**What is a Decision Tree?**

A **Decision Tree** is like a flowchart that helps you make decisions based on answers to a series of questions. Imagine you’re making a decision, like whether or not to go outside and play, and each question helps you narrow down the choices.

A computer screen shot of a diagram

Description automatically generated

**Key Parts of a Decision Tree:**

1. **Root Node**:
   * This is where everything starts. Think of it like the big decision you need to make, like "Should I go outside?"
   * It represents the whole set of possible choices and is split into smaller questions based on features of the situation (like "Is it sunny?" or "Is it raining?").
2. **Internal Nodes**:
   * These are like the decision points, where you ask questions like, "Is it too hot to go outside?" or "Is it windy?"
   * Each decision point splits the choices into different possibilities. Based on the answer (yes or no), you keep going down the tree to more questions or to the final decision.
3. **Leaf Nodes**:
   * These are the end points of the tree, where you get the final decision.
   * For example, after going through all the questions, you’ll either decide "Go outside" or "Stay inside" at the leaf nodes.
4. **Edges**:
   * These are like the paths between the questions. They represent the answers to the questions, such as "Yes" or "No," that lead you to the next decision point or the final outcome.

**How Does a Decision Tree Work?**

**Splitting:**

* Imagine you start at the root (the "Should I go outside?" question). Then, you split into different options (for example, "Is it sunny?" or "Is it raining?").
* The tree asks the questions that give you the best chance of narrowing things down. It picks the features (like "Is it sunny?" or "Is it windy?") that most help in making a good decision.

**Stopping Criteria:**

* The tree keeps asking questions until:
  + All the possible answers point to the same decision (like everyone agrees that it’s sunny and not windy, so you go outside).
  + The tree gets too deep (meaning you’ve asked too many questions).
  + There aren’t enough answers left to split anymore (you’ve run out of information to ask).

**Pruning:**

* Sometimes, a tree can get too big or too detailed and start making decisions based on irrelevant or unimportant things (this is called overfitting).
* **Pruning** is when we cut off some of those unnecessary branches to make the tree simpler and better at making decisions with new information.

**In Short:**

A **Decision Tree** is like a map that helps you make a series of decisions. It starts with a big question and then keeps asking smaller questions until it makes a final choice. It’s simple to understand, and the questions help you make decisions step by step!

**1. Entropy:**

In the context of decision trees, **entropy** is a measure of uncertainty or impurity in a dataset. It tells us how mixed or "disorderly" the data is. The lower the entropy, the more "pure" the data is. The higher the entropy, the more mixed or uncertain the data is.

* **High Entropy**: The data is mixed, meaning it has both classes (e.g., both "Yes" and "No").
* **Low Entropy**: The data is pure, meaning it mostly contains one class (e.g., mostly "Yes").

**Formula for Entropy:**

Entropy=−p1log2p1−p2log2p2

**Example:**

Imagine a group of 10 people, and we want to know if they like playing outside based on the weather:

* 6 say "Yes," and 4 say "No."

The **Entropy** would be higher because there is a mix of "Yes" and "No" answers, and it’s harder to predict what the next person will say.

**2. Gini Impurity:**

**Gini Impurity** is another way to measure how mixed the data is. It's used to decide which feature (question) should be asked next in a decision tree. A **lower Gini value** means the data is more "pure" (more similar answers), and a **higher Gini value** means the data is more "impure" (more mixed answers).

* **Low Gini Impurity**: The data is pure (most answers are the same).
* **High Gini Impurity**: The data is mixed (answers are spread out).

**Formula for Gini Impurity:**

Gini=1−∑(pi)^2 =

* Pi is the probability of each class in the dataset.

**Example:**

For the same group of 10 people:

* 6 say "Yes," and 4 say "No."

The **Gini Impurity** is calculated as: Gini=1-((6/10)^2+(4/10)^2)=0.48

This shows a moderate mix between the classes, meaning it’s not a perfect decision yet.

**3. Information Gain:**

In decision trees, we use **Information Gain** to choose which feature (question) to split on. **Information Gain** is the reduction in entropy after a split. The idea is that the more **Information Gain** we get, the better the feature is at splitting the data into pure groups.

* **High Information Gain**: A feature (question) that divides the data into pure groups.
* **Low Information Gain**: A feature that doesn’t divide the data well.

**Formula for Information Gain:**

Information Gain=Entropy (before split)−Weighted Entropy (after split)

| **Is it sunny?** | **Is it hot?** | **Play outside?** |
| --- | --- | --- |
|  |  |  |
| Yes | No | Yes |
| No | Yes | No |
| Yes | Yes | No |
| Yes | No | Yes |

For the **"Is it sunny?"** feature, the split would be:

* If **Yes**: 2 "Yes" and 1 "No" → relatively pure.
* If **No**: 1 "Yes" and 1 "No" → less pure.

**Summary:**

* **Entropy** measures disorder or uncertainty in the data. Lower entropy means the data is more pure.
* **Gini Impurity** is similar to entropy but focuses on how mixed the classes are.
* **Information Gain** tells us how much information we get by splitting the data based on a certain feature.