19K41A0416
M.shashank

Assignment - 7

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

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	48
Sunny	cool	normal	False	52
Overcat	cool	normal	~~False~~True	43
Rainy	mild	high	False	35
Rainy	cool	normal	False	38
Sunny	mild	normal	False	46
Rainy	mild	normal	~~False~~ True	48
Over cast	mild	high	True	52
Over Cast	Hot	nomal	false	44
Sunny	mild	high	True	30
Sunny	Cool	normal	True	23

Termination Criteria : CV <=10%. or minimum number of samples : 4

Calculating mean, standard derivation (SD), co-efficient of variation (CV)

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

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

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

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

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

and the result SDR (standard derivation reduction) is calculated

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

$$SD = 9.67$$

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

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

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

Temperatore :

Temperatore	mean	SD	CV	n	w(v)
Hot	36.25	10.34	30.6	4	$\frac{4}{14}$
Cool	39	12.14	31.1	4	$\frac{4}{14}$
mild	40.6	8.38	19.65	6	6/14

SD (temperatore) = 4/14 ^ 10.34 + 4/14 * 12.14 + 6/14 × 8.38 = 10.01

SDR (temperatore) = SD - SD (temperatore) = 9.67 - 10.01 = 0.34

Homidity :

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

SD(homidity) = 7/14 × 10.11 + $\frac{7}{14}$ × 9.4 = 9.77

= SD - SD (homidity)

= 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

SD (windy) = 6/14 * 11.6 + 8/14 * 8.41 = 9.77

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

SDR (outlook) = 1.08

SDR (temperature) = −0.34

SDR (Humidity) = −0.1

SDR (windy) = −0.1

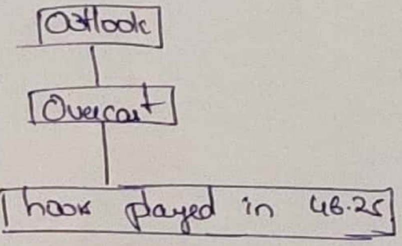
The value that has highest SDR is considered as root node (i.e decision node)

Considering termination criteria.

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

outlook

Overcast has CV of 8% which is less than threshold value therefore we need not go for further splitting



We need to split sunny and rainy columns

Sunny:

Outlook	Temperature	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

Temperatore :

Temperatore	mean	SD	CV	n	w(v)
mild	40.3	896	22.23	3	3/5
cold	35.5	20.50	54.66	2	2/5

$$SD(\text{temperatore}) = \frac{3}{5} \times 8.96 + \frac{2}{5} \times 20.50$$

$$= 13.576$$

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

$$= 12.2 - 13.576 - \cancel{1} - 1.37$$

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

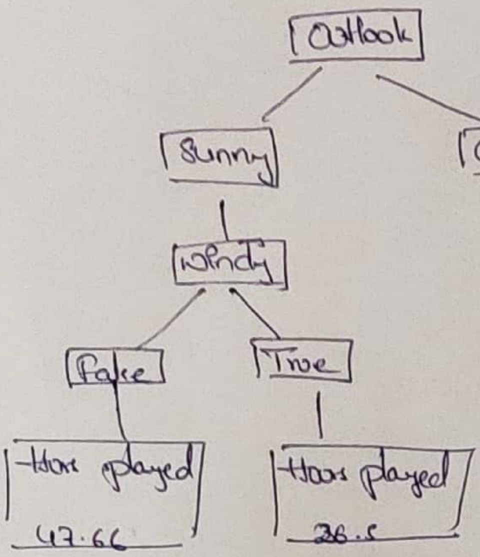
$$SD(\text{windy}) = \frac{3}{5} \times 3.78 + 2/5 \times 4.94$$

$$= 4.23$$

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

$$= 12.2 - 4.23 = 7.97$$

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 sedisfy the CV value



Rainy:

Outlook	Temperature	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

Temperatore:

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

$$SD(temp) = \frac{2}{5} \times 8.53 + \frac{2}{5} \times 9.19 + \frac{1}{5} \times 0$$

$$= 5.088$$

$$SDR(temperatore) = SD - SD(temp)$$

$$= 8.7 - 5.088$$

$$= 3.612$$

Humidity

Humidity	mean	SD	CV	n	w(v)
High	30	5	16.66	3	3/5
normal	43	7.07	1644	2	2/5

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

$$SDR(humidity) = SD - SD(humidity)$$

$$= 87 - 5.828$$

$$= 2872$$

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

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

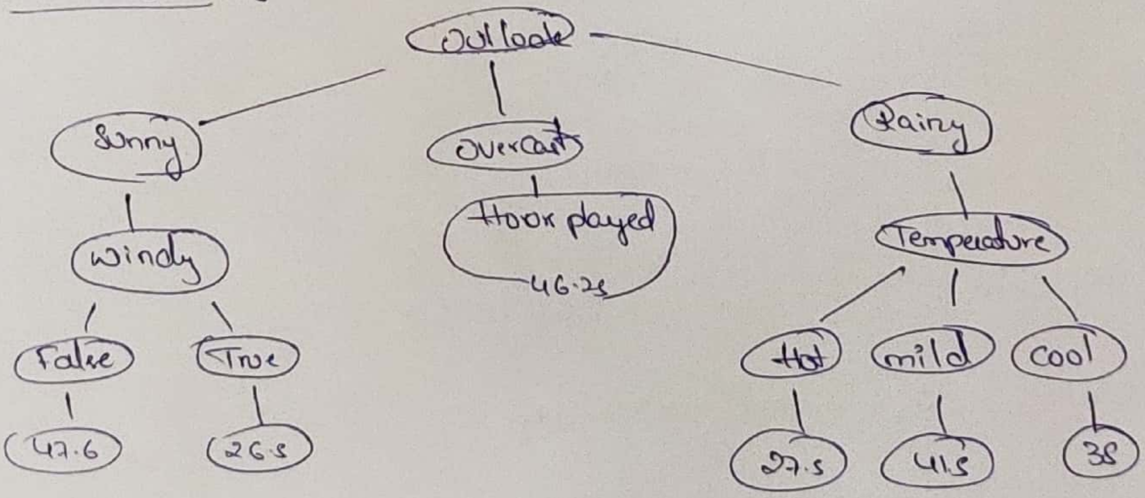
$$= 9.168$$

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

$$= 8.7 - 9.168 = -0.468$$

Among temperature homidity and windy the SDR value is high for temperature (ie 3.612). Then check for CV value of hot, mild, cold satisfy the CV value

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



Outlook
Sunny
Overcast
Rainy
Windy
Hour played
46.25
Temperature
False
True
47.6
26.5
Hot
mild
cool
27.5
41.5
38