
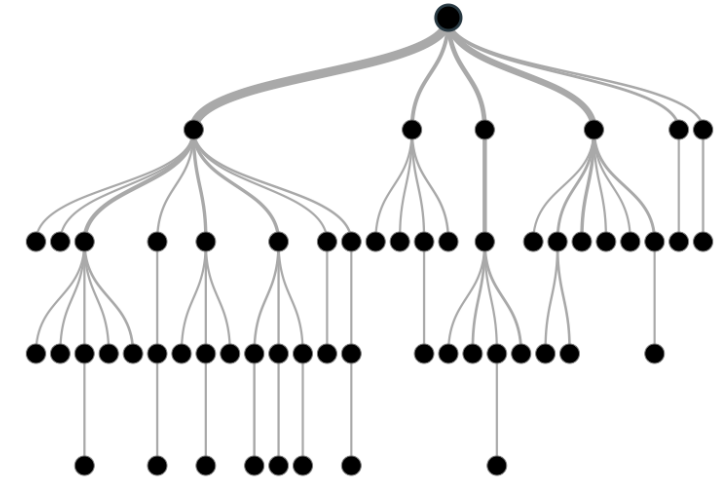


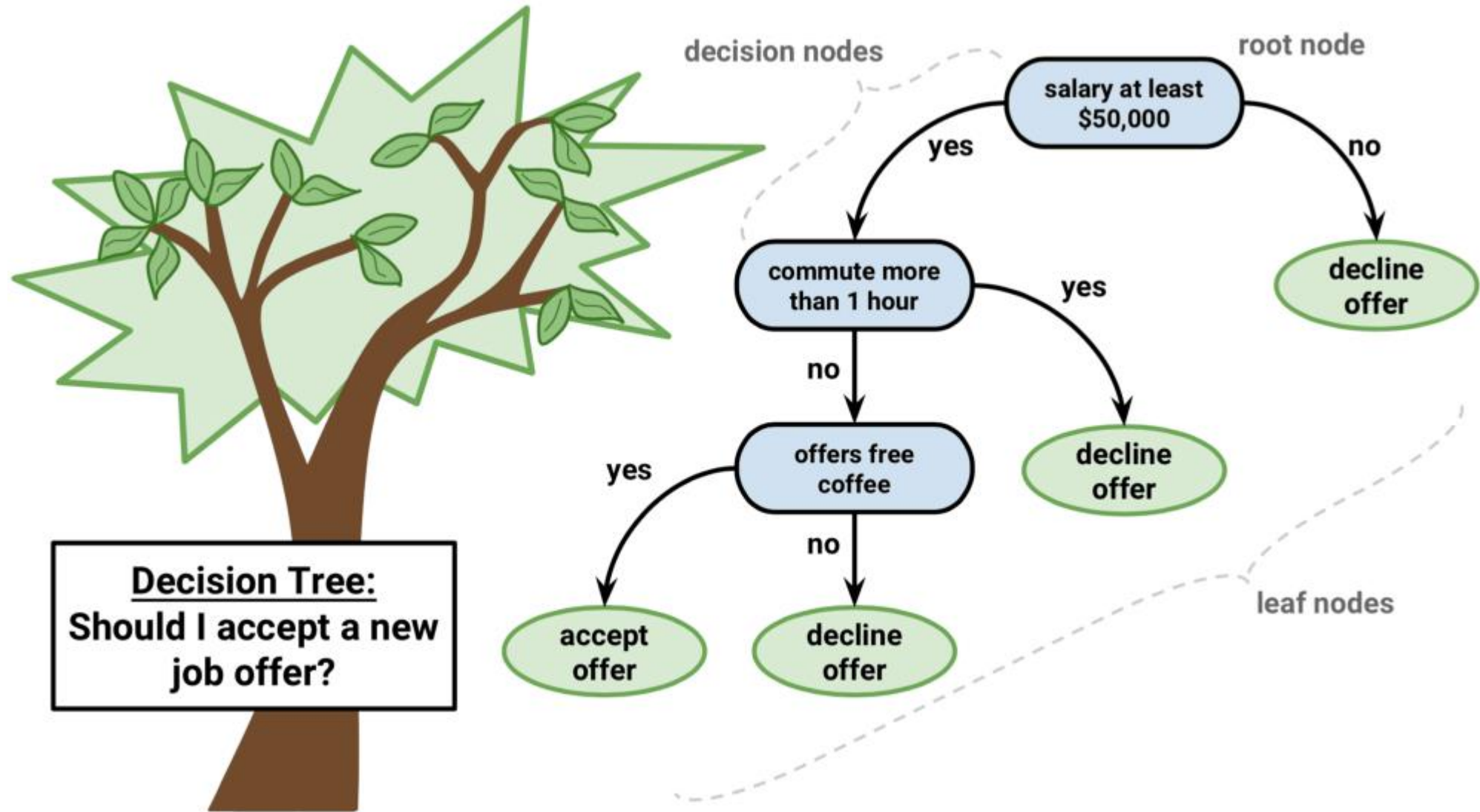
# Decision tree

# Decision tree

- Decision tree is a type of supervised learning algorithm that is mostly used in classification problems.
  - Decision trees often mimic the human level thinking so its so simple to understand the data and make some good interpretations.
- 

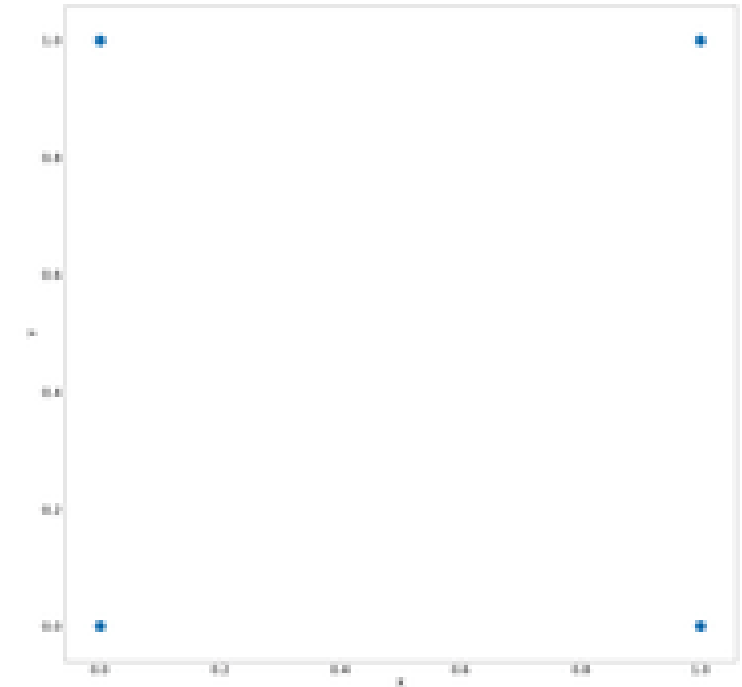


# Decision tree



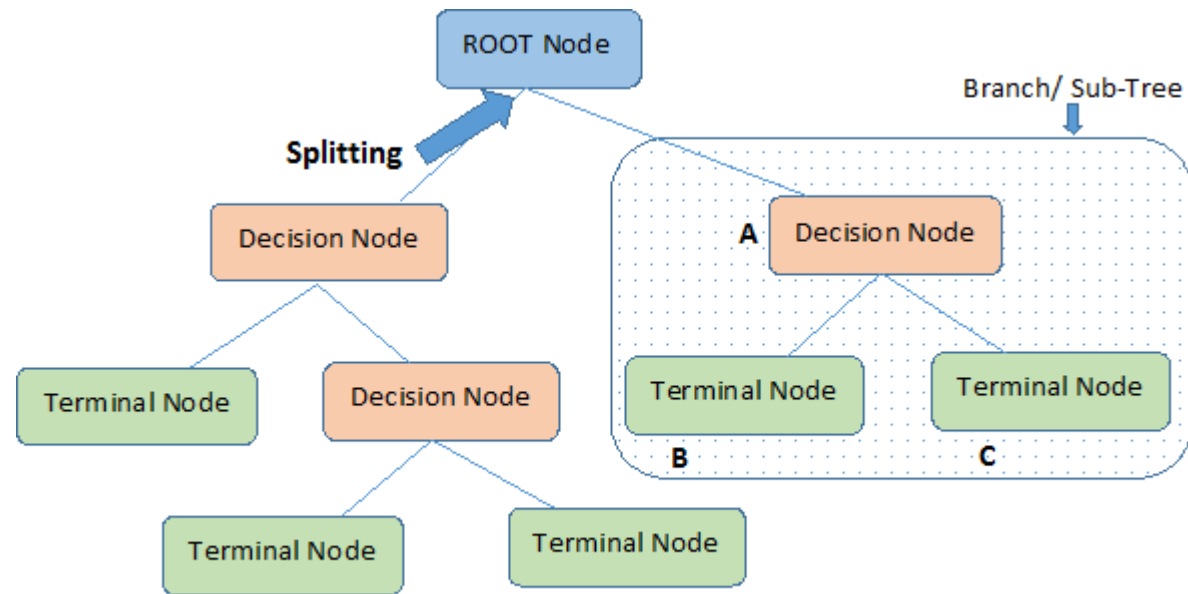
# Decision tree

| X | Y | X AND Y |
|---|---|---------|
| 0 | 0 | 0       |
| 0 | 1 | 0       |
| 1 | 0 | 0       |
| 1 | 1 | 1       |



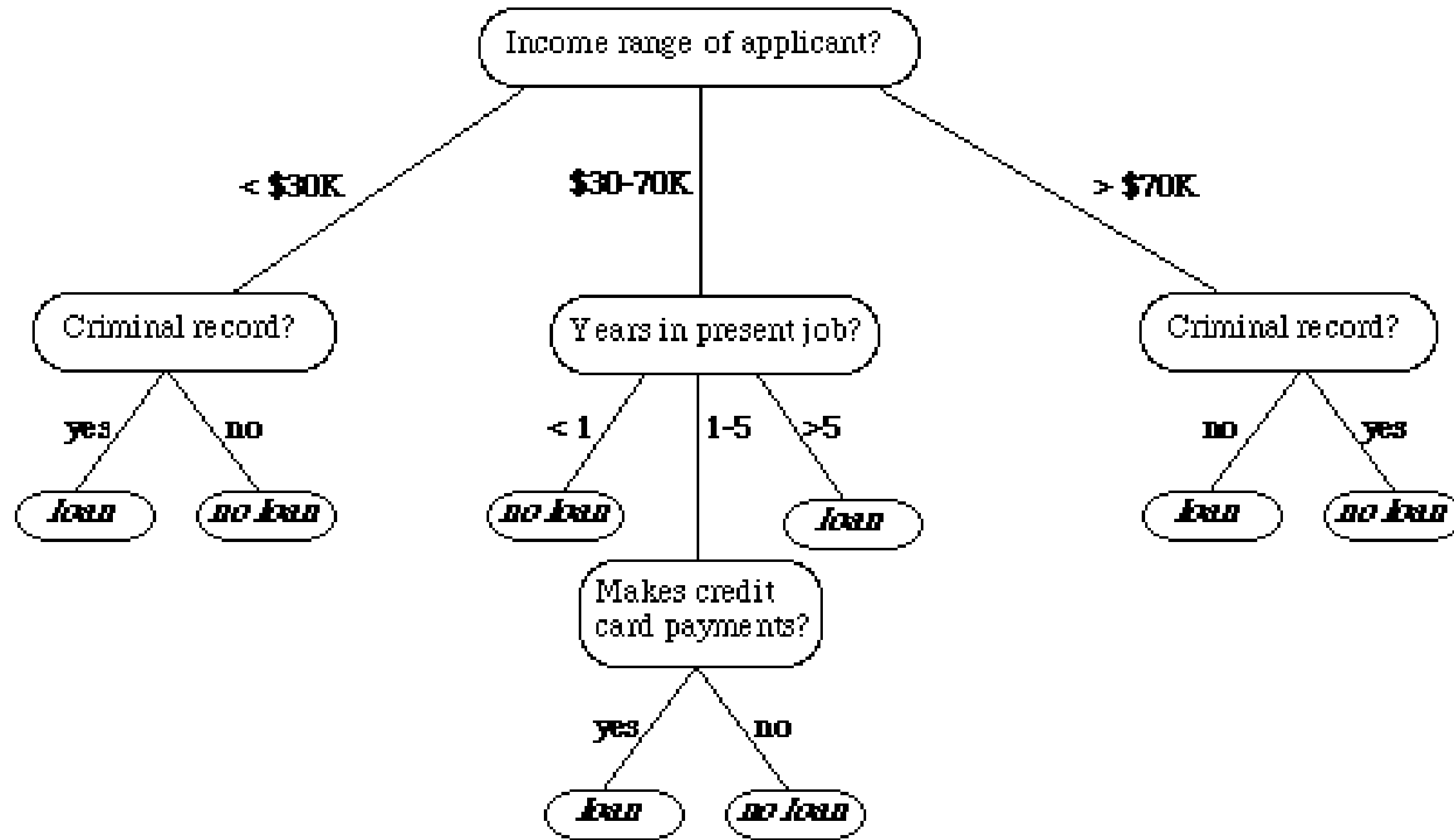
# Important Terminology

- Root Node: It represents entire population or sample
- Splitting: It is a process of dividing a node into two or more sub-nodes.
- Decision Node: When a sub-node splits into further sub-nodes, then it is called decision node.
- Leaf/ Terminal Node: Nodes do not split is called Leaf or Terminal node.

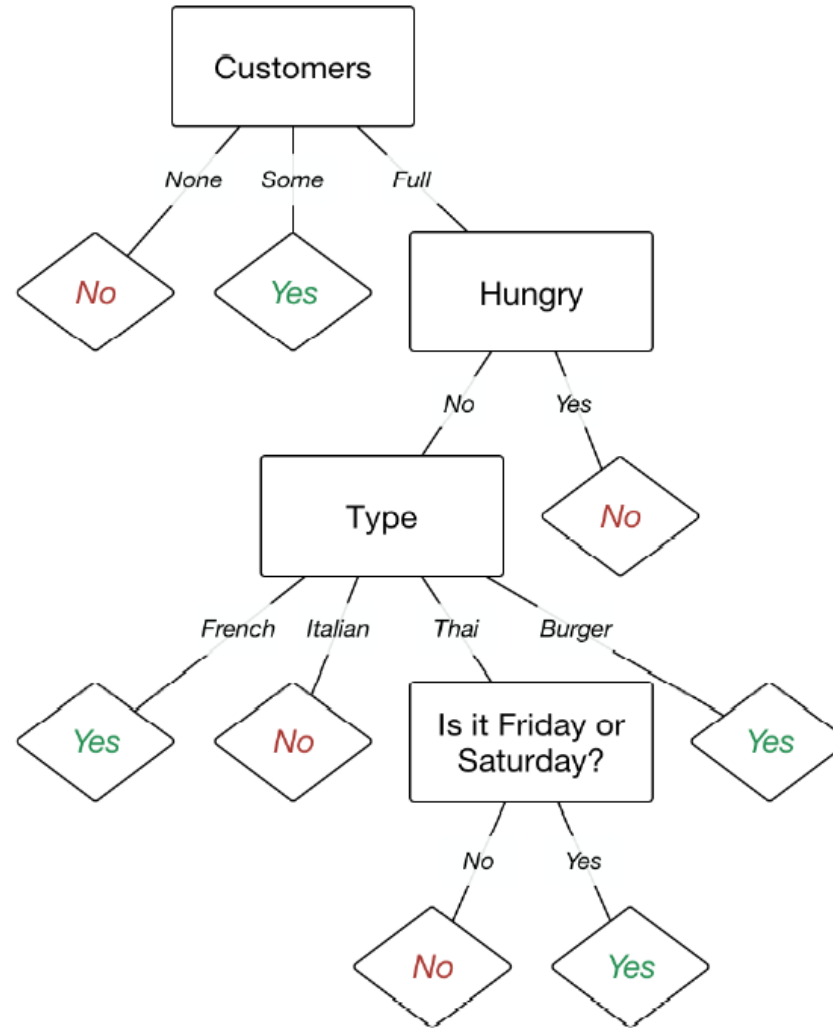


**Note:-** A is parent node of B and C.

# Example: bank loan application



# Example: Restaurent



# How it works?

- Suppose, you like to play tennis.
- On a particular day how would you decide whether or not you would play?
- Perhaps you would look outside and check to see if it's cloudy or raining.
- Maybe also see how hot (or cold) it is.
- Then, you'd use all of this information to inform your decision.
- Suppose you recorded the choices you made on different days into a table like the one in next slide:

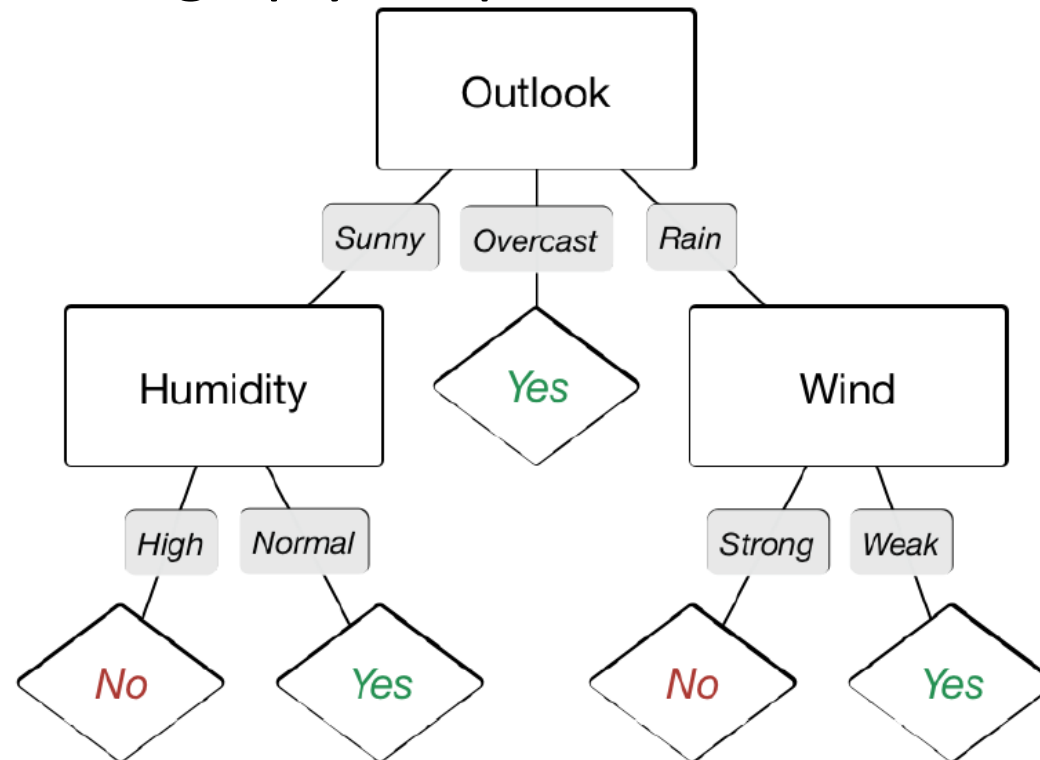


# How it works?

| Day | Outlook  | Temp. | Humidity | Wind   | Play? |
|-----|----------|-------|----------|--------|-------|
| 1   | Sunny    | Hot   | High     | Weak   | No    |
| 2   | Sunny    | Hot   | High     | Strong | No    |
| 3   | Overcast | Hot   | High     | Weak   | Yes   |
| 4   | Rain     | Mild  | High     | Weak   | Yes   |
| 5   | Rain     | Cool  | Normal   | Weak   | Yes   |
| 6   | Rain     | Cool  | Normal   | Strong | No    |
| 7   | Overcast | Cool  | Normal   | Strong | Yes   |
| 8   | Sunny    | Mild  | High     | Weak   | No    |
| 9   | Sunny    | Cool  | Normal   | Weak   | Yes   |
| 10  | Rain     | Mild  | Normal   | Weak   | Yes   |
| 11  | Sunny    | Mild  | Normal   | Strong | Yes   |
| 12  | Overcast | Mild  | High     | Strong | Yes   |
| 13  | Overcast | Hot   | Normal   | Weak   | Yes   |
| 14  | Rain     | Mild  | High     | Strong | No    |

# How it works?

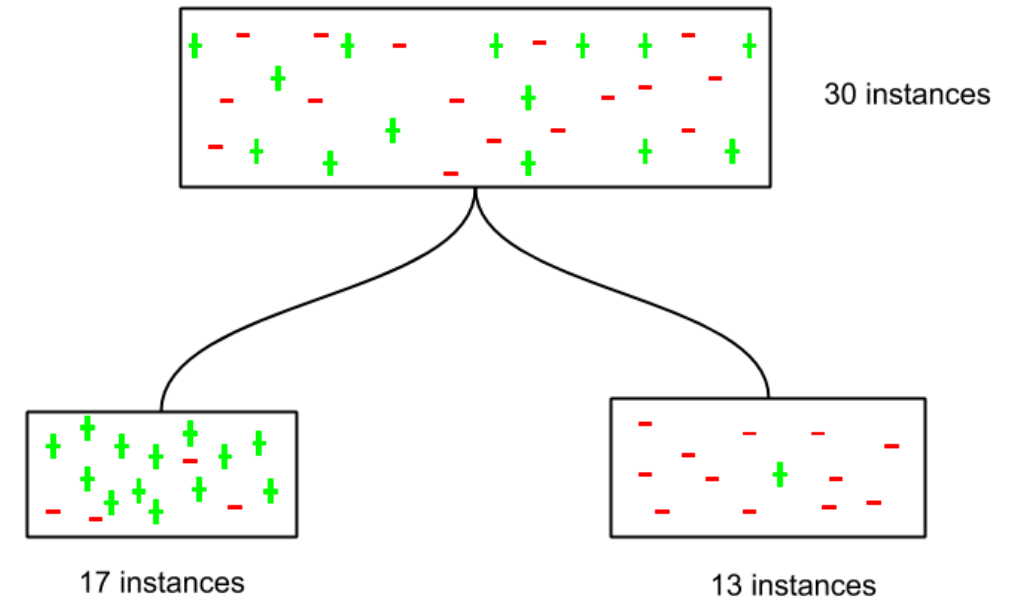
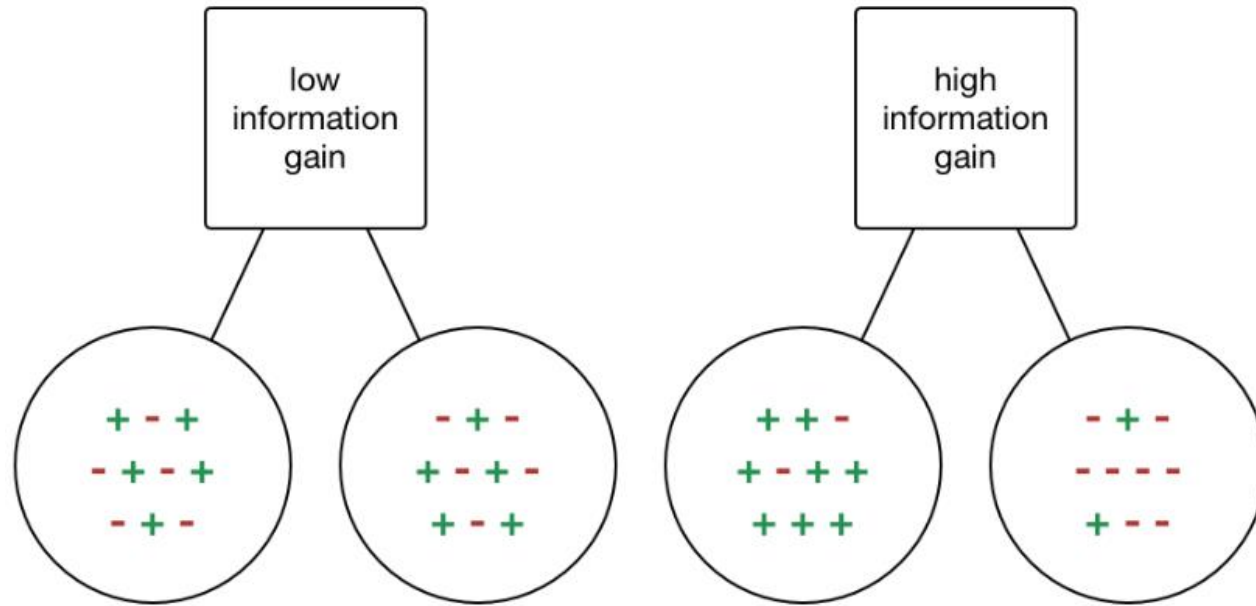
- With this table, other people would be able to use your intuition to decide whether they should play tennis.
- A decision tree would be a great way to represent data like this.
- A decision tree for this data allows you to make a decision by following a graph, rather than by looking up your particular situation in a table.



Decision tree

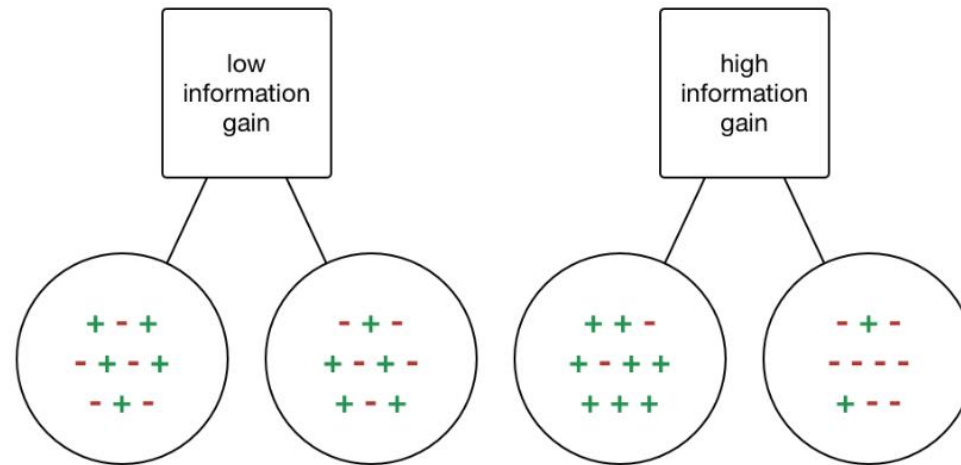
# Information Gain

- We need to split the data in table using the input variable that gives maximum information gain.



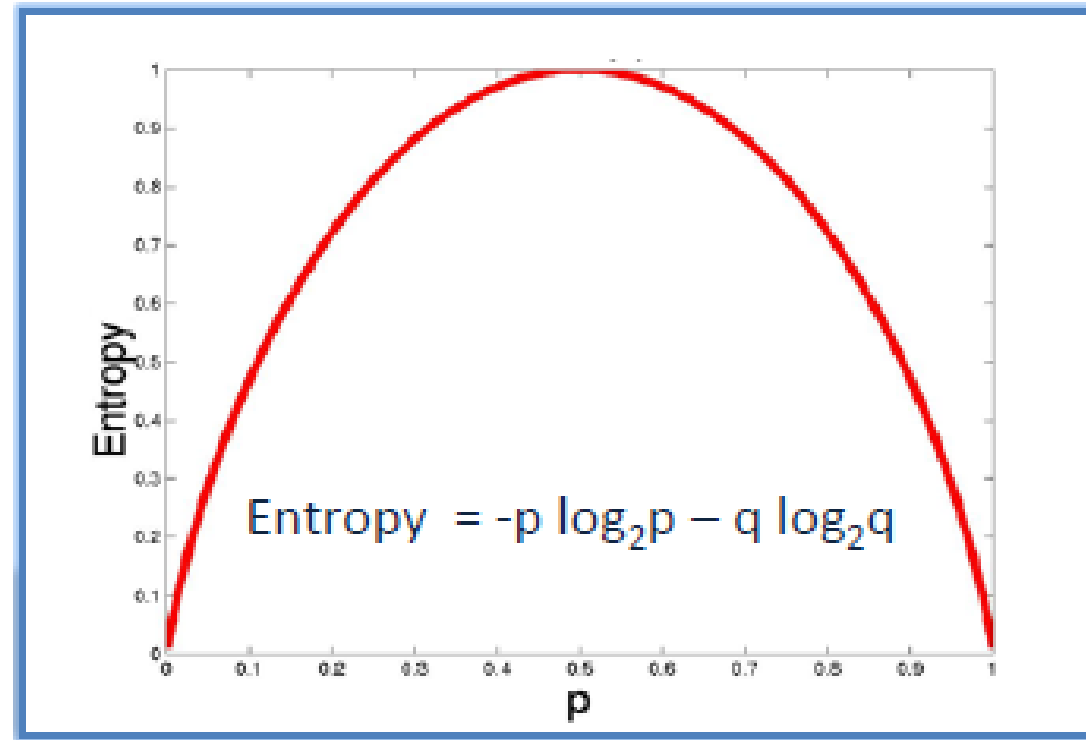
# Calculating information gain

- Information gain is a statistical property that measures how well a given attribute separates the training examples according to their target classification.
- We can see that an attribute with low information gain splits the data such that it doesn't bring us any closer to a decision.
- Whereas, an attribute with high information gain splits the data such that it helps in separating the two from each other.



# Entropy

- Entropy is the measure of homogeneity in the data. Its value is ranges from 0 to 1.



$$\text{Entropy} = -0.5 \log_2 0.5 - 0.5 \log_2 0.5 = 1$$

Thanks