Decision Tree Classifier
(CART)

CART Algorithm - Introduction

- CART stands for Classification and Regression Trees.
- It is a term introduced by Leo Breiman to refer to Decision Tree algorithms that can be used for classification or regression predictive modeling problems.
- Like, C4.5 algorithm, CART algorithm can also be used for all type of features i.e. categorical and continuous valued. It can also handle incomplete data.
- For Classification, CART uses Gini index metric to evaluate the goodness of split at each point of the decision tree.

Gini Index

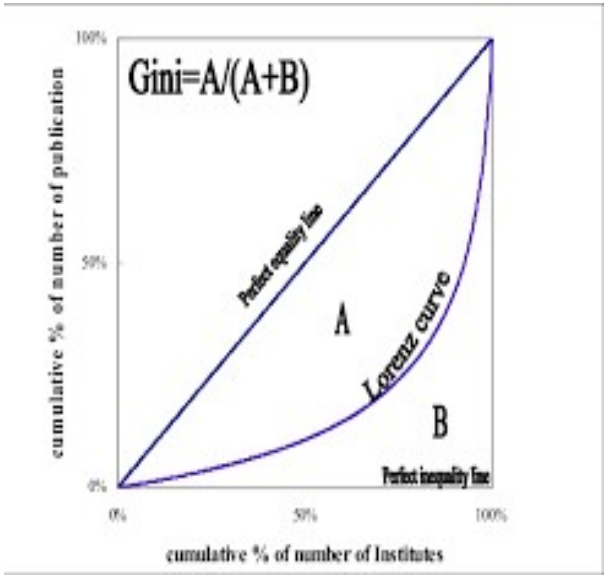
- The concept of Gini Index had been introduced in reference to Economics degree of inequality in a distribution of income/wealth.
- In 1905, Max Lorenz, an American economist a graphical representation of the distribution of income or wealth within a population.
- Lorenz curves graph percentiles of the population against cumulative income or wealth of people at or below that percentile.
- The Lorenz curve is often accompanied by a straight diagonal line with a slope of 1, which represents perfect equality in income or wealth distribution; the Lorenz curve lies beneath it, showing the observed or estimated distribution.

Gini Index (Contd….)

- The area between the straight line and the curved line, expressed as a ratio of the area under the straight line, is the **Gini coefficient**, a scalar measurement of inequality.

$$Gini\ Index = \frac{A}{A + B}$$

- If A=0, i.e. the Lorenz curve coincides with line of perfect inequality then Gini Coefficient is 0.
- If Lorenz curve coincides with x-axis (i.e., line of perfect inequality) then Gini Coefficient is 1 (because A=A+B).
- Therefore, 0 is the ideal value of Gini Coefficient when there is perfect equality and 1 is the worse value of Gini coefficient when there is perfect inequality.
- Hence, smaller is the value of Gini Coefficient, better is the distribution.



Gini Index (Contd.....)

- In Machine Learning, Gini Coefficient measures the inequality between the cumulative feature values and the target values.
- In other words, it measures degree of impurity i.e., the degree to which a feature is misclassified.

$$Gini\ Coefficient(S) = 1 - \sum_{i=1}^{n} p_i^2$$

where p_i is the probability of i^{th} label and n are the number of labels.

$$Gini\ Coefficient(S, A) = \frac{|S_v|}{|S|} * Gini\ Coefficient\ (S_v)$$

where $\mathbf{S_v}$ is the set of rows in **S** for which the feature column **A** has value **v**, $|\mathbf{S_v}|$ is the number of rows in $\mathbf{S_v}$ and likewise $|\mathbf{S}|$ is the number of rows in **S.**

CART Classification Algorithm

1. Check for the base cases (as discussed in ID3 algorithm).
2. For each attribute a, find the Gini Coefficient from splitting on a.
3. Let a_best be the attribute with the lowest Gini Coefficient.
4. Create a decision node that splits on a_best.
5. Recur on the sublists obtained by splitting on a_best, and add those nodes as children of node.

Numerical Example 1

Consider the **weather dataset** in which we have to decide that whether the player should play golf or not on the basis of weather conditions (shown in figure).

Train a decision tree classifier (using CART algorithm) that classifies any new test case according to given weather conditions.

S. No.	Outlook	Temperature	Humidity	Windy	PlayTennis
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rainy	Mild	High	Weak	Yes
5	Rainy	Cool	Normal	Weak	Yes
6	Rainy	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rainy	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rainy	Mild	High	Strong	No

Example 1- Solution

Gini index is a metric for classification tasks in CART. It stores sum of squared probabilities of each class.

$$Ginni\ Index = 1 - \sum_{i=1}^{n} p_i^2$$

So, in the first step, we will compute Gini Index of each attribute.

Example 1- Solution (Contd….)

Gini Index of Outlook

Gini(Outlook=Sunny) = $1 – (2/5)^2 – (3/5)^2 = 1 – 0.16 – 0.36 = 0.48$

Gini(Outlook=Overcast) = $1 – (4/4)^2 – (0/4)^2 = 0$

Gini(Outlook=Rain) = $1 – (3/5)^2 – (2/5)^2 = 1 – 0.36 – 0.16 = 0.48$

Then, we will calculate weighted sum of Gini indexes for outlook feature.

Gini(Outlook) = $(5/14) \times 0.48 + (4/14) \times 0 + (5/14) \times 0.48 = 0.171 + 0 + 0.171 = 0.342$

Outlook	Yes	No	Number of instances
Sunny	2	3	5
Overcast	4	0	4
Rain	3	2	5

Example 1- Solution (Contd….)

Temperature

Similarly, temperature is a nominal feature and it could have 3 different values: Cool, Hot and Mild. Let's summarize decisions for temperature feature.

Temperature	Yes	No	Number of instances
Hot	2	2	4
Cool	3	1	4
Mild	4	2	6

Gini(Temp=Hot) = $1 – (2/4)^2 – (2/4)^2 = 0.5$

Gini(Temp=Cool) = $1 – (3/4)^2 – (1/4)^2 = 1 – 0.5625 – 0.0625 = 0.375$

Gini(Temp=Mild) = $1 – (4/6)^2 – (2/6)^2 = 1 – 0.444 – 0.111 = 0.445$

We'll calculate weighted sum of gini index for temperature feature

Gini(Temp) = $(4/14) \times 0.5 + (4/14) \times 0.375 + (6/14) \times 0.445 = 0.142 + 0.107 + 0.190 = 0.439$

Example 1- Solution (Contd….)

Humidity

Humidity is a binary class feature. It can be high or normal.

Gini(Humidity=High) = $1 – (3/7)^2 – (4/7)^2 = 1 – 0.183 – 0.326 = 0.489$

Gini(Humidity=Normal) = $1 – (6/7)^2 – (1/7)^2 = 1 – 0.734 – 0.02 = 0.244$

Weighted sum for humidity feature will be calculated next

Gini(Humidity) = $(7/14) \times 0.489 + (7/14) \times 0.244 = 0.367$

Humidity	Yes	No	Number of instances
High	3	4	7
Normal	6	1	7

Example 1- Solution (Contd….)

Wind

Wind is a binary class similar to humidity. It can be weak and strong.

Gini(Wind=Weak) = $1 – (6/8)^2 – (2/8)^2 = 1 – 0.5625 – 0.062 = 0.375$

Gini(Wind=Strong) = $1 – (3/6)^2 – (3/6)^2 = 1 – 0.25 – 0.25 = 0.5$

Gini(Wind) = $(8/14) \times 0.375 + (6/14) \times 0.5 = 0.428$

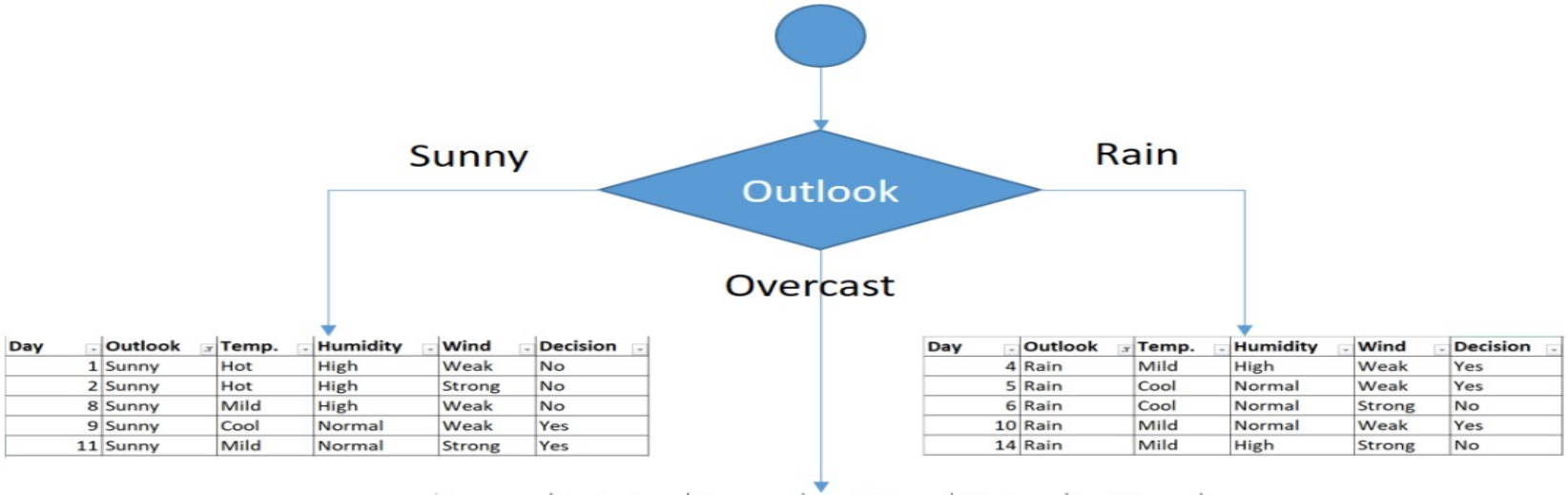
Wind	Yes	No	Number of instances
Weak	6	2	8
Strong	3	3	6

Example 1- Solution (Contd….)

We've calculated Gini index values for each feature. The winner will be outlook feature because its cost is the lowest.

Feature	Gini index
Outlook	0.342
Temperature	0.439
Humidity	0.367
Wind	0.428

Example 1- Solution (Contd….)



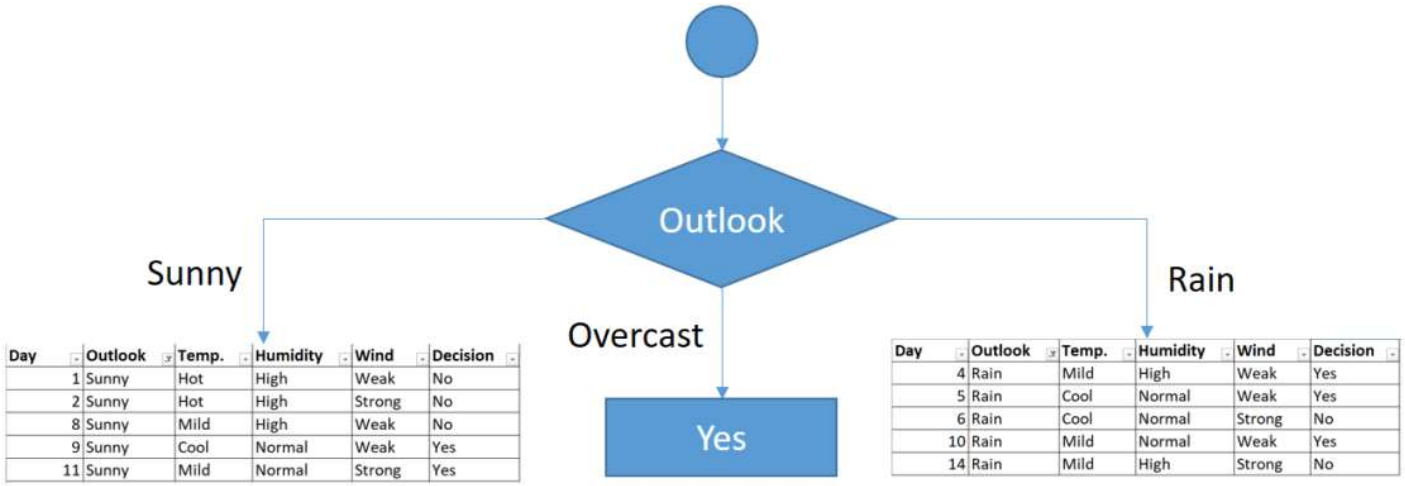
Day	Outlook	Temp.	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes

Day	Outlook	Temp.	Humidity	Wind	Decision
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
10	Rain	Mild	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Day	Outlook	Temp.	Humidity	Wind	Decision
3	Overcast	Hot	High	Weak	Yes
7	Overcast	Cool	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes

Example 1- Solution (Contd….)

The sub dataset in the overcast leaf has only yes decisions. This means that overcast leaf is over.



Day	Outlook	Temp.	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes

Day	Outlook	Temp.	Humidity	Wind	Decision
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
10	Rain	Mild	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Example 1- Solution (Contd….)

Focus on the sub dataset for sunny outlook. We need to find the gini index scores for temperature, humidity and wind features respectively.

Day	Outlook	Temp.	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes

Example 1- Solution (Contd….)

Gini of temperature for sunny outlook

Gini(Outlook=Sunny and Temp.=Hot) = $1 – (0/2)^2 – (2/2)^2 = 0$

Gini(Outlook=Sunny and Temp.=Cool) = $1 – (1/1)^2 – (0/1)^2 = 0$

Gini(Outlook=Sunny and Temp.=Mild) = $1 – (1/2)^2 – (1/2)^2 = 1 – 0.25 – 0.25 = 0.5$

Gini(Outlook=Sunny and Temp.) = $(2/5)x0 + (1/5)x0 + (2/5)x0.5 = 0.2$

Temperature	Yes	No	Number of instances
Hot	0	2	2
Cool	1	0	1
Mild	1	1	2

Example 1- Solution (Contd….)

Gini of humidity for sunny outlook

Gini(Outlook=Sunny and Humidity=High) = $1 – (0/3)^2 – (3/3)^2 = 0$

Gini(Outlook=Sunny and Humidity=Normal) = $1 – (2/2)^2 – (0/2)^2 = 0$

Gini(Outlook=Sunny and Humidity) = $(3/5)x0 + (2/5)x0 = 0$

Humidity	Yes	No	Number of instances
High	0	3	3
Normal	2	0	2

Example 1- Solution (Contd….)

Gini of wind for sunny outlook

Gini(Outlook=Sunny and Wind=Weak) = $1 - (1/3)^2 - (2/3)^2 = 0.266$

Gini(Outlook=Sunny and Wind=Strong) = $1- (1/2)^2 - (1/2)^2 = 0.2$

Gini(Outlook=Sunny and Wind) = $(3/5)x0.266 + (2/5)x0.2 = 0.466$

Wind	Yes	No	Number of instances
Weak	1	2	3
Strong	1	1	2

Example 1- Solution (Contd….)

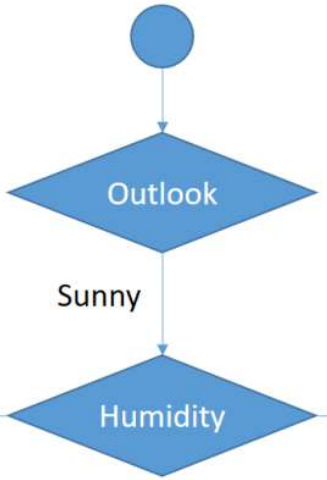
Decision for sunny outlook

We've calculated gini index scores for feature when outlook is sunny. The winner is humidity because it has the lowest value.

Feature	Gini index
Temperature	0.2
Humidity	0
Wind	0.466

Example 1- Solution (Contd….)

We'll put humidity check at the extension of sunny outlook.

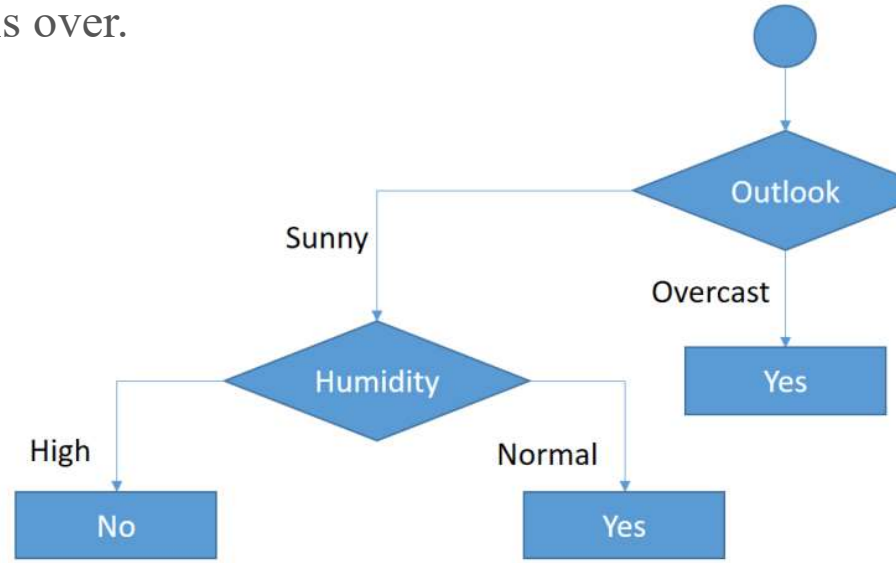


Day	Outlook	Temp.	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
8	Sunny	Mild	High	Weak	No

Day	Outlook	Temp.	Humidity	Wind	Decision
9	Sunny	Cool	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes

Example 1- Solution (Contd….)

As seen, decision is always no for high humidity and sunny outlook. On the other hand, decision will always be yes for normal humidity and sunny outlook. This branch is over.



Day	Outlook	Temp.	Humidity	Wind	Decision
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
10	Rain	Mild	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Example 1- Solution (Contd….)

Now, we need to focus on rain outlook.

Rain outlook

Day	Outlook	Temp.	Humidity	Wind	Decision
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
10	Rain	Mild	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Example 1- Solution (Contd….)

We'll calculate Gini index scores for temperature, humidity and wind features when outlook is rain.

Gini of temperature for rain outlook

Gini(Outlook=Rain and Temp.=Cool) = $1 - (1/2)^2 - (1/2)^2 = 0.5$

Gini(Outlook=Rain and Temp.=Mild) = $1 - (2/3)^2 - (1/3)^2 = 0.444$

Gini(Outlook=Rain and Temp.) = $(2/5)x0.5 + (3/5)x0.444 = 0.466$

Temperature	Yes	No	Number of instances
Cool	1	1	2
Mild	2	1	3

Example 1- Solution (Contd….)

Gini of humidity for rain outlook

Gini(Outlook=Rain and Humidity=High) = $1 - (1/2)^2 - (1/2)^2 = 0.5$

Gini(Outlook=Rain and Humidity=Normal) = $1 - (2/3)^2 - (1/3)^2 = 0.444$

Gini(Outlook=Rain and Humidity) = $(2/5)x0.5 + (3/5)x0.444 = 0.466$

Humidity	Yes	No	Number of instances
High	1	1	2
Normal	2	1	3

Example 1- Solution (Contd….)

Gini of wind for rain outlook

Gini(Outlook=Rain and Wind=Weak) = $1 - (3/3)^2 - (0/3)^2 = 0$

Gini(Outlook=Rain and Wind=Strong) = $1 - (0/2)^2 - (2/2)^2 = 0$

Gini(Outlook=Rain and Wind) = $(3/5)x0 + (2/5)x0 = 0$

Wind	Yes	No	Number of instances
Weak	3	0	3
Strong	0	2	2

Example 1- Solution (Contd….)

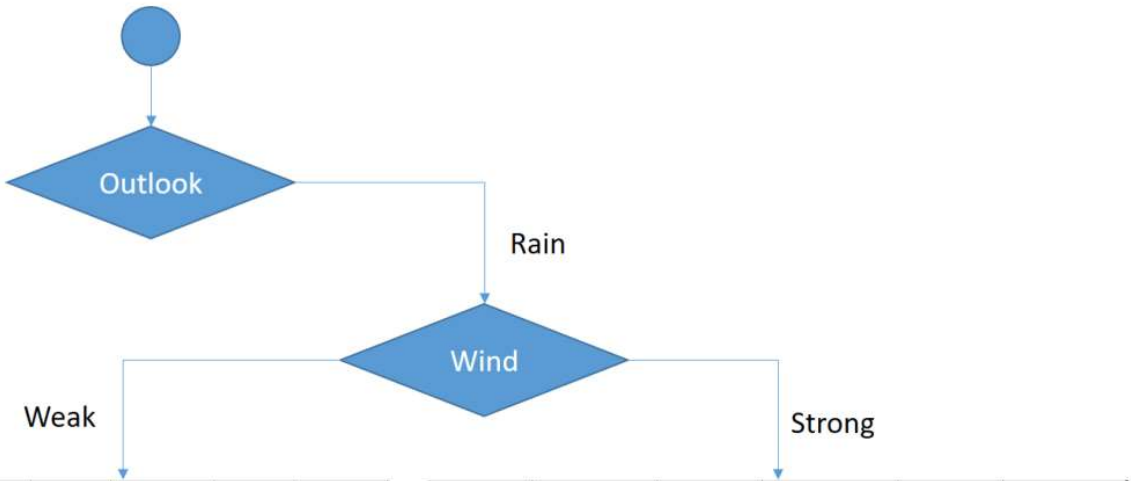
Decision for rain outlook

The winner is wind feature for rain outlook because it has the minimum gini index score in features.

Feature	Gini index
Temperature	0.466
Humidity	0.466
Wind	0

Example 1- Solution (Contd….)

Put the wind feature for rain outlook branch and monitor the new sub data sets.



Day	Outlook	Temp.	Humidity	Wind	Decision
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes

Day	Outlook	Temp.	Humidity	Wind	Decision
6	Rain	Cool	Normal	Strong	No
14	Rain	Mild	High	Strong	No

Example 1- Solution (Contd….)