A1 Assignment

- 19K41A0469
Nada Tabani

→ Draw a decision tree diagram to predict number of hours to play based on weather conditions like Outlook, temparature, humidity, windy, consider dataset shown below:

Outlook	Temparature	Humiddy	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
Overcast	Cool	Normal	True	43
Rainy	Mild	High	False	35
Rainy	~~Mild~~	Normal	False	38
Sunny	Mild	Normal	True	
Rainy	Mild	Normal	True	46
Overcast	Mild	High	True	48
Overcast	Hot	Normal	False	52
Sunny	Mild	High	True	44
Sunny	Cool	Normal	True	30
				23

Termination criteria : CV <= 10% or minimum no. of same

2. calculating mean, Standard deviation (SD), coefficient of variation (CV)

$$\text{mean} = \frac{\Sigma x}{n} = \frac{557}{14} = 39.78$$

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

$$CV = \frac{SD}{mean} \times 100 = \frac{9.67}{39.78} \times 100 = 34.50$$

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

SD(after) = Σw(branch) · SD(branch)

and the result SDR (Standard deviation reduction) is calculated SDR = SD − SD(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(Outlook) = $\frac{5}{4} * 8.7 + \frac{4}{14} * 4.03 + \frac{5}{14} * 12.2 = 8.59$

SDR(Outlook) = SD − SD(Outlook) = 9.67 − 8.59 = 1.08

Temparature :

Temparature	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	8.38	19.65	6	6/14

3.

SD(temparature) = $\frac{4}{14} * 10.34 + \frac{4}{14} * 12.14 + \frac{6}{14} * 8.36 = 10.01$

SDR(temparature) = SD − SD(temparature) = 9.67 − 10.01 = 0.34

Humidity :

Humidity	mean	SD	CV	n	w(7V)
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) = $\frac{6}{14} * 11.6 + \frac{8}{14} * 8.41 = 9.77$

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

SDR(outlook) = 1.08

SDR(Temparature) = -0.34

SDR(Humidity) = -0.1

SDR(windy) = -0.1

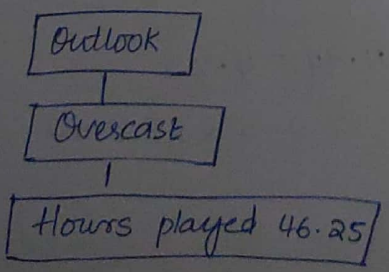
The value that has highest SDR is considered as root node (i.e decision node)

Considering termination criteria

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

Over cast has CV of 8% which is less than threshold.

4. Value therefore, we need not go for further spliting.



We need to Split Sunny & rainy columns.

Outlook	Temparature	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	50

$\therefore$ mean = 39.2, SD = 12.2; CV = 31.0

Temparature:

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

SD(temparature) = $\frac{3}{5} * 8.96 + \frac{3}{5} * 20.5 = 13.576$

SDR(temparature) = SD - SD(temparature) = 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	15.30	37.96	3	3/5

SD(humidity) = $2/5 * 10.6 + \frac{3}{5} * 15.30 = 13.44$

SDR(humidity) = SD - SD(Humidity) = 12.2 - 13.42 = -1.22

3. 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) = $\frac{3}{5} * 3.78 + \frac{2}{5} * 4.94 = 4.23$

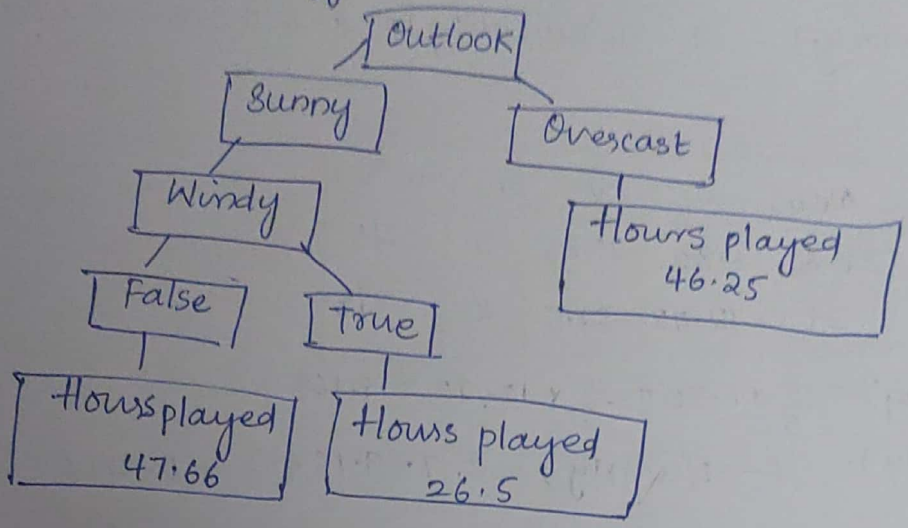
SDR(windy) = SD - SD(windy) = 12.2 - 4.23 = 7.97

In Outlook

among, Temparature, 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 :

Outlook	Temparature	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

6. Mean = 35.2 SD = 8.7, CV = 24.7

Temparature	Mean	SD	CV	n	w(v)
Hot	29.5	3.53	12.83	2	2/5
Mild	41.5	9.19	22.144	2	2/5
cool	38	0	0	1	1/5

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

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

Humidity:

Humidity	Mean	SD	CV	n	w(v)
High	30	5	16.66	3	3/5
Normal	43	7.07	16.64	2	2/5

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

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) = $\frac{3}{5} * 6.80 + \frac{2}{5} * 12.72 = 9.168$

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

Among, Temparature & windy, SDR value for Temp (3.612) is high. check CV value of hot, mild & cold satisfy the cv value.

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