Machine Learning

Lecture 8-9: Classification and Decision Trees

COURSE CODE: CSE451

2023

Course Teacher

Dr. Mrinal Kanti Baowaly

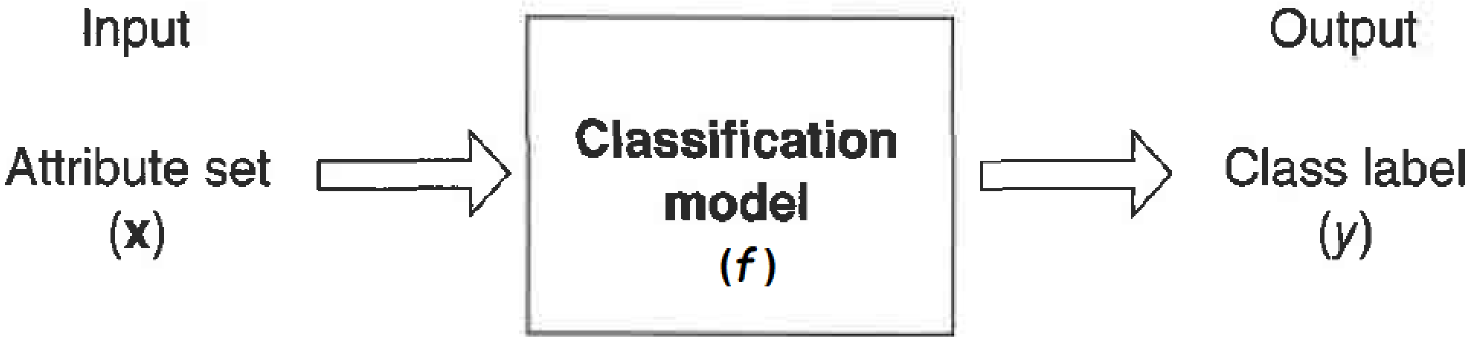
Associate Professor

Department of Computer Science and Engineering, Bangabandhu Sheikh Mujibur Rahman Science and Technology University, Bangladesh.

Email: mkbaowaly@gmail.com

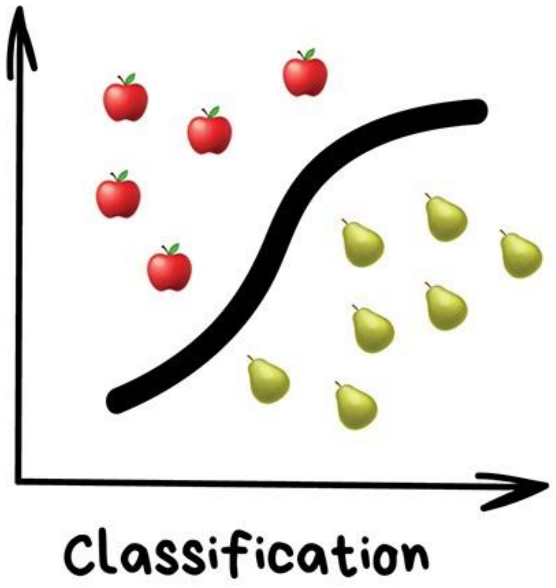
Classification: Definition

- Classification is the task of learning a target function *f that maps* each input *x* to one of the predefined class labels *y*.
  - *x: attribute, predictor, independent variable, input*
  - *y: class, category, response, dependent variable, target variable, output*
- The target function *f* is also known informally as a classification model.



Examples of Classification Task

- Email Spam filtering
- Language detection
- A search of similar documents
- Sentiment analysis
- Recognition of handwritten characters and numbers
- Fraud detection etc.



Types of Classification

- Binary Classification: Classifying instances into one of two class labels/categories
- Multiclass Classification: Classifying instances into one of three or more class labels/categories
- Multi-Label Classification: Multiple class labels or categories are to be predicted for each instance

Classification Techniques / Algorithms

Base Classifiers

- Decision Tree based Methods
- Rule-based Methods
- Nearest-neighbor
- Neural Networks
- Deep Learning
- Naïve Bayes and Bayesian Belief Networks
- Support Vector Machines

Ensemble Classifiers

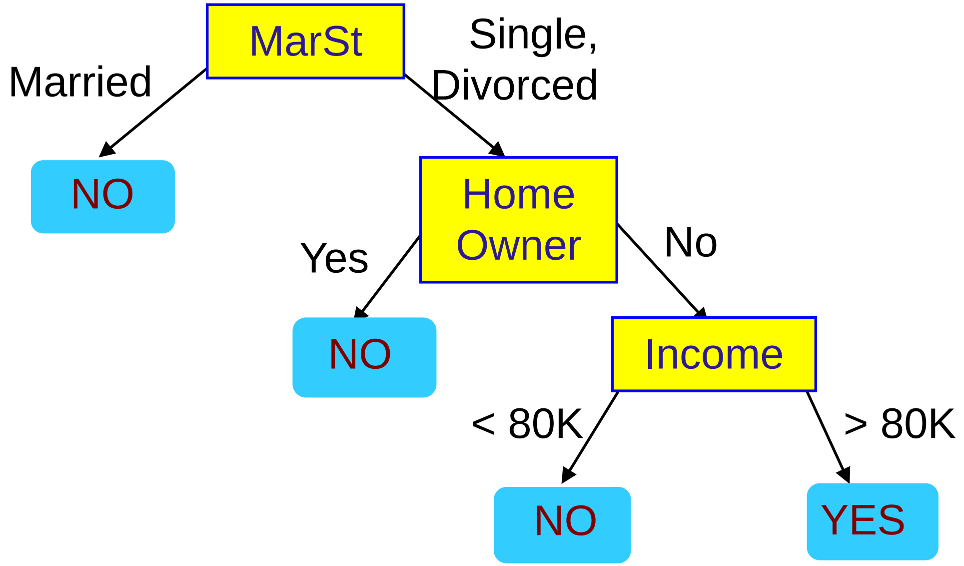
- Boosting, Bagging, Random Forests

Decision Tree Classification

- Decision tree is a type of supervised learning algorithm that is mostly used in classification problems.
- It works for both categorical and continuous input and output variables.
- This technique splits the population or data set into two or more homogeneous sets (or sub-populations) based on most significant splitter / differentiator in input variables.

An Example of Decision Tree

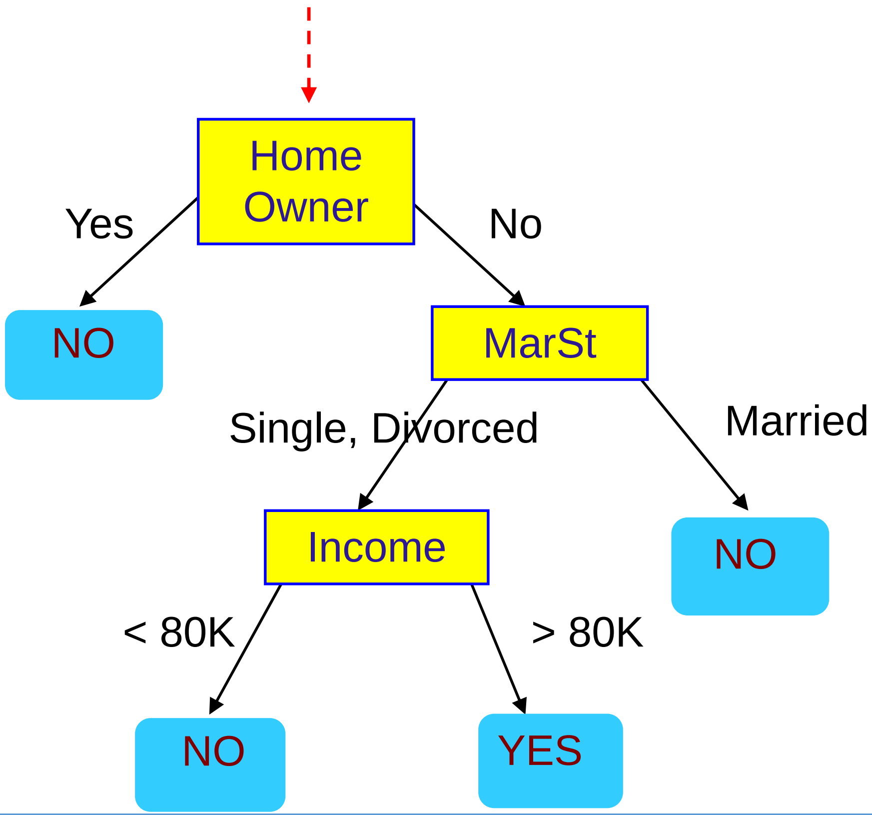
ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
	categorical	categorical	continuous	class
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single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



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

Apply Model to Test Data

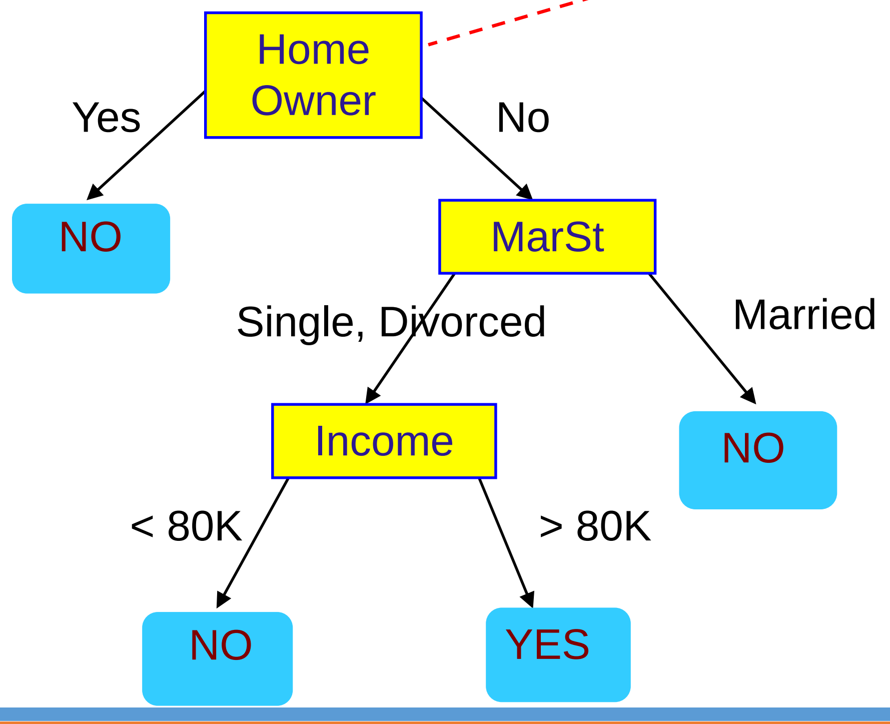
Start from the root of tree.



Test Data

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

Apply Model to Test Data



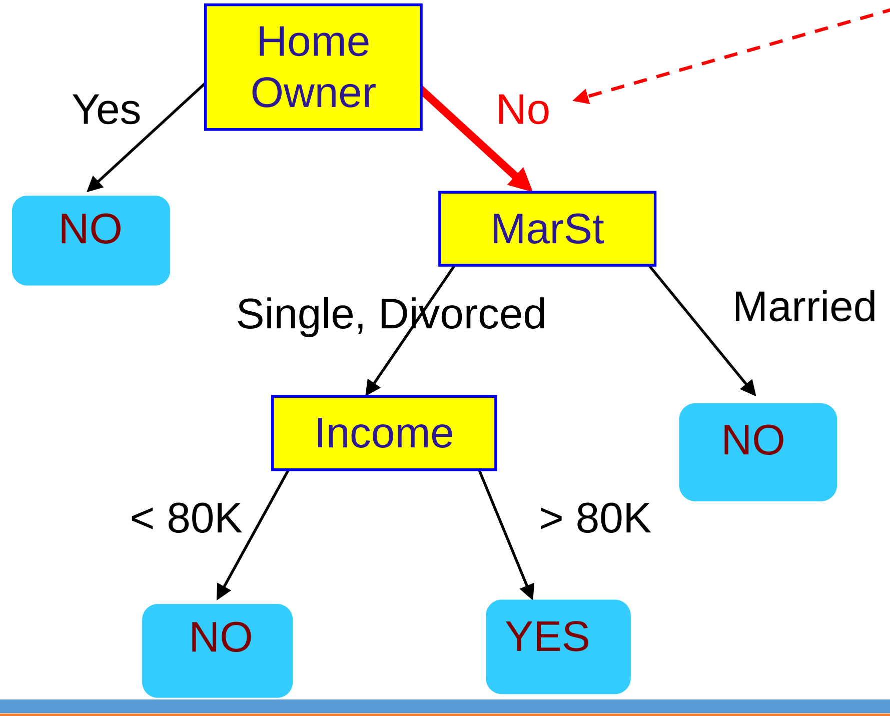
Test Data

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

Apply Model to Test Data

Test Data

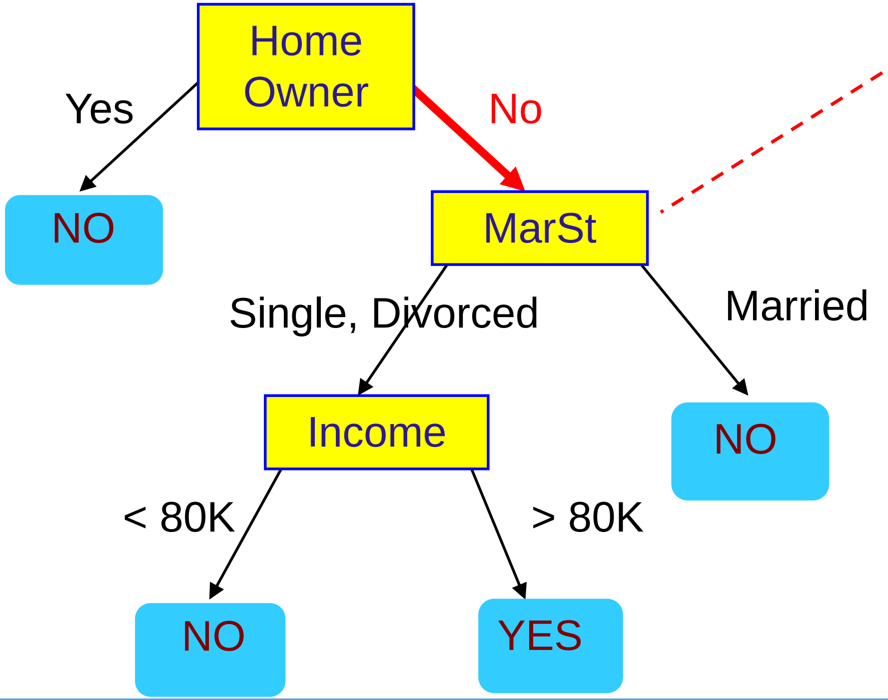
Home Owner	Marital Status	Annual Income	Defaulted Borrower
No	Married	80K	?



Apply Model to Test Data

Test Data

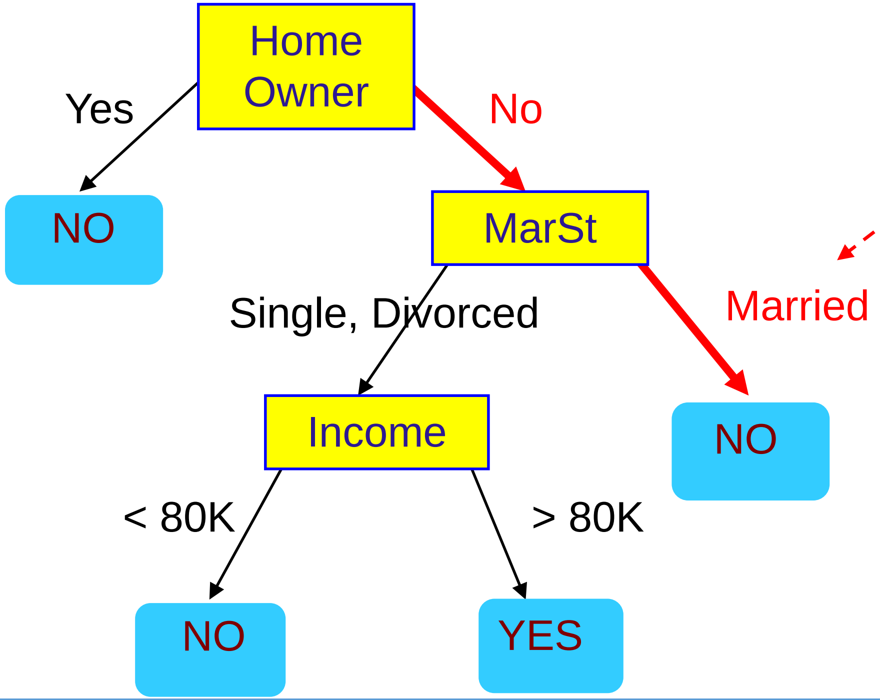
Home Owner	Marital Status	Annual Income	Defaulted Borrower
No	Married	80K	?



Apply Model to Test Data

Test Data

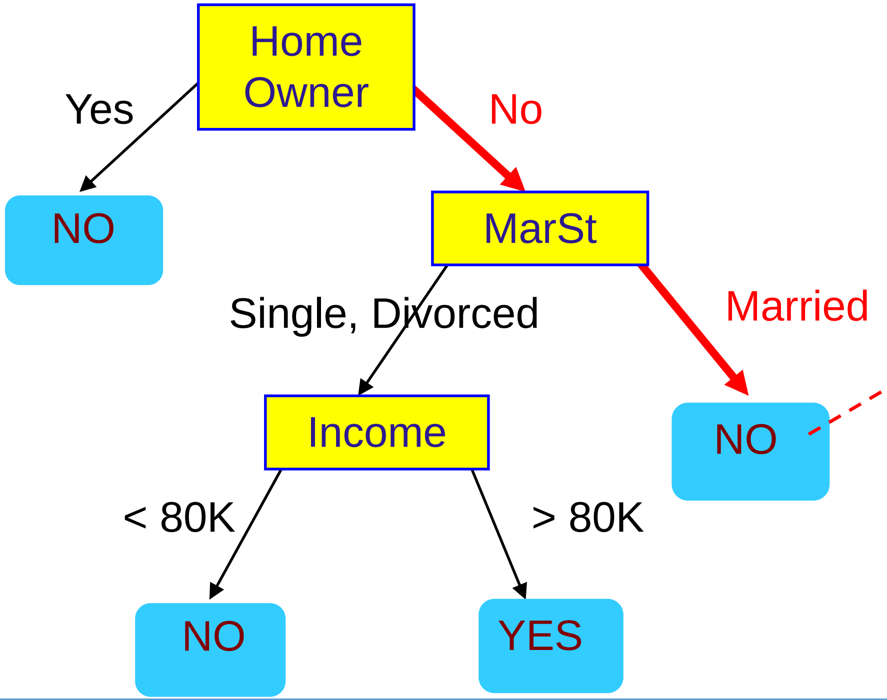
Home Owner	Marital Status	Annual Income	Defaulted Borrower
No	Married	80K	?



Apply Model to Test Data

Test Data

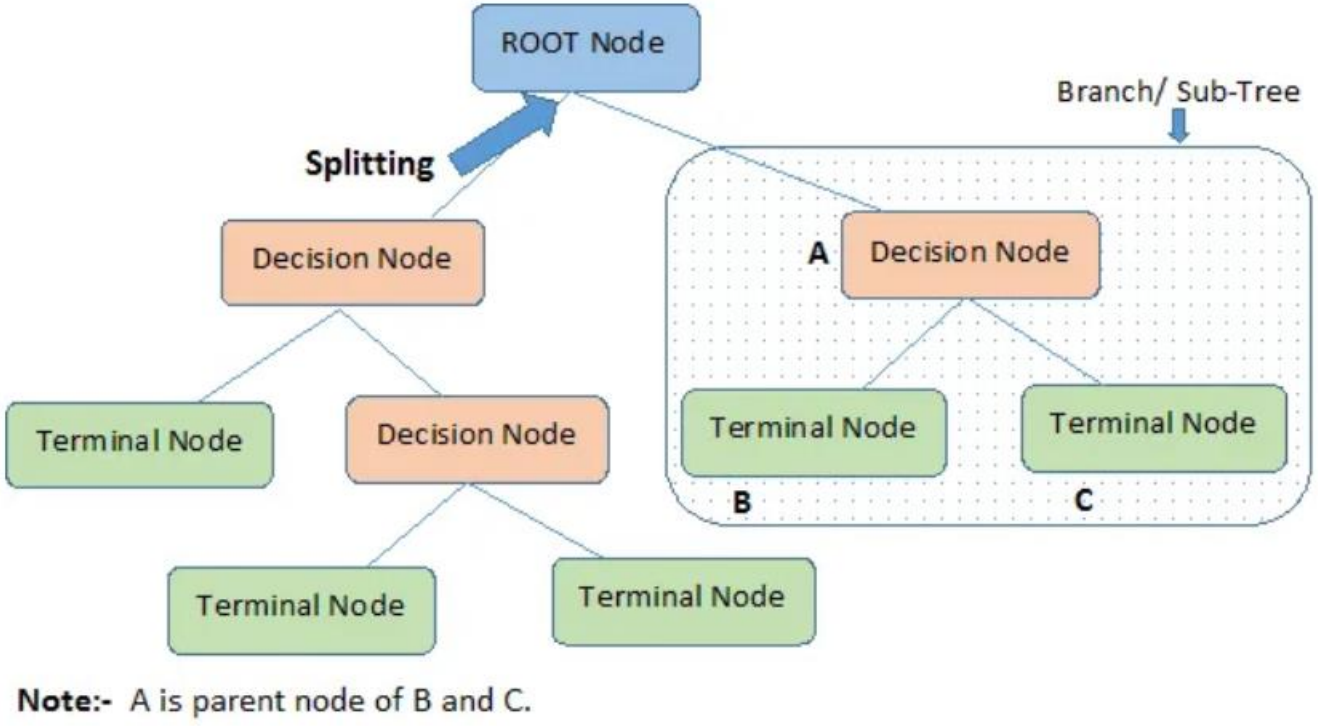
Home Owner	Marital Status	Annual Income	Defaulted Borrower
No	Married	80K	?



Assign Defaulted to "No"

What is a Decision Tree ?

A decision tree is a tree where each internal node represents a feature/attribute and acts as decision making, each link/branch represents a decision/rule and each leaf/terminal node represents an outcome(categorical or continues value). The topmost decision node in a decision tree is known as the root node.



Types of Decision Trees

Types of decision tree is based on the type of **target variable**:

1. Categorical/Classification decision Tree: Target variable is categorical
2. Continuous/Regression decision Tree: Target variable is continuous

How to construct a Decision Tree?

- A recursive fashion by partitioning the training records into successively purer subsets.
- The basic idea behind any decision tree algorithm is as follows:
  1. Select the best attribute using Attribute Selection Measures (e.g. information gain or Gini impurity) to split the records.
  2. Make that attribute a decision node and breaks the dataset into smaller subsets.
  3. Start tree building by repeating this process recursively for each subset until one of the condition will match:
     - All the records belong to the same class label (make it as a leaf node).
     - Maximum tree depth or minimum records per leaf is reached.
     - No more attributes to split on.
     - No more records.
  4. Assign class labels to each leaf node based on the majority class of the records in that node.
  5. Prune the tree to avoid overfitting.

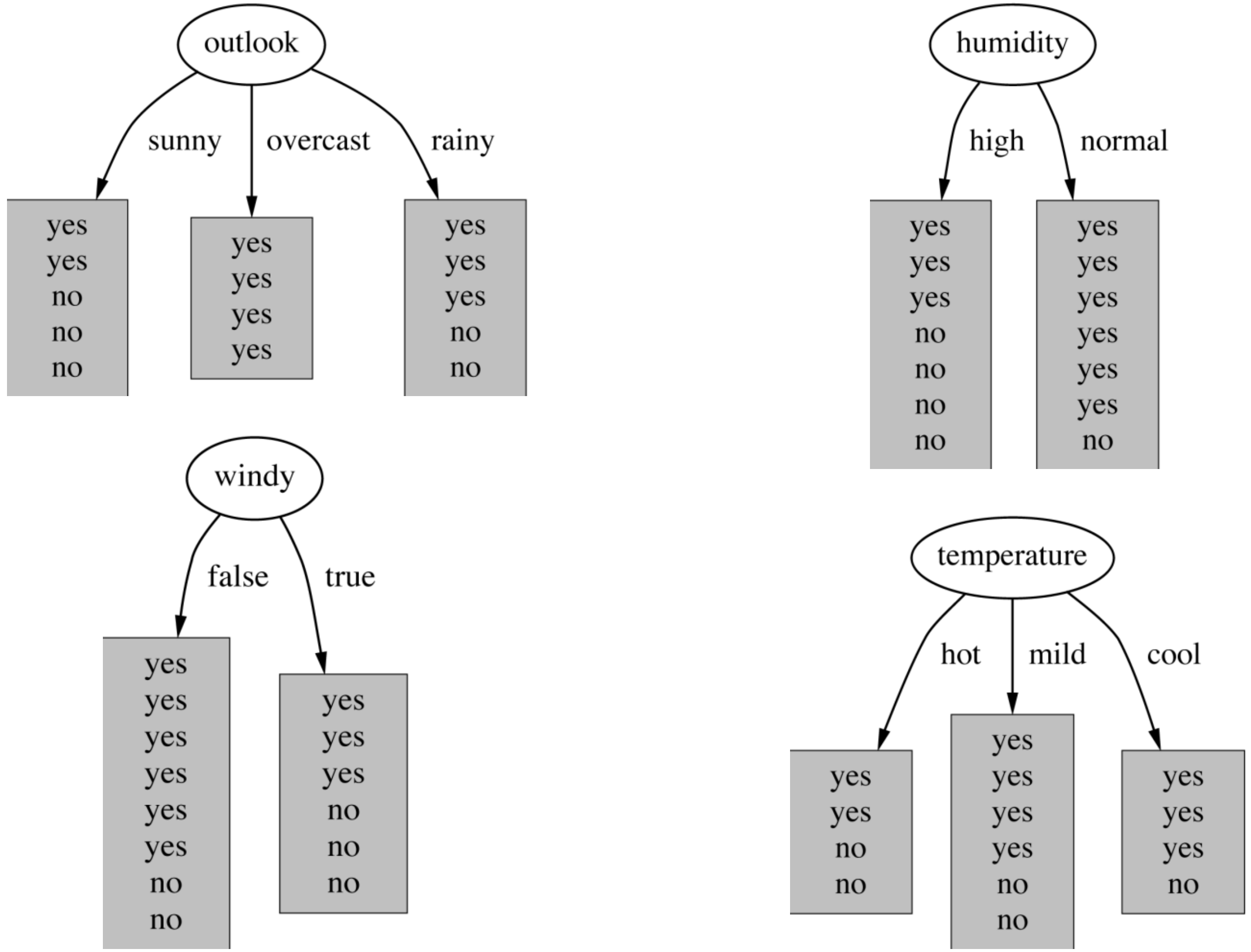
Decision Tree Algorithms

- Many Algorithms:
  - ID3 (Iterative Dichotomiser 3) - An early Decision Tree algorithm that uses information gain to determine the best split. It can only handle categorical variables and can lead to overfitting.
  - C4.5 - An improvement over ID3 that uses gain ratio to handle continuous variables and address overfitting issues.
  - CART (Classification and Regression Trees) - A Decision Tree algorithm for both classification and regression problems that uses Gini index for classification or mean squared error for regression to determine the best split.
  - C5.0 - A further improvement of C4.5 that uses boosting to improve accuracy and handle overfitting.
  - CHAID (Chi-squared Automatic Interaction Detection) - A Decision Tree algorithm that uses the chi-squared statistic to determine the best split for categorical variables.

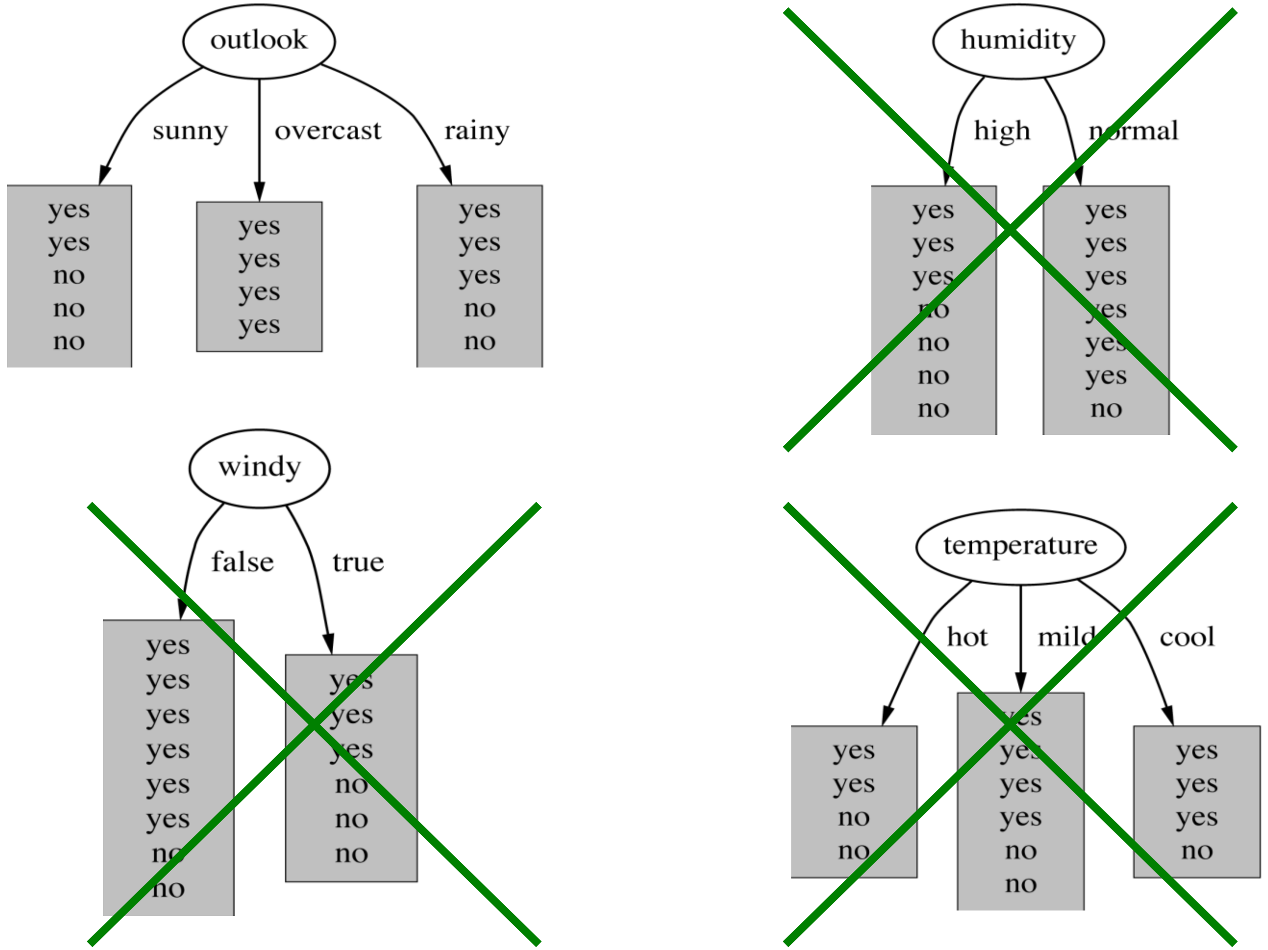
An Example of Attribute Selection. Tennis Weather: Can I play tennis today?

Outlook	Temperature	Humidity	Windy	Play
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
overcast	hot	high	FALSE	yes
rainy	mild	high	FALSE	yes
rainy	cool	normal	FALSE	yes
rainy	cool	normal	TRUE	no
overcast	cool	normal	TRUE	yes
sunny	mild	high	FALSE	no
sunny	cool	normal	FALSE	yes
rainy	mild	normal	FALSE	yes
sunny	mild	normal	TRUE	yes
overcast	mild	high	TRUE	yes
overcast	hot	normal	FALSE	yes
rainy	mild	high	TRUE	no

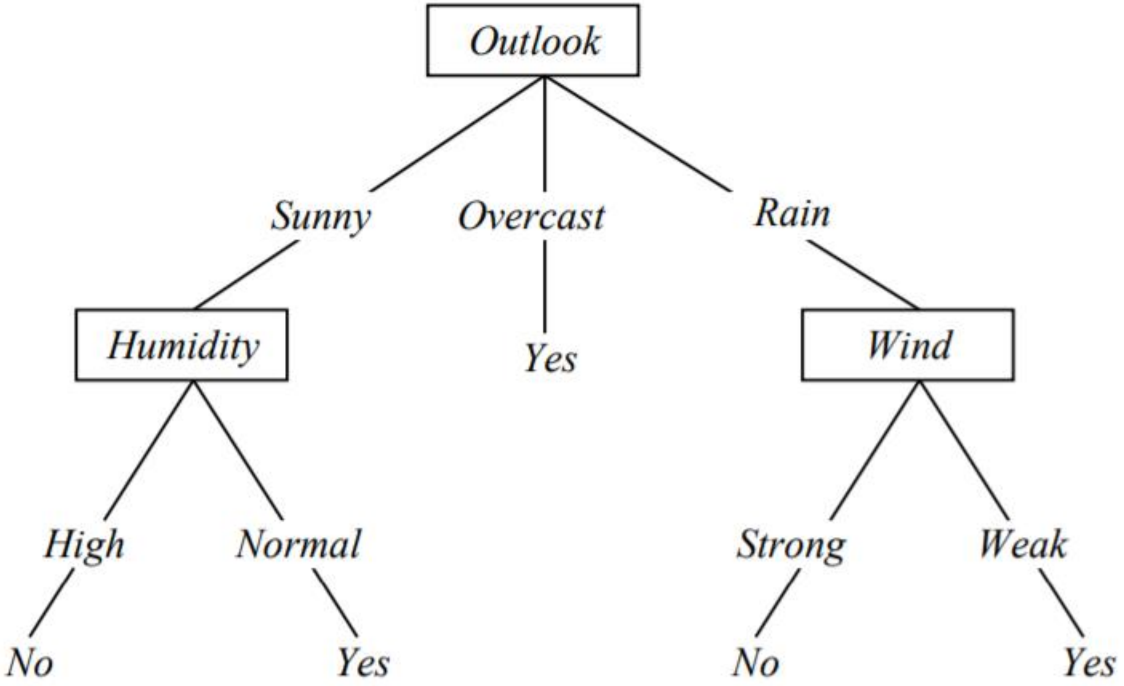
Which attribute to select?



Which attribute to select?



Decision Tree for Play Tennis



How to determine the Best Split

Attribute Selection Measures (ASM):

- Select an attribute to split the training records that increases the homogeneity of the resultant sub-nodes with respect to the target variable.
- In other words, we will split so that purity of the node increases with respect to the target variable.

Two popular methods:

- Gini index: used by CART
- Information gain: used by ID3 and C4.5

➢ Study detail from here.

Gini Index

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 works with categorical target variable “Success” or “Failure”.
- It performs only Binary splits.
- **Higher the value of Gini, higher the homogeneity**.
- CART (Classification and Regression Tree) uses Gini method to create binary splits.

Steps to Calculate Gini for a Split

1. Calculate Gini for sub nodes, using the following formula:

$$p^2 + q^2$$

   Here,

   p = probability for success

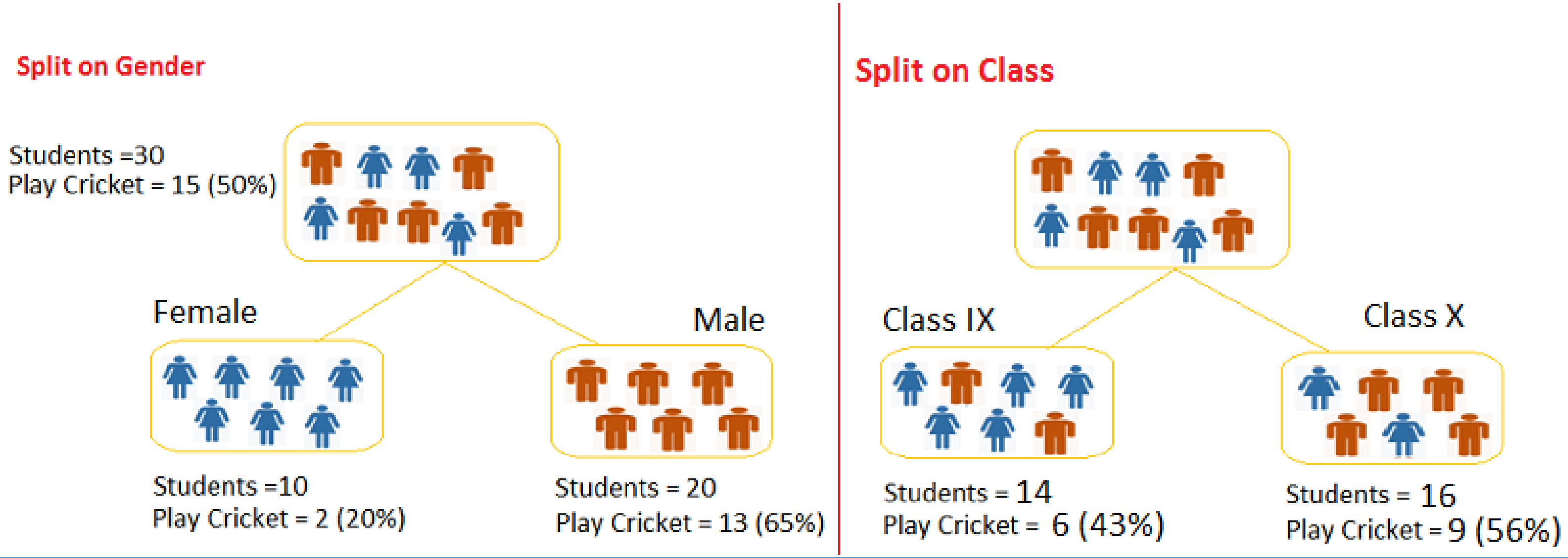
   q = probability for failure

2. Calculate Gini for split, using weighted Gini score of each node of that split.

Gini Index : Example

We have a sample of 30 students where:

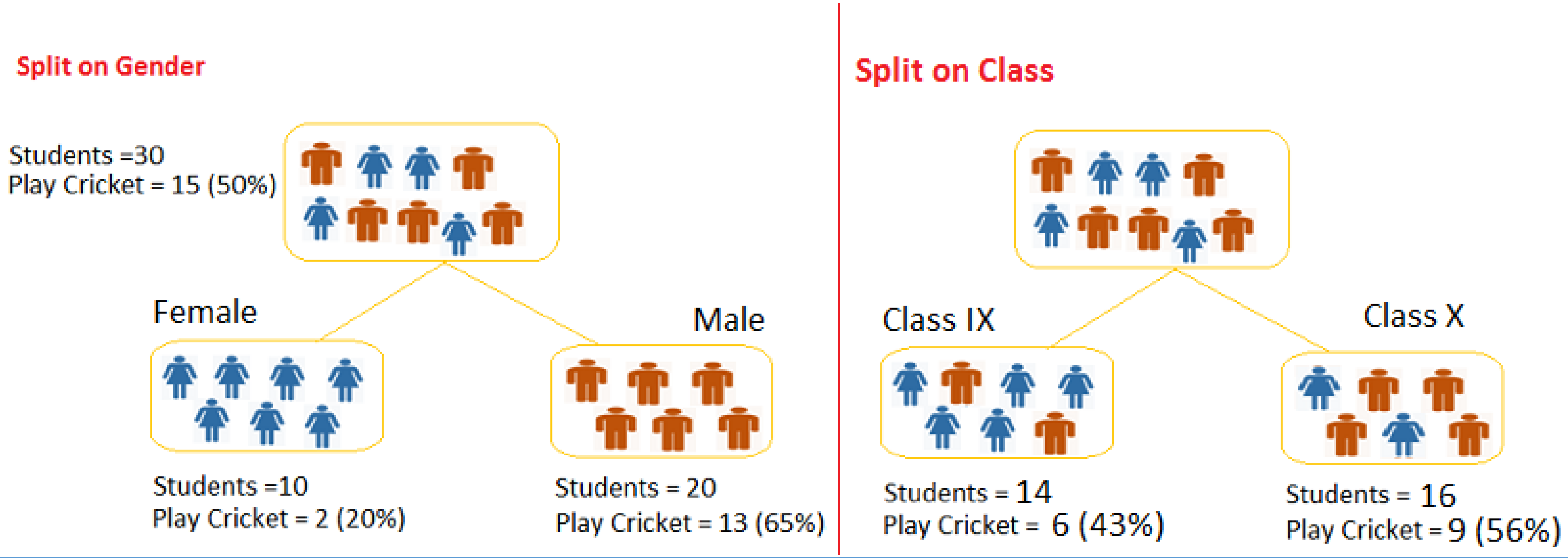
- Variable-1: Gender (Boy/ Girl), Variable-2: Class (IX/ X)
- Create a model to predict who will play cricket
- Identify which variable splits best



❖ Gender is able to identify best homogeneous sets

Gini Index: Split on Gender

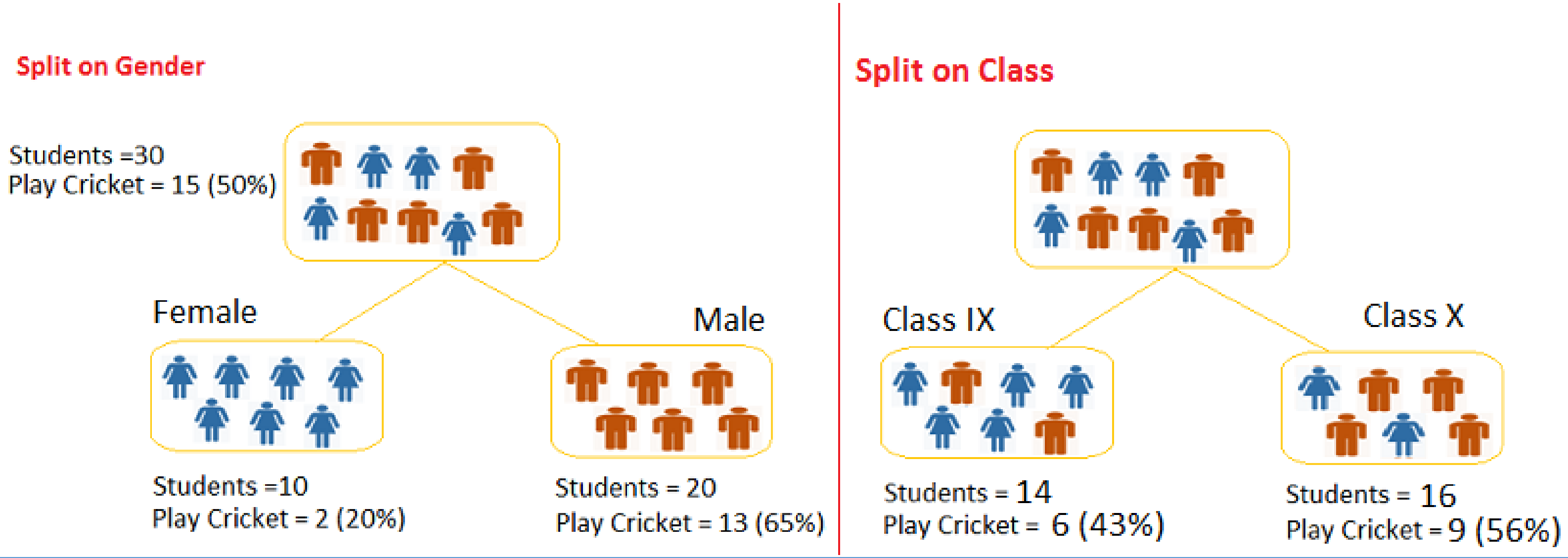
1. Gini for sub-node Female: $0.2^2 + 0.8^2 = 0.68$
2. Gini for sub-node Male: $0.65^2 + 0.35^2 = 0.55$
3. Weighted Gini for Split Gender: $\left(\frac{10}{30}\right)$*0.68 + $\left(\frac{20}{30}\right)$*0.55 = 0.59



Gini Index: Split on Class

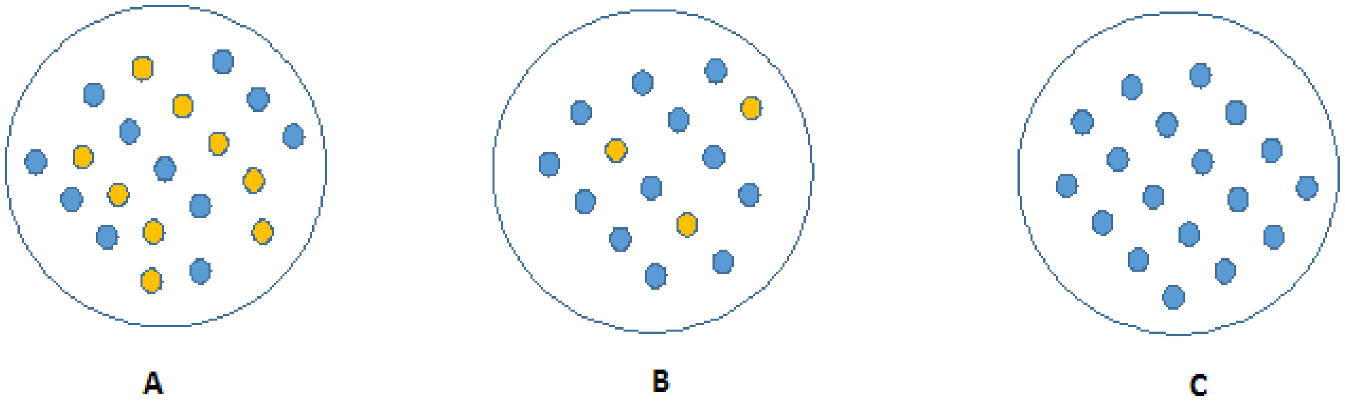
1. Gini for sub-node Class IX: $0.43^2 + 0.57^2 = 0.51$
2. Gini for sub-node Class X: $0.56^2 + 0.44^2 = 0.51$
3. Weighted Gini for Split Class: $\left(\frac{14}{30}\right)$*0.51 + $\left(\frac{16}{30}\right)$*0.51 = 0.51

❖ Gini score for *Split on* Gender is higher than *Split on Class*, hence, the node split will take place on Gender.



Information Gain

Look at the image below and think which node can be described easily?



- It is node C because it requires less information as all values are similar
- In other words, we can say that C is a pure/homogeneous node
- Degree of disorganization in a system known as Entropy.
- Entropy is zero when the sample is completely homogeneous, and it is 1 when the sample is an equally divided (50% – 50%)

Steps to Calculate Entropy for a Split

1. Formula to calculate entropy = $-p\ log_2p\ -qlog_2q$
2. Calculate entropy of parent node
3. Calculate entropy of each individual node of split
4. Calculate weighted average entropy of all sub nodes available in split.

*****The lesser the entropy, the better it is.**

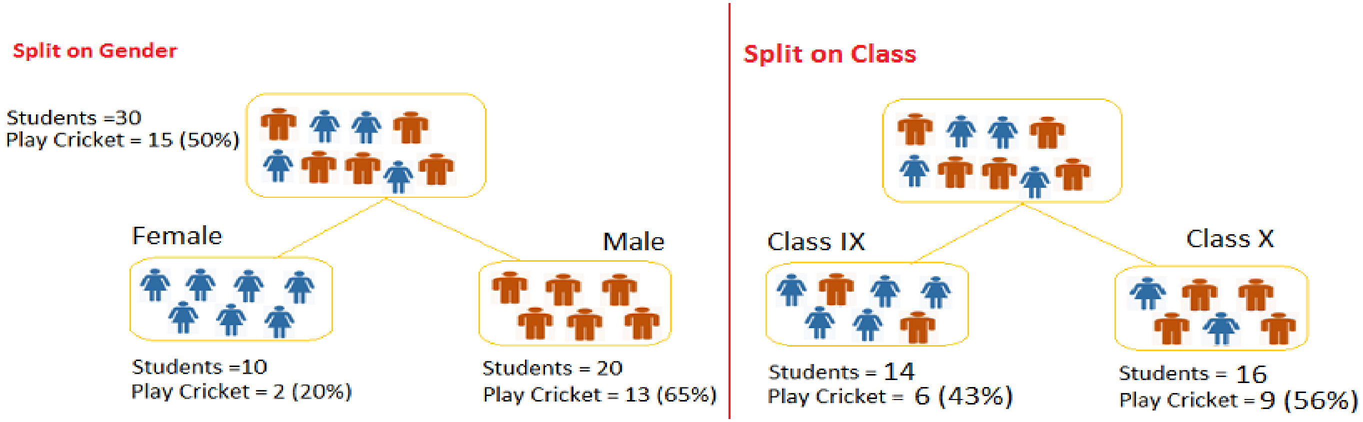
- We can derive information gain from entropy as 1- Entropy.

Information Gain: An Example

1. Entropy for Parent node: $-\left(\frac{15}{30}\right)\log_2\left(\frac{15}{30}\right) - \left(\frac{15}{30}\right)\log_2\left(\frac{15}{30}\right) = 1$

2. Entropy for split Gender:

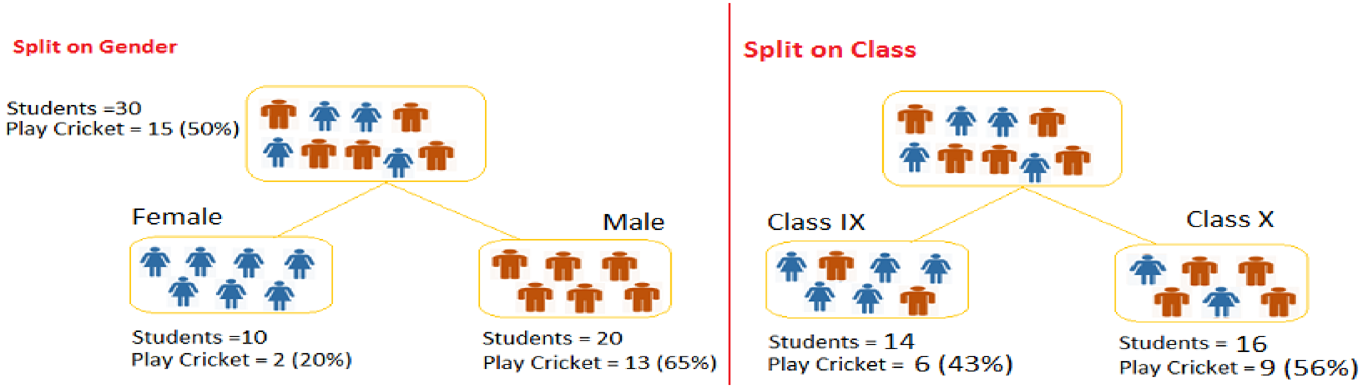
   - Entropy for Female node: $-\left(\frac{2}{10}\right)\log_2\left(\frac{2}{10}\right) - \left(\frac{8}{10}\right) log_2\left(\frac{8}{10}\right) = 0.72$
   - Entropy for Male node: $-\left(\frac{13}{20}\right)\log_2\left(\frac{13}{20}\right) - \left(\frac{7}{20}\right)\log_2\left(\frac{7}{20}\right) = 0.93$
   - Weighted entropy of Gender: $\left(\frac{10}{30}\right)*0.72 + \left(\frac{20}{30}\right)*0.93 = 0.86$



Information Gain: An Example

3. Entropy for split Class:
   - Entropy for Class IX node: $-\left(\frac{6}{14}\right)\log_2\left(\frac{6}{14}\right) - \left(\frac{8}{14}\right)\,log_2\left(\frac{8}{14}\right) = 0.99$
   - Entropy for Class X node: $-\left(\frac{9}{16}\right)\log_2\left(\frac{9}{16}\right) - \left(\frac{7}{16}\right)\log_2\left(\frac{7}{16}\right) = 0.99$
   - Weighted entropy of Class: $\left(\frac{14}{30}\right)$*0.99 + $\left(\frac{16}{30}\right)$*0.99 = 0.99

- ❖ Entropy for *Split on Gender* is the lowest, so the tree will split on Gender.
- ❖ We can derive information gain from entropy as **1-Entropy.**



Adv. & Disadv. of Decision Trees

Advantage

- Easy to Understand
- Less data cleaning required
- Can handle both numerical and categorical variables
- Useful in Data exploration

Disadvantage

- May contain lots of layers, which makes it complex
- May have an overfitting issue

Some Learning Materials

[AnalyticsVidhya: A Complete Tutorial on Tree Based Modeling from Scratch (in R & Python)]

[JavaTPoint: Decision Trees Algorithms]

[DataCamp: Decision Tree Classification in Python]

End of Lecture-8,9