AI

Roll No: 19K91A0450
ECE sec A
Year: III

Assignment - 7

Draw a decision tree diagram to predict no. of hours to play based on weather conditions like outlook, temp., humidity, windy. Consider dataset shown below.

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	cool	Normal	False	38
Sunny	Mild	Normal	False	46
Rainy	Mild	Normal	True	48
overcast	Mild	High	True	52
Overcast	Hot	Normal	False	44
Sunny	Mild	High	True	30

Termination criteria: CV <= 10% or min. no. of samples = 4

Cal. of mean, standard deviation, coefficient of variation

$$\text{mean} = \frac{\Sigma x}{n} = \frac{551}{14} = 39.78$$

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

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

Now, data set is split into different attributes. The SD of each brach is cal.

$$SD(\text{altr}) = \Sigma\, w(\text{branch}) \cdot SD(\text{branch})$$

and the result SDR (standard deviation reduction) is cal.

$$SDR = SD - SD(\text{altr})$$

$$SD = 9.67$$

Outlook:

outlook	mean	SD	CV	n	w(v)
Rainy	35.2	8.7	24.7	5	5/14
overcast	46.25	4.03	8.72	4	4/14
Sunny	39.2	12.2	81.0	5	5/14

$$\therefore SD(\text{outlook}) = \frac{5}{14} \times 8.7 + \frac{4}{14} \times 4.03 + \frac{5}{14} \times 12.2 = 8.59$$

$$SDR(\text{outlook}) = SD - SD(\text{out look}) = 9.67 - 8.59 = 1.08$$

Temperature:

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	43.6	8.38	19.65	6	6/14

$$\therefore SD(\text{temp.}) = \frac{4}{14} \times 10.34 + \frac{4}{14} \times 12.14 + \frac{6}{14} \times 8.38 = 10.01$$

$$SDR(\text{temp.}) = SD - SD(\text{temp.}) = 9.67 - 10.01 = -0.34$$

Humidity:

Humidity	mean	SD	CV	n	w(Iv)
High	37.51	10.11	26.92	7	7/14
Normal	42	9.4	22.4	7	7/14

$\therefore$ SD(humidity) $= \frac{7}{14} \times 10.11 + \frac{7}{14} \times 9.4 = 9.77$

SDR (humidity) = SD - SD (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	8.41	20.3	8	8/14

$\therefore$ SD(windy) $= \frac{6}{14} \times 11.6 + \frac{8}{14} \times 8.41 = 9.77$

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

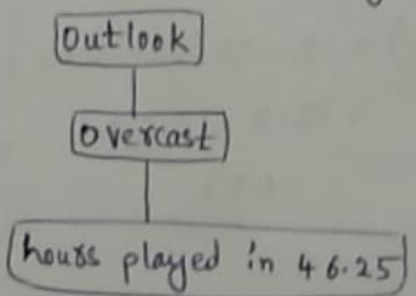
The value that has highest SDR is considered as root node (i.e, decision node)

Considering termination criteria

CV is 10% or CV is ($n \leq 4$)

Outlook:

overcast has CV of 8% which is less than threshold value therefore, we need not go to further splitting.



We need to split Sunny and rainy columns

Sunny:

outlook	Temp.	Humidity	windy	hours played
Sunny	mild	high	False	45
Sunny	cool	normal	False	52
Sunny	cool	normal	True	23
Sunny	mild	normal	False	46
Sunny	mild	high	True	30

∴ mean = 39.2 , SD = 12.2, CV = 31.0

Temperature:

Temp.	mean	SD	CV	n	w(v)
mild	40.3	8.96	22.73	3	3/5
cold	35.5	20.50	54.66	2	2/5

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

$SDR(temp.) = SD - SD(temp) = 12.2 - 13.576 = -1.37$

Humidity:

Humidity	mean	SD	CV	n	w(v)
High	37.5	10.6	28.26	2	2/5
normal	40.3	15.3	37.96	3	3/5

$SD(Humidity) = \frac{2}{5} \times 10.6 + \frac{3}{5} \times 15.3 = 13.42$

$SDR(Humidity) = SD - SD(Humidity)$

$= 12.2 - 13.42$

$= -1.22$

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

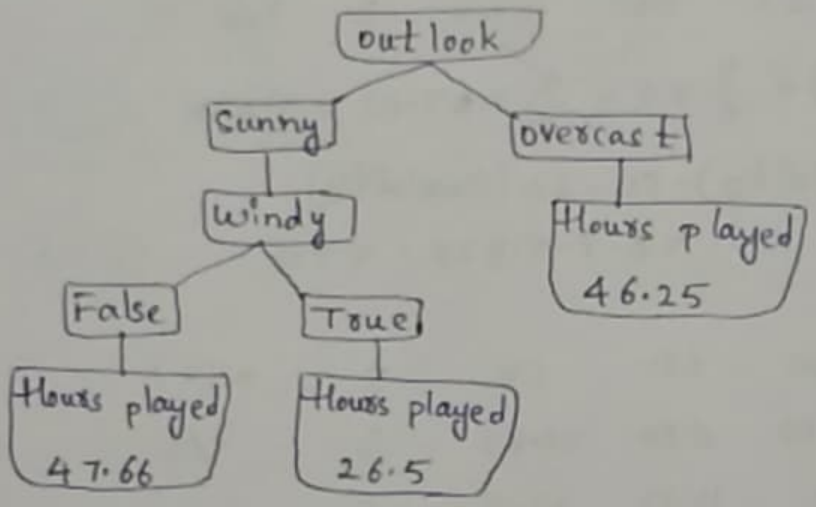
$SD(windy) = \frac{3}{5} \times 3.78 + \frac{2}{5} \times 4.94 = 4.23$

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

In outlook, among temp, humidity and windy SDR value is high for windy SDR = 7.97

Then, check for CV value

Both true & false satisfy the CV value



Rainy:

Out look	temp.	Humidity	windy	Hours played
Rainy	Hot	High	False	25
Rainy	Hot	High	True	30
Rainy	mild	High	False	35
Rainy	cool	Normal	False	38
Rainy	Mild	Normal	True	48

Mean = 35.2, SD = 8.7, CV = 24.7

Temp.:

Temp.	mean	SD	CV	n	w(v)
Hot	27.5	3.53	12.83	2	2/5
Mild	41.5	9.19	22.144	2	2/5
cool	38	0	0	1	1/5

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

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

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

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

$SDR(Humidity) = SD - SD(Humidity)$

$= 8.7 - 5.828 = 2.872$

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

$SD(windy) = \frac{3}{5} \times 6.80 + \frac{2}{5} \times 12.72 = 9.168$

$SDR(windy) = SD - SD(windy) = 8.7 - 9.168 = -0.468$

Among temp., Humidity and windy the SDR value is high for temp. (i.e, 3.612). Then check for cv value of hot, mild, cold satisfy the cv value.

Decesion tree diagram to predict number of hours to play based on weather conditions: