ASSIGNMENT-7

Roll no: 19K41A0575

Draw a decision tree diagram to predict Number of hours to play based on weather conditions like outlook, temperature, humidity, windy, consider dataset shown below

Step-1 :-

Roll NO. missing

Outlook	Temperature	Humidity	Windy	hours to play
Rainy	Hot	high	False	25
Rainy	Hot	High	True	30
Overcast	Hot	High	False	46
Sunny	mild	High	False	45
Sunny	Cool	Normal	False	52
sunny	Cool	Normal	True	23
Overcast	cool	Normal	True	43
Rainy	mild	high	false	35
Rainy	cool	Normal	False	38
Sunny	Mild	Normal	False	46
Rainy	Mild	Normal	False	48
Overcast	Mild	high	True	52
Overcast	hot	Normal	True	44
sunny	Mild	high	True	30

Step : 2 :- Calculate SD, CN, mean.

10

18(12)

$$\text{mean} = \frac{\sum x}{n}$$

$$= \frac{25+30+46+45+52+23+43+35+38+46+48+52+44+30}{14}$$

$= \frac{557}{14} = 39.78$

$SD = \sqrt{\frac{\Sigma(x-mean)^2}{n}} = 9.67$

$CN = \frac{SD}{mean} \times 100 = \frac{9.67}{39.78} \times 100 = 24.30$

Step 3:- Dataset is split on different attributes the SD of each branch is calculated

$SD(attr) = \Sigma w(branch)\ SD(branch)$

and the result is standard deviation reduction

$SDR = SD - SD(attr)$

$\therefore SD(target) = 9.67$

Outlook :-

	Mean	SD	CV	n	w(v)
Rainy	39.2	24.7	5	5/14	8.7
Overcast	46.25	8.72	4	4/14	4.03
Sunny	39.2	31.0	5	5/14	12.2

$SD(outlook) = \frac{5}{14}(8.7) + \frac{4}{14}(4.03) + \frac{5}{14}(12.2)$

$= 8.59$

$SDR(outlook) = SD(target) - SD(outlook)$

$= 9.67 - 8.59$

$= 1.08$

19K41A05775

Temp:-

Temp	Mean	SD	CV	n	w (v)
hot	36.25	10.34	30.6	4	4/14
Cool	39	12.14	31.1	4	4/14
Mild	42.6	3.38	19.65	6	6/14

$$SD(Temp) = \frac{4}{14}(10.34) + \frac{4}{14}(12.14) + \frac{6}{14}(3.38)$$

$$= 10.01$$

$$SDR(Temp) = 9.67 - 10.01$$

$$= -0.34$$

Humidity

Humidity	Mean	SD	CV	n	w(v)
High	37.57	10.11	26.92	7	7/14
Normal	42	9.4	27.4	7	7/14

$$SD(Humidity) = \frac{7}{14} \times 10.11 + \frac{7}{14} \times 9.14$$

$$= 9.77$$

$$SDR(Humidity) = 9.67 - 9.77 = -0.1$$

Windy:-

Windy	Mean	SD	CV	n	w(v)
True	37.6	11.6	30.8	6	6/14
False	41.3	8.11	20.3	8	8/14

$$SD(windy) = \frac{6}{14} \times 11.6 + \frac{7}{14} \times (7.41) = 9.77$$

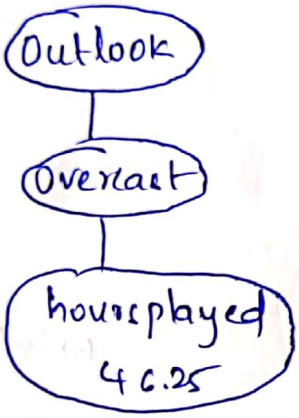
$$SDR(windy) = 9.67 - 9.77 = 0.1$$

The value that has highest SDR is considered as root node i.e (decision node).

Considering termination criteria

CV is 10%. (or) CV is (n ≤ 4)

Overcast has CV of 8% which is less than threshold value therefore we need not to further split.



We need to split node sunny and Rainy.

Outlook	Temp	Humidity	Windy	Hours played
sunny	mild	Normal	True	30
sunny	mild	high	false	45
sunny	cool	Normal	False	52
sunny	cool	Normal	False	46
sunny	mild	Normal	True	23

$Mean = 39.2$

$SD = 12.2$

$CV = 31.0$

Temp:-

Temp	mean	SD	CV	n	w(v)
Mild	40.3	8.96	22.23	3	3/5
COOL	37.5	20.50	54.66	2	2/5

$SD(Temp) = \frac{3}{5}(8.96) + \frac{2}{5}(20.50) = 13.576$

$= 12.2 - 13.576 = 1.37$

Humidity.

Humid	mean	SD	CV	n	w(v)
high	37.5	10.6	28.26	2	2/5
Normal	40.3	15.3	37.9	3	3/5

$SD(Humid) = \frac{2}{5}(10.6) + \frac{3}{5}(15.30)$

$= 64(10.6) + 0.6(15.30)$

$= 13.42.$

$SD(Humid) = 12.2 - 13.42 = -1.22.$

Windy:-

Windy	Mean	SD	CV	n	w(v)
false	47.66	3.78	7.94	3	3/5
True	26.5	4.94	18.65	2	2/5

$SD(windy) = \frac{3}{5}(3.78) + \frac{2}{5}(4.94) = 4.23$

$SD(windy) = 12.2 - 4.23 = 7.97$

Then check for highest SDR

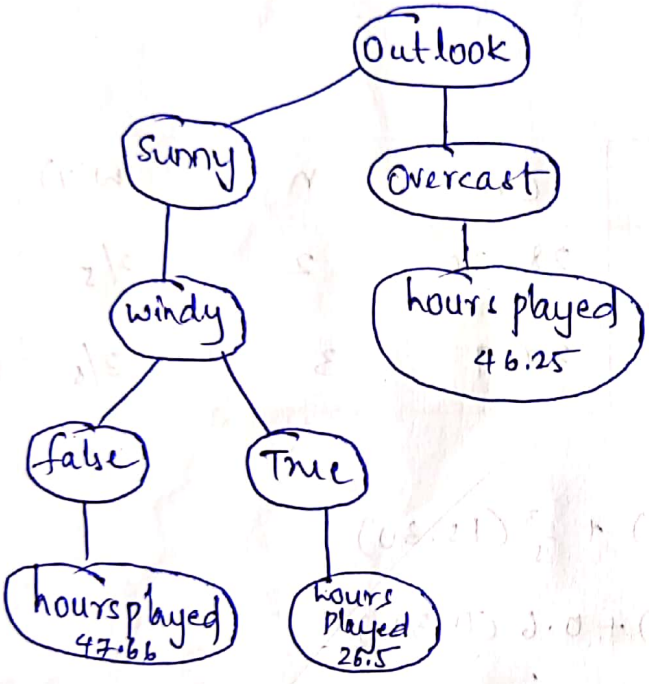
In outlook, among Temp, humidity and windy

SDR value is high for windy.

SDR = 7.97

Then, check for CV value..

Both True and False satisfy the CV value



Rainy:-

outlook	Temp	humdity	Windy	hours to play
Rainy	hot	high	False	25
Rainy	hot	high	False	35
Rainy	hot	high	True	30
Rainy	Kotd	Normal	False	38
Rainy	hrild	Nomal	True	48.

mean = 35.2

SD = 8.7

CV = 24.7

Temperature :-

Temp	mean	SD	CV	n	W(V)
hot	27.5	3.53	12.13	2	2/5
mild	41.5	9.19	22.164	2	2/5
cool	38	0	0	1	1/5

$$SD(Temp) = \frac{2}{5}(3.53) + \frac{2}{5}(9.19) + \frac{1}{5} \times 0 = 5.088$$

$$SDR(Temp) = SD - SD(Temp) = 8.7 - 5.088 = 3.612$$

Humidity :-

Humidity	mean	SD	CV	n	W(V)
high	30	5	16.66	3	3/5
Normal	43	7.07	16.44	2	2/5

$$SD(Humidity) = \frac{3}{5}(5) + \frac{2}{5}(7.07) = 5.828$$

$$SDR(Humidity) = SD - SD(Humidity) = 8.7 - 5.828 = 2.872$$

Windy :-

Windy	Mean	SD	CV	n	W(V)
False	32.66	6.80	20.85	3	3/5
True	39	12.72	32.5	2	2/5

$$SD(windy) = 3/5(6.80) + 2/5(12.72) = 9.168$$

$$SDR(windy) = SD - (SD(windy)) = 87 - 9.168 = -0.468.$$

among Temp, humidity and windy the SDR value is high for temperature (ie, 3.612) Then check for CV value of HOT, mild & cool satisfy the CV value.