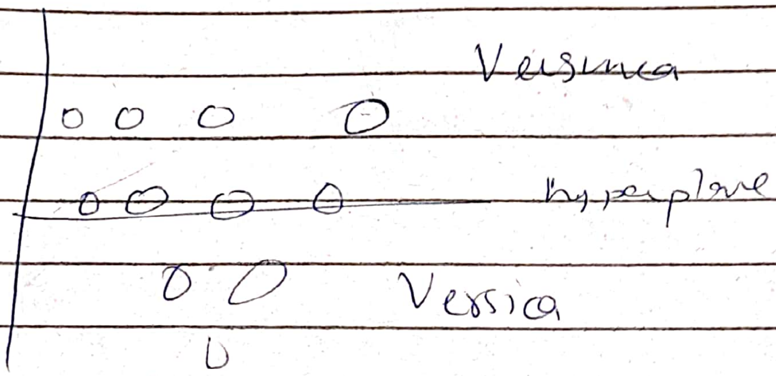
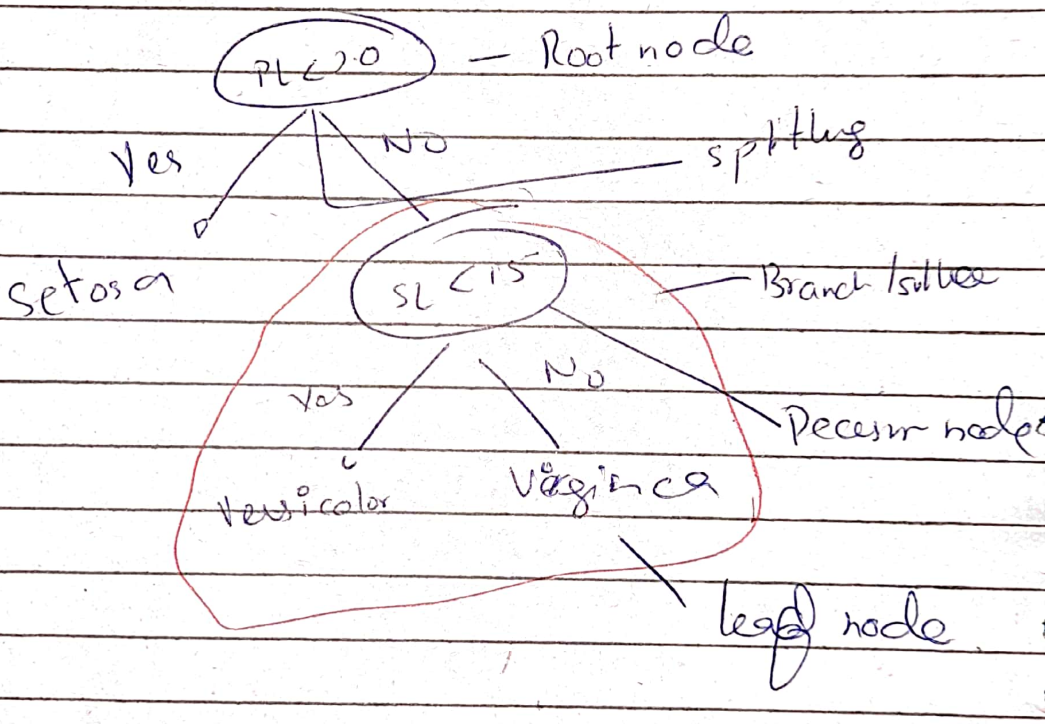
Day: Date:

Decision Tree





$-\sum p_i$ log

Day: Date:

Entropy

measure of disorder

$$E(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$$

$$E(D) = -P_{yes} \log_2 (P_{yes}) - P_{no} \log_2 (P_{no})$$

	Salary	Age	Purchase
1	20000	21	Yes
2	10000	45	No
3	60000	27	Yes
4	150000	31	No
5	120000	18	No

$$H(d) = -P_y \log_2 (P_y) - P_n \log_2 (P_n)$$

$$= -2/5 \log_2 (2/5) - 3/5 \log_2 (3/5)$$

$$H(d) = 0.97$$

Date:

3 terms

$$-P_x \log_2 P_x - P_y \log_2 P_y - P_z \log_2 P_z$$

- More the uncertainty more is entropy
- for a 2 class problem the min entropy is 0 and the max is 1
- for more than 2
  min → 0
  max can be greater than 1
- Both $\log_2$ or $\log_e$ can be used to calculate entropy

Information Gain

Decrease in Entropy

$IG = E(\text{Parent}) - \{\text{Weighted Avg}\}\ E\{\text{Children}\}$

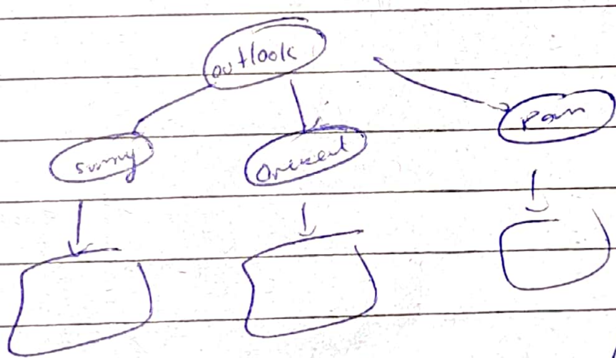
PlayTennis dataset

Entropy of Parent

$E(P) = -P_y \log_2 P_y - P_n \log_2 P_n$

$= -\frac{9}{14} \log_2 \frac{9}{14} - \frac{5}{14} \log_2 \frac{5}{14}$

$= 0.94$



$E(O) =$

$-\frac{4}{4} \log(\frac{5}{5})$

$E(R)$

$= 0.97$

$E(S) =$

$-\frac{2}{5} \log \frac{2}{5} - \frac{3}{5} \log \frac{3}{5}$

$E(O) = 0$

$= 0.97$

Calculate weighted Entropy

$W.E = \frac{5}{14} * 0.97 + \frac{4}{14} * 0 + \frac{5}{14} * 0.97$

Weights

$W.E = 0.67$

Day:________ Date:________

When Entropy is 0
than its a leaf node

I.G = 0.97 - 0.68
= 0.28

→ Gini Impurity

$G_I = 1 - (P_y^2 + P_N^2)$

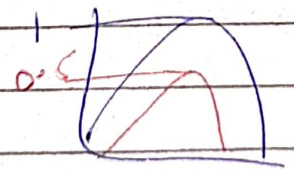
Salary Dataset

$G = 1 - (4/25 + 9/25)$

$G = 0.48$

I.G = P_G - W.T x gini Child

$P_x = 0.5$ ← Entropy = 1
$P_y = 0.5$ Gini = 0.5



Day:________ Date:________

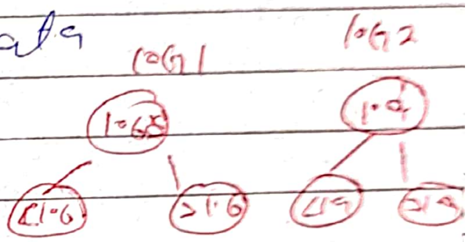
Handling Numerical Value

1	3.5	Yes
2	4.6	Yes
2	2.2	No
4	1.6	Yes
5	1.5	

13

first sort the data

S.No	UserRating	Downloaded
1	1.6	Yes
2	1.9	Yes
3	2.2	No
4	2.5	Yes
5	2.9	Yes
6	3.2	No
7	3.3	No
8	3.5	Yes
9	3.9	No
10	4.1	No
11	4.6	Yes
12		
13	4.8	Yes

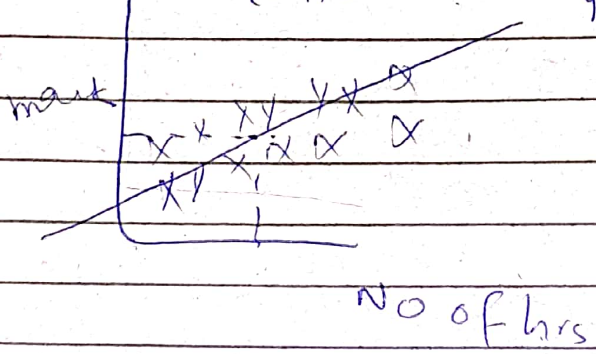


IG

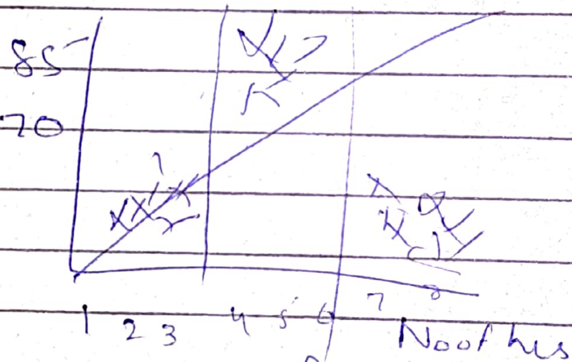
Max {IG1, IG2, IG3, ... IGn}

Regression Trees

$(m, b)\ x$ $\quad y = mx + b$



Case



first cut

if $x < 6$

yes | NO

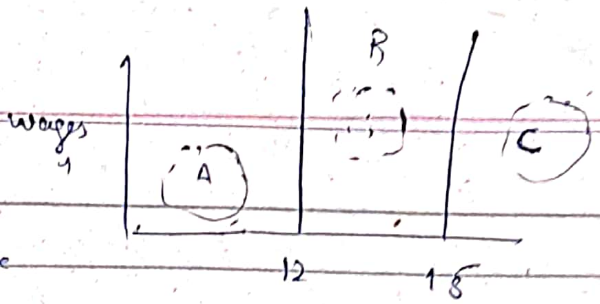
NO → mean score

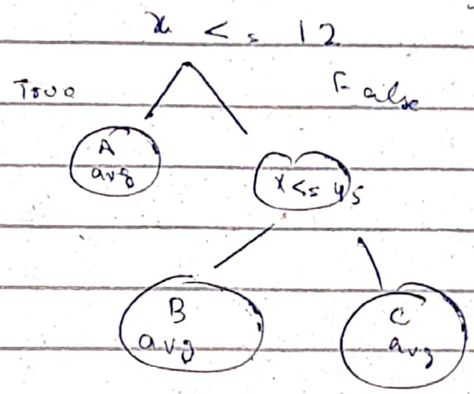
yes → if $x < 3$

yes → mean

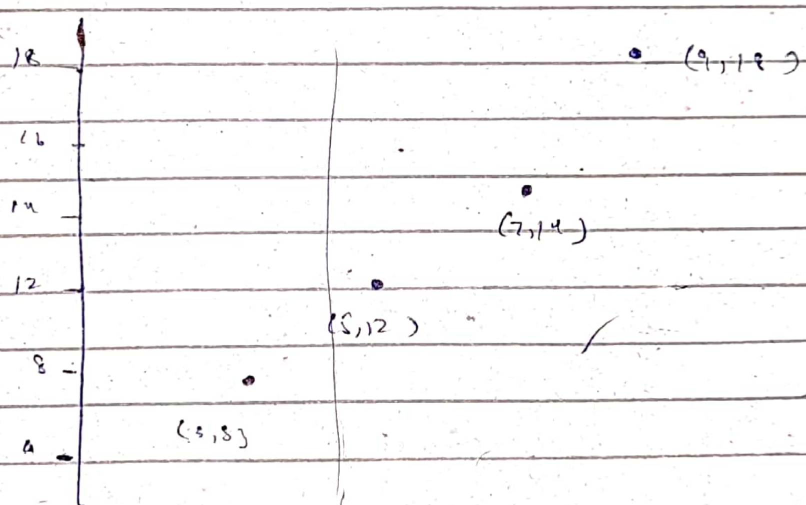
mean

ML





$$\frac{3+5}{2} = 4$$



$$\frac{12+14+18}{3} = 14.6$$

$$= 18.68 \quad (14.6-12)^2 + (14.6-14)^2 + (14.6-18)^2$$

$= (10-8)^2 + (10-12)^2$

$= 8$

error $8/2$

$\frac{8+12}{2} = 10$

$(16-14)^2 + (16-18)^2$

8

error $8/2$

$\frac{14+18}{2}$

$= 16$

6

$8+8 = 16$

1	y
	8
	12
	14
	18

$\frac{8+12+14+18}{4} = 13$

$$SSE = \frac{(13-8)^2 + (13-12)^2 + (13-14)^2 + (13-18)^2}{4}$$

$= 13$

9	18
9.5	18.5
1.0	18.6

$$\frac{18+18.5+18.6}{3}$$

$= 18.36$

$$= \frac{8+12+14}{3}$$

8

$= 11.33$

$$= \frac{(11.33-8)^2 + (11.33-12)^2 + (11.33-14)^2}{3}$$

$= \cancel{18.67}$

$= 6.22$

$$\frac{(18.36-18)^2 + (18.36-18.5)^2 + (18.36-18.6)}{3}$$

$= 0.068$