ASSIGNMENT - 7.

Step 1:

outlook	Temperature	Humidity	Windy	hoursb 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

Step 2:- calculate SD, CN, mean

$$mean = \varepsilon \, x/n$$

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

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

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

$$S.D. = 9.67$$

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

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

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

and result is SDR (standard deviation Reduction).

$$SDR = S.D. - SD\,(attr)$$

$$\therefore SD\,(target) = 9.67$$

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	31.0	5	5/14

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

$$= 8.59$$

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

$$= 9.67 - 8.59 = 1.08$$

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:

	mean	SD	CV	n	W(v)
high	37.51	10.1	26.92	7	7/14
normal	42	9.4	27.4	7	7/14

$$SD(\text{humidity}) = \frac{7}{14} \times 10 \cdot 11 + \frac{7}{14} \times 9.14$$

$$= 9.77$$

$$SDR(\text{humidity}) = 9.67 - 9.77 = -0.1$$

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

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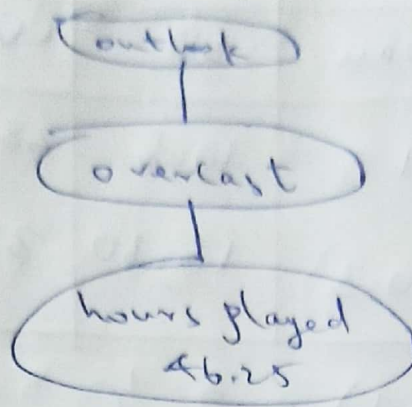
$$= 9.77$$

$$SDR(\text{windy}) = 9.67 - 9.77 = -0.1$$

→ The value having highest SDR is considered as root node (ie., 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, so we need not further split.

```
          ┌─────────┐
          │ outlook │
          └────┬────┘
               │
         ┌───────────┐
         │ overcast  │
         └─────┬─────┘
               │
        ┌──────────────┐
        │ hours played │
        │    46.25     │
        └──────────────┘
```

→ we need to split node sunny & rainy

outlook	Temp	humidity	windy	hrs 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

$$mean = 39.2$$
$$SD = 12.2$$
$$CV = 31.0$$

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

$$SD(Temp) = \frac{3}{5}(8.96) + \frac{2}{5}(20.50) = 13.576$$
$$= 12.2 - 13.576 = -1.37.$$

humid	mean	SD	cv	n	w(v)
high	37.5	10.6	28.26	2	2/5
normal	40.3	15.30	37.96	3	3/5

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

$$= 6.4\,(10.6) + 0.6\,(15.30)$$

$$= 13.42$$

$$SD\,(humid) = 12.2 - 13.42 = -1.22$$

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

$$SD\,(windy) = 3/5\,(3.78) + 2/5\,(4.94)$$
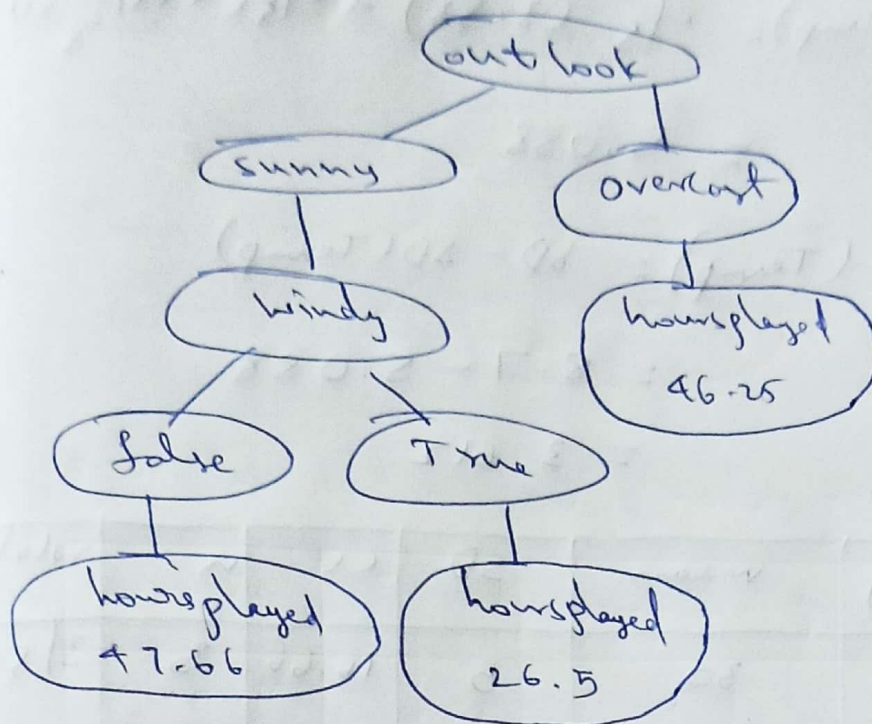
$$= 4.23$$

$$SD\,(windy) = 12.2 - 4.23 = 7.97.$$

then check for highest SDR

In outlook, among temp, humidity & windy SDR value is high for windy

SDR = 7.97

Then, check for cv value both True & false satisfy cv value.

outlook

- sunny
  - windy
    - false → hoursplayed 47.66
    - True → hoursplayed 26.5
- Overcast
  - hoursplayed 46.25

Rainy:

outlook	Temperature	humidity	windy	hrstoplay
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

$$\text{mean} = 35.2$$
$$SD = 8.7$$
$$CV = 24.7$$

Temperature	mean	SD	CV	n	w(v)
hot	24.5	3.53	12.83	2	2/5
mild	41.5	9.19	22.14	2	2/5
cool	38	0	0	1	1/5

$$SD(Temp) = \frac{2}{5}(3.55) + \frac{2}{5}(2.12) + \frac{1}{5} \times 0$$

$$= 5.088$$

$$SDR(Temp) = SD - SD(Temp)$$

$$= 8.7 - 5.088$$

$$= 3.612$$

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

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

$$= 9.168$$

$$SDR(windy) = SD - SD(windy)$$

$$= 8.7 - 9.168 = 0.468$$

*Among Temp, humidity & windy SDR value is high for Temperature (i.e., 3.612) Then, check for cn value of hot, mild & cool satisfy cn value.

outlook

Sunny   Overcast   Rainy

Windy   46.25 hrs played   Temp

False   True   hot   mild   cool

47.6 hrs played   26.5 hrs played   27.5 hrs playd   41.5 hrs pld   38 hrs played