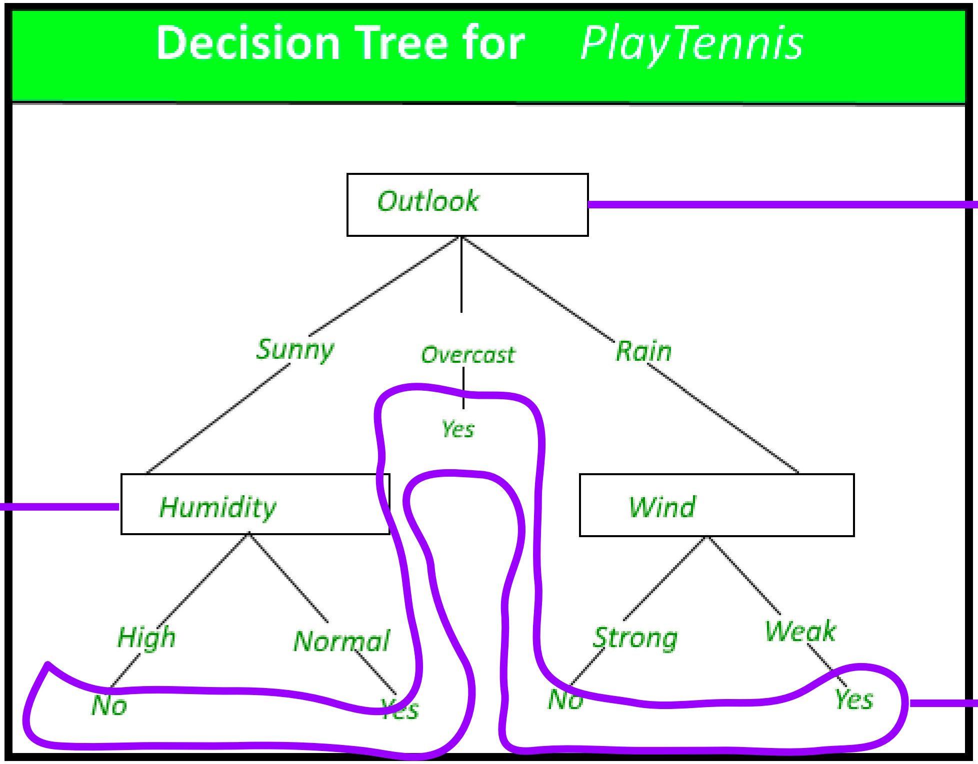
Decision Tree

Suresh Kumar Choudhary

Assistant Professor,
Department of Data Science & Analytics
Central University of Rjasthan
sureshdewenda@gmail.com

What is Decision Tree?

- **Decision tree is a N-ary Tree, that means it can have at most N childs**



Root Node

Intermediate or Decision Nodes

Leaf Nodes

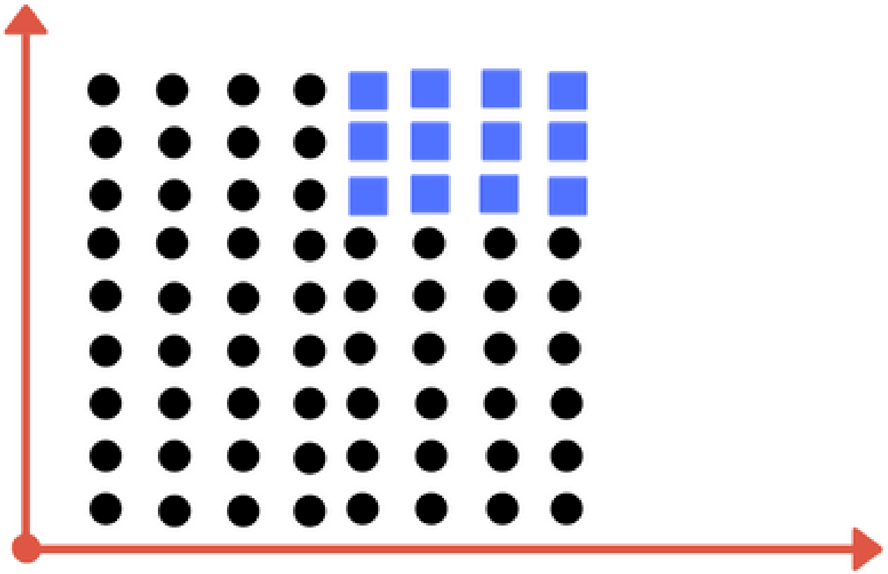
Note: 1. Edges of the tree respresnt the attribute values.
2. Leaf Nodes contains the class information.
3. Intermediate Nodes are the features/attributes of the Input Dataset

Decision Tree Definition

- **A decision tree is a tree where each node represents a feature(attribute), each link(branch) represents a decision(rule) and each leaf represents an outcome(categorical or continues value).**

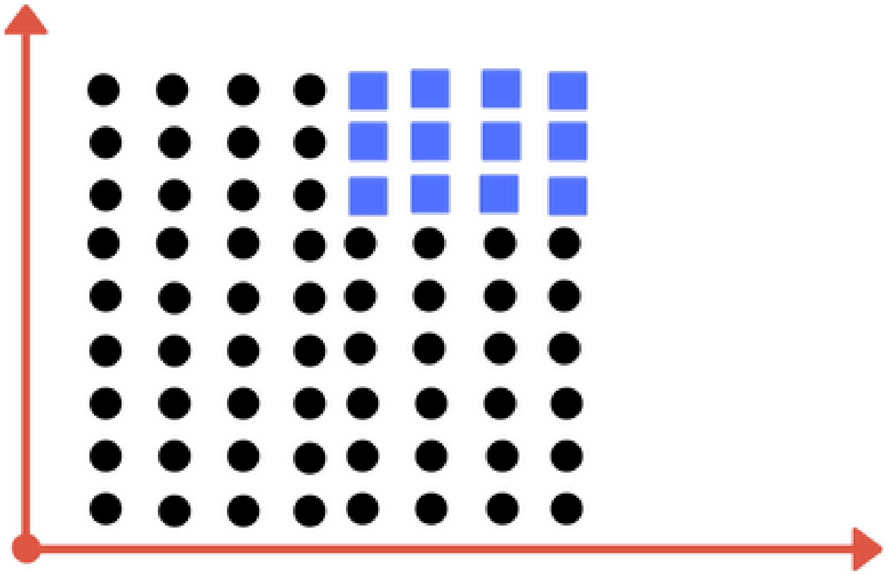
Motivation

- **Plot for two classes represented by black circle and blue squares**



Motivation

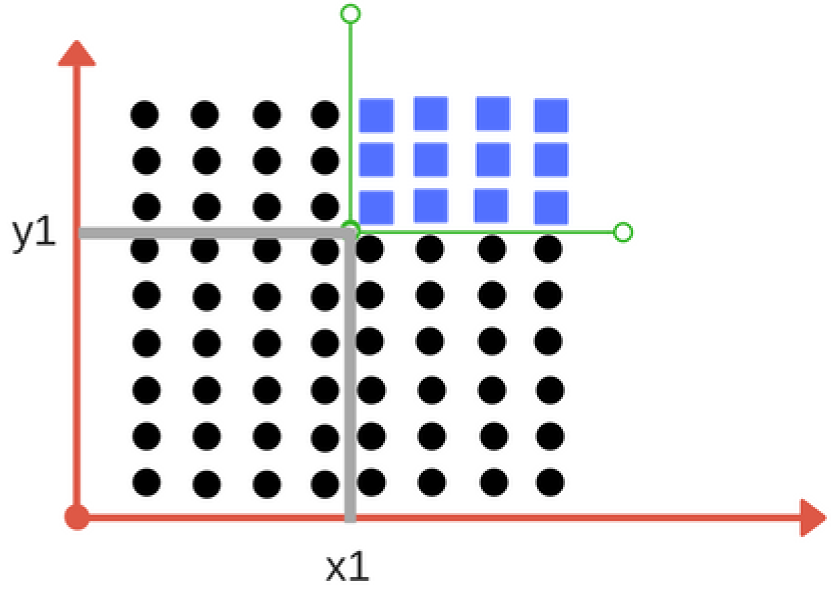
- Plot for two classes represented by black circle and blue squares



Is it possible to draw a single separation line ?

Motivation

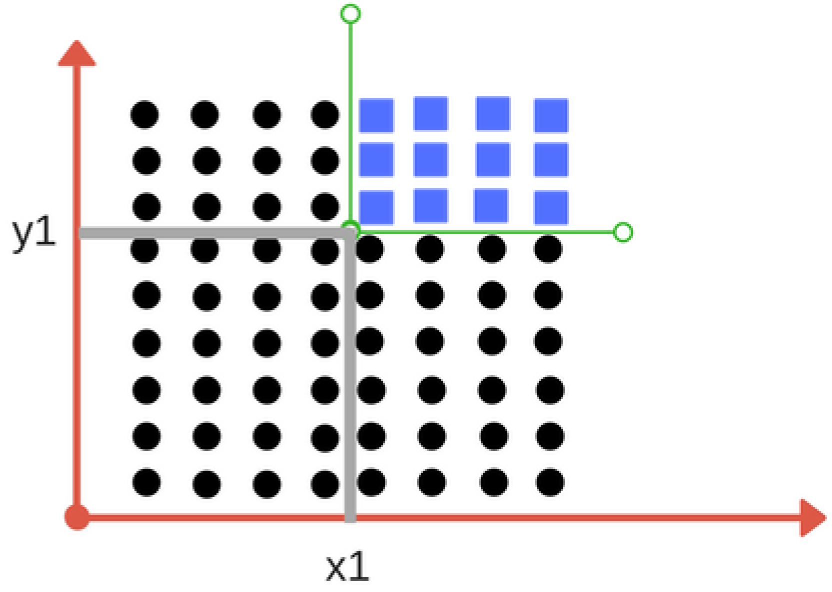
- **Plot for two classes represented by black circle and blue squares**



No, You need Two lines

Motivation

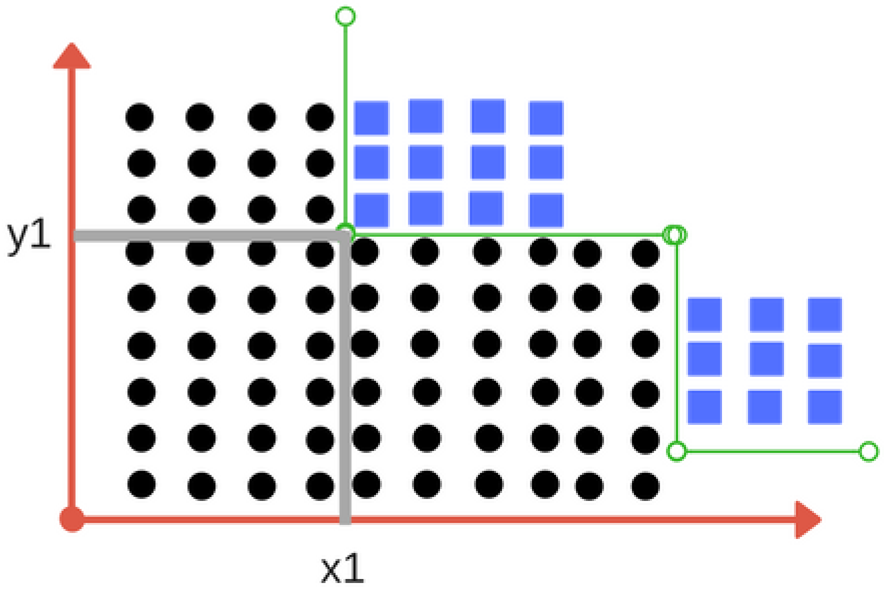
- **Plot for two classes represented by black circle and blue squares**



Two lines one for threshold of x and threshold for y

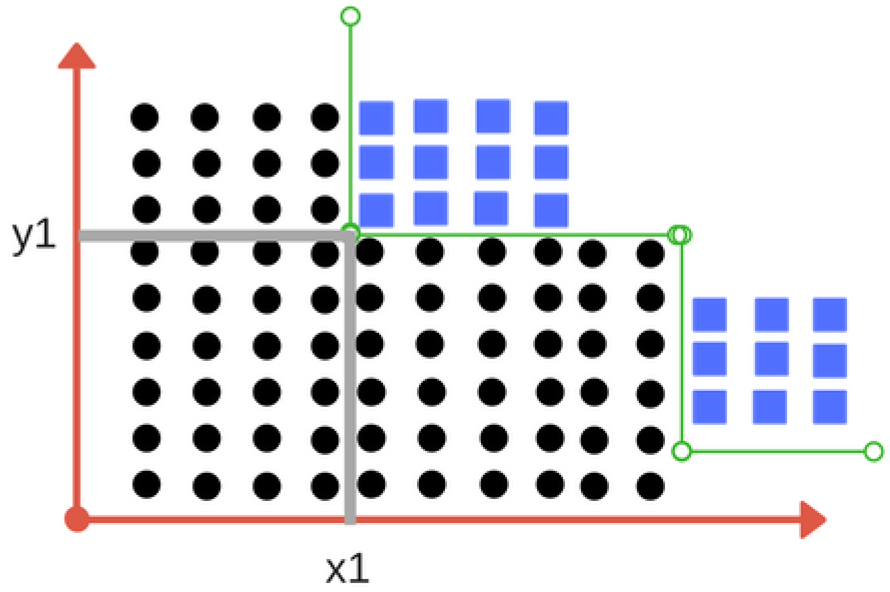
Decision Tree Classifier

- **Decision Tree Classifier, repetitively divides the working area(plot) into sub part by identifying lines**



Decision Tree Classifier

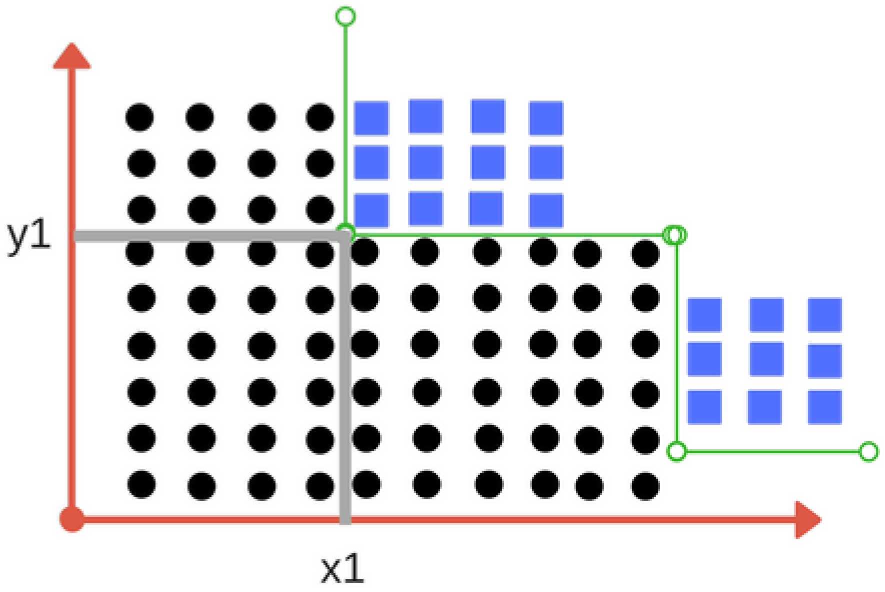
- **Decision Tree Classifier, repetitively divides the working area(plot) into sub part by identifying lines**



So when does it terminate?

When does it terminate?

- Either it has divided into classes that are pure (only containing members of single class)
- Some criteria of classifier attributes are met.



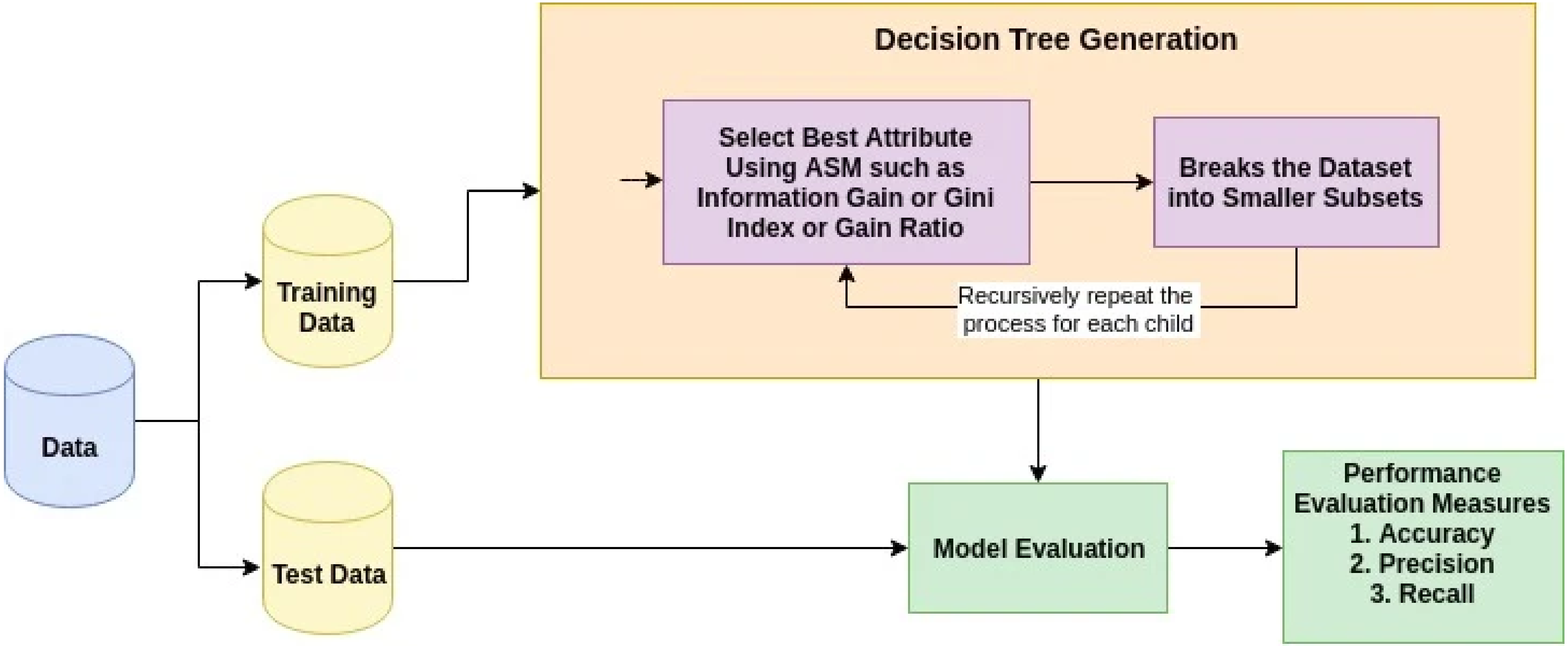
How does the Decision Tree algorithm work?

The basic idea behind any decision tree algorithm is as follows:

Step 1: Select the best attribute using Attribute Selection Measures(ASM) to split the records.
Stpe 2: Make that attribute a decision node and breaks the dataset into smaller subsets.
Step 3: Starts tree building by repeating this process recursively for each child until one of the condition will match:

1). All the tuples belong to the same attribute value.
2). There are no more remaining attributes.
3). There are no more instances.

How does the Decision Tree algorithm work?



Attribute Selection Measures

- Information Gain(Using Entropy)
- Gini Gain(Using Gini Index)
- Gain Ratio
- Classification Error
- Chi-square

Gini Index

- It gives the probability of incorrectly labeling a randomly chosen element from the dataset if we label it according to the distribution of labels in the subset.
- Gini Index is a metric to measure how often a randomly chosen element would be incorrectly identified.
- Gini index says, if we select two items from a population at random then they must be of same class and probability for this is 1 if population is pure.
- It performs only Binary splits
- CART (Classification and Regression Tree) uses Gini method to create binary splits.
- It means an attribute with lower Gini index should be preferred for split.

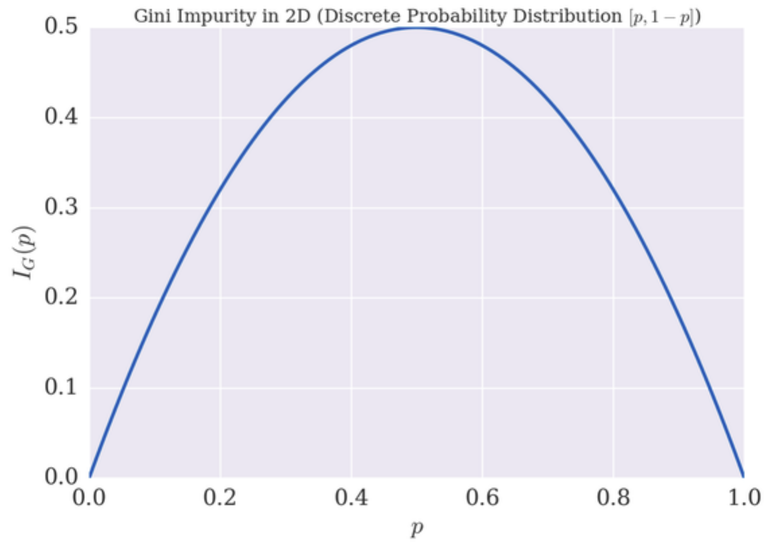
Gini Index

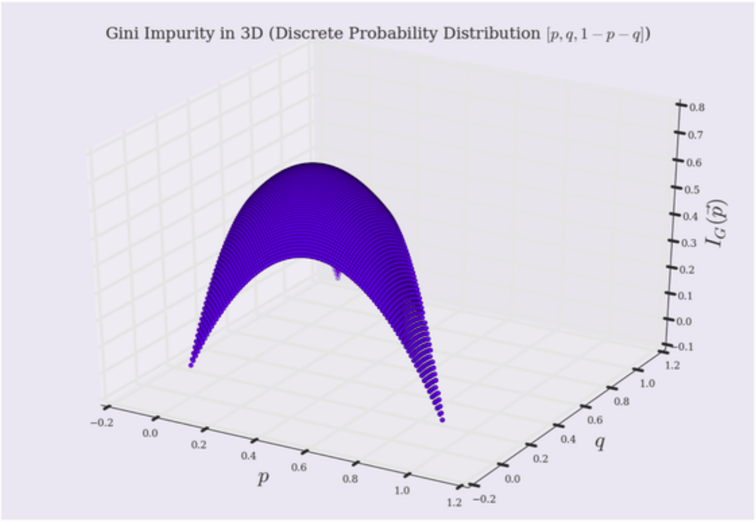
$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

Where Pi is the probability of class i

Gini Index

- Gini is maximum when Probabilty of both class is 1/2
  Gini=1-(1/2)^2-(1/2)^2
  Gini=.5
- Gini is minimum when all samples belong to same class
- Gini=1-(1)^2-(0)^2
  Gini=0



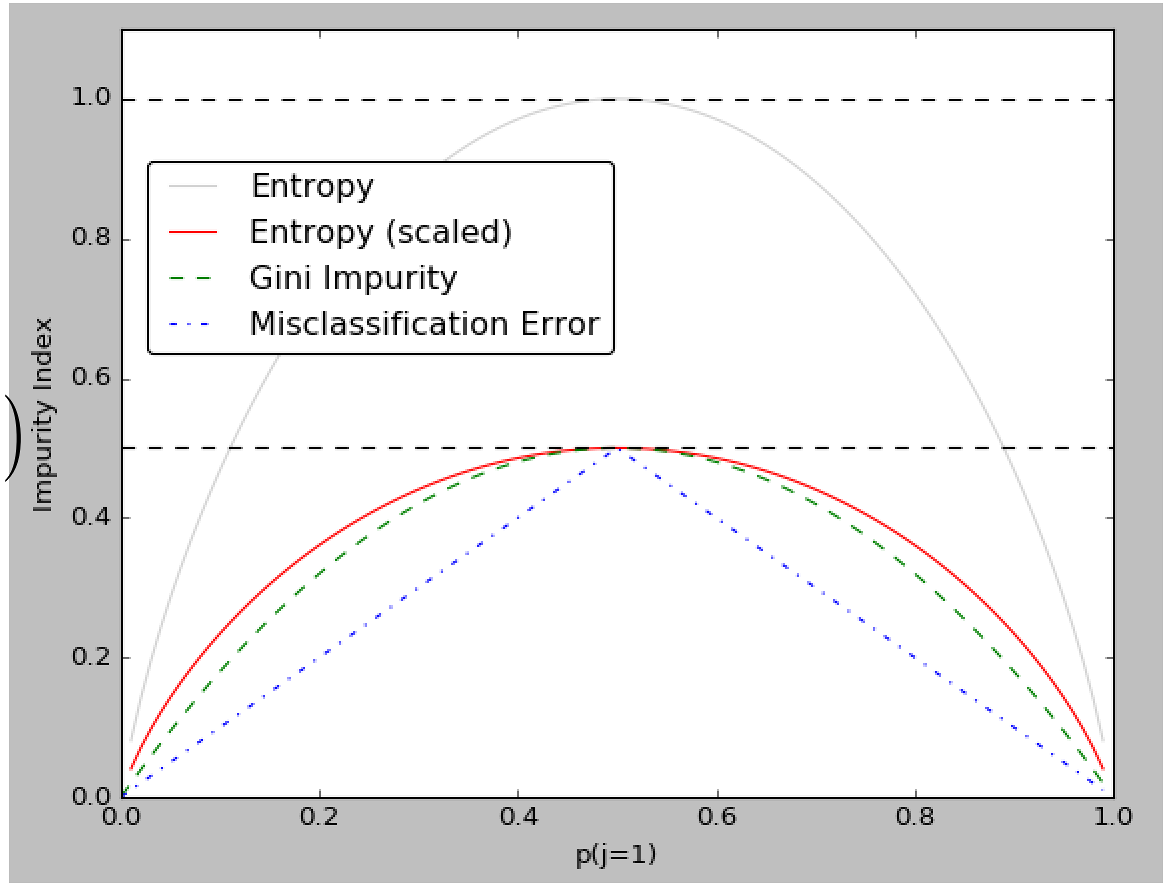


Comparison among Splitting Criteria

$$Entropy = -\sum_{i=0} P_i \log(P_i)$$

$$Gini\ Index = 1 - \sum_{i=0}^{C} P_i^2$$

$$Classification\ Error = 1 - Max(P_i)$$



Decision Tree Construction

- We have four X values (outlook,temp,humidity and wind)
- one y value (play Y or N) also categorical
- This is a binary classification problem
- we need to learn the mapping (what machine learning always does) between X and y
- To create a tree, we need to have a root node first and we know that nodes are features/attributes(outlook,temp,humidity and wind)
- so which one do we need to pick first?

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Day1	Sunny	Hot	High	Weak	*No*
Day2	Sunny	Hot	High	Strong	*No*
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	*No*
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	*No*
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	*No*

Dataset

Steps to Pick Best Node using Gini

Step 1: compute the gini index for data-set

Step 2: for every attribute/feature:

1. calculate gini index for all categorical values

$$Gini(S) = 1 - \sum_{i=0}^{C} P_i^2$$

2. take average Gini index for the current attribute

$$Gini(S, A) = \sum_i \frac{|S_i|}{|S|} . Gini(S_i)$$

3. calculate the gini gain

$$Gini\,Gain = Gini(S) - Gini(S, A)$$

Step 3: pick the best gini gain attribute.

Step 4: Repeat until we get the tree we desired.

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Day1	Sunny	Hot	High	Weak	*No*
Day2	Sunny	Hot	High	Strong	*No*
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	*No*
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	*No*
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	*No*

Dataset S

Compute Gini Index for Dataset S

Step 1: compute the gini index for data-set

$$Gini(S) = 1 - \sum_{i=0}^{C} P_i^2$$

$$Gini(S) = 1 - (9/14)^2 - (5/14)^2$$

$$Gini(S) = 0.46$$

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Day1	Sunny	Hot	High	Weak	*No*
Day2	Sunny	Hot	High	Strong	*No*
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	*No*
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	*No*
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	*No*

Dataset S

Compute Gini Index for Dataset S

Step 2: for every attribute/feature:

1. calculate gini index for all categorical values

$$Gini(Outlook_{sunny}) = [2(yes), 3(No)]$$

$$Gini(Outlook_{sunny}) = 1 - (2/5)^2 - (3/5)^2$$

$$Gini(Outlook_{sunny}) = 0.48$$

$$Gini(Outlook_{overcast}) = [4(yes), 0(No)]$$

$$Gini(Outlook_{sunny}) = 1 - (4/4)^2 - (0/4)^2$$

$$Gini(Outlook_{sunny}) = 0.0$$

$$Gini(Outlook_{rain}) = [2(yes), 3(No)]$$

$$Gini(Outlook_{rain}) = 1 - (2/5)^2 - (3/5)^2$$

$$Gini(Outlook_{rain}) = 0.48$$

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Day1	Sunny	Hot	High	Weak	No
Day2	Sunny	Hot	High	Strong	No
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	No
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	No
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	No

Dataset S

Compute Gini Index for Dataset S

Step 2: for every attribute/feature:

1. Done
2. take average Gini index for the current attribute

$$Gini(S,A)=\sum_i \frac{|S_i|}{|S|}.Gini(S_i)$$

$$Gini(S,Outlook)=5/14*0.48+4/14*0+5/14*0.48$$

$$Gini(S,outlook)=0.34$$

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Day1	Sunny	Hot	High	Weak	*No*
Day2	Sunny	Hot	High	Strong	*No*
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	*No*
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	*No*
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	*No*

Dataset S

Compute Gini Index for Dataset S

Step 2: for every attribute/feature:

1. **Done**
2. **Done**
3. **calculate the gini gain**

$$Gini\,Gain(S,A)=Gini(S)-Gini(S,A)$$

$$Gini\,Gain(S,Outlook)=Gini(S)-Gini(S,Outlook)$$

$$Gini\,Gain(S,Outlook)=0.46-0.34$$

$$Gini\,Gain(S,Outlook)=0.12$$

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Day1	Sunny	Hot	High	Weak	*No*
Day2	Sunny	Hot	High	Strong	*No*
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	*No*
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	*No*
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	*No*

Dataset S

Compute Gini Index for Dataset S

Step 2: for every attribute/feature:

1. calculate gini index for all categorical values

$$Gini(Temprature=Hot)=[2(yes),2(No)]$$

$$Gini(Temprature=Hot)=1-(2/4)^2-(2/4)^2$$

$$Gini(Temperature=Hot)=0.5$$

$$Gini(Temprature=Mild)=[4(yes),2(No)]$$

$$Gini(Temprature=Mild)=1-(4/6)^2-(2/6)^2$$

$$Gini(Temprature=Mild)=0.44$$

$$Gini(Temprature=Cool)=[3(yes),1(No)]$$

$$Gini(Temprature=Cool)=1-(3/4)^2-(1/4)^2$$

$$Gini(Temprature=Cool)=0.375$$

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Day1	Sunny	Hot	High	Weak	No
Day2	Sunny	Hot	High	Strong	No
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	No
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	No
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	No

Dataset S

Compute Gini Index for Dataset S

Step 2: for every attribute/feature:

1. **Done**
2. **take average Gini index for the current attribute**

$$Gini(S, A) = \sum_i \frac{|S_i|}{|S|} . Gini(S_i)$$

$$Gini(S, Temprature) = 4/14*0.5 + 6/14*0.44 + 4/14*0.375$$

$$Gini(S, Temprature) = 0.438$$

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Day1	Sunny	Hot	High	Weak	*No*
Day2	Sunny	Hot	High	Strong	*No*
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	*No*
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	*No*
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	*No*

Dataset S

Compute Gini Index for Dataset S

Step 2: for every attribute/feature:

1. Done
2. Done
3. calculate the gini gain

$$Gini\,Gain(S,A)=Gini(S)-Gini(S,A)$$

$$Gini\,Gain(S,Temprature)=Gini(S)-Gini(S,Temperature)$$

$$Gini\,Gain(S,Temprature)=0.46-0.438$$

$$Gini\,Gain(S,Outlook)=0.022$$

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Day1	Sunny	Hot	High	Weak	No
Day2	Sunny	Hot	High	Strong	No
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	No
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	No
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	No

Dataset S

Compute Gini Index for Dataset S

Step 2: for every attribute/feature:

1. calculate gini index for all categorical values

$$Gini(Humidity=High)=[3(yes),4(No)]$$

$$Gini(Humidity=High)=1-(3/7)^2-(4/7)^2$$

$$Gini(Humidity=High)=0.489$$

$$Gini(Humidity=Normal)=[6(yes),1(No)]$$

$$Gini(Humidity=Normal)=1-(6/7)^2-(1/7)^2$$

$$Gini(Humidity=Normal)=0.244$$

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Day1	Sunny	Hot	High	Weak	*No*
Day2	Sunny	Hot	High	Strong	*No*
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	*No*
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	*No*
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	*No*

Dataset S

Compute Gini Index for Dataset S

Step 2: for every attribute/feature:

1. **Done**
2. **take average Gini index for the current attribute**

$$Gini(S,A)=\sum_i \frac{|S_i|}{|S|}.Gini(S_i)$$

$$Gini(S,Humidity)=7/14*0.489+7/14*0.244$$

$$Gini(S,Humidity)=0.366$$

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Day1	Sunny	Hot	High	Weak	*No*
Day2	Sunny	Hot	High	Strong	*No*
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	*No*
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	*No*
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	*No*

Dataset S

Compute Gini Index for Dataset S

Step 2: for every attribute/feature:

1. Done
2. Done
3. calculate the gini gain

$$Gini\,Gain(S,A)=Gini(S)-Gini(S,A)$$

$$Gini\,Gain(S,Humidity)=Gini(S)-Gini(S,Humidity)$$

$$Gini\,Gain(S,Humidity)=0.46-0.366$$

$$Gini\,Gain(S,Humidity)=0.094$$

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Day1	Sunny	Hot	High	Weak	*No*
Day2	Sunny	Hot	High	Strong	*No*
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	*No*
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	*No*
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	*No*

Dataset S

Compute Gini Index for Dataset S

Step 2: for every attribute/feature:

1. calculate gini index for all categorical values

$$Gini(Wind=Weak)=[6(yes),2(No)]$$

$$Gini(Wind=Weak)=1-(6/8)^2-(2/8)^2$$

$$Gini(Wind=Weak)=0.375$$

$$Gini(Wind=Strong)=[3(yes),3(No)]$$

$$Gini(Wind=Strong)=1-(3/6)^2-(3/6)^2$$

$$Gini(Wind=Strong)=0.5$$

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Day1	Sunny	Hot	High	Weak	*No*
Day2	Sunny	Hot	High	Strong	*No*
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	*No*
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	*No*
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	*No*

Dataset S

Compute Gini Index for Dataset S

Step 2: for every attribute/feature:

1. Done
2. take average Gini index for the current attribute

$$Gini(S,A)=\sum_i \frac{|S_i|}{|S|}.Gini(S_i)$$

$$Gini(S,Wind)=8/14*0.375+6/14*0.5$$

$$Gini(S,Wind)=0.428$$

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Day1	Sunny	Hot	High	Weak	No
Day2	Sunny	Hot	High	Strong	No
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	No
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	No
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	No

Dataset S

Compute Gini Index for Dataset S

Step 2: for every attribute/feature:

1. Done
2. Done
3. calculate the gini gain

$$Gini\,Gain(S,A)=Gini(S)-Gini(S,A)$$

$$Gini\,Gain(S,Wind)=Gini(S)-Gini(S,Wind)$$

$$Gini\,Gain(S,Wind)=0.46-0.428$$

$$Gini\,Gain(S,Wind)=0.032$$

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Day1	Sunny	Hot	High	Weak	No
Day2	Sunny	Hot	High	Strong	No
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	No
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	No
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	No

Dataset S

Compute Gini Index for Dataset S

Step 2: for every attribute/feature:

1. Done
2. Done
3. Done

Step 3: pick the best gini gain attribute.
Best Gini Gain= Max(Gini Gain(Ai))
Where Ai is i th attribute

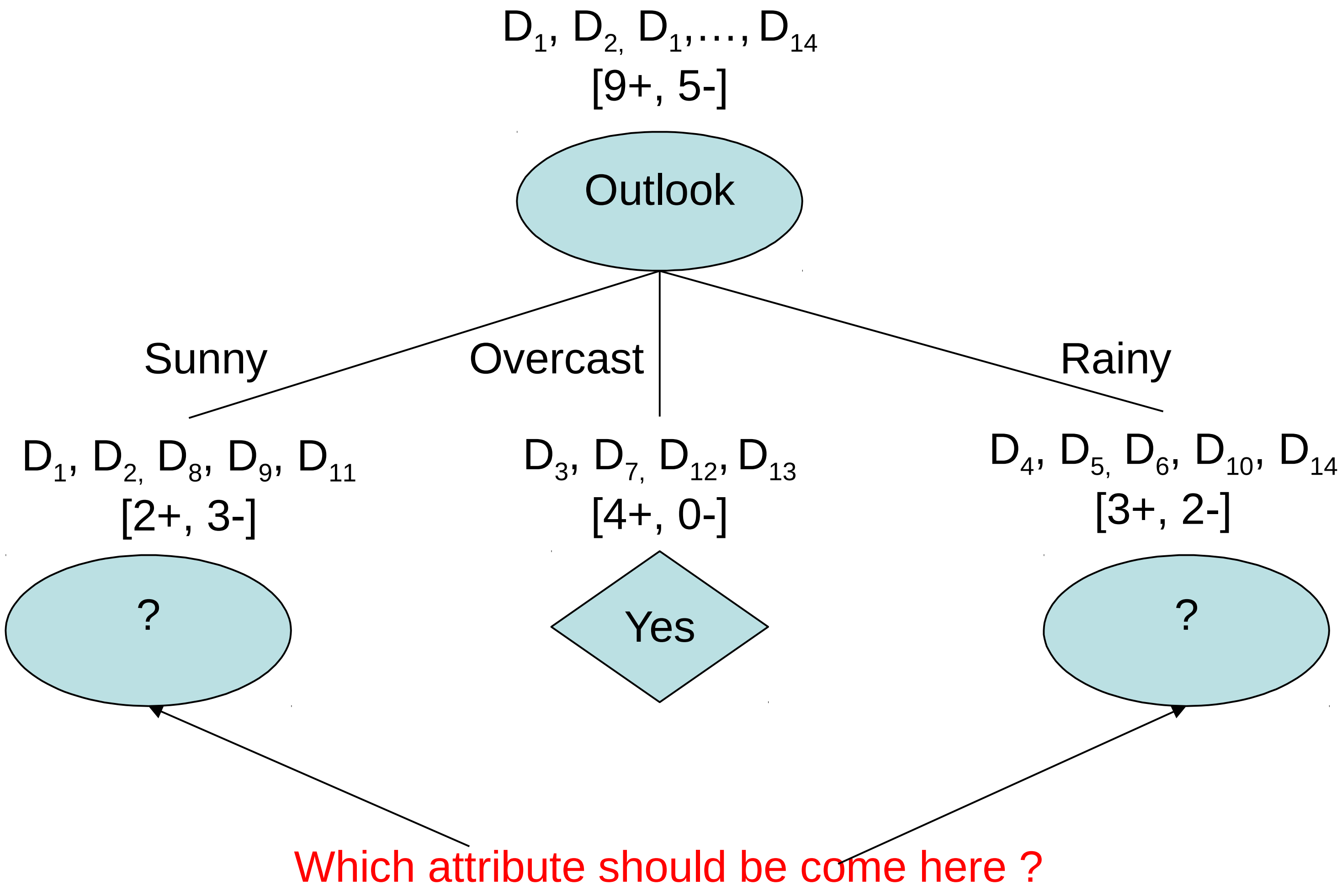
- Best Gini Gain=Max(Gini Gain(Outlook),Gini Gain(Temp.),Gini Gain(Humidity),Gini Gain(Wind))
  Best Gini Gain=Max(0.12,0.022,0.094,0.032)

  Best Gini Gain=0.12
- Best Node/Attribute to split is Outlook

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Day1	Sunny	Hot	High	Weak	No
Day2	Sunny	Hot	High	Strong	No
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	No
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	No
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	No

Dataset S

Decision Tree(CART)



Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Day1	Sunny	Hot	High	Weak	No
Day2	Sunny	Hot	High	Strong	No
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	No
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	No
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	No

S: Data set

Decision Tree(CART)

$$Gini(Outlook=sunny)=1-(2/5)^2-(3/5)^2$$
$$Gini(Outlook=sunny)=0.48$$

$$Gini(Outlook=sunny \wedge Temprature=Hot)=1-(0/2)^2-(2/2)^2$$
$$Gini(Outlook=sunny \wedge Temprature=Hot)=0$$

$$Gini(Outlook=sunny \wedge Temprature=Mild)=1-(1/2)^2-(1/2)^2$$
$$Gini(Outlook=sunny \wedge Temprature=Mild)=0.5$$

$$Gini(Outlook=sunny \wedge Temprature=cool)=1-(1/1)^2-(0/1)^2$$
$$Gini(Outlook=sunny \wedge Temprature=cool)=0$$

$$Gini(Outlook=sunny,Temperature)=2/5*0+2/5*0.5+1*0$$
$$Gini(Outlook=sunny,Temperature)=0.2$$

$$Gini\,Gain(Outlook=sunny,Temperature)=0.48-0.2=0.28$$

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Day1	Sunny	Hot	High	Weak	No
Day2	Sunny	Hot	High	Strong	No
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	No
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	No
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	No

S: Data set

Decision Tree(CART)

$$Gini(Outlook=sunny)=1-(2/5)^2-(3/5)^2$$
$$Gini(Outlook=sunny)=0.48$$

$$Gini(Outlook=sunny \wedge Humidity=High)=1-(0/3)^2-(3/3)^2$$
$$Gini(Outlook=sunny \wedge Humidity=High)=0.0$$

$$Gini(Outlook=sunny \wedge Humidity=Normal)=1-(2/2)^2-(0/2)^2$$
$$Gini(Outlook=sunny \wedge Humidity=High)=0.0$$

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Day1	Sunny	Hot	High	Weak	No
Day2	Sunny	Hot	High	Strong	No
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	No
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	No
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	No

S: Data set

$$Gini(Outlook=sunny, Humidity)=3/5*0+2/5*0.0$$
$$Gini(Outlook=sunny, Humidity)=0.0$$

$$Gini\,Gain(Outlook=sunny, Humidity)=0.48-0.0=0.48$$

Decision Tree(CART)

$$Gini(Outlook=sunny)=1-(2/5)^2-(3/5)^2$$
$$Gini(Outlook=sunny)=0.48$$

$$Gini(Outlook=sunny \wedge Wind=Weak)=1-(1/3)^2-(2/3)^2$$
$$Gini(Outlook=sunny \wedge Wind=Weak)=0.44$$

$$Gini(Outlook=sunny \wedge Wind=Strong)=1-(1/2)^2-(1/2)^2$$
$$Gini(Outlook=sunny \wedge Wind=Strong)=0.5$$

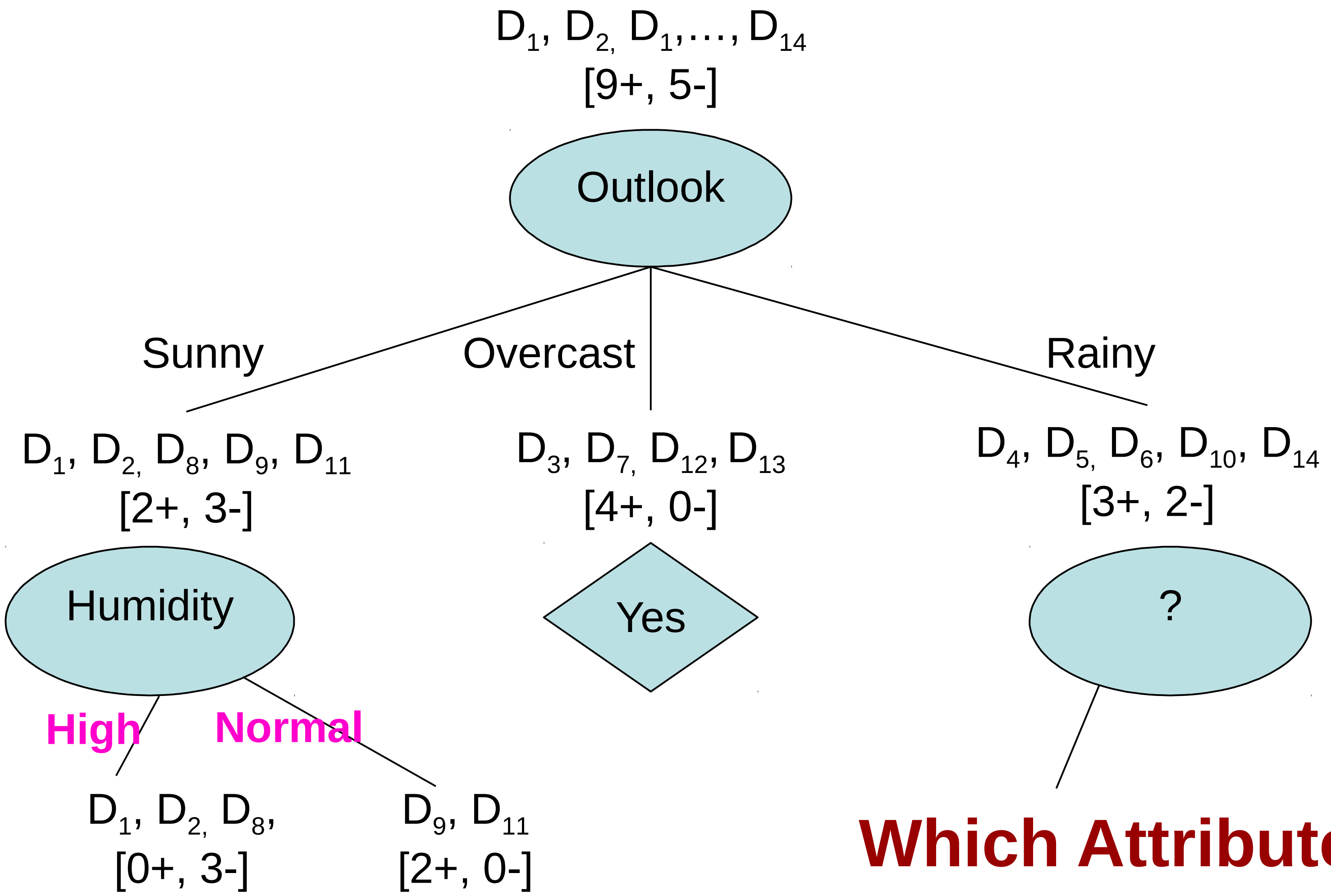
$$Gini(Outlook=sunny, Wind)=3/5*0.44+2/5*0.5$$
$$Gini(Outlook=sunny, Wind)=0.464$$

$$Gini\,Gain(Outlook=sunny, Wind)=0.48-0.464=0.016$$

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Day1	Sunny	Hot	High	Weak	No
Day2	Sunny	Hot	High	Strong	No
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	No
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	No
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	No

S: Data set

Decision Tree(CART)



Which Attribute?

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Day1	Sunny	Hot	High	Weak	*No*
Day2	Sunny	Hot	High	Strong	*No*
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	*No*
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	*No*
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	*No*

S: Data set

Decision Tree(CART)

$$Gini(Outlook=Rain)=1-(3/5)^2-(2/5)^2$$
$$Gini(Outlook=Rain)=0.48$$

$$Gini(Outlook=rain \wedge Temprature=Mild)=1-(2/3)^2-(1/3)^2$$
$$Gini(Outlook=rain \wedge Temprature=Mild)=0.44$$

$$Gini(Outlook=rain \wedge Temprature=cool)=1-(1/2)^2-(1/2)^2$$
$$Gini(Outlook=rain \wedge Temprature=cool)=0.5$$

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Day1	Sunny	Hot	High	Weak	No
Day2	Sunny	Hot	High	Strong	No
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	No
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	No
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	No

S: Data set

$$Gini(Outlook=rain, Temperature)=0/5*0+3/5*0.44+2/5*0.5$$
$$Gini(Outlook=rain, Temperature)=0.464$$
$$Gini\, Gain(Outlook=rain, Temperature)=0.48-0.464=0.016$$

Decision Tree(CART)

$$Gini(Outlook=rain)=1-(3/5)^2-(2/5)^2$$
$$Gini(Outlook=rain)=0.48$$

$$Gini(Outlook=rain \wedge Wind=Weak)=1-(3/3)^2-(0/3)^2$$
$$Gini(Outlook=rain \wedge Wind=Weak)=0.0$$

$$Gini(Outlook=sunny \wedge Wind=Strong)=1-(0/2)^2-(2/2)^2$$
$$Gini(Outlook=rain \wedge Wind=Strong)=0.0$$

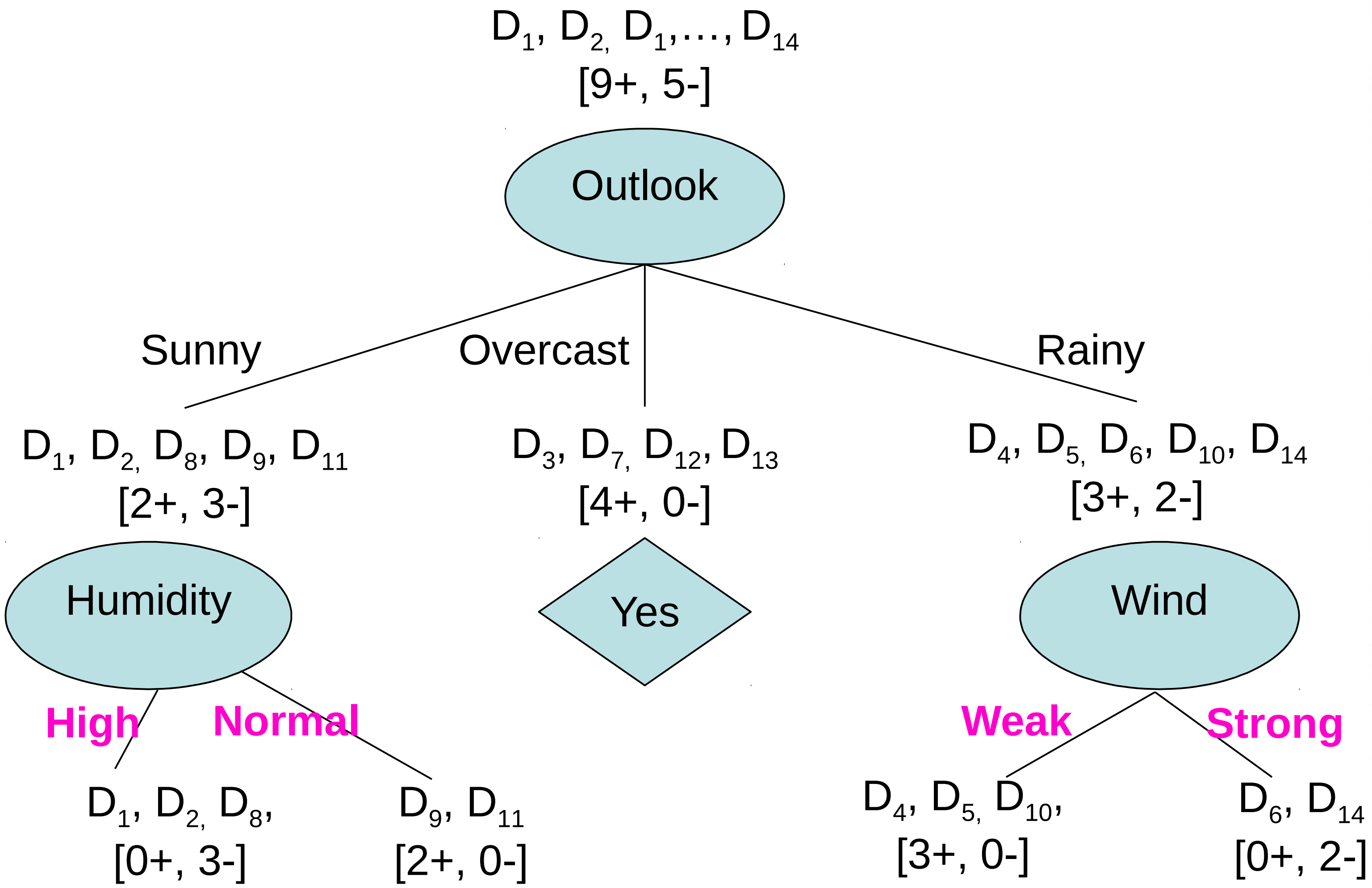
$$Gini(Outlook=rain, Wind)=3/5*0.0+2/5*0.0$$
$$Gini(Outlook=rain, Wind)=0.464$$

$$Gini\,Gain(Outlook=rain, Wind)=0.48-0.0=0.48$$

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Day1	Sunny	Hot	High	Weak	No
Day2	Sunny	Hot	High	Strong	No
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	No
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	No
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	No

S: Data set

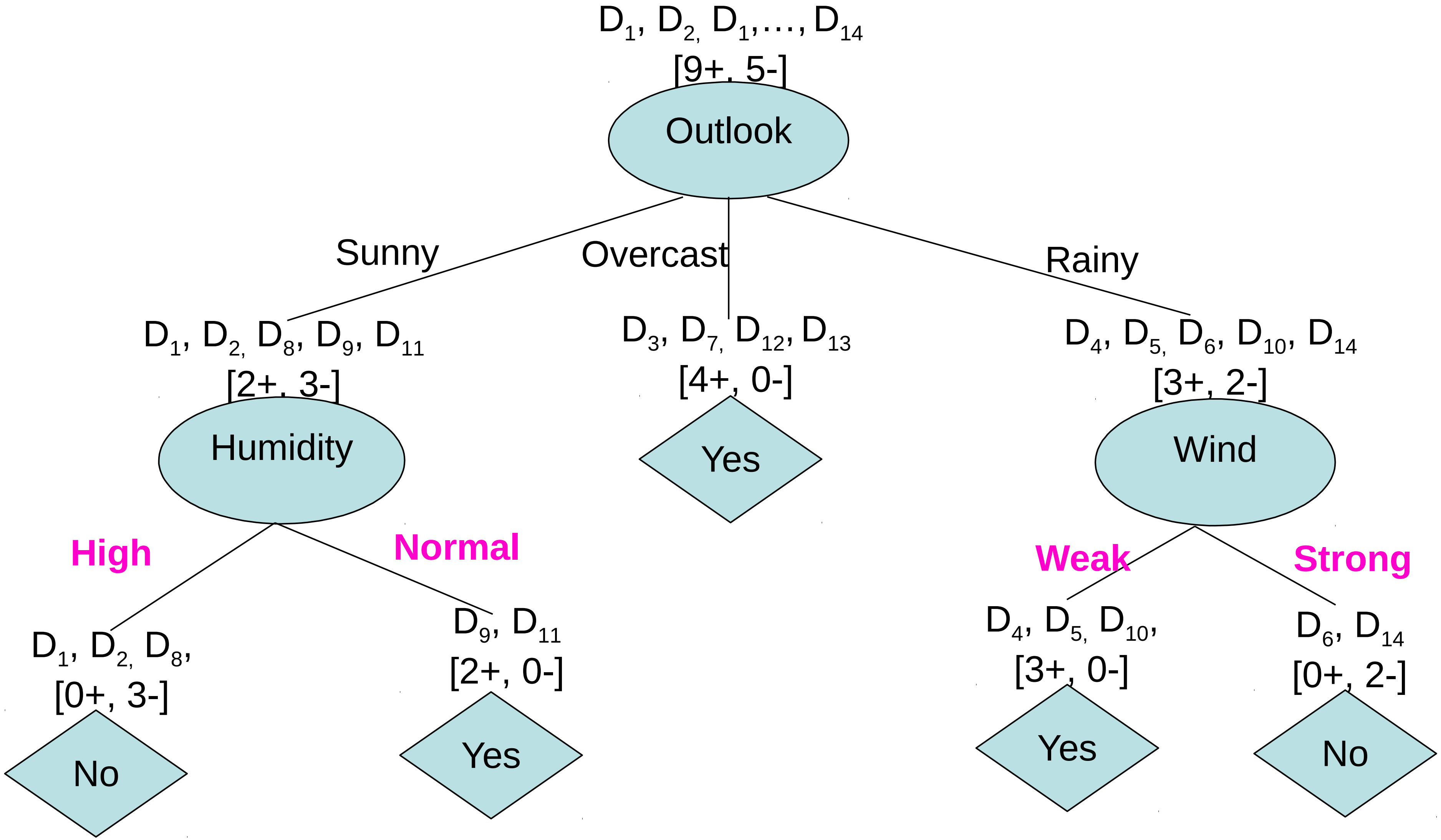
Decision Tree(CART)



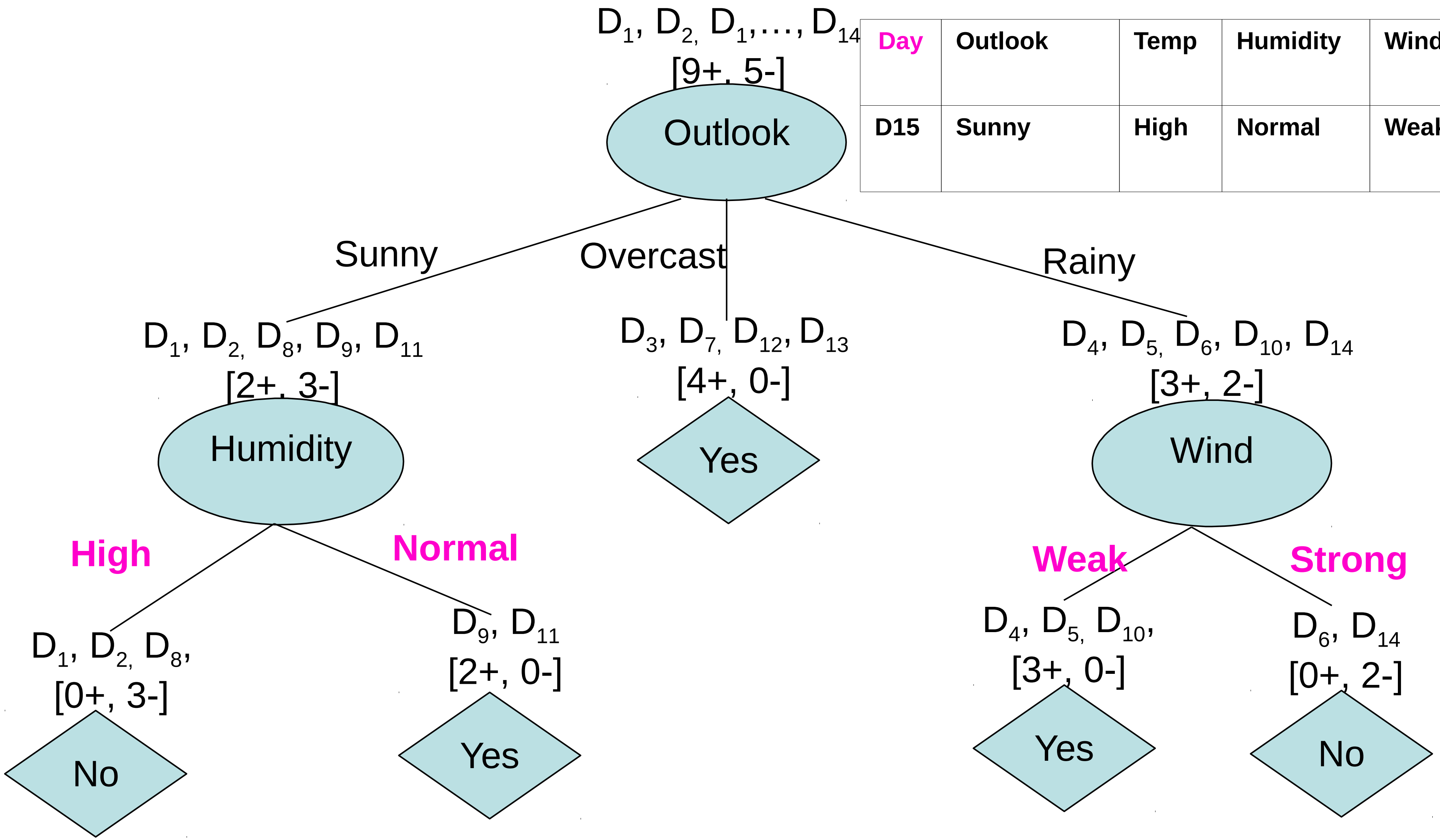
Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Day1	Sunny	Hot	High	Weak	*No*
Day2	Sunny	Hot	High	Strong	*No*
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	*No*
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	*No*
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	*No*

S: Data set

Decision Tree(CART)

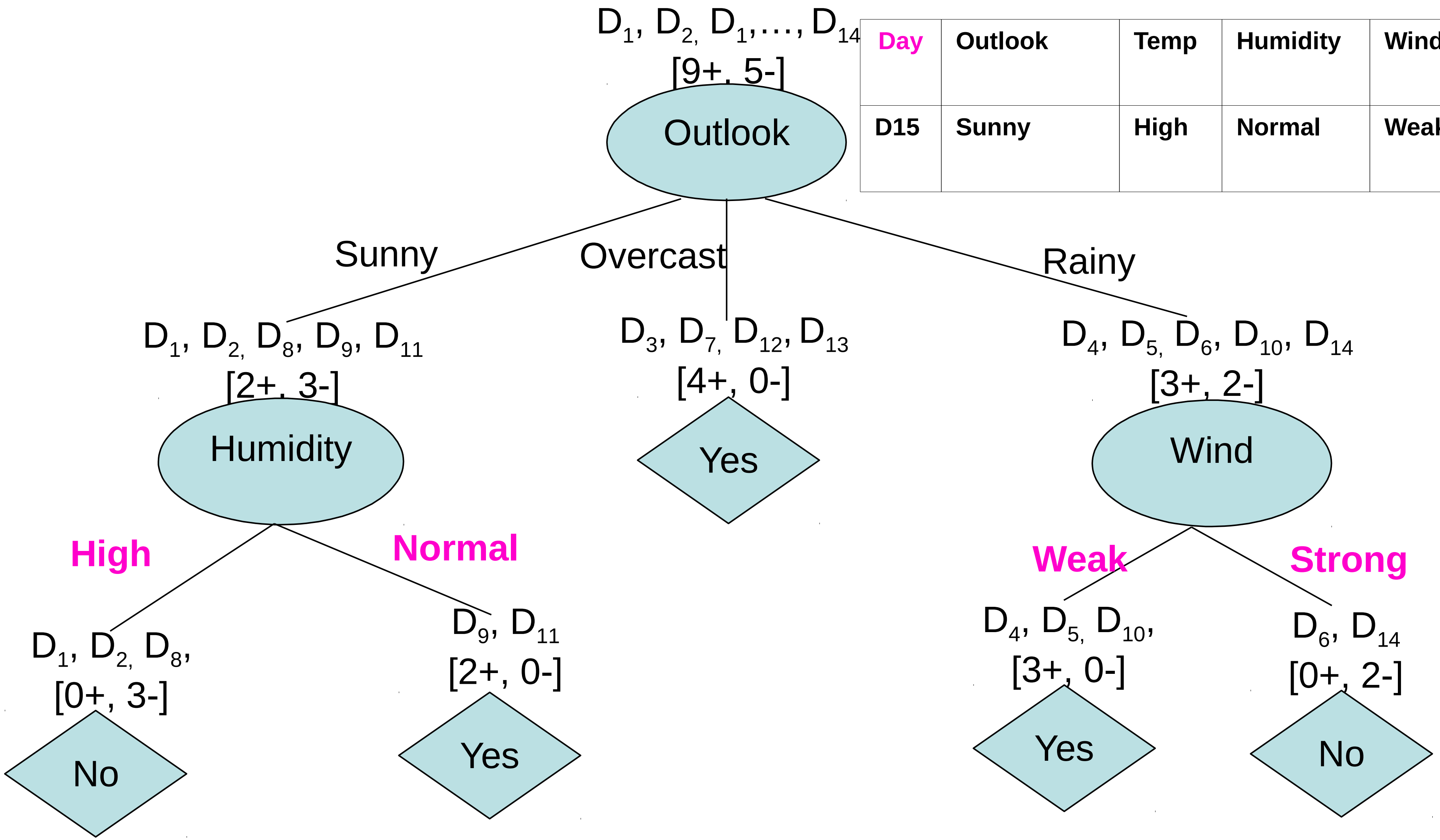


Decision Tree(CART)



Day	Outlook	Temp	Humidity	Wind	Play or not
D15	Sunny	High	Normal	Weak	?

Decision Tree(CART)



Day	Outlook	Temp	Humidity	Wind	Play or not
D15	Sunny	High	Normal	Weak	Yes

Highly-branching attributes

- **Problematic: attributes with a large number of values (extreme case: ID code)**
- **Subsets are more likely to be pure if there is a large number of values**
  - **Information gain is biased towards choosing attributes with a large number of values**
  - **This may result in *overfitting* (selection of an attribute that is non-optimal for prediction)**
- **Another problem: *fragmentation***

The Gain Ratio

- *Gain ratio*: a modification of the information gain that reduces its bias on high-branch attributes
- Gain ratio takes number and size of branches into account when choosing an attribute
  - It corrects the information gain by taking the *intrinsic information* of a split into account
  - Also called split ratio
- Intrinsic information: entropy of distribution of instances into branches
  - (i.e. how much info do we need to tell which branch an instance belongs to)

The Gain Ratio

- ***Gain ratio should be***
  - **Large when data is evenly spread**
  - **Small when all data belong to one branch**
- ***Gain ratio*** **(Quinlan'86) normalizes info gain by this reduction:**

$$IntrinsicInfo(S,A) \equiv -\sum \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|}.$$

$$GainRatio(S,A) = \frac{Gain(S,A)}{IntrinsicInfo(S,A)}.$$

Computing the gain ratio

- Example: intrinsic information for ID code

$$\text{info}([1,1,\ldots,1])=14\times(-1/14\times\log 1/14)=3.807\text{ bits}$$

- **Importance of attribute decreases as intrinsic information gets larger**
- Example of gain ratio:

$$\text{gain\_ratio}(\textit{Attribute})=\frac{\text{gain}(\textit{Attribute})}{\text{intrinsic\_info}(\textit{Attribute})}$$

- Example:

$$\text{gain\_ratio}(ID_{code})=\frac{0.940\text{ bits}}{3.807\text{ bits}}=0.246$$

Gain ratios for weather data

Outlook		Temperature	
Info:	0.693	Info:	0.911
Gain: 0.940-0.693	0.247	Gain: 0.940-0.911	0.029
Split info: info([5,4,5])	1.577	Split info: info([4,6,4])	1.362
Gain ratio: 0.247/1.577	0.156	Gain ratio: 0.029/1.362	0.021
Humidity		Windy	
Info:	0.788	Info:	0.892
Gain: 0.940-0.788	0.152	Gain: 0.940-0.892	0.048
Split info: info([7,7])	1.000	Split info: info([8,6])	0.985
Gain ratio: 0.152/1	0.152	Gain ratio: 0.048/0.985	0.049

More on the gain ratio

- “ Outlook” still comes out top
- However: “ID code” has greater gain ratio
  - Standard fix: *ad hoc* test to prevent splitting on that type of attribute
- Problem with gain ratio: it may overcompensate
  - May choose an attribute just because its intrinsic information is very low
  - Standard fix:
    - First, only consider attributes with greater than average information gain
    - Then, compare them on gain ratio

Any Question?