Machine Learning

Lecture 9: Classification and Decision Trees

COURSE CODE: CSE451

2021

Course Teacher

Dr. Mrinal Kanti Baowaly

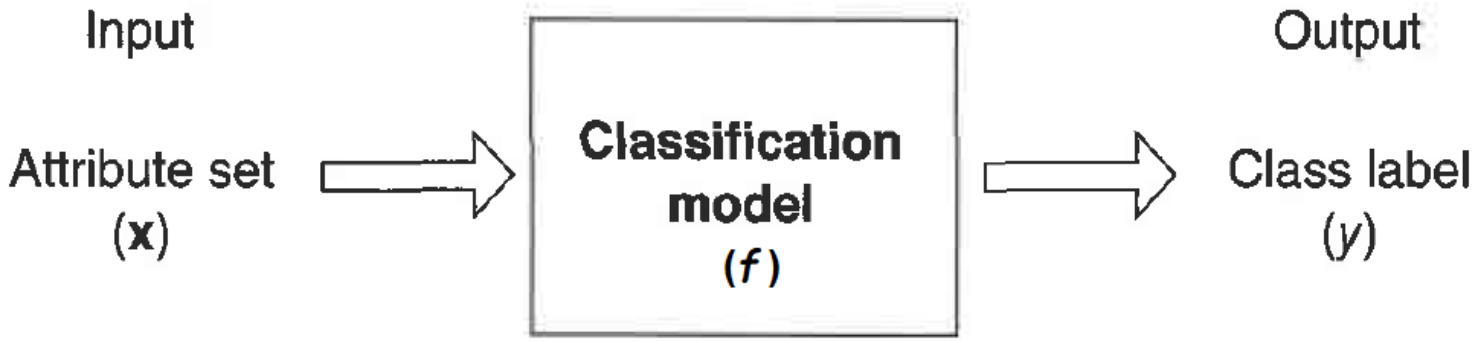
Associate Professor

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

Email: baowaly@gmail.com

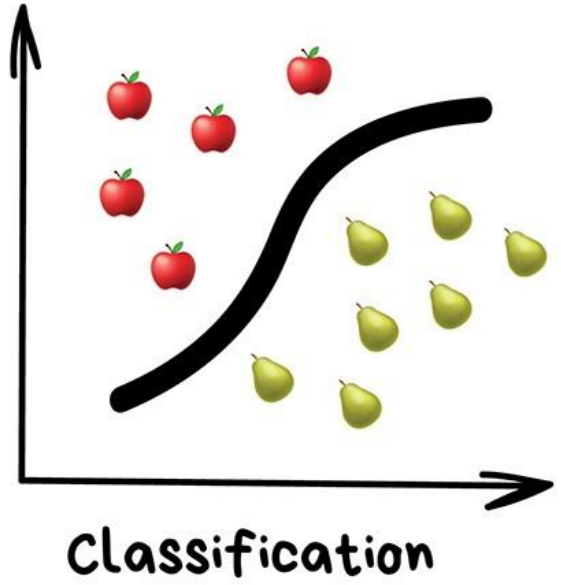
Classification: Definition

- 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, output*
- The target function *f* is also known informally as a classification model.



Examples of Classification Task

- 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

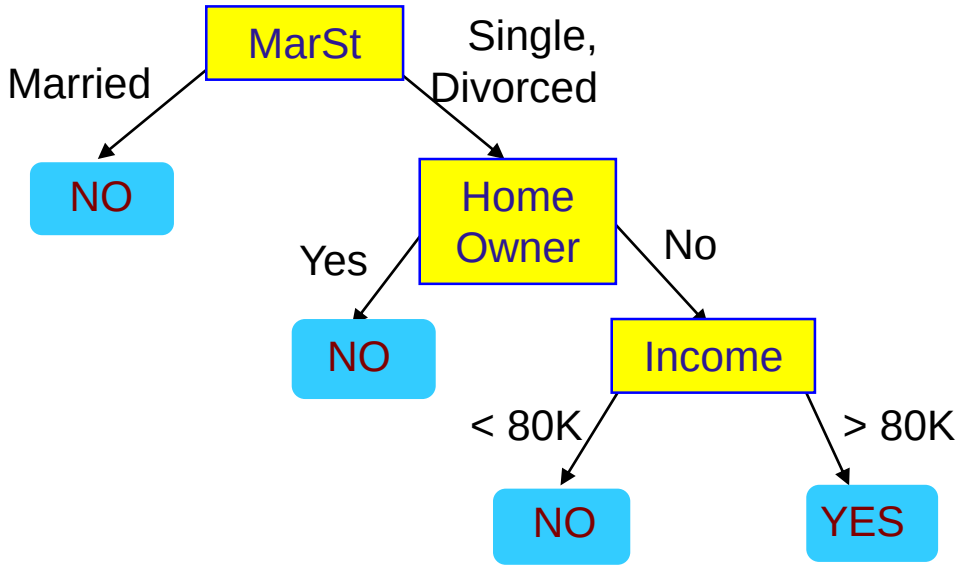
- Boosting, Bagging, Random Forests

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.

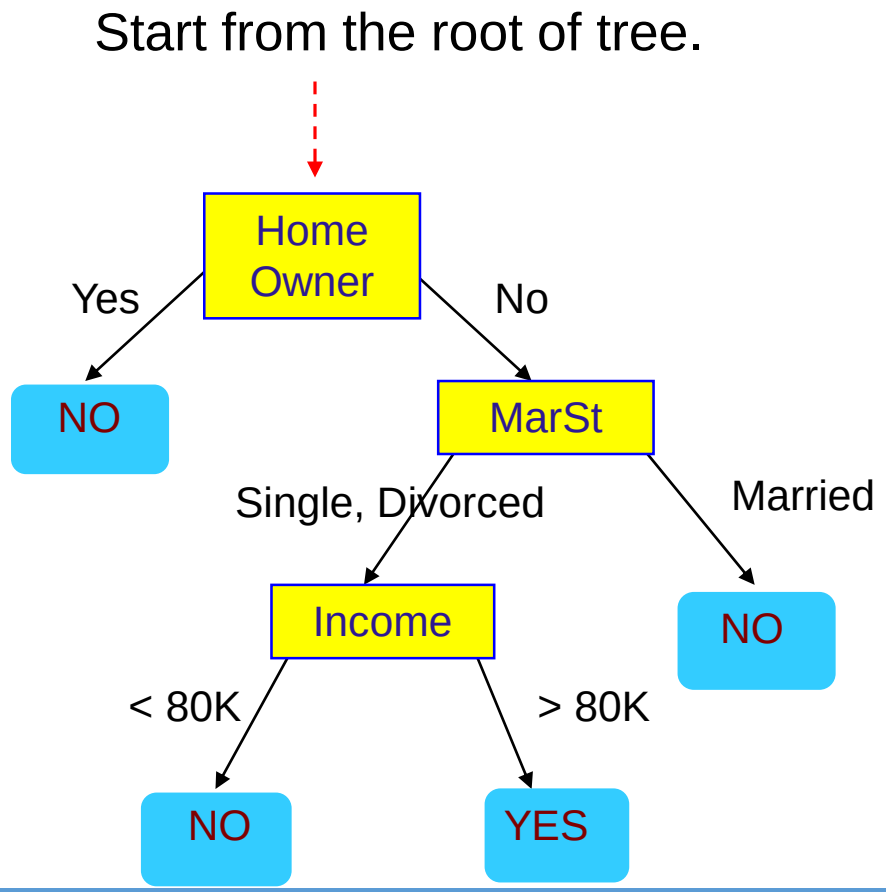
An Example of Decision Tree

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



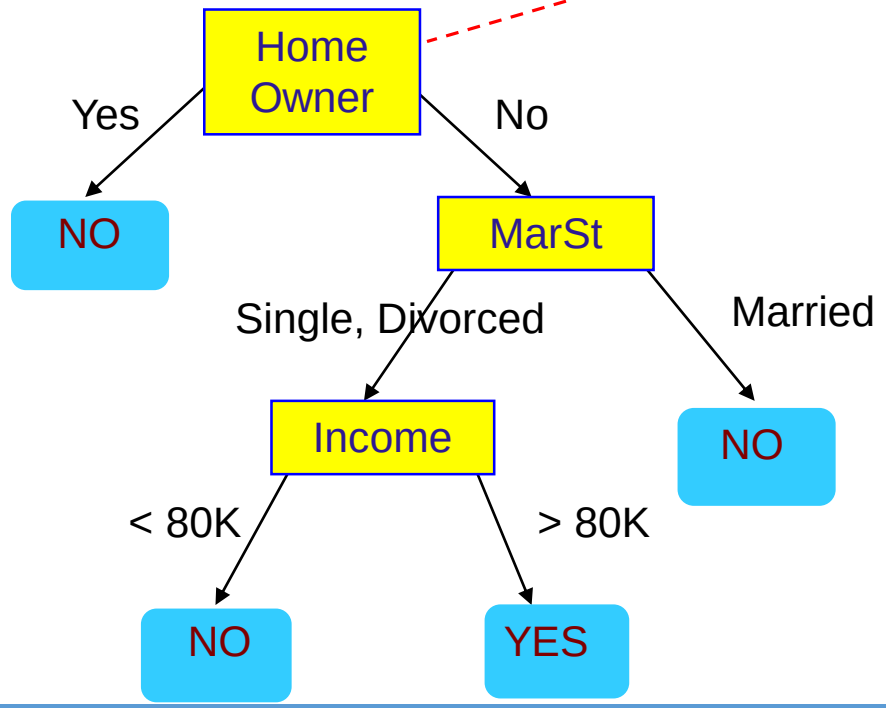
Test Data

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

Apply Model to Test Data

Test Data

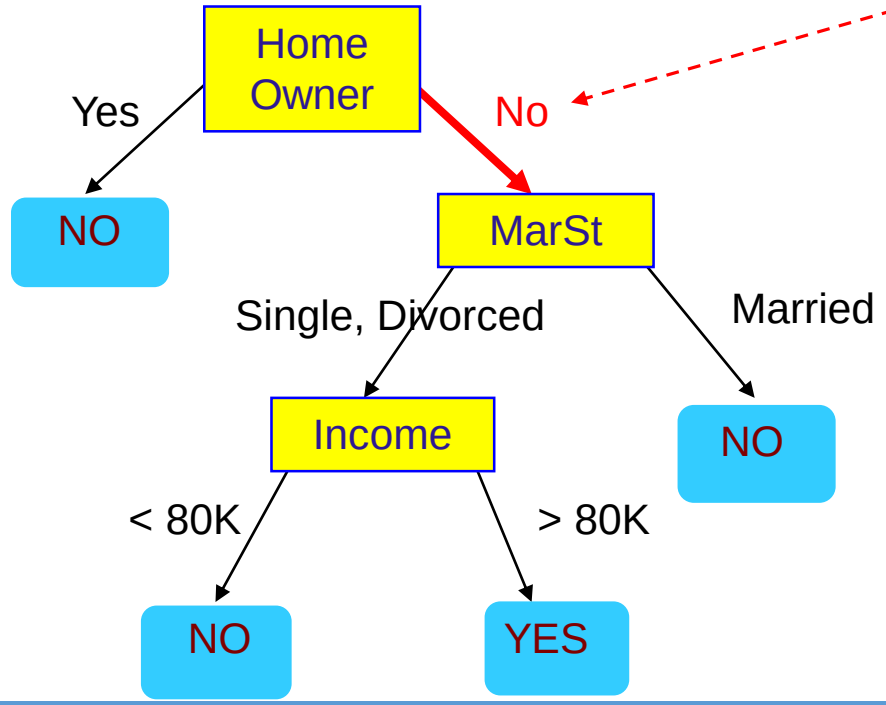
Home Owner	Marital Status	Annual Income	Defaulted Borrower
No	Married	80K	?



Apply Model to Test Data

Test Data

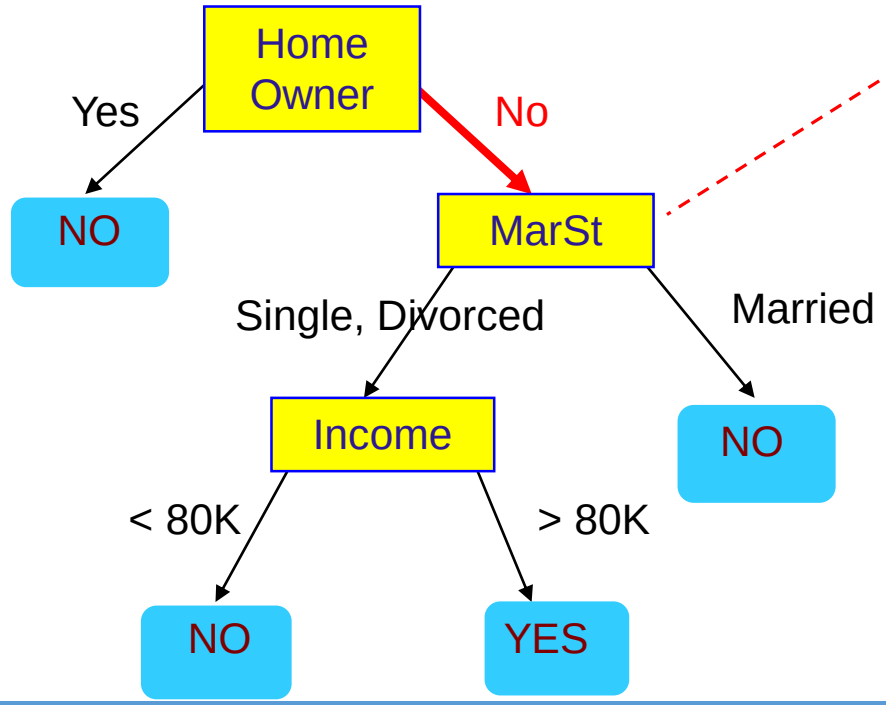
Home Owner	Marital Status	Annual Income	Defaulted Borrower
No	Married	80K	?



Apply Model to Test Data

Test Data

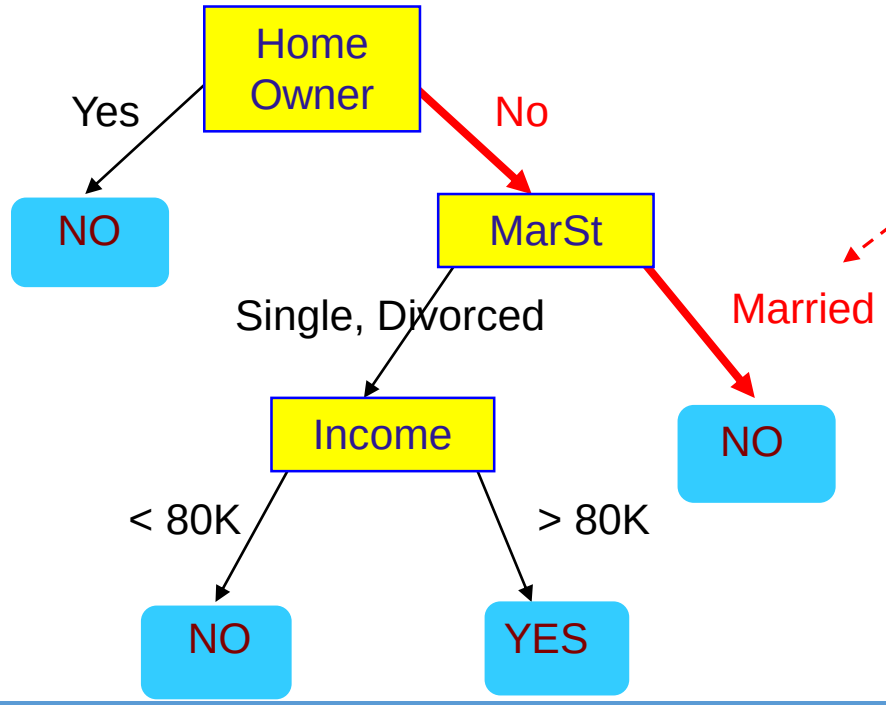
Home Owner	Marital Status	Annual Income	Defaulted Borrower
No	Married	80K	?



Apply Model to Test Data

Test Data

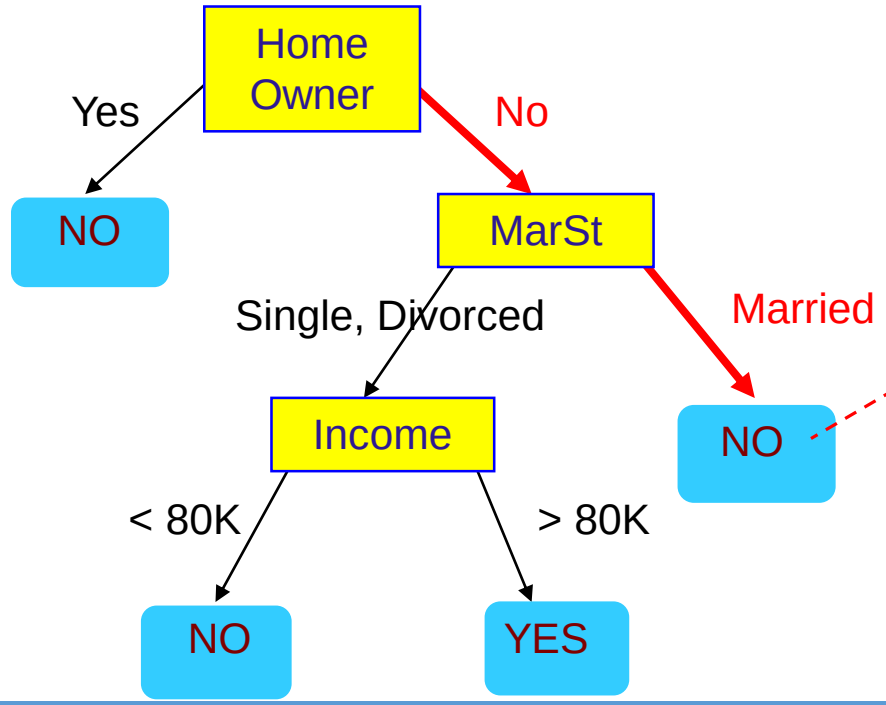
Home Owner	Marital Status	Annual Income	Defaulted Borrower
No	Married	80K	?



Apply Model to Test Data

Test Data

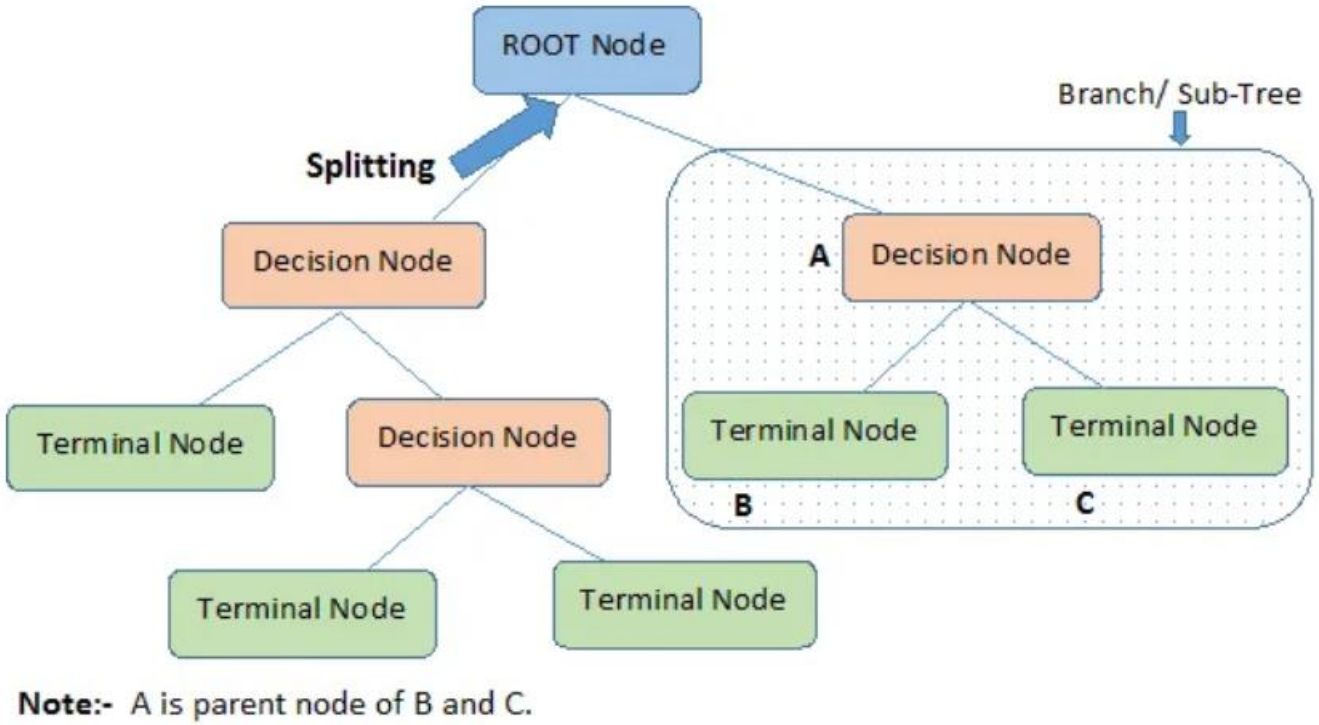
Home Owner	Marital Status	Annual Income	Defaulted Borrower
No	Married	80K	?



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?

- Many Algorithms:
  - Hunt's Algorithm (one of the earliest)
  - CART (Classification And Regression Tree)
  - ID3 (Iterative Dichotomiser 3)
  - C4.5 (Successor of ID3)

Hunt's Algorithm

A recursive fashion by partitioning the training records into successively purer subsets.

Let D_t be the set of training records that reach a node t and y = {y_1, y_2, …., y_c} be the class labels.

General Procedure:

- If D_t contains records that belong the same class y_t, then t is a leaf node labeled as y_t
- If D_t contains records that belong to more than one class, use an attribute test to split the data into smaller subsets. Recursively apply the procedure to each subset.

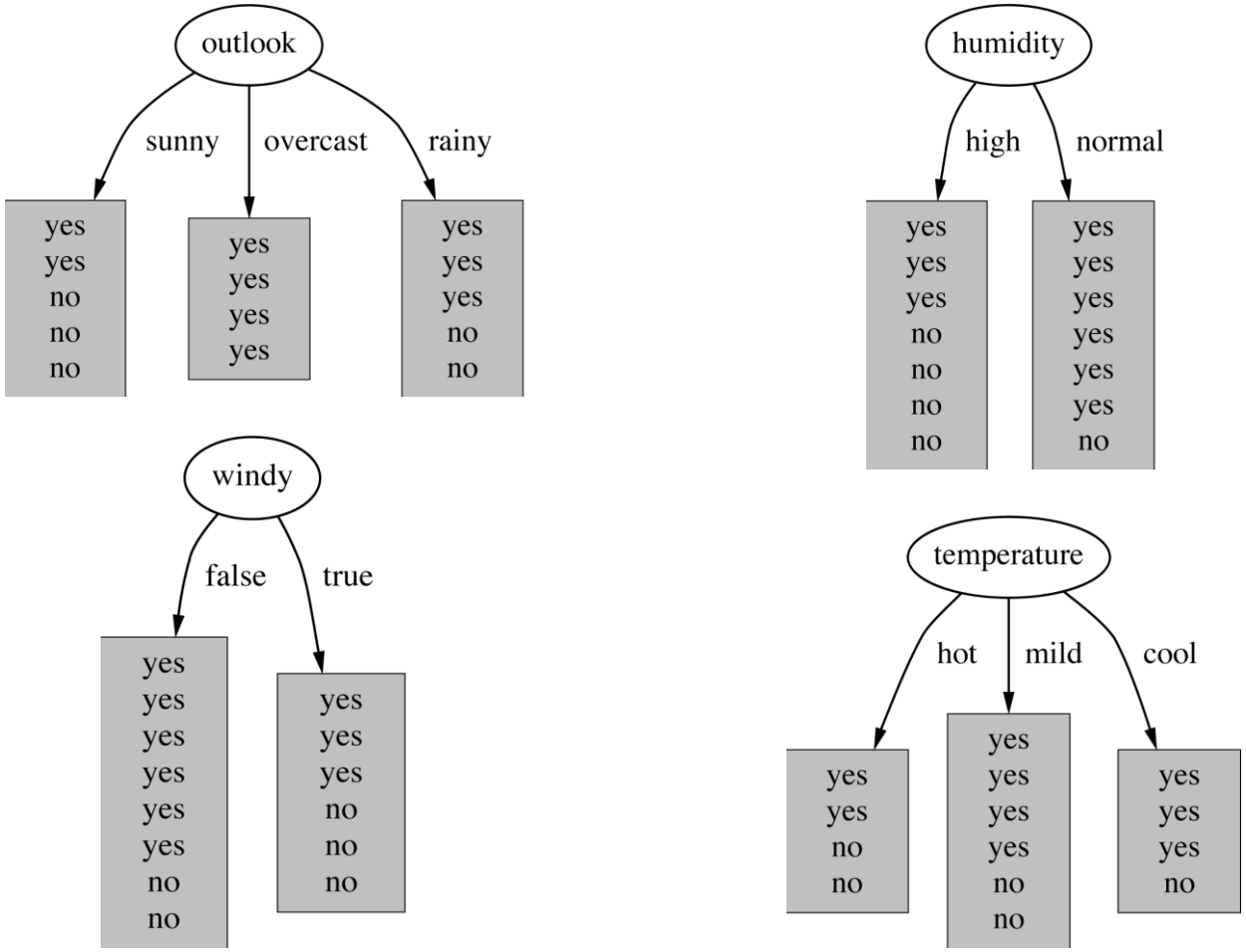
How to determine the Best Split

Attribute Selection Measures (ASM): Select an attribute to split the training records that increases the homogeneity of resultant sub-nodes with respect to the target variable

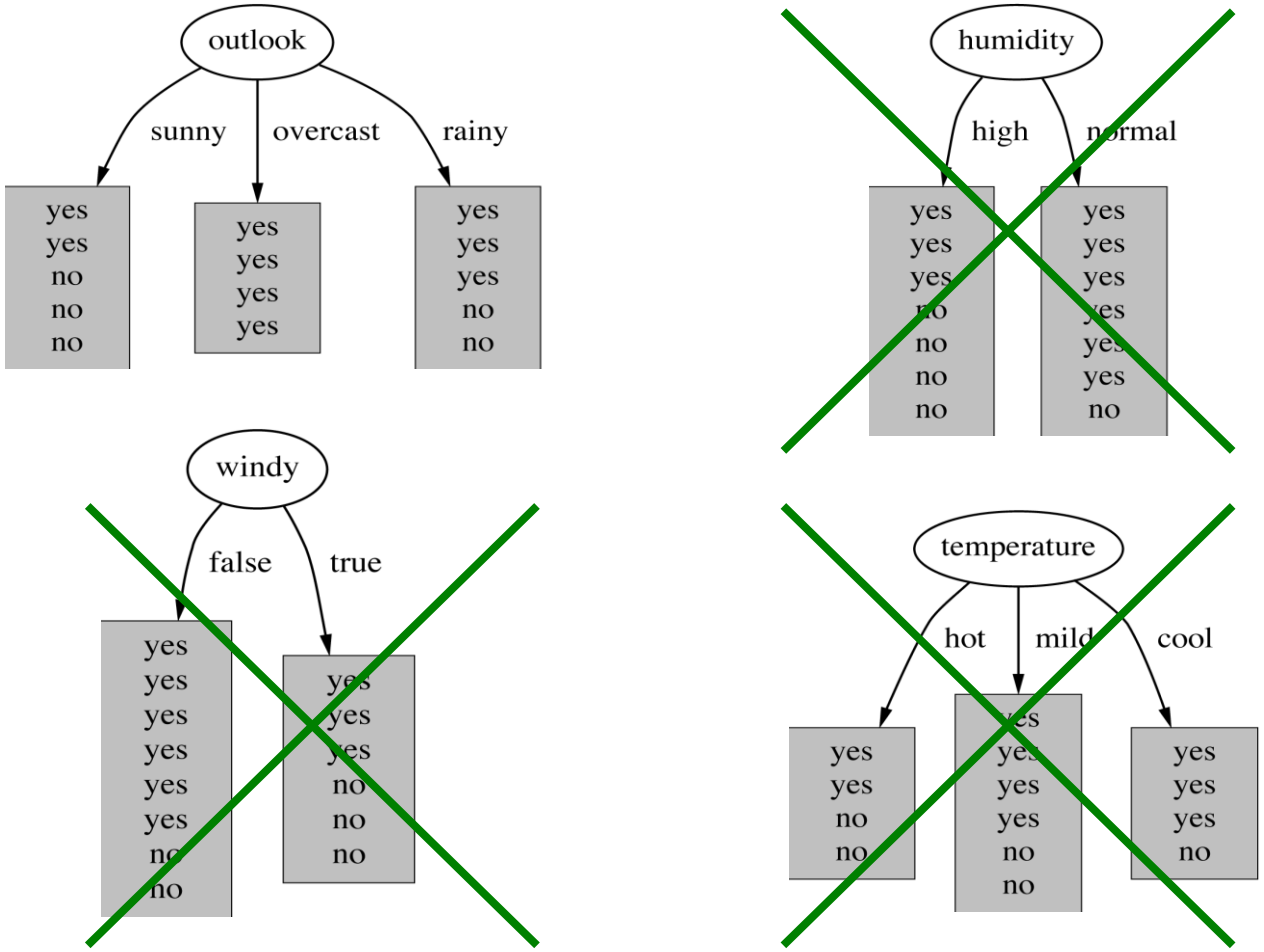
Tennis Weather: Can I play tennis today?

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

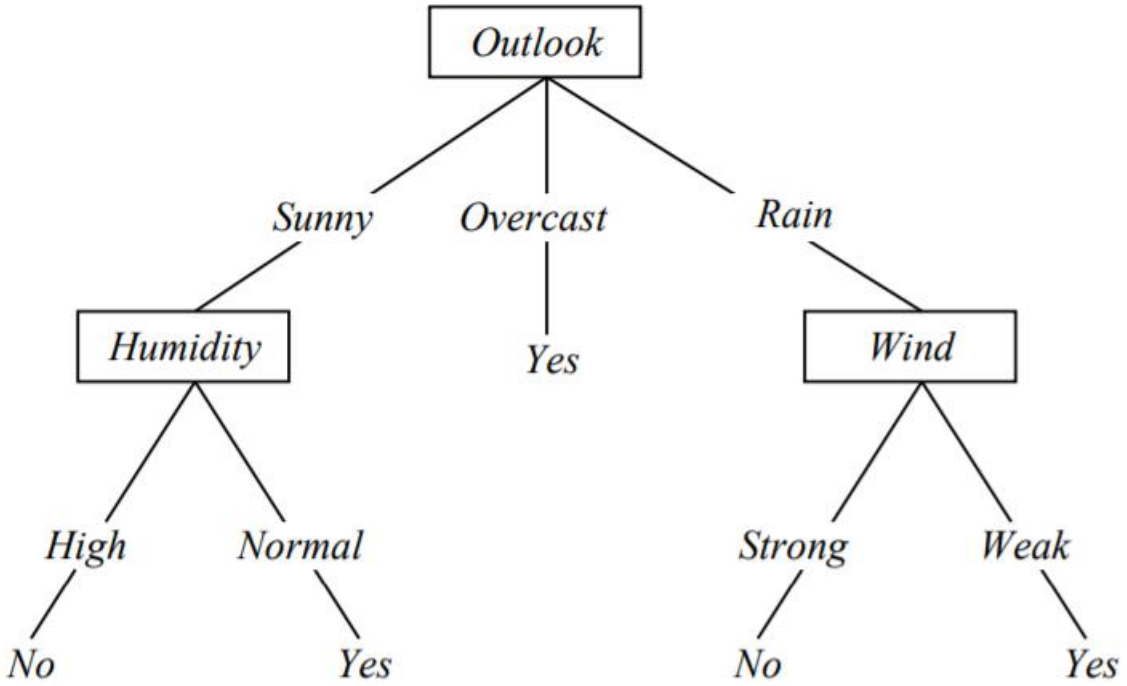
Which attribute to select?



Which attribute to select?



Decision Tree for Play Tennis



Attribute Selection Methods

Two popular methods:

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

➢Study detail from here.

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