Decision Tree for Regression Problems

Problem Statement:

A sport hosting company would like to decide to host a cricket match between India and South Africa based on whether data.

Whether data that is available has attributes like outlook, temperature, humidity and wind. And has a decision variable how many hours were played.

We need to build a decision tree model to predict based on whether data how many hours will be played?

Data:

Outlook	Temperature	Humidity	Windy	Hours Played
Sunny	Hot	High	FALSE	25
Sunny	Hot	High	TRUE	30
Overcast	Hot	High	FALSE	46
Rainy	Mild	High	FALSE	45
Rainy	Cool	Normal	FALSE	52
Rainy	Cool	Normal	TRUE	23
Overcast	Cool	Normal	TRUE	43
Sunny	Mild	High	FALSE	35
Sunny	Cool	Normal	FALSE	38
Rainy	Mild	Normal	FALSE	46
Sunny	Mild	Normal	TRUE	48
Overcast	Mild	High	TRUE	52
Overcast	Hot	Normal	FALSE	44
Rainy	Mild	High	TRUE	30

Solution:

The ID3 algorithm can be used to construct a decision tree for regression type problems by replacing Information Gain with Standard Deviation Reduction – SDR

A decision tree is built top down from a root node and involves partitioning the data into subsets that contain instances with similar values mean homogeneous data.

Here, standard deviation is used to calculate the homogeneity of a numerical sample (target variable). If the numerical sample (more specific target variable) is completely homogeneous with respect to independent variable (we check for each independent variable separately), then its standard deviation is zero.

Will use the same concept of standard deviation and will check on each split how much reduction in standard deviation is there, and the node (independent variable) which has more reduction in standard deviation that will be declared as a decision node. (In first iteration this node will be called as root node). If not clear at this point, no worries, follow the complete article and I am sure it will be clear.

Reduction in Variance or Standard Deviation Reduction (SDR):

Variance tells us the average distance of all data points from the mean point. Standard deviation is just the square root of the variance. As variance is calculated in squared unit and hence to come up a value having unit equal to the data points, we take square root of the variance and it is called as Standard Deviation.

Formulas:

$$\text{Variance} = \sum_{i=1}^{n}(x_i - \mu)^2/n$$

$$\text{Standard Deviation} = \sqrt{\sum_{i=1}^{n}(x_i - \mu)^2/n} \;\; , \mu \; is \; mean$$

$$\text{Mean} = \sum_{i=1}^{n} x_i \, / \sum_{i=1}^{n} f_i \quad \text{(Sum of all scores / sum of frequencies)}$$

So, what is standard deviation reduction. So basically, first we calculate the standard deviation of the target variable. And then calculate the weighted standard deviation of target with respect to each independent variable. Then take a difference. And this is known as reduction in standard deviation.

Step 1: First will calculate the total standard deviation with respect to target variable:

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

$$Count = n = 14$$

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

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

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

Coefficient of variation is ratio of standard deviation divide by the average value and take the percentage of it. This will be used as the stopping criteria for further split. Will discuss this point at the time of using it so it will be clearer there.

Step 2: will calculate SDR with respect to independent variables and decide on root node, decision nodes and leaf nodes.

<u>Deciding the root node</u>

SDR for Outlook:

Standard Deviation Reduction if we make Outlook as Root Node = 1.66

Outlook column values	Hours Played	Count
	Standard Deviation for respective outlook condition	
rainy	10.87	5
overcast	3.49	4
sunny	7.78	5

Add caption

$$Weighted\ SD_{Hours,Outlook} = P(rainy) * SD_{Hours,Rainy} + P(Overcast) * SD_{Hours,Overcast} + P(Sunny) * SD_{Hours,Sunny}$$

$$SD: Standard\ Deviation$$

$$Weighted\ SD_{Hours,Outlook} = \left(\frac{5}{14}\right) * 10.87 + \left(\frac{4}{14}\right) * 3.49 + \left(\frac{5}{14}\right) * 7.78 = 7.66$$

$$SDR = SD(Hours) - SD(Hours, Outlook)$$

$$SDR: Standard\ Deviation\ Reduction$$

$$SDR = 9.32 - 7.66 = 1.66$$

www.ashutoshtripathi.com

SDR for Temperature: (use same formulars as mentioned above)

Standard Deviation Reduction if we make Temperature as Root Node = 0.39

Temperature	Hours Played	Standard Deviation	Count	Probability = (count/total rows)	Weighted SD = (SD*Probability)
Hot	25	8.954747344	4	4/14 = 0.29	8.95*0.29 = 2.59
Hot	30				
Hot	46				
Hot	44				
Mild	45	7.652160189	6	6/14 = 0.43	7.65*0.43 = 3.28
Mild	35				
Mild	46				
Mild	48				
Mild	52				
Mild	30				
Cool	52	10.51189802	4	4/14 = 0.29	10.51*0.29 = 3.04
Cool	23				
Cool	43				
Cool	38				
Weighted SD(Hours, Temperature) = 2.59+3.28+3.04 = 8.93					
Standard Deviation Reduction = Total SD(Hours) - weighted SD(Hours, Temperature) = 9.32-8.93 = 0.39					

*Though utmost care has been taken while calculation, however focus on concept, not the exact calculated number

www.ashutoshtripathi.com

SDR for Humidity:

Standard Deviation Reduction if we make Humidity as Root Node = 0.09

Humidity	Hours Played	Standard Deviation	Count	Probability = (count/total rows)	Weighted SD = (SD*Probability)
High	25	9.73104996	7	7/14 = 0.5	9.73*0.5 = 4.865
High	30				
High	46				
High	45				
High	35				
High	52				
High	30				
Normal	52	8.734169353	7	7/14 = 0.5	8.73*0.5 = 4.365
Normal	23				
Normal	43				
Normal	38				
Normal	46				
Normal	48				
Normal	44				
Weighted SD(Hours, Humidity) = 4.865+4.365 = 9.23					
Standard Deviation Reduction = Total SD(Hours) - weighted SD(Hours, Humidity) = 9.32-9.23 = 0.09					

*Though utmost care has been taken while calculation, however focus on concept, not the exact calculated number

www.ashutoshtripathi.com

SDR for Windy:

Standard Deviation Reduction if we make Windy as Root Node = 0.39

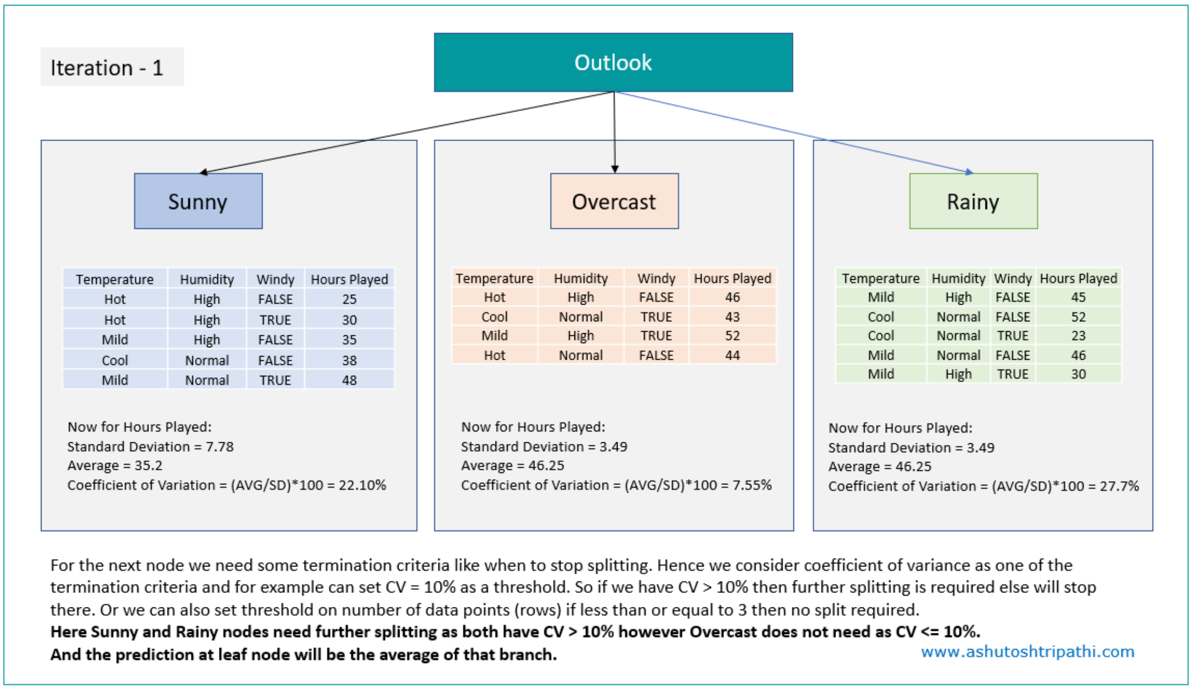
Windy	Hours Played	Standard Deviation	Count	Probability = (count/total rows)	Weighted SD = (SD*Probability)
FALSE	25	7.873015623	8	8/14 = 0.57	7.87*0.57 = 4.48
FALSE	46				
FALSE	45				
FALSE	52				
FALSE	35				
FALSE	38				
FALSE	46				
FALSE	44				
TRUE	30	10.59349905	6	6/14 = 0.42	10.59*0.42 = 4.44
TRUE	23				
TRUE	43				
TRUE	48				
TRUE	52				
TRUE	30				
Weighted SD(Hours, Windy) = 4.48+4.44 = 8.93					
Standard Deviation Reduction = Total SD(Hours) - weighted SD(Hours, Windy) = 9.32-8.93 = 0.39					

*Though utmost care has been taken while calculation, however focus on concept, not the exact calculated number

www.ashutoshtripathi.com

1. SDR (Hours, Outlook) = 1.66
2. SDR (Hours, Temperature) = 0.39
3. SDR (Hours, Humidity) = 0.09
4. SDR (Hours, Windy) = 0.39

Outlook has most reduction in standard deviation hence outlook will be the root node.

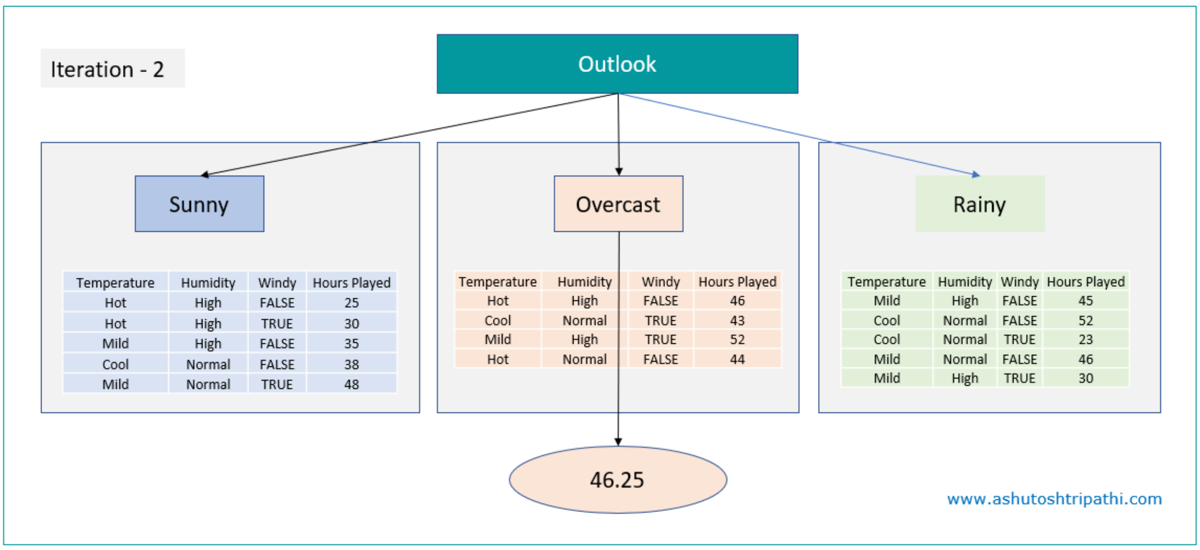


For the next node we need some termination criteria like when to stop splitting. Hence, we consider coefficient of variance as one of the termination criteria and for example can set CV = 10% as a threshold. So, if we have CV > 10% then further splitting is required else will stop there. Or we can also set threshold on number of data points (rows) if less than or equal to 3 then no split required.

And the prediction at leaf node will be the average of that branch.

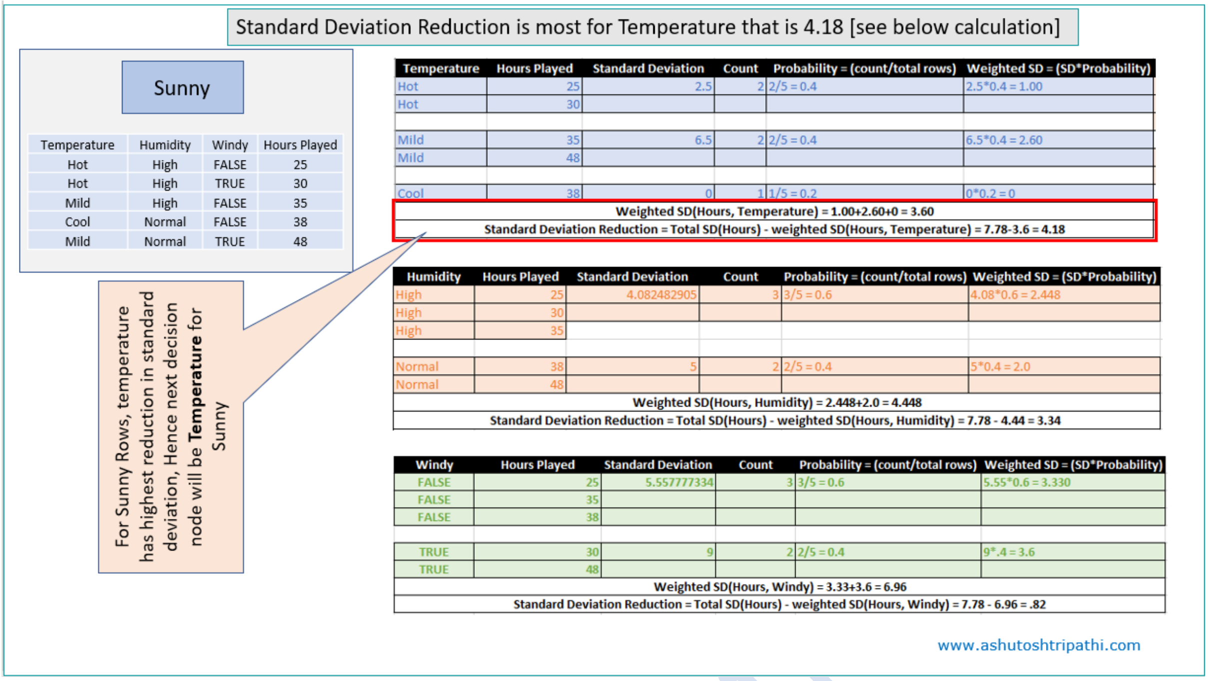
Here Sunny and Rainy nodes need further splitting as both have CV > 10% however Overcast does not need splitting as CV for overcast <= 10%.

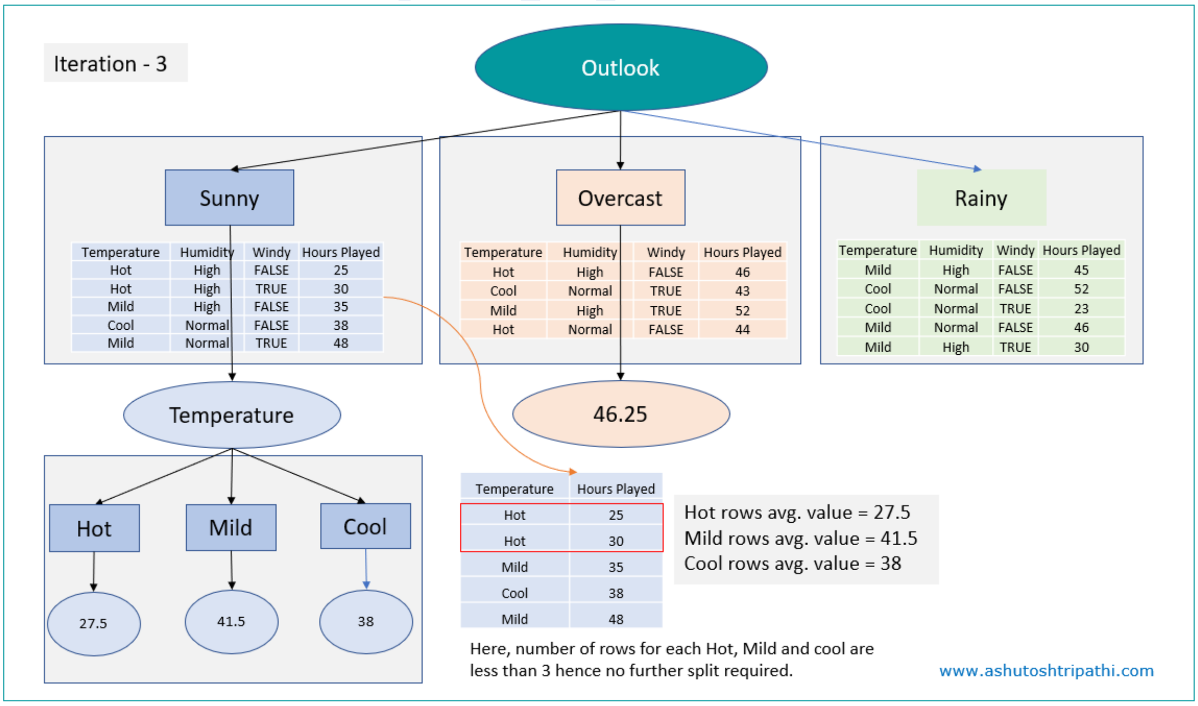
Average value of overcast branch = 46.25 will be the prediction at overcast leaf.



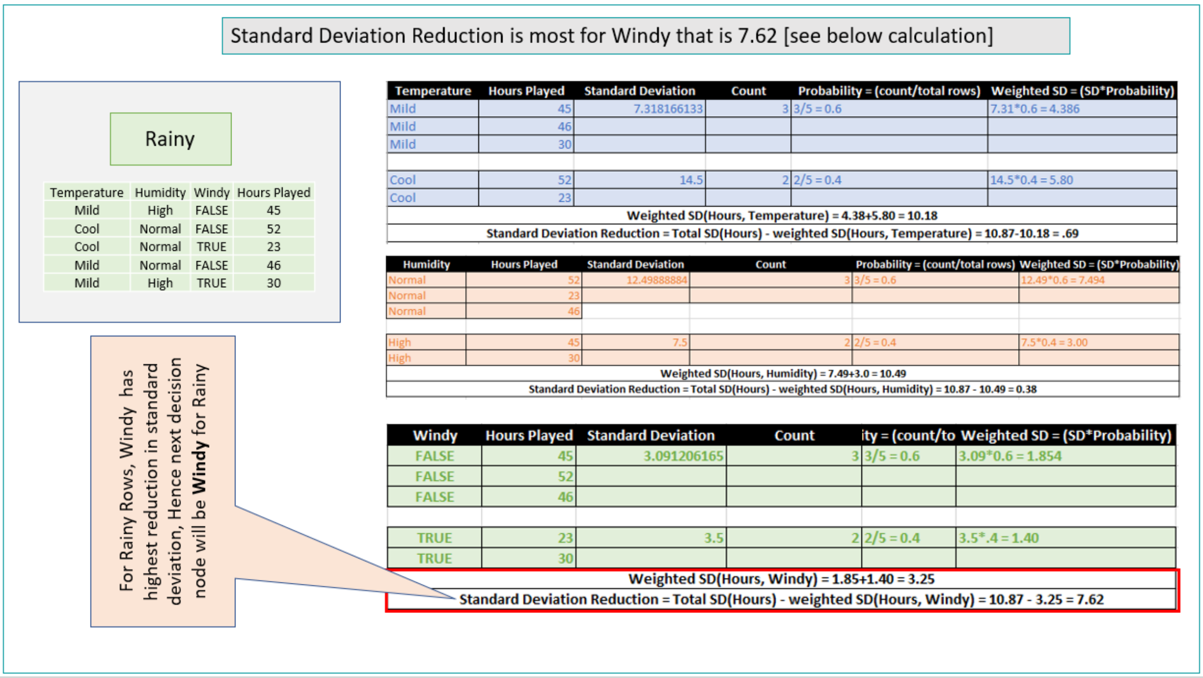
Now will repeat the same SDR check for data filtered for Sunny and Rainy to get other decision nodes and subsequent leaf nodes.

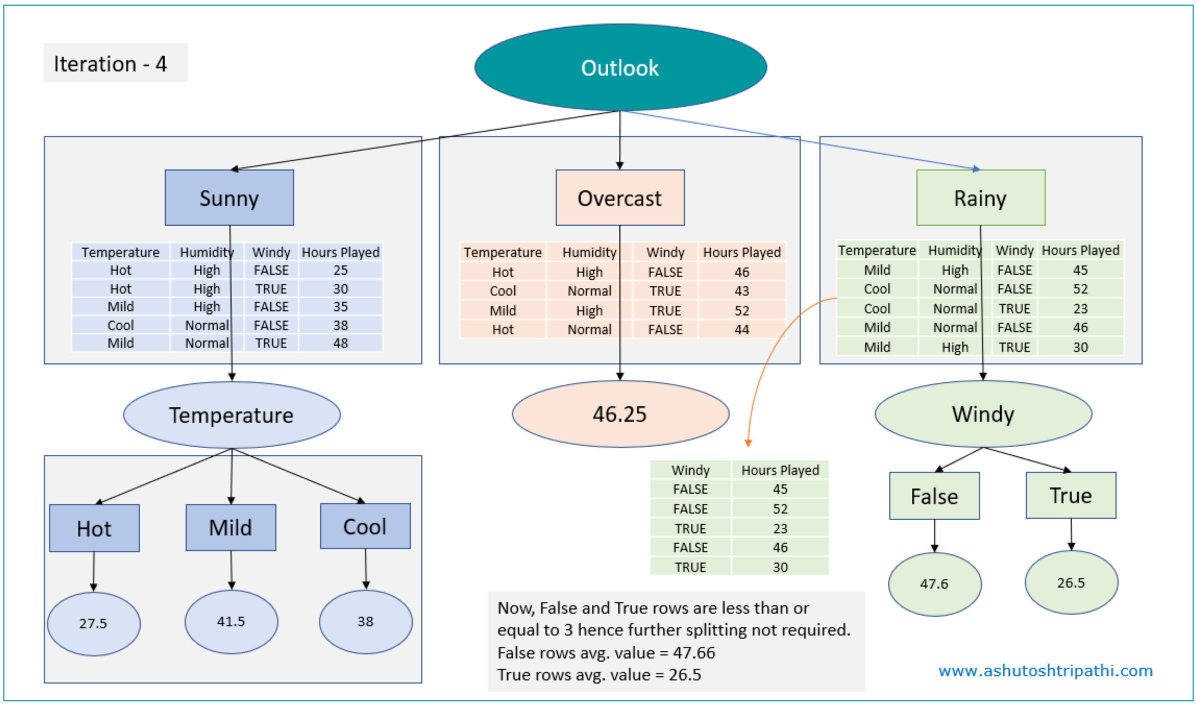
SDR check for Sunny data rows:

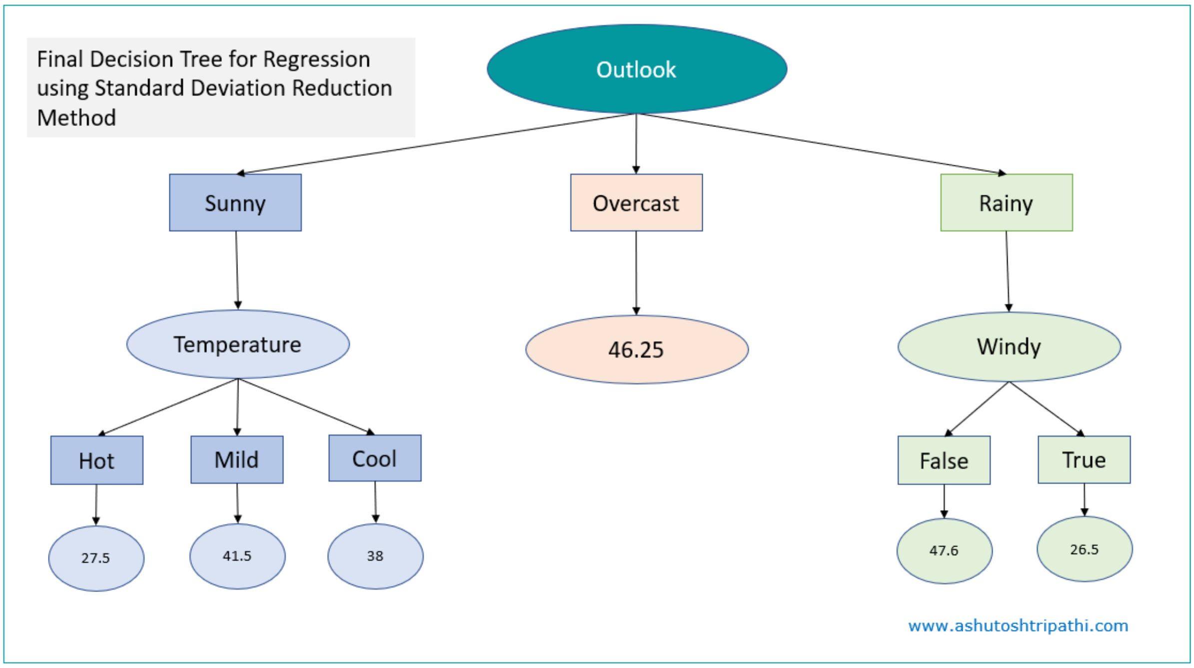




SDR check for Rainy Data rows:







Decision Tree Interpretation:

1. If outlook condition is sunny and temperature is mild then prediction on number of hours match can be played is 41.5 hours irrespective of other conditions.
2. If outlook is overcast then irrespective of other conditions, prediction is 46.25 hours.
3. If outlook is rainy then if it is windy then prediction is 26.5 hours and if it is not windy then prediction is 47.6 hours.

If you have any doubts, then feel free to get in touch with me:

https://ashutoshtripathi.com/contact/

Thank You!!!