

Assignment - 7

19K41A0447

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

dataset:

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	True	48
overcast	Mild	High	True	52
overcast	Hot	normal	False	44
Sunny	mild	High	True	30

Termination criteria: $CV \leq 10\%$ or minimum number of sample 4.

447
10
18/11/21

calculating mean, standard deviation (SD), co-efficient of variation (cv)

$$\text{mean} = \frac{\sum x}{n} = 39.78$$

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

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

Now, data set is split into different attributes. The SD of each branch is calculated

$$SD(\text{attr}) = \sum w(\text{branch}) \cdot sd(\text{branch})$$

and the result SDR (Standard deviation reduction) is calculated

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

$$\therefore SD = 9.67$$

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

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

$$SD(\text{outlook}) = SD - SD(\text{outlook})$$

$$= 9.67 - 8.59$$

$$= 1.08$$

Temperature:-

Temperature	mean	SD	CV	n	w(v)
Hot	36.5	10.34	30.6	4	4/14
Cool	39	12.14	31.1	4	4/14
Mild	42.6	8.38	19.65	6	6/14

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

$$SDR(\text{temperature}) = SD - SD(\text{temperature}) = 9.67 - 10.01 = -0.344$$

Humidity:

Humidity	mean	SD	CV	n	w(H)
High	37.31	10.11	26.92	7	7/14
Normal	42	9.4	22.4	7	7/14

$$\therefore SD(\text{humidity}) = \frac{7}{14} \times 10.11 + \frac{7}{14} \times 9.4 = 9.85$$

$$SDR(\text{humidity}) = SD - SD(\text{humidity}) = 9.67 - 9.85 = -0.18$$

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

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

$$SDR(\text{windy}) = 9.67 - 9.77 = -0.10$$

$$SDR(\text{outlook}) = 1.08$$

$$SDR(\text{temperature}) = -0.3$$

$$SDR(\text{humidity}) = -0.1$$

$$SDR(\text{windy}) = -0.1$$

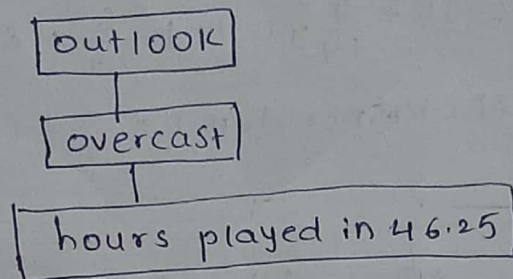
The value that has highest SDR is considered as root node

(i.e., decision node)

consider termination criteria

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

overcast has CV of 0% which is less than threshold value therefore, we need not go to further splitting



We need to split sunny and rainy columns

Sunny:

outlook	temperature	humidity	windy	hoursplayed
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

$$\text{mean} = 39.2$$

$$SD = 12.2$$

$$CV = 31.0$$

Temperature:

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

$$SD(\text{temperature}) = \frac{3}{5} \times 8.96 + \frac{2}{5} \times 20.50 = 9.99 \approx 13.57$$

$$\begin{aligned} SDR(\text{temperature}) &= SD - SD(\text{temperature}) \\ &= 12.2 - 13.57 \\ &= -1.37 \end{aligned}$$

Humidity:

Humidity	mean	SD	CV	n	w(CV)
high	37.5	10.6	28.26	2	2/5
Normal	40.3	15.30	37.96	3	3/5

$$SD(\text{humidity}) = \frac{2}{5} \times 10.6 + \frac{3}{5} \times 15.30 = 13.42$$

$$SDR(\text{humidity}) = SD - SD(\text{humidity}) = 12.2 - 13.42 = -1.22$$

Windy:

windy	mean	SD	CV	n	w(CV)
False	47.66	3.78	7.94	3	3/5
True	26.5	4.94	18.65	2	2/5

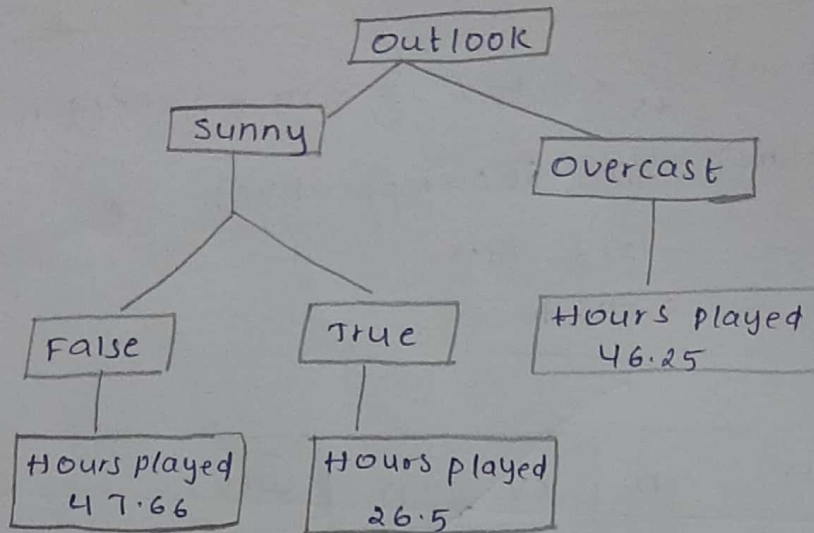
$$SD(\text{windy}) = \frac{3}{5} \times 3.78 + \frac{2}{5} \times 4.94 = 4.24$$

$$SDR(\text{windy}) = SD - SD(\text{windy}) = 12.2 - 4.24 = 7.95$$

In outlook among temperature, 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	Temperature	humidity	windy	hoursplayed
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

temperature:

Temperature	mean	SD	CV	n	w(CV)
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

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

$$= 5.088$$

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

$$= 8.7 - 5.088$$

$$= 3.612$$

Humidity:

Humidity	mean	SD	CV	n	w(CV)
High	30	5	16.66	3	3/5
normal	43	7.07	16.44	2	2/5

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

$$SDR(\text{humidity}) = SD - SD(\text{humidity}) = 8.7 - 5.828 = 2.872$$

Windy

windy	mean	SD	CV	n	w(CV)
false	32.66	6.80	20.85	3	3/5
True	39	12.72	32.5	2	2/5

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

$$SDR(\text{windy}) = SD - SD(\text{windy}) = 8.7 - 9.168 = -0.468$$

Among temperature humidity and windy the SDR value is high for temperature (i.e., 3.612). Then check for CV value of hot, mild, cool.
satisfy the CV value

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

