Prepared and Written By
RAZEEN AHMED

CSE-437

[DATA SCIENCE]

HANDWRITTEN NOTE

Computer Science and Engineering
BRAC University

Github: github.com/razeen
LinkedIn: linkedin.com/in/razeenahmed

MINING ASSOCIATION RULES

* Degree of association can be different

RULES :-

→ Assume all data are categorical

* MARKET BASKET ANALYSIS

I = item database
T = transaction database

The model :-

(i) $I = \{ i_1, i_2, i_3 \ldots i_m \}$

(ii) $t \subseteq I$ (t is a subset of I)

(iii) $T = \{ t_1, t_2 \ldots t_n \}$

$I = \{ A, B, X, Y \}$

$t_1 = \{ A, X, Y \}$

$t_2 = \{ B, X, Y \}$

$t_3 = \{ X, Y \}$

$T = \{ t_1, t_2, t_3 \}$

The model rules :-

$I_1 = \{ A, B, X \}$

$A, B \longrightarrow X$

[If someone buys A & B he will purchase X]

$A, B \longrightarrow X, B$ [X / Violation of rule ! — not possible.]

Application of Association rule

Example → efficient keyboard

⇒ Put the keys which have higher association together

→ Items ⇒ keys

transaction ⇒ most frequent words

transaction ⇒ All the database dictionary

Measure of strength of association :-

[There will be some amount of threshold]

* Support = $P_r(X \cup Y)$ / sup

↑ ↑ → Probability of finding X & Y together.

* Confidence

Conf = ~~$P_r(X$~~ $P_r(Y|X)$

↑ → If X is there then what is the probality that Y will also be there.

Support = $\frac{(X \cup Y).\text{count}}{n}$ ← no. of times they were together

← no. of total transaction

Confidence = $\frac{(X \cup Y).\text{count}}{X.\text{count}}$

GOAL → Find all association rules that support "support threshold" and "confidence threshold."

EXAMPLE :-

minsup = 30 %.

min conf = 80 %.

sup { chicken, clothes, milk } = $\frac{3}{7}$

= 0.4...

This can be considered as association rule as it surpasses minsup threshold 0.3

$$Conf = \frac{(X \cup Y).count}{X.count}$$

$$= \frac{3}{3}$$

$$= 1$$

$$= 100\ \%$$

This can be considered as association rule as it surpasses min conf threshold.

$$\text{Confidence} = \frac{sup(X \cup Y)}{sup(X)} \qquad X \longrightarrow Y$$

APRIORI ALGORITHM.

apriori property :- Any subsets of a frequent itemset that is also a frequent itemset.

"we have to maintain this property"

feasible itemset ⟶ candidate itemset.

K=2

find out, ⟶

C_k ← F_{k-1}

candidate itemset

Problem

It needs to scan the big dataset again & again.

ALGORITHM's 1st step :- finding frequent itemsets,

EXAMPLE :-

2. scan T ⟶

C_3 : Join operation ⟶ (X, Y), (X, Z)

if this condition holds true then we can apply join operation

F_3 ⟶ {123, 135, 234, 124, 134}

After join :- {1, 3, 4}, {1, 3, 5}, {1, 4, 5}, {3, 4, 5}

possible subsets,

C_4 = {{1, 2, 3, 4}, {1, 3, 4, 5}}

subsets from candidate set

{1, 2, 3}, {1, 2, 4}, {1, 3, 4}, {2, 3, 4}

now search this subset in F_3

{1, 4, 5}

no present

If one of the subset is missing in F_3 then we can discard the

candidate set ... or ... prune ... candidate set.
because it the violates the apriori algorithm.

ALGORITH'S 2ND STEP :- GENERATING RULES

Association rules for F_3 → {2, 3, 5}

Conf thres = 80%

1. (i) 2 → 3, 5 | (i) 2 → 3, 5 conf = $\frac{2}{3}$ = 67%
   (ii) 3 → 2, 5
   (iii) 5 → 2, 3
   (iv) 2, 3 → 5
   (v) 2, 5 → 3
   (vi) 3, 5 → 2

for F_2 :-

(i) {1, 3} (ii) {2, 3} (iii) {2, 5} (iv) {3, 5}

1 → 3 | 2 → 3 | 2 → 5 | 3 → 5
3 → 1 | 3 → 2 | 5 → 2 | 5 → 3

find conf.

LIFT THRESHOLD :-

$$\star \text{ Lift} = \frac{\sup(X \cup Y)}{\sup(X) * \sup(Y)} = \frac{n * (X \cup Y).\text{count}}{X.\text{count} * Y.\text{count}}$$

if X & Y has some dependency

$$\star \text{ Lift} = \frac{P(X \cup Y)}{P(X)\,P(Y)} = \frac{P(Y|X)\,P(X)}{P(X)\,P(Y)} = A$$

$$\frac{P(Y)\,P(X)}{P(X)\,P(Y)} = 1$$ ← if X & Y are independent.

PRINCIPAL COMPONENT ANALYSIS.

28/02/25

STATISTICS

BASICS :-

(i) MEAN

Population → n

Sample → n-1

(ii) STANDARD DEVIATION

(iii) VARIANCE → $s^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})}{n-1}$ ⇒ 1 feature

(iv) CO-VARIANCE

↳ $s' \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$ ⇒ 2 feature

(how one feature varies with another feature)

Covariance matrix properties.

$x, y, z > 0$ → 3 dimension

$cov(x,x)$ → covariance of x

$cov(y,y)$ → " " y

$cov(x,y) = cov(y,x)$

MATHEMATICS

BASICS

(i) Eigenvector & Eigen value.

$$A = \begin{bmatrix} 7 & 3 \\ 3 & -1 \end{bmatrix} \quad \& \quad I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Step 1

$$\lambda I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \lambda$$

$$= \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}$$

λ is the eigen value

Step 2

$$A - \lambda I = \begin{bmatrix} 7 & 3 \\ 3 & -1 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 7-\lambda & 3 \\ 3 & -1-\lambda \end{bmatrix}$$

Step-3

$$|A - \lambda I| = \begin{vmatrix} 7-\lambda & 3 \\ 3 & -1-\lambda \end{vmatrix}$$

$$= (7-\lambda)(-1-\lambda) - 9$$

$$= -7 - 7\lambda + \lambda + \lambda^2 - 9$$

$$= -7 - 7\lambda + \lambda^2 - 9$$

$$= -16 - 6\lambda + \lambda^2$$

$$\Rightarrow \lambda^2 - 6\lambda - 16 = 0$$

$$\therefore \lambda = 8 \quad | \quad \therefore \lambda = 2$$

Step-4

Substituting $\lambda = 8$ in

$$B = \begin{bmatrix} -1 & 3 \\ 3 & -9 \end{bmatrix}$$

Property of eigenvector —

$$B\bar{v} = \bar{0}$$

3×3

$B\bar{v} = \bar{0}$

$$\begin{bmatrix} -1 & 3 \\ 3 & -9 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

E collar form transformation

$$\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}$$

↓

$$\begin{bmatrix} a' & b' & c' \\ 0 & e' & f' \\ 0 & 0 & 1 \end{bmatrix}$$

B transform into E collar form →

$R_2 = 3 * R_1 + R_2$

$$B \Rightarrow \begin{bmatrix} -1 & 3 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Now,

$\Rightarrow -1v_1 + 3v_2 = 0 \longrightarrow (ii)$

$\Rightarrow 0 * v_1 + 0 * v_2 = 0$

$0 = 0$

from (i) we get,

$\Rightarrow -1v_1 + 3v_2 = 0$

$\Rightarrow -v_1 = -3v_2$

$\therefore v_1 = 3v_2$

$\begin{bmatrix} 6 \\ 2 \end{bmatrix} = 2\begin{bmatrix} 3 \\ 1 \end{bmatrix}$ if $v_2 = 2$, $v_1 = 6$

if $v_2 = 1$

$v_1 = 3$

why ?

Eigenvector →

$$V \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 3 \\ 1 \end{bmatrix}$$

Covariance matrix will be a square matrix

Test →

$$A\bar{v} = \lambda \bar{v}$$

Random vector × eigen vector = eigen vector × eigen value.

$$\Rightarrow \begin{bmatrix} 7 & 3 \\ 3 & -1 \end{bmatrix} \begin{bmatrix} 3 \\ -1 \end{bmatrix} = 8 \begin{bmatrix} 3 \\ 1 \end{bmatrix}$$

$$\Rightarrow \begin{bmatrix} 7*3 + 3*1 \\ 3*3 + -1*1 \end{bmatrix} = \begin{bmatrix} 24 \\ 8 \end{bmatrix} \text{ (valid)}$$

PCA → (Unsupervised method)

Feature vector, $FV^* = (\bar{V}_2, \bar{V}_1)$ eigen vector = $\begin{bmatrix} V_1 & V_2 \\ \cdots & \cdots \end{bmatrix}$

In a way that J^* → largest eigenvalue, least eigenvalue

Final data, $FD = FV^T * AdjData^T$

$PC = FD^T$

$\begin{bmatrix} PC_1 & PC_2 \\ \cdots & \cdots \end{bmatrix}$

How can we select 99% variability?

→

eigenvalues = $\begin{bmatrix} 0.049083 \\ 1.28403 \end{bmatrix}$

$\lambda_1 = 0.049$

$\lambda_2 = 1.284$

$\lambda_1 + \lambda_2 = 1.333$

$FV = [v_2 \quad v_1]$

↑ PC1 ↑ PC2

PC1 holds the variability

$PC_1 = \frac{\lambda_2}{\lambda_2 + \lambda_2} = \frac{1.284}{1.333} = 0.963 \approx 96.3\%$

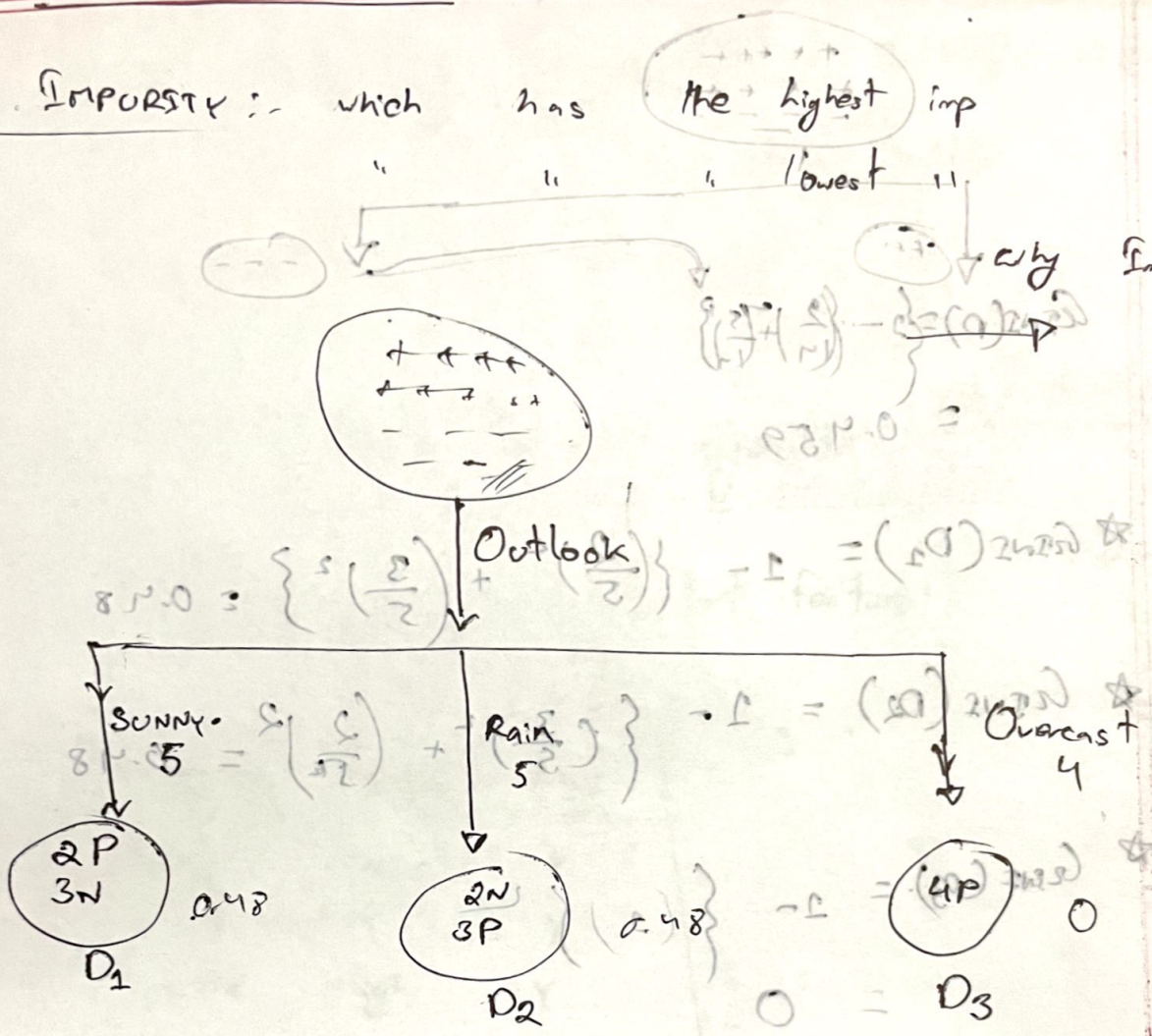
$PC_2 = 3.7\%$

→ $PC_1 \rightarrow 96.3\%$

→ If it is said we can take 96% of the variability then we can take PC1 and drop PC2

RANDOM FOREST

GINI IMPURITY :- which has the highest imp / lowest ... why Impurity?



D

$$GINI(D) = \frac{4}{6} \times \frac{2}{6} + \frac{2}{6} \times \frac{4}{6}$$

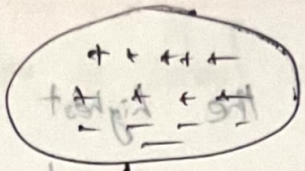
GINI Another way :-

$$GINI(D) = \frac{4}{6} \times \left(1 - \frac{4}{6}\right) + \frac{2}{6} \times \left(1 - \frac{2}{6}\right)$$

eqn →

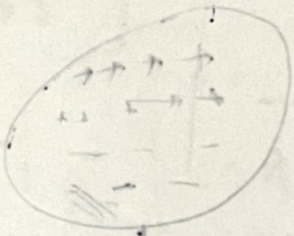
$$\text{Eq}(P) = 1 - \sum_{i=1}^{n} P_i^2$$

$$GINI(D) = 1 - \left\{\left(\frac{4}{6}\right)^2 + \left(\frac{2}{6}\right)^2\right\}$$



$$GINI(D) = 1 - \left\{\left(\frac{2}{14}\right)^2 + \left(\frac{5}{14}\right)^2\right\}$$

$$= 0.459$$



★ $GINI(D_1) = 1 - \left\{\left(\frac{2}{5}\right)^2 + \left(\frac{3}{5}\right)^2\right\} = 0.48$

★ $GINI(D_2) = 1 - \left\{\left(\frac{3}{5}\right)^2 + \left(\frac{2}{5}\right)^2\right\} = 0.48$

★ $GINI(D_3) = 1 - \{(1)\}$

$= 0$

$$\rightarrow GINI(D_{outlook}) = \frac{5}{14} * GINI(D_1) + \frac{5}{14} * GINI(D_2) + \frac{4}{14} * GINI(D_3)$$

$$= \frac{5}{14} * 0.48 + \frac{5}{14} * 0.48 + \frac{4}{14} * 0$$

$$= 0.343$$

☆ Most Reduction will be the discriminative.

Reduction of Impurity by outlook = 0.459 - 0.343

= 0.116

Decision tree construction use cart algorithm 01/03/25

How to find out the most important feature?

Day	Temperature	Humidity	Play
10	Mild	Normal	Y
11	Mild	N	Y
12	Mild	High	Y
13	Hot	N	Y
14	Mild	IA	N
2	Hot	H	N

D (+++ / +--) 0.44 + GINI(D) = Impurity $= 1 - \{P_+^2 + P_-^2\}$

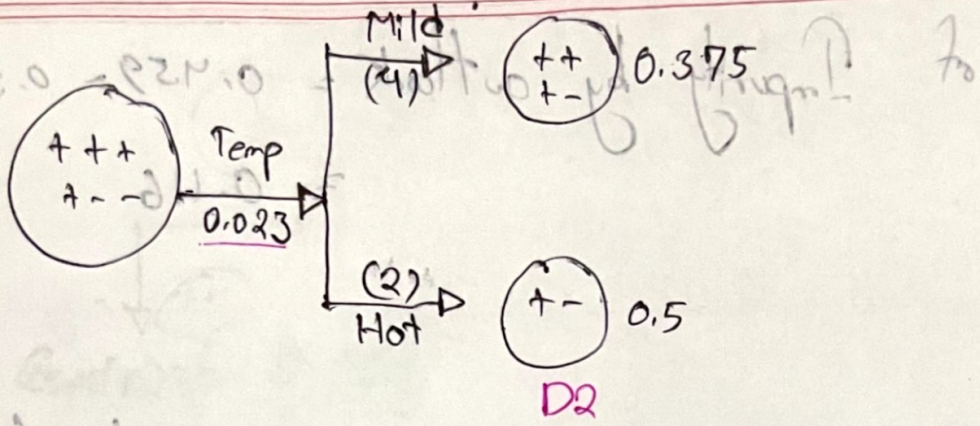
$= 1 - \left\{\left(\frac{4}{6}\right)^2 + \left(\frac{2}{6}\right)^2\right\}$

$= 1 - \frac{16}{36} + \frac{4}{36}$

$= 1 - \frac{20}{36}$

$= \frac{16}{\quad} = 4/9 = 0.44$

Random forest uses → Classification, Regression, feature selection



$$GINI(D1) = 1 - (P_+^2 + P_-^2)$$

$$= 1 - \left\{ \left(\frac{3}{4}\right)^2 + \left(\frac{1}{4}\right)^2 \right\}$$

$$= 1 - \left(\frac{9}{16} + \frac{1}{16} \right)$$

$$= 1 - \frac{10}{16}$$

$$= \frac{6}{16}$$

$$= \frac{3}{8} = 0.375$$

$$GINI(D2) = 0.5$$

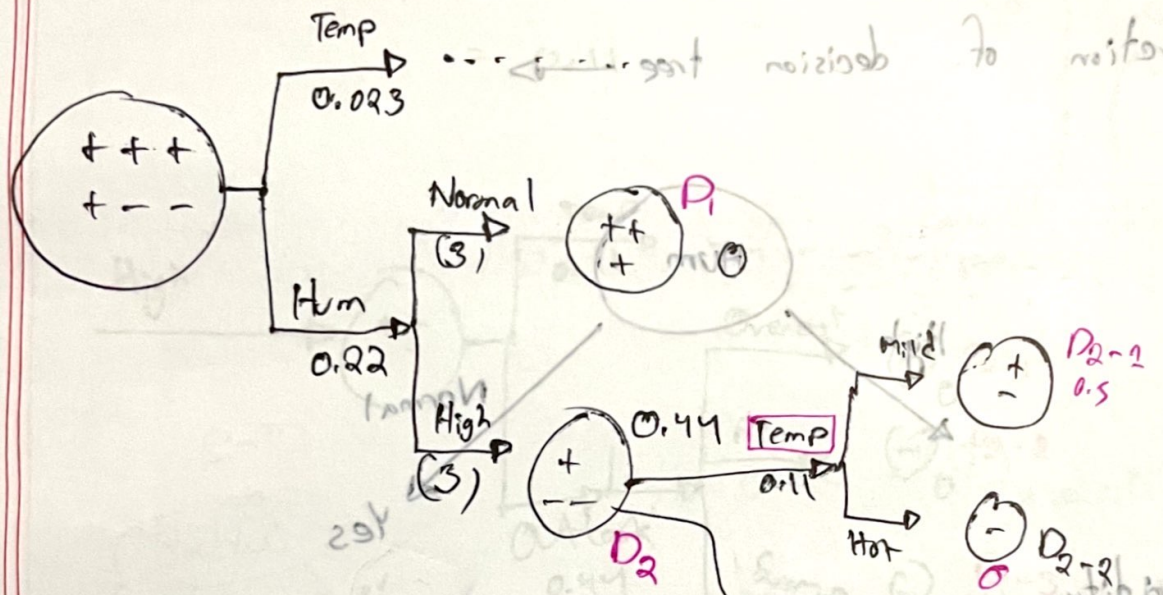
$$\Rightarrow GINI_{Temp}(D) = \frac{|D_1|}{|D|} GINI(D_1) + \frac{|D_2|}{|D|} GINI(D_2)$$

$$= \frac{4}{6} * 0.375 + \frac{2}{6} * 0.5$$

$$= 0.417$$

Reduction of GINI by Temp = 0.44 - 0.417

= 0.023



$$GINI(D_2) = 1 - \left\{\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2\right\}$$ outlook <0.44>

$$= 1 - \frac{1}{9} - \frac{4}{9}$$

$$= 0.44$$

$$GINI_{Hum}(D) = \frac{|D_1|}{|D|} GINI(D_1) + \frac{|D_2|}{|D|} GINI(D_2)$$

$$= \frac{5}{6} * 0 + \frac{5}{6} * 0.44$$

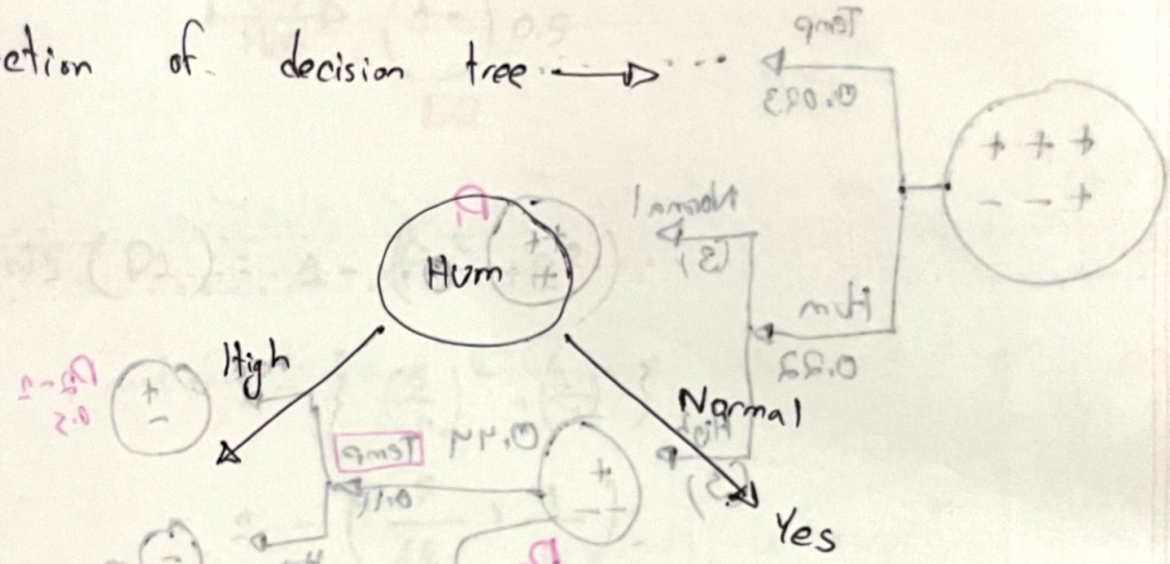
$$= 0.22$$

have to take at least two col...

Reduction of GINI by humidity = 0.44 − 0.22

= 0.22.

Construction of decision tree →



For humidity we are selecting another two column from the bootstrapped dataset -1

Day	Outlook	Temperature	Play
12	Overcast	mild	Y
14	Rain		
2	Sunny		

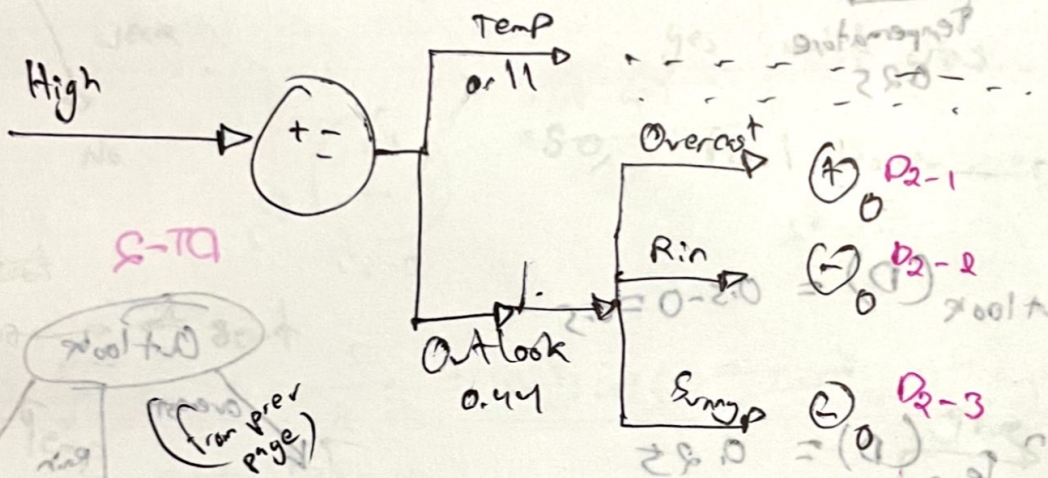
How many bootstrap do we need?

→ No. of trees we need

$$GINI_{TEMP}(D2) = \frac{2}{3} \times 0.5 + 1 + \frac{3}{3} \times 0 = 0.33$$

Reduction Temp = 0.44 − 0.33

= 0.11

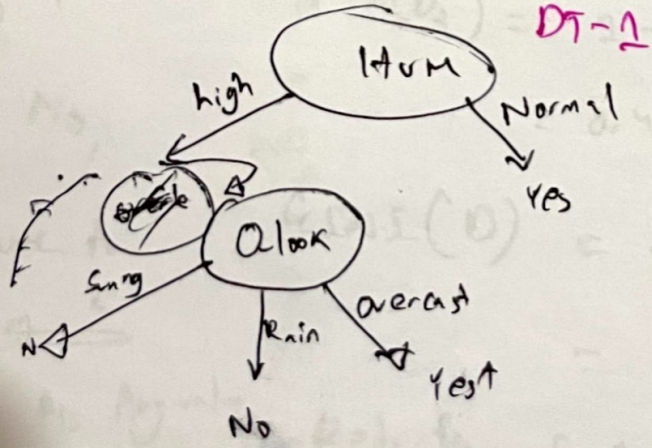


$$GINI_{Outlook}(D2) = \frac{1}{3} * 0 + \frac{1}{3} * 0 + \frac{1}{3} * 0$$

= 0

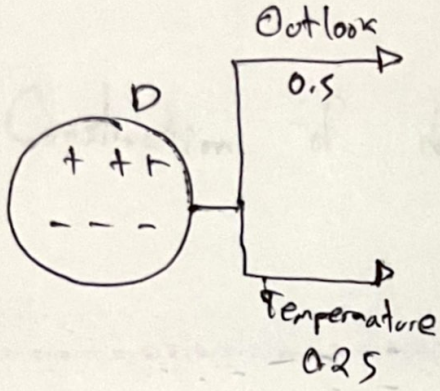
Reduction outlook = 0.44 − 0 = 0.44

DT-1



Bootstrapped - 3
(H.W.)

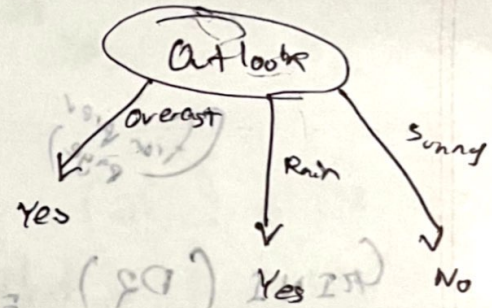
BOOTSTRAP DATASET -2



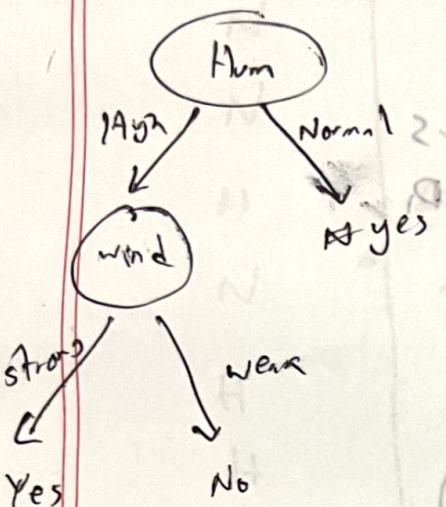
$Reduction_{Outlook}(D) = 0.5 - 0 = 0.5$

$Reduction_{Temp}(D) = 0.25$

DT-2



DT-3

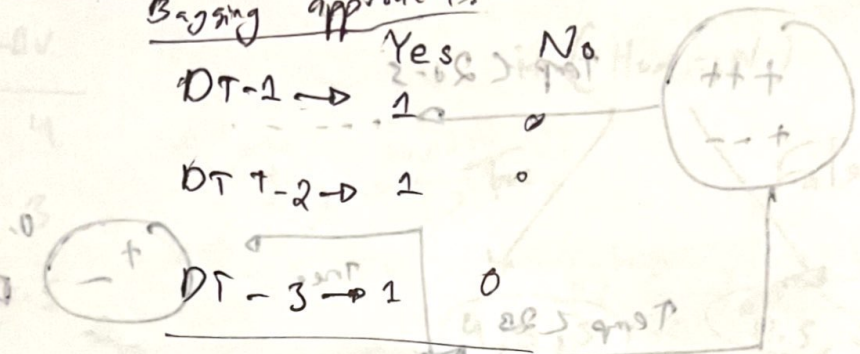


Bagging approach:-

	Yes	No
DT-1 →	1	0
DT-2 →	1	0
DT-3 →	1	0
	3	0

yes will wins

So, final decision → yes



first

sort

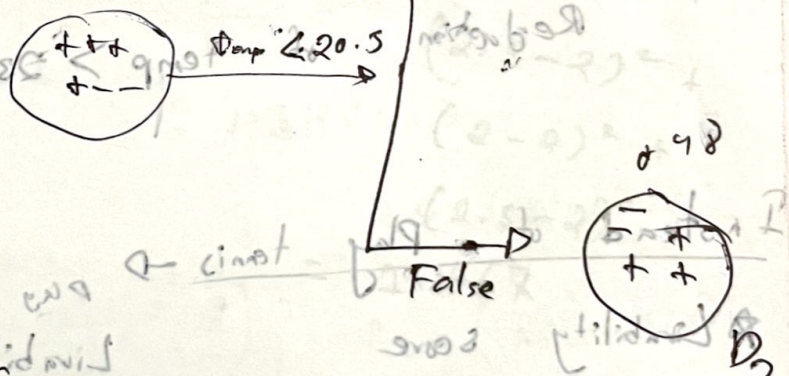
Temp	Avg	PT
20	20.5	Y
21		
25	23	N
27	26	Y
41	34	Y
42	41.5	Y

No

we have to

~~this Avg~~ as

Use this Avg value as a splitter



$$\text{GINI}(D_2) = 1 - \left\{\left(\frac{3}{5}\right)^2 + \left(\frac{2}{5}\right)^2\right\}$$

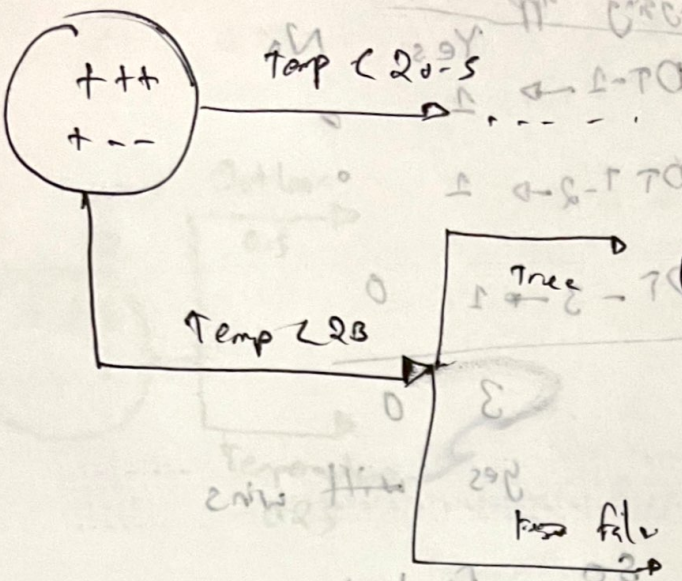
$$= 0.48$$

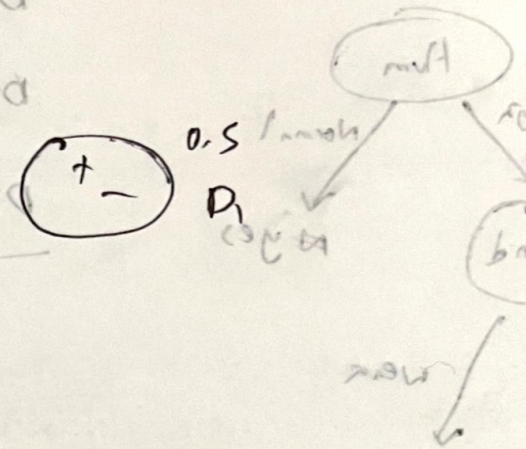
$$\text{GINI}(D) = \frac{1}{6}(0) + \frac{5}{6} * 0.48$$

$$= 0.4$$

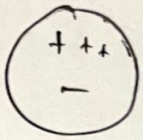
Reduction of Temp < 20.5 = 0.44 − 0.4

= 0.04





0.5 D_1



D_2

0.37

$GINI(D_2) = 0.37$

Gini Impurity $GINI(D) = \frac{2}{6} \times 0.5 + \frac{4}{6} \times 0.37$

$= 0.162 + 0.2468 = 0.4288$

Reduction of temp > 28 = 0.4287

Instead of Play tennis → Play tennis

Livability score

Livability → regression problem

4

3

2.5

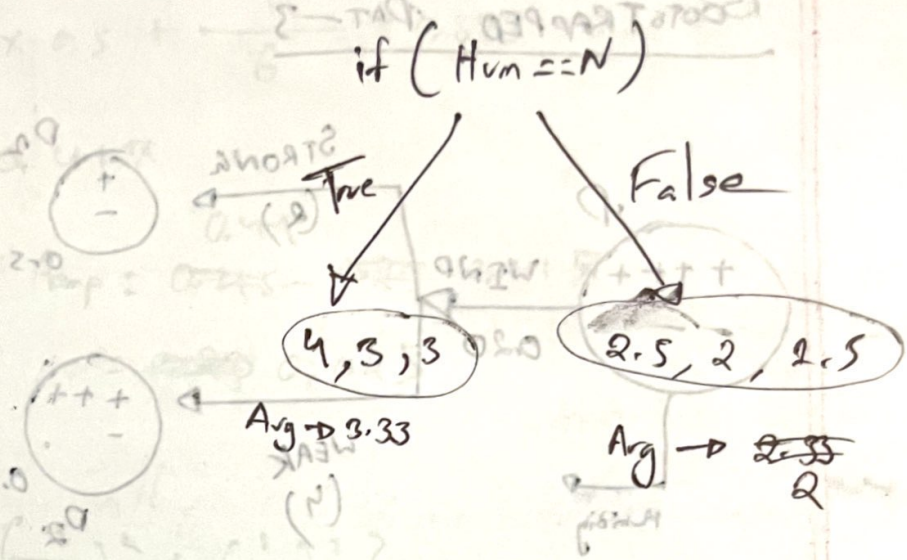
2.5

3

2

1.5

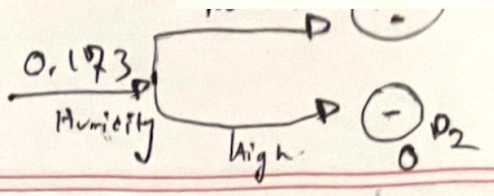
Hum	Lav
N	4
N	3
H	2.5
N	3
H	2
H	1.5



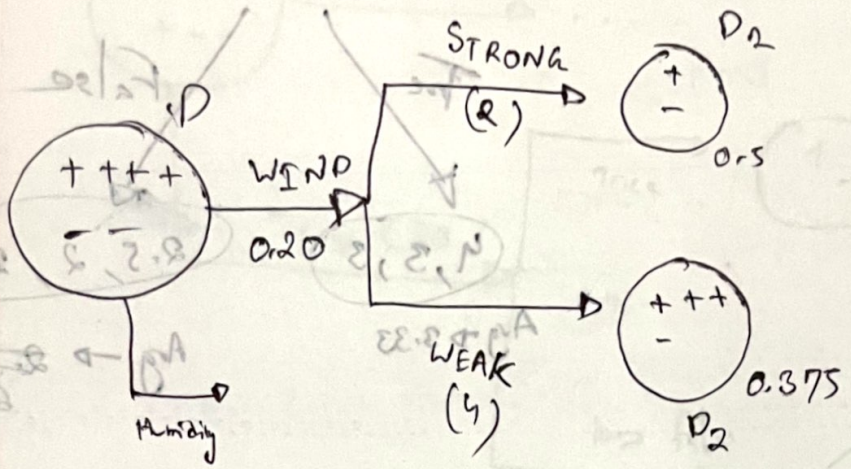
$$SSR_{(Hum == N)} = (4 - 3.3)^2 + (3 - 3.33)^2 + (3 - 3.33)^2 + (2.5 - 2)^2 + (2 - 2)^2 + (2.5 - 2)^2$$

(4 → Actual, 3.3 → read)

$$= 1.1667$$



BOOTSTRAPPED PAT-3



WIND	Humidity
W	H
S	H
S	H
W	H
S	H
S	H
W	H

$$GINI(D) = 1 - \left\{\left(\frac{4}{6}\right)^2 + \left(\frac{2}{6}\right)^2\right\}$$

$$= 1 - \frac{4^2+2^2}{36}$$

$$= 1 - \frac{20}{6}$$

$$= 0.444$$

$$GINI(D_1) = 0.5$$

$$GINI(D_2) = 1 - \left\{\left(\frac{3}{4}\right)^2 + \left(\frac{1}{4}\right)^2\right\}$$

$$= 1 - \frac{10}{16}$$

$$= 1 - 0.625$$

$$= 0.375$$

$$Gini_{WIND}(D) = \frac{2}{6} \times 0.5 + \frac{4}{6} \times 0.375$$

$$= 0.41\overline{6}$$

Reduction of Gini by temp $= 0.44 - 0.41\overline{6}$

$$= 0.023$$

$$Gini_{hum}(D_1) = 1 - \left\{\left(\frac{4}{5}\right)^2 + \left(\frac{1}{5}\right)^2\right\}$$

$$= 0.32$$

$$Gini_{hum}(D_2) = 0$$

$$Gini_{hum}(D) = \frac{5}{6} \times 0.32 + 0$$

$$= 0.267$$

Reduction of Gini by hum $= 0.44 - 0.267$

$$= 0.173$$

We need to visualize the data → If follows linear trend → linear algorithm curve

→ Non linear algorithm curve
If doesn't follow linear trend

SSR instead of GINI for Regression tree.

loss function

[Considering Age]

Age ≤ 9 using split criteria

T → Avg = 2 → 2

F → 3, 2, 4, 5 → Avg = 3

Residual = Original − Predicted

$= (2-2)^2 + (3-3)^2 + (2-3)^2 + (4-3)^2 + (5-3)^2$

$= 0 + 0 + 1 + 1 + 4$

$= 6$

Age < 12

T → 2, 3

F → 4, 7, 5, 2 → Avg = 3

$SSR = (2-2)^2 + (3-2)^2 + (2-3)^2 + (7-3)^2 + (5-3)^2$

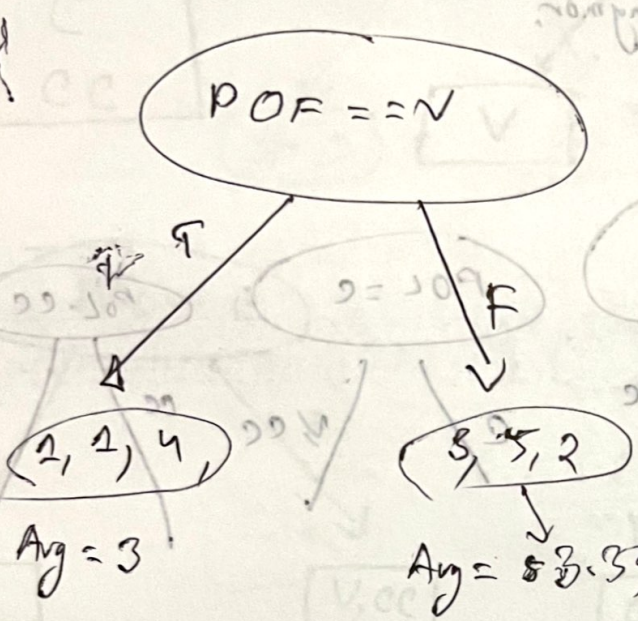
$= 0 + 1 + 1 + 16 + 4$

$= 22$

Age — 2 provides the lowest SSR.

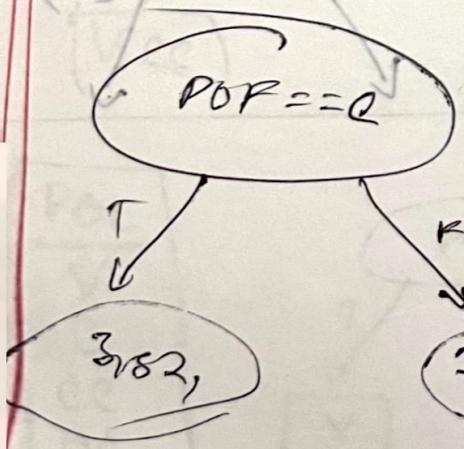
~~First we~~

, First we have to split



$SSR_{village} = (2-3)^2 + (1-2)^2 + (4-2)^2 + (5-3.3)^2 + (5-3.3)^2 + (2-3.33)^2$

$= 10.68$



If we get something as mirror we don't calculate SSR as they are same

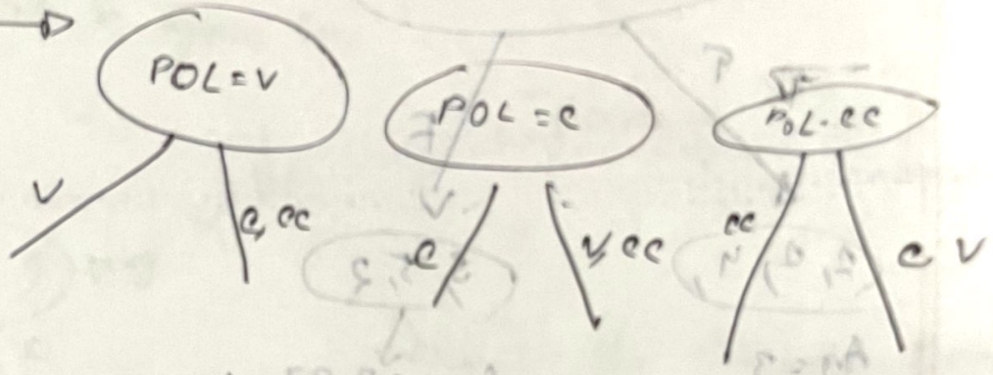
$SSR_{age} = 7.933$

$SSR_{village} = 10.68$

If there is only 1 instances we don't need to expand the tree anymore.

POL → Place of living.

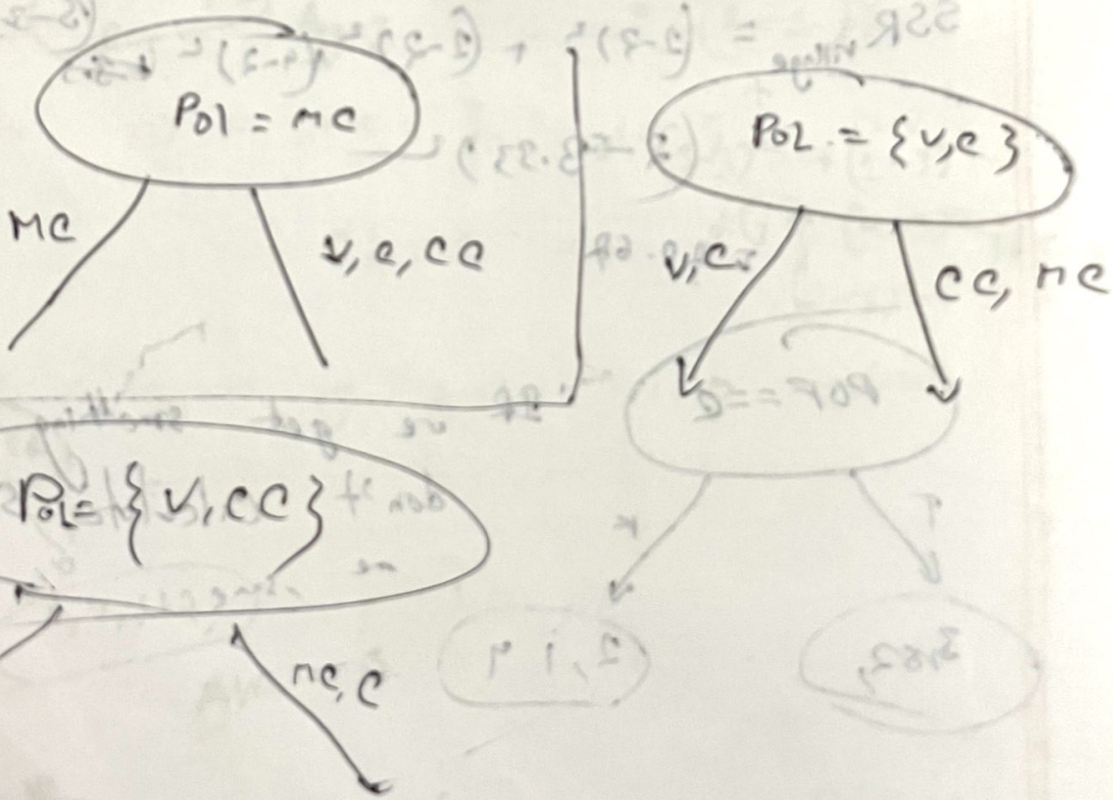
POL
V
C
CC



POL

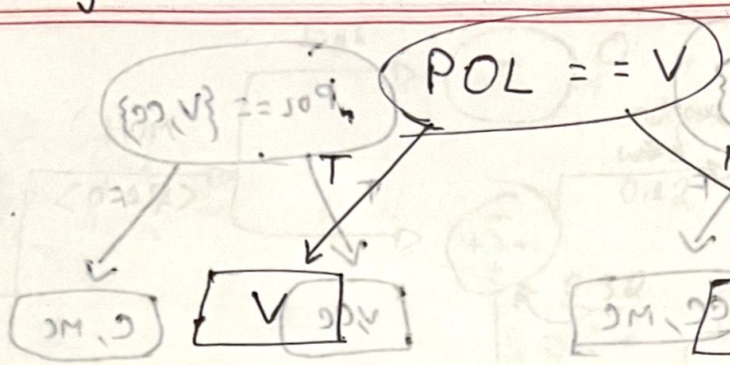
no. of splitter $= 2^{n-1} + 1$

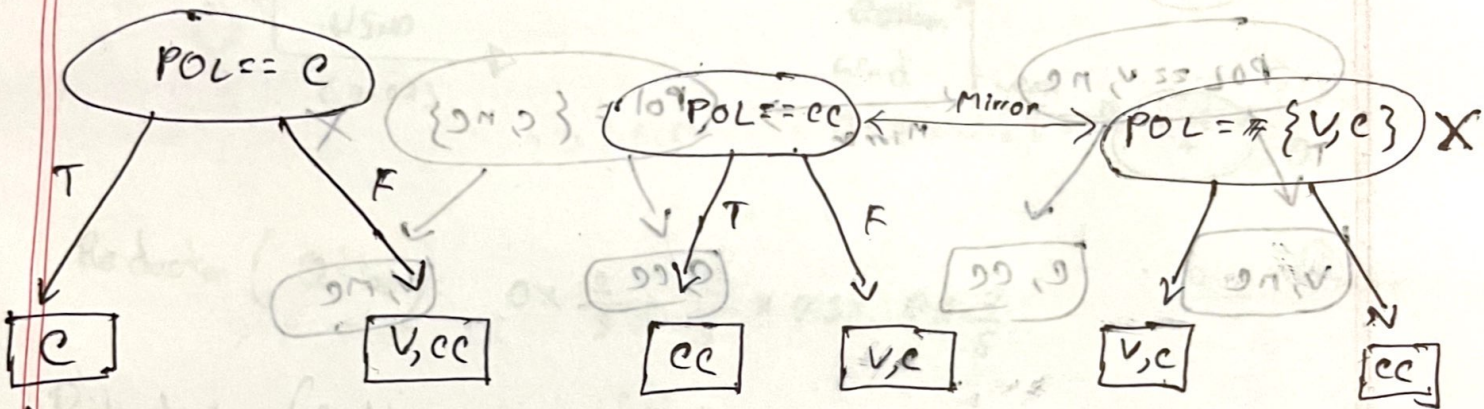
V
C
CC
MC

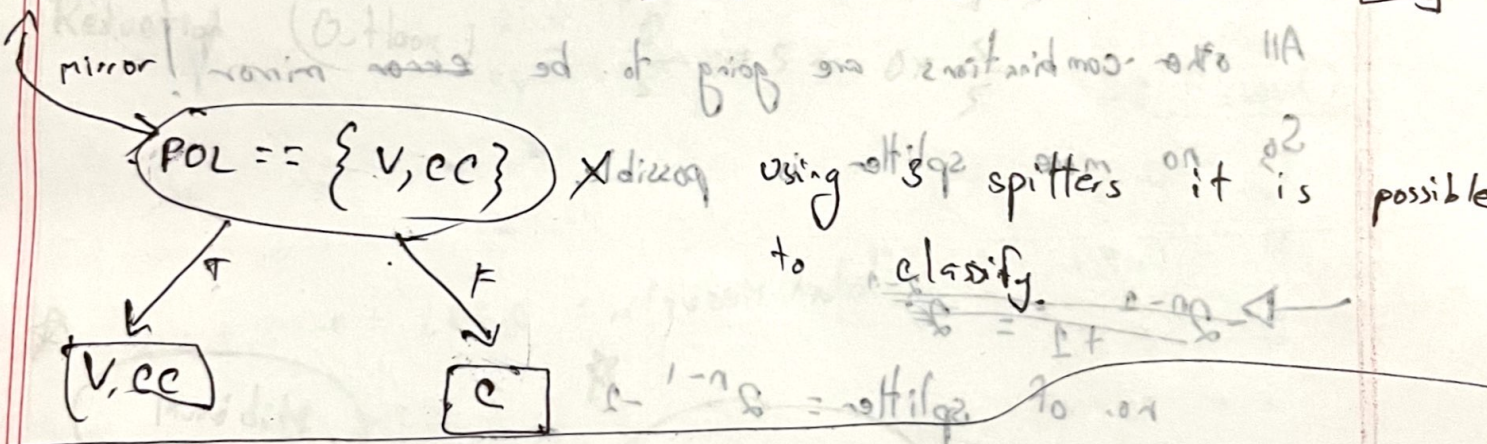


Place of Living

POL
V
C
CC

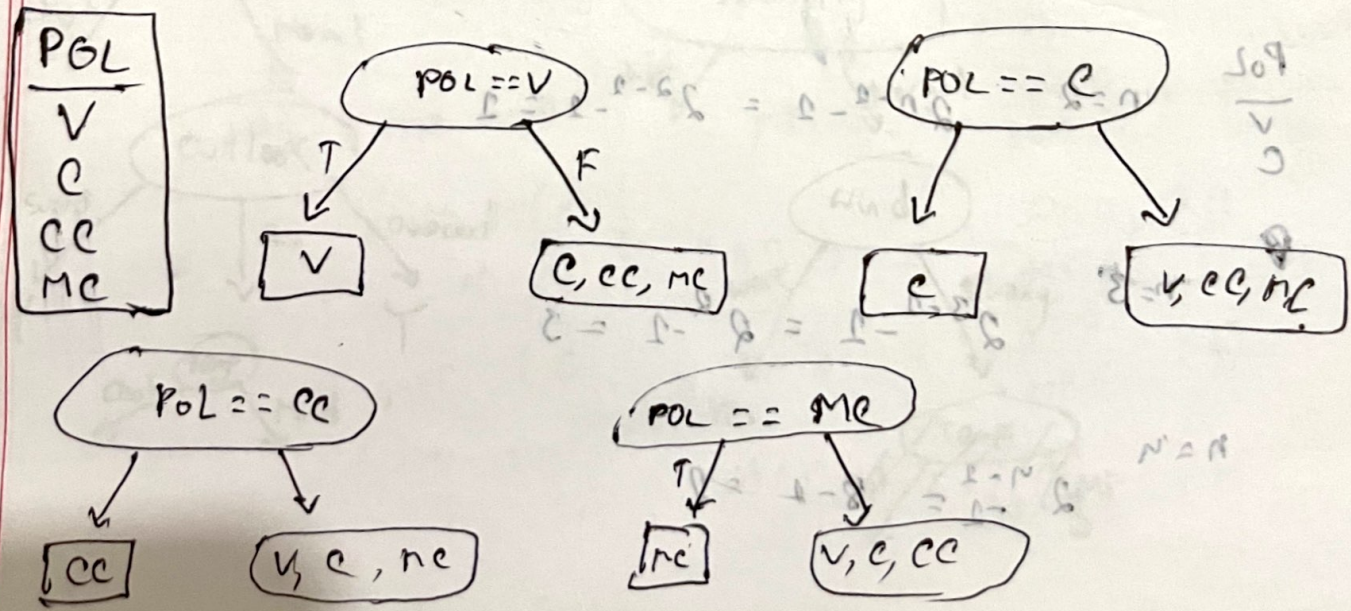


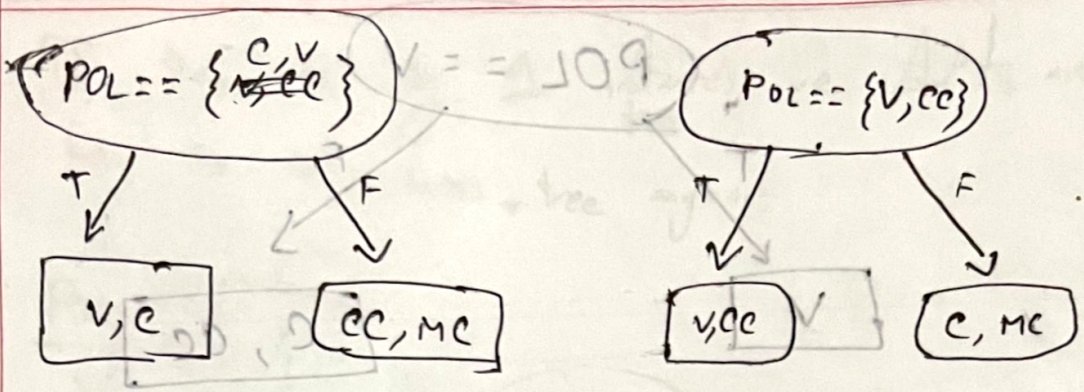




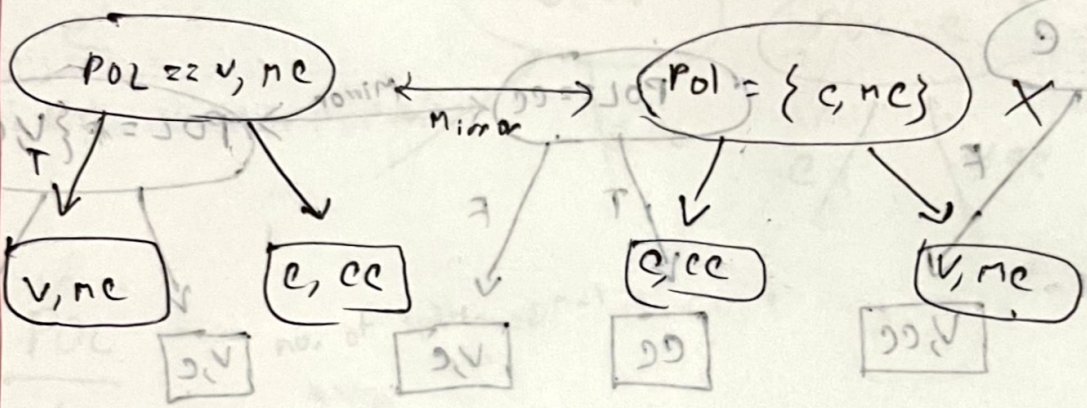
using splitters it is possible to classify.

POL
V
C
CC
MC





POL
V
C
CC



All other combinations are going to be ~~error~~ mirror !

So, no more splitter is possible.

$2^{n-1} + 1 = 2^{2-1}$

no. of splitter = $2^{n-1} - 1$

POL
V
C
CC
MC

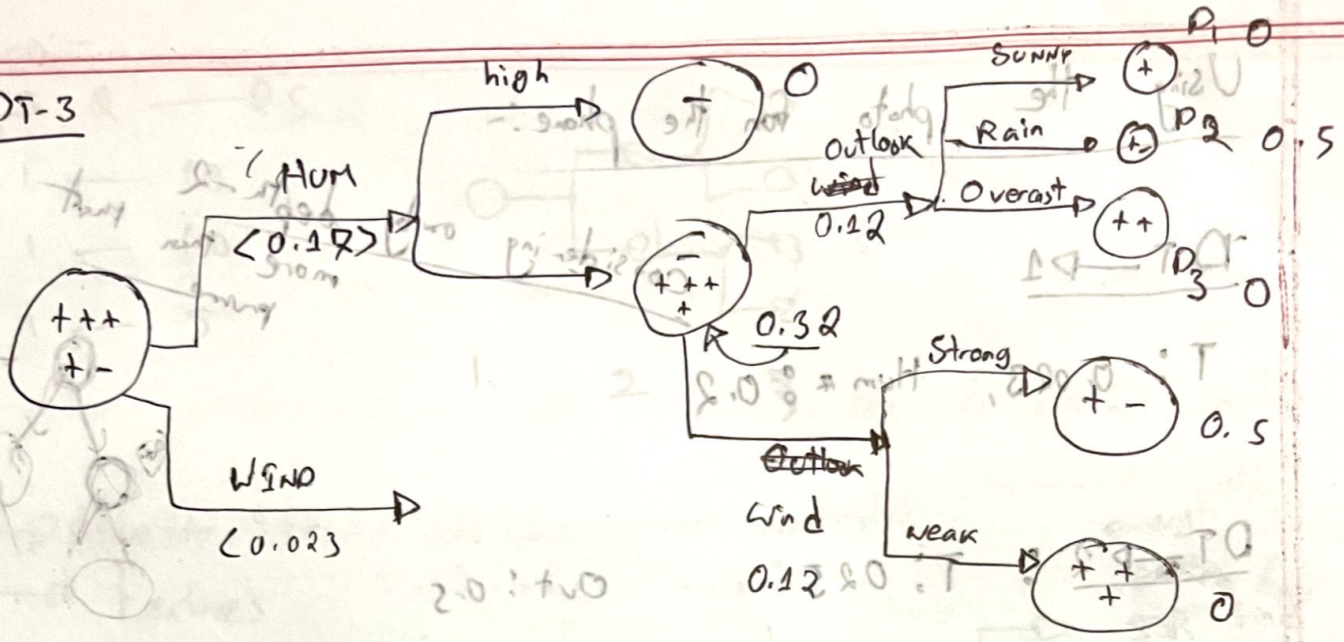
POL
V
C

$n = 2$ $\quad 2^{n-1} - 1 = 2^{2-1} - 1 = 1$

$n = 3$ $\quad 2^{3-1} - 1 = 2^2 - 1 = 3$

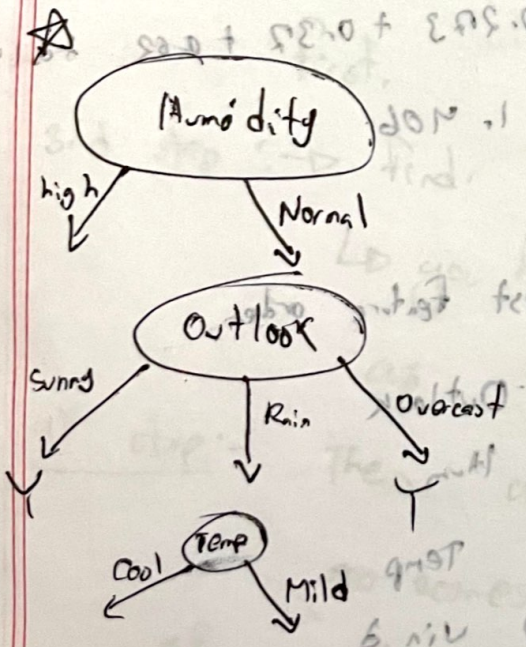
$n = 4$ $\quad 2^{4-1} - 1 = 8 - 1 = 7$

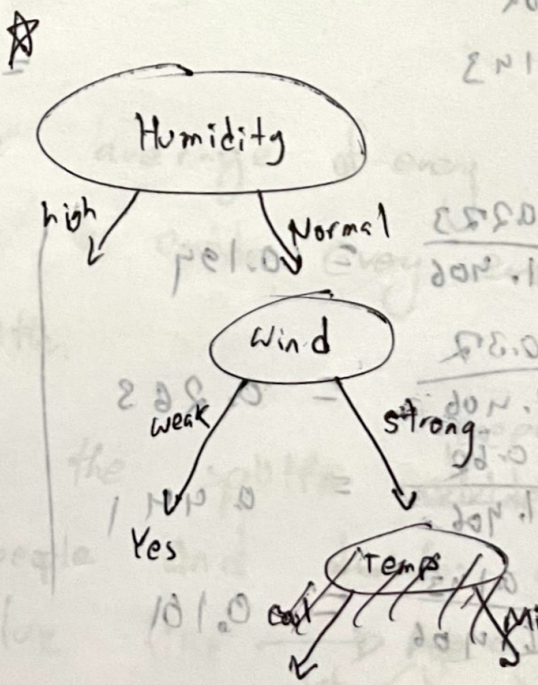
DT-3



Reduction (Outlook) = $0 \times \frac{1}{5} + \frac{2}{5} \times 0.5 + 0 \times \frac{2}{5}$

Reduction (Outlook) = $\frac{2}{5} \times 0.5 + 0 \times \frac{3}{5}$

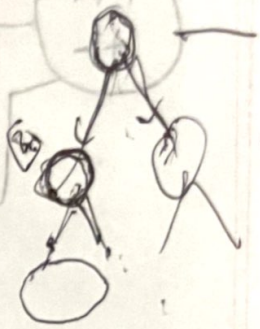




Using the photo for the phone :-

DT → D1

considering only depth-2 tree we will get more than that prune



T : 0.023, Hum : 0.2

DT → D2 : T : 025 Out : 0.5

DT → D3 :

Hum : 0.17 × Wind : 0.023 ~~Out : 0.12~~

Out : 0.12

Wind : 0.12

Overall :-

T = 0.273

H = 0.37

O = 0.62

W = 0.143

Total Reduction = 0.273 + 0.37 + 0.62 + 0.143

= 1.406

Now,

T : $\frac{0.273}{1.406}$ = 0.194

H : $\frac{0.37}{1.406}$ = 0.263

O : $\frac{0.62}{1.406}$ = 0.441

W : $\frac{0.143}{1.406}$ = 0.101

Best feature order

1) Outlook

2) Hum

3) Temp

4) Wind

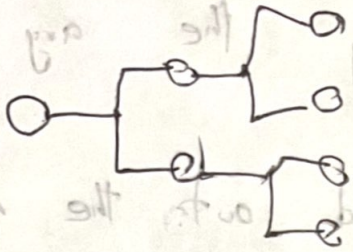
MSD:-

5 Q — 25

1 — 10

1 — 10

1 — 5



1 2 3

→ Regression tree

→ features

↓ type of features → categorical → every splitter find SSR

→ numerical → every splitter find SSR

REGRESSION TREE:-

1st step → identify the column type features low

2nd step → for continuous values (numerical) sort the numbers first.

3rd step :→ find the average of every two consecutive data

→ you have to consider every average as a splitter.

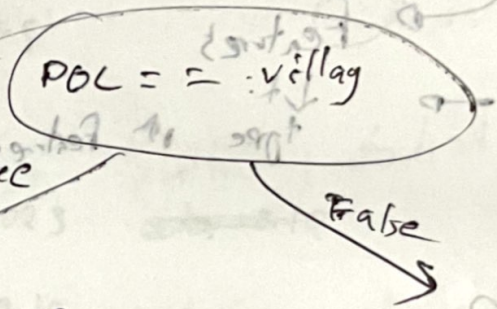
4th step:- Then using the splitter seperate the IQ scores people and find the average of the value like → Age < 9 → left → avg → right → avg

step 5 → using the avg find the SSR value.

step-6 → Find out the minimum SSR value (which has low error)

For Categorical data:-

step-1 → selecting any one of sector apply True or False every data ex-

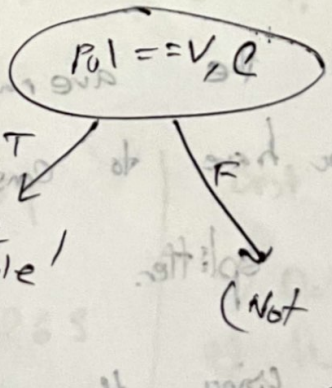


1+1+1=2

3+5+2 = 3.33

Repeat → find avg of IG, classify, find the SSR value

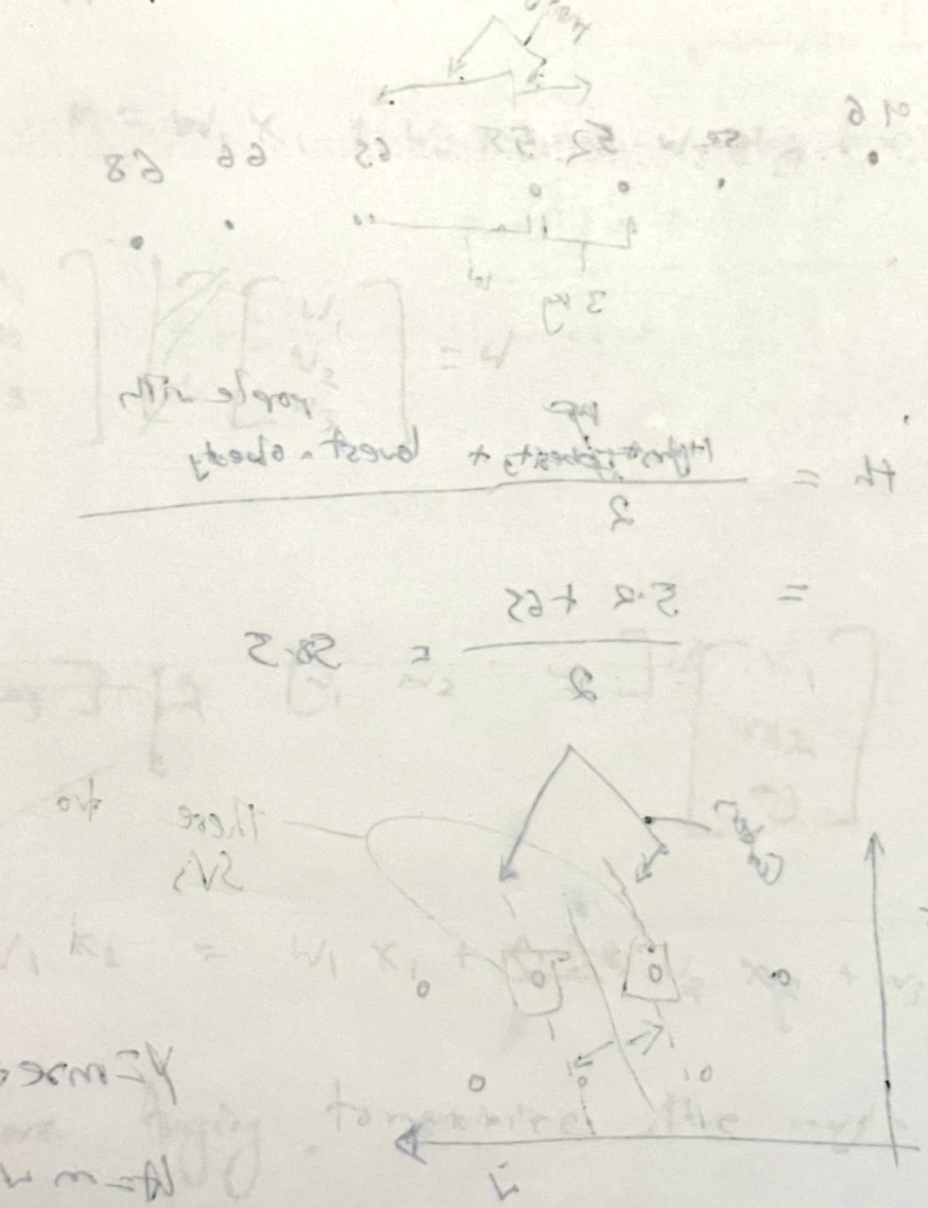
exception :- If there are more than 4 categorical value



Columns final

Step-7: Comparing the two of SSR values the one which has less will become the splitter

Step-8: You have to expand the regression tree up to expansion threshold.



INFERENTIAL STATISTICS

We want to infer something ~~from~~ from the data.

We " " prove a hypothes.

★ It is random sample of data taken from population to describe and make inferences ~~about~~ the population.

Discrete

→ Continuous probability

discrete

N, i

HYPOTHESIS TESTING :-

→ It can really be anything as long as you put it in the test

Hypothesis testing

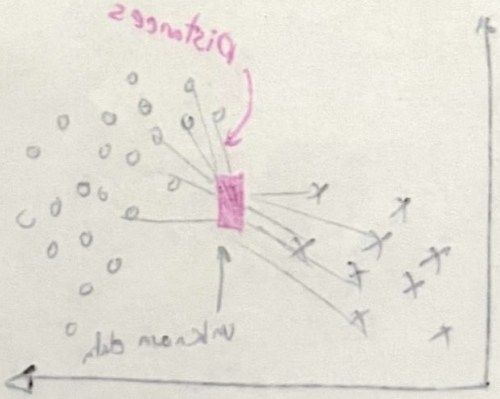
(i) One tailed test

(ii) two " "

Standard score / Z score

(i) substract the mean

(ii) calculate the Standard D.



KNN Algorithm

For large dataset

K ⟶ hyperparameter.

High space complexity	doesn't not require a training
Distance calculation time consuming	

Rules ⟶

1. Odd number
2. sqrt(n) where n = total no. of data.

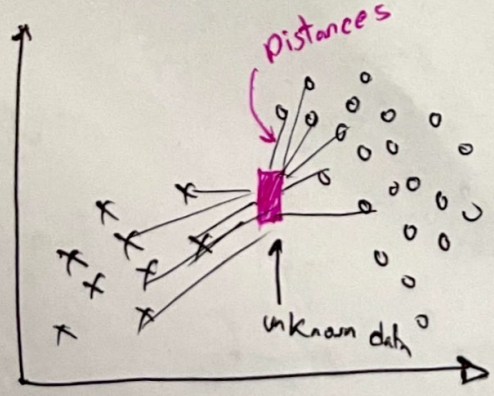
Ex:⟶ n = 100

K = $\sqrt{100}$ = 10

↳ 3, 5, 7, 9 (odd no. except 1.)

STEPS:-

1. Choose the no. of K.
2. Calculate the distance between the new points and all points in the data.
3. Sort the distance values.
4. Assign the class with most common among the K neighbours.



D1, D2, D3, D4, D5 D6

sort ⟶

D4 → O		D2
D6 → X		D3
D1 → X		D5

[K = 3]

classify ⟶

majority is 'X' so, classify ⟶

EXAMPLE:-

	X	Y	label
A.	1	2	Red
B.	2	3	Red
C.	3	3	Blue
D.	6	5	Blue
P	3	2	?

$n = 4$

$\rightarrow \sqrt{n} = \sqrt{4} = 2$

$K = 3$

$D(P, A) = \sqrt{(3-1)^2 + (2-2)^2} = 2$

$D(P, B) = \sqrt{(3-2)^2 + (2-3)^2} = \sqrt{2} = 1.42$

$D(P, C) = \sqrt{(3-3)^2 + (2-3)^2} = 1$

$D(P, D) = \sqrt{(3-6)^2 + (2-5)^2} = 9.24$

Sorting →

$D(P, C)$

$D(P, B)$

$D(P, A)$

$D(P, D)$

$K = 3$

$\hookrightarrow D(P, C) = B$

$D(P, B) = R$

$D(P, A) = R$

→ Majority is RED.

Regression.

EXAMPLE-2

	X	Y
A	1	1
B	2	2
C	3	1.3
D	4	3.75
E	5	2.25
P	3.5	

$K = \sqrt{5}$ $n = 5$

$= 2.5$

$K = 3$

$D(P,A) = |3.5 - 1| = 2.5$

$D(P,B) = |3.5 - 2| = 1.5$

$D(P,C) = |3.5 - 3| = 0.5$

$D(P,D) = |3.5 - 4| = 0.5$

$D(P,E) = |3.5 - 5| = 1.25$

Sort →

$D(P,C) = 0.5$

$D(P,D) = 0.5$

$D(P,B) = 1.15$

Mean =>

$\frac{0.5 + 0.5 + 1.5}{3} \Rightarrow \frac{1.3 + 3.75 + 2}{3}$ ← (y values)

$= \frac{2.5}{3}$

$\Rightarrow 2.35$

$=$

Example-3

	Length	width	weight	label
A	5	3	150	Apple
B	6	3.5	160	Apple
C	7	4	180	Apple
D	10	6	300	Mango
E	11	6.5	320	Mango
P	6.5	3.8	170	?

$n = 5 \Rightarrow K = \sqrt{5} = 2.5 \approx 3$

$D(P, A) = \sqrt{(6.5-5)^2 + (3.8-3)^2 + (170-150)^2}$

$D(P, B) = \sqrt{(6.5-6)^2 + (3.8-3.5)^2 + (170-160)^2}$

$D(P, C) = \sqrt{(6.5-7)^2 + (3.8-4)^2 + (170-180)^2}$

$D(P, D) = \sqrt{(6.5-10)^2 + (3.8-6)^2 + (170-300)^2}$

$D(P, E) = \sqrt{(6.5-11)^2 + (3.8-6.5)^2 + (170-300)^2}$

Unsupervised learning algorithm 23/05/25

K-Means Clustering :-

K → the Number of clusters

STEPS of ALGORITHM :-

1. Initialize the K-centroids randomly (for each clusters)
2. Assign the datapoints to the nearest centroid.
3. Recalculating centroid.

Repeat

↳ when to stop →

(i) The centroids are not changing

(ii) The clusters

(iii) If the centroids & clusters are changing continuously then run upto fixed number of iteration.

EXAMPLE:-

$c_1 \rightarrow (1,2)$; $c_2 \rightarrow (10,0)$

	x_1	x_2	K=2	
			c_1	c_2
$c_1 \rightarrow$	1	2	0	
A	1	4		
B	1	0		
C	10	2		
D	10	4		
$c_2 \rightarrow$	10	0		0

← calculate all the distances

← find out the shortest distance,

c_2 ↑ Random centroid

$$d_1(c_1, c_2) = \sqrt{(1-10)^2 + (2-0)^2} = 9.22$$

$C_1(A,B)$ | $c_2 = (C, D)$

$A = (1,4)$ | $B = (1,0)$ | $c_1 = (1,2)$

x_1	x_2
$\frac{1+1+1}{3} = 1$	$\frac{4+0+2}{3} = 2$

∴ $NC_1 = (1,2)$ ← centroid unchanged

EXAMPLE-2

Iteration - 1

$C_1 = A$ | $C_2 = C$

	x	y	C_1	C_2	selected centroids
$C_1 \rightarrow$ A	2	10	0	8.49	a_1 ← observe which one is the closest of them.
B	2	5	5	6.08	C_1
$C_2 \rightarrow$ C	8	4	8.49	0	C_2
D	5	8	3.61	5	C_1
E	7	5	7.02	4.12	1.41	C_2

$C_1 \rightarrow A$

$C_2 \rightarrow A \Rightarrow \sqrt{(2-8)^2 + (10-4)^2} = 8.49$ ← distance between C_1 & C_2

Cluster 1 :- A, B, D

Cluster 2 :- C, E

From C_2 new cluster =>

$C_1 \Rightarrow \frac{2+2+5}{3}, \frac{10+5+8}{3}$

$= 3, 7.67$

↳ then again find the distance

From C_2 new cluster =>

$C_2 \Rightarrow \frac{8+7}{2}, \frac{4,5}{2}$

$= 7.5 : 4.5, 4.5$

New points

$C_1 \rightarrow (3, 7.67)$ $C_2 \rightarrow (7.5, 4.5)$

For iteration -2

	x	y	c1	c2	selected cluster
A	2	10	2.45	9.28	c_1
B	2	5	2.83	5.5	c_1
C	8	4	6.01	0.71	c_2
D	5	8	2.04	4.5	c_1
E	2	5	4.72	0.21	c_2
F	3	7	0.67	5.5	c_1

New point

$C_1 = (3, 7.67)$ $C_2 = (7.5, 4.5)$

Cluster - 1 → A, B, D

Cluster - 2 → C, E

Clusters are same as before

Stop

New feature → $\sqrt{(x_2-x_1)^2 + (y_2-y_1)^2 + \cdots (z_2-z_1)^2}$

For F point →

$d(F, c_1) \rightarrow \sqrt{(3-3)^2 + (7.67-5)^2}$

$d(F, c_2) \rightarrow 5.5$

SUPPORT VECTOR MACHINE (Machine learning Model)

Concept :- SVM aims to find the optimal hyperplane that best seperates data into different classes.

This hyperplane is a line [In two dimensions]

This hyperplane " " plane [In 3 "]

COMMON TERMS NEEDED :-

(i) Margin :- It is the distance between the hyperplane and the ~~clas~~ closest data points from both the classes

margin depends only on the closest points

These closest points are called

(ii) support vectors. → The data points that lie closest to the hyperplane.

hyperplane is a decision boundary that SVM uses to seperate two classes

~~Place the~~

MATHEMATICAL TERMS

Equation of hyperplane :-

$$H \Rightarrow w^T(x) + b = 0$$

Distance from the hyperplane :-

$$d = \frac{|a(x_0) + b(y_0) + c|}{\sqrt{a^2 + b^2}}$$

$$d_H(x_0) = \frac{w^T(x_0) + b}{\|w\|_2}$$

Place the hyperplane is such a position that from the positive samples and negative samples will be in the farthest possible distances. In other words, we have to maximise the margin.

Why maximise margin is necessary?

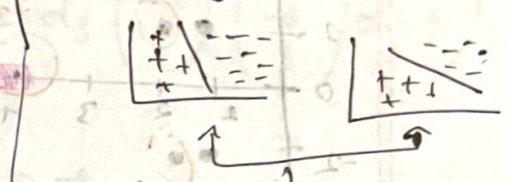
→ If the point is on more distant then we will be more confident to say ~~that~~ their class labels.

$||w||_2$ is the euclidean norm.

→ $\sqrt{w_1^2 + w_2^2 + \dots w_n^2}$

GOAL is to maximize the minimum distance

we need to find the slope that can maximize the margin



different slope

If we insert a point $p(x, y)$ on a hyperplane equation and we get positive value. then the point is from the positive group.

$$w^T(x) + b > 0$$

And, if we get negative value then the point from the negative group.

$$w^T(x) + b < 0$$

VERIFICATION:

$y_n [w^T \phi(x) + b]$

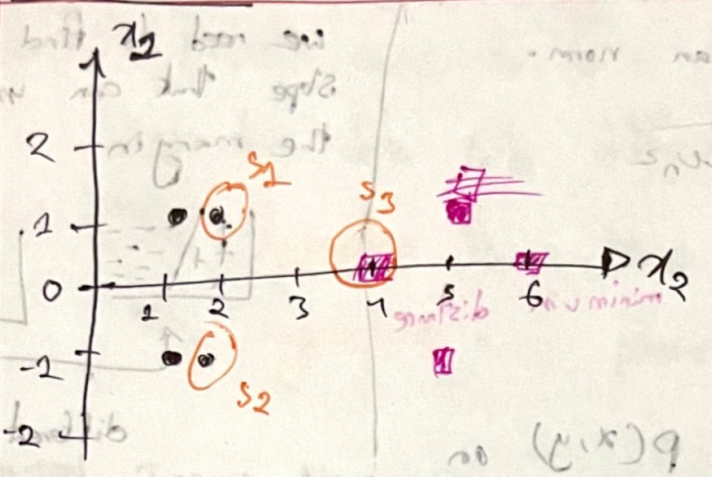
↑ original — ↓ predicted

+1	×	+1	= +1	→ Correct
-1	×	-1	= +1	
+1	×	-1	= -1	→ Incorrect
-1	×	+1	= -1	

The no. of support vectors we need to try —

→ $d+1$ (d = dimension of data)

EXAMPLE :-



step-1

Dimension is $x_1, x_2 = 2$

Required support vectors = 2+1 = 3

$S_1 (2, 1)$; $S_2 (2, -1)$

$S_3 (4, 0)$

step2 :- Augment the support vectors

$S_1 = (2, 1, 1)$, $S_2 = (2, -1, 1)$, $S_3 = (4, 0, 1)$

Step-3 : find the 3 parameters α_1, α_2 & α_3

$$\alpha_1 S_1 \cdot S_1 + \alpha_2 S_2 \cdot S_1 + \alpha_3 S_3 \cdot S_1 = -1 \quad \text{(i)}$$

$$\alpha_1 S_2 \cdot S_2 + \alpha_2 S_2 \cdot S_2 + \alpha_3 S_3 \cdot S_2 = -1 \quad \text{(ii)}$$

$$\alpha_1 S_1 \cdot S_3 + \alpha_2 S_2 \cdot S_3 + \alpha_3 S_3 \cdot S_3 = +1 \quad \text{(iii)}$$

From (i) Substituting S_1, S_2 & S_3

$$\alpha_1 \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix} \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix} + \alpha_2 \begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix} \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix} + \alpha_3 \begin{pmatrix} 4 \\ 0 \\ 1 \end{pmatrix} \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix} = -1$$

$$\alpha_2 \times 6 + \alpha_2 \times 4 + \alpha_3 \times 9 = -1 \quad \text{(iv)}$$

From (ii) we get,

$$\alpha_1 \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix}\begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix} + \alpha_2 \begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix}\begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix} + \alpha_3 \begin{pmatrix} 4 \\ 0 \\ 1 \end{pmatrix}\begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix} = -1$$

$$\alpha_1 \times 4 + \alpha_2 \times 6 + \alpha_3 \times 9 = -1 \quad \text{---(iv)}$$

From (iii) we get,

$$\alpha_1 \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix}\begin{pmatrix} 4 \\ 0 \\ 1 \end{pmatrix} + \alpha_2 \begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix}\begin{pmatrix} 4 \\ 0 \\ 1 \end{pmatrix} + \alpha_3 \begin{pmatrix} 4 \\ 0 \\ 1 \end{pmatrix}\begin{pmatrix} 4 \\ 0 \\ 1 \end{pmatrix} = 1$$

$$\alpha_1 \times 9 + \alpha_2 9 + \alpha_3 17 = 1 \quad \text{---(v)}$$

Step-4

$\alpha_1 = \alpha_2 = -3.25$ & $\alpha_3 = 3.5$

find α_1, α_2, α_3 using the eqn (iv), (v), (vi)

Step-5

$$w = \sum_i \alpha_i s_i$$

$$w = \alpha_1 \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix} + \alpha_2 \begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix} + \alpha_3 \begin{pmatrix} 4 \\ 0 \\ 1 \end{pmatrix}$$

$$w = -3.25 \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix} + \begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix}(-3.23) + (3.5)\begin{pmatrix} 4 \\ 0 \\ 1 \end{pmatrix}$$

$$w^* = \begin{pmatrix} 2 \\ 0 \\ -3 \end{pmatrix} \rightarrow w^*$$

b

$w^* = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ $b = -3$

$w^T = [1 \; 0]$, $b = 3$

The equation is

$$y = [1 \; 0] \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - 3$$

(equation of hyperplane)

Testing hyperplane

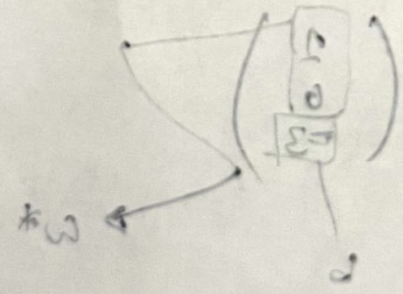
Test point :- $\begin{bmatrix} 1 \\ 2 \end{bmatrix}$

Here,

$$[1 \; 0] \begin{bmatrix} 1 \\ 2 \end{bmatrix} - 3$$

$$= 1 - 3$$

$$= -2$$ (negative class)



"INFERENTIAL STATISTICS"

→ Inferential statistics allows you to make predictions/inferences from chart or graph. It allows You take data from samples and make generalizations about a population.

→ Inferential statistics are concerned with making inferences based on the relation we found from the sample data.

ONE METHOD OR TESTS : Z score / standard score

$$Z\text{ score} = \frac{X - \text{mean}(\mu)}{\text{Standard deviation}(\sigma)}$$

value

highlow / 2

→ Sample standard deviation s:

$$\sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}}$$

68% → 1 standard deviation of the mean

95% → 2 standard deviation of the mean

99.7% → 3 standard deviation of the mean

EXAMPLE →

sample data: 26, 33, 65, 28, 34, 55, 25, ...

calculate z score!

Given

mean = 38.8

standard deviation, σ = 11.4

Value	calc	z-score
26	$\frac{26-38.8}{11.4}$	-1.12
33	$\frac{33-38.8}{11.4}$	-0.51
65	$\frac{65-38.8}{11.4}$	2.29
28	$\frac{28-38.8}{11.4}$	-0.947
34	$\frac{34-38.8}{11.4}$	-0.42

how much deviated from the mean.

HYPOTHESIS TESTING

STEPS: Hp Hypothesis

NULL HYPOTHESIS (H_0)

Propose a base line statement

STATEMENT which states that there is no relationship between the variables

ALTERNATE HYPOTHESIS (H_A)

STATEMENT THAT STATES THERE IS RELATIONSHIP BETWEEN THE VARIABLES

HYPOTHESIS TESTING

Parametric

(i) When the population is Normally distributed

(ii) Where the data is quantitative

z-test / T test

NON PARAMETRIC

(i) Where the population is not Normally distributed

(ii) Where the data is qualitative.

Chi squared Test

Compute P value

$P \leq \alpha$

Reject H_0

$P > \alpha$

Accept H_0

CONFIDENCE INTERVAL VS CONFIDENCE LEVEL

US. Census bureau's use: 90% confidence level

→ If ~~bureau~~ Census bureau repeated the survey using the same techniques 90% of the time the result would fall between 35,534 to 37,315 people in poverty.

→ 35,534 to 37,315 is the confidence interval.

Confidence Interval :- $\bar{x} \pm z \frac{\sigma}{\sqrt{n}}$ ← standard deviation Population

[z-score]

z-score

For, Sample size ≥ 30 (z test)

Confidence Interval : $\bar{x} \pm t \frac{s}{\sqrt{n}}$ ← sample standard deviation of sample size < 30

(t score) (t test)

COMPUTING T-SCORE:-

~~def~~ degrees of freedom, df = $n - 1 = 10 - 1 = 9$

Significance level, α = 1 − conf level = $1 - 0.95 = 0.025$ | s = 25

COMPUTING CONFIDENCE INTERVAL:-

Standard Error, $SE = \frac{s}{\sqrt{n}}$

Upper Confidence interval = $240 + 2.262 \times \frac{25}{\sqrt{10}}$

= 257.883

using df & significance level find the t-score from t-table

$t_{0.025} = 2.262$

Lower confidence interval $= \bar{x} - t \times \frac{s}{\sqrt{n}}$

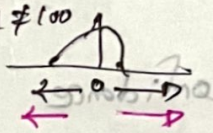
$= 240 - 2.262 \times \frac{25}{\sqrt{10}}$

$= 222.117$

Confidence Interval: 257.88 ± 222.117

Testing → two tailed test (if it can be on both side)

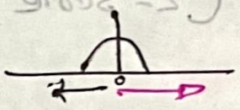
$\mu = 100 \quad \mu \neq 100$



(can be either)

→ Single tailed test

it can on one side

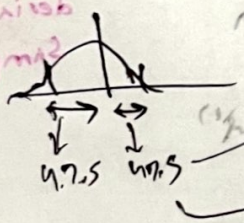


(can be any one of them.)

Computing z-score from table :-

$\mu > 100$

$\mu < 100$



How? → 50% − α = 47.5

47.5 47.5 → 0.4975 → find this value in z table

the odd row & column value

1.9 + 0.06 → 1.96.

Z test statistics :- $\frac{\bar{x} - \mu}{\sigma}$

$$Z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

$\bar{x}$ ← mean which will be null hypothesis

μ_0

σ ← standard deviation

n

$n = \frac{z^2 p(1-p)}{e^2}$

Chi Square Test

chi squared statistics

FORMULA :-

$$\chi^2 = \sum \frac{(\text{Number of Observed } (O) - \text{Expected } (E))^2}{\text{Expected}(E)}$$

$$= \sum \frac{(O-E)^2}{E}$$

→ use this value to find a value from the table.

→ then we get P value

CONDITIONS

$P > .1$ not significant [null hypothesis can't be rejected]

$P \leq .1$ marginally "

$P \leq 0.5$ significant

$P \leq .01$ highly ~~sig~~ significant.

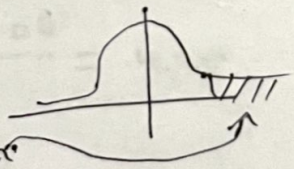
If this value is not present on the table then find the range!

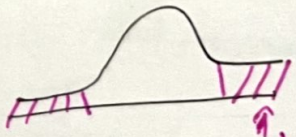
Expected = $\frac{\text{Total artist}}{\text{no. of category}}$

(For 1 category) = $\frac{256}{12}$

= 21.33

For 2 category Expected = $\frac{\text{row total} \times \text{Calm total}}{\text{All total}}$





α level → significance value

If Z-score from test statistics is greater than rejection area then reject the null hypothesis.

T-test statistics → $t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$

$6 - x_1 + (x_1 - x_2)^2$

NON LINEAR SVM

STEP-1 → Transform

$$\phi\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{cases} \begin{pmatrix} 6 - x_1 + (x_1 - x_2)^2 \\ 6 - x_2 + (x_1 - x_2)^2 \end{pmatrix} & \text{if } \sqrt{x_1^2 + x_2^2} \geq 2 \\ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} & \text{else} \end{cases}$$

TRANSFORM

blue class → $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ $\begin{pmatrix} -1 \\ 1 \end{pmatrix}$ → $\sqrt{x_1^2 + x_2^2} < 2$

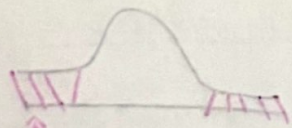
Red class → $\begin{pmatrix} 2 \\ 0 \end{pmatrix}$ $\begin{pmatrix} 0 \\ 2 \end{pmatrix}$ → for $\sqrt{x_1^2 + x_2^2} \geq 2$

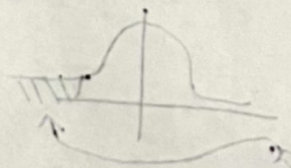
$$\phi\begin{pmatrix} 2 \\ 0 \end{pmatrix} = \begin{pmatrix} 8 \\ 10 \end{pmatrix}$$

$$\phi\begin{pmatrix} 0 \\ 2 \end{pmatrix} = \begin{pmatrix} 10 \\ 8 \end{pmatrix}$$

(transformed)

STEP-2





$$z_{test} = \frac{\bar{x} - \mu_0}{\sigma}$$