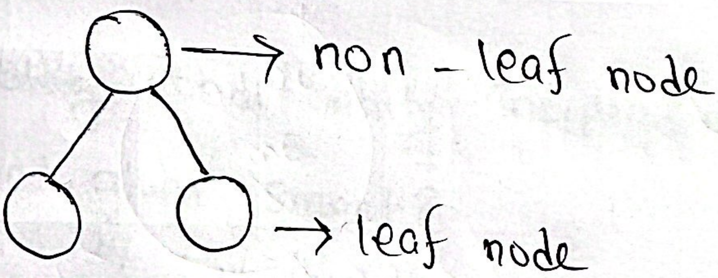
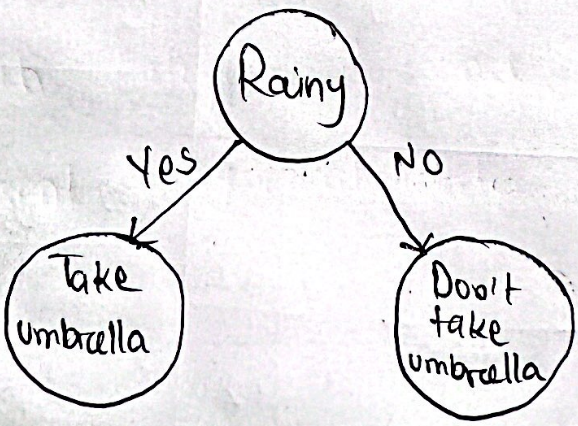
Video 16

Learning (Decision Tree)

in real life, if we are planning to get out of home, we take a look outside if we have to carry our umbrella or not. So, watching / observing the weather we are taking a decision.



সেই criteria অনুযায়ী আমাদের Decision নিচ্ছি সেগুলো leaf node এ রাখছি।



Basic tree

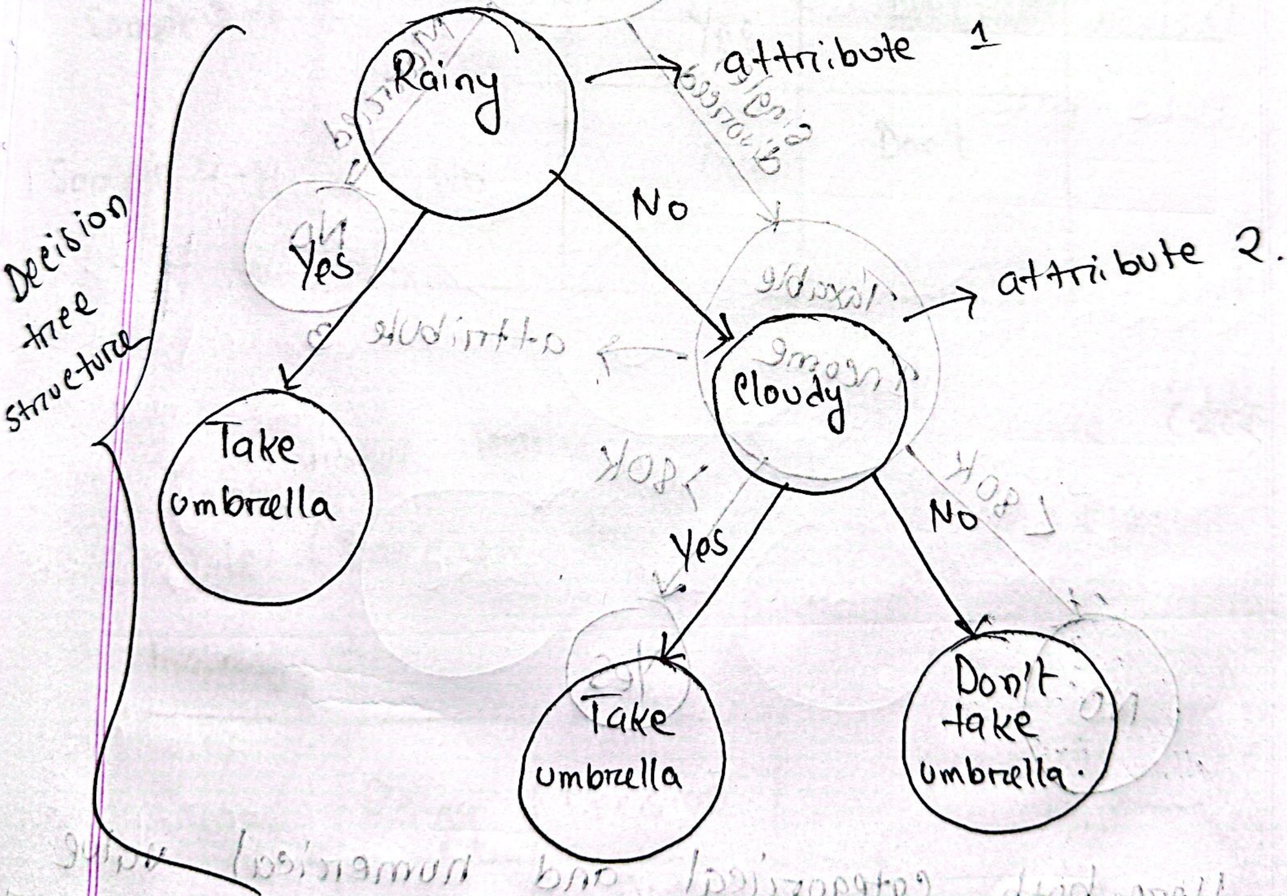
1 attribute

2 value

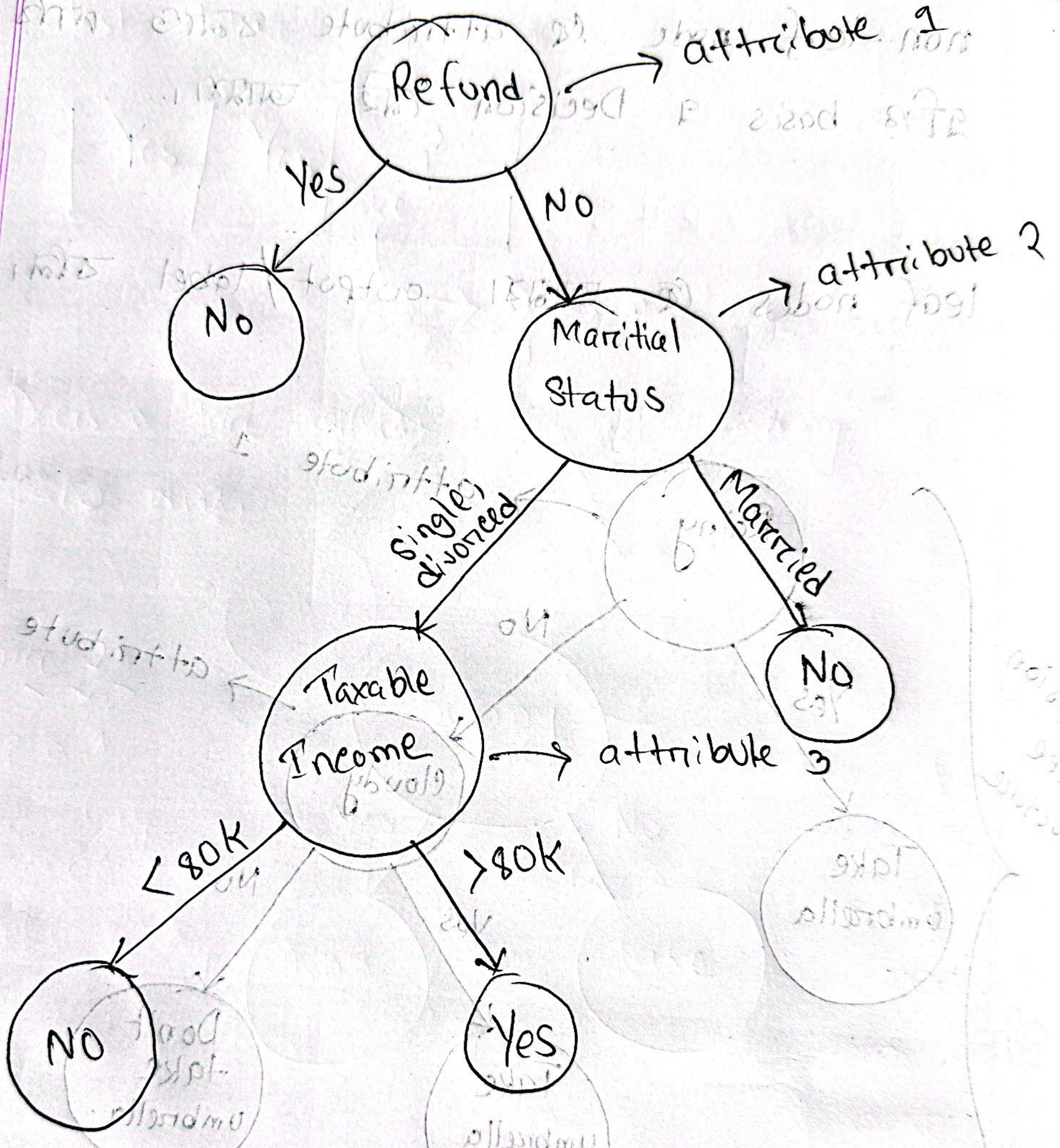
2 level on basis of value.

non-leaf node কে attribute বলতে পারি as এটির basis এ Decision নেই আমরা।

leaf nodes কে আমরা output/label বলি।



Yes, No → categorical value



Here, both categorical and numerical value are present. Categorical data is converted to numerical.

1st tree into tabular form

	Rainy	Cloudy	Decision
Sample 1 →	Yes	Yes	Take
Sample 2 →	Yes	No	Take
Sample 3 →	No	Yes	Take
Sample 4 →	No	No	Don't

feature

label

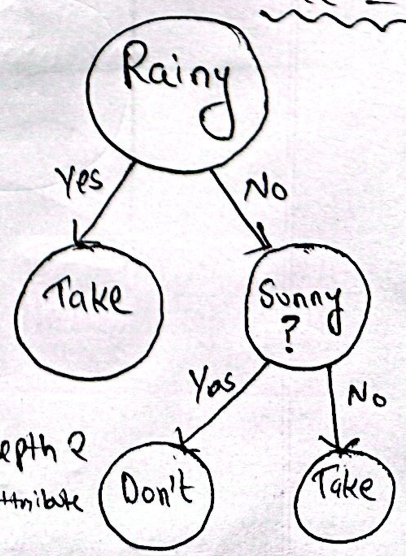
Supervised learning এ আমরা table থেকে model (Decision tree) এ convert করি।

training data

Sunny	Rainy	Decision
Yes	No	Don't
No	Yes	Take
No	No	Take

training phase → (learning phase)

tree 1



depth 2 attribute

test data

Yes	Yes	? take

→ Decision Tree থেকে দেখে।

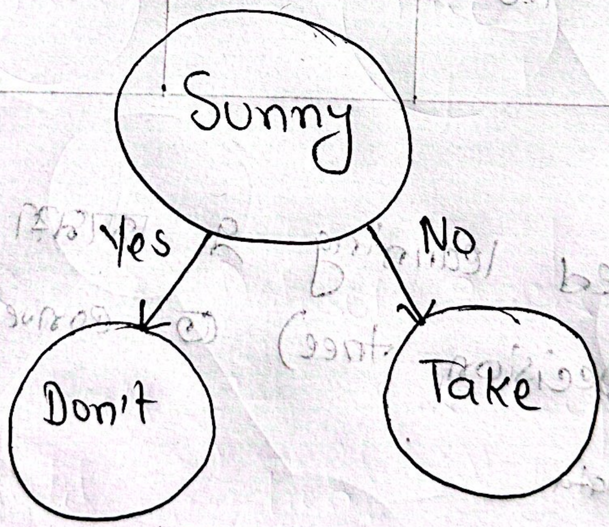
Decision tree আমরা supervised learning এ কাজে করতে পারি।

tree 2

depth 1

comparatively smaller tree

single attribute



যখন আমরা, non leaf node/ root node এ Sunny (attribute) ইউজ করেছি তখন এতে দিয়ে আমাদের rainy consider ই করা লাগতেছে না।

আমরা concise decision tree form করার try করবো। { কিছু criteria }.

কথা হচ্ছে কোন criteria decision tree কে concise করে।

Root node Sunny হইলে Decision tree concise.

Root node Rainy হইলে Decision tree long.

We describe it using

Information Purity:

Yes	No
−	+ +

Sunny

Yes	No
+	+ −

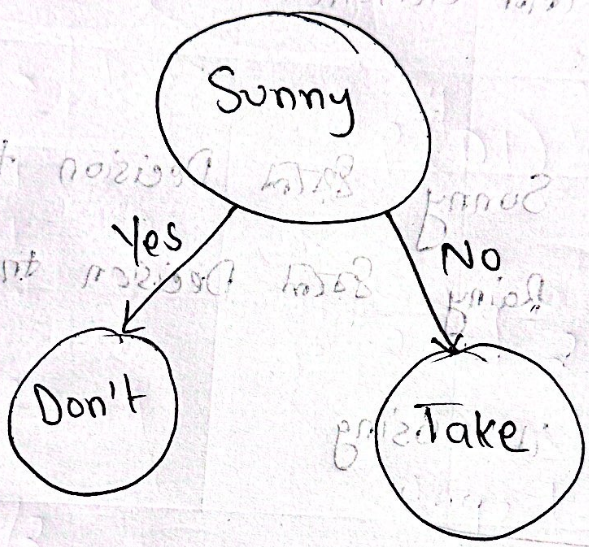
Rainy

mix of information এর জন্যে কোন conclusion এ পৌঁছানো যাচ্ছে না।

Don't → −

Take → +

∴ Sunny is more pure than Rainy. No mix of information. And it has more information. And, Sunny should have more priority.



As, purity Sunny এর বেশি, so এটি দিয়ে graph (decision tree) বানালে concise.

যখন অনেক বড় Dataset থাকবে, Graph plot করে purity check করা অনেক tough. Then we do mathmetical operation.

→ Entropy: Value High → less pure

Value Low → high purity

0 ←——— entropy ———→ 10

more pure ... less pure

Entropy = ~~[illegible]~~ $\sum_{i=n}^{n} -p_i \log p_i$

i = number of labels.

entropy attribute এর each unique value দিয়ে বের করি।

↳ entropy (Sunny = Yes)

↳ entropy (Sunny = No)

↳ entropy (Rainy = Yes)

↳ entropy (Rainy = No)

Sunny	Rainy	Decision
Yes	No	Don't
No	Yes	Take
No	No	Take

label 2 → take
don't

entropy(Sunny = Yes) = $- p(\text{Don't take}) \log_2 P(DT)$
$- p(\text{Take}) \log_2 (TU)$
$= - 1/1 \log_2 1/1 - 0/1 \log_2 0/1$

$= -\log_2 1$

$= 0$. (very pure).

Entropy (Sunny = NO) $= -P(DT) \log P(DT) - P(TU) \log_2 (TU)$

$= -\frac{0}{2} \log \frac{0}{2} - \frac{2}{2} \log_2 \frac{2}{2}$

$P(DT) = 0/2$

$P(TU) = 2/2$

$= -\log_2 1$

$= 0$ (very pure).

Entropy (Rainy = Yes) $= -P(DT) \log P(DT) - P(TU) \log_2 (TU)$

$P(DT) =$

$= -\frac{0}{1} \log \frac{0}{1} - \frac{1}{1} \log_2 \frac{1}{1}$

$P(TU)$

$= 0$ (Very pure).

Entropy (Rainy = No) = $-p(DT)\log_2 P(DT) - p(TU)\log_2 (TU)$

p(DT) =

$$= -\frac{1}{2}\log_2 \frac{1}{2} - \frac{1}{2}\log_2 \frac{1}{2}$$

$$= .5\log_2 0.5 - 0.5\log_2 0.5$$

= 1 (High entropy = less pure).

Entropy (Sunny) < Entropy (Rainy)

Sunny is more pure.

Video-17

Entropy → Information Gain → Select feature

Decision tree algorithm → ID3

Day	Outlook	Temp	Humidity	Wind	Decision
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

↳ Decision tree.

↳ → root node (anything).
→ but কোনটা আমরা বসাবো।

Entropy → Information Gain → feature selection.

Information Gain (Outlook)

$$\text{Entropy of Decision} = -p(yes)\log_2(yes) - p(no)\log_2(no)$$

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

$$= .940$$

$$\text{Entropy of (Sunny)} = -p(yes\backslash sunny)\log_2(yes\backslash sunny) - p(No\backslash sunny)\log_2(No\backslash sunny)$$

$$= -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}$$

$$= .971.$$

$$\text{Entropy}(\text{Rain}) = -p(\text{yes}\backslash\text{rain})\log_2(\text{Yes}\backslash\text{Rain}) - p(\text{No}\backslash\text{rain})\log_2(\text{No}\backslash\text{Rain})$$

$$= -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}$$

$$= .971$$

$$\text{Entropy}(\text{Overcast}) = -p(\text{Yes}/\text{overcast})\log_2(\text{Yes}|\text{overcast}) - p(\text{No}|\text{overcast})\log_2(\text{No}/\text{overcast})$$

$$= -\frac{4}{4}\log_2\frac{4}{4} - \frac{0}{4}\log_2\frac{0}{4}$$

$$= 0.$$

$$\text{Information Gain}(\text{Outlook}) = E(\text{Decision}) - E(\text{Sunny}) - E(\text{Rain}) - E(\text{overcast}).$$

এখানে Sunny and Rain 5 বার appear করেছে overcast appear করেছে 4 বার।

So, information gain calculate করার সময়,
যেই particular category একটু কম appear করবে তাকে একটু কম weight দেয়া and বেশি যারা appear করবে তাদেরকে বেশি weight দেওয়া।

So,

$$IG(outlook) = E(Decision) - P(sunny) * E(sunny)$$
$$- P(Rain) * E(Rain)$$
$$- P(overcast) * E(overcast).$$

$$= .940 - \left(\frac{5}{14} \times .971\right) - \left(\frac{5}{14} \times .971\right)$$
$$- \left(\frac{4}{14} \times 0\right)$$

Information gain = .246 → so outlook থেকে আমরা এই amount of information পাবো।

So, Temp, Humidity and wind এর জন্যও same ভাবে information gain বের করতে হবে and then যার gain বেশি তাকে root node বানিয়ে decision tree বানাতে হবে।

Entropy (Humidity = High):

$$= -\frac{3}{7}\log_2\frac{3}{7} - \frac{4}{7}\log_2\frac{4}{7}$$

$$= .985$$

Entropy (Humidity = Normal)

$$= -\frac{6}{7}\log_2\frac{6}{7} - \frac{1}{7}\log_2\frac{1}{7}$$

$$= .592.$$

Information gain (Humidity) = E(Decision) − P(High) * E(High) − P(Normal) * E(Normal)

$$= .940 - \left(\frac{7}{14} \times .985\right) - \left(\frac{7}{14} * .592\right)$$

Information Gain (Humidity) = .151

$\Rightarrow IG(\text{Outlook}) = .246$

$\Rightarrow IG(\text{Humidity}) = .151$

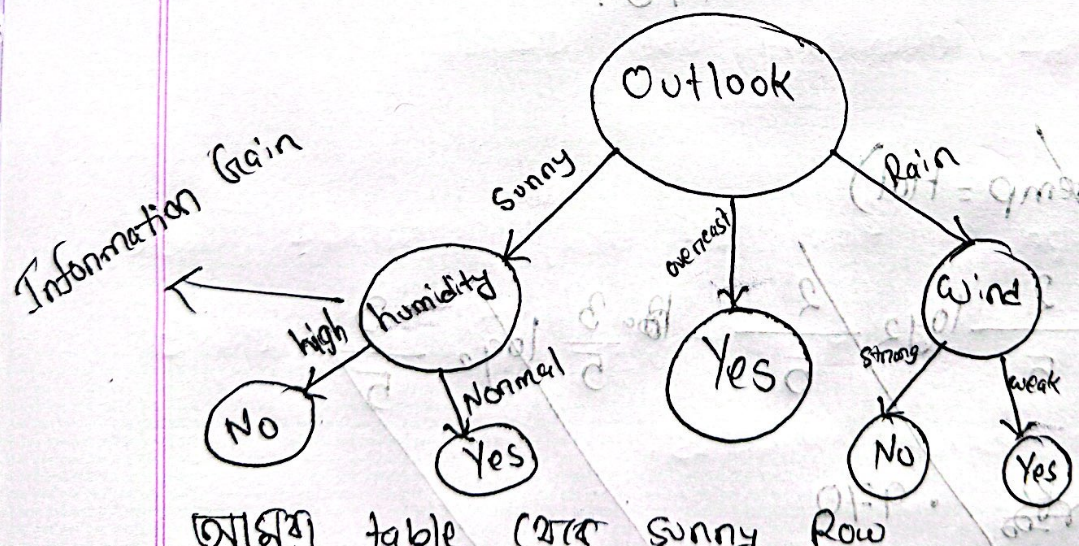
$\Rightarrow IG(\text{Wind}) = .048$

$\Rightarrow IG(\text{Temp}) = .029$

আমাদের যদি Outlook and Humidity থেকে select করতে হয় যে কোনটি root node নিবো তাহলে Outlook নিবো as, এখান থেকে বেশি information পাই।

More information Gain = more priority

Outlook has the most information gain



আমরা outlook থেকে overcast পেলে Yes দিচ্ছি, Rain & Sunny পেলে কিছু দিতে পারতেছি না as mix info. So,

আমরা table থেকে sunny Row আলাদা করলাম then,

Day	Outlook	Temp	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
8	Sunny	mild	High	Weak	Yes
9	Sunny	cool	Normal	Weak	Yes
11	Sunny	mild	Normal	Strong	Yes.

$$\text{Entropy (Decision)} = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}$$

$$= \cdot 970.$$

~~Entropy (temp = Hot)~~

$$\cancel{= -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}}$$

$$\cancel{= \cdot 970}$$

Entropy (Hot)

$= -p(no/hot)\log_2 p(no/hot) - p(yes/hot)\log_2 p(yes/hot)$

$= -2/2 \log 2/2 - 0/2 \log_2 0/2$

$= 0.$

Entropy (Mild)

$= -\frac{1}{2}\log_2\frac{1}{2} - 1/2 \log_2 1/2$

$= 1.$

Entropy (Cool)

$= -\frac{0}{1}\log_2 0/1 - 1/1 \log_2 1/1$

$= 0.$

$IG(Temp) = E(D) - P(Hot) * E(Hot) - P(Mild) * E(Mild) - P(Cool) * E(Cool)$

$= .970 - (2/5 * 0) - (2/5 * 1) - (1/5 * 0)$

= ·570.

∴ Information gain (Temp) = ·570

Same ভাবে Humidity and wind এর জন্য করতে হবে।

IG (Humidity) = ·970

IG (Wind) = ·019.

So, আমরা humidity use করবো।

* last Graph a Rain এর জন্য আমাদের table থেকে Rain এর Row গুলো isolate করতে হবে। Then Same process.

$$\text{Entropy (Decision)} = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}$$

$$= .970.$$

Entropy (~~Rain =~~ ~~Temp~~ Wind) =