Classification

Decision Trees

Mr.Gangadhar Immadi

immadi.gangadhar@gmail.com

9986789040

Agenda

- Introduction to Classification
- Classification Techniques
- Decission Trees
  - ID3
  - CART
  - CHAID
  - C4.5
- Performance Metrics

CLASSIFICATION PROBLEMS

- Classification problems are an important category of problems in analytics in which the response variable (Y) takes a discrete value.
- The primary objective is to predict the class of a customer (or class probability) based on the values of explanatory variables or predictors.
- Complexity – Linearly separable v/s Non linearly separable

Given a collection of records (training set)

- Each record is by characterized by a tuple ($\boldsymbol{x}$,y), where $\boldsymbol{x}$ is the attribute set and y is the class label
  - $\boldsymbol{x}$: attribute, predictor, independent variable, input
  - y: class, response, dependent variable, output

Task:

- Learn a model that maps each attribute set $\boldsymbol{x}$ into one of the predefined class labels y

Classification Problems

- Classification is an important category of problems in which the decision maker would like to classify the case/entity/customers into two or more groups.

- Examples of Classification Problems:
  - ✓Customer profiling (customer segmentation)
  - ✓Customer Churn.
  - ✓Credit Classification (low, high and medium risk)
  - ✓Employee attrition.
  - ✓Fraud (classification of transaction to fraud/no-fraud)
  - ✓Stress levels
  - ✓Text Classification (Sentiment Analysis)
  - ✓Outcome of any binomial and multinomial experiment.

Examples of Classification Task

Task	Attribute set, $\boldsymbol{x}$	Class label, y
Categorizing email messages	Features extracted from email message header and content	spam or non-spam
Identifying tumor cells	Features extracted from x-rays or MRI scans	malignant or benign cells
Cataloging galaxies	Features extracted from telescope images	Elliptical, spiral, or irregular-shaped galaxies

Classification Techniques

- Base Classifiers
  - Decision Tree based Methods
  - Logistic Regression
  - Rule-based Methods
  - Nearest-neighbor
  - Naïve Bayes and Bayesian Belief Networks
  - Support Vector Machines
  - Neural Networks, Deep Neural Nets

- Ensemble Classifiers
  - Boosting, Bagging, Random Forests

DECISION TREES: INTRODUCTION

Decision trees (also known as decision tree learning or classification trees) are a collection of predictive analytics techniques that use tree-like graphs for predicting the value of a response variable (or target variable) based on the values of explanatory variables (or predictors).

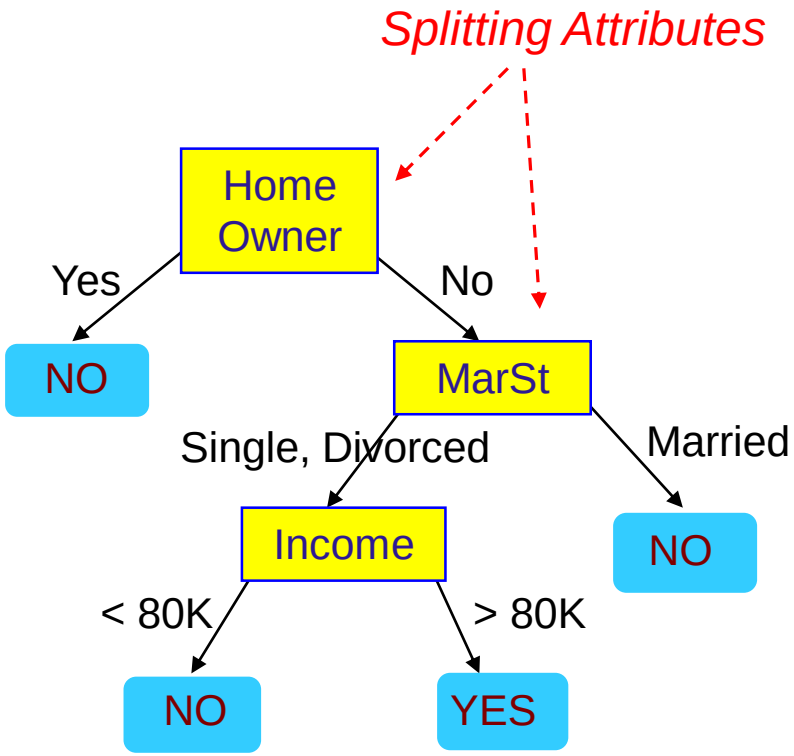
Decision trees use the following criteria to develop the tree

Splitting Criteria
Merging Criteria
Stopping Criteria

Example of a Decision Tree

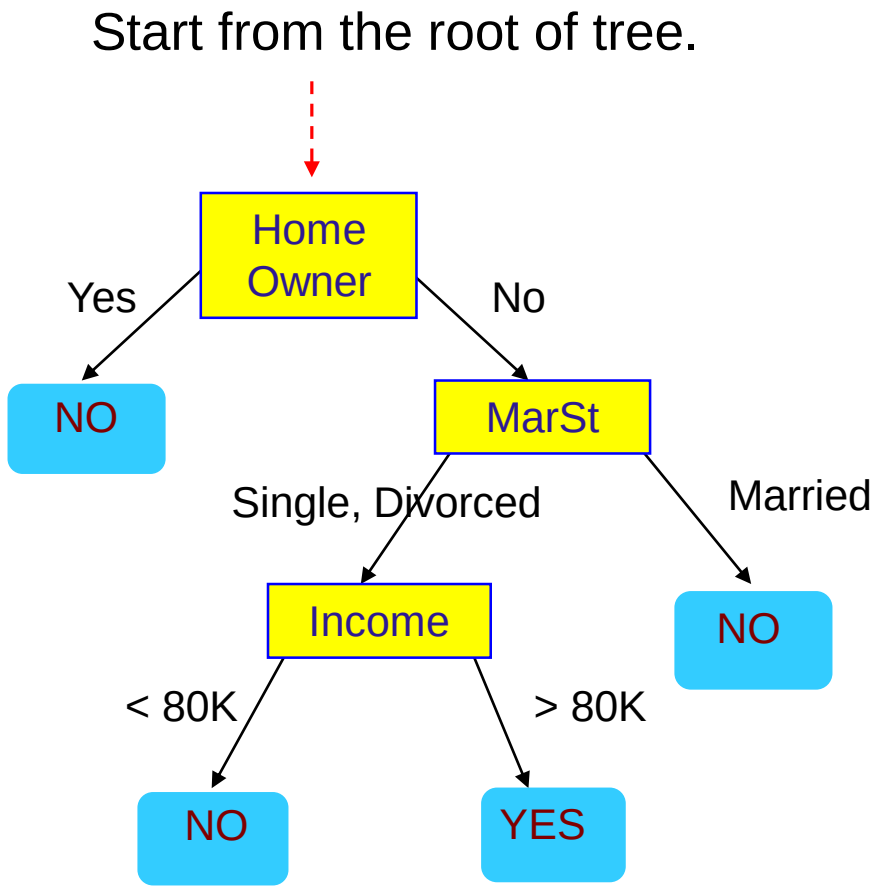
ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
	categorical	categorical	continuous	class
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Training Data



Model: Decision Tree

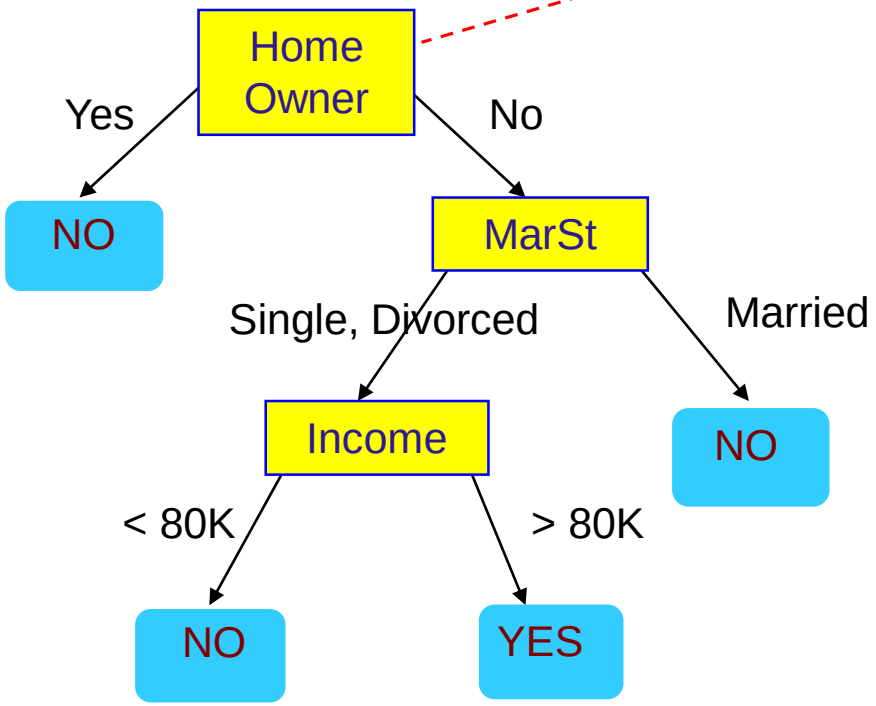
Apply Model to Test Data



Test Data

Home Owner	Marital Status	Annual Income	Defaulted Borrower
No	Married	80K	**?**

Apply Model to Test Data



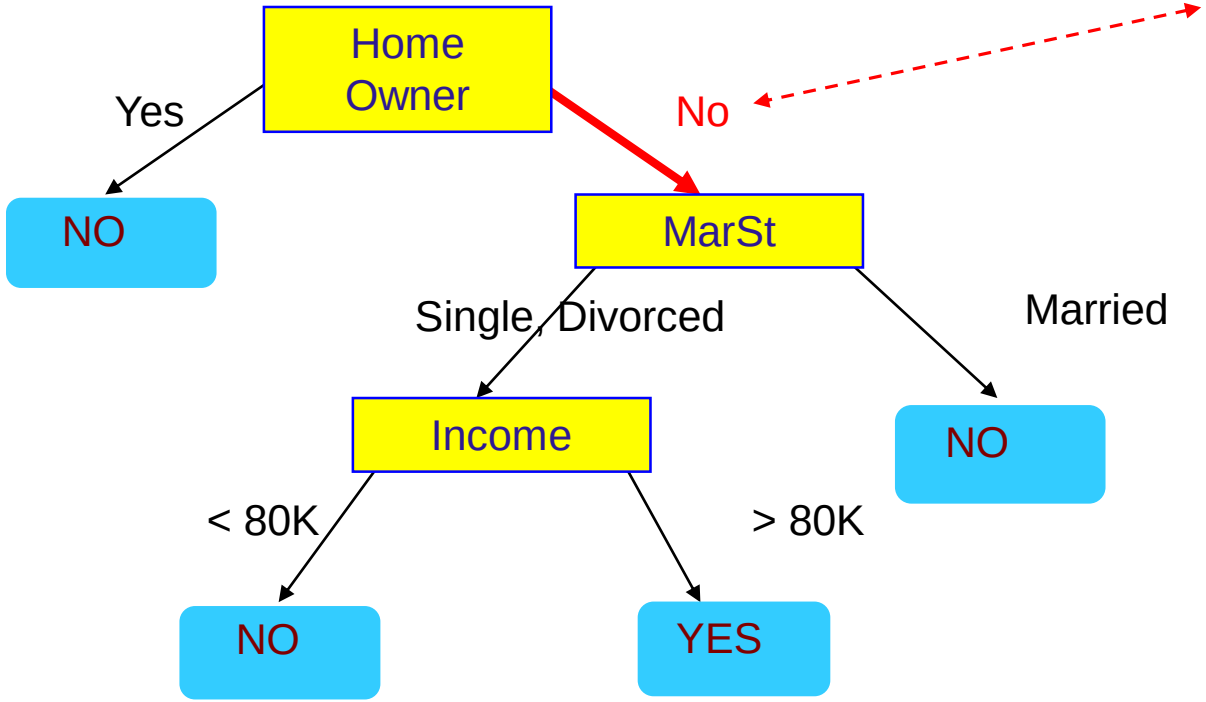
Test Data

Home Owner	Marital Status	Annual Income	Defaulted Borrower
No	Married	80K	?

Apply Model to Test Data

Test Data

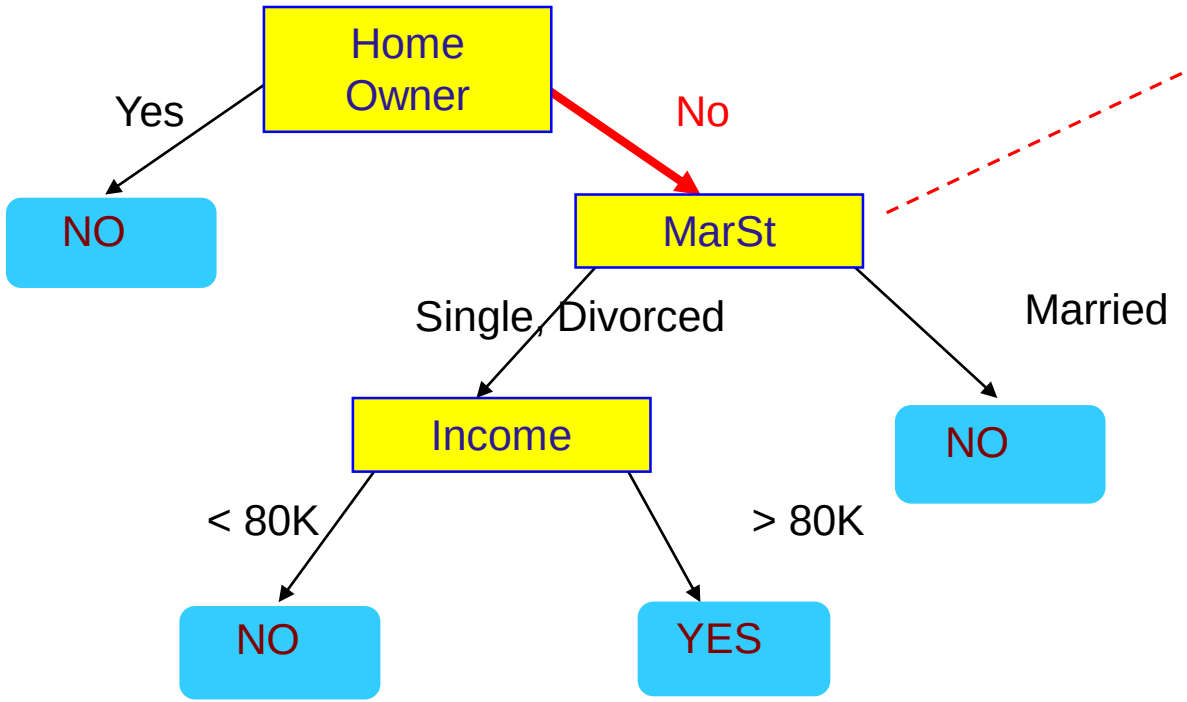
Home Owner	Marital Status	Annual Income	Defaulted Borrower
No	Married	80K	?



Apply Model to Test Data

Test Data

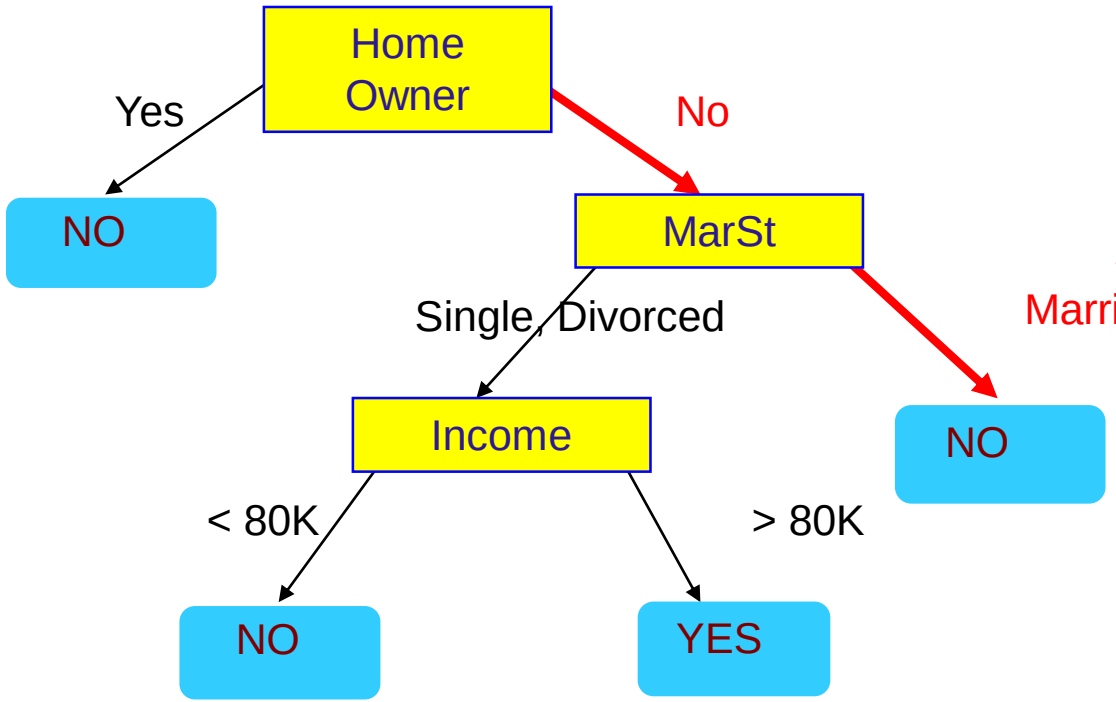
Home Owner	Marital Status	Annual Income	Defaulted Borrower
No	Married	80K	?



Apply Model to Test Data

Test Data

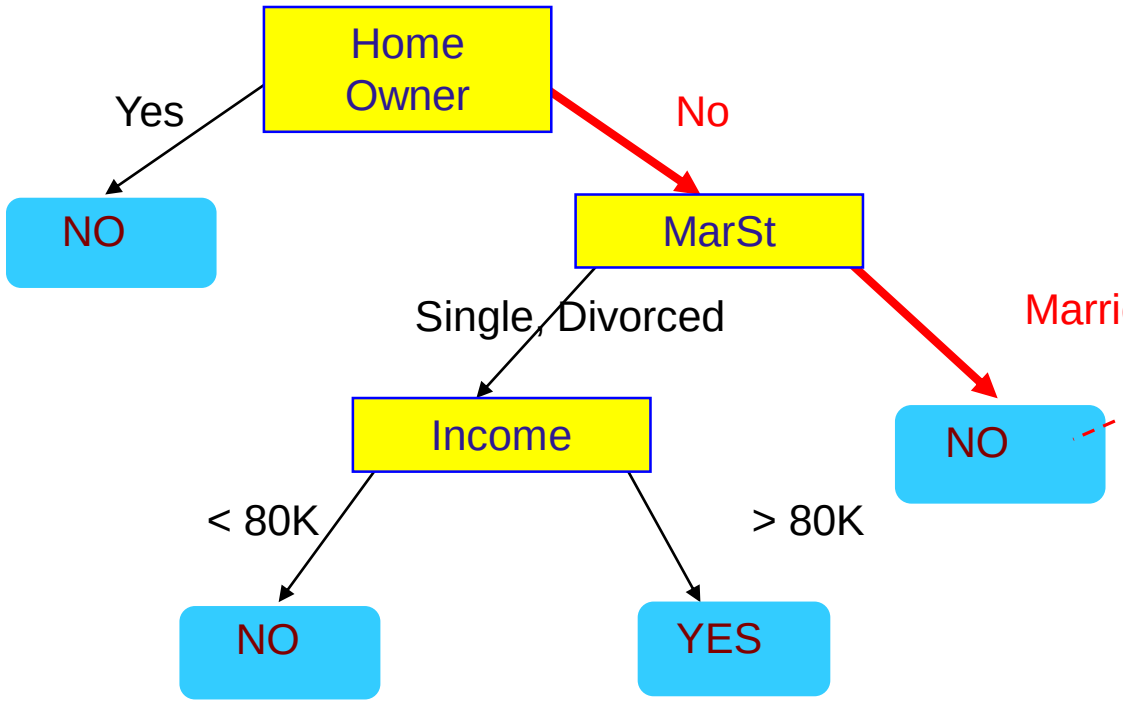
Home Owner	Marital Status	Annual Income	Defaulted Borrower
No	Married	80K	?



Apply Model to Test Data

Test Data

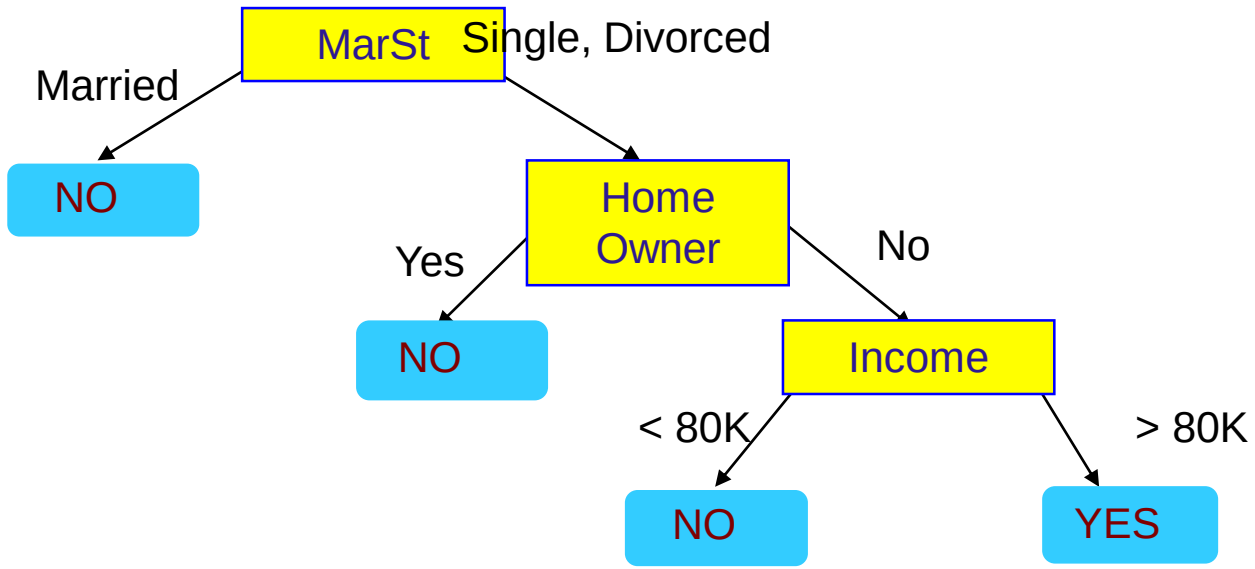
Home Owner	Marital Status	Annual Income	Defaulted Borrower
No	Married	80K	?



Assign Defaulted to “No”

Another Example of Decision Tree

ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
	categorical	categorical	continuous	class
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



There could be more than one tree that fits the same data!

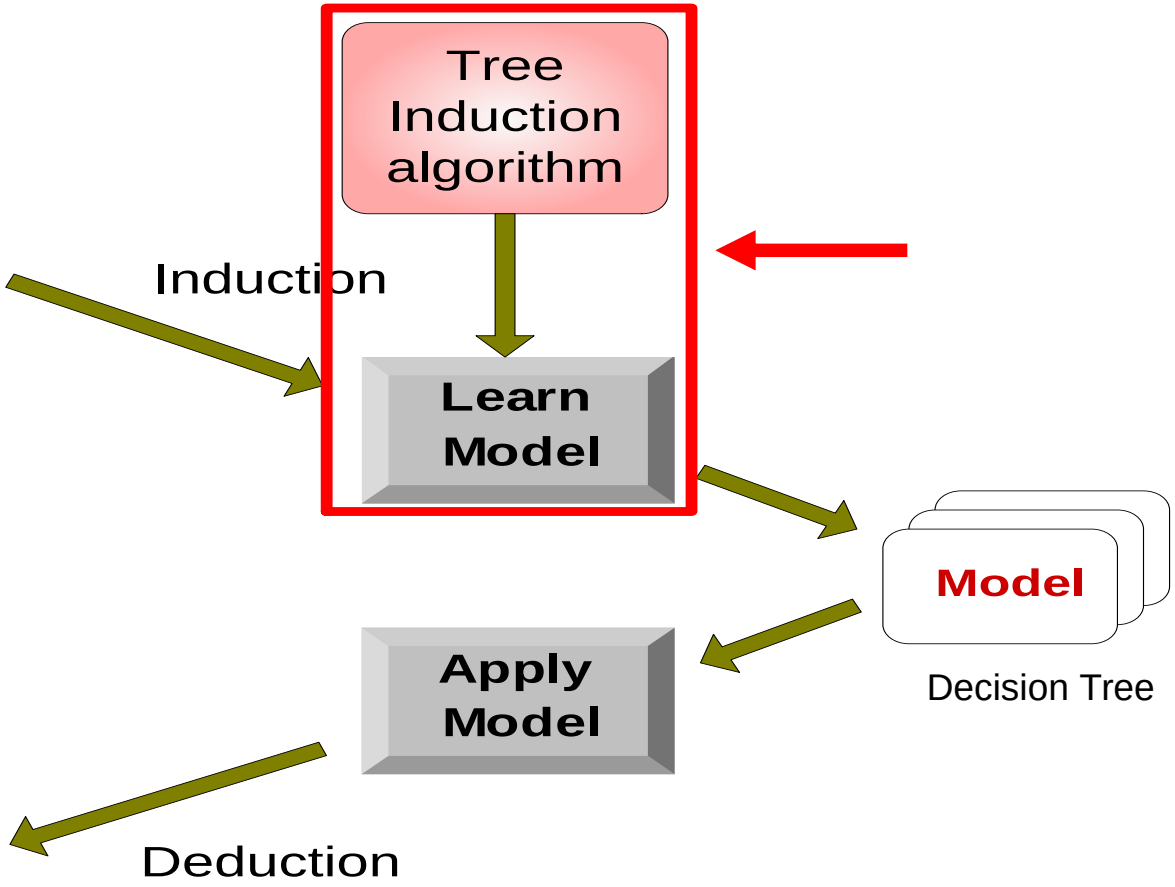
Decision Tree Classification Task

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set



Decision Trees - STEPS

- **Step1 :** Start with the **root node** in which all the data is present.
- **Step2:** Decide on a splitting criterion and stopping criteria.The root node is then split into two or more subsets leading to tree branches (called edges) using the splitting criterion. These are known as **internal nodes**. Each internal node has exactly one incoming edge.
- **Step3 :** Further divide each internal node until no further splitting is possible or the stopping criterion is met. The **terminal nodes** (aka **leaf nodes**) will not have any outgoing edges.
- **Step4 :** Terminal nodes are used for generating business rules.
- **Step 5 :** **Tree pruning** (a process for restricting the size of the tree) is used to avoid large trees and over-fitting the data. Tree pruning is achieved through different stopping criteria

Algorithm / steps

1.) Find the Entropy of class attribute

Entropy – class = -(P/P+N)log2(P/P+N) – N/(P+N)log2(N/P+N)

2.) Find the root Node

a) Information Gain

Information Gain= -(P/P+N)log2(P/P+N) – N/(P+N)log2(N/P+N)

b) Entropy (Attribute)

Entropy(Attribute)=SUM[(Pi+Ni)/(P+N) * I.G(Pi,Ni)]

c) Gain = Entropy(Class) – Entropy (Attribute)

3.) Find for all attributes and consider the attribute which has max value of gain ad assign as root note

4.) Repeat until the decision tree is complete

Construct a Decision using the following Training Dataset

Day	Outlook	Temp.	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Entropy of class attribute :

$$-[(9)/(9+5)]\log2\,[(9)/(9+5)] - [(5)/(9+5)]\log2\,[(5)/(9+5)] = 0.94$$

[Gain of outlook]

Outlook	P(Yes)	N(No)	I.G
Sunny	2	3	0.971
Rainy	3	2	0.971
overcast	4	0	0

I.G(sunny) = $-[(2)/(2+3)]\log2[(2)/(2+3)]-[(3)/(2+3)]\log2[(3)/(2+3)] = 0.97$

I.G(Rainy) = $-[(3)/(2+3)]\log2[(3)/(2+3)]-[(2)/(2+3)]\log2[(2)/(2+3)] = 0.97$

Entropy(outlook) = $[(2+3)/(9+5)] \times (0.971)+[(3+2)/(9+5)]\times(0.971) + [(4+0)/(9+5)] \times (0) = 0.693$

Gain(outlook) = 0.94 - 0.693 = **0.247**

[Gain for Temperature]

Temperature	P(Yes)	N(No)	I.G
Hot	2	2	1
Mild	4	2	0.918
Cold	3	1	0.811

I.G(Mild) = -[(4)/(4+2)]log2[(4)/(4+2)]-[(2)/(4+2)]log2[(2)/(4+2)] = 0.918

I.G(Cold) = -[(3)/(3+1)]log2[(3)/(3+1)]-[(1)/(3+1)]log2[(1)/(3+1)] = 0.811

Entropy(Temperature) = [(2+2)/(9+5)]x(1)+[(4+2)/(9+5)]x(0.918)+[(3+1)/(9+5)]x(0.811) = 0.911

Gain(Temperature) = 0.94 – 0.911 = **0.029**

[Gain for Humidity]

Humidity	P(Yes)	N(No)	I.G.
High	3	4	0.985
Normal	6	1	0.591

I.G(High) = -[(3)/(3+4)]log2[(3)/(3+4)]-[(4)/(3+4)]log2[(4)/(3+4)] = 0.985

I.G(Normal) = -[(6)/(6+1)]log2[(6)/(6+1)]-[(1)/(6+1)]log2[(1)/(6+1)] = 0.591

Entropy(Humidity) = [(3+4)/(9+5)](-.985)+[(6+1)/(9+5)](0.591) = 0.788

Gain(Humidity) = 0.94 -0.788 = **0.152**

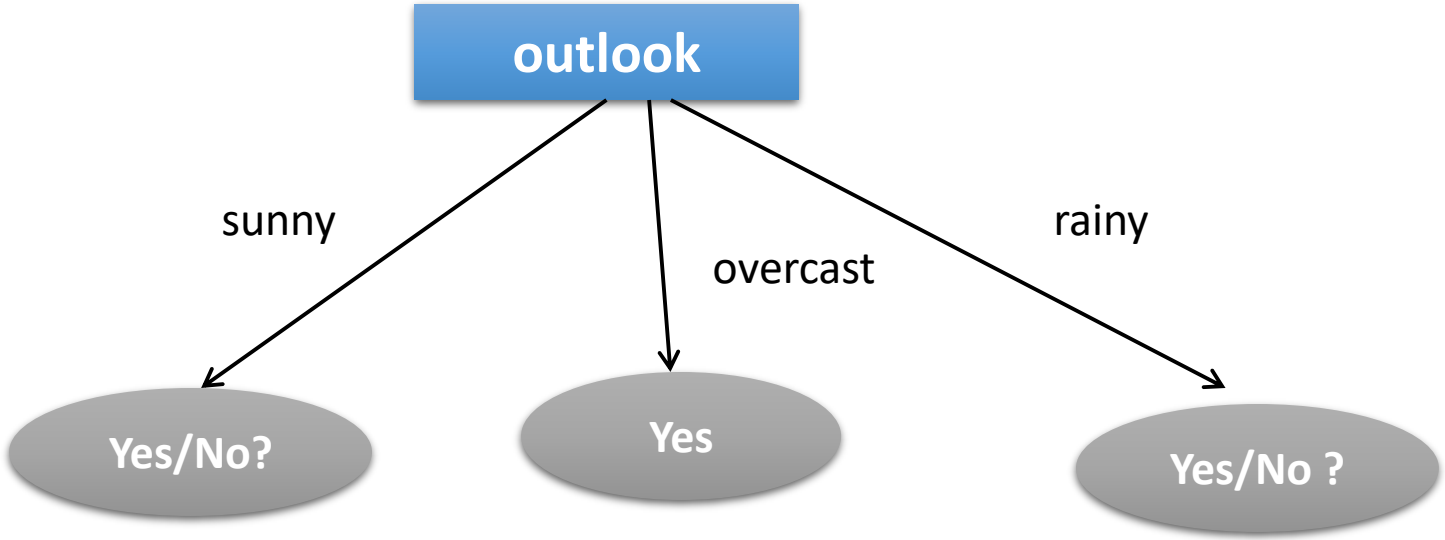
[Gain for windy]

wind	P(Yes)	N(No)	I.G.
Strong	3	3	1
weak	6	2	0.811

I.G(weak) = -[(6)/(6+2)]log2[(6)/(6+2)] -[(2)/(6+2)]log2[(2)/(6+2)] = 0.811

Entropy(wind) = [(3+3)/(9+5)]x(1)+[(6+2)/(9+5)]x(0.811) = 0.892

Gain(wind) = 0.94 – 0.892 = **0.048**



Sub table when Outlook is Sunny

outlook	Temp.	Humidity	Windy	Playtennis
sunny	Hot	High	Weak	No
sunny	Hot	High	Strong	No
sunny	Mild	High	Weak	No
sunny	Cool	Normal	Weak	Yes
sunny	mild	normal	strong	Yes

Entropy(sunny) = -[(2)/(2+3)]log2[(2)/(2+3)]-[(3)/(2+3)]log2[(3)/(2+3)] = **0.971**

Sub table when Outlook is overcast

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

Sub table when Outlook is Rainy

outlook	temp	humidity	windy	playtennis
Rainy	Mild	High	Weak	Yes
Rainy	Cool	Normal	Weak	Yes
Rainy	Cool	Normal	Strong	No
Rainy	Mild	Normal	Weak	Yes
Rainy	Mild	High	Strong	No

Gain of Temp(sunny)

Temp(sunny)	P(yes)	N(No)	I.G
Cool	1	0	0
Hot	0	2	0
mild	1	1	1

Entropy(Temp(sunny)) = [(1+0)/(2+3)]x(0)+[(0+2)/(2+3)]x(0)+[(1+1)/(2+3)]x(1) = 0.4

Gain(Temp(sunny)) = 0.971 – 0.4 = **0.571**

Gain of Humidity(sunny)

Humidity(sunny)	P(yes)	N(No)	I.G
High	0	3	0
Normal	2	0	0

Entropy(Humidity(sunny)) = 0

Gain(Humidity(sunny)) = 0.971 – 0 = **0.971**

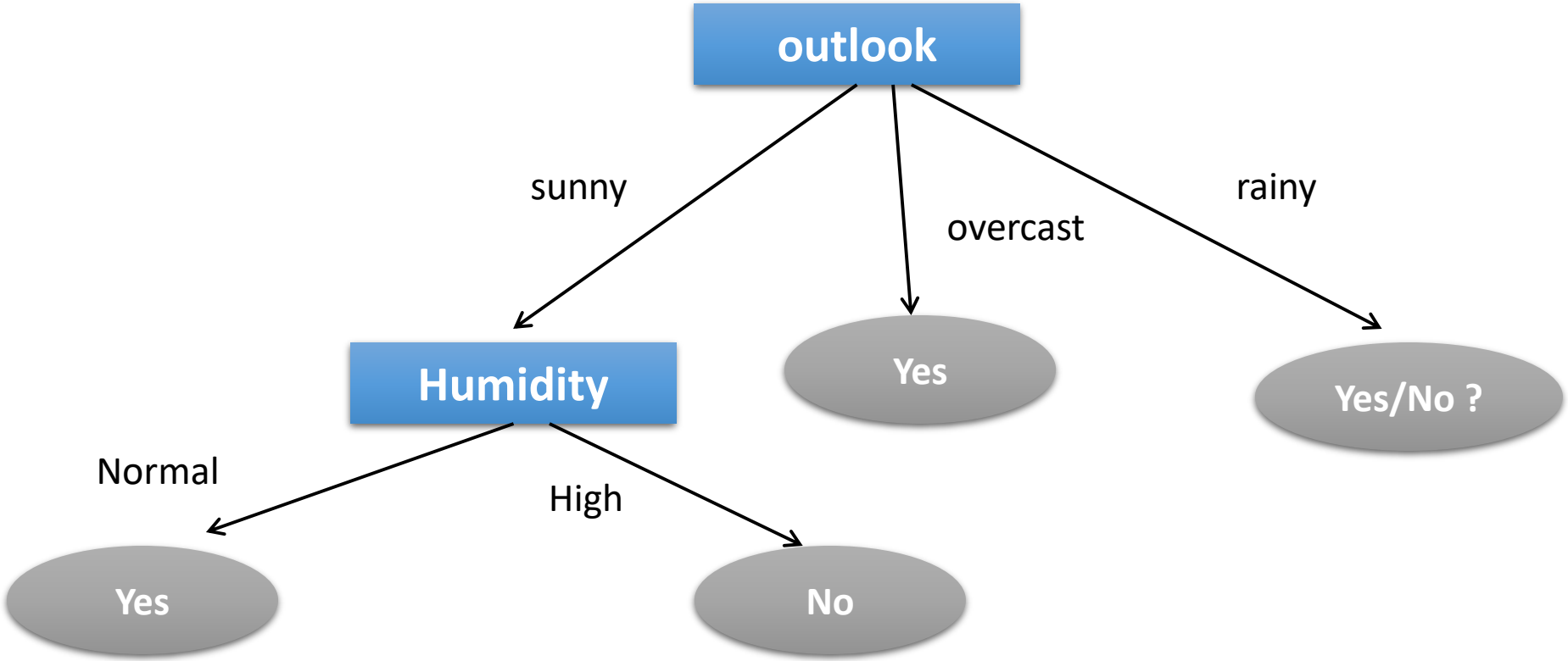
Gain(windy(sunny))

windy(sunny)	P(yes)	N(No)	I.G
Strong	1	1	1
weak	1	2	0.981

IG(weak) = -[(1)/(1+2)]log2[(1)/(1+2)] -[(2)/(1+2)]log2[(2)/(1+2)] = 0.981

Entropy(windy(sunny)) = [(1+1)/(2+3)]x(1)+[(1+2)/(2+3)]x(0.981) = 0.951

Gain(windy(sunny)) = 0.971 – 0.951 = **0.020**



outlook	temp	humidity	windy	playtennis
Rainy	Mild	High	Weak	Yes
Rainy	Cool	Normal	Weak	Yes
Rainy	Cool	Normal	Strong	No
Rainy	Mild	Normal	Weak	Yes
Rainy	Mild	High	Strong	No

Entropy(outlook(rainy)) = -[(3)/(3+2)]log2[(3)/(3+2)]-[(2)/(3+2)]log2[(2)/(3+2)] = **0.971**

Gain(Humidity(Rainy))

Humidity(rainy)	**P(yes)**	**N(No)**	**I.G**
High	1	1	1
Normal	2	1	0.918

Entropy(Humidity(Rainy)) = [(1+1)/(3+2)]x(1)+[(2+1)/(3+2)]x(0.918) = 0.951

Gain(Humidity(Rainy)) = 0.971 – 0.951 = **0.020**

Gain of windy(rainy)

windy(sunny)	P(yes)	N(No)	I.G
Strong	0	2	0
weak	3	0	0

Entropy(windy(rainy)) = 0

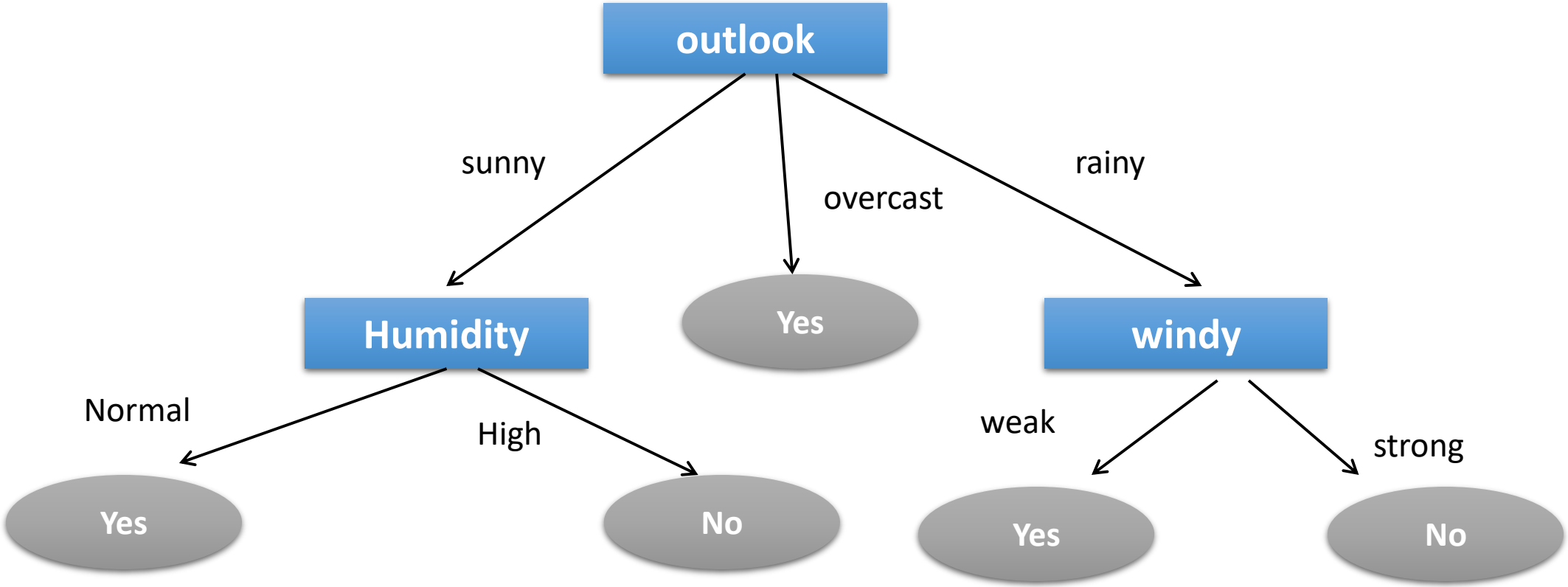
Gain(windy(rainy)) = 0.971 – 0 = **0.971**

Gain of temperature(rainy)

Temp(sunny)	P(yes)	N(No)	I.G
Mild	2	1	0.918
cool	1	1	1

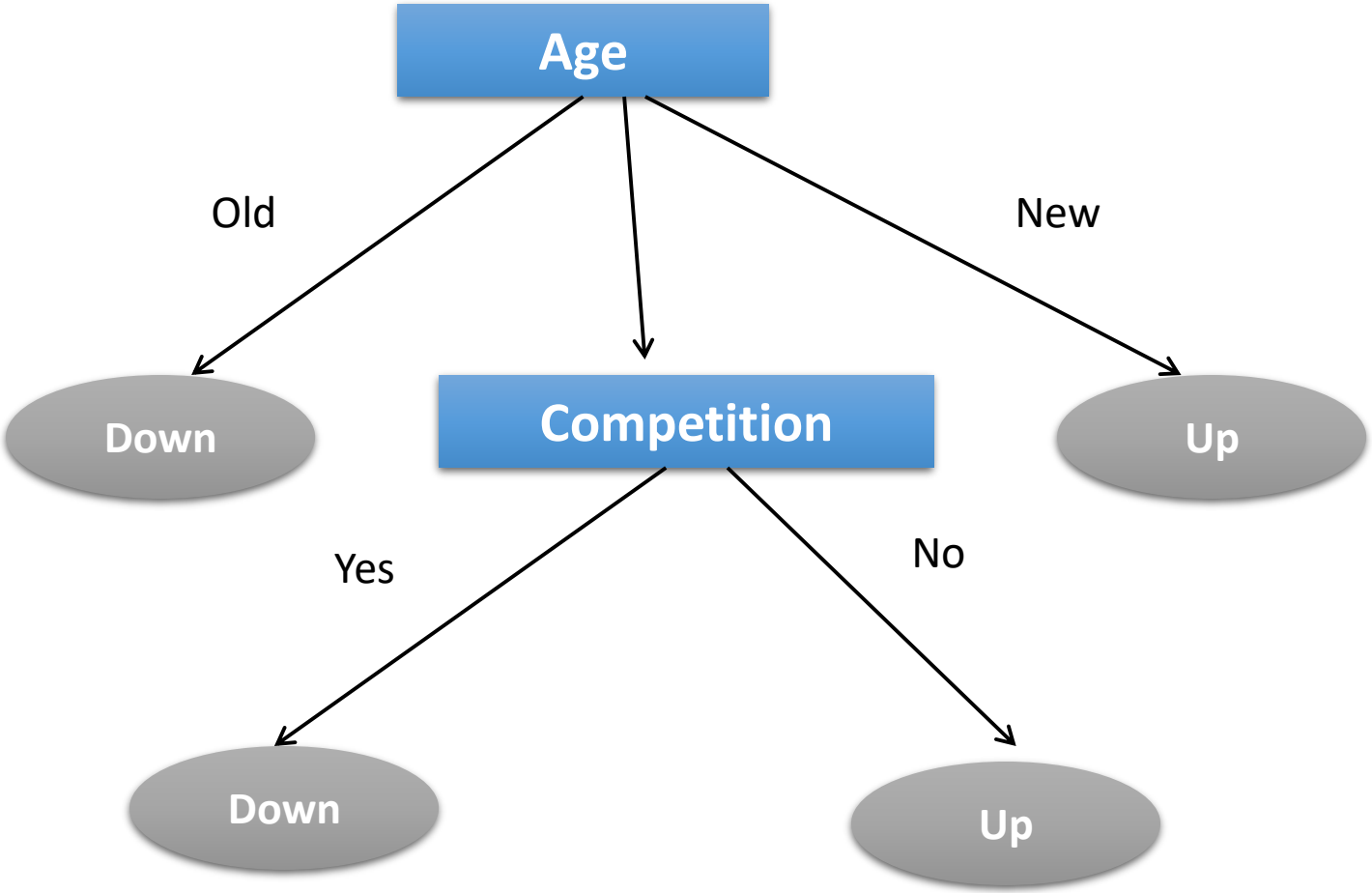
Entropy(temperature(rainy)) = [(2+1)/(3+2)]x(0.918)+[(1+1)/(3=2)]x(1) =0.951

Gain(temperature(rainy)) = 0.971 – 0.951 = **0.02**


outlook
sunny
overcast
rainy
Humidity
Yes
windy
Normal
High
weak
strong
Yes
No
Yes
No

Construct a Decision Tree for the following dataset

Age	Competition	Type	Profit
Old	Yes	Software	Down
Old	No	Software	Down
Old	No	Hardware	Down
Mid	Yes	Software	Down
Mid	Yes	Hardware	Down
Mid	No	Hardware	Up
Mid	No	Software	Up
New	Yes	Software	Up
New	No	Hardware	Up
New	No	Software	Up
Mid	No	Hardware	?


Age
Old
New
Down
Competition
Up
Yes
No
Down
Up

CHI-SQUARE AUTOMATIC INTERACTION DETECTION (CHAID)

- CHAID is an extension of Automatic Interaction Detection (AID), which is designed to categorize the dependent variable using categorical predictors.
- CHAID trees use statistical significance of independent variables to split the subset of the data (represented by nodes of the tree).

$$E_i = \frac{(\text{row total} \times \text{column total})}{\text{Table total}}$$

The formula for Chi-Square t-statistic is defined as:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Here, O_i = observed frequency and E_i = expected frequency.

Construct a Decision using the following Training Dataset using CHAID

Day	Outlook	Temp.	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Weak	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cold	Normal	Weak	Yes
D10	Rain	Mild	Normal	Strong	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Contingency table for outlook

	Observed			Expected	
	Yes	No	Total	Yes	No
Overcast	4	0	4	2.57	1.42
Sunny	2	3	5	3.21	1.78
rainy	3	2	5	3.21	1.78
Total	9	5	14	9	5

$$\chi^2 = \frac{(4-2.57)^2}{2.57} + \frac{(0-1.42)^2}{1.42} + \frac{(2-3.21)^2}{3.21} + \frac{(3-1.78)^2}{1.78} + \frac{(3-3.21)^2}{3.21} + \frac{(2-1.78)^2}{1.78}$$

= 0.795+1.42+0.456+0.836+0.013+0.027 = 3.547

Contingency table for temperature

	Observed			Expected	
	Yes	No	Total	Yes	No
Hot	2	2	4	2.57	1.42
Mild	4	2	6	3.85	2.14
Cool	3	1	4	2.57	1.42
Total	9	5	14	9	5

$$\chi^2 = \frac{(2-2.57)^2}{2.57} + \frac{(2-1.42)^2}{1.42} + \frac{(4-3.85)^2}{3.85} + \frac{(2-2.14)^2}{2.14} + \frac{(3-2.57)^2}{2.57} + \frac{(1-1.42)^2}{1.42}$$

=0.126+0.2336+0.005+0.009+0.0.71+0.142 = 0.589

Contingency table for humidity

	Observed			Expected	
	Yes	No	Total	Yes	No
High	3	4	7	4.5	2.5
Normal	6	1	7	4.5	2.5
Total	9	5	14	9	5

$$\chi^2 = \frac{(3-4.5)^2}{4.5} + \frac{(4-2.5)^2}{2.5} + \frac{(6-4.5)^2}{4.5} + \frac{(1-2.5)^2}{2.5}$$

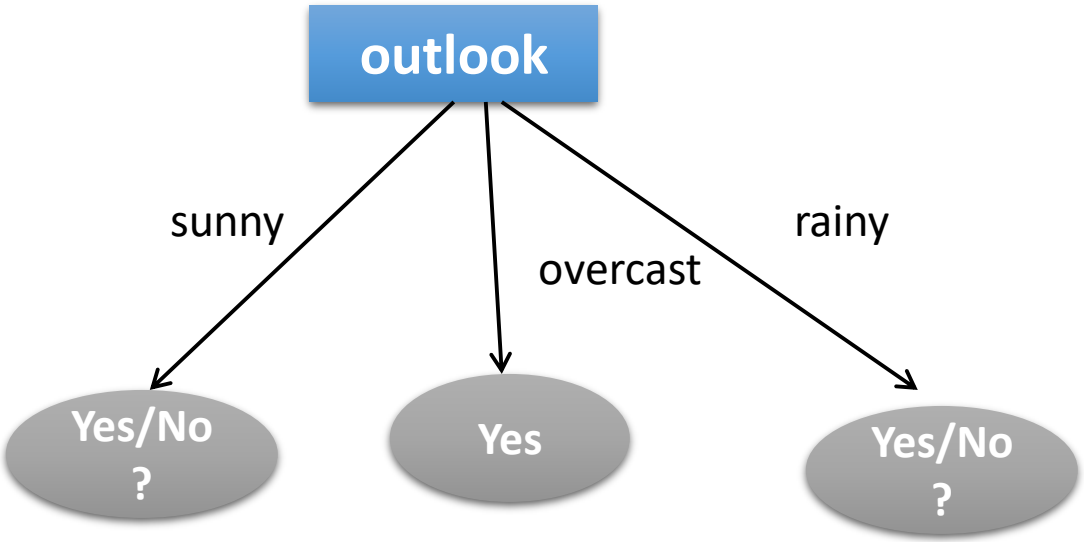
=0.5+0.9+0.5+0.9 = 2.8

Contingency table for Wind

	Observed			Expected	
	Yes	No	Total	Yes	No
Weak	6	2	8	5.14	2.85
STrong	3	3	6	3.85	2.14
Total	9	5	14	9	5

$$\chi^2 = \frac{(6-5.14)^2}{5.14} + \frac{(2-2.85)^2}{2.85} + \frac{(3-3.85)^2}{3.85} + \frac{(3-2.14)^2}{2.14}$$

=0.143+0.253+0.187+0.345= 0.928



Outlook	3.547
Temp	0.589
Humidity	2.8
wind	0.928

outlook	Temp.	Humidity	Windy	Playtennis
sunny	Hot	High	Weak	No
sunny	Hot	High	Strong	No
sunny	Mild	High	Weak	No
sunny	Cool	Normal	Weak	Yes
sunny	mild	normal	strong	Yes

outlook	temp	humidity	windy	playtennis
Rainy	Mild	High	Weak	Yes
Rainy	Cool	Normal	Weak	Yes
Rainy	Cool	Normal	Strong	No
Rainy	Mild	Normal	Weak	Yes
Rainy	Mild	High	Strong	No

Outlook ,sunny = playtennis

outlook	Temp.	Humidity	Windy	Playtennis
sunny	Hot	High	Weak	No
sunny	Hot	High	Strong	No
sunny	Mild	High	Weak	No
sunny	Cool	Normal	Weak	Yes
sunny	mild	normal	strong	Yes

Contingency table for temperature, when outlook = sunny

	Observed			Expected	
	Yes	No	Total	Yes	No
Hot	0	2	2	0.8	1.2
Mild	1	1	2	0.8	1.2
Cool	1	0	1	0.4	0.6
Total	2	3	5	2	3

$$\chi^2 = \frac{(0-0.8)^2}{0.8} + \frac{(2-1.2)^2}{1.2} + \frac{(1-0.8)^2}{0.8} + \frac{(1-1.2)^2}{1.2} + \frac{(1-0.4)^2}{0.4} + \frac{(0-0.6)^2}{0.6}$$

=0.8+0.53+0.05+0.03+0.9+0.6 = 2.91

Outlook ,sunny = playtennis

outlook	Temp.	Humidity	Windy	Playtennis
sunny	Hot	High	Weak	No
sunny	Hot	High	Strong	No
sunny	Mild	High	Weak	No
sunny	Cool	Normal	Weak	Yes
sunny	mild	normal	strong	Yes

Contingency table for humidity, when outlook = sunny

	Observed			Expected	
	Yes	No	Total	Yes	No
High	0	3	3	1.2	1.8
Normal	2	0	2	0.5	1.2
Total	2	3	5	2	3

$$\chi^2 = \frac{(0-1.2)^2}{1.2} + \frac{(3-1.8)^2}{1.8} + \frac{(2-0.5)^2}{0.5} + \frac{(0-1.2)^2}{1.2}$$

=1.2+0.8+4.5+1.2= 7.7

Outlook ,sunny = playtennis

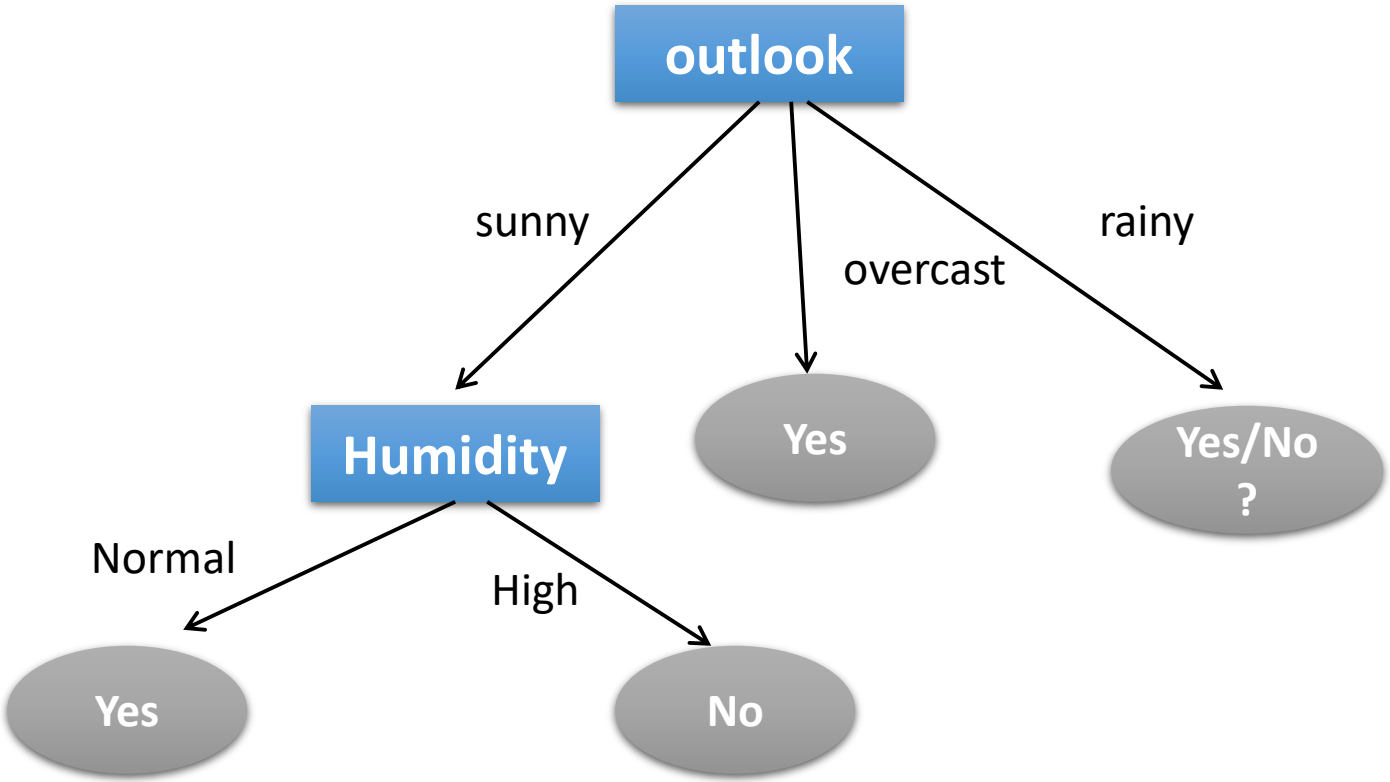
outlook	Temp.	Humidity	Windy	Playtennis
sunny	Hot	High	Weak	No
sunny	Hot	High	Strong	No
sunny	Mild	High	Weak	No
sunny	Cool	Normal	Weak	Yes
sunny	mild	normal	strong	Yes

Contingency table for wind, when outlook = sunny

	Observed			Expected	
	Yes	No	Total	Yes	No
Weak	1	2	3	1.2	1.8
Strong	1	1	2	0.8	1.2
Total	2	3	5	2	3

$$\chi^2 = \frac{(1-1.2)^2}{1.2} + \frac{(2-1.8)^2}{1.8} + \frac{(1-0.8)^2}{0.8} + \frac{(1-1.2)^2}{1.2}$$

=0.03+0.02+0.05+0.03= 0.13


outlook
sunny
overcast
rainy
Humidity
Yes
Yes/No
?
Normal
High
Yes
No

outlook	temp	humidity	windy	playtennis
Rainy	Mild	High	Weak	Yes
Rainy	Cool	Normal	Weak	Yes
Rainy	Cool	Normal	Strong	No
Rainy	Mild	Normal	Weak	Yes
Rainy	Mild	High	Strong	No

Contingency table for temp, when outlook = rainy

	Observed			**Expected**	
	Yes	No	Total	Yes	No
Mild	2	1	3	1.8	1.2
Cool	1	1	2	1.2	0.8
Total	3	2	5	3	2

$$\chi^2 = \frac{(2-1.8)^2}{1.8} + \frac{(1-1.2)^2}{1.2} + \frac{(1-1.2)^2}{1.2} + \frac{(1-0.8)^2}{0.8}$$

=0.02+0.03+0.03+0.05= 0.13

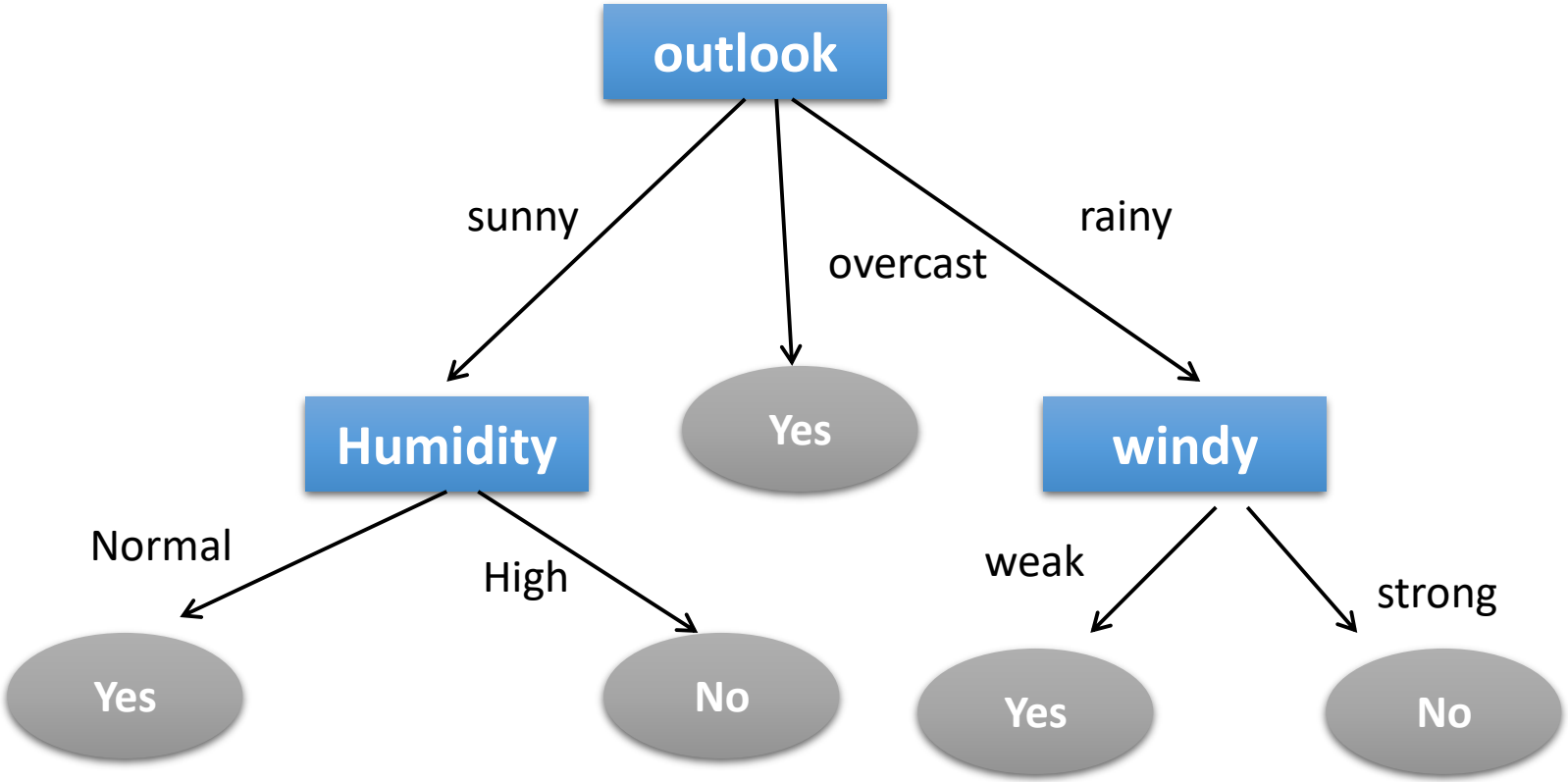
outlook	temp	humidity	windy	playtennis
Rainy	Mild	High	Weak	Yes
Rainy	Cool	Normal	Weak	Yes
Rainy	Cool	Normal	Strong	No
Rainy	Mild	Normal	Weak	Yes
Rainy	Mild	High	Strong	No

Contingency table for wind, when outlook = rainy

	Observed			**Expected**	
	Yes	No	Total	Yes	No
Weak	3	0	3	1.8	1.2
Strong	0	2	2	1.2	0.8
Total	3	2	5	3	2

$$\chi^2 = \frac{(3-1.8)^2}{1.8} + \frac{(0-1.2)^2}{1.2} + \frac{(0-1.2)^2}{1.2} + \frac{(2-0.8)^2}{0.8}$$

=0.8+1.2+1.2+1.8= 5


outlook
sunny
overcast
rainy
Humidity
Yes
windy
Normal
High
weak
strong
Yes
No
Yes
No

CLASSIFICATION AND REGRESSION TREE (CART)

- Classification and Regression Tree (CART) is a common terminology that is used for a **Classification Tree** (used when the dependent variable is discrete) and a **Regression Tree** (used when the dependent variable is continuous).

- Classification tree uses various impurity measures such as the Gini Impurity Index

- Regression Tree, on the other hand, splits the node that minimizes the Sum of Squared Errors

ALGORITHM - STEPS

The following steps are used to generate a classification and a regression tree (Breiman *et al.* 1984)

- **Step 1 :** Start with the complete training data in the **root node**.
- **Step 2 :** Decide on the **measure of impurity** (usually Gini impurity index or Entropy). Choose a predictor variable that minimizes the impurity when the parent node is split into **children nodes**

This happens when the original data is divided into two subsets using a predictor variable such that it results in the maximum reduction in the impurity in the case of discrete dependent variable or the maximum reduction in SSE in the case of a continuous dependent variable.

❑ **Step 3 :** Repeat step 2 for each subset of the data (for each **internal node**) using the independent variables until:

- ✓ All the dependent variables are exhausted .
- ✓ The stopping criteria is met. Few stopping criteria used are number of levels of tree from the root node, minimum number of observations in parent/child node (eg. 10% of the training data) and minimum reduction in impurity index.

❑ **Step 4 :** Generate business rules for the **leaf (terminal) nodes** of the tree.

Gini Impurity Index

Gini impurity index is one of the measures of impurity that is used by classification trees to split the nodes.

$$GI(t) = \sum_{i=1}^{K} \sum_{j=1, j \neq i}^{K} P(C_i|t)P(C_j|t) = \sum_{i=1}^{K} P(C_i \mid t)(1 - P(C_i \mid t) = 1 - \sum_{i=1}^{K} [P(C_i \mid t)]^2$$

where

$GI(t)$ = Gini index at node t

$P(C_i/t)$ = Proportion of observations belonging to class C_i in node t

Construct a Decision using the following Training Dataset using CART

Day	Outlook	Temp.	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Outlook	Yes	No	Total
Sunny	2	3	5
Overcast	4	0	4
Rain	3	2	5

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) =

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

Temp	Yes	No	Total
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

Humidity	Yes	No	Total
High	3	4	7
Normal	6	1	7

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

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

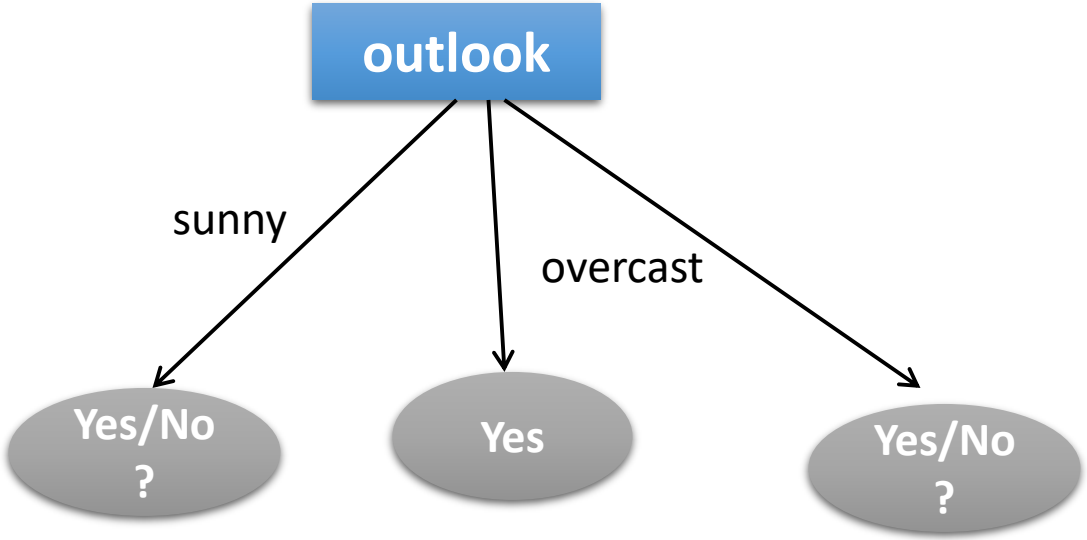
Weighted sum for humidity feature will be calculated next

Gini(Humidity) = (7/14) x 0.489 + (7/14) x 0.244 = 0.367

Wind	Yes	No	Total
Weak	6	2	8
Strong	3	3	6

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) x 0.375 + (6/14) x 0.5 = 0.428

Feature	Gini index
Outlook	0.342
Temperature	0.439
Humidity	0.367
Wind	0.428



outlook	Temp.	Humidity	Windy	Playtennis
sunny	Hot	High	Weak	No
sunny	Hot	High	Strong	No
sunny	Mild	High	Weak	No
sunny	Cool	Normal	Weak	Yes
sunny	mild	normal	strong	Yes

outlook	temp	humidity	windy	playtennis
Rainy	Mild	High	Weak	Yes
Rainy	Cool	Normal	Weak	Yes
Rainy	Cool	Normal	Strong	No
Rainy	Mild	Normal	Weak	Yes
Rainy	Mild	High	Strong	No

outlook	Temp.	Humidity	Windy	Playtennis
sunny	Hot	High	Weak	No
sunny	Hot	High	Strong	No
sunny	Mild	High	Weak	No
sunny	Cool	Normal	Weak	Yes
sunny	mild	normal	strong	Yes

Temp	Yes	No	Total
Hot	0	2	2
Cool	1	0	1
Mild	1	1	2

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.) = $(2/5)x0 + (1/5)x0 + (2/5)x0.5 = 0.2$

Humidity	Yes	No	Total
High	0	3	3
Normal	2	0	2

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) = $(3/5)x0 + (2/5)x0 = 0$

outlook	Temp.	Humidity	Windy	Playtennis
sunny	Hot	High	Weak	No
sunny	Hot	High	Strong	No
sunny	Mild	High	Weak	No
sunny	Cool	Normal	Weak	Yes
sunny	mild	normal	strong	Yes

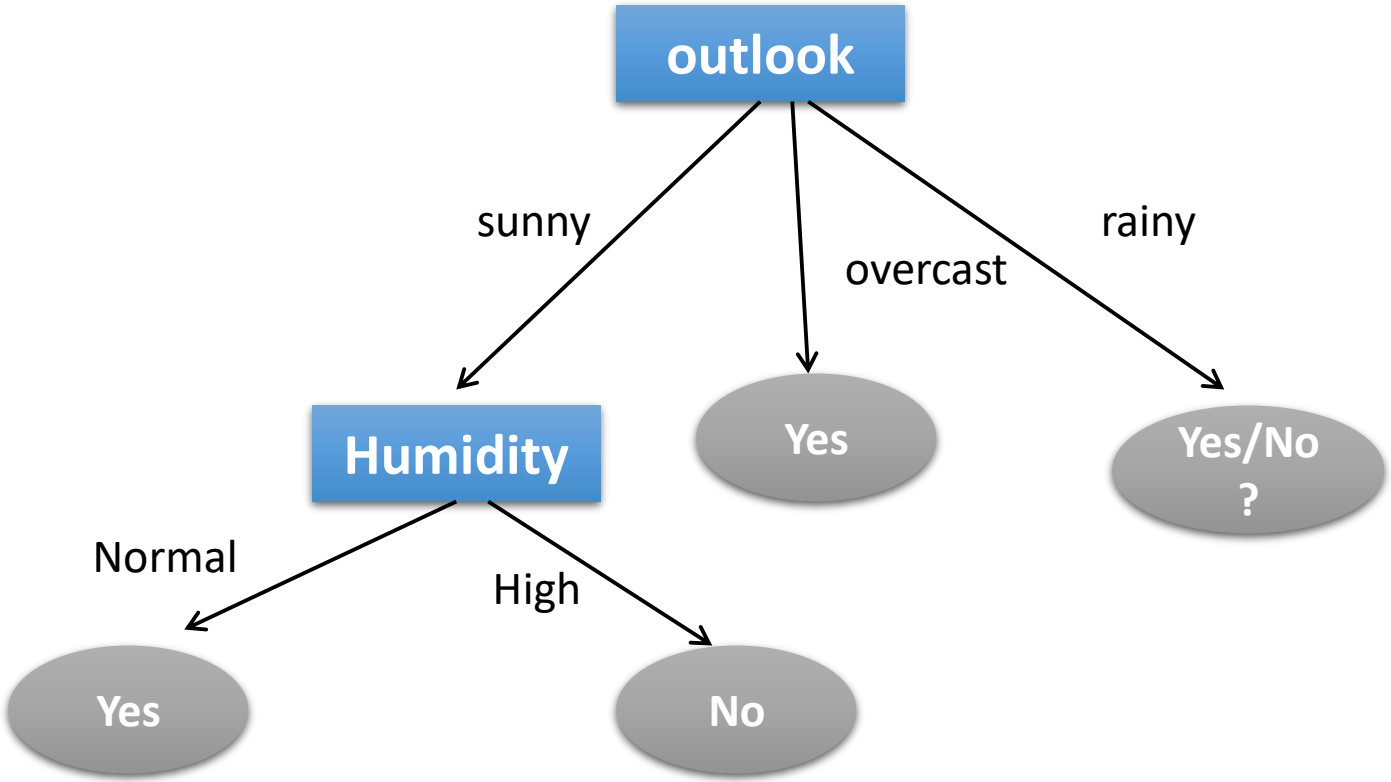
Wind	Yes	No	Total
Weak	1	2	3
Strong	1	1	2

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) = $(3/5)x0.266 + (2/5)x0.2 = 0.466$

Feature	Gini index
Temperature	0.2
Humidity	0
Wind	0.466



outlook	temp	humidity	windy	playtennis
Rainy	Mild	High	Weak	Yes
Rainy	Cool	Normal	Weak	Yes
Rainy	Cool	Normal	Strong	No
Rainy	Mild	Normal	Weak	Yes
Rainy	Mild	High	Strong	No

Temp	Yes	No	Total
Cool	1	1	2
Mild	2	1	3

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.) = $(2/5)x0.5 + (3/5)x0.444 = 0.467$

Humidity	Yes	No	Total
High	1	1	2
Normal	2	1	3

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) = $(2/5)x0.5 + (3/5)x0.444 = 0.467$

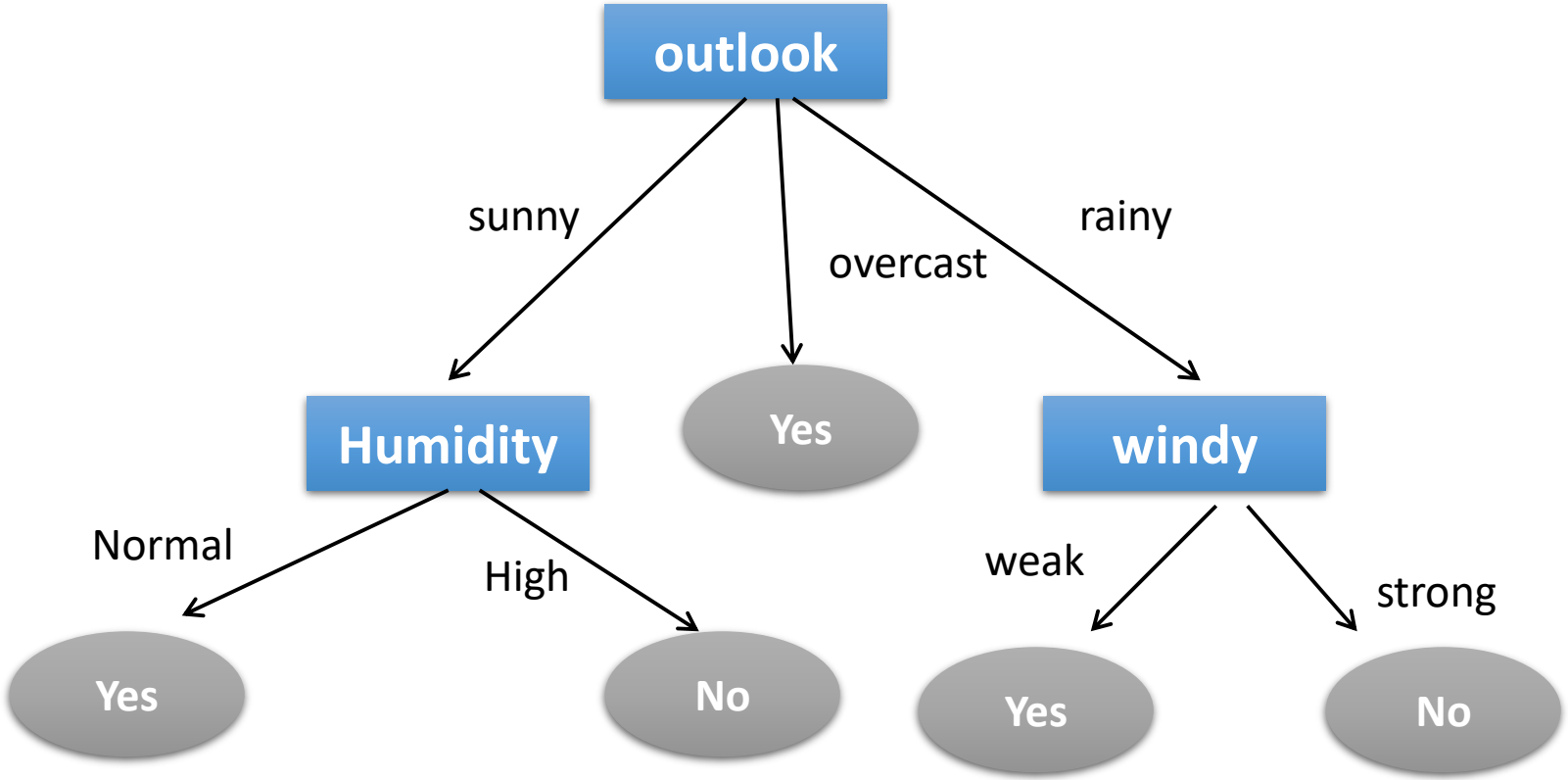
Wind	Yes	No	Total
Weak	3	0	3
Strong	0	2	2

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) = $(3/5)x0 + (2/5)x0 = 0$

Temperature	Gini index
Temperature	0.466
Humidity	0.466
Wind	0



Construct a Decision using the following Training Dataset using CART

Day	Outlook	Temp	Humidity	Wind	Tennis?
D1	Sunny	85	85	Weak	No
D2	Sunny	90	90	Strong	No
D3	Overcast	83	86	Weak	Yes
D4	Rain	70	96	Weak	Yes
D5	Rain	68	80	Weak	Yes
D6	Rain	65	70	Strong	No
D7	Overcast	64	65	Strong	Yes
D8	Sunny	72	95	Weak	No
D9	Sunny	69	70	Weak	Yes
D10	Rain	75	80	Weak	Yes
D11	Sunny	75	70	Strong	Yes
D12	Overcast	72	90	Strong	Yes
D13	Overcast	81	75	Weak	Yes
D14	Rain	71	91	Strong	No

Outlook	Yes	No	Total
Sunny	2	3	5
Overcast	4	0	4
Rain	3	2	5

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) =

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

64	Yes
65	No
68	Yes
69	Yes
70	Yes
71	No
72	No
72	Yes
75	Yes
75	Yes
81	Yes
83	Yes
85	No
90	No

Temp	Yes	No	Total
<=64	1	0	1
>64	8	5	13

Gini(<=64) = $1 - (1/1)^2 - (0/1)^2 = 0$

Gini(>64) = $1 - (8/13)^2 - (5/13)^2 = 0.473$

Gini(64) = (1/14) x 0 + (13/14) x 0.473 = **0.439**

Temp	Yes	No	Total
<=65	1	1	2
>65	8	4	12

Gini(<=65) = $1 - (1/2)^2 - (1/2)^2 = 0.5$

Gini(>65) = $1 - (8/12)^2 - (4/12)^2 = 0.44$

Gini(65) = (2/14) x 0.5 + (12/14) x 0.44 =**0.448**

Temp	Yes	No	Total
<=68	2	1	3
>68	7	4	11

Gini(<=68) = $1 - (2/3)^2 - (1/3)^2 = 0.44$

Gini(>68) = $1 - (7/11)^2 - (4/11)^2 = 0.46$

Gini(68) = (3/14) x 0.4 + (11/14) x 0.46 =**0.455**

Temp	Yes	No	Total
<=69	2	1	3
>69	7	4	11

Gini(<=69) = $1 - (3/4)^2 - (1/4)^2 = 0.375$

Gini(>69)= $1 - (6/10)^2 - (4/10)^2 = 0.48$

Gini(69)=(4/14) x 0.375 + (10/14) x 0.48 = =**0.45**

64	Yes
65	No
68	Yes
69	Yes
70	Yes
71	No
72	No
72	Yes
75	Yes
75	Yes
81	Yes
83	Yes
85	No
90	No

Temp	Yes	No	Total
<=70	4	1	5
>70	5	4	9

Gini(<=70) = $1 - (4/5)^2 - (1/5)^2 = 0.32$

Gini(>70) = $1 - (5/9)^2 - (4/9)^2 = 0.49$

Gini(70) = (5/14) x 0.23 + (9/14) x 0.49 **= 0.429**

Temp	Yes	No	Total
<=71	2	4	6
>71	5	3	8

Gini(<=71) = $1 - (2/6)^2 - (4/6)^2 = 0.44$

Gini(>71) = $1 - (5/8)^2 - (3/8)^2 = 0.468$

Gini(71) = (6/14) x 0.44 + (8/14) x 0.468 **= 0.456**

Temp	Yes	No	Total
<=72	5	3	8
>72	4	2	6

Gini(<=72) = $1 - (5/8)^2 - (3/8)^2 = 0.468$

Gini(>72) = $1 - (4/6)^2 - (2/6)^2 = 0.44$

Gini(72) = (8/14) x 0.468 + (6/14) x 0.44 **= 0.457**

Temp	Yes	No	Total
<=75	7	3	10
>75	2	2	4

Gini(<=75) = $1 - (7/10)^2 - (3/10)^2 = 0.42$

Gini(>75) = $1 - (2/4)^2 - (2/4)^2 = 0.5$

Gini(75) = (10/14) x 0.42 + (4/14) x 0.5 **= 0.442**

64	Yes
65	No
68	Yes
69	Yes
70	Yes
71	No
72	No
72	Yes
75	Yes
75	Yes
81	Yes
83	Yes
85	No
90	No

Temp	Yes	No	Total
<=81	8	3	11
>81	1	2	3

Gini(<=81) = $1 - (8/11)^2 - (3/11)^2 = 0.396$

Gini(>81) = $1 - (1/3)^2 - (2/3)^2 = 0.44$

Gini(81) = (11/14) x 0.396 + (3/14) x 0.44 **= 0.405**

Temp	Yes	No	Total
<=83	9	3	12
>83	0	2	2

Gini(<=83) = $1 - (9/12)^2 - (3/12)^2 = 0.375$

Gini(>83) = $1 - (0/2)^2 - (2/2)^2 = 0$

Gini(83) = (12/14) x 0.375 + (2/14) x 0 **= 0.321**

Temp	Yes	No	Total
<=85	9	4	13
>85	0	2	2

Gini(<=85) = $1 - (9/13)^2 - (4/13)^2 = 0.426$

Gini(>85) = $1 - (0/1)^2 - (1/1)^2 = 0$

Gini(85) = (13/14) x 0.426 + (1/14) x 0 **= 0.395**

Temp	Yes	No	Total
<=90	9	5	14
>90	0	0	0

Gini(<=90) = $1 - (9/14)^2 - (5/14)^2 = 0.459$

Gini(>90) = $1 - (0/0)^2 - (0/0)^2 = 1$

Gini(90) = (14/14) x 0.459 + (0/14) x 1 **= 0.459**

Humidity	Yes	No	Total
<=65	1	0	1
>65	8	5	13

Gini(<=65) = $1 - (1/1)^2 - (0/1)^2 = 0$

Gini(>65) = $1 - (8/13)^2 - (5/13)^2 = 0.473$

Gini(65) = (1/14) x 0 + (13/14) x 0.473 **= 0.439**

65	Yes
70	No
70	Yes
70	Yes
75	Yes
80	Yes
80	Yes
85	No
86	Yes
90	No
90	Yes
91	No
95	No
96	Yes

Humidity	Yes	No	Total
<=70	3	1	4
>70	6	4	10

Gini(<=70) = $1 - (3/4)^2 - (1/4)^2 = 0.375$

Gini(>70) = $1 - (6/10)^2 - (4/10)^2 = 0.48$

Gini(70) = (4/14) x 0.375 + (10/14) x 0.48 **= 0.45**

Humidity	Yes	No	Total
<=75	4	1	5
>75	5	4	9

Gini(<=75) = $1 - (4/5)^2 - (1/5)^2 = 0.32$

Gini(>75) = $1 - (5/9)^2 - (4/9)^2 = 0.493$

Gini(75) = (5/14) x 0.32 + (9/14) x 0.493 **= 0.431**

Humidity	Yes	No	Total
<=80	6	1	7
>80	3	4	7

Gini(<=80) = $1 - (6/7)^2 - (1/7)^2 = 0.244$

Gini(>80) = $1 - (3/7)^2 - (4/7)^2 = 0.489$

Gini(80) = (7/14) x 0.244 + (7/14) x 0.489 **= 0.366**

Humidity	Yes	No	Total
<=85	6	2	8
>85	3	3	6

Gini(<=85) = $1 - (6/8)^2 - (2/8)^2 = 0.375$

Gini(>85) = $1 - (3/6)^2 - (3/6)^2 = 0.5$

Gini(85) = (8/14) x 0.375 + (6/14) x 0.5 **= 0.428**

65	Yes
70	No
70	Yes
70	Yes
75	Yes
80	Yes
80	Yes
85	No
86	Yes
90	No
90	Yes
91	No
95	No
96	Yes

Humidity	Yes	No	Total
<=86	7	2	9
>86	2	3	5

Gini(<=86) = $1 - (7/9)^2 - (2/9)^2 = 0.34$

Gini(>86) = $1 - (2/5)^2 - (3/5)^2 = 0.48$

Gini(86) = (9/14) x 0.34 + (5/14) x 0.48 **= 0.39**

Humidity	Yes	No	Total
<=90	8	3	11
>90	1	2	3

Gini(<=90) = $1 - (8/11)^2 - (3/11)^2 = 0.396$

Gini(>90) = $1 - (1/3)^2 - (2/3)^2 = 0.44$

Gini(90) = (11/14) x 0.396 + (3/14) x 0.44 **= 0.405**

Humidity	Yes	No	Total
<=91	8	4	12
>91	1	1	2

Gini(<=91) = $1 - (8/12)^2 - (4/12)^2 = 0.44$

Gini(>91) = $1 - (1/2)^2 - (1/2)^2 = 0.5$

Gini(91) = (12/14) x 0.44 + (2/14) x 0.5 **= 0.44**

Humidity	Yes	No	Total
<=95	8	4	13
>95	1	0	1

Gini(<=95) = $1 - (8/13)^2 - (4/13)^2 = 0.526$

Gini(>95) = $1 - (1/1)^2 - (0/1)^2 = 0$

Gini(95) = (13/14) x 0.526 + (1/14) x 0 **= 0.488**

65	Yes
70	No
70	Yes
70	Yes
75	Yes
80	Yes
80	Yes
85	No
86	Yes
90	No
90	Yes
91	No
95	No
96	Yes

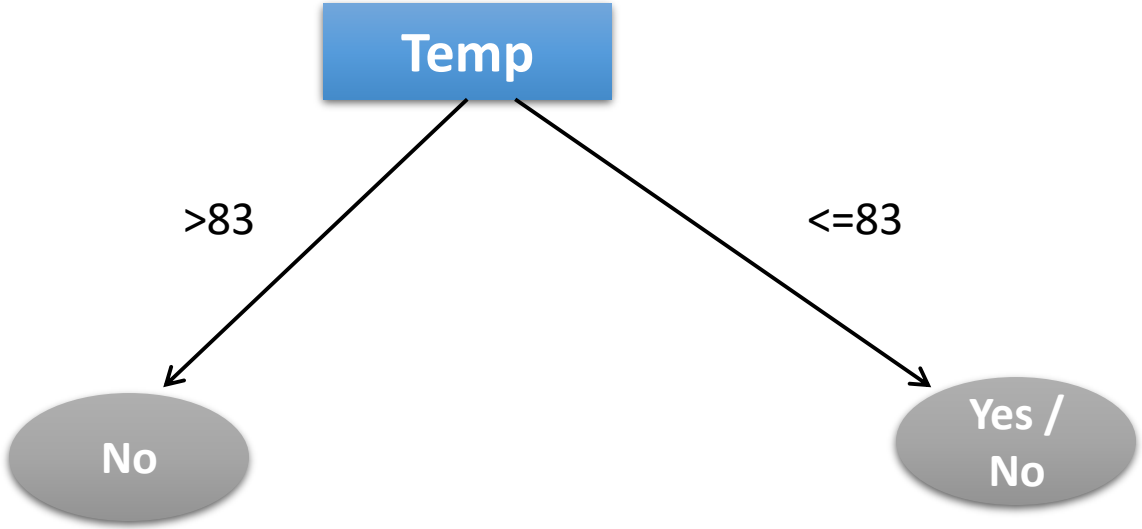
Humidity	Yes	No	Total
<=96	9	5	14
>96	1	0	1

Gini(<=96) = $1 - (9/14)^2 - (5/14)^2 = 0.459$

Gini(>96) = $1 - (0/14)^2 - (0/14)^2 = 1$

Gini(96) = (14/14) x 0.459 + (0/14) x 1 **= 0.459**

Outlook	**0.342**
Temperature	0.321
Humidity	0.366
Wind	0.428



D1	Sunny	85	85	Weak	No
D2	Sunny	90	90	Strong	No

D3	Overcast	83	86	Weak	Yes
D4	Rain	70	96	Weak	Yes
D5	Rain	68	80	Weak	Yes
D6	Rain	65	70	Strong	No
D7	Overcast	64	65	Strong	Yes
D8	Sunny	72	95	Weak	No
D9	Sunny	69	70	Weak	Yes
D10	Rain	75	80	Weak	Yes
D11	Sunny	75	70	Strong	Yes
D12	Overcast	72	90	Strong	Yes
D13	Overcast	81	75	Weak	Yes
D14	Rain	71	91	Strong	No

Construct a Decision using the following Training Dataset using CART –Regression Trees

Day	Outlook	Temp.	Humidity	Wind	Golf Players
1	Sunny	Hot	High	Weak	25
2	Sunny	Hot	High	Strong	30
3	Overcast	Hot	High	Weak	46
4	Rain	Mild	High	Weak	45
5	Rain	Cool	Normal	Weak	52
6	Rain	Cool	Normal	Strong	23
7	Overcast	Cool	Normal	Strong	43
8	Sunny	Mild	High	Weak	35
9	Sunny	Cool	Normal	Weak	38
10	Rain	Mild	Normal	Weak	46
11	Sunny	Mild	Normal	Strong	48
12	Overcast	Mild	High	Strong	52
13	Overcast	Hot	Normal	Weak	44
14	Rain	Mild	High	Strong	30

Standard deviation

Golf players = {25, 30, 46, 45, 52, 23, 43, 35, 38, 46, 48, 52, 44, 30}

Average of golf players = (25 + 30 + 46 + 45 + 52 + 23 + 43 + 35 + 38 + 46 + 48 + 52 + 44 + 30)/14 = 39.78

Standard deviation of golf players = SQRT[($(25 - 39.78)^2 + (30 - 39.78)^2 + (46 - 39.78)^2 + (45 - 39.78)^2 + (52 - 39.78)^2 + (23 - 39.78)^2 + (43 - 39.78)^2 + (35 - 39.78)^2 + (38 - 39.78)^2 + (46 - 39.78)^2 + (48 - 39.78)^2 + (52 - 39.78)^2 + (44 - 39.78)^2 + (30 - 39.78)^2$)/14] = **9.32**

Outlook

Sunny outlook

Day	Outlook	Temp.	Humidity	Wind	Golf Players
1	Sunny	Hot	High	Weak	25
2	Sunny	Hot	High	Strong	30
8	Sunny	Mild	High	Weak	35
9	Sunny	Cool	Normal	Weak	38
11	Sunny	Mild	Normal	Strong	48

Golf players for sunny outlook = {25, 30, 35, 38, 48}

Average of golf players for sunny outlook = (25+30+35+38+48)/5 = 35.2
Standard deviation of golf players for sunny outlook = $\sqrt{(((25 - 35.2)^2 + (30 - 35.2)^2 + \ldots)/5)} = 7.78$

Overcast outlook

Day	Outlook	Temp.	Humidity	Wind	Golf Players
3	Overcast	Hot	High	Weak	46
7	Overcast	Cool	Normal	Strong	43
12	Overcast	Mild	High	Strong	52
13	Overcast	Hot	Normal	Weak	44

Golf players for overcast outlook = {46, 43, 52, 44}

Average of golf players for overcast outlook = (46 + 43 + 52 + 44)/4 = 46.25
Standard deviation of golf players for overcast outlook = $\sqrt{((46-46.25)^2+(43-46.25)^2+\ldots)}$= **3.49**

Rainy outlook

Day	Outlook	Temp.	Humidity	Wind	Golf Players
4	Rain	Mild	High	Weak	45
5	Rain	Cool	Normal	Weak	52
6	Rain	Cool	Normal	Strong	23
10	Rain	Mild	Normal	Weak	46
14	Rain	Mild	High	Strong	30

- Golf players for overcast outlook = {45, 52, 23, 46, 30}
- Average of golf players for overcast outlook = (45+52+23+46+30)/5 = 39.2
- Standard deviation of golf players for rainy outlook = $\sqrt{(((45 - 39.2)^2+(52 - 39.2)^2+\ldots)/5)}$=**10.87**

Outlook	Stdev of Golf Players	Instances
Overcast	3.49	4
Rain	10.87	5
Sunny	7.78	5

Weighted standard deviation for outlook = (4/14)x3.49 + (5/14)x10.87 + (5/14)x7.78 = **7.66**

Standard deviation reduction for outlook = 9.32 – 7.66 = 1.66

Temperature

Hot temperature

Day	Outlook	Temp.	Humidity	Wind	Golf Players
1	Sunny	Hot	High	Weak	25
2	Sunny	Hot	High	Strong	30
3	Overcast	Hot	High	Weak	46
13	Overcast	Hot	Normal	Weak	44

Golf players for hot temperature = {25, 30, 46, 44}
Standard deviation of golf players for hot temperature = **8.95**

Cool temperature

Day	Outlook	Temp.	Humidity	Wind	Golf Players
5	Rain	Cool	Normal	Weak	52
6	Rain	Cool	Normal	Strong	23
7	Overcast	Cool	Normal	Strong	43
9	Sunny	Cool	Normal	Weak	38

Golf players for cool temperature = {52, 23, 43, 38}
Standard deviation of golf players for cool temperature = **10.51**

Mild temperature

Day	Outlook	Temp.	Humidity	Wind	Golf Players
4	Rain	Mild	High	Weak	45
8	Sunny	Mild	High	Weak	35
10	Rain	Mild	Normal	Weak	46
11	Sunny	Mild	Normal	Strong	48
12	Overcast	Mild	High	Strong	52
14	Rain	Mild	High	Strong	30

Golf players for mild temperature
= {45, 35, 46, 48, 52, 30}
Standard deviation of golf players for mild temperature
= **7.65**

Summarizing standard deviations for temperature feature

Temperature	Stdev of Golf Players	Instances
Hot	8.95	4
Cool	10.51	4
Mild	7.65	6

Weighted standard deviation for temperature
= (4/14)x8.95 + (4/14)x10.51 + (6/14)x7.65 = **8.84**
Standard deviation reduction for temperature
= 9.32 – 8.84 = 0.47

Humidity

High humidity

Day	Outlook	Temp.	Humidity	Wind	Golf Players
1	Sunny	Hot	High	Weak	25
2	Sunny	Hot	High	Strong	30
3	Overcast	Hot	High	Weak	46
4	Rain	Mild	High	Weak	45
8	Sunny	Mild	High	Weak	35
12	Overcast	Mild	High	Strong	52
14	Rain	Mild	High	Strong	30

Golf players for normal humidity
= {52, 23, 43, 38, 46, 48, 44}
Standard deviation for golf players for normal humidity
= **8.73**

Golf players for high humidity
= {25, 30, 46, 45, 35, 52, 30}
Standard deviation for golf players for high humidity
= **9.36**

Normal humidity

Day	Outlook	Temp.	Humidity	Wind	Golf Players
5	Rain	Cool	Normal	Weak	52
6	Rain	Cool	Normal	Strong	23
7	Overcast	Cool	Normal	Strong	43
9	Sunny	Cool	Normal	Weak	38
10	Rain	Mild	Normal	Weak	46
11	Sunny	Mild	Normal	Strong	48
13	Overcast	Hot	Normal	Weak	44

Summarizing standard deviations for humidity feature

Humidity	Stdev of Golf Player	Instances
High	9.36	7
Normal	8.73	7

Weighted standard deviation for **humidity**
= (7/14)x9.36 + (7/14)x8.73 = 9.04
Standard deviation reduction for humidity
= 9.32 – 9.04 = 0.27

Wind

Strong Wind

Day	Outlook	Temp.	Humidity	Wind	Golf Players
2	Sunny	Hot	High	Strong	30
6	Rain	Cool	Normal	Strong	23
7	Overcast	Cool	Normal	Strong	43
11	Sunny	Mild	Normal	Strong	48
12	Overcast	Mild	High	Strong	52
14	Rain	Mild	High	Strong	30

Golf players for strong wind
= {30, 23, 43, 48, 52, 30}
Standard deviation for golf players for strong wind
= **10.59**

Weak Wind

1	Sunny	Hot	High	Weak	25
3	Overcast	Hot	High	Weak	46
4	Rain	Mild	High	Weak	45
5	Rain	Cool	Normal	Weak	52
8	Sunny	Mild	High	Weak	35
9	Sunny	Cool	Normal	Weak	38
10	Rain	Mild	Normal	Weak	46
13	Overcast	Hot	Normal	Weak	44

Golf players for weakk wind
= {25, 46, 45, 52, 35, 38, 46, 44}
Standard deviation for golf players for weak wind
= **7.87**

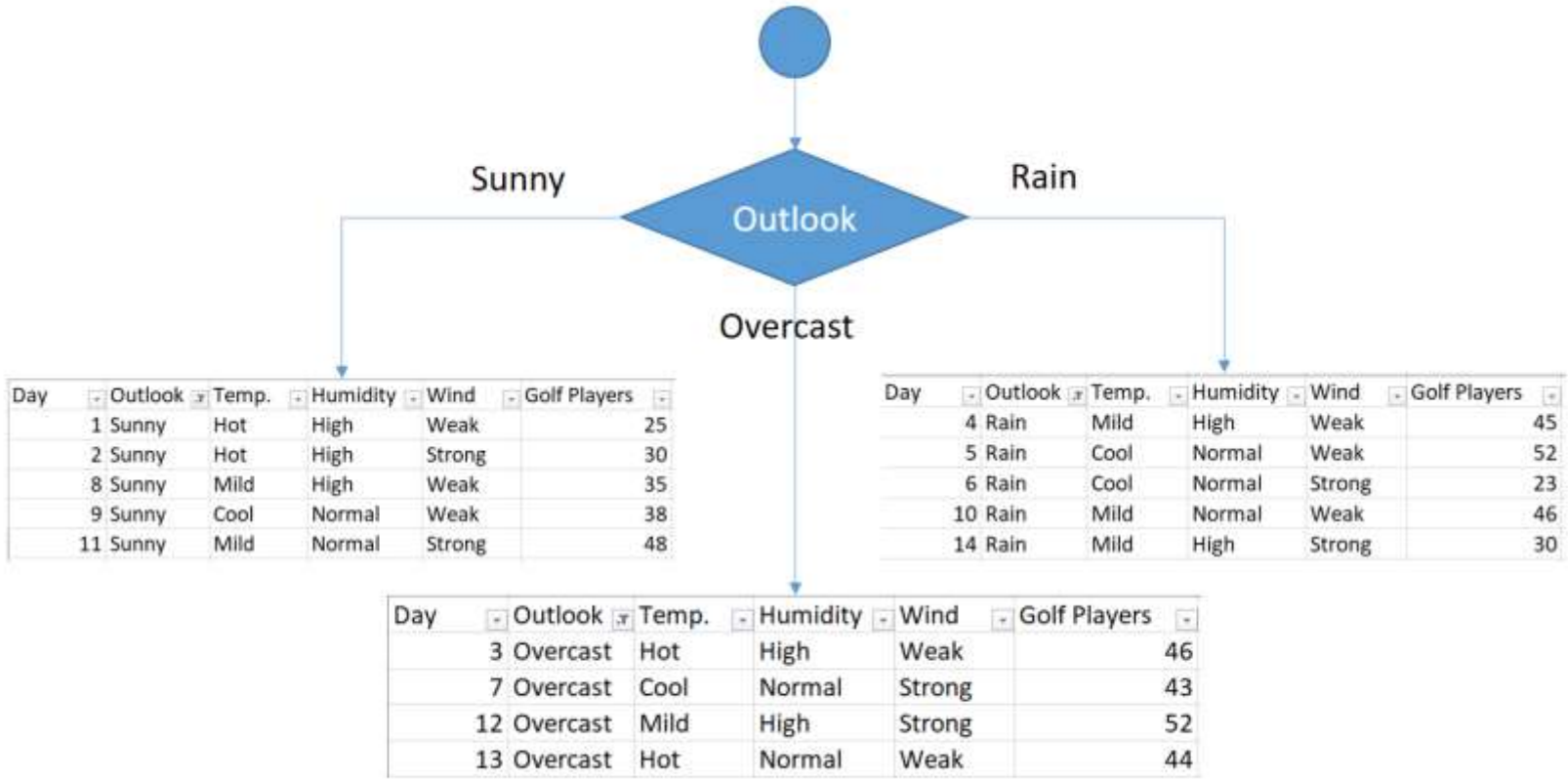
Summarizing standard deviations for wind feature

Wind	Stdev of Golf Player	Instances
Strong	10.59	6
Weak	7.87	8

Weighted standard deviation for wind = (6/14)x10.59 + (8/14)x7.87 = **9.03**

Standard deviation reduction for wind = 9.32 – 9.03 = 0.29

Feature	Standard Deviation Reduction
Outlook	1.66
Temperature	0.47
Humidity	0.27
Wind	0.29



Day	Outlook	Temp.	Humidity	Wind	Golf Players
1	Sunny	Hot	High	Weak	25
2	Sunny	Hot	High	Strong	30
8	Sunny	Mild	High	Weak	35
9	Sunny	Cool	Normal	Weak	38
11	Sunny	Mild	Normal	Strong	48

Day	Outlook	Temp.	Humidity	Wind	Golf Players
4	Rain	Mild	High	Weak	45
5	Rain	Cool	Normal	Weak	52
6	Rain	Cool	Normal	Strong	23
10	Rain	Mild	Normal	Weak	46
14	Rain	Mild	High	Strong	30

Day	Outlook	Temp.	Humidity	Wind	Golf Players
3	Overcast	Hot	High	Weak	46
7	Overcast	Cool	Normal	Strong	43
12	Overcast	Mild	High	Strong	52
13	Overcast	Hot	Normal	Weak	44

Sunny Outlook

Day	Outlook	Temp.	Humidity	Wind	Golf Players
1	Sunny	Hot	High	Weak	25
2	Sunny	Hot	High	Strong	30
8	Sunny	Mild	High	Weak	35
9	Sunny	Cool	Normal	Weak	38
11	Sunny	Mild	Normal	Strong	48

Golf players for sunny outlook
= {25, 30, 35, 38, 48}
Standard deviation for sunny outlook
= **7.78**

Sunny outlook and Hot Temperature

Day	Outlook	Temp.	Humidity	Wind	Golf Players
1	Sunny	Hot	High	Weak	25
2	Sunny	Hot	High	Strong	30

Standard deviation for sunny outlook and hot temperature = **2.5**

Sunny outlook and Cool Temperature

Day	Outlook	Temp.	Humidity	Wind	Golf Players
9	Sunny	Cool	Normal	Weak	38

Standard deviation for sunny outlook and cool temperature = **0**

Sunny outlook and Mild Temperature

Day	Outlook	Temp.	Humidity	Wind	Golf Players
8	Sunny	Mild	High	Weak	35
11	Sunny	Mild	Normal	Strong	48

Standard deviation for sunny outlook and mild temperature = **6.5**

Summary of standard deviations for temperature feature when outlook is sunny

Temperature	Stdev for Golf Players	Instances
Hot	2.5	2
Cool	0	1
Mild	6.5	2

Weighted standard deviation for sunny outlook and temperature = (2/5)x2.5 + (1/5)x0 + (2/5)x6.5 = **3.6**

Standard deviation reduction for sunny outlook and temperature = 7.78 – 3.6 = 4.18

Sunny outlook and high humidity

Day	Outlook	Temp.	Humidity	Wind	Golf Players
1	Sunny	Hot	High	Weak	25
2	Sunny	Hot	High	Strong	30
8	Sunny	Mild	High	Weak	35

Standard deviation for sunny outlook and high humidity = **4.08**

Sunny outlook and normal humidity

Day	Outlook	Temp.	Humidity	Wind	Golf Players
9	Sunny	Cool	Normal	Weak	38
11	Sunny	Mild	Normal	Strong	48

Standard **deviation for sunny outlook and normal humidity =** 5

Summarizing standard deviations for humidity feature when outlook is sunny

Humidity	Stdev for Golf Players	Instances
High	4.08	3
Normal	5.00	2

Weighted standard deviations for sunny outlook and humidity
= (3/5)x4.08 + (2/5)x5 = **4.45**
Standard deviation reduction for sunny outlook and humidity
= 7.78 – 4.45 = 3.33

Sunny outlook and Strong Wind

Day	Outlook	Temp.	Humidity	Wind	Golf Players
2	Sunny	Hot	High	Strong	30
11	Sunny	Mild	Normal	Strong	48

Standard deviation for sunny outlook and strong wind = **9**

Sunny outlook and Weak Wind

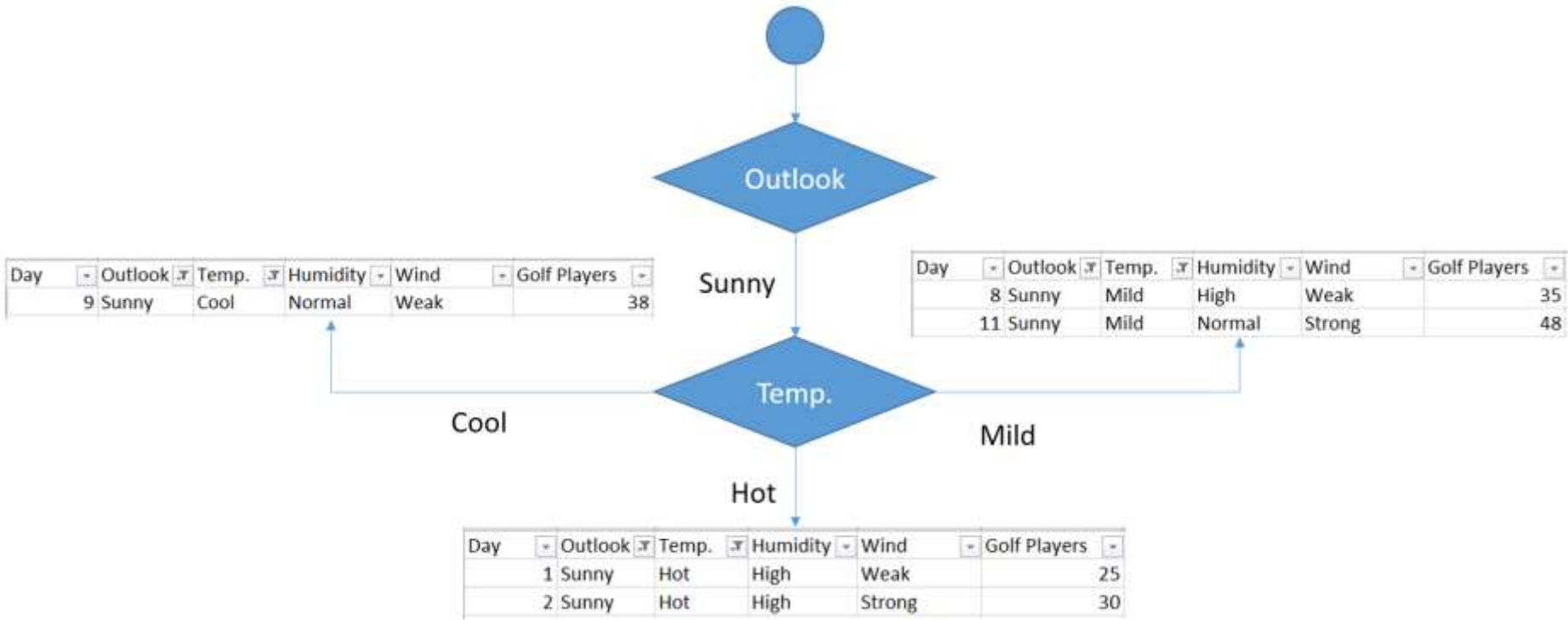
Day	Outlook	Temp.	Humidity	Wind	Golf Players
1	Sunny	Hot	High	Weak	25
8	Sunny	Mild	High	Weak	35
9	Sunny	Cool	Normal	Weak	38

Standard deviation for sunny outlook and weak wind = **5.56**

Wind	Stdev for Golf Players	Instances
Strong	9	2
Weak	5.56	3

Weighted standard deviations for sunny outlook and wind = (2/5)x9 + (3/5)x5.56 = **6.93**
Standard deviation reduction for sunny outlook and wind = 7.78 – 6.93 = 0.85

Feature	Standard Deviation Reduction
Temperature	4.18
Humidity	3.33
Wind	0.85



Pruning

- Should we add another branch for for each inividual value ? No, we should not. Because this causes over-fitting.
- We should terminate building branches
- for example if there are less than five instances in the sub data set. Or standard deviation of the sub data set can be less than 5% of the entire data set.
- Will terminate the branch if there are less than 5 instances in the current sub data set.
- If termination satisfied .calculate the average of the sub data set. This operation is called as pruning in decision tree trees.

Overcast outlook

Day	Outlook	Temp.	Humidity	Wind	Golf Players
3	Overcast	Hot	High	Weak	46
7	Overcast	Cool	Normal	Strong	43
12	Overcast	Mild	High	Strong	52
13	Overcast	Hot	Normal	Weak	44

If outlook is overcast, then there would be (46+43+52+44)/4 = **46.25** golf players.

Rainy Outlook

Day	Outlook	Temp.	Humidity	Wind	Golf Players
4	Rain	Mild	High	Weak	45
5	Rain	Cool	Normal	Weak	52
6	Rain	Cool	Normal	Strong	23
10	Rain	Mild	Normal	Weak	46
14	Rain	Mild	High	Strong	30

Standard deviation for rainy outlook = **10.87**

Rainy outlook and temperature

Temperature	Standard deviation for golf players	instances
Cool	**14.50**	2
Mild	**7.32**	3

Weighted standard deviation for rainy outlook and temperature = (2/5)x14.50 + (3/5)x7.32 = **10.19**
Standard deviation reduction for rainy outlook and temperature = 10.87 – 10.19 = 0.67

Rainy outlook and humidity

Humidity	Standard deviation for golf players	instances
High	**7.50**	2
Normal	**12.50**	3

Weighted standard deviation for rainy outlook and humidity = (2/5)x7.50 + (3/5)x12.50 = **10.50**
Standard deviation reduction for rainy outlook and humidity = 10.87 – 10.50 = 0.37

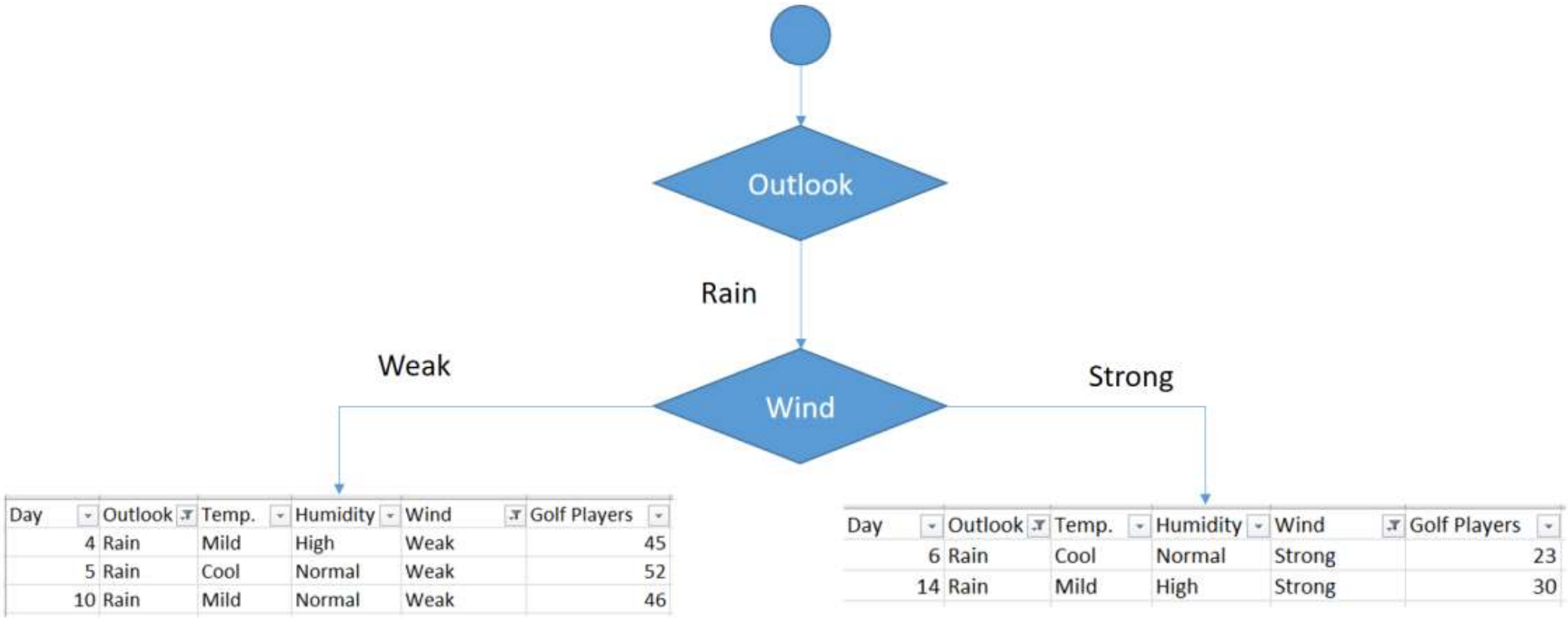
Rainy outlook and wind

Wind	Standard deviation for golf players	instances
Weak	**3.09**	3
Strong	**3.5**	2

Weighted standard deviation for rainy outlook and wind = (3/5)x3.09 + (2/5)x3.5 = **3.25**
Standard deviation reduction for rainy outlook and wind = 10.87 – 3.25 = 7.62

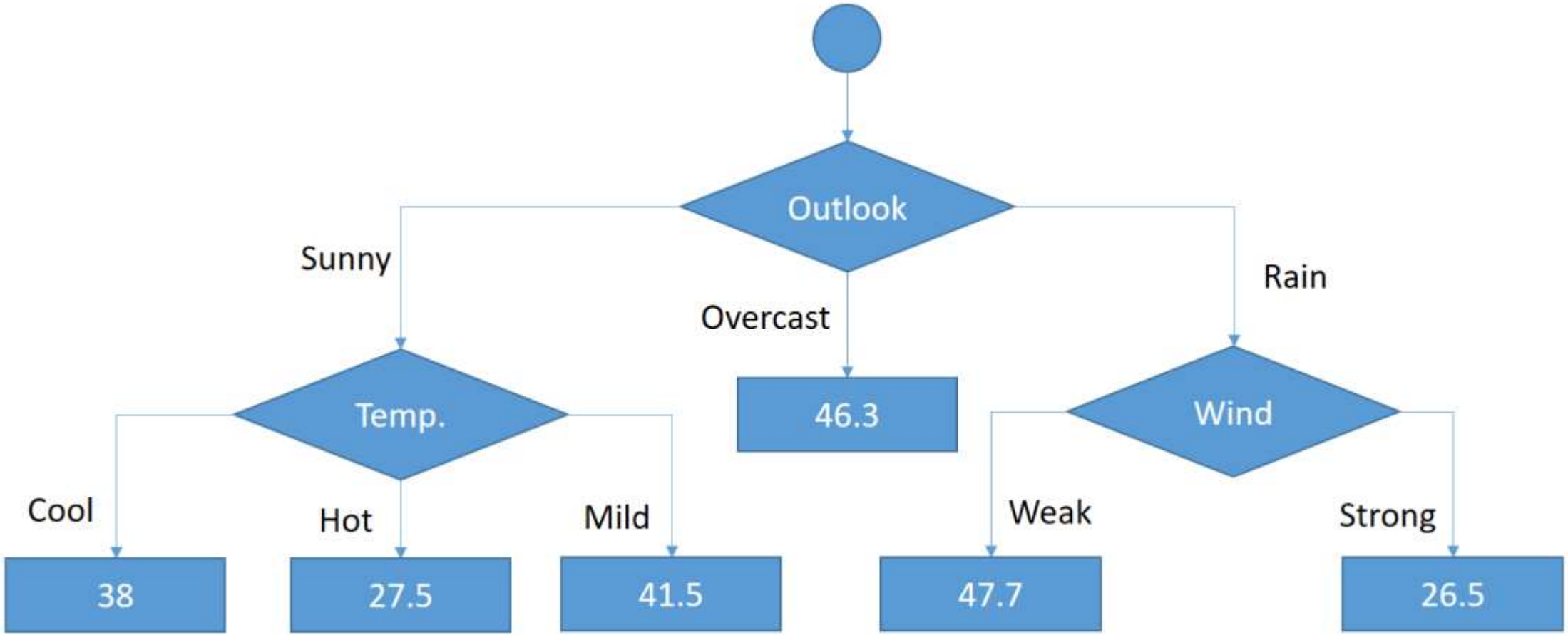
Summarizing rainy outlook

Feature	Standard deviation reduction
Temperature	0.67
Humidity	0.37
Wind	7.62



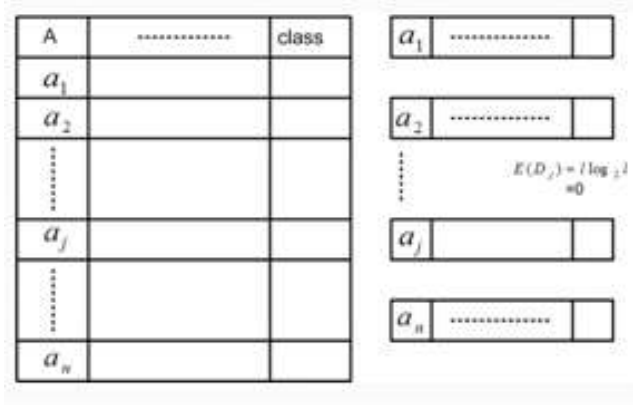
Day	Outlook	Temp.	Humidity	Wind	Golf Players
4	Rain	Mild	High	Weak	45
5	Rain	Cool	Normal	Weak	52
10	Rain	Mild	Normal	Weak	46

Day	Outlook	Temp.	Humidity	Wind	Golf Players
6	Rain	Cool	Normal	Strong	23
14	Rain	Mild	High	Strong	30


Outlook
Sunny
Overcast
Rain
Temp.
46.3
Wind
Cool
Hot
Mild
Weak
Strong
38
27.5
41.5
47.7
26.5

C4.5

- ID3 uses Information Gain measure which is in fact biased towards splitting attribute having large number of outcomes
- For eg. If attribute has distinct values for all tuples, then it would result in a large number of partitions, each one containing just one tuple. For such case E(D) = 0



$$E_A(D) = \sum_{j=1}^{n} \frac{|D_j|}{|D|} . E(D_j) = \sum_{j=1}^{n} \frac{1}{|D|} . 0 = 0$$

- ID3 favors splitting attribute having a large number of values
- Such partitions appears to be useless for classification
- This Situation is called as Overfitting, suffered by ID3

- In order to reduce the effect of this bias due to Information Gain , C 4.5 uses a different measure called Gain Ratio
- Gain Ratio is a kind of normalization to information gain using a split information

$$GainRatio(A) = Gain(A) / SplitInfo(A)$$

$$Entropy(S) = \sum - p(I) \cdot \log_2 p(I)$$
$$Gain(S, A) = Entropy(S) - \sum [\, p(S|A) \cdot Entropy(S|A) \,]$$

$$SplitInfo(A) = -\sum |Dj|/|D| \times \log_2 |Dj|/|D|$$

C4.5

1. Check **for** the above **base** cases.
2. For each attribute A, find the normalised information gain ratio **from** splitting on A.
3. Let A_best be the attribute **with** the highest normalized information gain.
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.

Advantages of C4.5 over other Decision Tree systems:

1. The algorithm inherently employs Single Pass Pruning Process to Mitigate overfitting.
2. It can work with both Discrete and Continuous Data
3. C4.5 can handle the issue of missing data / incomplete data very well

Day	Outlook	Temp.	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

$$\text{Entropy(Decision)} = - p(\text{Yes}) \cdot \log_2 p(\text{Yes}) - p(\text{No}) \cdot \log_2 p(\text{No})$$

$$\text{Entropy(Decision)} = - (9/14) \cdot \log_2 (9/14) - (5/14) \cdot \log_2 (5/14) = 0.940$$

[Gain Ratio of outlook]

Outlook	P(Yes)	N(No)	I.G
Sunny	2	3	0.971
Rainy	3	2	0.971
overcast	4	0	0

I.G(sunny) = -[(2)/(2+3)]log2[(2)/(2+3)]-[(3)/(2+3)]log2[(3)/(2+3)] =0.97
I.G(Rainy) = -[(3)/(2+3)]log2[(3)/(2+3)]-[(3)/(2+3)]log2[(3)/(2+3)] =0.97
Entropy(outlook) = [(2+3)/(9+5)] x (0.971)+[(3+2)/(9+5)]x(0.971) + [(4+0)/(9+5)] x (0) = 0.693
Gain(outlook) = 0.94 - 0.693 = 0.246

SplitInfo(Decision, Outlook) = $-(5/14).\log_2(5/14) -(4/14).\log_2(4/14) -(5/14).\log_2(5/14) = 1.577$

GainRatio(Decision, Outlook) = Gain(Decision, Outlook)/SplitInfo(Decision, Outlook)
= 0.246/1.577 = **0.155**

[Gain Ratio for windy]

wind	P(Yes)	N(No)	I.G.
Strong	3	3	1
weak	6	2	0.811

I.G(weak) = -[(6)/(6+2)]log2[(6)/(6+2)] -[(2)/(6+2)]log2[(2)/(6+2)] = 0.811
Entropy(wind) = [(3+3)/(9+5)]x(1)+[(6+2)/(9+5)]x(0.811) = 0.892
Gain(wind) = 0.94 – 0.892 = 0.049

SplitInfo(Decision, Wind) = $-(8/14).\log_2(8/14) – (6/14).\log_2(6/14) = 0.461 + 0.524 = 0.985$

GainRatio(Decision, Wind) = Gain(Decision, Wind) / SplitInfo(Decision, Wind)
= 0.049 / 0.985 = **0.049**

Day	Humidity	Decision
7	65	Yes
6	70	No
9	70	Yes
11	70	Yes
13	75	Yes
3	78	Yes
5	80	Yes
10	80	Yes
14	80	No
1	85	No
2	90	No
12	90	Yes
8	95	No
4	96	Yes

Entropy(Decision|Humidity<=65) = – p(No) . $\log_2$p(No) – p(Yes) . $\log_2$p(Yes) = -(0/1).$\log_2$(0/1) – (1/1).$\log_2$(1/1) = 0
Entropy(Decision|Humidity>65) = -(5/13).$\log_2$(5/13) – (8/13).$\log_2$(8/13) =0.530 + 0.431 = 0.961
Gain(Decision, Humidity<> 65) = 0.940 – (1/14).0 – (13/14).(0.961) = 0.048
SplitInfo(Decision, Humidity<> 65) = -(1/14).$\log_2$(1/14) -(13/14).$\log_2$(13/14) = 0.371
GainRatio(Decision, Humidity<> 65) = **0.126**

Entropy(Decision|Humidity<=70) = – (1/4).$\log_2$(1/4) – (3/4).$\log_2$(3/4) = 0.811
Entropy(Decision|Humidity>70) = – (4/10).$\log_2$(4/10) – (6/10).$\log_2$(6/10) = 0.970
Gain(Decision, Humidity<> 70) = 0.940 – (4/14).(0.811) – (10/14).(0.970) = 0.940 – 0.231 – 0.692 = 0.014
SplitInfo(Decision, Humidity<> 70) = -(4/14).$\log_2$(4/14) -(10/14).$\log_2$(10/14) = 0.863
GainRatio(Decision, Humidity<> 70) = **0.016**

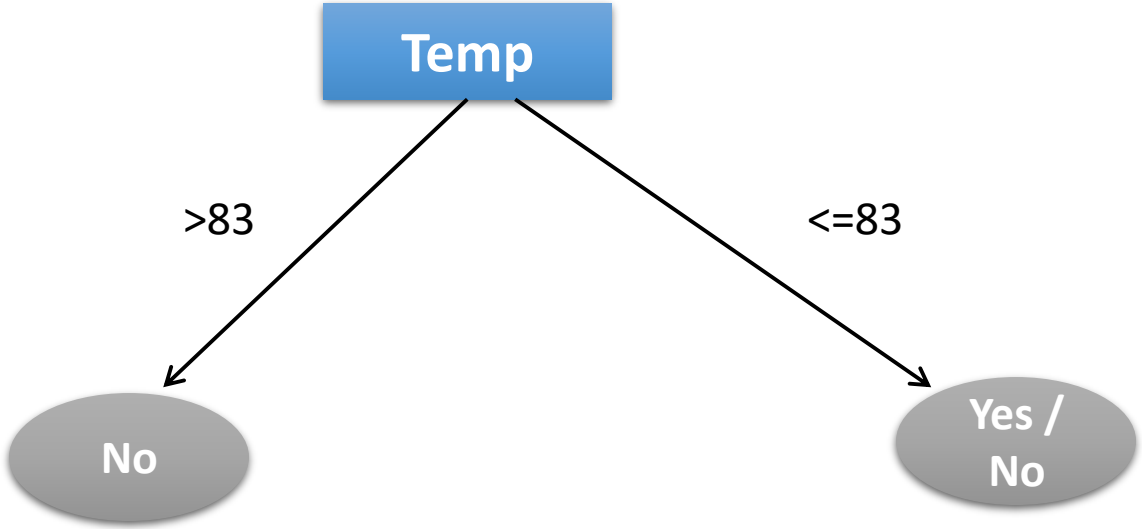
Entropy(Decision|Humidity<=75) = – (1/5).$\log_2$(1/5) – (4/5).$\log_2$(4/5) = 0.721
Entropy(Decision|Humidity>75) = – (4/9).$\log_2$(4/9) – (5/9).$\log_2$(5/9) = 0.991
Gain(Decision, Humidity<> 75) = 0.940 – (5/14).(0.721) – (9/14).(0.991) = 0.940 – 0.2575 – 0.637 = 0.045
SplitInfo(Decision, Humidity<> 75) = -(5/14).$\log_2$(4/14) -(9/14).$\log_2$(10/14) = 0.940
GainRatio(Decision, Humidity<> 75) = **0.047**

Day	Humidity	Decision
7	65	Yes
6	70	No
9	70	Yes
11	70	Yes
13	75	Yes
3	78	Yes
5	80	Yes
10	80	Yes
14	80	No
1	85	No
2	90	No
12	90	Yes
8	95	No
4	96	Yes

Gain(Decision, Humidity <> 78) =0.090, GainRatio(Decision, Humidity <> 78) =0.090
Gain(Decision, Humidity <> 80) = 0.101, GainRatio(Decision, Humidity <> 80) = 0.107
Gain(Decision, Humidity <> 85) = 0.024, GainRatio(Decision, Humidity <> 85) = 0.027
Gain(Decision, Humidity <> 90) = 0.010, GainRatio(Decision, Humidity <> 90) = 0.016
Gain(Decision, Humidity <> 95) = 0.048, GainRatio(Decision, Humidity <> 95) = 0.128

Gain(Decision, Temperature <> 83) = 0.113, GainRatio(Decision, Temperature<> 83) = 0.305

Attribute	Gain	GainRatio
Wind	0.049	0.049
Outlook	0.246	0.155
Humidity <> 80	0.101	0.107
Temperature <> 83	0.113	0.305



D1	Sunny	85	85	Weak	No
D2	Sunny	90	90	Strong	No

D3	Overcast	83	86	Weak	Yes
D4	Rain	70	96	Weak	Yes
D5	Rain	68	80	Weak	Yes
D6	Rain	65	70	Strong	No
D7	Overcast	64	65	Strong	Yes
D8	Sunny	72	95	Weak	No
D9	Sunny	69	70	Weak	Yes
D10	Rain	75	80	Weak	Yes
D11	Sunny	75	70	Strong	Yes
D12	Overcast	72	90	Strong	Yes
D13	Overcast	81	75	Weak	Yes
D14	Rain	71	91	Strong	No

Accuracy

		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a (TP)	b (FN)
	Class=No	c (FP)	d (TN)

$$\text{Accuracy} = \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+TN+FP+FN}$$

Limitation of Accuracy

- Consider a 2-class problem
  - Number of Class 0 examples = 9990
  - Number of Class 1 examples = 10

- If model predicts everything to be class 0, accuracy is 9990/10000 = 99.9 %
  - Accuracy is misleading because model does not detect any class 1 example

Sensitivity, Specificity and Precision

- The ability of the model to correctly classify positives and negatives are called sensitivity and specificity, respectively.
- The terminologies sensitivity and specificity originated in medical diagnostics.
- In generic case

Sensitivity = P(model classifies Y_i as positive | Y_i is positive)

Sensitivity is calculated using the following equation:

$$\text{Sensitivity} = \frac{\text{True Positive (TP)}}{\text{True Positive (TP)} + \text{False Negative (FN)}}$$

where True Positive (TP) is the number of positives correctly classified as positives by the model and False Negative (TN) is positives misclassified as negative by the model. Sensitivity is also called as **recall**.

Specificity

- ***Specificity*** is the ability of the diagnostic test to correctly classify the test as negative when the disease is not present. That is:

Specificity = *P*(diagnostic test is negative | patient has no disease)

- In general:

Sensitivity = *P*(model classifies Y_i as negative | Y_i is negative)

Specificity can be calculated using the following equation:

$$\text{Specificity} = \frac{\text{True Negative (TN)}}{\text{True Negative (TN)} + \text{False Positive (FP)}}$$

where True Negative (TN) is number of the negatives correctly classified as negatives by the model and False Positive (FP) is number of negatives misclassified as positives by the model.

- The decision maker has to consider the tradeoff between sensitivity and specificity to arrive at an optimal cut-off probability.
- Precision measures the accuracy of positives classified by the model.

Precision = *P*(patient has disease | diagnostic test is positive)

$$\text{Precision} = \frac{\text{True Positive (TP)}}{\text{True Positive (TP)} + \text{False Positive (FP)}}$$

- **F Score (F Measure)** is another measure used in binary logistic regression that combines both precision and recall and is given by:

$$\text{F} - \text{Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Observed			Predicted		
			Credit Rating		Percentage Correct
			0 (Negative)	1 (positive)	
Step 14	Credit Rating	0 (Negative) 561	507	54	90.4
		1 (Positive) 239	124	115	48.1
	Overall Percentage				77.8

$$\text{Sensitivit y} = \left(\frac{TP}{TP + FN}\right) = \left(\frac{115}{115 + 124}\right) = 48.1$$

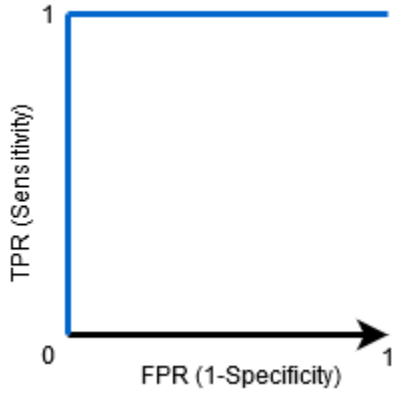
$$Specificity = \left(\frac{TN}{TN + FP}\right) = \left(\frac{507}{507 + 54}\right) = 90.4$$

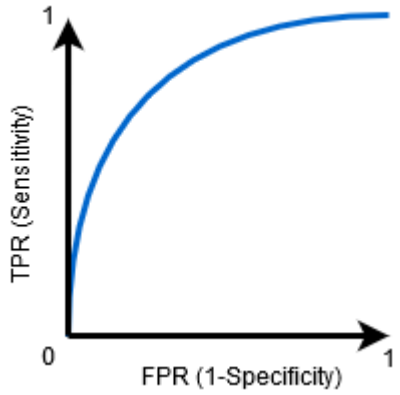
AUC-ROC Curve

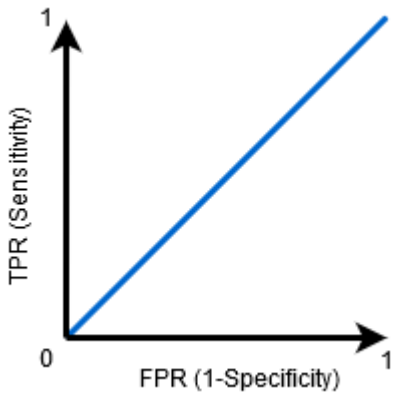
• **Receiver Operator Characteristic (ROC)** curve is an evaluation metric for binary classification problems

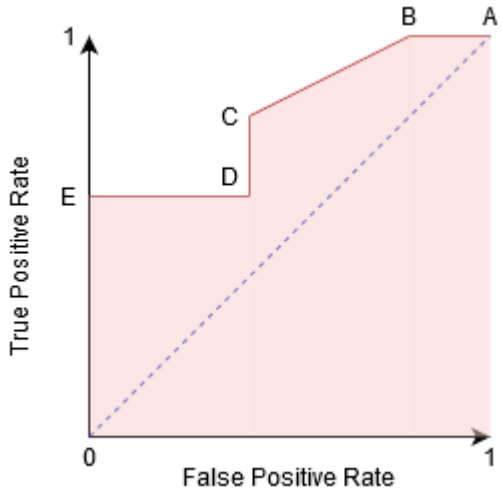
• probability curve that plots the **TPR** against **FPR**

• *The higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes.*









Find the class label for {Red, SUV, Bangalore} using the following dataset using Decision Tree(CART) classifier

Car No	Color	Type	City	Stolen
1	Red	Sports	Bangalore	Yes
2	Red	Sports	Bangalore	No
3	Red	Sports	Bangalore	Yes
4	Yellow	Sports	Bangalore	No
5	Yellow	Sports	Manipal	Yes
6	Yellow	SUV	Manipal	No
7	Yellow	SUV	Manipal	Yes
8	Yellow	SUV	Bangalore	No
9	Red	SUV	Manipal	No
10	Red	Sports	Manipal	No

Find the class label for {Old, Yes, Hardware} using the following dataset using ID3

Age	Competition	Type	Profit
Old	Yes	Software	Down
Old	No	Software	Down
Old	No	Hardware	Down
Mid	Yes	Software	Down
Mid	Yes	Hardware	Down
Mid	No	Hardware	Up
Mid	No	Software	Up
New	Yes	Software	Up
New	No	Hardware	Up
New	No	Software	Up