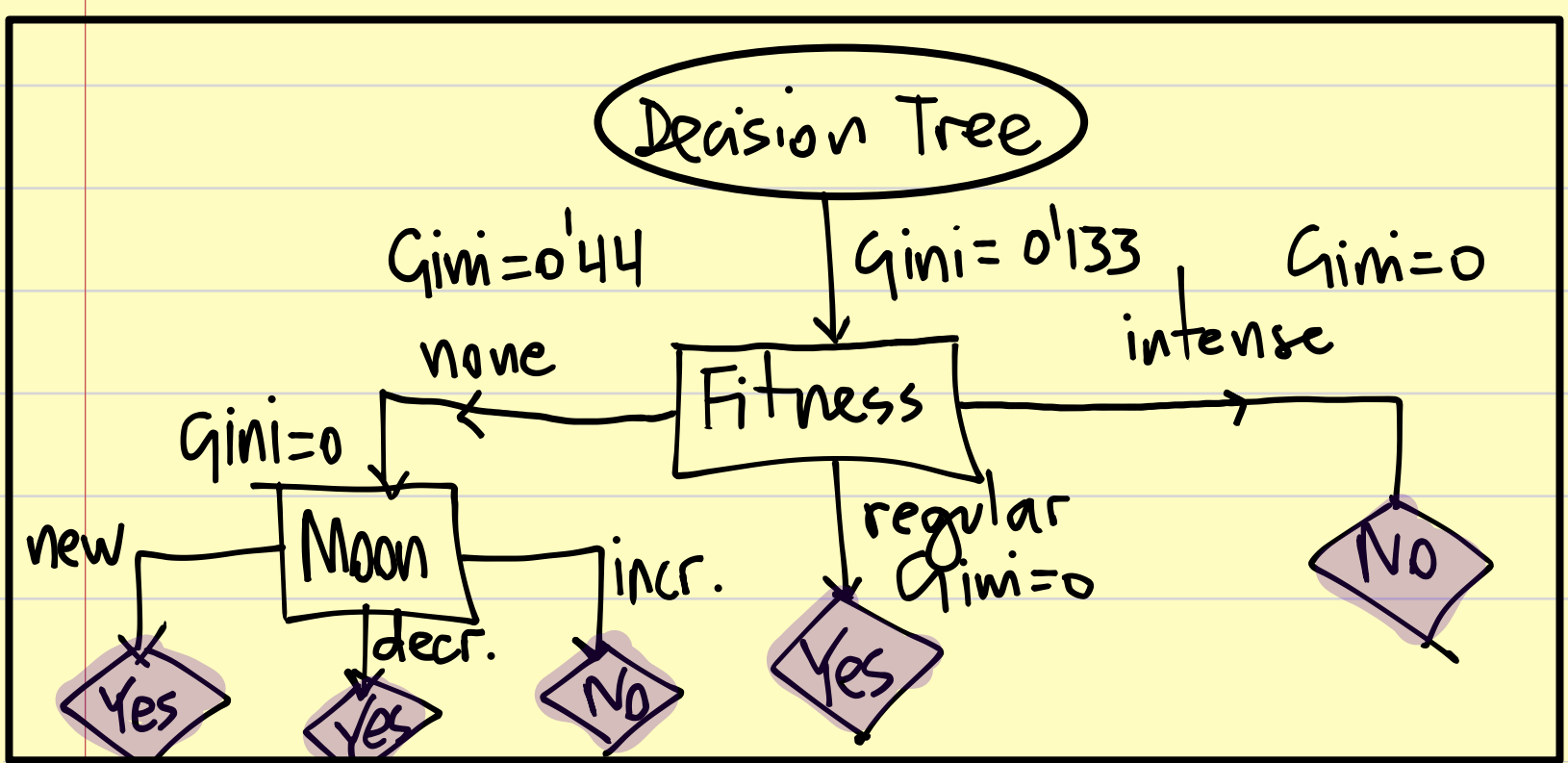
$$Gini(\text{Fitness None + ET often}) = 0'5$$

$$Gini(\text{Fitness None + ET seldom}) = 0$$

$$Gini(\text{Fitness None + ET}) = 0'33$$

Fitness None + Moon	Yes	No	#
new	1	0	1
decreasing	1	0	1
increasing	0	1	1

$$Gini(\text{Moon}) = 0$$



Day	outlook	temperature	humidity	wind	Decision
1	sunny	hot	high	weak	No
2	sunny	hot	high	strong	No
3	overcast	hot	high	weak	Yes
4	rainfall	mild	high	weak	Yes
5	rainfall	cool	normal	weak	Yes
6	rainfall	cool	normal	strong	No
7	overcast	cool	normal	wtrong	Yes
8	sunny	mild	high	weak	No
9	sunny	cool	normal	weak	Yes
10	rainfall	mild	normal	weak	Yes
11	sunny	mild	normal	strong	Yes
12	overcast	mild	high	strong	Yes
13	overcast	hot	normal	weak	Yes
14	rainfall	mild	high	strong	No

Excercise. Case is a decision to play a soccer game or not.