DECISION TREE

- **Cl**assification and **R**egression **T**ree (CART)

Recommendation System - 1

Gender	Occupation	App
F	Study	🔴
F	Work	🟢
M	Work	👻
F	Work	🟢
M	Study	🔴
M	Study	🔴

Quiz: Woman, works at an office. What app do we recommend?

○ 🔴 Pokémon Go

● 🟢 WhatsApp

○ 👻 Snapchat

Recommendation System - 2

Gender	Occupation	App
F	Study	Pokémon Go
F	Work	WhatsApp
M	Work	Snapchat
F	Work	WhatsApp
M	Study	Pokémon Go
M	Study	Pokémon Go

Quiz: Man, works at a factory. What app do we recommend?

- ○ Pokémon Go
- ○ WhatsApp
- ○ Snapchat

Gender	Occupation	App
F	Study	🔴 Pokémon Go
F	Work	🟢 WhatsApp
M	Work	👻 Snapchat
F	Work	🟢 WhatsApp
M	Study	🔴 Pokémon Go
M	Study	🔴 Pokémon Go

Quiz: Man, works at a factory. What app do we recommend?

- ○ 🔴 Pokémon Go
- ○ 🟢 WhatsApp
- ● 👻 Snapchat

Recommendation System - 3

Gender	Occupation	App
F	Study	Pokémon Go
F	Work	WhatsApp
M	Work	Snapchat
F	Work	WhatsApp
M	Study	Pokémon Go
M	Study	Pokémon Go

Quiz: Girl, goes to high school. What app do we recommend?

○ Pokémon Go
○ WhatsApp
○ Snapchat

Gender	Occupation	App
F	Study	Pokémon Go
F	Work	WhatsApp
M	Work	Snapchat
F	Work	WhatsApp
M	Study	Pokémon Go
M	Study	Pokémon Go

Quiz: Girl, goes to high school. What app do we recommend?

- ● Pokémon Go
- ○ WhatsApp
- ○ Snapchat

Way Machine approaches

Gender	Occupation	App
F	Study	
F	Work	
M	Work	
F	Work	
M	Study	
M	Study	

Gender	Occupation	App
F	Study	
F	Work	
M	Work	
F	Work	
M	Study	
M	Study	

Gender	Occupation	App
F	Study	(Pokéball)
F	Work	(WhatsApp)
M	Work	(Snapchat)
F	Work	(WhatsApp)
M	Study	(Pokéball)
M	Study	(Pokéball)

Gender	Occupation	App
F	Study	(Pokéball)
F	Work	(WhatsApp)
M	Work	(Snapchat)
F	Work	(WhatsApp)
M	Study	(Pokéball)
M	Study	(Pokéball)

Quiz: Between **Gender** and **Occupation**, which one seems more decisive for predicting what app will the users download?

○ Gender

● Occupation

Construction of a Tree

Continuous Data

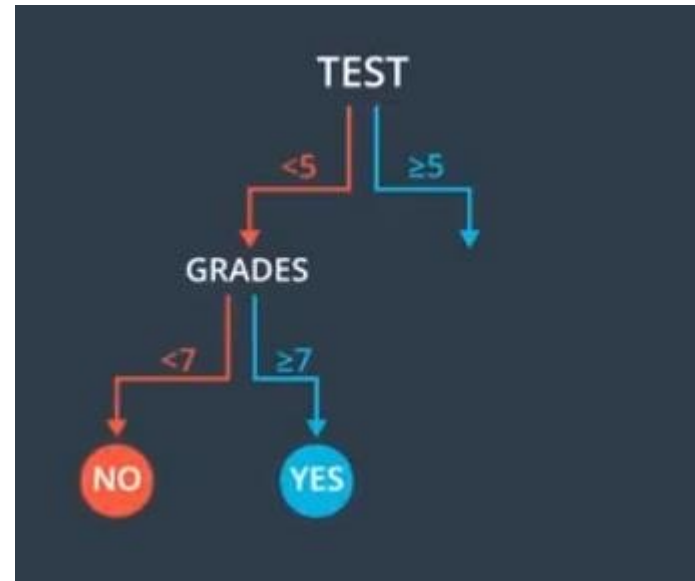**Quiz:** Between grades and test, which one determines student acceptance better?

Or

Quiz: Between a horizontal and a vertical line, which one would cut the data better?

○ Horizontal

○ Vertical
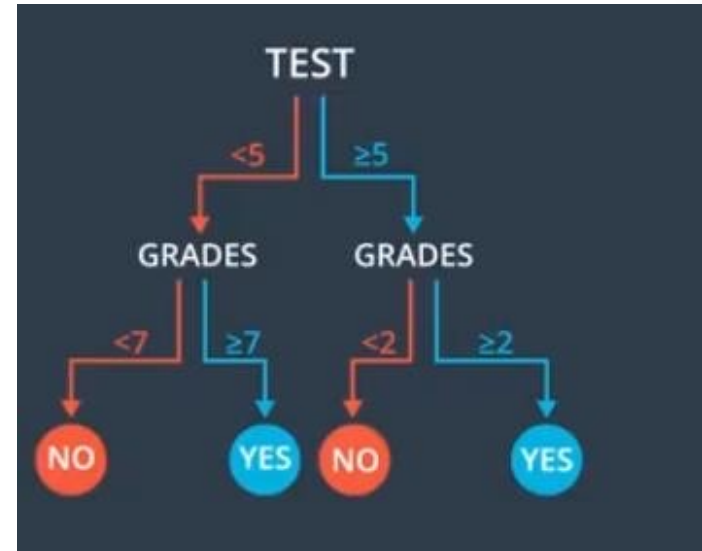
Horizontal vs Vertical

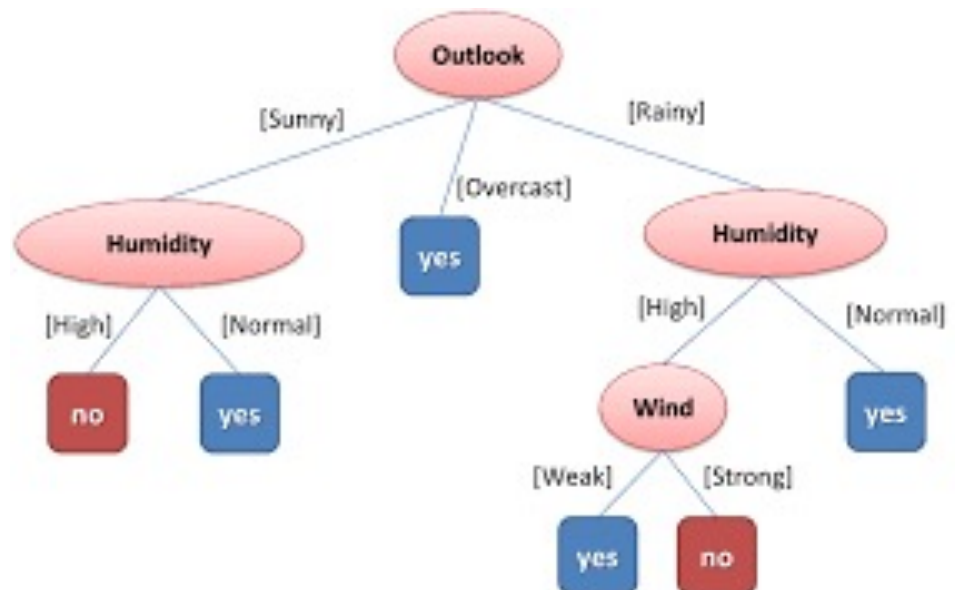# Construction of a Tree

Decision Tree – Manual Structure

Supervised learning algorithm

Root Node

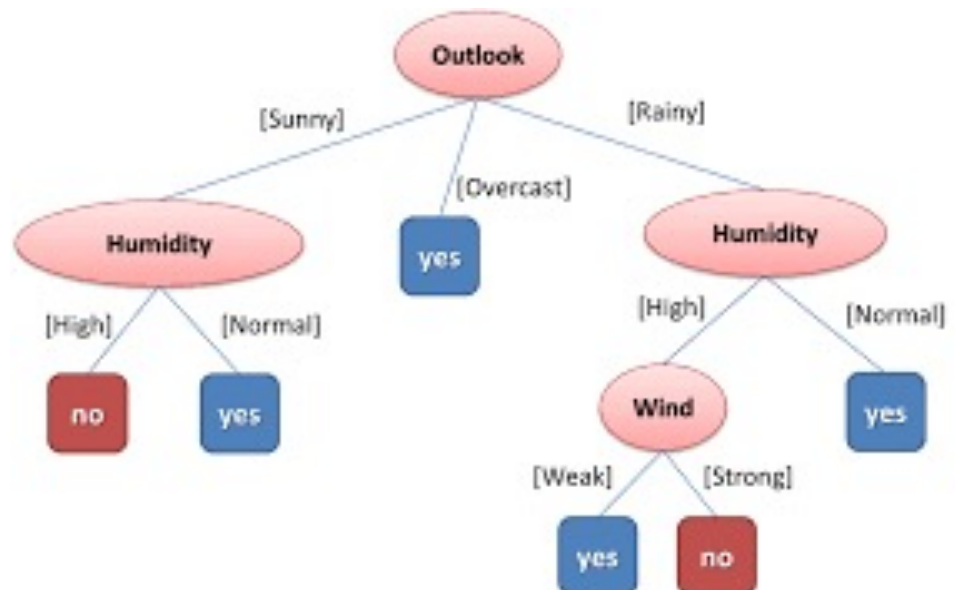**Decision node**

Leaves

Structure of a Tree

Supervised learning algorithm

Root Node - Outlook

Decision node - Humidity/Wind

Leaves - Yes/No

Structure of a Tree

HOW DECISION TREE ALGORITHM WORKS

HOW TO FIND ROOT (2 WAYS)

- **Information gain**
- **Gini index**

Information Gain & Entropy

Information Gain -> Information theory -> Entropy

Entropy = **Randomness** or **Uncertainty** of a random variable.

There are **2 steps for calculating information gain** for each attribute:
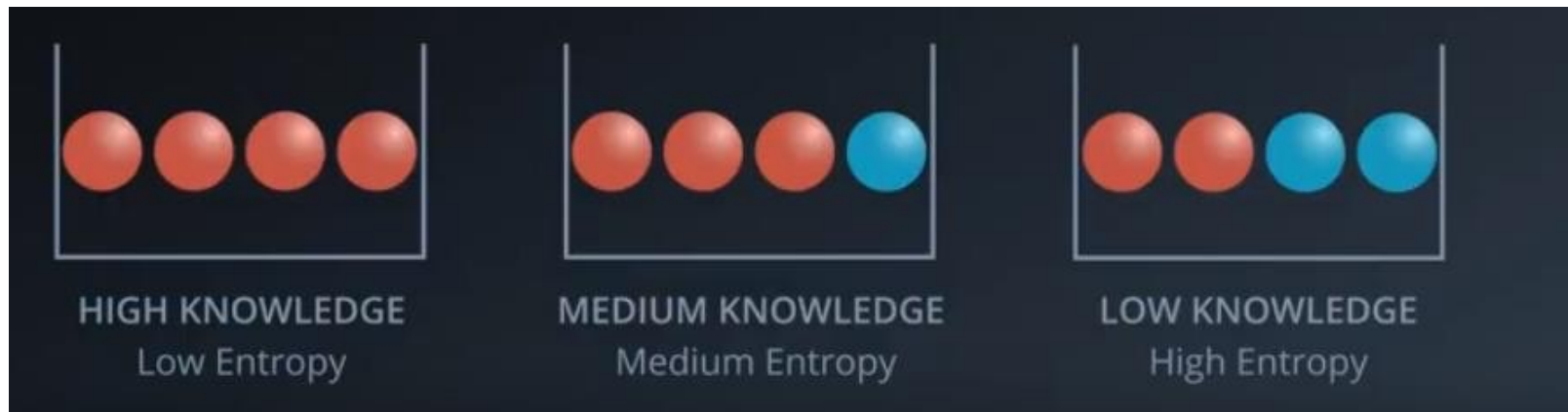
➢ Calculate entropy of Target.

➢ Calculate the Entropy for every attribute.

Information gain = Entropy of target - Entropy of attribute

Entropy - The measure of uncertainty

Entropy - The measure of uncertainty

HIGH KNOWLEDGE
Low Entropy

MEDIUM KNOWLEDGE
Medium Entropy
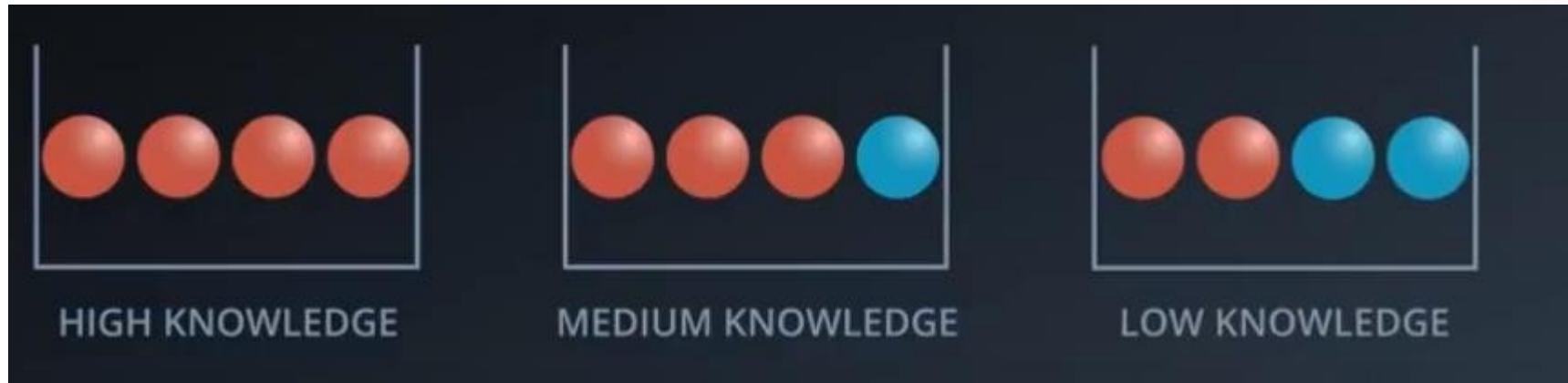
LOW KNOWLEDGE
High Entropy

$$H(X) = \mathbb{E}_X[I(x)] = -\sum_{x \in \mathbb{X}} p(x) \log p(x).$$

Case Study – Golf Play Dataset

Predictors   Target

Outlook	Temp.	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

Entropy of Target

Play Golf

Play Golf
No
No
Yes
Yes
Yes
No
Yes
No
Yes
Yes
Yes
Yes
Yes
No

Sort →

Play Golf
No
No
No
No
No
Yes
Yes
Yes
Yes
Yes
Yes
Yes
Yes
Yes

→ $5 / 14 = 0.36$

→ $9 / 14 = 0.64$

Entropy(PlayGolf) = Entropy (5,9)

= Entropy (0.36, 0.64)

= $- (0.36 \log_2 0.36) - (0.64 \log_2 0.64)$

= 0.94

Activate
Go to PC set

Frequency Table – 4 Attributes

		Play Golf	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3

		Play Golf	
		Yes	No
Temp.	Hot	2	2
	Mild	4	2
	Cool	3	1

		Play Golf	
		Yes	No
Humidity	High	3	4
	Normal	6	1

		Play Golf	
		Yes	No
Windy	False	6	2
	True	3	3

Entropy - Outlook

		Play Golf		
		Yes	No	
	Sunny	3	2	5
Outlook	Overcast	4	0	4
	Rainy	2	3	5
				14

E(PlayGolf, Outlook) = **P**(Sunny)***E**(3,2) + **P**(Overcast)***E**(4,0) + **P**(Rainy)***E**(2,3)

$= (5/14)*0.971 + (4/14)*0.0 + (5/14)*0.971$

$= 0.693$

Information Gain - Outlook

$$\mathbf{G}(\text{PlayGolf, Outlook}) = \mathbf{E}(\text{PlayGolf}) - \mathbf{E}(\text{PlayGolf, Outlook})$$

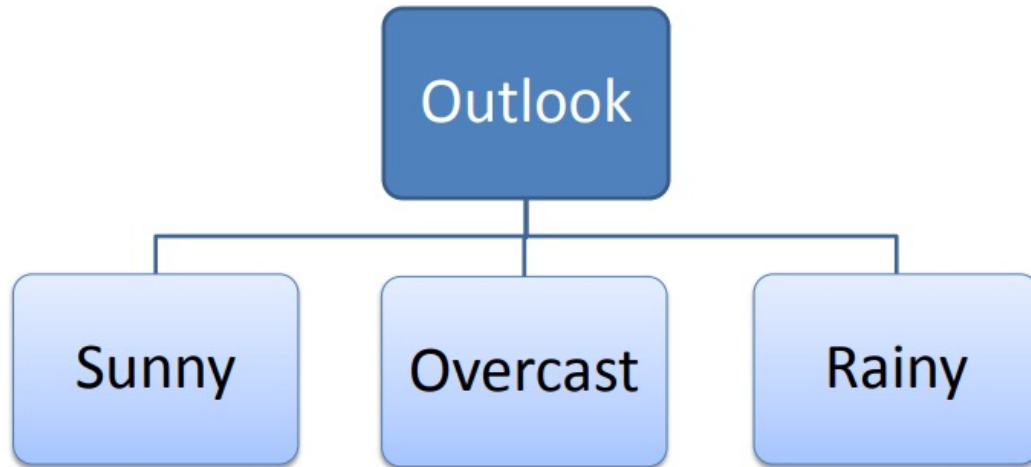$$= 0.940 - 0.693 = 0.247$$

Information Gain - All Attributes

Outlook		Play Golf	
		Yes	No
	Sunny	3	2
	Overcast	4	0
	Rainy	2	3
Gain = 0.247			

Temp.		Play Golf	
		Yes	No
	Hot	2	2
	Mild	4	2
	Cool	3	1
Gain = 0.029			

Humidity		Play Golf	
		Yes	No
	High	3	4
	Normal	6	1
Gain = 0.152			

Windy		Play Golf	
		Yes	No
	False	6	2
	True	3	3
Gain = 0.048			

Construction of Tree

Outlook	Temp.	Humidity	Windy	Play Golf
Sunny	Mild	High	FALSE	Yes
Sunny	Cool	Normal	FALSE	Yes
Sunny	Cool	Normal	TRUE	No
Sunny	Mild	Normal	FALSE	Yes
Sunny	Mild	High	TRUE	No

Rainy	Hot	High	FALSE	No
Rainy	Hot	High	TRUE	No
Rainy	Mild	High	FALSE	No
Rainy	Cool	Normal	FALSE	Yes
Rainy	Mild	Normal	TRUE	Yes

Overcast	Hot	High	FALSE	Yes
Overcast	Cool	Normal	TRUE	Yes
Overcast	Mild	High	TRUE	Yes
Overcast	Hot	Normal	FALSE	Yes

Overcast

Temp	Humidity	Windy	Play Golf
Hot	High	FALSE	Yes
Cool	Normal	TRUE	Yes
Mild	High	TRUE	Yes
Hot	Normal	FALSE	Yes
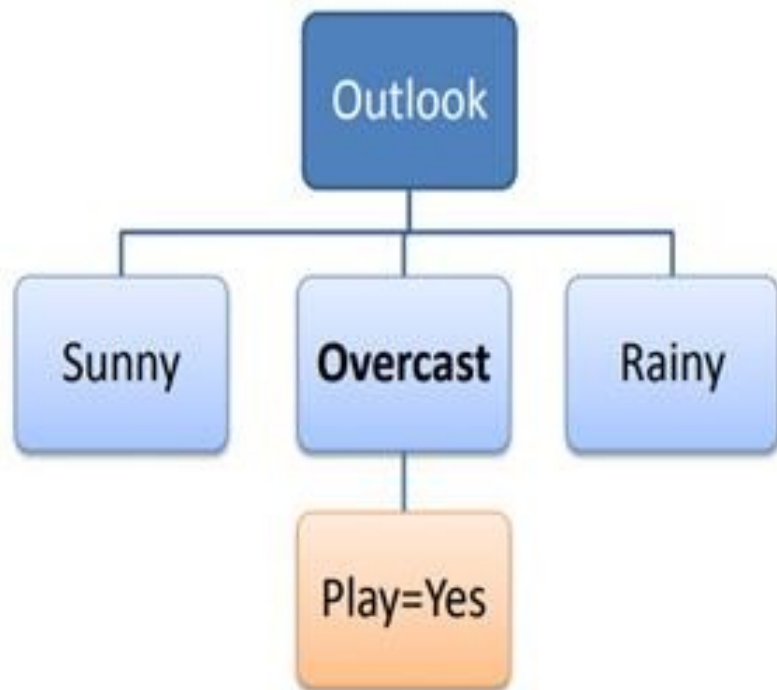
Sunny

Temp.	Humidity	Windy	Play Golf
Mild	High	FALSE	Yes
Cool	Normal	FALSE	Yes
Cool	Normal	TRUE	No
Mild	Normal	FALSE	Yes
Mild	High	TRUE	No

		Play Golf	
		Yes	No
Temp.	Mild	2	1
	Cool	1	1
Gain = 0.02			

		Play Golf	
		Yes	No
Humidity	High	1	1
	Normal	2	1
Gain = 0.02			

		Play Golf	
★		Yes	No
Windy	False	3	0
	True	0	2
Gain = 0.97			

Construction of Tree

Temp.	Humidity	Windy	Play Golf
Mild	High	FALSE	Yes
Cool	Normal	FALSE	Yes
Mild	Normal	FALSE	Yes
Cool	Normal	TRUE	No
Mild	High	TRUE	No

Outlook
- Sunny
  - Windy
    - FALSE → Play=Yes
    - TRUE → Play=No
- Overcast → Play=Yes
- Rainy
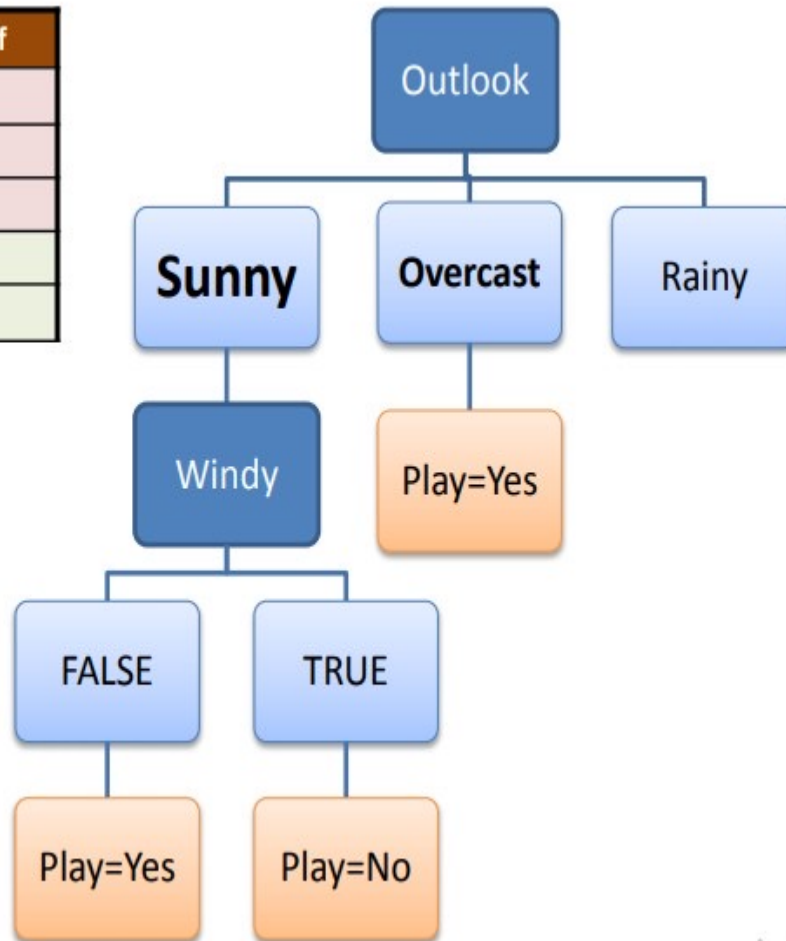
Rainy

Temp.	Humidity	Windy	Play Golf
Hot	High	FALSE	No
Hot	High	TRUE	No
Mild	High	FALSE	No
Cool	Normal	FALSE	Yes
Mild	Normal	TRUE	Yes

Temp.		Play Golf	
		Yes	No
	Hot	0	2
	Mild	1	1
	Cool	1	0
Gain = 0.57			

⭐		Play Golf	
		Yes	No
Humidity	High	0	3
	Normal	2	0
Gain = 0.97			

Windy		Play Golf	
		Yes	No
	False	1	2
	True	1	1
Gain = 0.02			

Final Tree Structure
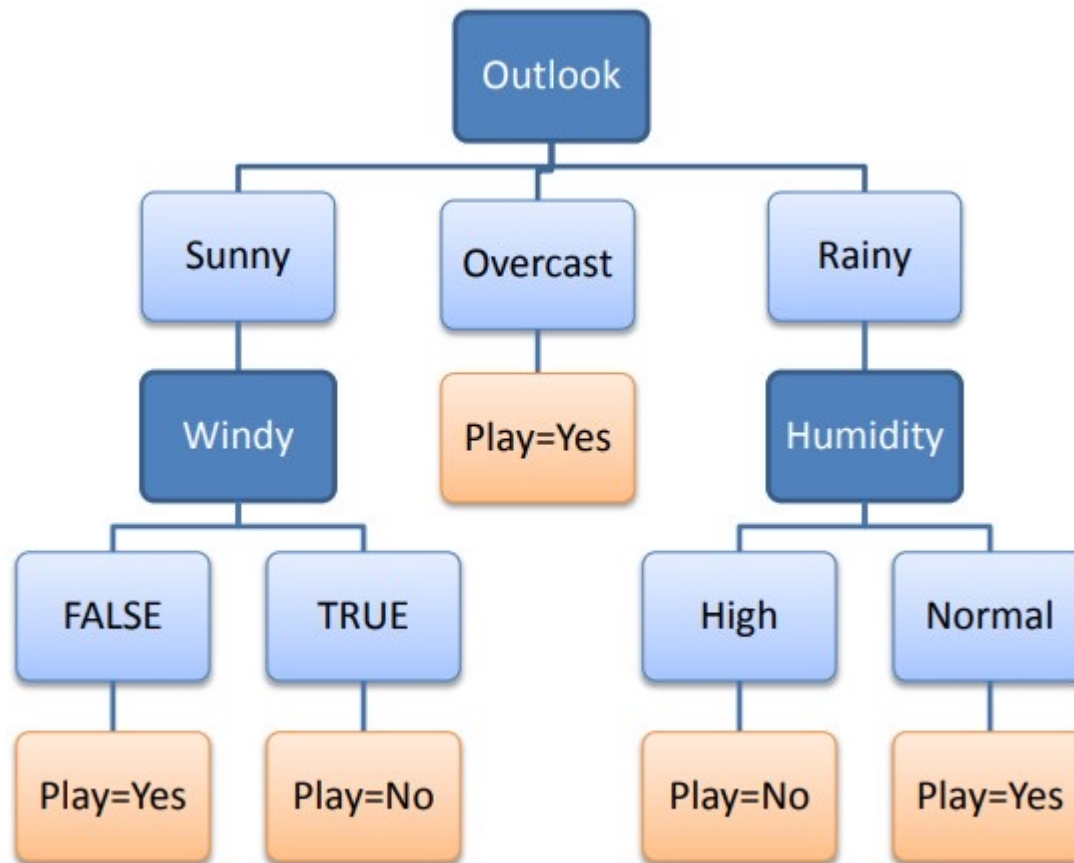
Predict the Play – D17 ?

Outlook	Temp	Humidity	Windy	Play Golf
Sunny	Cool	Normal	FALSE	?

Predict the Play – D17 ?

Outlook	Temp	Humidity	Windy	Play Golf
Sunny	Cool	Normal	FALSE	Yes
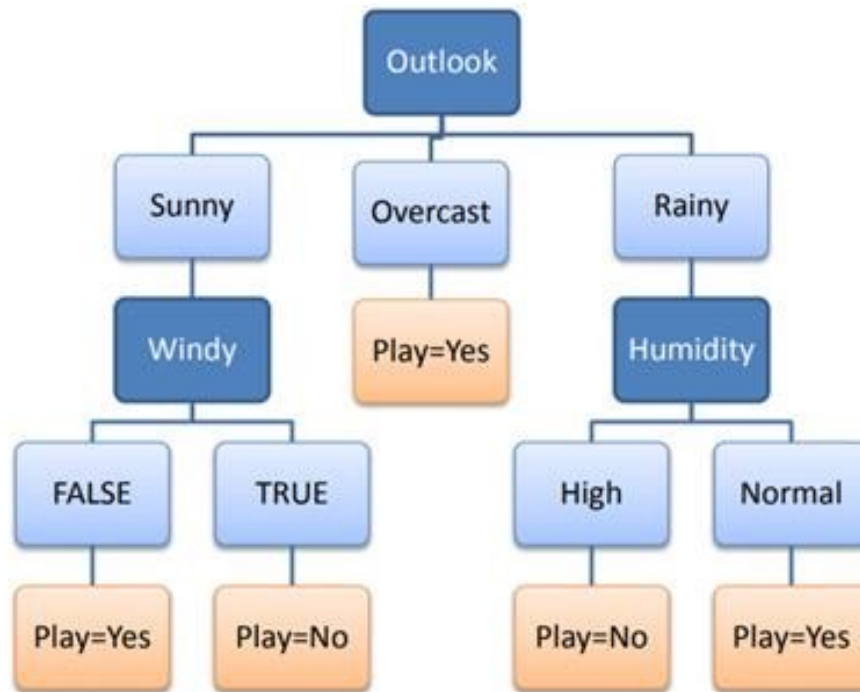
Decision Rules – Traditional approach

R₁: **IF** (Outlook=Sunny) AND (Windy=FALSE) **THEN** Play=Yes

R₂: **IF** (Outlook=Sunny) AND (Windy=TRUE) **THEN** Play=No

R₃: **IF** (Outlook=Overcast) **THEN** Play=Yes

R₄: **IF** (Outlook=Rainy) AND (Humidity=High) **THEN** Play=No

R₅: **IF** (Outlook=Rain) AND (Humidity=Normal) **THEN** Play=Yes
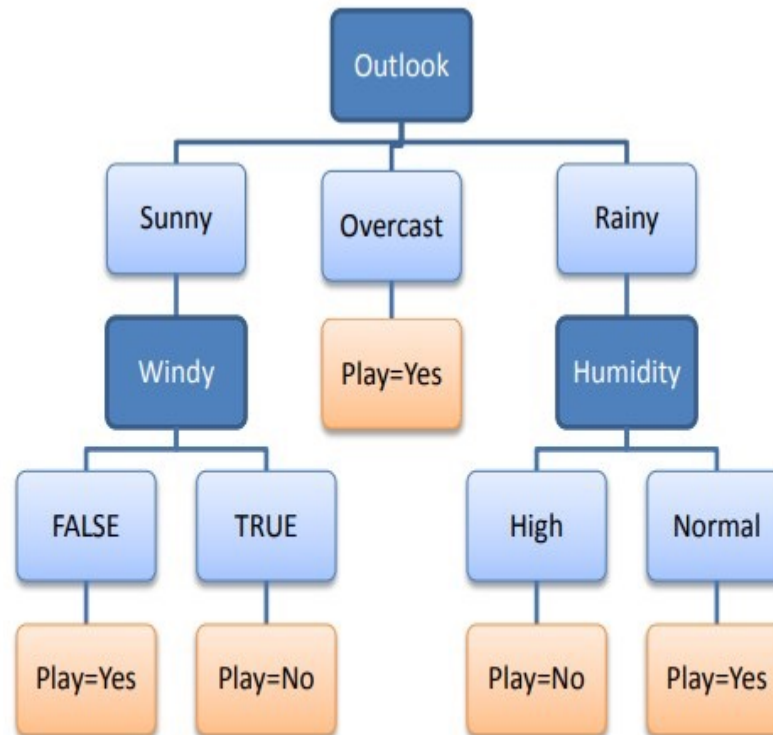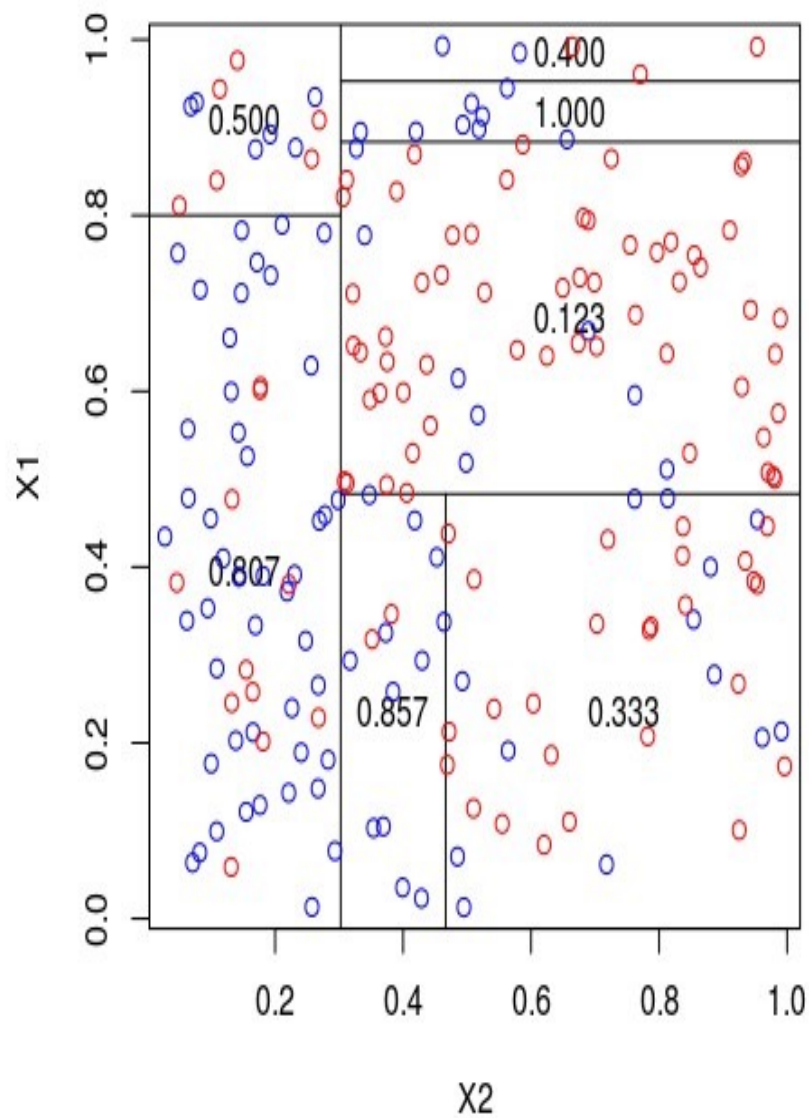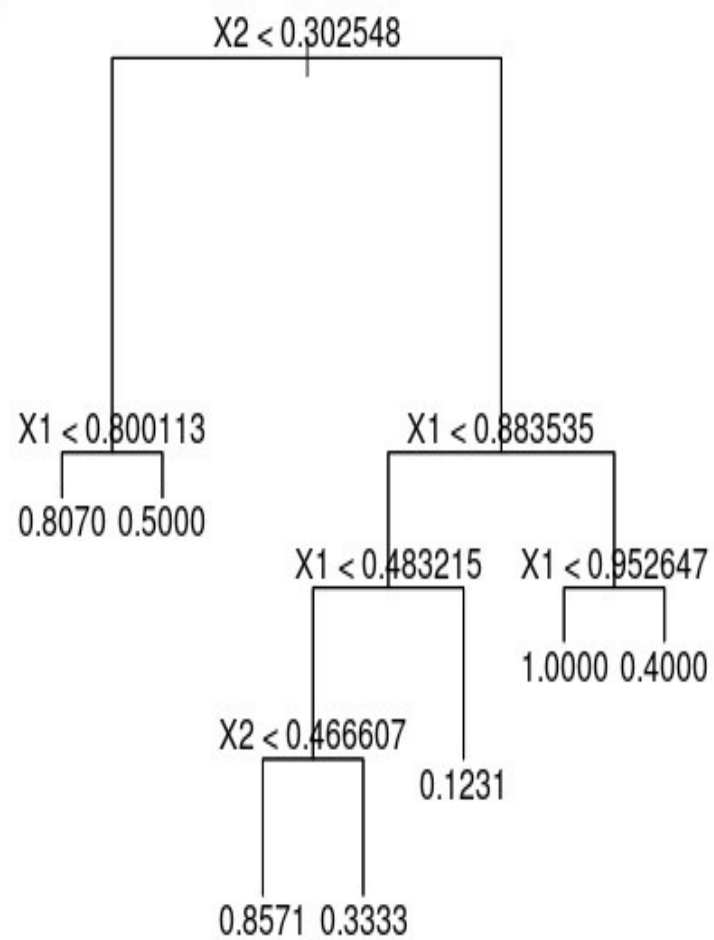
Finding Root using Gini Index

$$Gini\ Index = 1 - \sum_j p_j^2$$

1. The steps to build the tree using **Gini Index** approach is same as the Entropy with the only change in the Formula.

2. In Gini the attribute with the lowest Gini score is used as the ROOT

3. Gini Index is the default method of building the Decision Tree
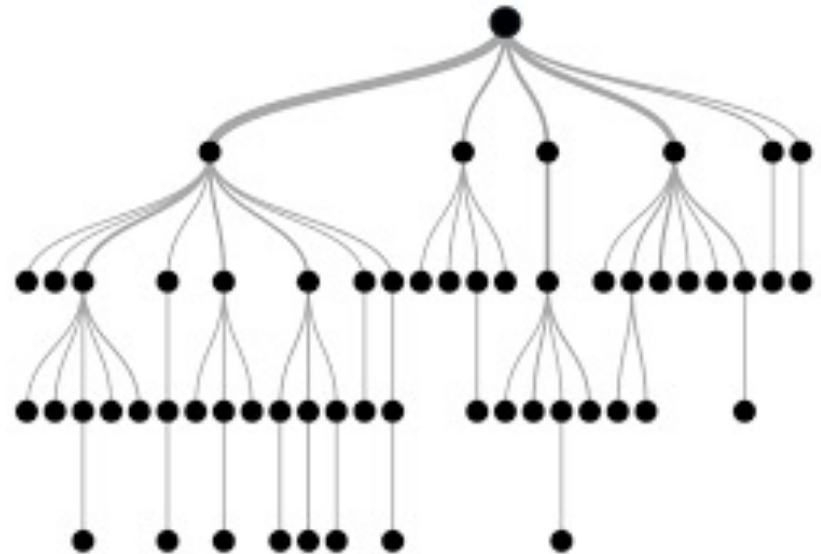
Disadvantage on using Continuous data
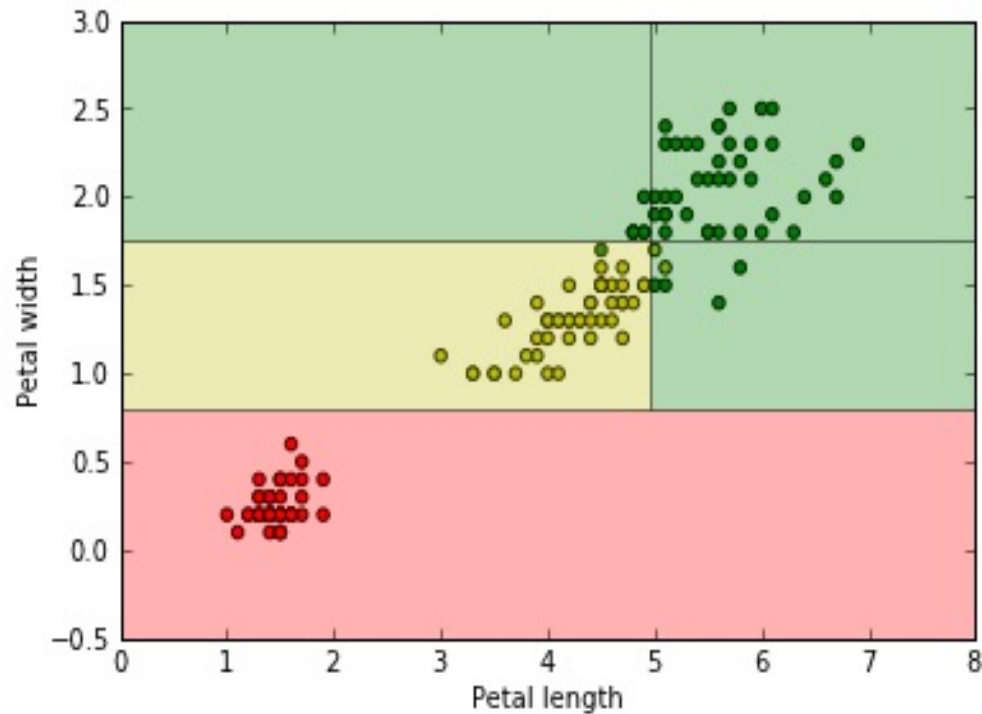
When to stop splitting?
Overfitting

How to overcome Overfitting?

Pruning

1. Pre-pruning
2. Post-pruning

Classification vs Regression Tree