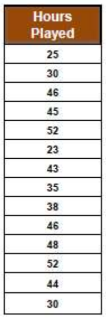
| Outlook | Temperature | Humidity | Windy | Hours to play |
|---|---|---|---|---|
| Rainy | Hot | High | False | 25 |
| Rainy | Hot | High | True | 30 |
| Overcast | Hot | High | False | 46 |
| sunny | Mild | High | False | 45 |
| sunny | Cool | Normal | False | 52 |
| sunny | Cool | Normal | True | 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 |

Step1: Calculate SD, count, mean, cv



| Hours Played |
|---|
| 25 |
| 30 |
| 46 |
| 45 |
| 52 |
| 23 |
| 43 |
| 35 |
| 38 |
| 46 |
| 48 |
| 52 |
| 44 |
| 30 |

$$\boldsymbol{Count} = n = 14$$

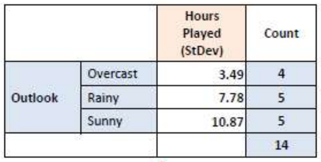
$$\boldsymbol{Average} = \bar{x} = \frac{\sum x}{n} = 39.8$$

$$\boldsymbol{Standard\ Deviation} = S = \sqrt{\frac{\sum(x-\bar{x})^2}{n}} = 9.32$$

$$\boldsymbol{Coeffeicient\ of\ Variation} = CV = \frac{S}{\bar{x}} * 100\% = 23\%$$

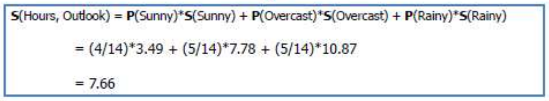
Step2: Calculate SD and SDR for two attributes

$$S(T,X) = \sum_{c \in X} P(c)S(c)$$



|  |  | Hours Played (StDev) | Count |
|---|---|---|---|
| Outlook | Overcast | 3.49 | 4 |
|  | Rainy | 7.78 | 5 |
|  | Sunny | 10.87 | 5 |
|  |  |  | 14 |

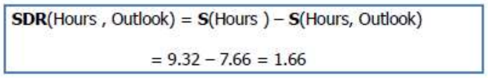
$$SDR(T,X) = S(T) - S(T,X)$$



**S**(Hours, Outlook) = **P**(Sunny)***S**(Sunny) + **P**(Overcast)***S**(Overcast) + **P**(Rainy)***S**(Rainy)

= (4/14)*3.49 + (5/14)*7.78 + (5/14)*10.87

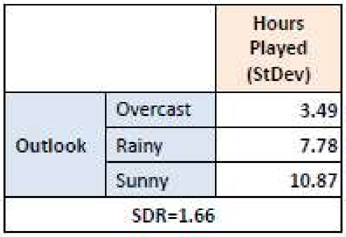
= 7.66



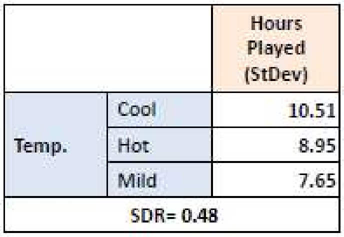
**SDR**(Hours , Outlook) = **S**(Hours ) – **S**(Hours, Outlook)

= 9.32 – 7.66 = 1.66

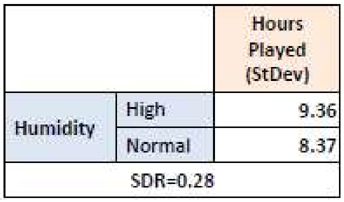
Step2: The dataset is then split on the different attributes. The standard deviation for each branch is calculated. The resulting standard deviation is subtracted from the standard deviation before the split. The result is the standard deviation reduction.



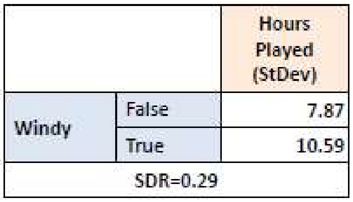
| | | Hours Played (StDev) |
|---|---|---|
| Outlook | Overcast | 3.49 |
| | Rainy | 7.78 |
| | Sunny | 10.87 |
| SDR=1.66 | | |



| | | Hours Played (StDev) |
|---|---|---|
| Temp. | Cool | 10.51 |
| | Hot | 8.95 |
| | Mild | 7.65 |
| SDR= 0.48 | | |

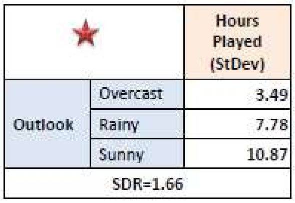


| | | Hours Played (StDev) |
|---|---|---|
| Humidity | High | 9.36 |
| | Normal | 8.37 |
| SDR=0.28 | | |



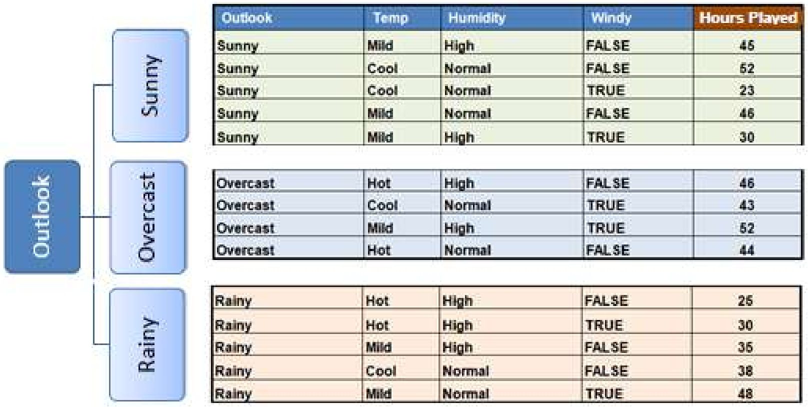
| | | Hours Played (StDev) |
|---|---|---|
| Windy | False | 7.87 |
| | True | 10.59 |
| SDR=0.29 | | |

***Step 3***: The attribute with the largest standard deviation reduction is chosen for the decision node.



| ★ | | Hours Played (StDev) |
|---|---|---|
| Outlook | Overcast | 3.49 |
| | Rainy | 7.78 |
| | Sunny | 10.87 |
| SDR=1.66 | | |

***Step 4a***: The dataset is divided based on the values of the selected attribute. This process is run recursively on the non-leaf branches, until all data is processed.



| 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 |

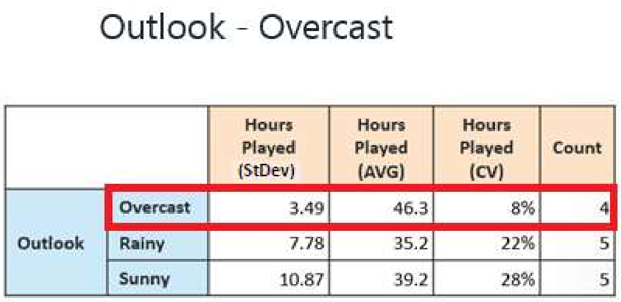
| | | | | |
|---|---|---|---|---|
| Overcast | Hot | High | FALSE | 46 |
| Overcast | Cool | Normal | TRUE | 43 |
| Overcast | Mild | High | TRUE | 52 |
| Overcast | Hot | Normal | FALSE | 44 |

| | | | | |
|---|---|---|---|---|
| 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 |

In practice, we need some termination criteria. For example, when coefficient of deviation (**CV**) for a branch becomes smaller than a certain threshold (e.g., 10%) and/or when too few instances (**n**) remain in the branch (e.g., 3).

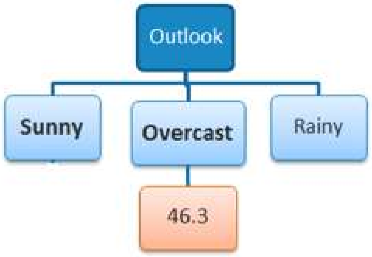
***Step 4b***: "Overcast" subset does not need any further splitting because its CV (8%) is less than the threshold (10%).

The related leaf node gets the average of the "Overcast" subset.

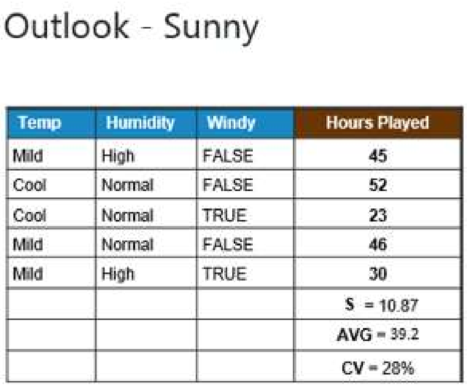


Outlook - Overcast

| | | Hours Played (StDev) | Hours Played (AVG) | Hours Played (CV) | Count |
|---|---|---|---|---|---|
| Outlook | Overcast | 3.49 | 46.3 | 8% | 4 |
| | Rainy | 7.78 | 35.2 | 22% | 5 |
| | Sunny | 10.87 | 39.2 | 28% | 5 |

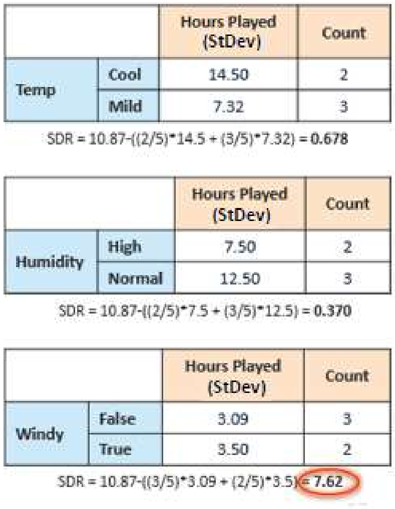


***Step 4c***: However, the "Sunny" branch has an CV (28%) more than the threshold (10%) which needs further splitting. We select "Windy" as the best best node after "Outlook" because it has the largest SDR.



Outlook - Sunny

| Temp | Humidity | Windy | Hours Played |
|---|---|---|---|
| Mild | High | FALSE | 45 |
| Cool | Normal | FALSE | 52 |
| Cool | Normal | TRUE | 23 |
| Mild | Normal | FALSE | 46 |
| Mild | High | TRUE | 30 |
| | | | S = 10.87 |
| | | | AVG = 39.2 |
| | | | CV = 28% |



| | | Hours Played (StDev) | Count |
|---|---|---|---|
| Temp | Cool | 14.50 | 2 |
| | Mild | 7.32 | 3 |

SDR = 10.87-((2/5)*14.5 + (3/5)*7.32) = **0.678**

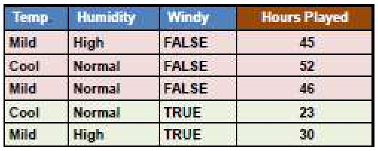
| | | Hours Played (StDev) | Count |
|---|---|---|---|
| Humidity | High | 7.50 | 2 |
| | Normal | 12.50 | 3 |

SDR = 10.87-((2/5)*7.5 + (3/5)*12.5) = **0.370**

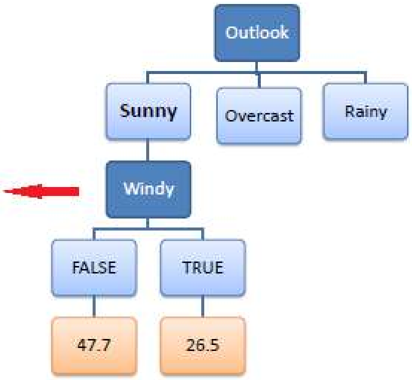
| | | Hours Played (StDev) | Count |
|---|---|---|---|
| Windy | False | 3.09 | 3 |
| | True | 3.50 | 2 |

SDR = 10.87-((3/5)*3.09 + (2/5)*3.5) = **7.62**

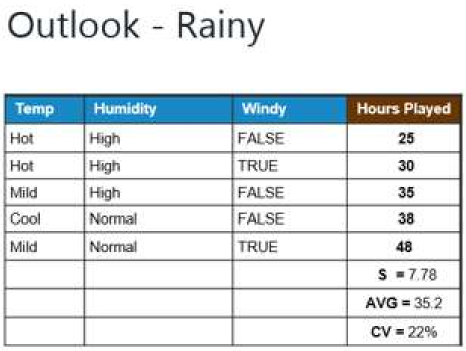
Because the number of data points for both branches (FALSE and TRUE) is equal or less than 3 we stop further branching and assign the average of each branch to the related leaf node.



| Temp | Humidity | Windy | Hours Played |
|---|---|---|---|
| Mild | High | FALSE | 45 |
| Cool | Normal | FALSE | 52 |
| Mild | Normal | FALSE | 46 |
| Cool | Normal | TRUE | 23 |
| Mild | High | TRUE | 30 |

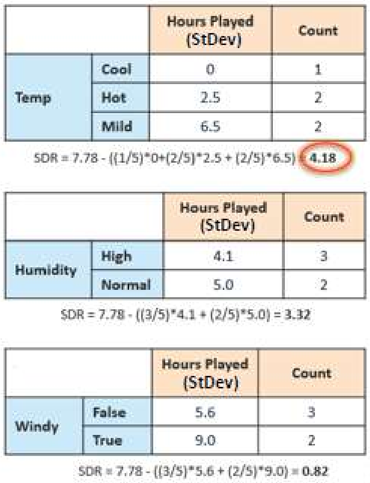


***Step 4d***: Moreover, the "rainy" branch has an CV (22%) which is more than the threshold (10%). This branch needs further splitting. We select "Temp" as the best best node because it has the largest SDR.



Outlook - Rainy

| Temp | Humidity | Windy | Hours Played |
|---|---|---|---|
| Hot | High | FALSE | 25 |
| Hot | High | TRUE | 30 |
| Mild | High | FALSE | 35 |
| Cool | Normal | FALSE | 38 |
| Mild | Normal | TRUE | 48 |
| | | | S = 7.78 |
| | | | AVG = 35.2 |
| | | | CV = 22% |



| | | Hours Played (StDev) | Count |
|---|---|---|---|
| Temp | Cool | 0 | 1 |
| | Hot | 2.5 | 2 |
| | Mild | 6.5 | 2 |

SDR = 7.78 - ((1/5)*0+(2/5)*2.5 + (2/5)*6.5) = 4.18

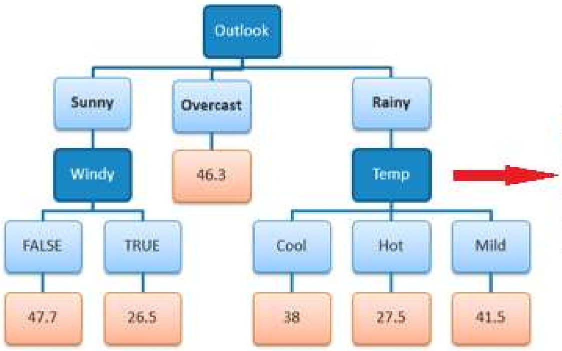
| | | Hours Played (StDev) | Count |
|---|---|---|---|
| Humidity | High | 4.1 | 3 |
| | Normal | 5.0 | 2 |

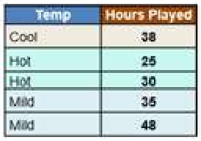
SDR = 7.78 - ((3/5)*4.1 + (2/5)*5.0) = 3.32

| | | Hours Played (StDev) | Count |
|---|---|---|---|
| Windy | False | 5.6 | 3 |
| | True | 9.0 | 2 |

SDR = 7.78 - ((3/5)*5.6 + (2/5)*9.0) = 0.82

Because the number of data points for all three branches (Cool, Hot and Mild) is equal or less than 3 we stop further branching and assign the average of each branch to the related leaf node. When the number of instances is more than one at a *leaf node* we calculate the *average* as the final value for the target.





| Temp | Hours Played |
|---|---|
| Cool | 38 |
| Hot | 25 |
| Hot | 30 |
| Mild | 35 |
| Mild | 48 |