Machine Learning

Decision Tree

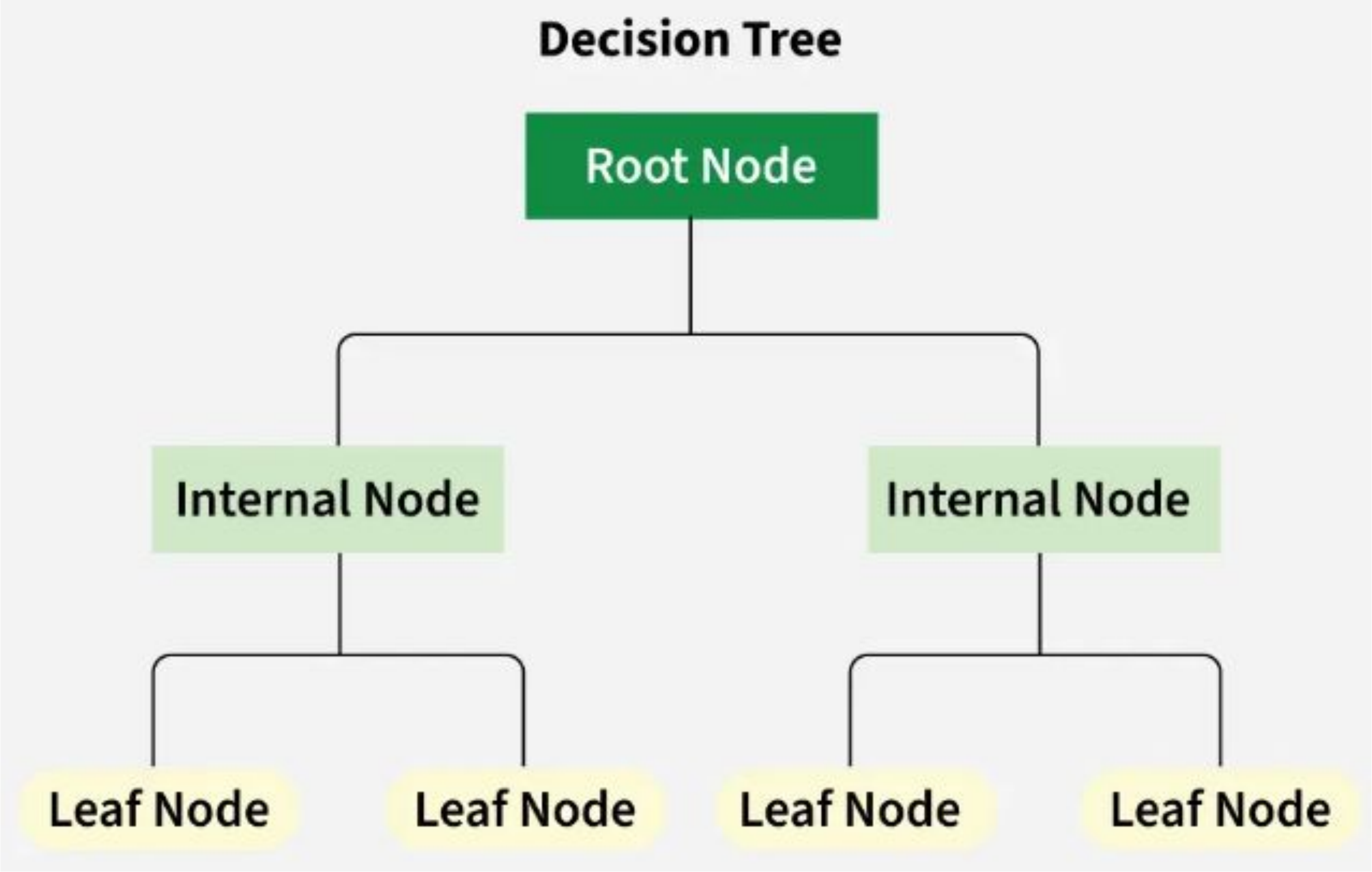
Decision Tree

- A decision tree is a supervised learning algorithm used for both classification and regression tasks.

Decision Tree

- It models decisions as a **tree-like structure** where:
  - **Root Node** is the starting point that represents the entire dataset.
  - **Branches**: These are the lines that connect nodes. It shows the flow from one decision to another.
  - **Internal Nodes** are points where decisions are made based on the input features.
  - **Leaf Nodes**: These are the terminal nodes at the end of branches that represent final outcomes or predictions

Decision Tree



Decision Tree

- mainly two types of decision tree based on the nature of the target variable:
  - **classification trees**
  - **regression trees**

Classification Tree

- **Classification trees** are designed to predict categorical outcomes.
  - means they classify data into different classes.
- They can determine whether an email is “spam” or “not spam” based on various features of the email.

Regression Tree

- **Regression trees** are used when the target variable is continuous.
- It predict numerical values rather than categories.
- For example a regression tree can estimate the price of a house based on its size, location, and other features.

Decision Tree

- How decision tree works step by step:
- **Step 1. Start with the Whole Dataset**

  We begin with all the data, which is treated as the root node of the decision tree.
- **Step 2. Choose the Best Question (Attribute)**

  Pick the best question to divide the dataset.
- **Step 3. Split the Data into Subsets**

  Divide the dataset into groups based on the question.

Decision Tree

- **Step 4. Split Further if Needed (Recursive Splitting)**

  For each subset, ask another question to refine the groups.

- **Step 5. Assign Final Decisions (Leaf Nodes)**

  When a subset contains only one activity, stop splitting and assign it a label.

- **Step 6. Use the Tree for Predictions**

  To predict an activity, follow the branches of the tree.

Decision Tree

- Two popular attribute selection measures used:
    - **Information Gain**
    - **Gini Index**

Information Gain

- **Information Gain** tells us how useful a question (or feature) is for splitting data into groups.
- It measures how much the uncertainty decreases after the split.
- A good question will create clearer groups, and the feature with the highest Information Gain is chosen to make the decision.

Gini Index

- **Gini Index** is a metric to measure how often a randomly chosen element would be incorrectly identified.
- It means an attribute with a lower Gini index should be preferred.

Uncertainty/Impurity/Entropy

- **Case Study: Ball in the Box**
- Dealing with a lack of information or uncertainty.

[WHITE BOARD]

Uncertainty/Impurity/Entropy

- In which case, the measure of entropy of a dataset must be zero?

Uncertainty/Impurity/Entropy

- In which case, the measure of entropy of a dataset must be zero?
    - when only one class is represented.

Uncertainty/Impurity/Entropy

- In which case, the measure of entropy of a dataset must be at a maximum?

Uncertainty/Impurity/Entropy

- In which case, the measure of entropy of a dataset must be at a maximum?
    - when all possible classes are equally represented.

Decision Tree

- There are essentially **two** questions that need to be answered at each step of the tree building process:
    - where to split the data
    - when to stop splitting

Where to split data?

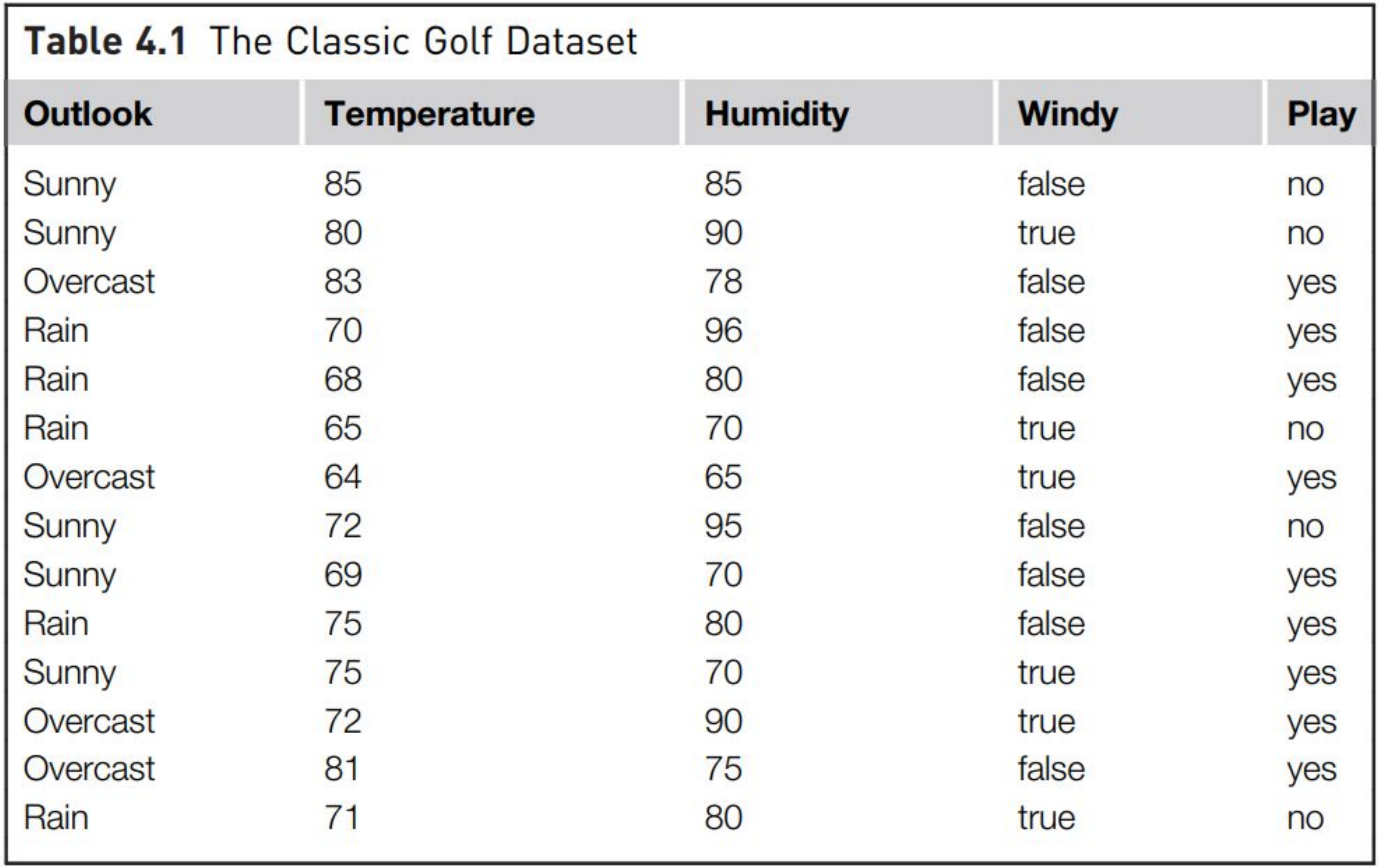


Table 4.1 The Classic Golf Dataset

Outlook	Temperature	Humidity	Windy	Play
Sunny	85	85	false	no
Sunny	80	90	true	no
Overcast	83	78	false	yes
Rain	70	96	false	yes
Rain	68	80	false	yes
Rain	65	70	true	no
Overcast	64	65	true	yes
Sunny	72	95	false	no
Sunny	69	70	false	yes
Rain	75	80	false	yes
Sunny	75	70	true	yes
Overcast	72	90	true	yes
Overcast	81	75	false	yes
Rain	71	80	true	no

Where to split data?

- Perform calculation for *Outlook* attribute.

[WHITE BOARD]

Where to split data?

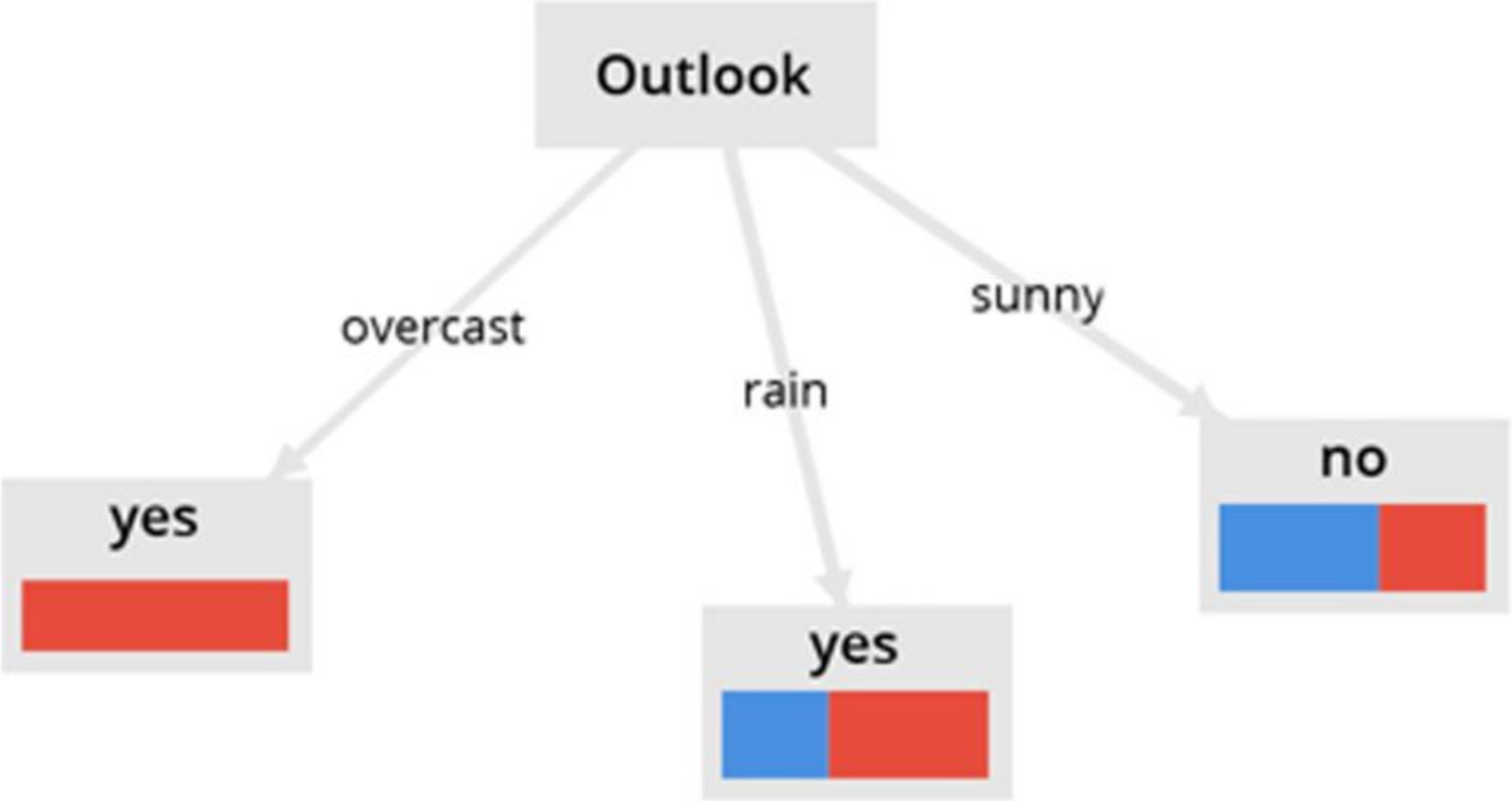


FIGURE 4.3

Splitting the Golf dataset on the Outlook attribute yields three subsets or branches. The middle and right branches may be split further.

Where to split data?

- Perform calculation for *Windy, Humidity, Temperature* attributes.
- [Shown in Excel file]

Where to split data?

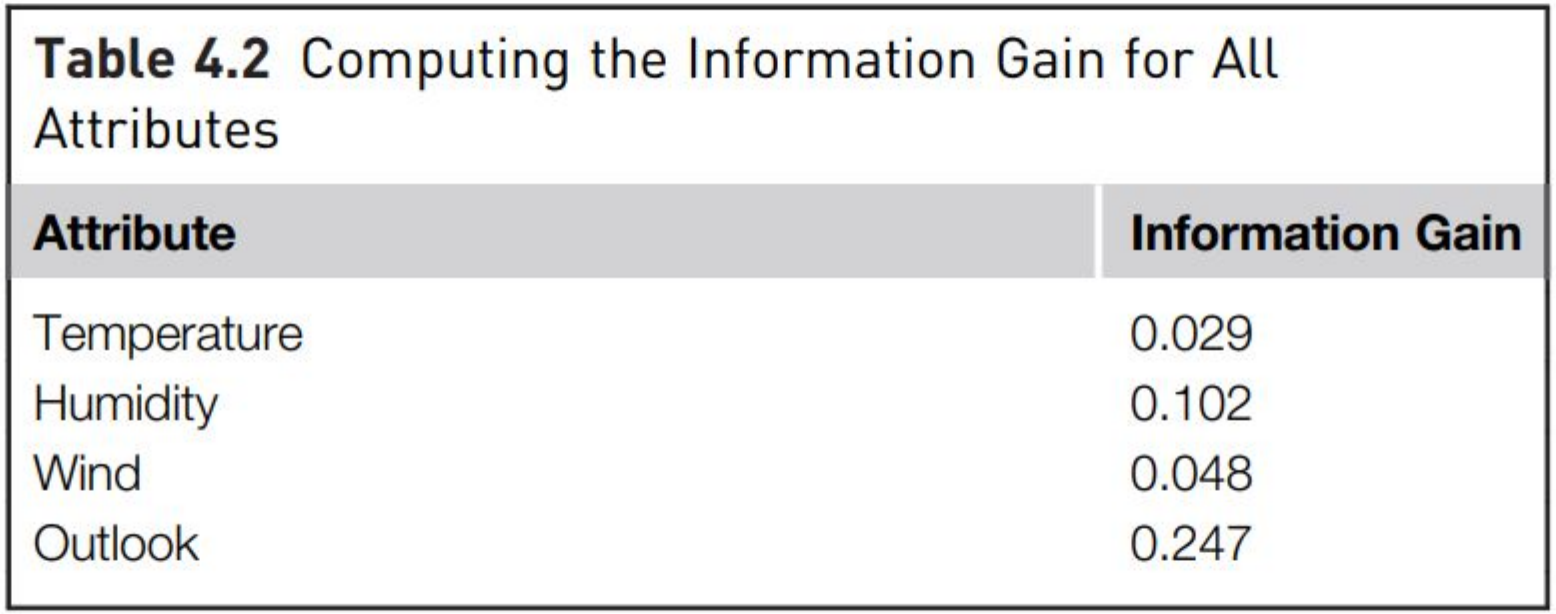


Table 4.2 Computing the Information Gain for All Attributes

Attribute	**Information Gain**
Temperature	0.029
Humidity	0.102
Wind	0.048
Outlook	0.247

Where to split data?

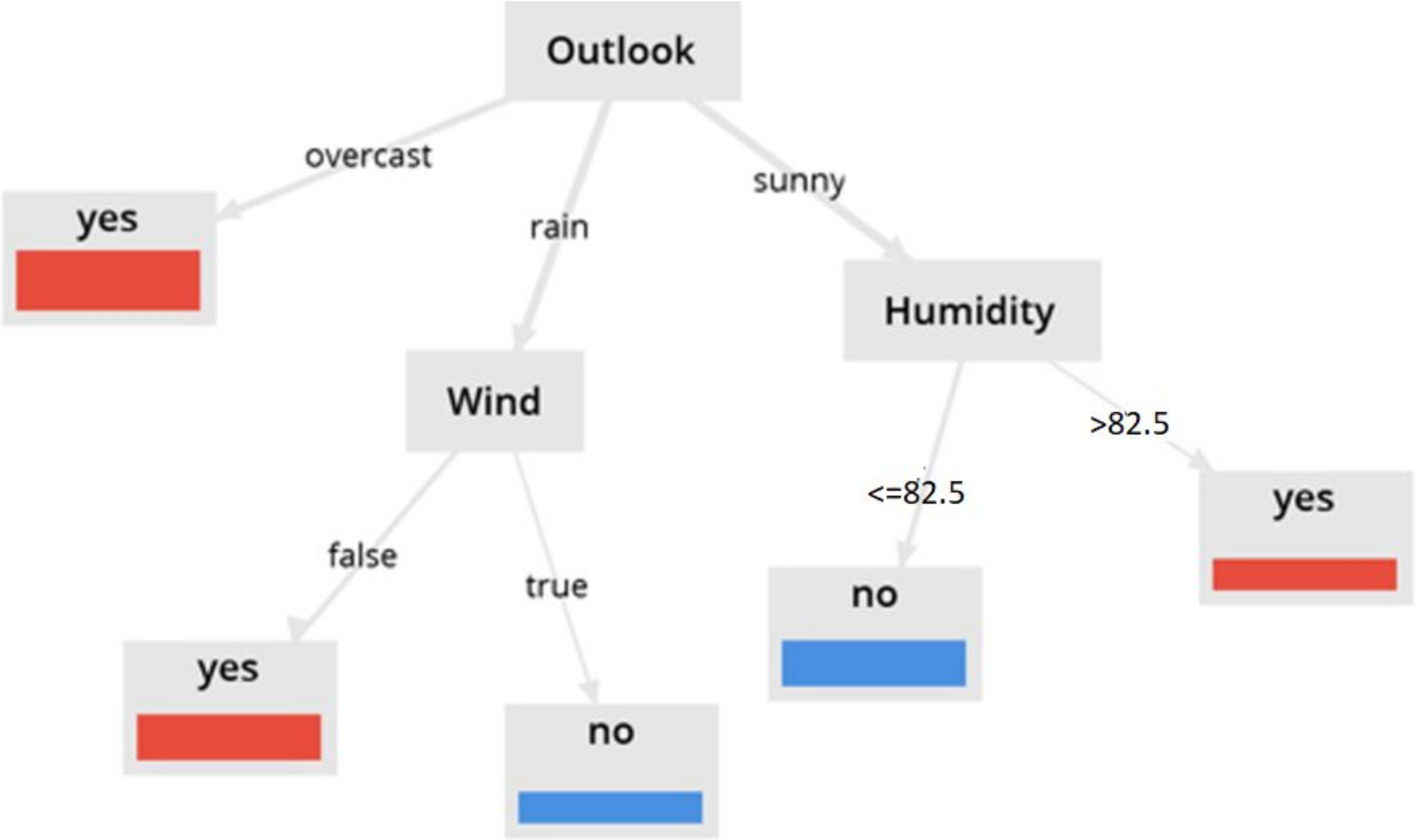


FIGURE 4.4

Decision Tree for the Golf data.

Where to stop splitting data?

- In real-world datasets, it is very unlikely that to get terminal nodes that are 100% homogeneous as was just seen in the golf dataset.
- In this case, the algorithm would need to be instructed when to stop.

Where to stop splitting data?

- There are several situations where the process can be terminated:
  - No attribute satisfies a minimum information gain threshold
  - A maximal depth is reached: as the tree grows larger, not only does interpretation get harder, but a situation called “overfitting” is induced.
  - There are less than a certain number of examples in the current subtree: again, a mechanism to prevent overfitting.

Where to stop splitting data?

- To prevent overfitting, tree growth may need to be restricted or reduced, using a process called *pruning*.

Where to stop splitting data?

- All three stopping techniques mentioned constitute what is known of as *pre-pruning* the decision tree, because the pruning occurs before or during the growth of the tree.

Where to stop splitting data?

- There are also methods that will not restrict the number of branches and allow the tree to grow as deep as the data will allow, and then trim or prune those branches that do not effectively change the classification error rates. This is called *post-pruning*.

Term Test

Term Test

X	Y	Z	O
1	1	1	A
1	1	0	A
0	0	1	B
1	0	0	B

- Which attribute to split first in building a decision tree according to the information gain of the attributes in the given dataset?