Cart (Classification and Regression Trees)

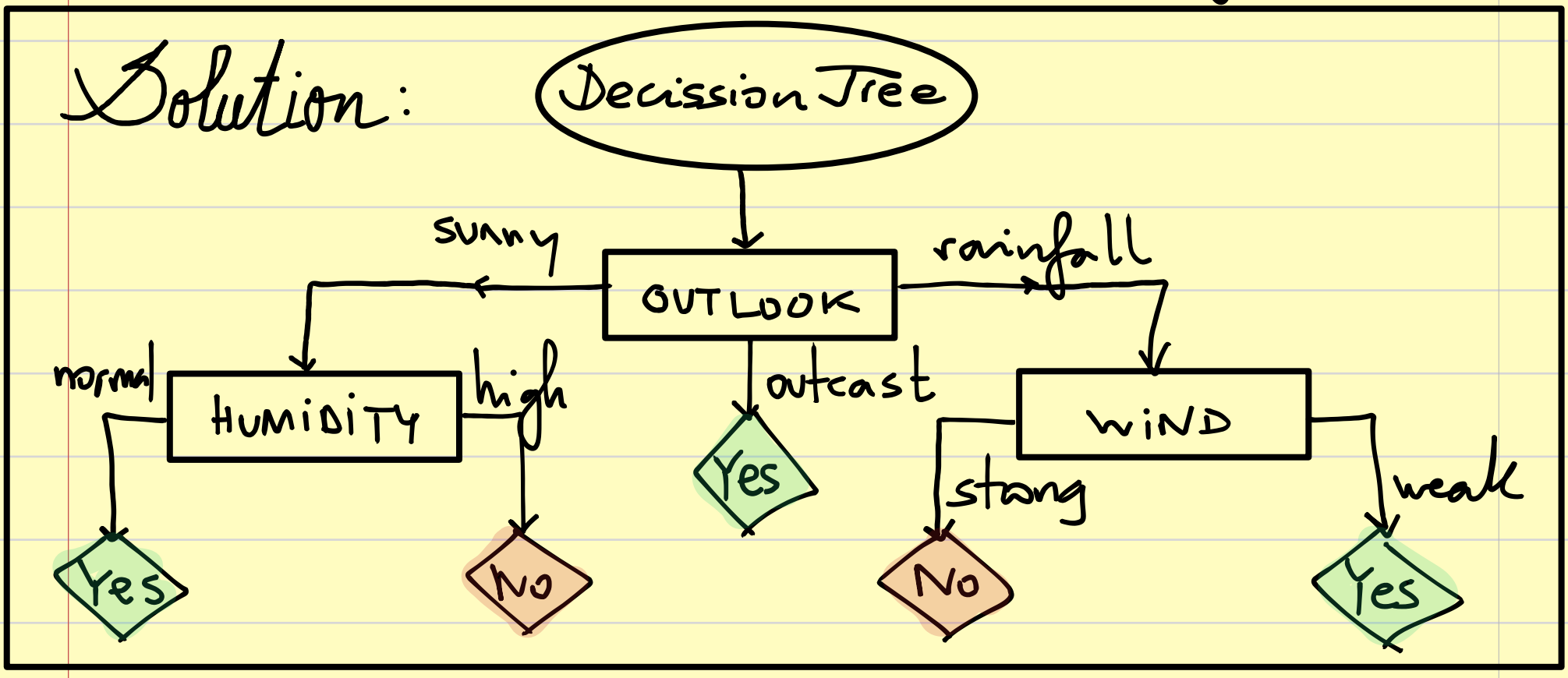
Day	outlook	temperature	humidity	wind	Decision
1	sunny	hot	high	weak	No
2	sunny	hot	high	strong	No
3	overcast	hot	high	weak	Yes
4	rainfall	mild	high	weak	Yes
5	rainfall	cool	normal	weak	Yes
6	rainfall	cool	normal	strong	No
7	overcast	cool	normal	wtrong	Yes
8	sunny	mild	high	weak	No
9	sunny	cool	normal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

Dirtyness of Information

The dirtyness meassures the homogeneity of the dataset.
We meassure it with the Gini-Index: $GINI = 1 - \sum_{i=1}^{n} p_i^2$

- $GINI = 0$ means that the data is pure.
- $GINI = 1$ means that the data is dirty.

Solution:



Goal: create a decission tree.

1 2 3 4

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

Step 1. Look at all criteria and pick the one with smallest Gini.

1. Outlook : sunny, overcast, rainfall.

OUTLOOK	Yes	No	#
sunny	2	3	5
overcast	4	0	4
rainfall	3	2	5

$P_{Yes}(\text{Outlook sunny})^2$

$$Gini(\text{Outlook sunny}) = 1 - \sum p_i^2 = 1 - \left[\left(\frac{2}{5}\right)^2 + \left(\frac{3}{5}\right)^2\right] = 0'48$$

$P_{No}(\text{Outlook sunny})^2$

$$Gini(\text{Outlook overcast}) = 1 - \left[\left(\frac{4}{4}\right)^2 - \left(\frac{0}{4}\right)^2\right] = 0$$

$$Gini(\text{Outlook rainfall}) = 1 - \left[\left(\frac{3}{5}\right)^2 + \left(\frac{2}{5}\right)^2\right] = 0'48$$

$$Gini(\text{Outlook}) = \frac{5}{14}\cdot 0'48 + \frac{4}{14}\cdot 0 + \frac{5}{14}\cdot 0'48 = 0'342$$

2. TEMPERATURE : ~~hot~~, mild, cool.

TEMP.	Yes	No	#
hot	2	2	4
cool	3	1	4
mild	4	2	6

$$Gini(\text{Temp hot}) = 1 - \left[\left(\frac{2}{4}\right)^2 + \left(\frac{2}{4}\right)^2\right] = 0'5$$

$$Gini(Temp\ cool) = 1-\left[\left(\frac{3}{4}\right)^2+\left(\frac{1}{4}\right)^2\right] = 0'375$$

$$Gini(Temp\ mild) = 1-\left[\left(\frac{4}{6}\right)^2+\left(\frac{2}{6}\right)^2\right] = 0'445$$

$$Gini(Temp) = \frac{4}{14}\cdot 0'5+\frac{4}{14}\cdot 0'375+\frac{6}{14}\cdot 0'445 = 0'439$$

3. HUMIDITY : high. normal

HUM.	Yes	No	#
high	3	4	7
normal	6	1	7

$$Gini(Hum.\ high) = 1-\left[\left(\frac{3}{7}\right)^2+\left(\frac{4}{7}\right)^2\right] = 0'489$$

$$Gini(Hum.\ normal) = 1-\left[\left(\frac{6}{7}\right)^2+\left(\frac{1}{7}\right)^2\right] = 0'244$$

$$Gini(Hum.) = \frac{7}{14}\cdot 0'489+\frac{7}{14}\cdot 0'244 = 0'367$$

4. WIND. weak, strong

WIND	YES	NO	#
weak	3	3	6
strong	6	2	8

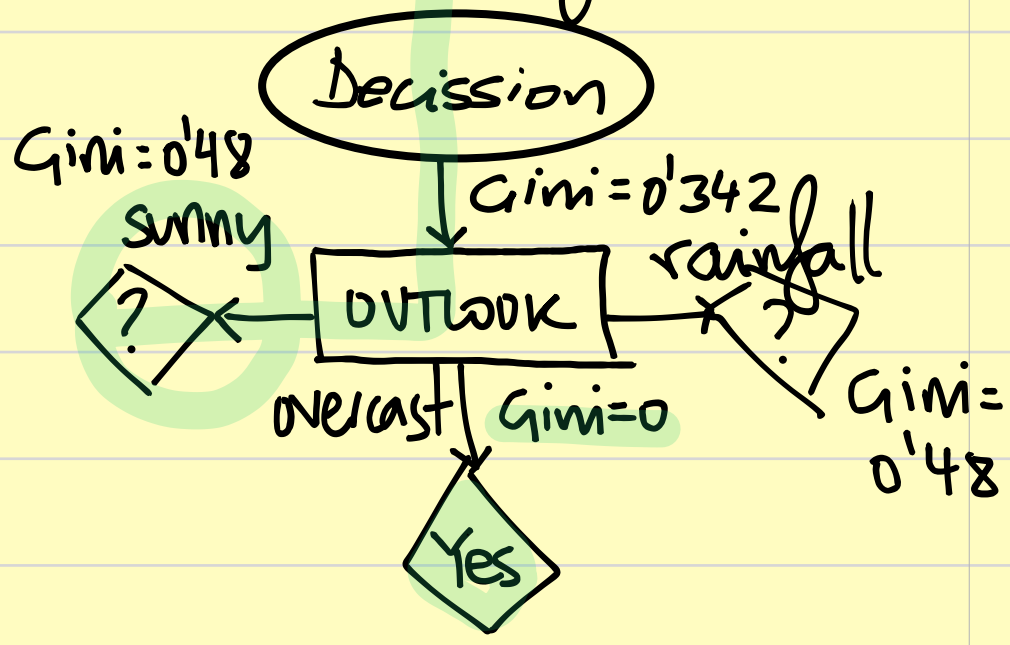
$$Gini(Wind\ Weak) = 0'5$$

$$Gini(Wind\ strong) = 1-\left[\left(\frac{6}{8}\right)^2+\left(\frac{2}{8}\right)^2\right] = 0'375$$

$$Gini(Wind) = \frac{6}{14}\cdot 0'5+\frac{8}{14}\cdot 0'375 = 0'428$$

The variable with the lowest GINI is our first node.

	Gini
OUTLOOK	0'342 →
TEMP	0'439
HUM	0'367
WIND	0'428



OUTLOOK (Sunny) : TEMP / HUM / WIND

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

1. OUTLOOK(Sunny) + Temp .

out Sun + Temp	Yes	No	#
hot	0	2	2
mild	1	1	2
cool	1	0	1
			5

Gini(Outlook sunny + hot) = 0
Gini(Outlook sunny + mild) = 0'5
Gini(Outlook sunny + cool) = 0

} Gini(Outlook sunny + Temp) = $\frac{2}{5} \cdot 0'5 = 0'2$

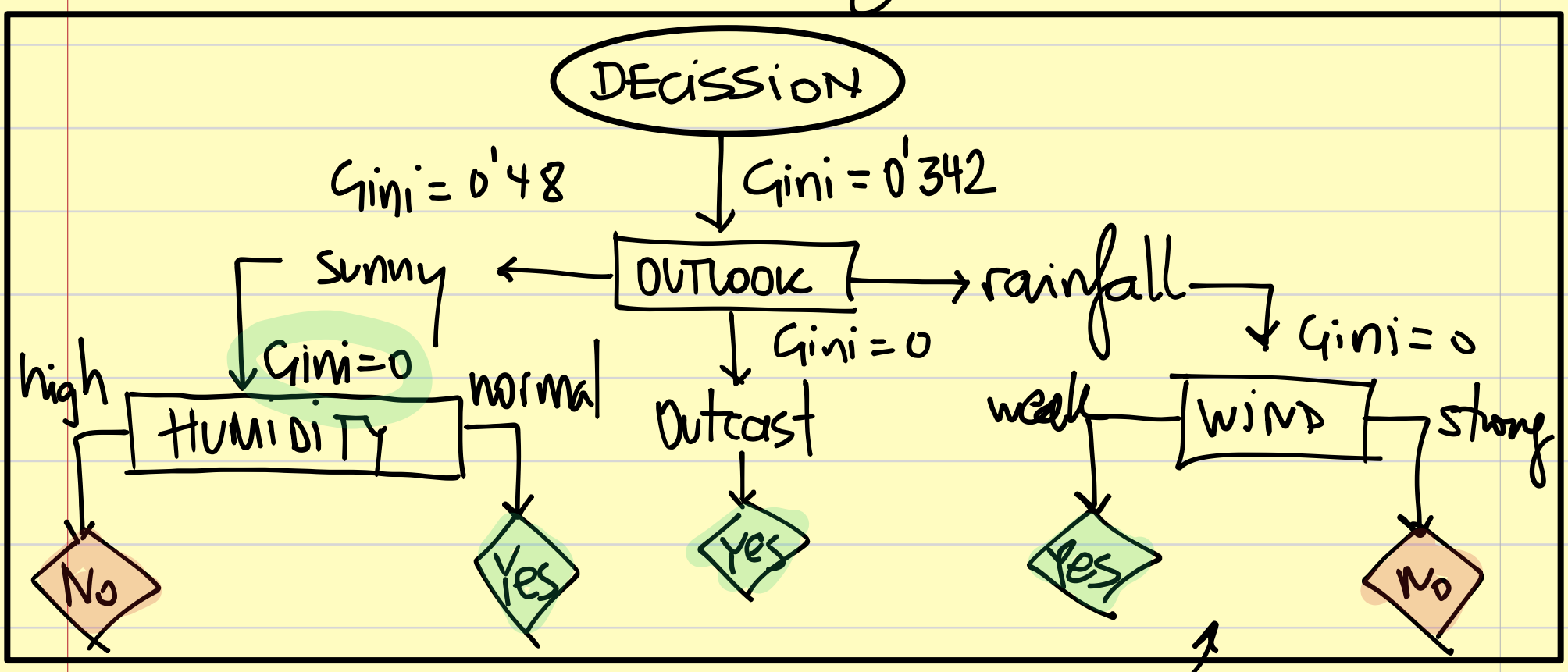
2. Outlook sunny + HUM

Outlook sunny + HUM	Yes	No	#
high	0	3	3
normal	2	0	2
			5

Gini(Outlook sunny + HUM) = 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

We have found a criteria with Gini = 0, so we do not have to search anymore.



Outlook Rainfall + Temp/wind

Outlook Rainfall + Wind	Yes	No	#
weak	3	0	3
strong	0	2	2
			5

Gini(Outlook Rain+Wind) = 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 #2.

SEX: · Yes · No

	Kitchen clean	Groceries	Stress @ Work	Time Together	D
1.	dirty	big	strong	long	Y
2.	very dirty	small	mild	short	N
3.	dirty	small	mild	short	Y
4.	clean	big	weak	long	Y
5.	very clean	small	strong	long	N
6.	dirty	small	weak	short	N
7.	dirty	big	strong	long	Y
8.	clean	small	weak	short	N

Kitchen.

$$Gini(Kitchen) = \frac{4}{8}\left[1-\left[\left(\frac{3}{4}\right)^2+\left(\frac{1}{4}\right)^2\right]\right] + \frac{1}{8}\cdot\left[1-\left(\frac{1}{1}\right)\right] + \frac{2}{8}\cdot\left[1-\left[\left(\frac{1}{2}\right)^2+\left(\frac{1}{2}\right)^2\right] + \frac{1}{8}\left[1-\left(\frac{1}{1}\right)^2\right] = 0'1875+0+0'125+0 = 0'3125$$

(dirty, very dirty, clean, very clean)

Groceries.

$$Gini(Groceries) = \frac{3}{8}\left[1-\left(\frac{3}{3}\right)^2\right] + \frac{5}{8}\left[1-\left[\left(\frac{4}{5}\right)^2+\left(\frac{1}{5}\right)^2\right] = 0'2$$

(big, small)

Stress @ Work

$$Gini(Stress) = \frac{3}{8}\left[1-\left[\left(\frac{2}{3}\right)^2+\left(\frac{1}{3}\right)^2\right]\right] + \frac{2}{8}\cdot 0'5 + \frac{3}{8}\left[1-\left[\left[\frac{2}{3}\right]^2+\left[\frac{1}{3}\right]^2\right]\right]$$

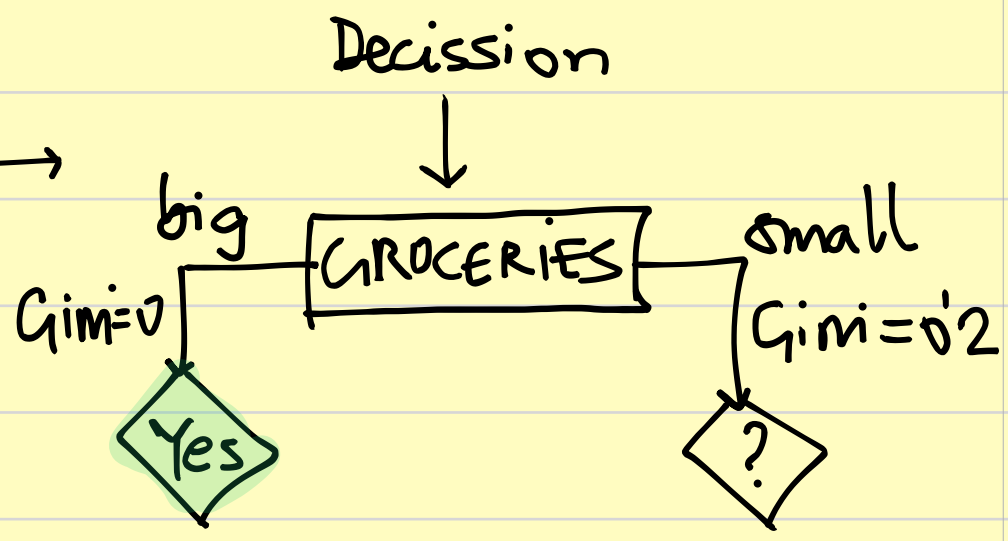
(strong, weak, mild)

$= 0'33$

$$\text{Time Together} \quad Gini(Time) = \frac{4}{8}\underbrace{\left[1-\left[\left(\frac{3}{4}\right)^2+\left(\frac{1}{4}\right)^2\right]\right]}_{\text{long}} + \frac{4}{8}\underbrace{\left[1-\left[\left(\frac{1}{4}\right)^2+\left(\frac{3}{4}\right)^2\right]\right]}_{\text{short}} =$$

$: 0'375$

	Gini
Kitchen	0'3125
Groceries	0'2
Stress@Work	0'33
Time Together	0'375



Groceries small + Stress@Work.

	Yes	No	#
weak	0	2	2
mild	1	1	2
strong	0	1	1/5

$Gini(\text{Groceries small} + \text{Stress@Work}) =$

$= \frac{2}{5}\cdot 0 + \frac{2}{5}\cdot 0'5 + \frac{1}{5}\cdot 0 = 0'2$

Groceries small + Kitchen

	Yes	No	#	Gini
dirty	1	1	2	0'5
very dirty	0	1	1	0
clean	0	1	1	0
very clean	0	1	1/5	0

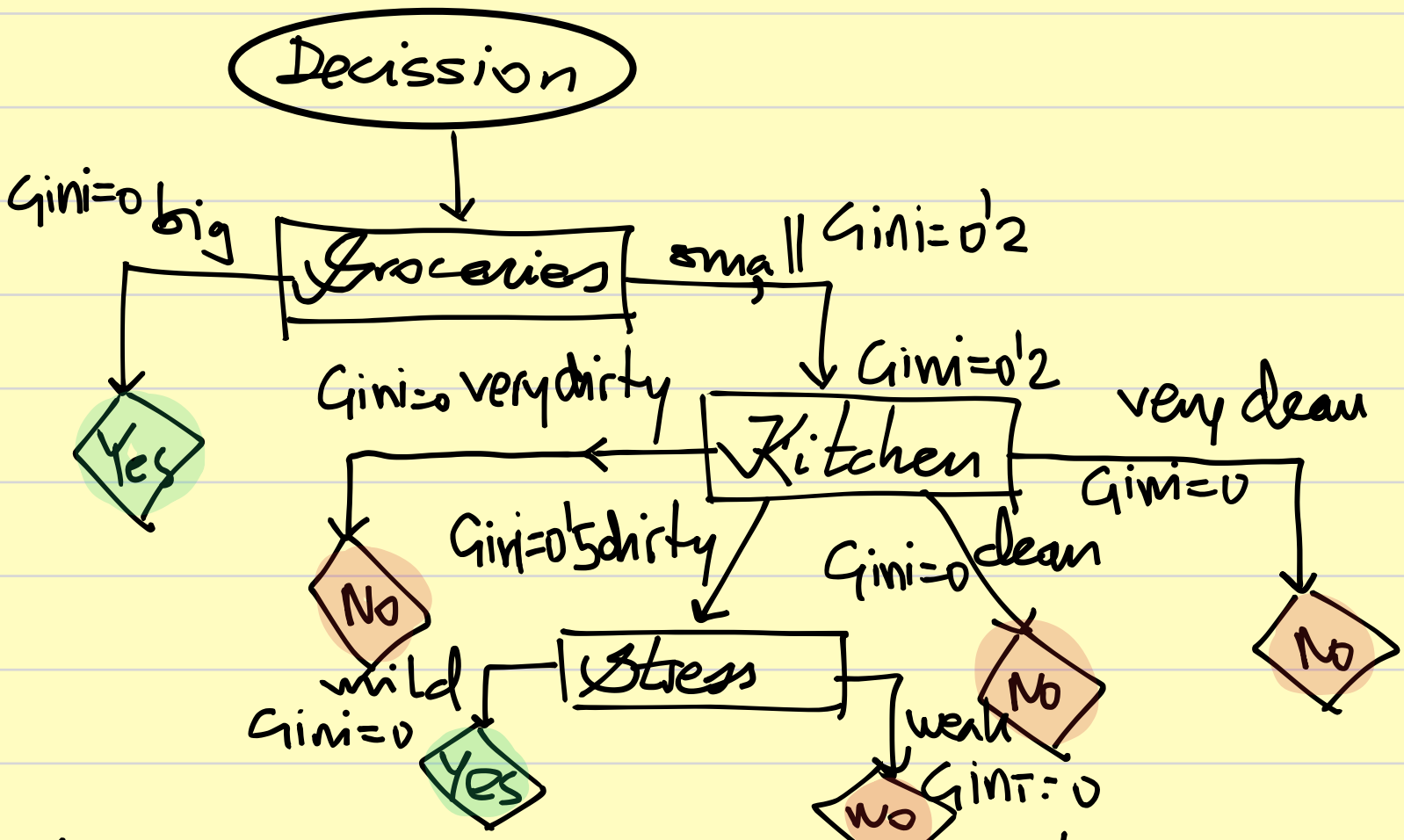
$$Gini(Groceries\ small + Kitchen) =$$

$$= \frac{2}{5} \cdot 0'5 + 0 + 0.0 = 0'2$$

Groceries small + Time Together	Yes	No	#
short	1	3	4
long	0	1	1/5

$$Gini(Groceries\ small + Time\ Together) =$$

$$= \frac{4}{5}\left[1 - \left(\left(\frac{1}{4}\right)^2 + \left(\frac{3}{4}\right)^2\right)\right] + 0 = 0'3$$



$$Gini(Groceries\ small + Kitchen\ dirty + Stress\ @\ Work) =$$

$$= \frac{1}{2} \cdot 0 + \frac{1}{2} \cdot 0 = 0$$