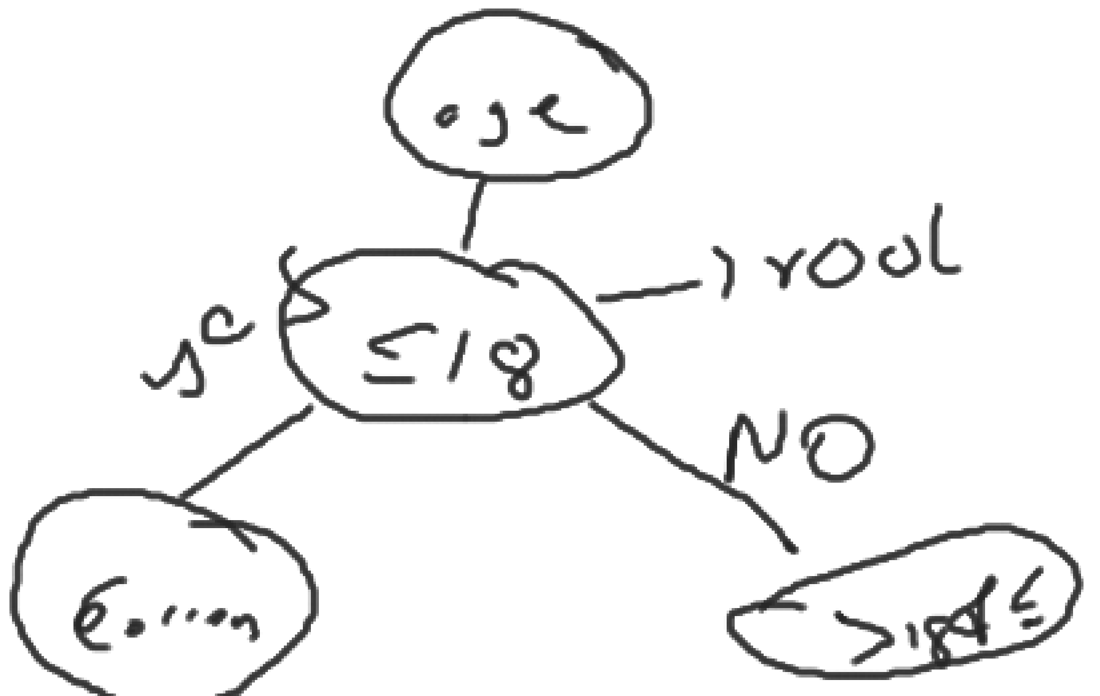
Decision Tree

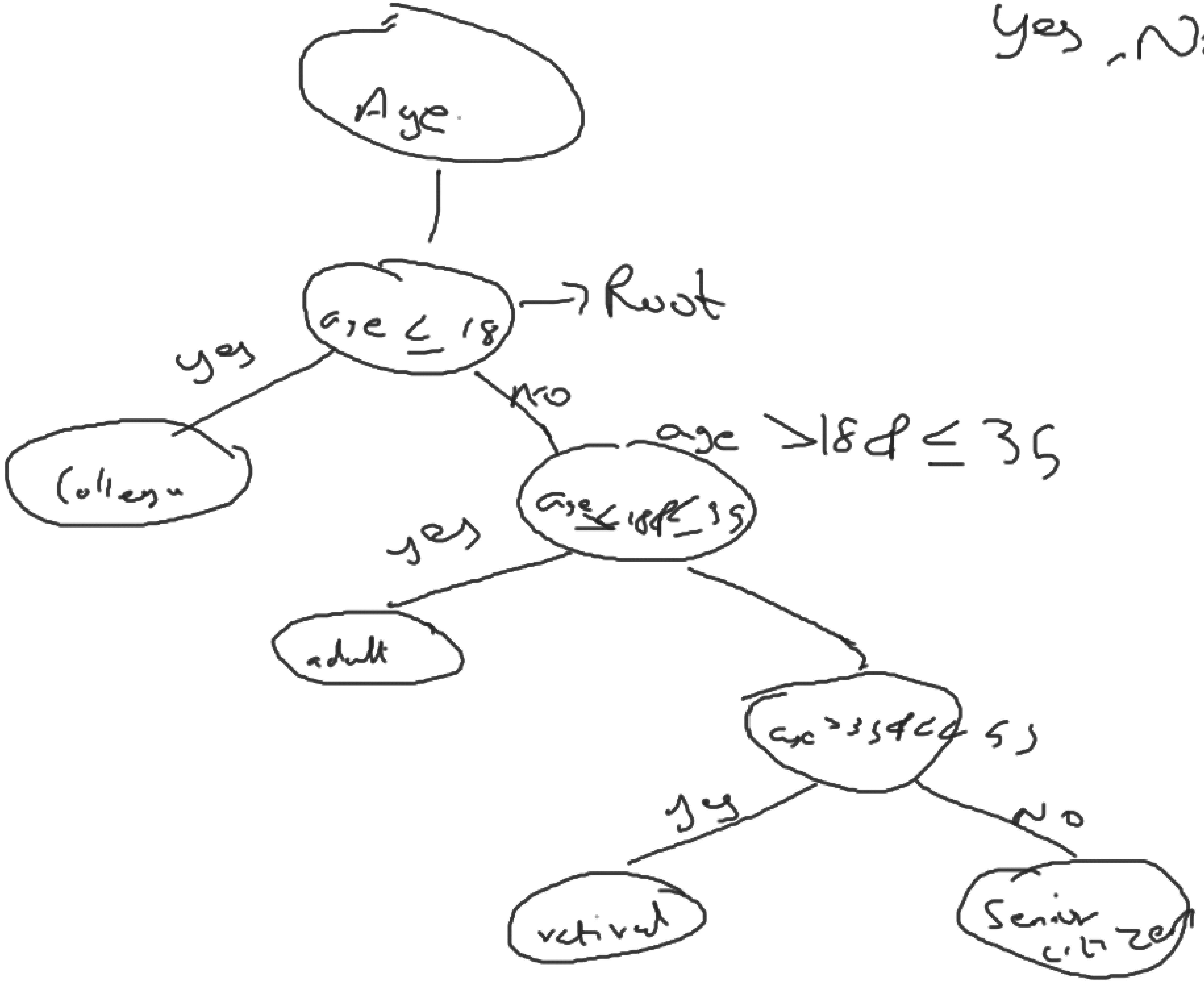
- Regression
- Classification

multiclassification



Age

if (age ≤ 18)
    print (college)
elif (age > 18 & ≤ 35)
    print (adult)
elif (age > 35)
    print (retired)



yes, No

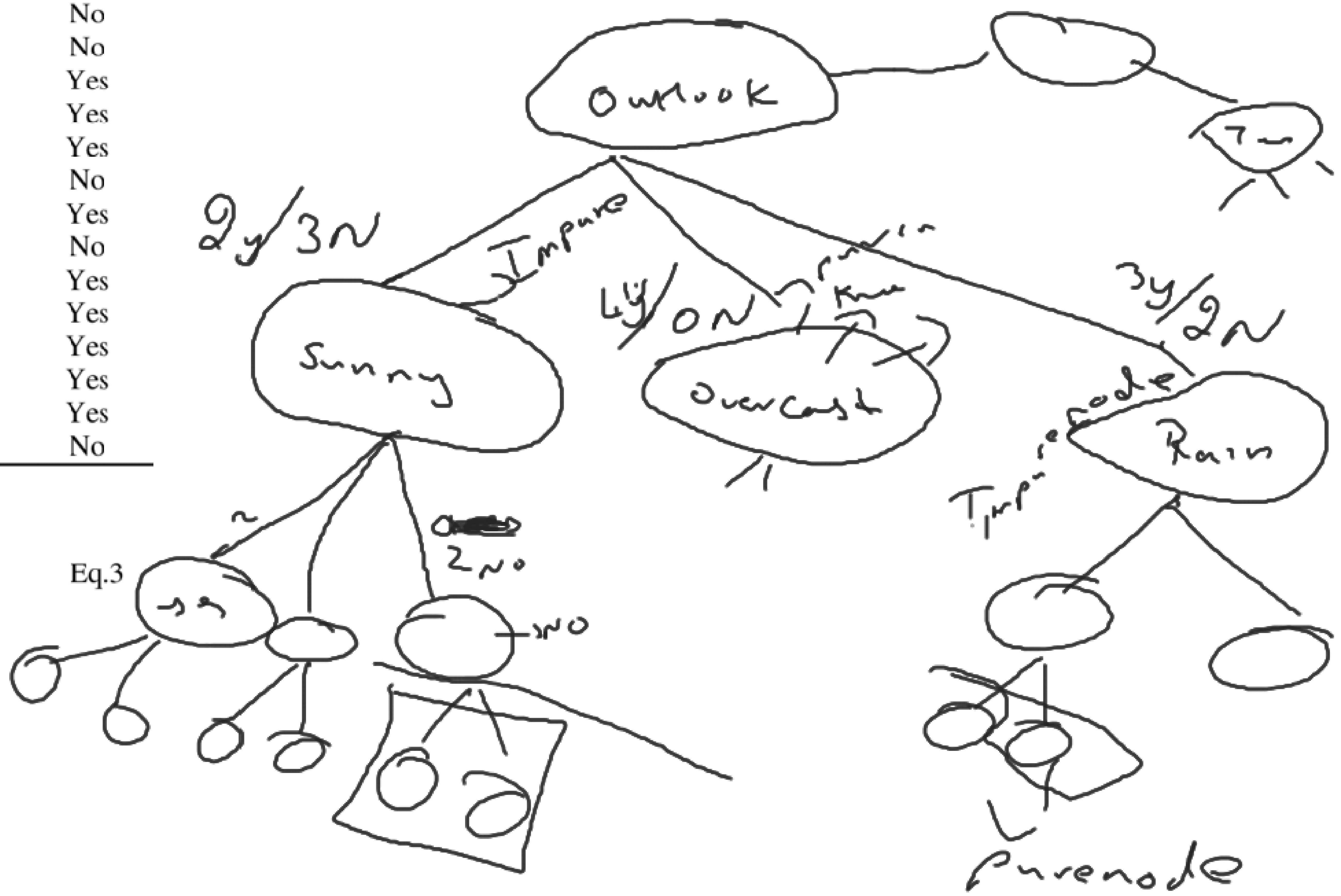
Pure Split

Impure split

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

In this example,

$$Entropy(S) = -\left(\frac{9}{14}\right)\log_2\left(\frac{9}{14}\right) - \left(\frac{5}{14}\right)\log_2\left(\frac{5}{14}\right) = 0.9450 \qquad \text{Eq.3}$$



Purity

-> Entropy ->

-> Gini Impurities (or) Gini Coefficient

This Select particular Column which has
Gain information → good knowlege (or) huge Regarding
data and the Output.

Entropy :

$$H(s) = -p_+ \log_2 p_+ - p_- \log_2 p_-$$

yes, NO

$H(s) = -P_{+} \log_2 P_{+} - P_{-} \log_2 P_{-}$

$$H(s) = -\frac{3}{3} \log_2 \frac{3}{3} - \frac{0}{3} \log_2 \frac{0}{3}$$

Purenode

6y/3No

f_1

3y/3No

c_1

c_2

3y/0n

c_1

$$H(s) = -\frac{3}{6} \log_2 \frac{3}{6} - \frac{3}{6} \log_2 \frac{3}{6}$$

$= \log_2 = 1$ impurenode

$$H(s) = -\frac{3}{3} \log_2 \frac{3}{3} - \frac{0}{3} \log_2 \frac{0}{3}$$

c_2

$= -1 \log_2 1$

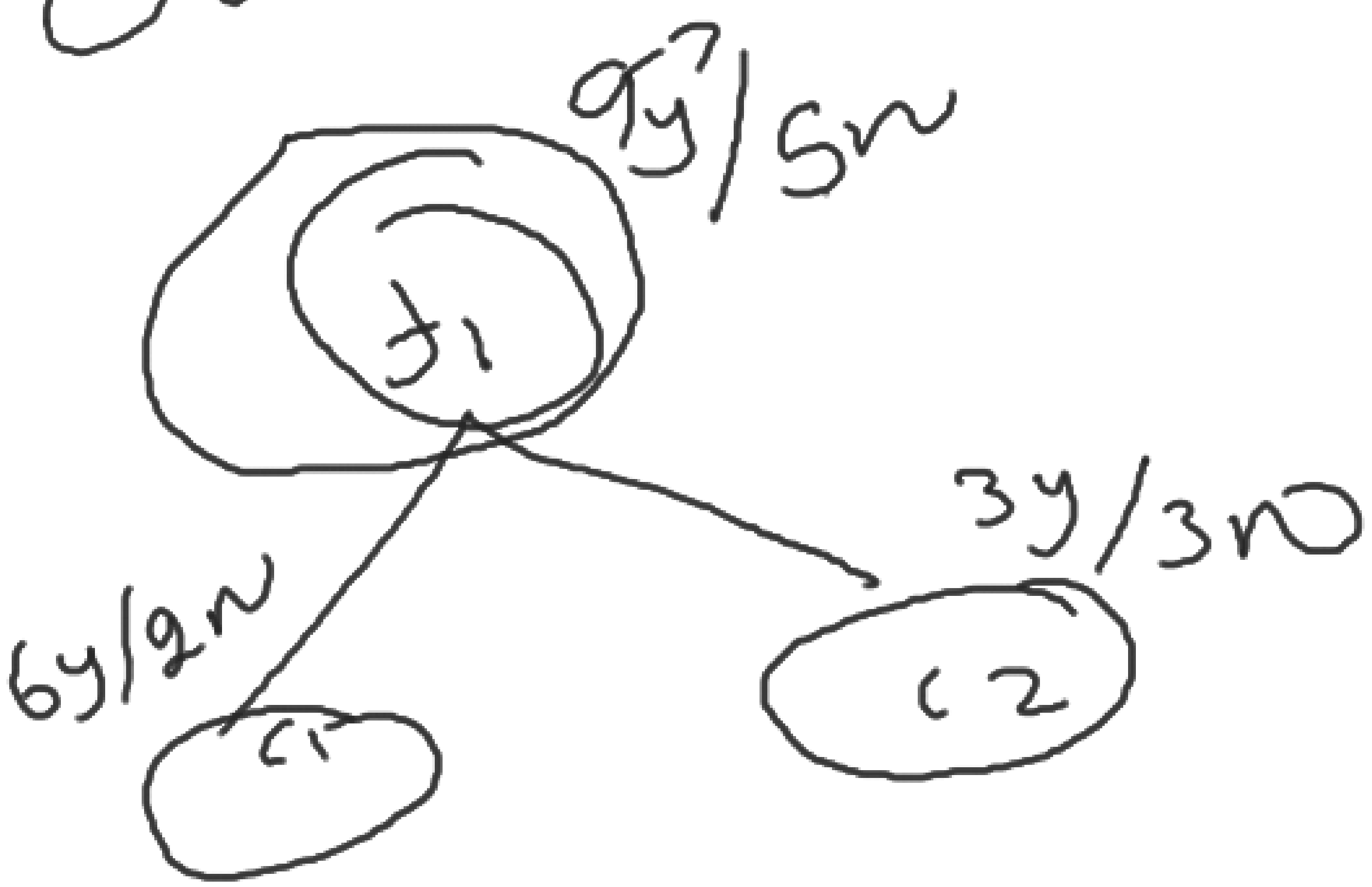
$= 0$ → purenode

Feature Selection

$$H(1) = -p_+ \log_2 p_+ - p_- \log_2 p_-$$

(EG)

Gain information



$$\text{Gain}(S, f_1) = H(S) - \sum_{v \in val} \frac{|S_v|}{|S|}$$

nth entropy term

No of sample data

overall No. of sample

$$H(S) - \sum_{v \in val} \frac{|S_v|}{|S|}$$

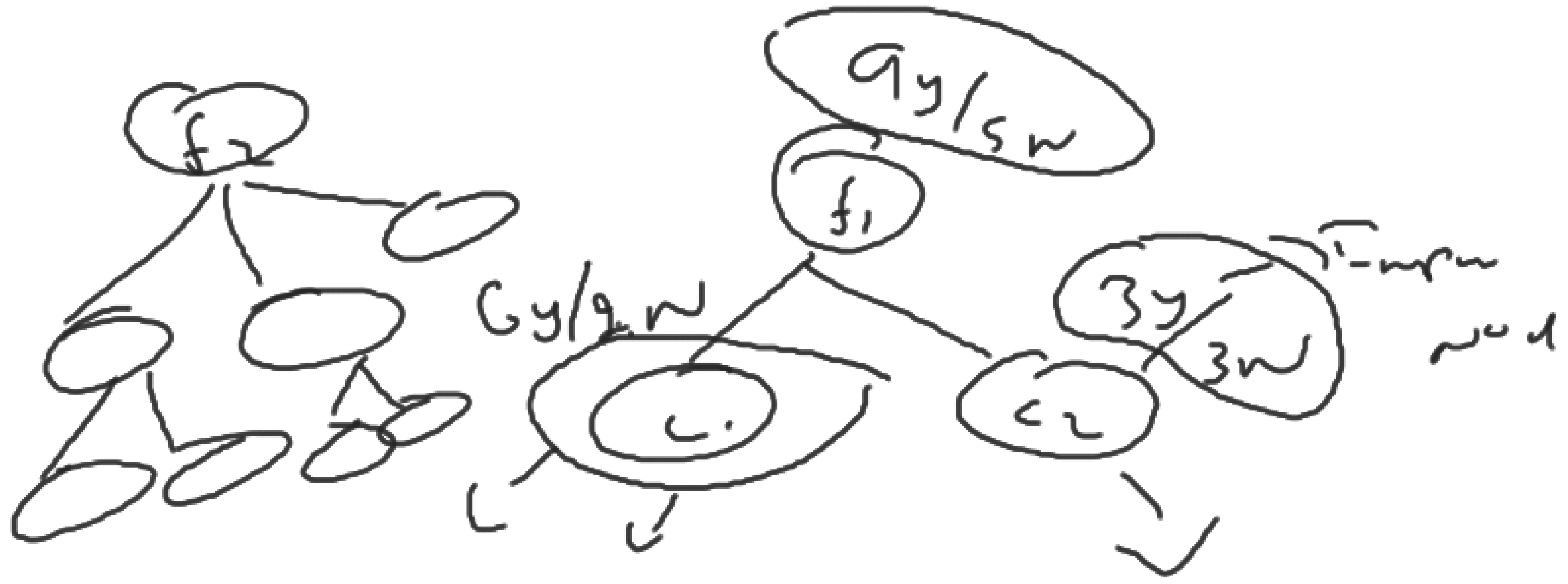
node value

overall sample

$$H(S) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14}$$

$$H(S) \simeq 0.94$$

$$G(S, A) = H(S) - \sum_{v \in \text{values}} \frac{|S_v|}{|S|}$$



$$H(S_v\, c_1) = -\frac{6}{8} \log_2 \frac{6}{8} - \frac{2}{8} \log_2 \frac{2}{8}$$

$$H(S_v\, c_1) = 0.81$$

$$= -\frac{3}{6} \log_2 \frac{3}{6} - \frac{3}{6} \log_2 \frac{3}{6}$$

$$H(S_v\, c_2) = 1$$

$H(S) = 0.94$

$H(S_{v}C_1) = 0.81$

$H(S_{v}C_2) = 1$

$= \underline{H(S)} - \sum_{v \in val} \frac{|S_v|}{|S|} \cdot (S_v)$

feature

$= \underline{\underline{0.94}} - \left[\frac{8}{14} \times 0.81 + \frac{6}{14} \times 1 \right]$

Gain

$\text{Gain}(S, f_1) = 0.049 \approx$

$\underline{\underline{\text{Gain}(S, f_2)}} = 0.072$ more knowledge

Nodes.

Gini Impurity

0 - pure
1 -> Impure

$$G.I = 1 - \sum_{i=1}^{n} (p)^2$$

Gini
0 - pure
0.5 -> Impure node

2y/2N

$$\frac{2}{4} = \frac{1}{2}$$

$$G.I = 1 - \left[(p_+)^2 + (p_-)^2 \right]$$

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

$$= 1 - 0.5$$

$$= 0.5$$ -> Impure node

0 - pure
Gini = 0.5

large dataset

log → much time

Entropy

Entropy

gini impurity ← → gini impurity

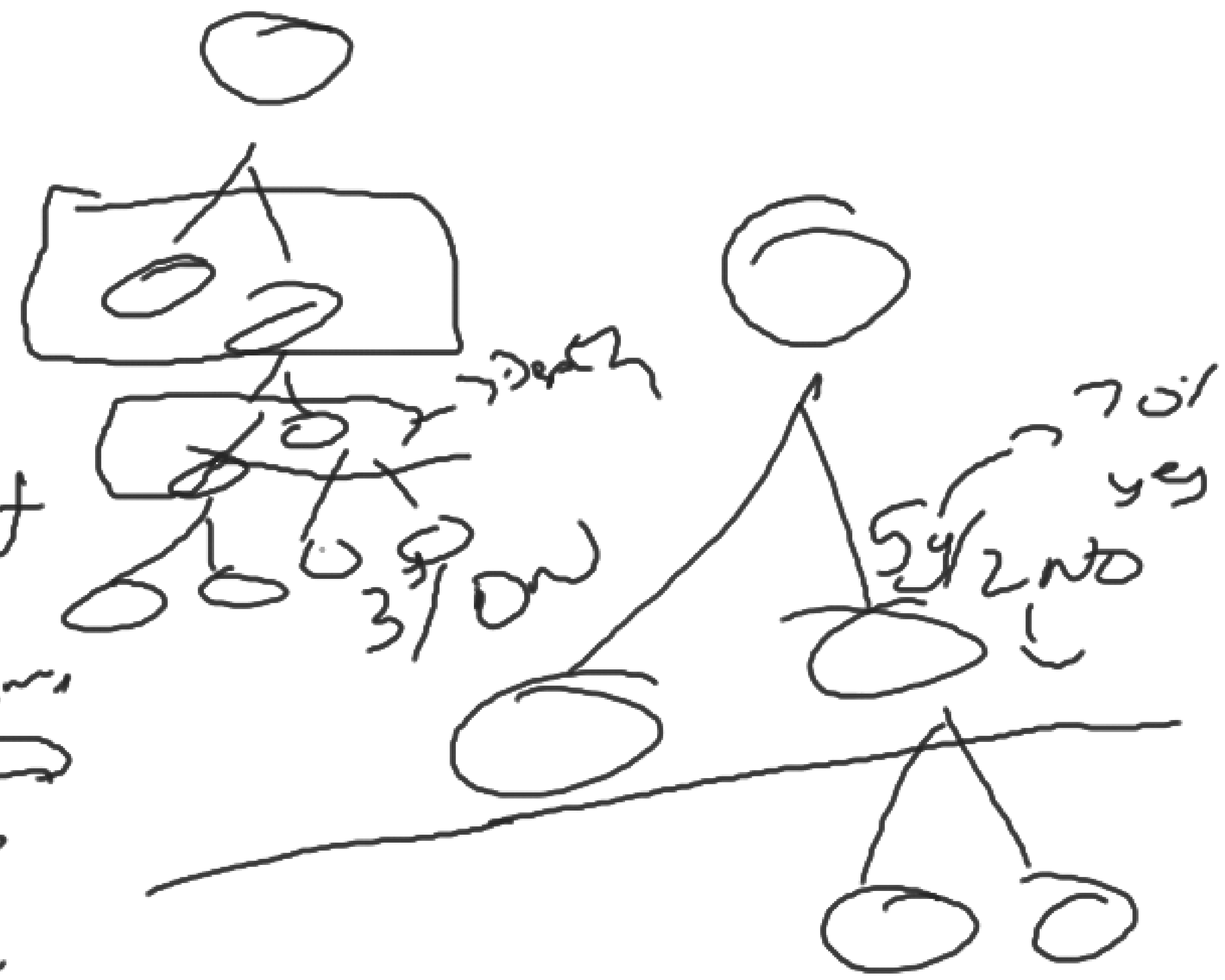
Regression

Continuous

Hyperparameters = Decision

6

Maxdepth; Max. leaf



post pruni

pre prune