Fundamentals of Artificial Intelligence

COMP307/AIML420
Tutorial 2: kNN and DTs

Dr. Heitor Murilo Gomes
heitor.gomes@vuw.ac.nz
http://www.heitorgomes.com

Some questions

1. Will supplying the .py file suffice or do I need to supply a .exe file generated from the .py file?

2. Should we use k-fold CV or other evaluation?

3. Extra points for COMP307 if you solve AIML420 questions?

4. Should normalise both train and test datasets?

4. Can I use libraries for reading, storing and accessing data (e.g. accessing particular indexes, splitting, …)?

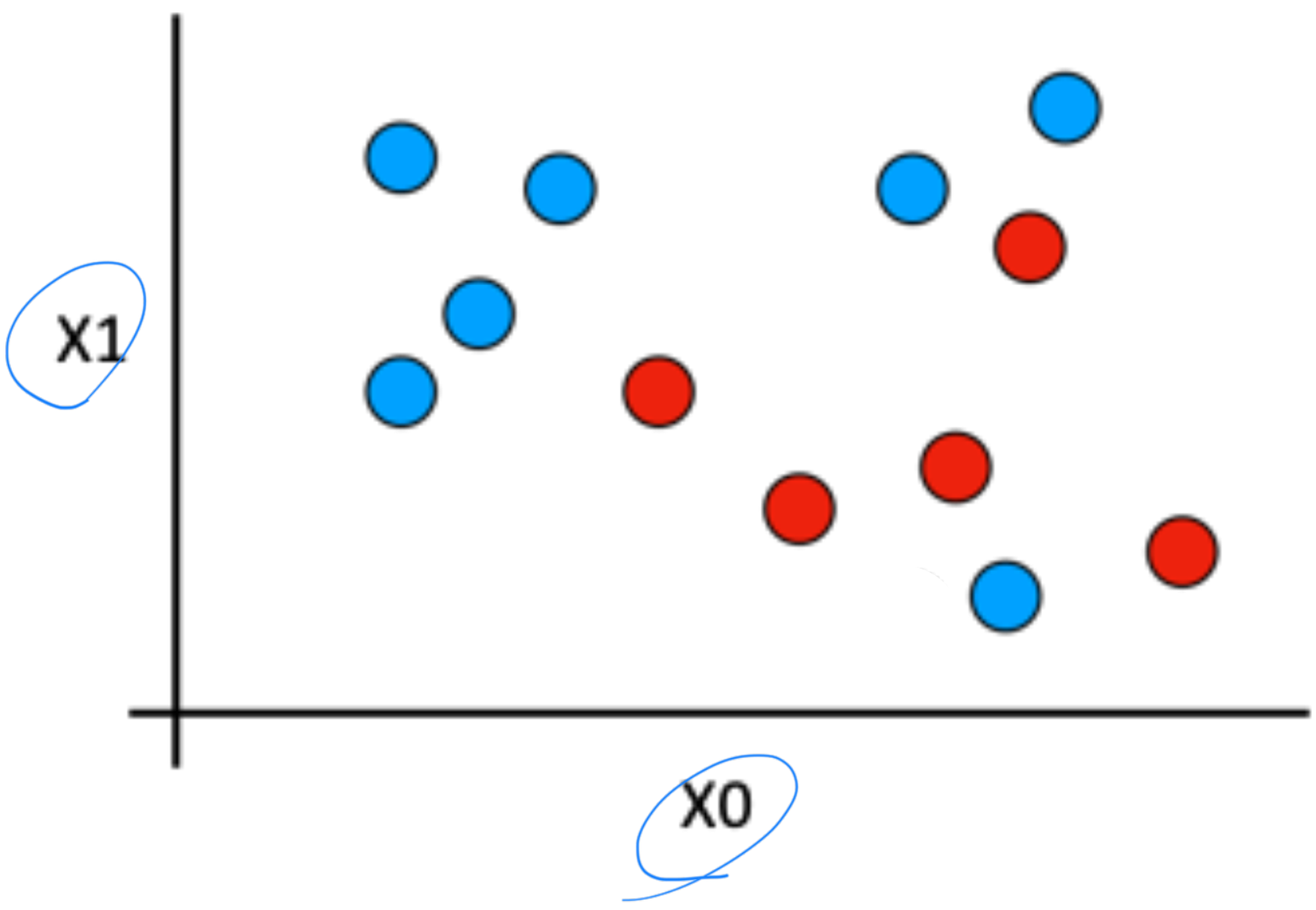
5. How do I organise my code?

6. When asked to calculate the accuracy of the training set, should modify the implementation?

7. How much details in the README?

8. Do we need to compute the Gini-index as well? in addition to the entropy and information gain

kNN practice

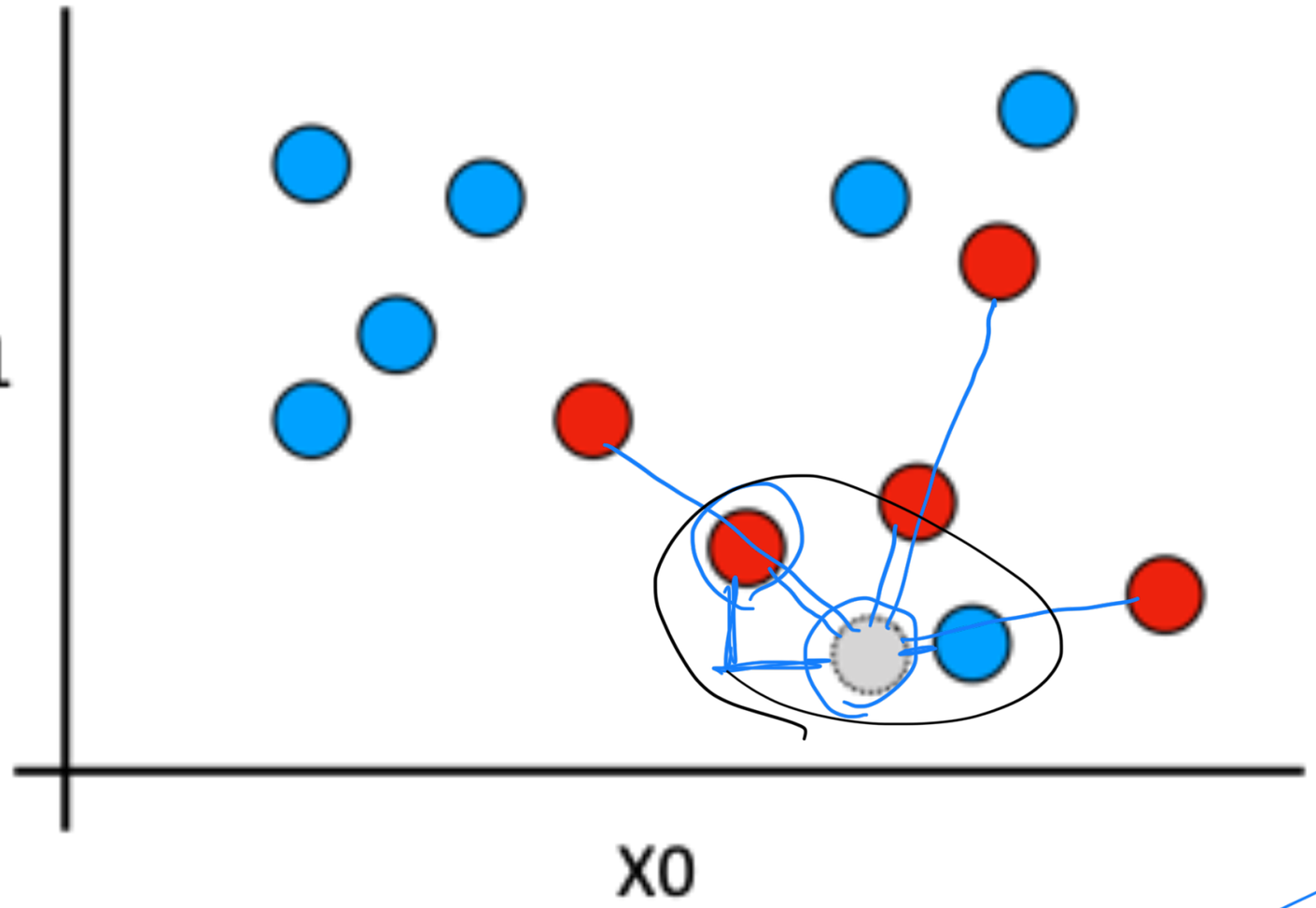


X0	X1	Class
6.5	1	Blue
8.5	1.2	Red
6.3	2	Red
5	1.8	Red
...	...	...
...	...	...

Training

SAVE

kNN practice



X0	X1	Class
6.5	1	Blue
8.5	1.2	Red
6.3	2	Red
5	1.8	Red
...	...	...
...	...	...

Predicting for (X0=6, X1=0.9)

$$d(a,b) = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2}$$

$$\sqrt{[(6-5)^2 + (0.9-1.8)^2]}$$

Preprocessing a dataset?

Example: min-max normalisation

IMPORTANT: learn on the training data, only apply on the test data

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Train

X0	X1	Class
6.5	1	Blue
8.5	1.2	Red
6.3	2	Red
5	1.8	Red
…	…	…
…	…	…

Test

X0	X1	Class
6	0.9	Blue
4.5	3.14	Red
2.4	1	Red
…	…	…

$X_{0_{max}}^{train}$

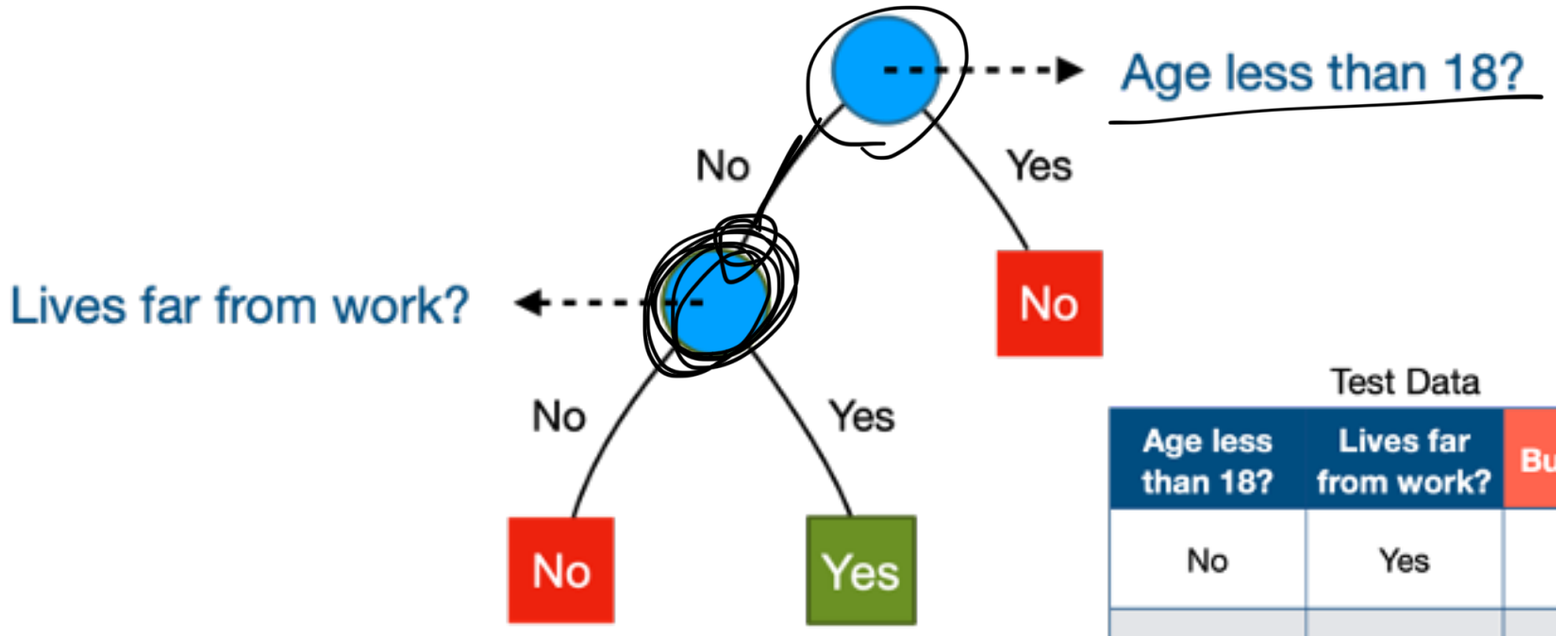
$X_{0_{min}}^{train}$

$X_{1_{max}}^{train}$

$X_{1_{min}}^{train}$

Decision trees

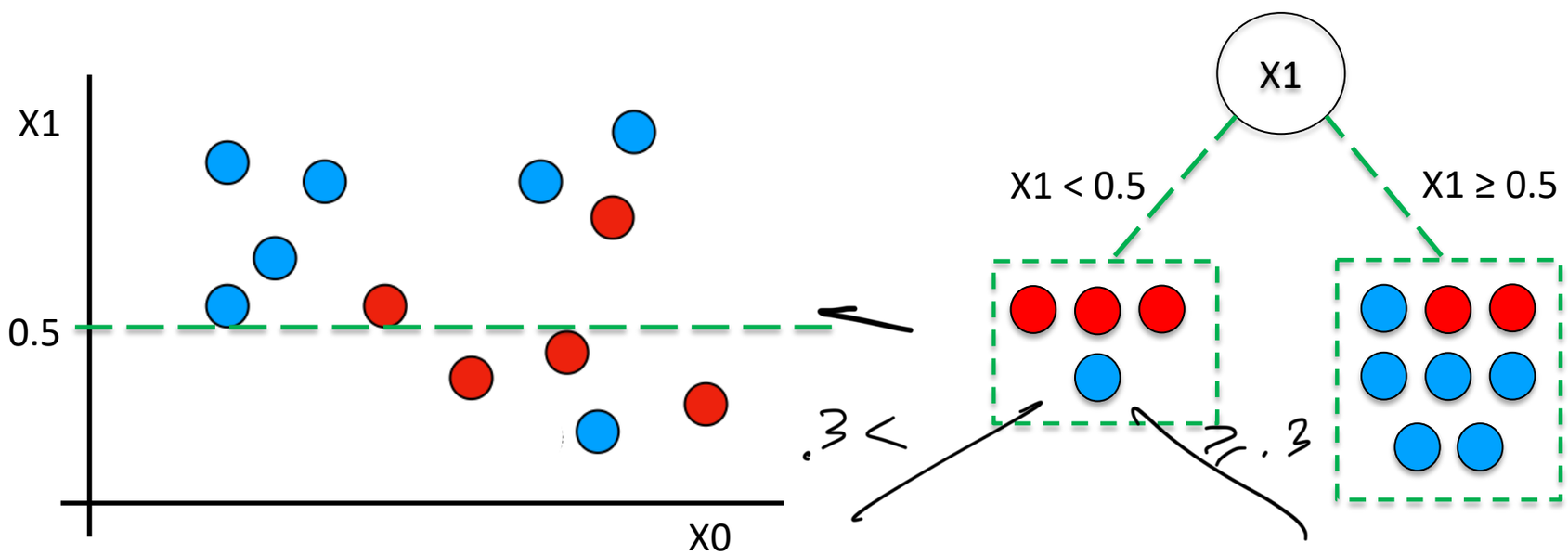
Structure



Test Data

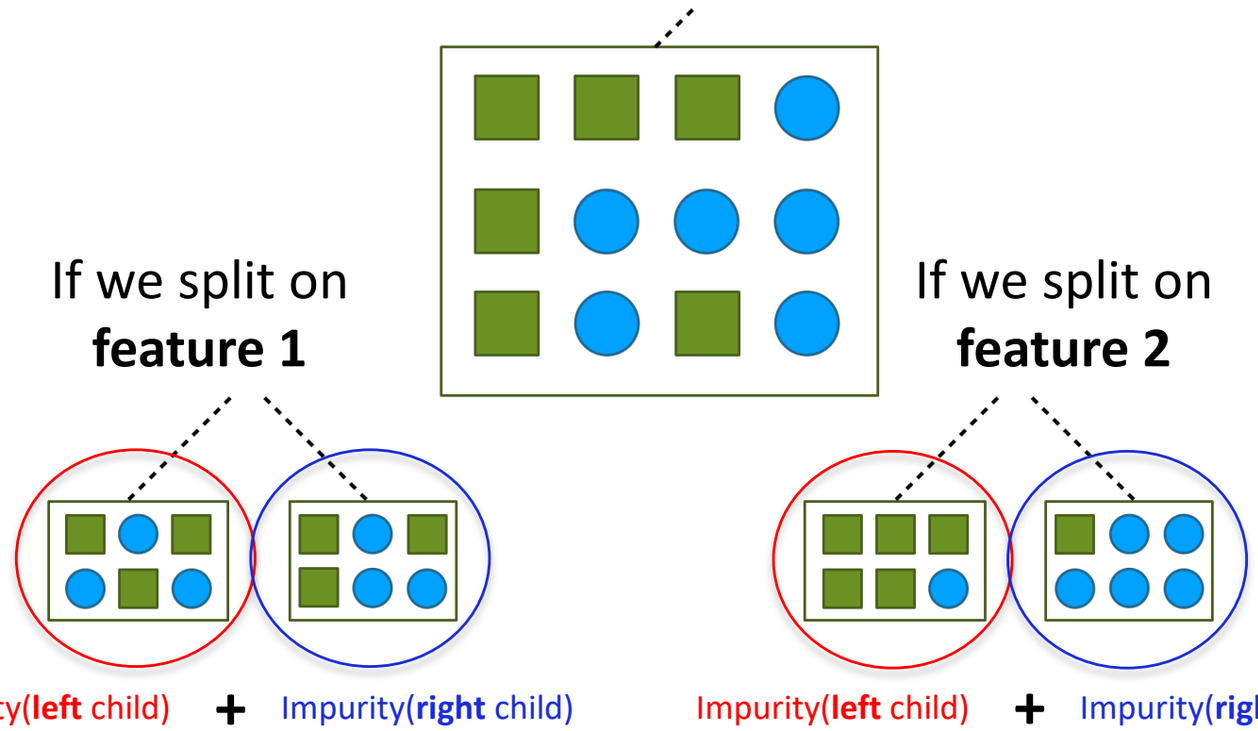
Age less than 18?	Lives far from work?	Buy a car?
No	Yes	
Yes	Yes	

Dividing the space



Impurity & Splitting

Should split?



Impurity measures

Entropy: $H = -\sum_{i=1}^{c} P(i) \cdot \log_2(P(i))$

Gini Impurity: $G = 1 - \sum P(i)^2$

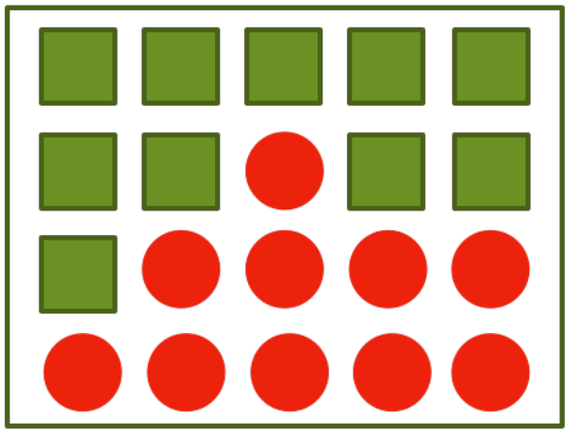
Information Gain

$$IG(F) = H(P) - \sum_{l=1}^{k} \left(\frac{N_l}{N_P}\right) \cdot H(l)$$

Gini Gain?

$$GG(F) = G(P) - \sum_{L=1}^{k} \left(\frac{N_L}{N_P}\right) G(L)$$

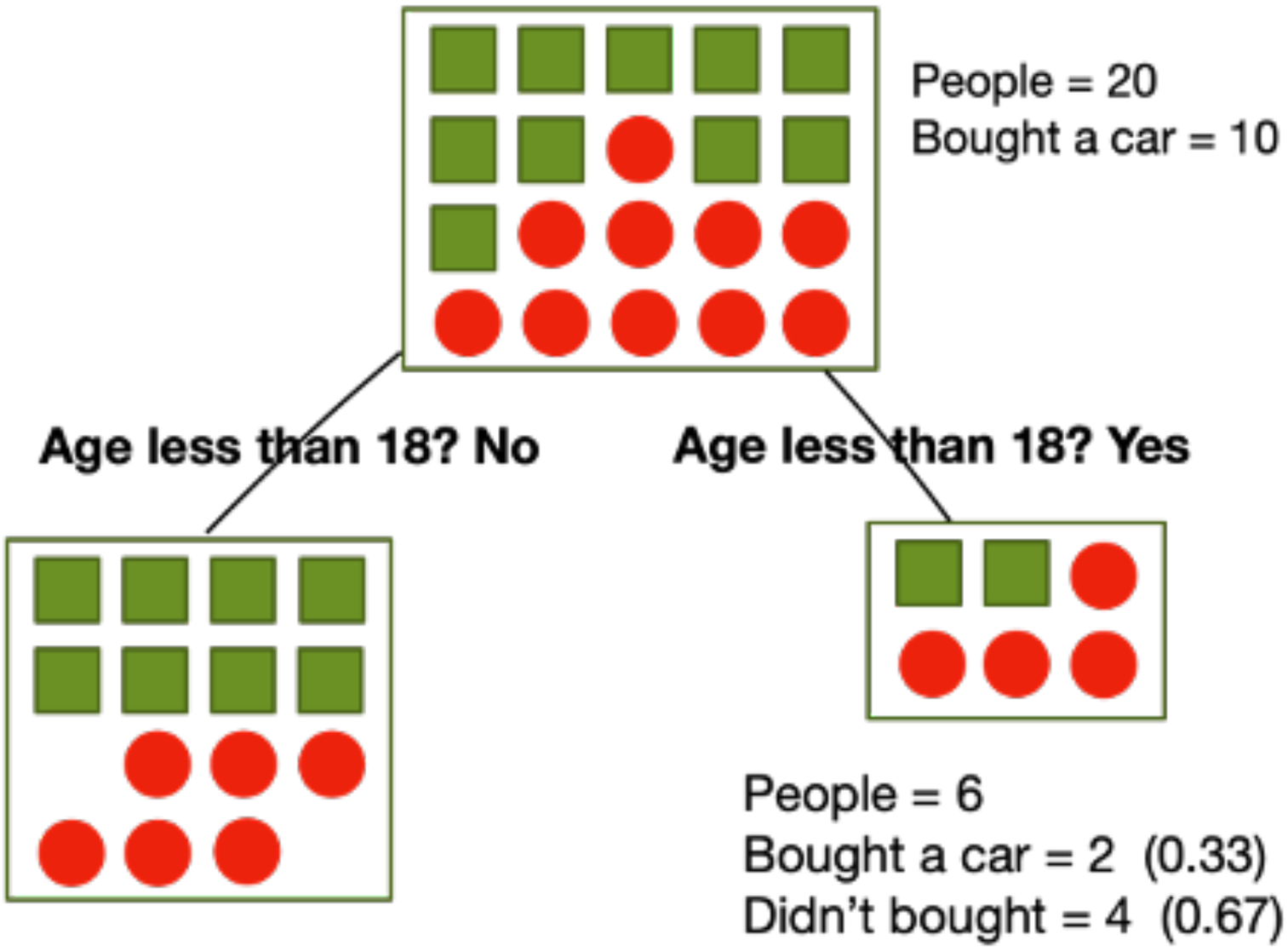
Gini Gain



People = 20
Bought a car = 10 (0.5)

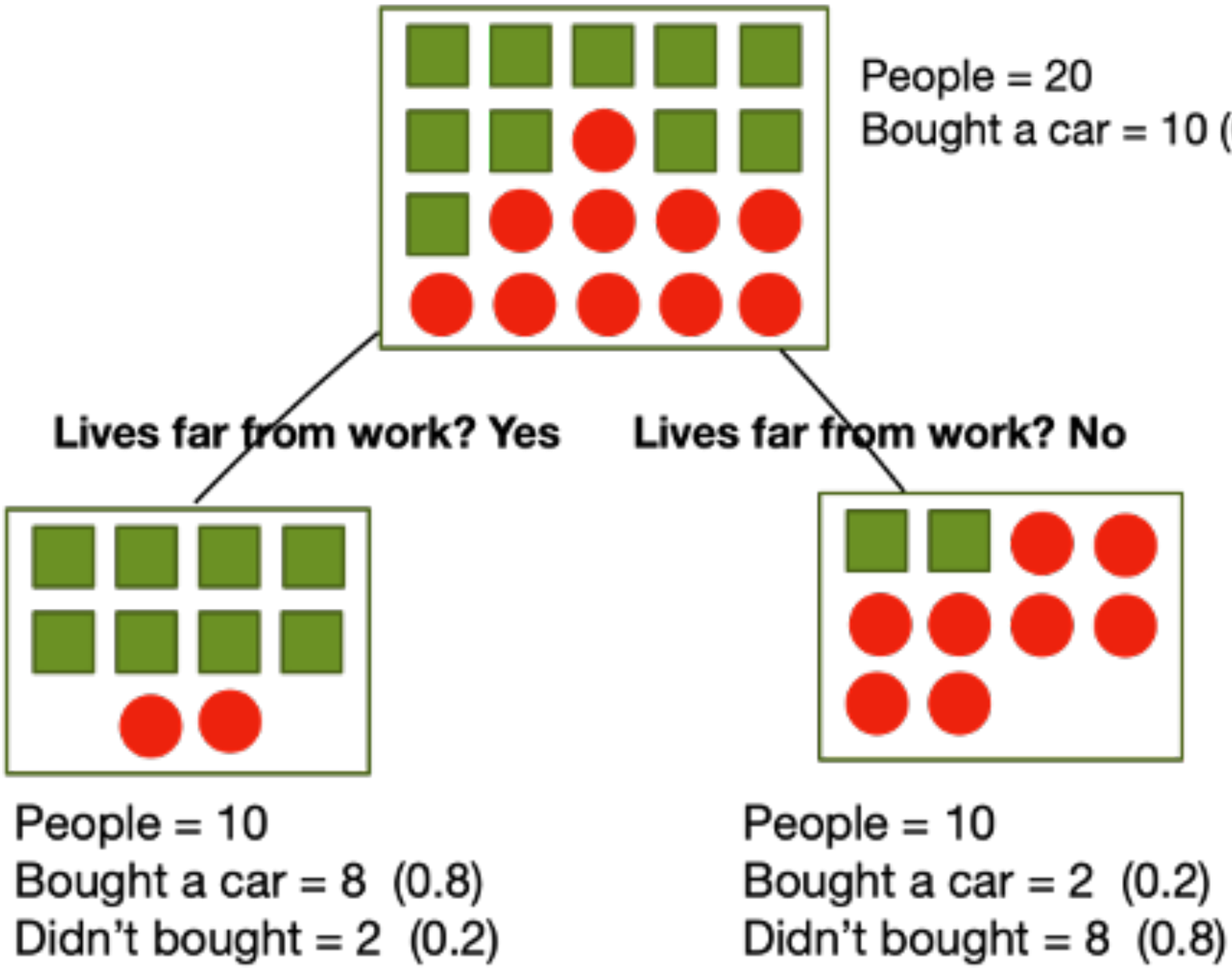
Age less than 18?	Lives far from work?	Buy a car?
Yes	Yes	Yes
Yes	Yes	Yes
No	Yes	Yes
No	Yes	Yes
No	Yes	Yes
No	Yes	Yes
No	Yes	Yes
No	Yes	Yes
No	Yes	Yes
No	No	Yes
No	No	Yes
Yes	No	No
Yes	No	No
Yes	No	No
Yes	No	No
No	No	No
No	No	No
No	No	No
No	No	No
No	Yes	No
No	Yes	No

Gini Gain



Age less than 18?	Lives far from work?	Buy a car?
Yes	Yes	Yes
Yes	Yes	Yes
No	Yes	Yes
No	Yes	Yes
No	Yes	Yes
No	Yes	Yes
No	Yes	Yes
No	Yes	Yes
No	Yes	Yes
No	No	Yes
No	No	Yes
Yes	No	No
Yes	No	No
Yes	No	No
Yes	No	No
No	No	No
No	No	No
No	No	No
No	No	No
No	Yes	No
No	Yes	No

Gini Gain



Age less than 18?	Lives far from work?	Buy a car?
Yes	Yes	Yes
Yes	Yes	Yes
No	Yes	Yes
No	Yes	Yes
No	Yes	Yes
No	Yes	Yes
No	Yes	Yes
No	Yes	Yes
No	Yes	Yes
No	No	Yes
No	No	Yes
Yes	No	No
Yes	No	No
Yes	No	No
Yes	No	No
No	No	No
No	No	No
No	No	No
No	No	No
No	Yes	No
No	Yes	No

Gini Gain: Age less than 18?

Parent node

People = 20
Bought a car = 10 (0.5)

Age less than 18? No

People = 14
Bought a car = 8 (0.57)
Didn't bought = 6 (0.43)

$$G = 1 - [(0.57^2) + (0.43^2)] = 0.49$$

Age less than 18? Yes

People = 6
Bought a car = 2 (0.33)
Didn't bought = 4 (0.67)

$$G = 1 - [(0.33^2) + (0.67^2)] = 0.44$$

$$(14/20) * 0.49 + (6/20) * 0.44 = 0.475$$

Gini Gain: Age less than 18?

Parent node

People = 20
Bought a car = 10 (0.5)

Lives far from work? Yes

People = 10
Bought a car = 8 (0.8)
Didn't bought = 2 (0.2)

Lives far from work? No

People = 10
Bought a car = 2 (0.2)
Didn't bought = 8 (0.8)

$$G = 1 - [(0.2^2) + (0.8^2)] = 0.32$$

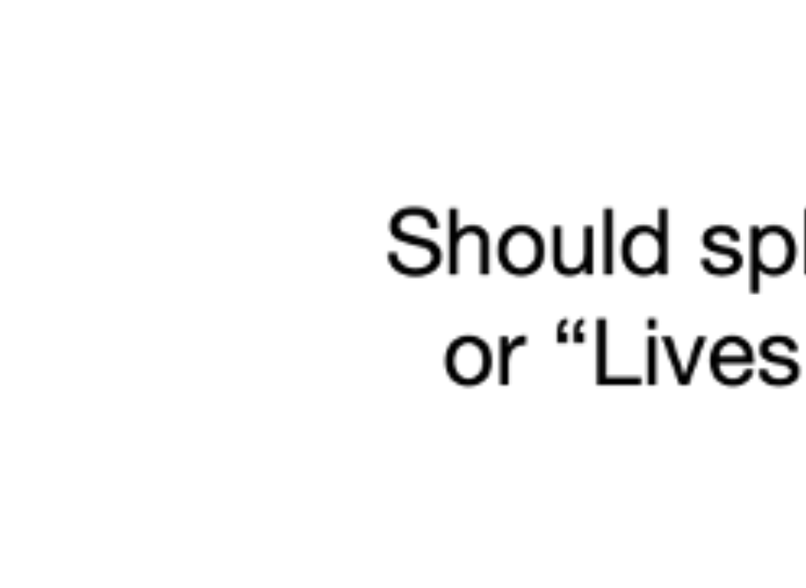
$$G = 1 - [(0.8^2) + (0.2^2)] = 0.32$$

$$(10/20) * 0.32 + (10/20) * 0.32 = 0.32$$

Should split on “Age < 18?”
or “Lives far from work?”
?

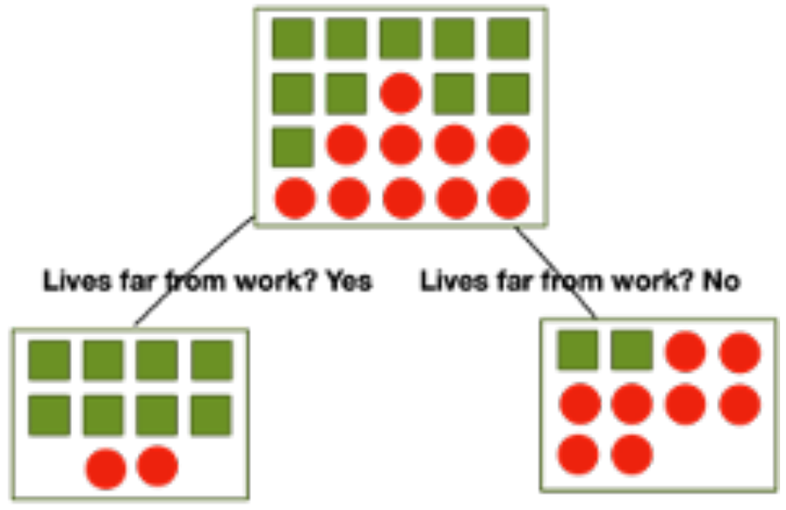
Age < 18?

Gini Gain = 0.5 - 0.475 =
0.025

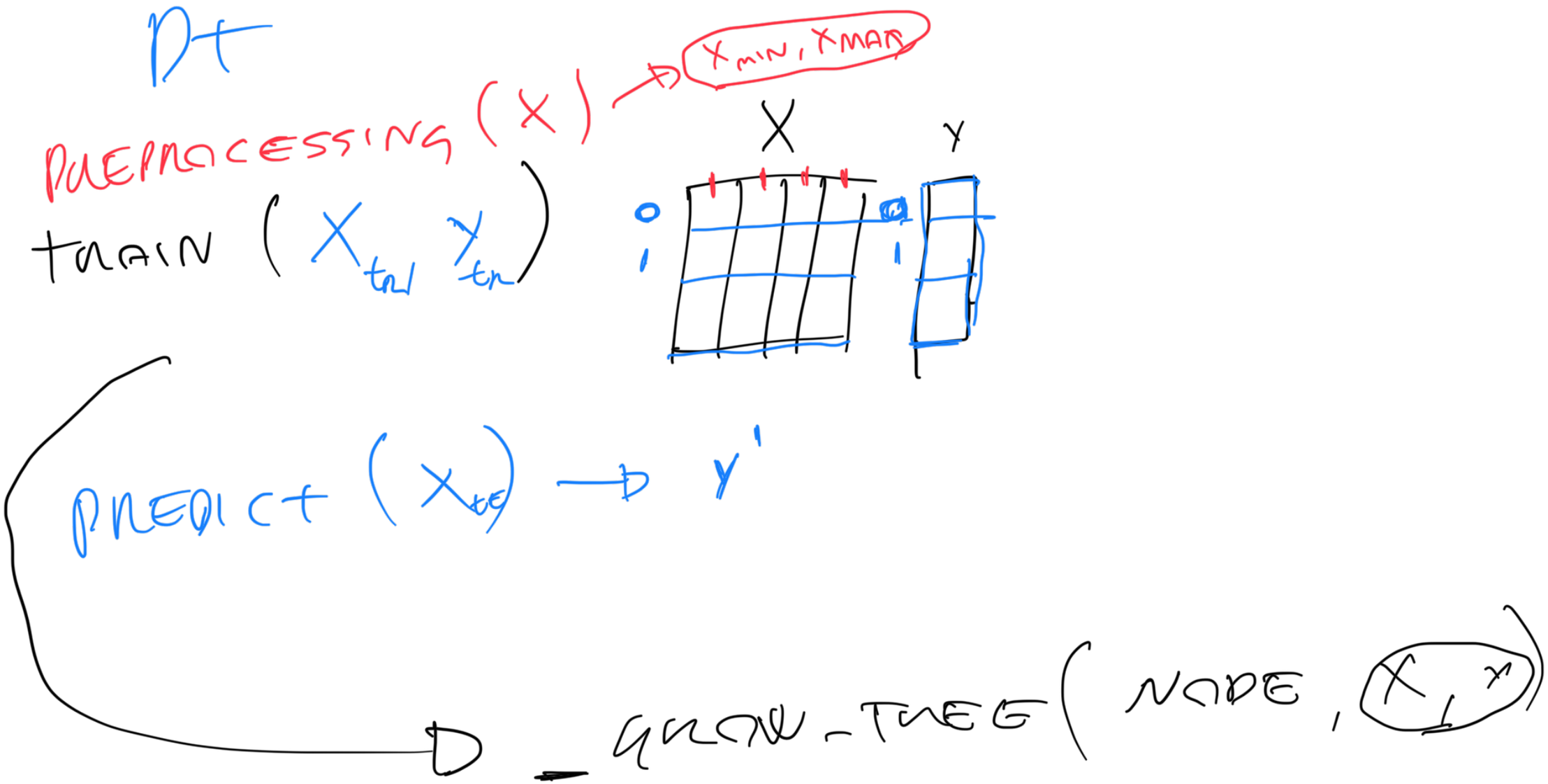


Lives far from work?

Gini Gain = 0.5 - 0.32 =
0.18



Creating a classifier (coding)



Wrap-up

- Chapters from [3]: 19.7 (kNN), 19.3 (DTs)

Next lecture:

- ensemble learning

[3] Russell, Stuart J., and Peter Norvig. *Artificial intelligence a modern approach*. 4th edition