CART (Classification and Regression Trees)

· **Concept. Dirtyness of Information.**

The dirtyness meassures the homogeneity of the data set. If the data set is homogeneous, the data belong to the same class and dirtyness is zero.

We meassure dirtyness with the **GINI-Index**:

$$GINI = 1 - \sum_{i=1}^{n} p_i^2$$

The GINI-Index is a KPI that meassures the dirtyness of data and is always between ZERO and ONE: $GINI \in [0,1]$

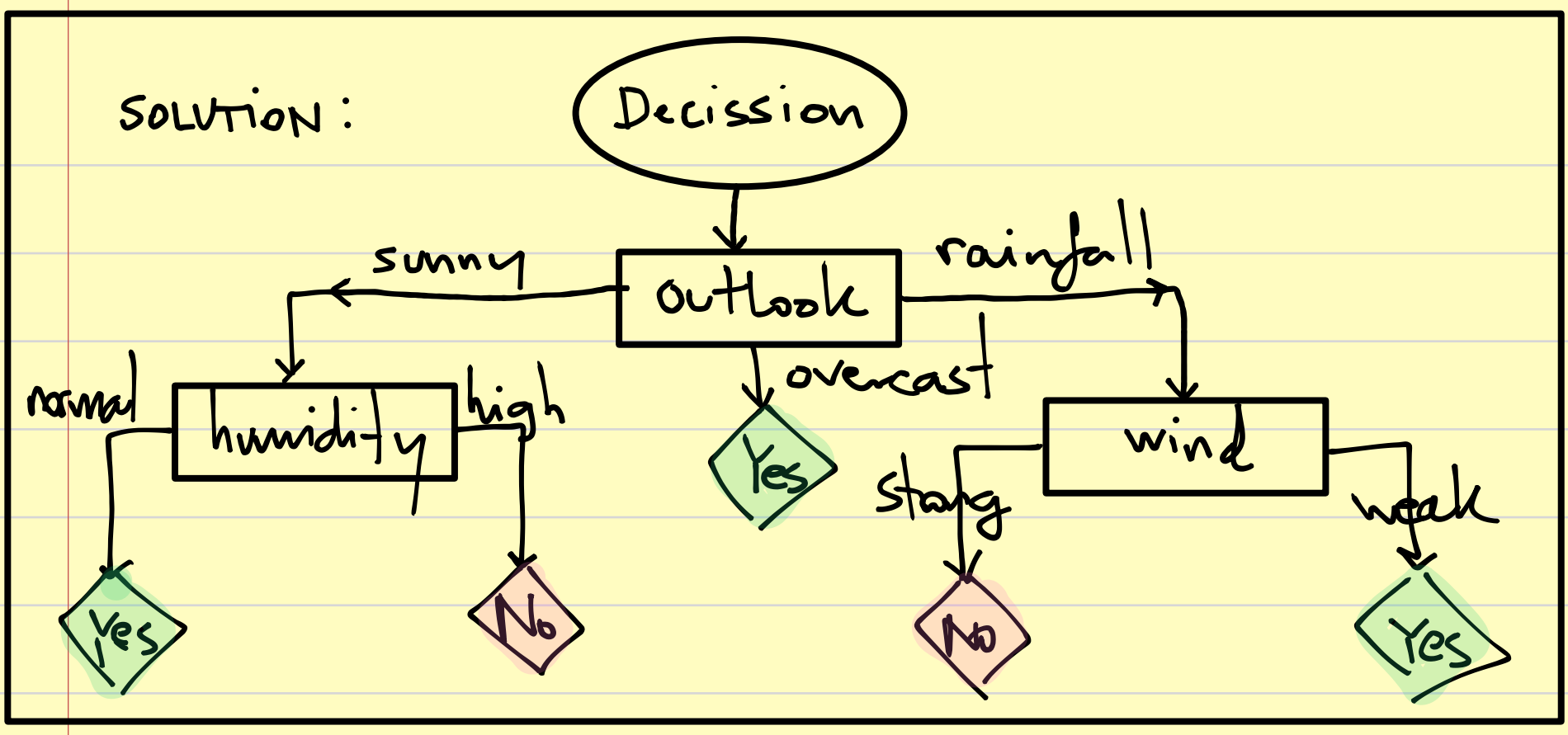
- GINI-Index = 0 means that the data is purely homogeneous and all elements belong to the same class.
- GINI-Index = 1 means that the data is maximally dirty and all elements belong to different classes.

Example. We have taken data of previous decissions made by soccer associations base on the weather about the realization of a soccer match.

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

Goal: create a decission tree.

SOLUTION :



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

OUTLOOK. Outlook is a nominal variable. It can take 3 values: sunny, overcast, rainfall.

OUTLOOK	Yes	No	#
sunny	2	3	5
overcast	4	0	4
rainfall	3	2	5
			$\Sigma = 14$

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

$$Gini(outlook\ Overcast) = 1 - \sum p_i^2 = 1 - \left[\left(\frac{4}{4}\right)^2 + \left(\frac{0}{4}\right)^2\right] = 0$$

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

The Gini Index for Outlook is the weighted sum:

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

The dirtyness of OUTLOOK is 0'342.

TEMPERATURE.

Temperature	Yes	No	#
hot	2	2	4
cool	3	1	4
mild	4	2	6
			$\Sigma=14$

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

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

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

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

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

HUMIDITY

HUMIDITY	Yes	No	#
high	3	4	7
normal	6	1	7
			$\Sigma 14$

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

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

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

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

WIND

WIND	Yes	No	#
strong	6	2	8
weak	3	3	6

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

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

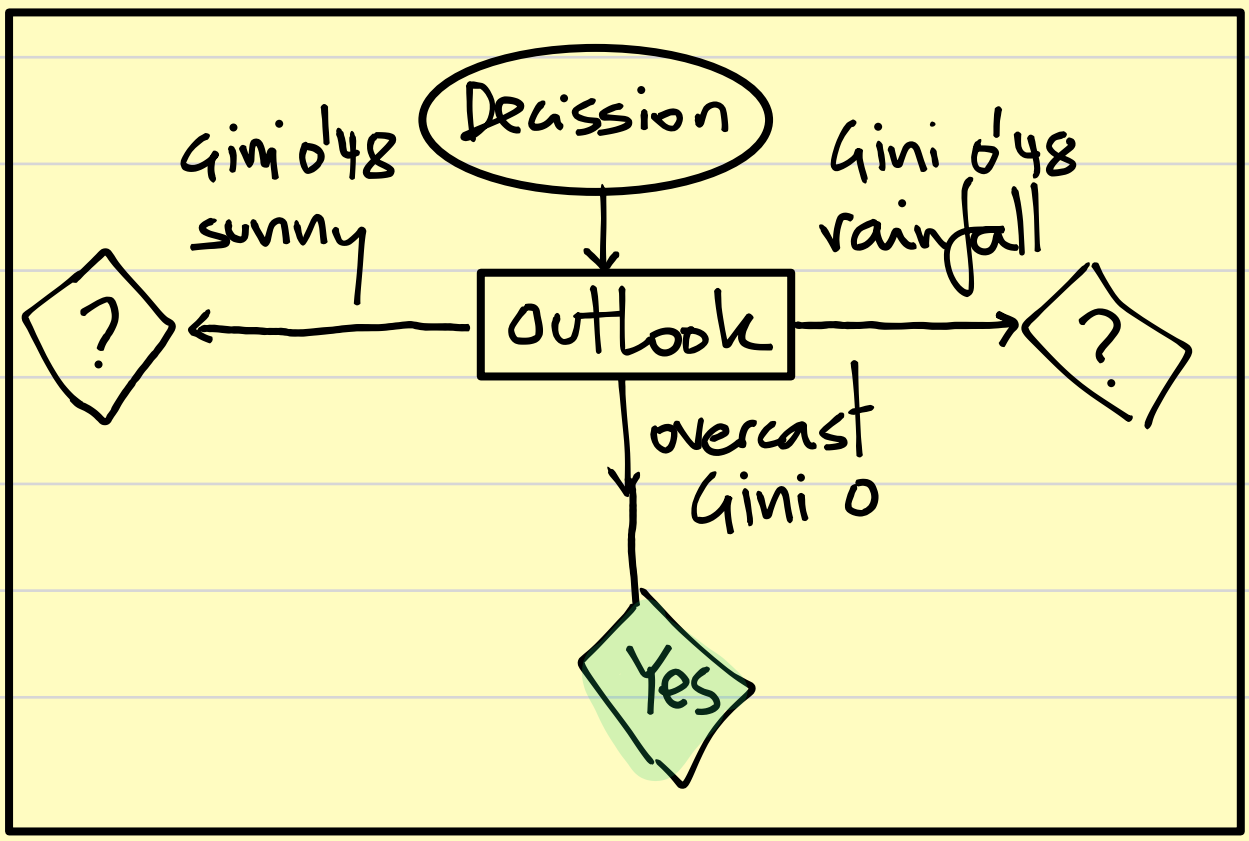
$$Gini(wind\ mild) = 1 - \left[\left(\frac{3}{6}\right)^2 + \left(\frac{3}{6}\right)^2\right] = 0'5$$

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

- Based on the values of the GINI for all variables, we decide which is our first node in our decission.

The variable with the lowest GINI is our first node.

Feature	Gini	
outlook	0'342	→ The outlook is the cleanest variable and delivers therefore a sharper decission.
Temp.	0'439	
Humidity	0'367	
Wind	0'428	



OUTLOOK SUNNY . TEMPERATURE
. HUMIDITY
. 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

OUTLOOK SUNNY	TEMPERATURE	Yes	No	#
sunny	hot	0	2	2
sunny	cold	1	0	1
sunny	mild	1	1	2
				Σ5

$$Gini(Sunny+hot) = 1-\left[\left(\frac{0}{2}\right)^2+\left(\frac{2}{2}\right)^2\right]=0$$

$$Gini(Sunny+cold) = 1-\left[\left(\frac{1}{1}\right)^2+\left(\frac{0}{1}\right)^2\right]=0$$

$$Gini(Sunny+mild) = 0'5$$

$$Gini(sunny+Temperature) = \frac{2}{5}\cdot 0+\frac{1}{5}\cdot 0+\frac{2}{5}\cdot 0'5 = 0'2$$

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

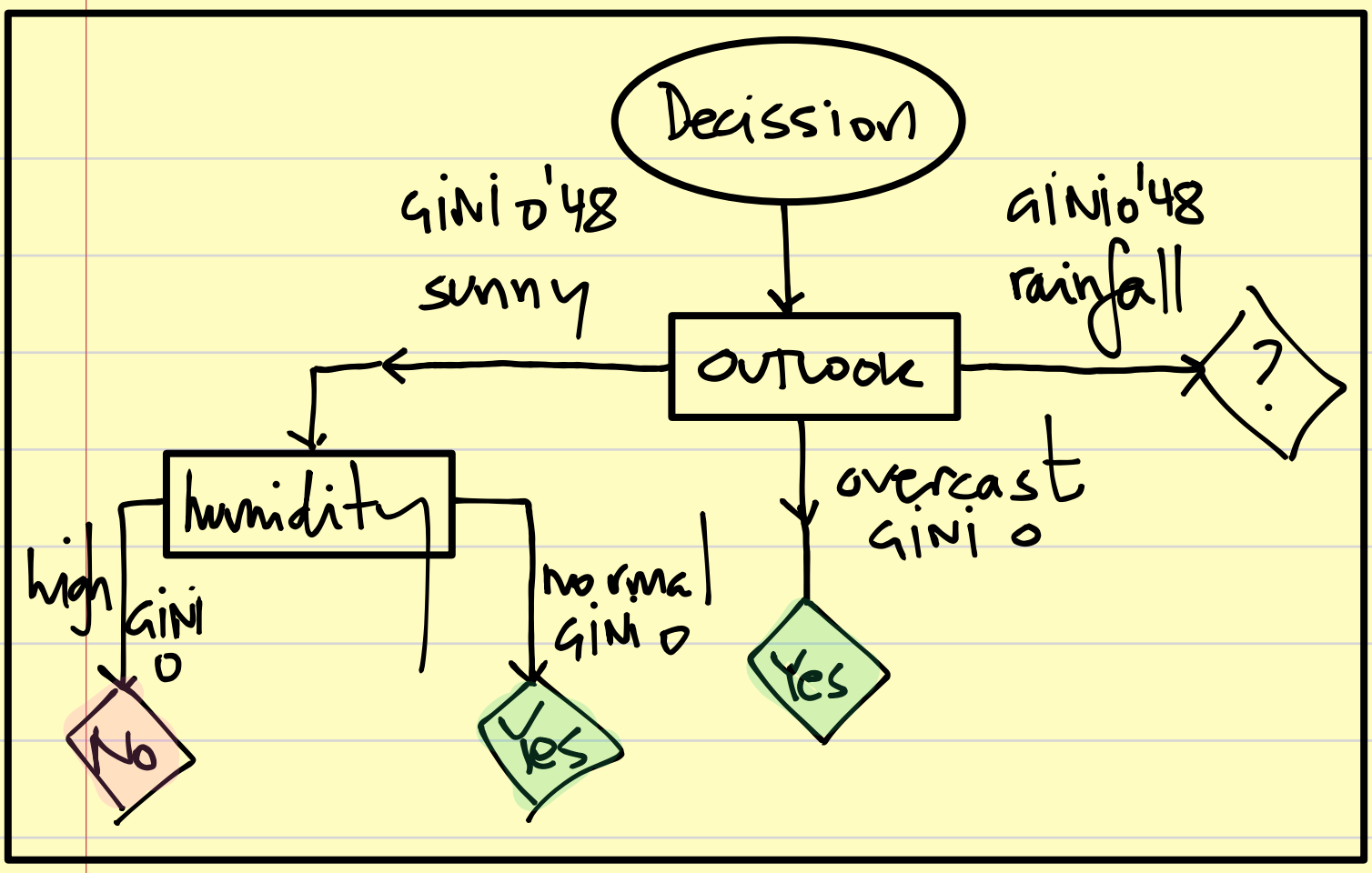
OUTLOOK SUNNY	HUMIDITY	Yes	No	#
sunny	high	0	3	3
sunny	normal	2	0	2

$$Gini(sunny+high) = 0$$

$$Gini(sunny+normal) = 0$$

$$Gini(Outlook\ Sunny+humidity) = 0$$

I have found a criteria (HUMIDITY) which combined with OUTLOOK sunny cuts perfectly the data GINI = 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

OUTLOOK RAINFALL . TEMPERATURE

. WIND

OUTLOOK	RAINFALL TEMP.	Yes	No	#
Rainfall	cool	1	1	2
Rainfall	mild	2	1	3
				5

$$Gini(Rainfall + Temp\,cool) = 1 - \left[\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2\right] = 0'5$$

$$Gini(Rainfall + Temp\,mild) = 1 - \left[\left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2\right] = 0'44$$

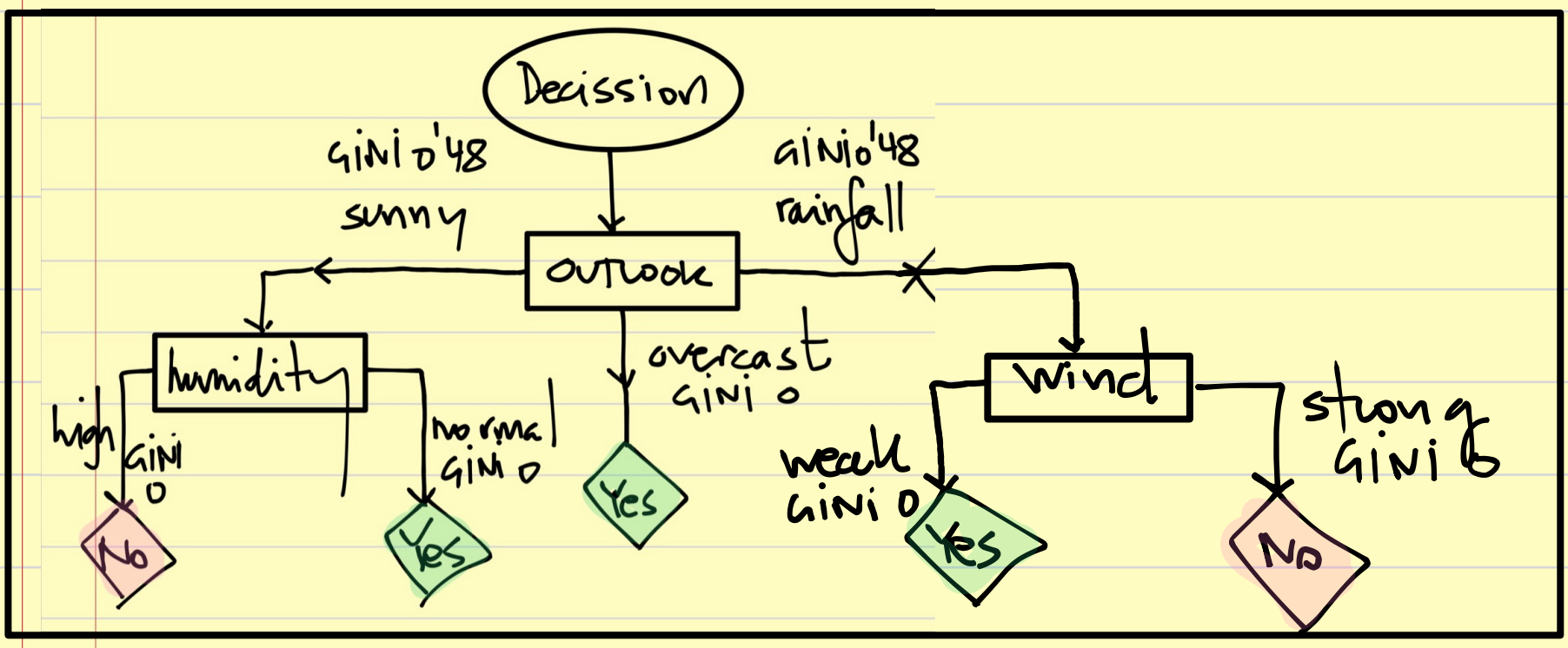
$$Gini(Rainfall + Temp) = \frac{2}{5} \cdot 0'5 + \frac{3}{5} \cdot 0'44 = 0'467$$

OUTLOOK	RAINFALL	WIND	Yes	No	#
	Rainfall	weak	3	0	3
	Rainfall	strong	0	2	2

Gini (Rainfall + Wind weak) =

= Gini (Rainfall + Wind strong) =

= Gini (Rainfall) = 0



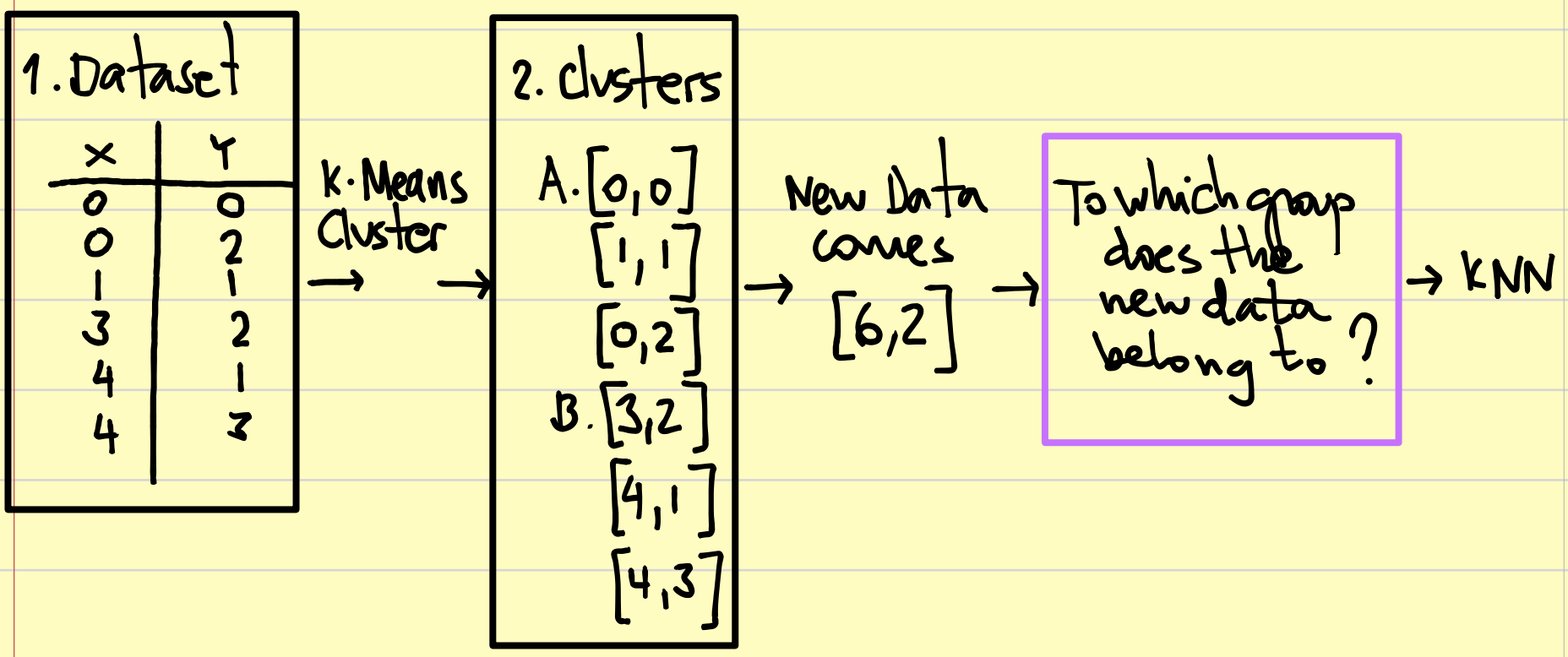
CART for SOCCER game Decission

Excercise . SEX . Yes or 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

K·NEAREST NEIGHBOUR (KNN)

Given an already clustered dataset (for instance with k.Means clustering), we are confronted with a new datapoint and are asked to which cluster the new point belongs to.



KNN is a supervised learning algorythm used for regression and classification. KNN tries to predict the correct class for the new data, then select the k·number of points which are closest to the test data. The KNN algorithm calculates the probability of the test data belonging to the classes and assigns the class to the highest probability.

Example from data above:

A. [0,0] [0,2] [1,1]

B. [3,2] [4,1] [4,3]

New data point ≡ [6,2] ≡ ∝

Step 1. calculate the distance from the new point to all the rest of the data.

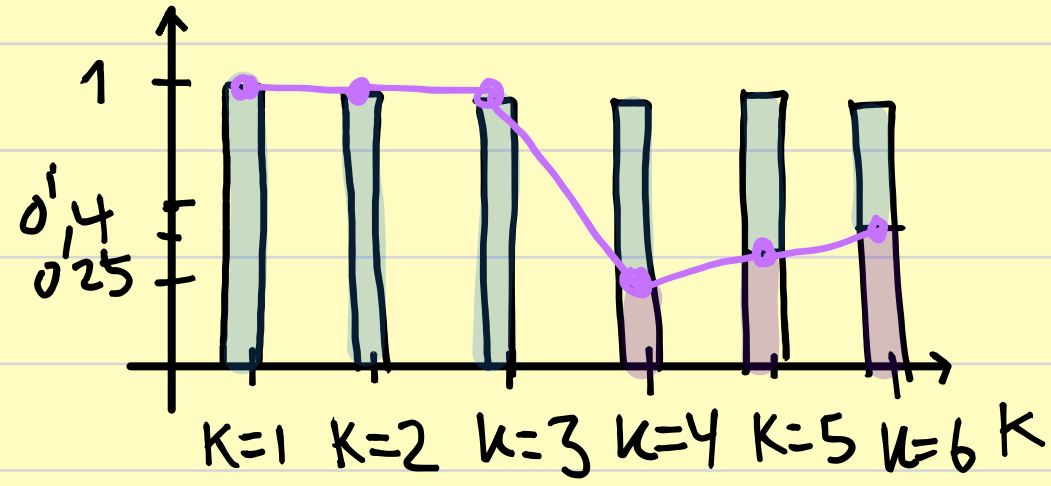
$d_{\alpha,A1} = \sqrt{(6-0)^2+(2-0)^2} = 6'32$ $\quad d_{\alpha,B1} = \sqrt{(6-3)^2+(2-2)^2} = 3$

$d_{\alpha,A2} = \sqrt{(6-0)^2+(2-2)^2} = 6$ $\quad d_{\alpha,B2} = \sqrt{(6-4)^2+(2-1)^2} = 2'23$

$d_{\alpha,A3} = \sqrt{(6-1)^2+(2-1)^2} = 5'1$ $\quad d_{\alpha,B3} = \sqrt{(6-4)^2+(2-3)^2} = 2'23$

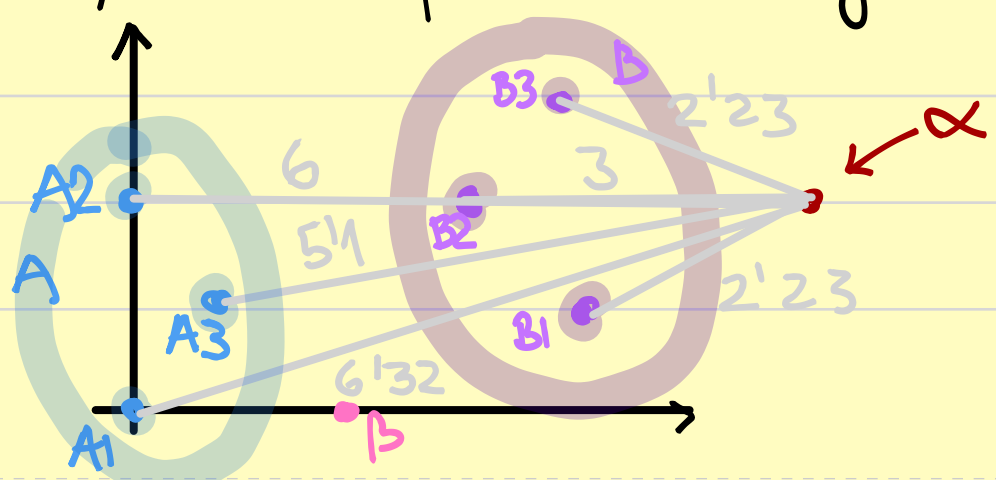
Step 2. Take different values of K (points in the group) and calculate the probability it belongs to one or other group.

B3, B2, B1, A3, A2, A1 . Hierarchy of distances



· For K=4 (the point of minimum error) the probability of belonging to group B is 75%.
· Hence, the new point belongs to group B.

GRAPHICALLY



Now, with α in group 3, a new point arises $B=[2,0]$

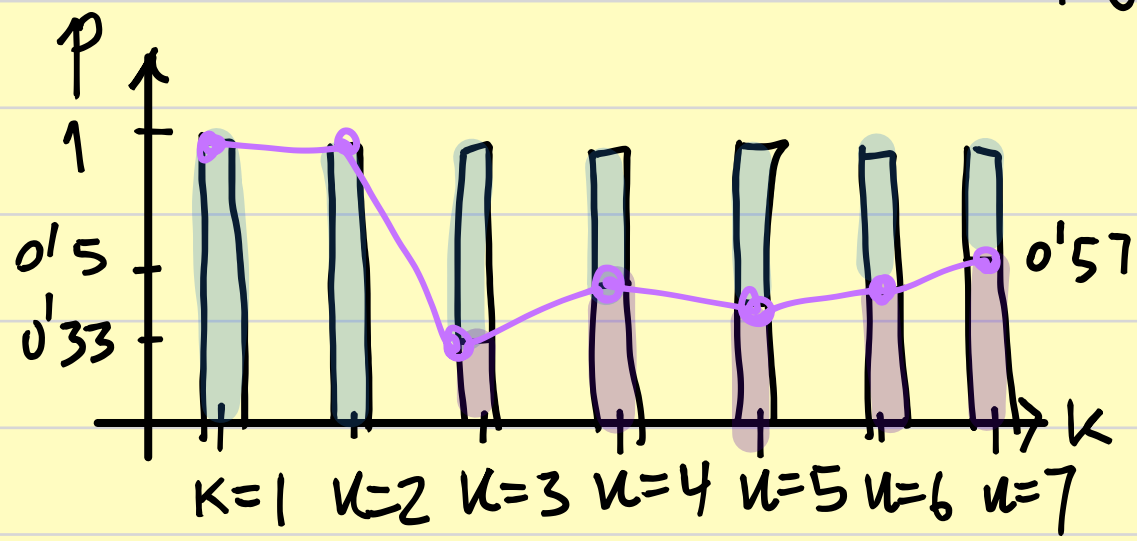
$$d_{B,A_1}=\sqrt{(2-0)^2+(0-0)^2}=2 \qquad d_{B,B1}=\sqrt{(2-3)^2+(2-0)^2}=2'23$$

$$d_{B,A_2}=\sqrt{(2-0)^2+(0-2)^2}=2'83 \qquad d_{B,B2}=\sqrt{(2-4)^2+(0-1)^2}=2'23$$

$$d_{B,A_3}=\sqrt{(2-1)^2+(0-1)^2}=1'414 \qquad d_{B,B3}=\sqrt{(2-4)^2+(0-3)^2}=3'6$$

$$d_{B,\alpha}=\sqrt{(2-6)^2+(0-2)^2}=4'47$$

A3, A1, B1, B2, A2, B3, α . Hierarchy of distances.



- K=3 delivers best probability
- B belongs to group A.