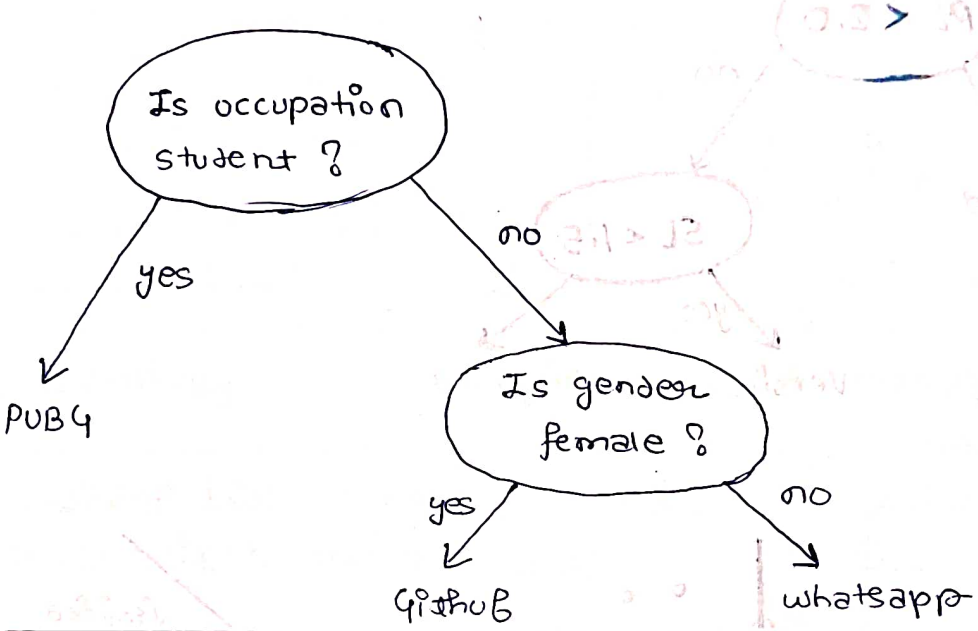
Decision Tree Intuition

Example 1)

Gender	Occupation	Suggestion
F	Student	PUBG
F	Programmer	Github
M	Programmer	Whatsapp
F	Programmer	Github
M	Student	PUBG
M	Student	PUBG

```
if occupation == student
    print(PUBG)
Else
    if gender == female
        print(Github)
    else
        print(Whatsapp)
```

Tree



Pseudo-code

- Begin with your trainning dataset, which should have some feature variables and classification or regression o/p.

determine the "Best feature" in the dataset to split the data on; (ie Root node)

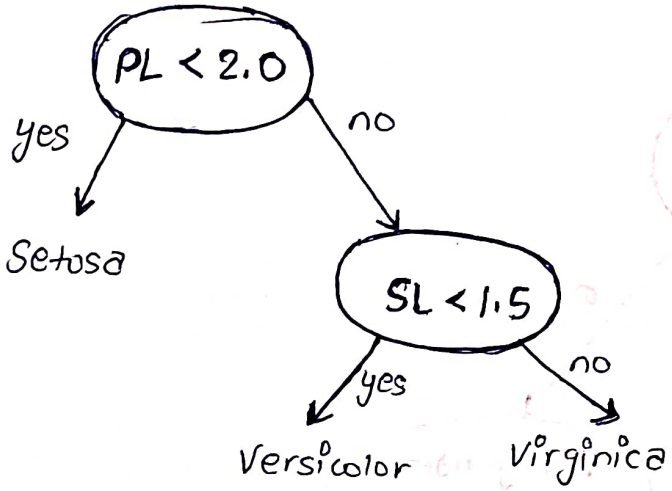
Split the data into subset that contain the correct values for the Best feature. This splitting basically defines a node on the tree (ie each node is a splitting point based on a certain feature from our data)

Recursively generate new tree nodes by using subset of data created from step 3

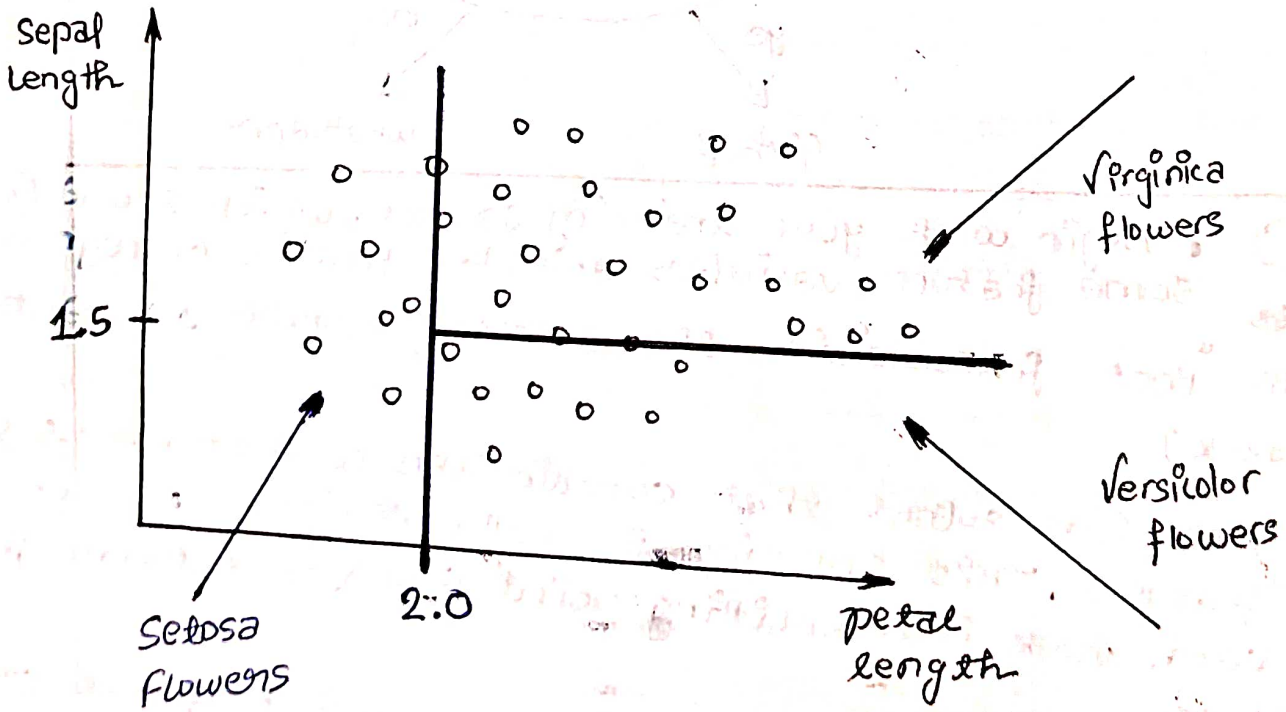
What if we have numerical data ?

Petal Length	sepal Length	size
1.34	0.34	Setosa
3.45	1.45	Versicolor
1.69	0.98	Setosa
2.56	1.79	Virginica
3.0	1.13	Versicolor
1.3	0.08	setosa

* Working well with Both either its —
(i) Categorical data or
(ii) numerical data.

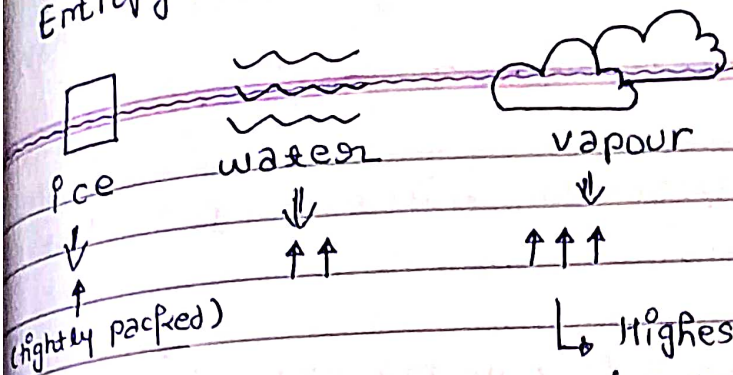


ie

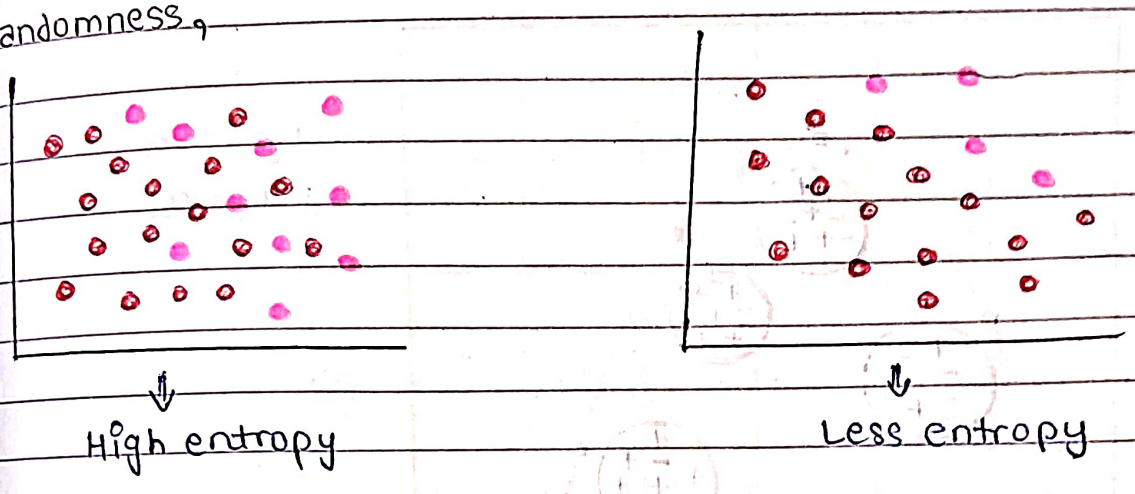


What is entropy?

Entropy is the measure of impurity / disorder



Randomness



More knowledge Less Entropy

Higher uncertainity Higher entropy

How to calculate Entropy?

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

where 'Pi' is the frequentist probability of an element / class 'i' in our data.

Eg our data has two labels Yes and No.

$$E(S) = -p_{yes} \log_2(p_{yes}) - p_{no} \log_2(p_{no})$$

for '3' labels (log) yes & No & maybe

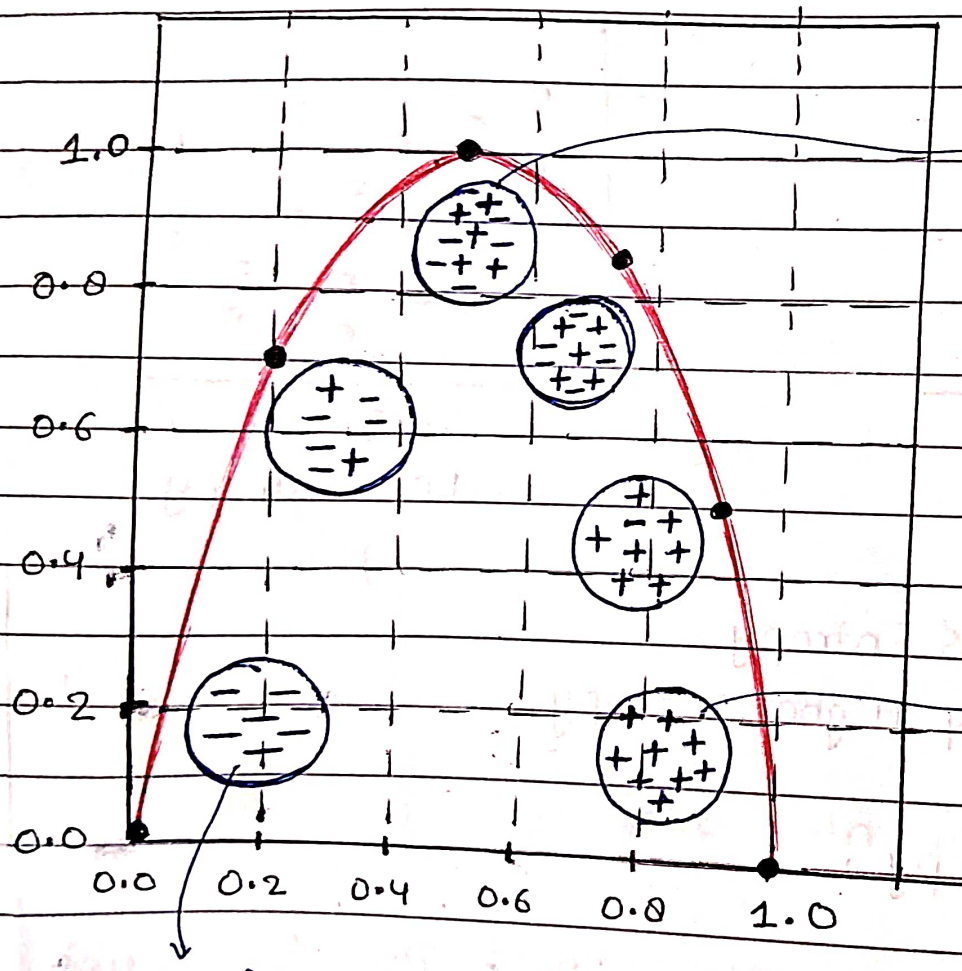
$$E(S) = -p_{yes} \log_2 p_{yes} - p_{No} \log_2 p_{No} - p_{maybe} \log_2 p_{maybe}$$

Summary
- More the uncertainty → Less Entropy.
- For a 2 class problem the min Entropy is zero (0) & max

yes → N or Yc
No → 0

yes →
No →

- For ~~Both~~ more than 2 classes the min entropy is 0 but max can be greater than 1.
- Both $\log_2$ or $\log_e$ can be used to calculate entropy.



$P_{yes} = 1/2$

$P_{No} = 1/2$

more knowledge
entropy = 0
every point is '1'
no negative '0'

more knowledge
↓
~~yes~~ entropy = 0
↓
every point is '0'
no positive '1'

Entropy v/s Probability

Entropy for continuous variables

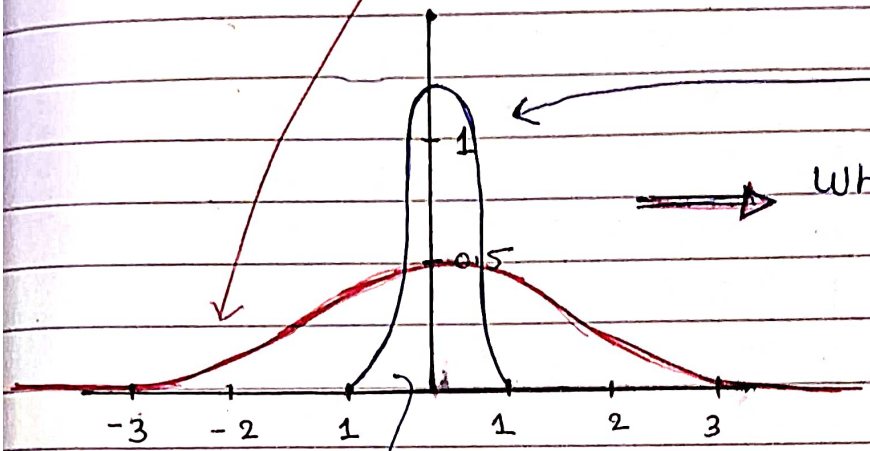
Q) Which of the below dataset have higher entropy ?

Area	Built-in	Price
1200	1999	3.5
1800	2011	5.6
1400	2000	7.3
....	....	....

Dataset 1

Area	Built-in	Price.
2200	1989	4.6
800	2018	6.5
1100	2005	12.8
....	....	....

Dataset 2



Whichever is less peaked
↓
High Entropy ...

→ prob of finding knowledge for small Area is more

Information Gain

- It's a metric used to train decision trees.

Specifically this metric measures the quality of a split.

The information gain is based on the decrease in entropy after a data-set is split on an attribute.

Constructing a decision tree is all about finding input that returns the highest information gain.

- $G = E\{parent\} - \{weighted\ Avg\} * E\{children\}$

Decision Trees!

Day	Outlook	Temp	Humidity	Wind	play
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	yes
D4	Rain	Mild	High	Weak	yes
D5	Rain	Cool	Normal	Weak	yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	yes
D10	Rain	Mild	Normal	Weak	yes
D11	Sunny	Mild	Normal	Strong	yes
D12	Overcast	Mild	High	strong	yes
D13	Overcast	Hot	Normal	Weak	yes
D14	Rain	Mild	High	Strong	No

Entropy measures homogeneity of examples,

- Entropy measures the impurity of a collection of examples It depends from the distribution of the random variable p.

→ S is a collection of trainning examples
→ $P_{\oplus}$ the proportion of positive examples in S.
→ $P_{\ominus}$ the proportion of negative examples in S.

- Entropy is the measure of disorder or impurity in a node

$$E = -\sum_{i=1}^{n} p_i \log_2 p_i$$

- node with a more variable composition considered to be more entropy than node with less variable composition

- Entropy of universal fact = 0;

High Entropy → less homogenous / more impurity
less Entropy → more homogenous / less impurity

ie, Attribute : Outlook
Values (Outlook) = Sunny, Overcast, Rain

$S = [9+, 5-]$ entropy(S) $= \frac{-9}{14} \log_2^{9/14} - \frac{5}{14} \log_2^{5/14} \Rightarrow 0.94$

$S_{sunny} \leftarrow [2+, 3-]$ entropy(S) $= \frac{-2}{5} \log_2^{2/5} - \frac{3}{5} \log_2^{3/5} \Rightarrow 0.971$
(Ssunny)

$S_{overcast} \leftarrow [4+, 0-]$ Entropy $= \frac{-4}{4} \log_2^{4/4} - \frac{0}{4} \log_2^{0/4} \Rightarrow 0$
(Sovercast)

$S_{rain} \leftarrow [3+, 2-]$ Entropy $= \frac{-3}{5} \log_2^{3/5} - \frac{2}{5} \log_2^{2/5} \Rightarrow 0.971$
(Srain)

- Information gain, measure of change in Entropy

$$Gain(S, outlook) = Entropy(S) - \sum_{v \in \{sunny, overcast, rain\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S, outlook) = Entropy(S) - \frac{5}{14} Entropy(S_{sunny}) - \frac{4}{14} Entropy(S_{overcast}) - \frac{5}{14} Entropy(S_{rain})$$

$$\Rightarrow 0.94 - \frac{5}{14}(0.971) - \frac{4}{14}(0) - \frac{5}{14}(0.971) \Rightarrow 0.7464$$

Attribute : Temp

Values(Temp) : Hot, Mild, Cold

$S = [9+, 5-]$ $\quad Entropy(S) \rightarrow -\frac{9}{14} \log_2^{9/14} - \frac{5}{14} \log_2^{5/14} \Rightarrow 0.94$

$S_{hot} \leftarrow [2+, 2-]$ $\quad Entropy(S_{hot}) \rightarrow -\frac{2}{4} \log_2^{2/4} - \frac{2}{4} \log_2^{2/4} \Rightarrow 1.00$

$S_{mild} \leftarrow [4+, 2-]$ $\quad Entropy(S_{mild}) \rightarrow -\frac{4}{6} \log_2^{4/6} - \frac{2}{6} \log_2^{2/6} \Rightarrow 0.9183$

$S_{cool} \leftarrow [3+, 1-]$ $\quad Entropy(S_{cool}) \rightarrow -\frac{3}{4} \log_2^{3/4} - \frac{1}{4} \log_2^{1/4} \Rightarrow 0.8132$

$$Gain(S, Temp) \Rightarrow Entropy(S) - \sum_{v \in (Hot, mild, cold)} \frac{|S_v|}{|S|} Entropy(S_v).$$

Gain (S, Temp) ⇒ Entropy (S) − $\frac{4}{15}$ Entropy (Shot) − $\frac{6}{14}$ Entropy(Smi
− $\frac{4}{14}$ Entropy (Scool)

⇒ 0.94 − $\frac{4}{14}$ (1.0) − $\frac{6}{14}$ (0.9183) − $\frac{4}{14}$ (0.8113) ⇒ 0.0289

Attribute : Humidity
Values (Humidity) = High, Normal

S = [9+, 5−] Entropy(S) = $\frac{-9}{14}\log_2^{9/14} - \frac{5}{14}\log_2^{5/14}$ ⇒ 0.94

S_{high} ← [3+, 4−] Entropy (Shigh) = $\frac{-3}{7}\log_2^{3/7} - \frac{4}{7}\log_2^{4/7}$ = 0.9852

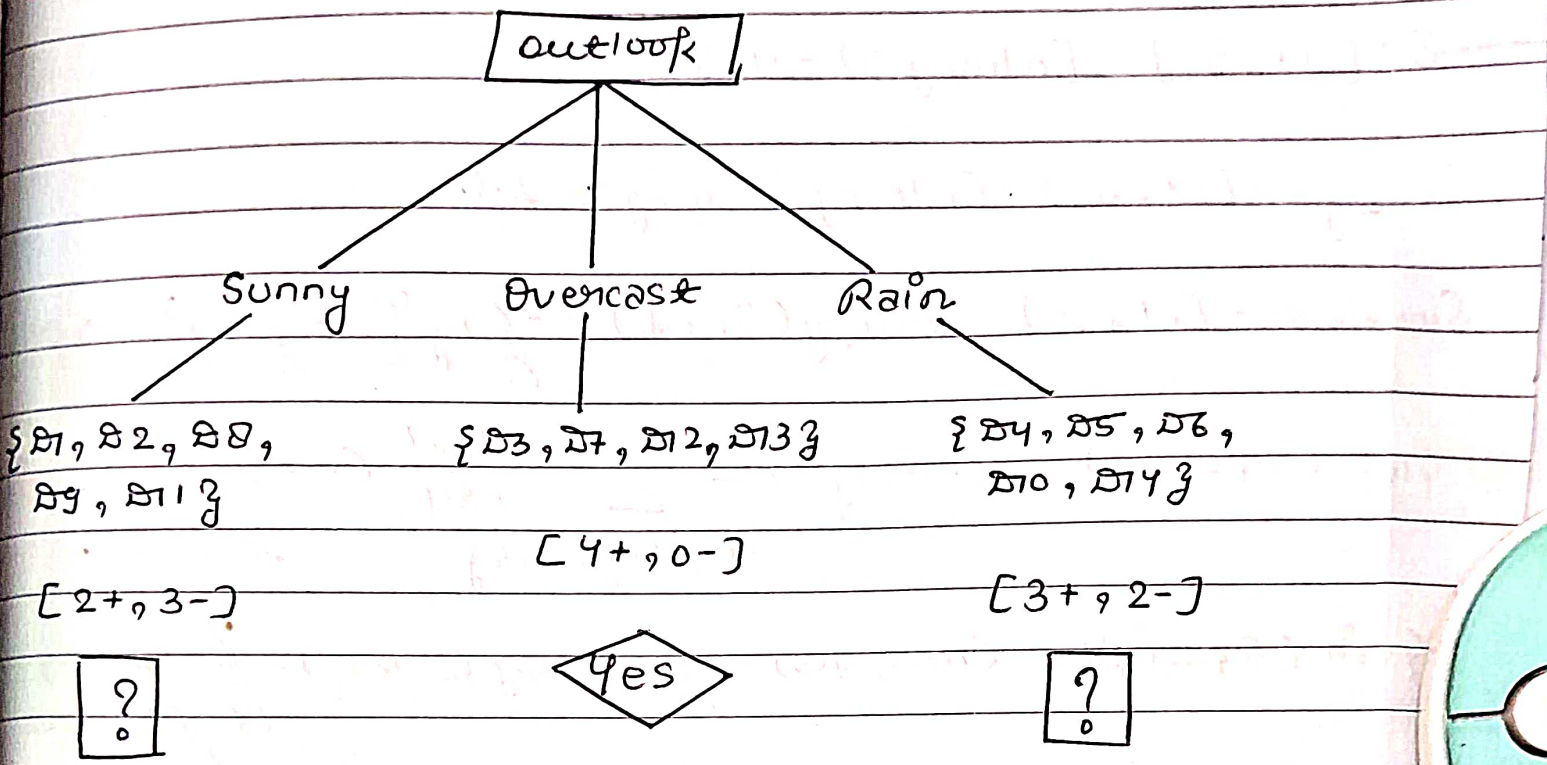
S_{normal} ← [6+, 1−] Entropy (SNormal) = $\frac{-6}{7}\log_2^{6/7} - \frac{1}{7}\log_2^{1/7}$ ⇒ 0.5916

Gain (S, Humidity) = Entropy(S) − $\sum_{v \in \{high, normal\}} \frac{|S_v|}{|S|}$ Entropy(S_v)

Gain (S, Humidity) ⇒ 0.94 − $\frac{7}{14}$ (0.9852) − $\frac{7}{14}$ (0.5916) = 0.1516

{D1, D2, ..., D14}

[9+, 5-]



Day	Temp	Humidity	Wind	Play
D1	Hot	High	Weak	No
D2	Hot	High	Strong	No
D8	Mild	High	Weak	No
D9	Cool	Normal	Weak	yes
D11	Mild	Normal	Strong	yes

Attribute : Temp

Values (Temp) = Hot, Mild, Cool

S_{sunny} = [2+, 3-] Entropy (S_{sunny}) = $-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}$ => 0.97

S_{Hot} = [0+, 2-] S_{Mild} = [1+, 1-] S_{cool} = [1+, 0-]

Entropy (S_{Hot}) ↓ Entropy (S_{Mild}) ↓ Entropy (S_{cool}) ↓

0.0 1.0 0.0

Attribute : Wind

Values (Wind) = Strong, Weak

S = [9+, 5-] Entropy(S) => 0.94

$S_{strong} \leftarrow$ [3+, 3-] Entropy (S_{strong}) = 1.0

$S_{weak} \leftarrow$ [6+, 2-] Entropy (S_{weak}) $\Rightarrow -\frac{6}{8}\log_2^{6/8} - \frac{2}{8}\log_2^{2/8} \Rightarrow$

$$Gain(S, wind) = Entropy(S) - \sum_{v \in \{strong, weak\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S, wind) = Entropy(S) - \frac{6}{14} Entropy(S_{strong}) - \frac{8}{14} Entropy(S_{weak})$$

$$Gain(S, wind) = 0.94 - \frac{6}{14}(1.0) - \frac{8}{14}(0.8113) \Rightarrow 0.0478$$

Gain (S, Outlook) → 0.2464

Gain (S, temp) → 0.0289

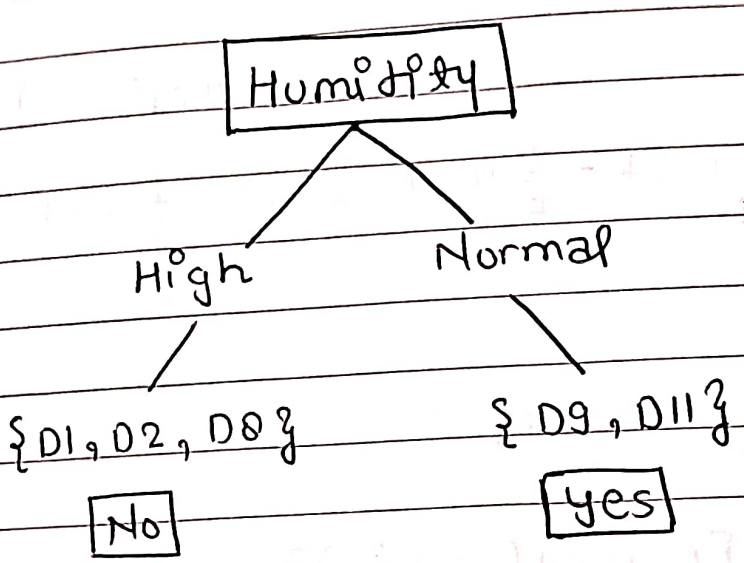
Gain (S, Humidity) → 0.1516

Gain (S, Wind) → 0.0478

Gain (Ssunny, temp) = 0.570

Gain (Ssunny, Humidity) = 0.97 ← node

Gain (Ssunny, Wind) = 0.0192



Day	Temp	Humidity	wind	play
D4	Mild	High	Weak	yes
D5	Cool	Normal	Weak	yes
D6	Cool	Normal	Strong	No
D10	Mild	Normal	Weak	yes
D14	Mild	High	Strong	No

Attribute : Humidity

Values (Humidity) = High, Normal

S_{sunny} ← [3+, 2-] Entropy ⇒ 0.97 (Ssunny)

S_{high} ← [1+, 1-] Entropy (S_{high}) ⇒ 1.0

S_{normal} ← (2+, 1-) Entropy (S_{normal}) ⇒ 0.9183

Gain (S_{rain}, humidity) ⇒ 0.0192

Gain (Ssunny, Temp) = Entropy(S) - $\sum_{V \in \{hot, cool, mild\}} \frac{|S_v|}{|S|}$ Entropy(S_v)

= Entropy(S) - $\frac{2}{5}$ Entropy(Shot) - $\frac{2}{5}$ Entropy(Smild) - $\frac{1}{5}$ Entropy(Scool)

Gain (Ssunny, Temp) => 0.97 - $\frac{2}{5}$(0.0) - $\frac{2}{5}$(1) - $\frac{1}{5}$(0.0) => 0.570

Attribute : Humidity
Values(Humidity) : High, Normal

Ssunny = [2+, 3-] Entropy(S) = 0.97

Shigh ← [0+, 3-] Entropy(Shigh) = 0.0
Snormal ← [2+, 0-] Entropy(Snormal) = 0.0

Gain (Ssunny, Humidity) => 0.97

Attribute : Wind
Values(Wind) : Strong, Weak

Ssunny = [2+, 3-] Entropy(S) = 0.97

Sstrong ← [1+, 1-] Entropy(Sstrong) = 1.0

Sweak ← [1+, 2-] Entropy(Sweak) = 0.9183

Gain(Ssunny, Wind) => 0.0192

Attribute : Wind

Values (Wind) : strong, weak

$S_{Rain} \leftarrow [3+, 2-]$ Entropy (S) = 0.97

$S_{strong} \leftarrow [0+, 2-]$ Entropy (S_{strong}) = 0.0

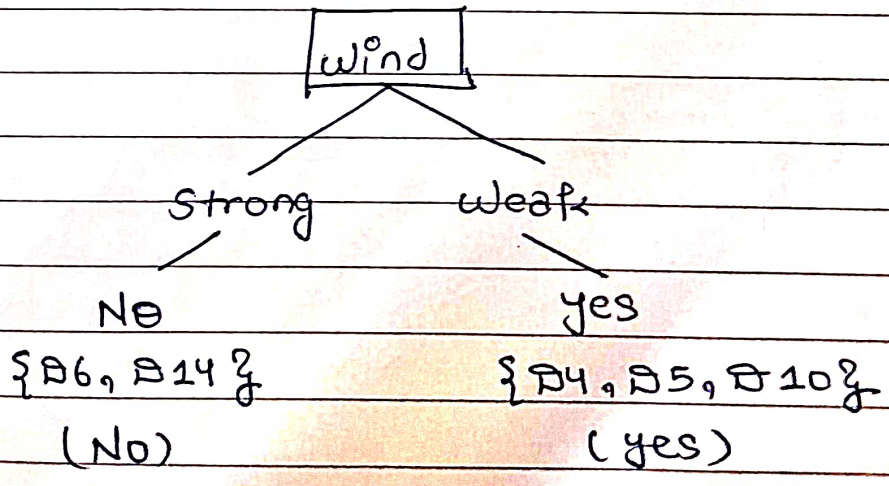
$S_{weak} \leftarrow [3+, 0-]$ Entropy (S_{weak}) = 0.0

Gain (S_{Rain}, Wind) = 0.97

Gain (S_{Rain}, Temp) = 0.0192

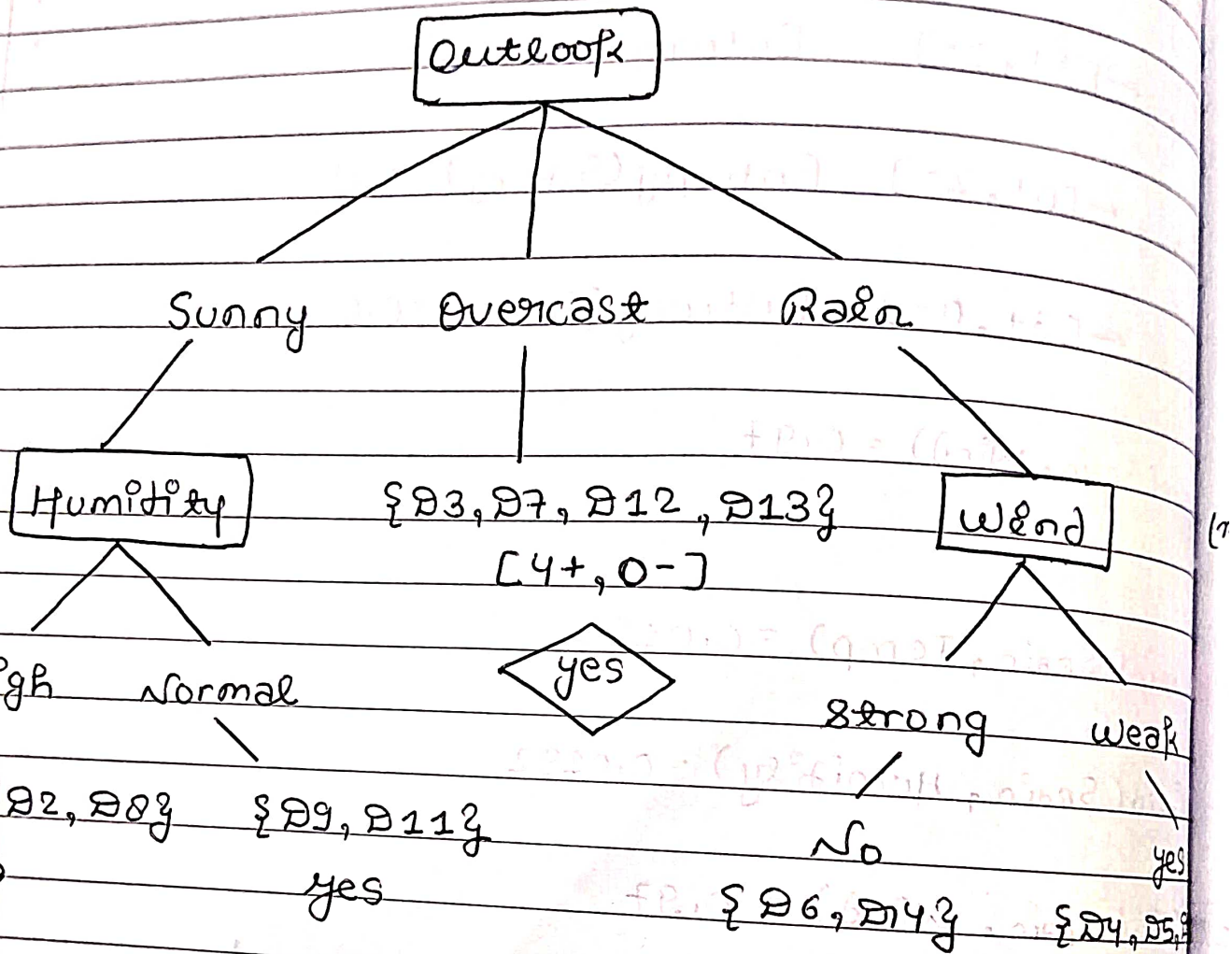
Gain (S_{Rain}, Humidity) = 0.0192

Gain (S_{Rain}, Wind) = 0.97

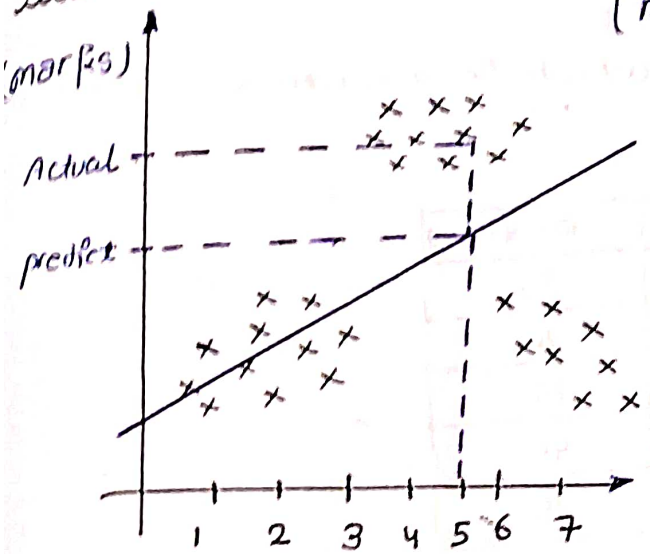


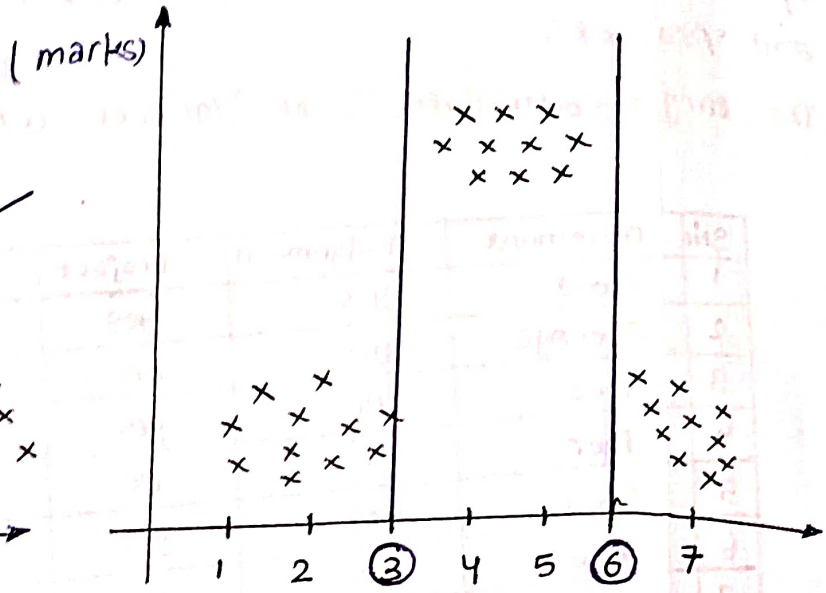
{D1, D2, ..., D14}

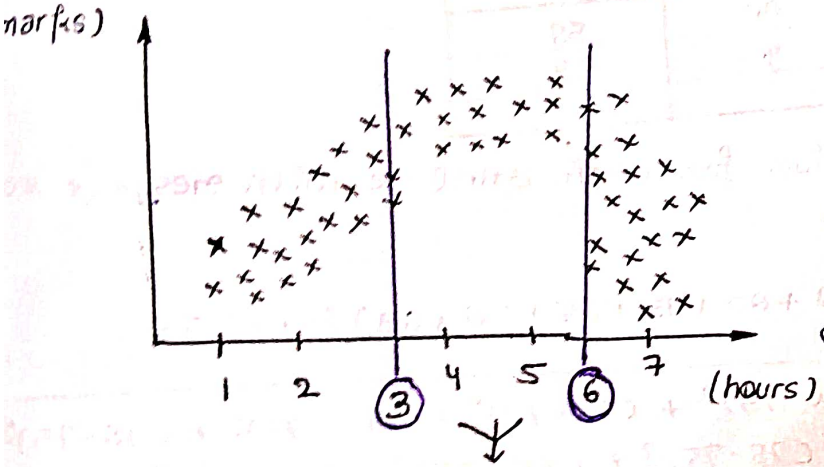
[9+, 5−]

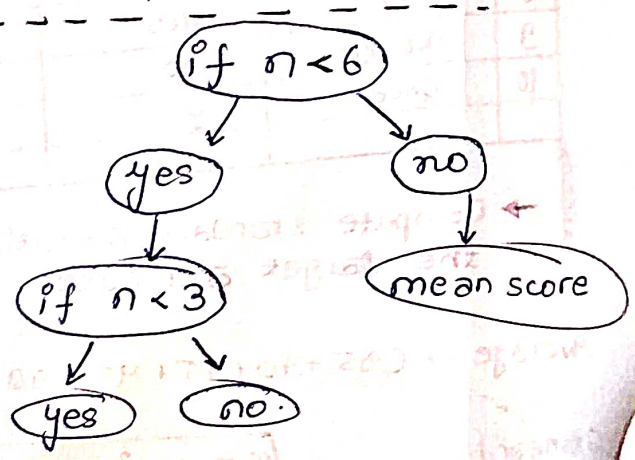

Outlook
Sunny
Overcast
Rain
Humidity
{D3, D7, D12, D13}
[4+, 0−]
Wind
High
Normal
yes
Strong
Weak
{D1, D2, D8}
No
{D9, D11}
yes
No
{D6, D14}
yes

Regression Tree





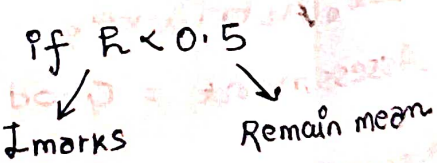


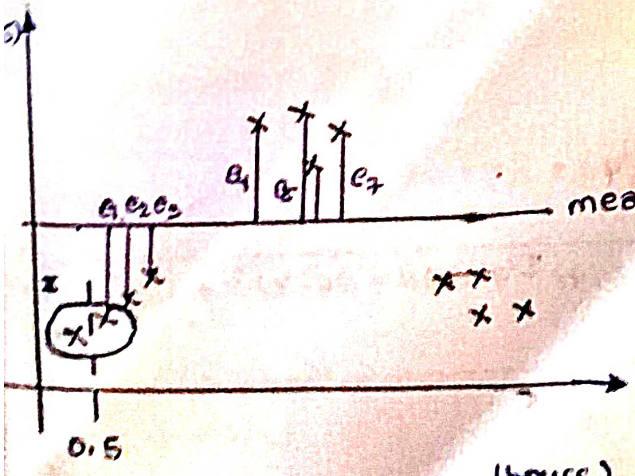


How to find these Threshold values?

I) selecting a group, find mean

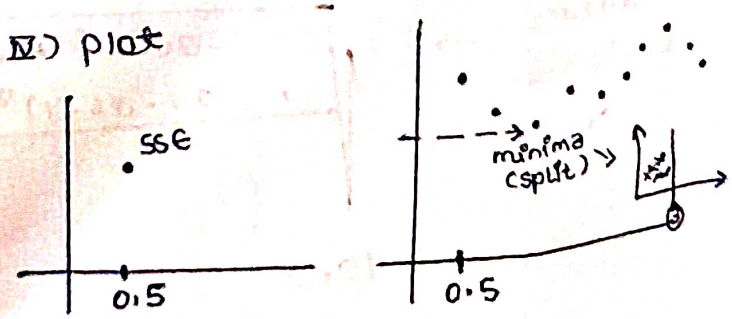
II) perform splitting





III) $SSE \Rightarrow e_1^2 + e_2^2 + \dots + e_n^2$

IV) plot



V) So pt by grouping others-

Construct a regression tree using the following data which consists of 10 data instances & 3 attributes 'Assessment', 'Assignment' and 'project'.

The target attribute is the 'result' which is a continuous attribute.

SNo	Assesment	Assignment	project	Result(%)
1	Good	yes	yes	95
2	Average	yes	no	70
3	Good	no	yes	75
4	Poor	no	no	45
5	Good	yes	yes	98
6	Average	no	yes	80
7	Good	no	no	75
8	Poor	yes	yes	65
9	Average	no	no	58
10	Good	yes	yes	89

→ Compute standard deviation for each attribute with respect to the target attribute

Average → (95+70+75+45+98+80+75+65+58+89)/10 → 75

Standard Deviation →

$$\sqrt{\frac{(95-75)^2 + (70-75)^2 + (75-75)^2 + (45-75)^2 + (98-75)^2 + (80-75)^2 + (75-75)^2 + (65-75)^2 + (58-75)^2 + (89-75)^2}{10}}$$

⇒ 16.55

* Let, Attribute,

Assessment = Good

Average → (95+75+98+75+89) → 86.4

Standard Deviation →

$$\sqrt{\frac{(95-86.4)^2 + (75-86.4)^2 + (98-86.4)^2 + (75-86.4)^2 + (89-86.4)^2}{5}}$$

⇒ 10.9

Let Attribute,
Assesment = Average

Average → (70+80+50) → 69.3

Std → $\sqrt{\frac{(70-69.3)^2 + (80-69.3)^2 + (50-69.3)^2}{3}}$

→ 11.01

Let Attribute,
Assessment = Poor

Average → (45+65)/2 → 55

Std → $\sqrt{\frac{(45-55)^2 + (65-55)^2}{2}}$

⇒ 14.14

Assesment	Sta. Dev	Data Instances
Good	10.9	5
Average	11.01	3
Poor	14.14	2

- Weighted Standard Deviation for Assessment

→ $(\frac{5}{10}) \times 10.9 + (\frac{3}{10}) \times 11.01 + (\frac{2}{10}) 14.14$ → 11.58

- Standard deviation reduction for Assessment

→ 16.55 − 11.58 → 4.97

Let Attribute,

Assignment = yes

Average → (95+70+98+65+89)/5 → 83.4

Std. Dev → 14.98

Let Attribute,

Assignment = No

Average → (75+45+80+75+50)/5 → 66.6

Std. Dev → 14.7

Construct a regression

Standard Deviation for Assignment		
Assignment	Std. Dev.	Data Instance.
Yes	14.98	5
No	14.7	5

Weighted standard deviation for assignment

$= \left(\frac{5}{10}\right) \times 14.98 + \left(\frac{5}{10}\right) \times 14.7 \rightarrow 14.84$

Std. Dev. reduction for assignment $\rightarrow 16.55 - 14.84$

$\rightarrow 1.71.$

Standard Deviation for Project		
Project	Std. Dev	Data Instance
yes	12.6	6
No	13.39	4

Weighted Std. Dev. for Assessment $\Rightarrow \left(\frac{6}{10}\right) \times 2.6 + \left(\frac{4}{10}\right) \times 13.39$

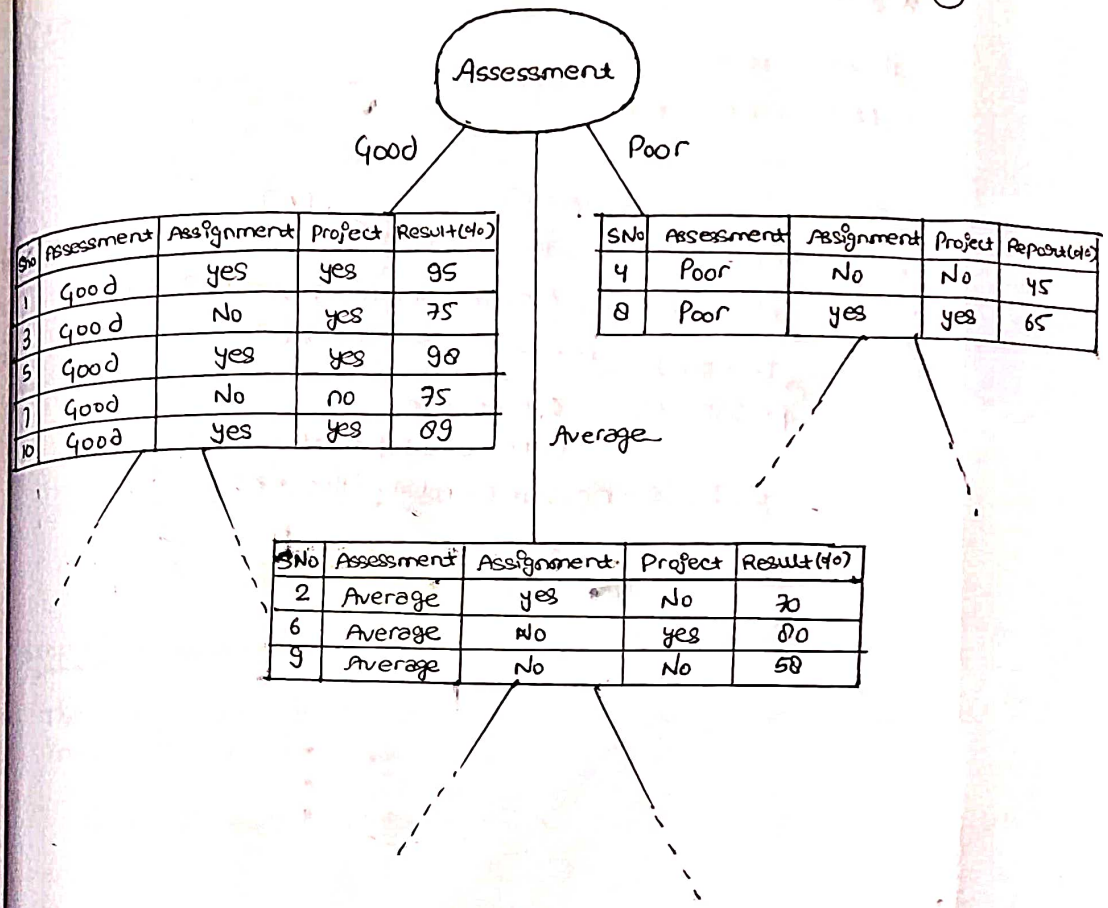
$\rightarrow 12.92$

Std. Dev. reduction for Assessment

$\rightarrow 16.55 - 12.92 \rightarrow 3.63$

Attribute	Standard dev. reduction
Assessment	4.97
Assignment	1.71
Project	3.63

→ make highest as node



Sno	Assessment	Assignment	Project	Result(40)
1	Good	yes	yes	95
3	Good	No	yes	75
5	Good	yes	yes	98
7	Good	No	no	75
10	Good	yes	yes	89

SNo	Assessment	Assignment	Project	Report(40)
4	Poor	No	No	45
8	Poor	yes	yes	65

SNo	Assessment	Assignment	Project	Result(40)
2	Average	yes	No	70
6	Average	No	yes	80
9	Average	No	No	58

★ function to calculate gini-impurity :

example labels

labels = [0, 1, 1, 0, 1, 0].

```
def gini_impurity(labels):                                  → [6].
    total_samples = len(labels)
    label_counts = np.unique(labels,                        → [3, 3]
                             return_counts = True)[1].
    probabilities = label_counts / total_samples            ← [0.5 0.5]
    gini = 1 - np.sum(probabilities ** 2)
```

Classification & Regression Trees!

CGPA	Interactive	Practical knowledge	Communication skills	Job offers
>=9	yes	Very Good	Good	yes
>=8	no	Good	Moderate	yes
>=9	no	Average	Poor	no
<8	no	Average	Good	no
>=8	yes	Good	Moderate	yes
>=9	yes	Good	Moderate	yes
<8	yes	Good	Poor	no
>=9	no	Very good	Good	yes
>=8	yes	Good	Good	yes
>=8	yes	Average	Good	yes

Step 1: Calculate the Gini index for the dataset

The target attribute "job offer" has '7' instances as 'yes' & '3' instances as 'no'.

$$\text{Gini\_Index}(T) = 1 - \sum_{i=1}^{m} P_i^2$$

$$\text{Gini\_Index}(T) = 1 - \left(\frac{7}{10}\right)^2 - \left(\frac{3}{10}\right)^2$$

$$\Rightarrow 1 - 0.49 - 0.09$$

$$\Rightarrow 1 - 0.58$$

$$\text{Gini\_Index}(T) \Rightarrow 0.42$$

Step 2: Compute Gini-Index for each of the attribute and each of the subset in the attribute

Categories of CGPA

CGPA	Job offer = "yes"	Job offer = "No"
≥9	3	1
≥8	4	0
<8	0	2

possible subsets ↓

2^3

↓

{ }, { >= 9 }, { >= 8 }, { > 8 }, { >= 9, >= 8 }, { >= 9, > 8 }, { >= 8, > 8 } and { >= 9, >= 8 and > 8 }

- Gini_Index (T) = $1 - \sum_{i=1}^{m} p_i^2$
- Gini_Index (T, A) = $\frac{|S_1|}{|T|}$ Gini(S1) + $\frac{|S_2|}{|T|}$ Gini(S2)

* Gini_Index (T, CGPA ∈ { >= 9, >= 8 }) => $1 - \left(\frac{7}{8}\right)^2 - \left(\frac{1}{8}\right)^2$

=> 1 - 0.7806 => 0.2194

* Gini_Index (T, CGPA ∈ { < 8 }) => $1 - (0/2)^2 - (2/2)^2$

=> 1 - 1 => 0

* Gini_Index (T, CGPA ∈ { { >= 9, >= 8 }, < 8 })

→ $\left(\frac{8}{10}\right)$ 0.2194 + $\left(\frac{2}{10}\right)$ 0 => 0.17552

⇓ similarly

* Gini_Index (T, CGPA ∈ { >= 9, < 8 }) => $1 - \left(\frac{3}{6}\right)^2 - \left(\frac{3}{6}\right)^2$

* Gini_Index (T, CGPA ∈ { >= 8 }) => $1 - \left(\frac{4}{4}\right)^2 - \left(\frac{0}{4}\right)^2$

=> 1 - 1 => 0

* Gini_Index (T, CGPA ∈ { { >= 9, < 8 }, >= 8 })

=> $\left(\frac{6}{10}\right) \times 0.5 + \left(\frac{4}{10}\right) \times 0$ => 0.3

Gini-Index (T, CGPA ∈ {>=8, <8}) ⇒ $1-\left(\frac{4}{6}\right)^2-\left(\frac{2}{6}\right)^2$

⇒ 1-0.55 ⇒ 0.445

* Gini-Index (T, CGPA ∈ {≥9}) ⇒ $1-\left(\frac{3}{4}\right)^2-\left(\frac{1}{4}\right)^2$

⇒ 1-0.625 ⇒ 0.375

* Gini-Index (T, CGPA ∈ {(>=8, <8), >=9}

⇒ $\left(\frac{6}{10}\right)$ (0.445) + $\left(\frac{4}{10}\right)$ (0.375) ⇒ 0.417

Step III) Choose the best splitting subset which has minimum Gini-Index for an attribute

The subset {(>=9, >=8), <8} has lowest Gini-Index value as 0.1755 is chosen as best splitting subset

Gini-Index of CGPA

Subsets		Gini-Index
>=9, >=8	<8	0.1755
>=9, <8	>=8	0.3
>=8, <8	>=9	0.417

Step IV) Compute ΔGini or best splitting subset of that attribute

$$\Delta Gini(CGPA) = Gini(T) - Gini(T, CGPA)$$

→ 0.42 - 0.1755 → 0.2445

Repeat the same process for the remaining attributes in the dataset such as for interactiveness, practical knowledge & communication skills

Categories for practical knowledge

Practical knowledge	Job offer = Yes	Job offer = No
Very good	2	0
good	4	1
average	1	2

- Gini-Index CT, Praco know. ∈ {Very good, good}
  $\Rightarrow 1-\left(\frac{6}{7}\right)^2-\left(\frac{1}{7}\right)^2 \Rightarrow 0.7544$
- Gini-Index CT, Praco know. ∈ {Average}) $= 1-\left(\frac{1}{3}\right)^2-\left(\frac{2}{3}\right)^2$
  $\Rightarrow 1-0.55 \Rightarrow 0.445$
- Gini-Index CT, praco know. ∈ ({Very good, good}, Avg))
  $\rightarrow \left(\frac{7}{10}\right)^2 \times 0.2456 + \left(\frac{3}{10}\right) \times 0.445 \Rightarrow 0.3054$
- Gini-Index CT, praco know. ∈ {Very good, average})
  $\Rightarrow 1-\left(\frac{3}{5}\right)^2-\left(\frac{2}{5}\right)^2 \Rightarrow 1-0.52 \Rightarrow 0.48$
- Gini-Index CT, praco know. ∈ {Good})
  $\Rightarrow 1-\left(\frac{4}{5}\right)^2-\left(\frac{1}{5}\right)^2 \Rightarrow 1-0.68 \Rightarrow 0.32$
- Gini-Index CT, praco know. ∈ ({Very good, avg}, good)
  $\rightarrow \left(\frac{5}{10}\right) \times 0.48 + \left(\frac{5}{10}\right) \times 0.32 \Rightarrow 0.40.$
- Gini-Index CT, praco know. ∈ {Good, Avg})
  $\Rightarrow 1-(5/8)^2-(3/8)^2 \Rightarrow 1-0.5312 \Rightarrow 0.4688$
- Gini-Index CT, praco know. ∈ {Good})
  $\rightarrow 1-(2/2)^2-(0/2)^2 \Rightarrow 1-1 \Rightarrow 0$
- Gini-Index CT, praco know. ∈ ({Good, Average}, very good)
  $\rightarrow \left(\frac{8}{10}\right) 0.4688 + \left(\frac{2}{10}\right) \times 0 \Rightarrow 0.3750$

Gini_Index for practical knowledge

Subsets		Gini-Index
(Very good, Good)	Average	0.3054
(Very good, Average)	Good	0.40
(Good, Average)	Very Good	0.3750

$\Delta Gini(praco\ knowo) = Gini(T) - Gini(T, praco\ knowo)$

$\Rightarrow 0.142 - 0.3054 \Rightarrow 0.1146.$

Categories for communication skills

Commo skills	Job offer =yes	NO
Good	4	1
Moderate	3	0
Poor	0	2

Gini-Index for subset commo skills

Subsets		Gini-Index
{Good, Moderate}	Poor	0.1755
{Good, Poor}	Moderate	0.3429
{Moderate, Poor}	Good	0.40

$\Delta Gini(Commo\ skills) = Gini(T) - Gini(T, Commo\ skills)$

$\Rightarrow 0.142 - 0.1755$

$\Rightarrow 0.2445$

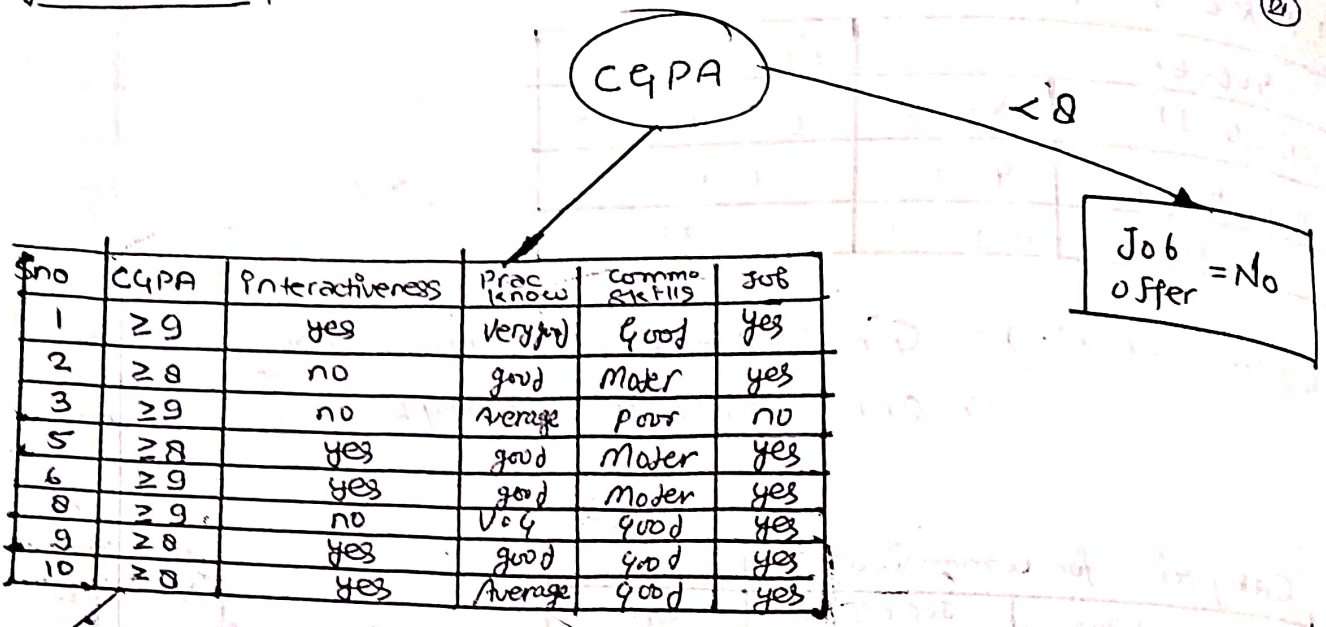
Gini Index & Δ Gini for all attributes

Attributes	Gini_index	ΔGini
CGPA	0.1755	0.2445
Interactiveness	0.360	0.052
praco knowo	0.3054	0.1146
commo skills	0.1755	0.2445

Subset CGPA = {(>=9, >=8), <8}

is the best splitting subset

From table,



Sno	CGPA	Interactiveness	Prac knowledge	Commo skills	Job
1	≥9	yes	Very good	Good	yes
2	≥8	no	good	Moder	yes
3	≥9	no	Average	poor	no
5	≥8	yes	good	Moder	yes
6	≥9	yes	good	Moder	yes
8	≥9	no	V.G	good	yes
9	≥8	yes	good	good	yes
10	≥8	yes	Average	good	yes