Complete Working of Decision Trees for Classification

A **Decision Tree** is a supervised machine learning algorithm used for classification and regression tasks. In classification, a decision tree splits the dataset into branches based on the feature that best divides the data into different classes. Each internal node in the tree represents a decision based on one of the features, and each leaf node represents a class label (the predicted output).

Let's walk through the **complete working of a Decision Tree** for classification in a step-by-step process with **formulas** and **numerical examples**. I will break down each step and explain it in easy-to-understand terms.

Step-by-Step Process for Building a Decision Tree for Classification

Step 1: Understanding the Data

Let's say we have a dataset of animals with features like "Has Fur" and "Is Aquatic" and a target variable "Class" (Mammal or Not Mammal).

Animal	Has Fur	Is Aquatic	Class
Dog	Yes	No	Mammal
Cat	Yes	No	Mammal
Whale	No	Yes	Not Mammal
Fish	No	Yes	Not Mammal
Tiger	Yes	No	Mammal

Step 2: Calculate the Entropy of the Entire Dataset

The first thing we do is calculate the **entropy** of the whole dataset. Entropy measures how "mixed" the data is in terms of the class labels (Mammal or Not Mammal).

- We have 3 Mammals (Dog, Cat, Tiger) and 2 Not Mammals (Whale, Fish).

Entropy (H(S)) is calculated using the formula:

$$H(S) = -\sum_{i=1}^{k} p_i \log_2(p_i)$$

Where:

- p_i is the probability of class i in the set S,
- k is the total number of distinct classes.

For our dataset:

- The probability of being a Mammal is $p_{\text{Mammal}} = \frac{3}{5}$,
- The probability of being a Not Mammal is $p_{\text{Not Mammal}} = \frac{2}{5}$.

Now, plug these values into the entropy formula:

$$H(S) = -\left(\frac{3}{5}\log_2\frac{3}{5} + \frac{2}{5}\log_2\frac{2}{5}\right)$$

First, calculate the log values:

- $\log_2\frac{3}{5} \approx -0.737$
- $\log_2\frac{2}{5} \approx -1.322$

Now, calculate the entropy:

$$H(S) = -\left(\frac{3}{5} \times -0.737 + \frac{2}{5} \times -1.322\right) = 0.442 + 0.528 = 0.971$$

Interpretation:

- The entropy of the dataset is 0.971, which means there is some uncertainty in the data because we have both classes (Mammal and Not Mammal).

Step 3: Calculate the Information Gain for Each Feature

Next, we calculate the **Information Gain (IG)** for each feature (Has Fur, Is Aquatic). The goal is to see which feature provides the best way to split the data, reducing uncertainty (entropy).

Information Gain for a feature A is given by:

$$IG(S, A) = H(S) - \sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|} H(S_v)$$

Where:

- $H(S)$ is the entropy of the whole dataset,
- S_v is the subset of data where feature A has value v,
- $|S_v|$ is the number of elements in S_v,
- $|S|$ is the total number of elements in the original dataset.

Information Gain for "Has Fur":

Let's split the data based on the "Has Fur" feature:

- **Has Fur = Yes**: Dog, Cat, Tiger (all Mammals)
- **Has Fur = No**: Whale, Fish (all Not Mammals)

Now calculate the entropy for each subset:

- For **Has Fur = Yes**, all 3 animals are Mammals, so the entropy is 0 (pure).
- For **Has Fur = No**, all 2 animals are Not Mammals, so the entropy is 0 (pure).

The **weighted entropy** for the "Has Fur" split is:

$$H(S_{\text{Has Fur}}) = \frac{3}{5} \times 0 + \frac{2}{5} \times 0 = 0$$

Thus, the **Information Gain for "Has Fur"** is:

$$IG(S, \text{Has Fur}) = H(S) - H(S_{\text{Has Fur}}) = 0.971 - 0 = 0.971$$

Information Gain for "Is Aquatic":

Now, split the data based on the "Is Aquatic" feature:

- **Is Aquatic = Yes**: Whale, Fish (Not Mammals)
- **Is Aquatic = No**: Dog, Cat, Tiger (Mammals)

Now calculate the entropy for each subset:

- For **Is Aquatic = Yes**, all 2 animals are Not Mammals, so the entropy is 0.
- For **Is Aquatic = No**, all 3 animals are Mammals, so the entropy is 0.

The **weighted entropy** for the "Is Aquatic" split is:

$$H(S_{\text{Is Aquatic}}) = \frac{2}{5} \times 0 + \frac{3}{5} \times 0 = 0$$

Thus, the **Information Gain for "Is Aquatic"** is:

$$IG(S, \text{Is Aquatic}) = H(S) - H(S_{\text{Is Aquatic}}) = 0.971 - 0 = 0.971$$

Step 4: Choose the Best Split

From the calculations, we can see that **both features ("Has Fur" and "Is Aquatic") give the same Information Gain** of 0.971. In cases like this, the algorithm can choose either feature to split first. However, to keep things simple, we'll choose **"Has Fur"** as the first split.

Step 5: Recursively Build the Tree

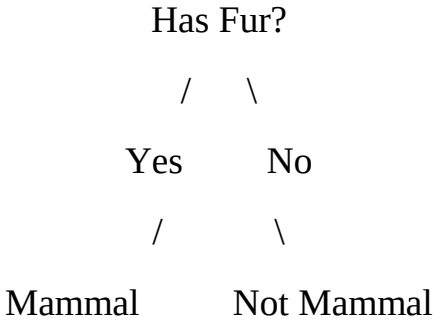
After the first split (based on "Has Fur"), we look at the resulting subsets:

- **Subset 1 (Has Fur = Yes)**: Dog, Cat, Tiger → All Mammals (pure node).
- **Subset 2 (Has Fur = No)**: Whale, Fish → All Not Mammals (pure node).

Since both subsets are pure (they contain only one class), there's no need for further splitting. These are the leaf nodes of the tree.

Final Tree Structure

The decision tree will look like this:



Interpretation:

- If an animal has fur, it is classified as a **Mammal**.
- If an animal does not have fur, it is classified as **Not Mammal**.

Step 6: Classifying a New Sample

Now let's classify a new sample, say a **Lion**, which has fur but is not aquatic:

- First, we check the **"Has Fur"** feature.
- Since the Lion has fur, it will be classified as a **Mammal**.

Conclusion

To summarize, building a decision tree for classification involves the following steps:

1. **Calculate the entropy** of the entire dataset to measure how "mixed" the classes are.
2. **Calculate Information Gain** for each feature, which tells us how much uncertainty (entropy) is reduced by splitting the data based on that feature.
3. **Choose the feature with the highest Information Gain** to split the data.
4. **Recursively repeat** the process for each subset until the data in each subset is pure (or other stopping criteria are met).
5. The resulting tree can then be used to classify new data by following the splits based on feature values.

Example of Multiclass Classification Decision Tree

Let's consider an example where we have a dataset of **animals**, and the task is to classify them into one of three classes: **Mammal**, **Bird**, or **Reptile**.

Animal	Has Fur	Can Fly	Cold-blooded	Class
Dog	Yes	No	No	Mammal
Cat	Yes	No	No	Mammal
Sparrow	No	Yes	No	Bird
Lizard	No	No	Yes	Reptile
Whale	Yes	No	No	Mammal
Parrot	No	Yes	No	Bird
Snake	No	No	Yes	Reptile

Step 1: Calculate the Entropy of the Entire Dataset

To begin, we calculate the **entropy** of the whole dataset. Entropy measures the disorder or uncertainty in the dataset, and it helps us determine how "mixed" the classes are.

Entropy Formula:

Entropy Formula:

$$H(S) = -\sum_{i=1}^{k} p_i \log_2(p_i)$$

Where:

- p_i is the probability of class i in the dataset,
- k is the number of classes in the dataset.

Step 1.1: Calculate the class probabilities

We have 7 total samples, with 3 Mammals, 2 Birds, and 2 Reptiles.

$$p_{\text{Mammal}} = \frac{3}{7}, \quad p_{\text{Bird}} = \frac{2}{7}, \quad p_{\text{Reptile}} = \frac{2}{7}$$

Step 1.2: Compute the entropy

Now, we calculate the entropy of the dataset:

$$H(S) = -\left(\frac{3}{7}\log_2\frac{3}{7} + \frac{2}{7}\log_2\frac{2}{7} + \frac{2}{7}\log_2\frac{2}{7}\right)$$

Let's calculate each log term:

- $\log_2 \frac{3}{7} \approx -1.222$
- $\log_2 \frac{2}{7} \approx -1.807$

Now, substitute the values:

$$H(S) = -\left(\frac{3}{7} \times -1.222 + \frac{2}{7} \times -1.807 + \frac{2}{7} \times -1.807\right)$$

$$H(S) = 0.522 + 0.515 + 0.515 = 1.552$$

So, the **entropy** of the entire dataset is approximately **1.552**.

Step 2: Calculate the Information Gain for Each Feature

We now need to calculate the **Information Gain** for each feature ("Has Fur," "Can Fly," and "Cold-blooded"). The feature that provides the highest Information Gain will be selected for the first split.

Information Gain Formula:

$$IG(S, A) = H(S) - \sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|} H(S_v)$$

Where:

- $H(S)$ is the entropy of the original dataset,
- S_v is the subset of the dataset where feature A takes value v,
- $|S_v|$ is the number of samples in subset S_v,
- $|S|$ is the total number of samples in the dataset.

Step 2.1: Information Gain for "Has Fur"

Split the dataset based on "Has Fur":

- **Has Fur = Yes**: Dog, Cat, Whale (All Mammals)
  - This subset has 3 samples, and all are Mammals.
  - The **entropy** of this subset is 0 because it's pure (all samples belong to one class).
- **Has Fur = No**: Sparrow, Lizard, Parrot, Snake (Birds and Reptiles)
  - This subset has 4 samples: 2 Birds and 2 Reptiles.
  - We need to calculate the entropy of this subset.

Entropy of the "Has Fur = No" subset:

For the "Has Fur = No" subset, we have 2 Birds and 2 Reptiles. The probabilities are:

- $p_{\text{Bird}} = \frac{2}{4} = 0.5,$
- $p_{\text{Reptile}} = \frac{2}{4} = 0.5.$

Now, compute the entropy for this subset:

$$H(S_{\text{Has Fur = No}}) = -\left(0.5 \log_2 0.5 + 0.5 \log_2 0.5\right)$$

$$H(S_{\text{Has Fur = No}}) = -(0.5 \times -1 + 0.5 \times -1) = 1$$

Calculate Information Gain for "Has Fur":

Now, we calculate the Information Gain for "Has Fur" by subtracting the weighted entropy of the subsets from the entropy of the original dataset.

$$IG(S, \text{Has Fur}) = H(S) - \left(\frac{3}{7} \times 0 + \frac{4}{7} \times 1\right)$$

$$IG(S, \text{Has Fur}) = 1.552 - (0 + 0.571) = 1.552 - 0.571 = 0.981$$

Thus, the **Information Gain for "Has Fur"** is **0.981**.

Step 2.2: Information Gain for "Can Fly"

Now, let's compute the Information Gain for the feature **"Can Fly"**.

Split the dataset based on "Can Fly":

- **Can Fly = Yes**: Sparrow, Parrot (Birds)
  - This subset has 2 samples, both belonging to the "Bird" class.
  - The entropy of this subset is 0 because it's pure (all samples belong to one class).
- **Can Fly = No**: Dog, Cat, Whale, Lizard, Snake (Mammals and Reptiles)
  - This subset has 5 samples: 3 Mammals and 2 Reptiles.

- We need to calculate the entropy of this subset.

Entropy of the "Can Fly = No" subset:

For the "Can Fly = No" subset, we have 3 Mammals and 2 Reptiles. The probabilities are:

- $p_{\text{Mammal}} = \frac{3}{5} = 0.6$,
- $p_{\text{Reptile}} = \frac{2}{5} = 0.4$.

Now, compute the entropy for this subset:

$$H(S_{\text{Can Fly = No}}) = -\left(0.6 \log_2 0.6 + 0.4 \log_2 0.4\right)$$

$$H(S_{\text{Can Fly = No}}) = -(0.6 \times -0.736 + 0.4 \times -1.322) = 0.442 + 0.529 = 0.971$$

Calculate Information Gain for "Can Fly":

Now, we calculate the Information Gain for "Can Fly":

$$IG(S, \text{Can Fly}) = H(S) - \left(\frac{2}{7} \times 0 + \frac{5}{7} \times 0.971\right)$$

$$IG(S, \text{Can Fly}) = 1.552 - (0 + 0.693) = 1.552 - 0.693 = 0.859$$

Thus, the **Information Gain for "Can Fly"** is **0.859**.

Step 2.3: Information Gain for "Cold-blooded"

Let's compute the Information Gain for **"Cold-blooded"**.

Split the dataset based on "Cold-blooded":

- **Cold-blooded = Yes**: Lizard, Snake (Reptiles)
  - This subset has 2 samples, both belonging to the "Reptile" class.
  - The entropy of this subset is 0 because it's pure (all samples belong to one class).
- **Cold-blooded = No**: Dog, Cat, Sparrow, Parrot, Whale (Mammals and Birds)
  - This subset has 5 samples: 3 Mammals and 2 Birds.
  - We need to calculate the entropy of this subset.

Entropy of the "Cold-blooded = No" subset:

For the "Cold-blooded = No" subset, we have 3 Mammals and 2 Birds. The probabilities are:

- $p_{\text{Mammal}} = \frac{3}{5} = 0.6$,
- $p_{\text{Bird}} = \frac{2}{5} = 0.4$.

Now, compute the entropy for this subset:

$$H(S_{\text{Cold-blooded = No}}) = -\left(0.6 \log_2 0.6 + 0.4 \log_2 0.4\right)$$

$$H(S_{\text{Cold-blooded = No}}) = -(0.6 \times -0.736 + 0.4 \times -1.322) = 0.442 + 0.529 = 0.971$$

Calculate Information Gain for "Cold-blooded":

Now, we calculate the Information Gain for "Cold-blooded":

$$IG(S, \text{Cold-blooded}) = H(S) - \left(\frac{2}{7} \times 0 + \frac{5}{7} \times 0.971\right)$$

$$IG(S, \text{Cold-blooded}) = 1.552 - (0 + 0.693) = 1.552 - 0.693 = 0.859$$

Thus, the **Information Gain for "Cold-blooded"** is **0.859**.

tep 3: Choose the Best Feature for the First Split

Now we compare the Information Gains for each feature:

- **"Has Fur"**: Information Gain = 0.981
- **"Can Fly"**: Information Gain = 0.859
- **"Cold-blooded"**: Information Gain = 0.859

1. We found that **"Has Fur"** provided the highest Information Gain, so it is selected as the first feature to split on.
2. **First Split**: Split the dataset based on **"Has Fur"**:
    - **Has Fur = Yes** → All samples are **Mammals**.
    - **Has Fur = No** → Split further using **"Can Fly"** or **"Cold-blooded"**.

Step 4: Build the Tree (Continued)

Since we split on **Has Fur** first, the tree will have two branches: one for **Has Fur = Yes** and another for **Has Fur = No**.

Branch 1: Has Fur = Yes

- All animals in this subset are **Mammals** (Dog, Cat, Whale).
- This is a **pure node**, so no further splits are needed, and we assign the class **Mammal**.

Branch 2: Has Fur = No

We will now split this subset based on the next best feature, which is **"Can Fly"** (because it had the same Information Gain as "Cold-blooded" but we can choose either).

- **Can Fly = Yes**: These animals are **Birds** (Sparrow, Parrot).
  - This is a **pure node** because all samples in this subset are **Birds**.
- **Can Fly = No**: These animals are **Reptiles** (Lizard, Snake).
  - This is a **pure node** because all samples in this subset are **Reptiles**.

Final Decision Tree Structure