Decision Tree

Class 11 • ID3 Algorithm

Class Objective

By the end of this class, students will be able to:

- Understand what a Decision Tree is
- Explain the ID3 algorithm
- Calculate Entropy
- Calculate Information Gain
- Select the root node manually

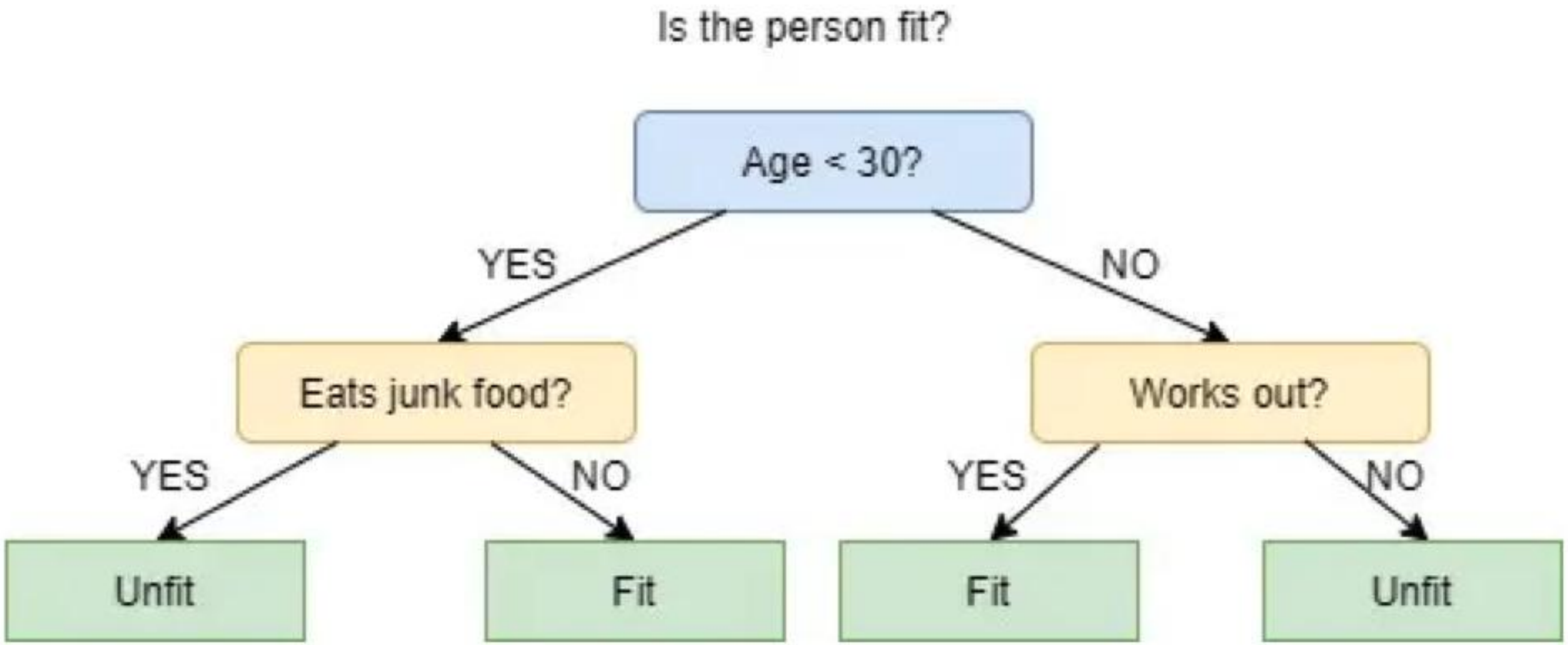
What Is a Decision Tree?

A Decision Tree is a supervised learning model that:

- Predicts output using decision rules
- Splits data based on attribute values
- Has:
  - **Decision nodes** → attributes
  - **Leaf nodes** → class labels

Used mainly for classification.

Example


Is the person fit?
Age < 30?
YES
NO
Eats junk food?
Works out?
YES
NO
YES
NO
Unfit
Fit
Fit
Unfit

ID3 Algorithm

ID3 stands for **Iterative Dichotomiser 3**

Key points:

- Top-down approach
- Greedy algorithm
- Selects attribute with maximum Information Gain
- Works best with categorical data

How Does ID3 Choose a Split?

ID3 uses:

Information Gain

The attribute that reduces uncertainty the most is chosen as the node.

To calculate Information Gain, we first need **Entropy**.

Entropy (Theory)

Entropy measures impurity / disorder in data.

- Entropy = 0 → Pure (all same class)
- Entropy = 1 → Maximum disorder (equal classes)

Formula:

$$\text{Entropy}(S) = -\sum p_i \log_2 p_i$$

Where:

- **S** = Dataset
- **p_i** = Proportion/probability of class i (e.g., p_yes = 9/14)
- **$\log_2$** = Logarithm base 2
- **Σ** = Sum over all classes

- All students pass → **No confusion** → **Entropy = 0**
- Half pass, half fail → **Maximum confusion**

Information Gain (Theory)

Information Gain measures:

Reduction in entropy after splitting on an attribute

Formula:

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum \frac{|S_v|}{|S|} \text{Entropy}(S_v)$$

Where:

- **S** = Original dataset
- **A** = Attribute to split on
- **S_v** = Subset v after splitting by attribute A
- **|S_v|** = Number of records in subset v
- **|S|** = Total number of records in S
- **Σ** = Sum over all subsets created by splitting on A

In short:

1. **Find entropy before split**
2. **Find entropy after split**
3. **Subtract**

Numerical Example Dataset

Target: Play Tennis (Yes / No)

S. No.	Outlook	Temperature	Humidity	Windy	PlayTennis
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rainy	Mild	High	Weak	Yes
5	Rainy	Cool	Normal	Weak	Yes
6	Rainy	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rainy	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rainy	Mild	High	Strong	No

Step 1: Entropy of Dataset

Total records = **14**

- Yes = **9**
- No = **5**

$$\text{Entropy}(S) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14}$$

Entropy(S) = 0.94

Step 2: Split on Attribute – Outlook

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

Entropy for Each Split

Sunny:

$$\text{Entropy} = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.97$$

Overcast:

$$\text{Entropy} = 0 \quad \text{(pure)}$$

Rain:

$$\text{Entropy} = 0.97$$

Step 3: Information Gain (Outlook)

$$\text{Gain}(S, \text{Outlook}) = 0.94 - \left(\frac{5}{14} \times 0.97 + \frac{4}{14} \times 0 + \frac{5}{14} \times 0.97\right)$$

Gain(S, Outlook) = 0.246

Root Node Selection

- Calculate Information Gain for all attributes
- Attribute with highest gain becomes:

🌳 **Root Node**

➡️ **Outlook**

Building the Decision Tree

```
         Outlook (Root)
          /    |    \
         /     |     \
  Sunny  Overcast  Rain
   /       |        /  \
  /        |       /    \
?        YES     ?      ?
         /   \
        /     \
     NO     YES
```

After splitting on **Outlook**:

- **Overcast** → Pure (all YES) ✓
- **Sunny** & **Rain** → Need further splits

Next Step: Analyze Sunny Node

Sunny subset: 5 records (2 Yes, 3 No)
Entropy(Sunny) = 0.97

Need to check other attributes:

- Temperature, Humidity, Wind, etc.

Calculate Information Gain for each attribute within Sunny subset.

Next Step: Analyze Rain Node

Rain subset: 5 records (3 Yes, 2 No)
Entropy(Rain) = 0.97

Need to check other attributes to further split.

Example: Wind might be the best split here

ID3 Algorithm Steps (Exam Answer)

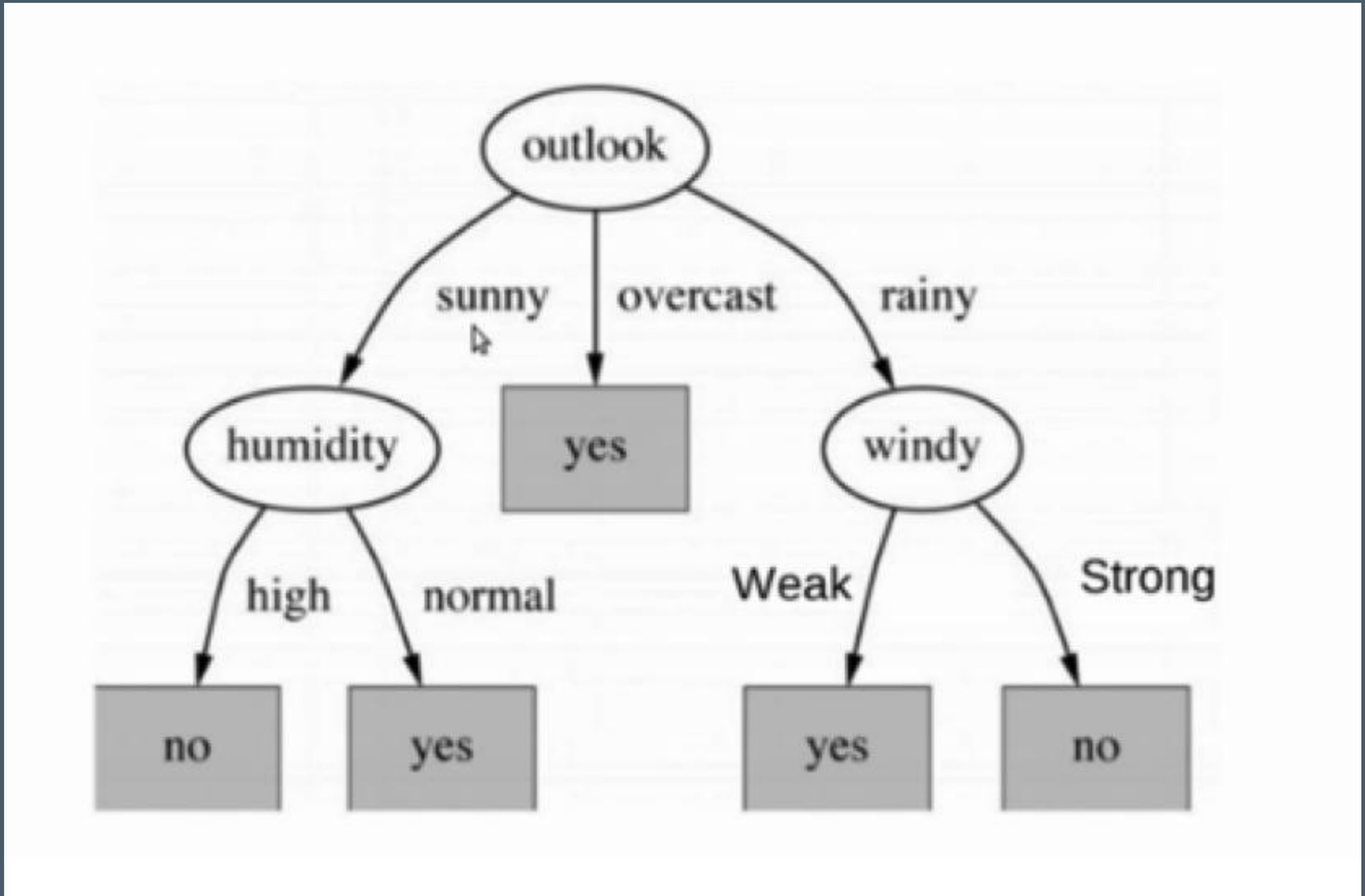
1. Calculate entropy of dataset
2. Calculate information gain for each attribute
3. Select attribute with highest gain
4. Split dataset
5. Repeat until:
    - All nodes are pure OR
    - No attributes left

Partial Decision Tree Summary

```
              Outlook [Gain = 0.246]
             /           |           \
            /            |            \
      Sunny         Overcast         Rain
    (2Y, 3N)        (4Y, 0N)       (3Y, 2N)
        |              |               |
        ?             YES              ?
                                  (Need split)
```

- **Overcast** is pure (leaf node)
- **Sunny** and **Rain** need further analysis

Final Decision Tree



Key Takeaways

✓ ID3 builds trees top-down using greedy approach
✓ Information Gain determines which attribute to split on
✓ Process repeats recursively on subsets
✓ Stops when all nodes are pure or no attributes remain

Next class: Complete decision tree + Handling continuous attributes