**Decision Tree**

**1.Introduction:**

A decision tree is a popular machine learning algorithm used for both classification and regression tasks, based on a tree-like structure.

Visual archirtecture of decision tree:

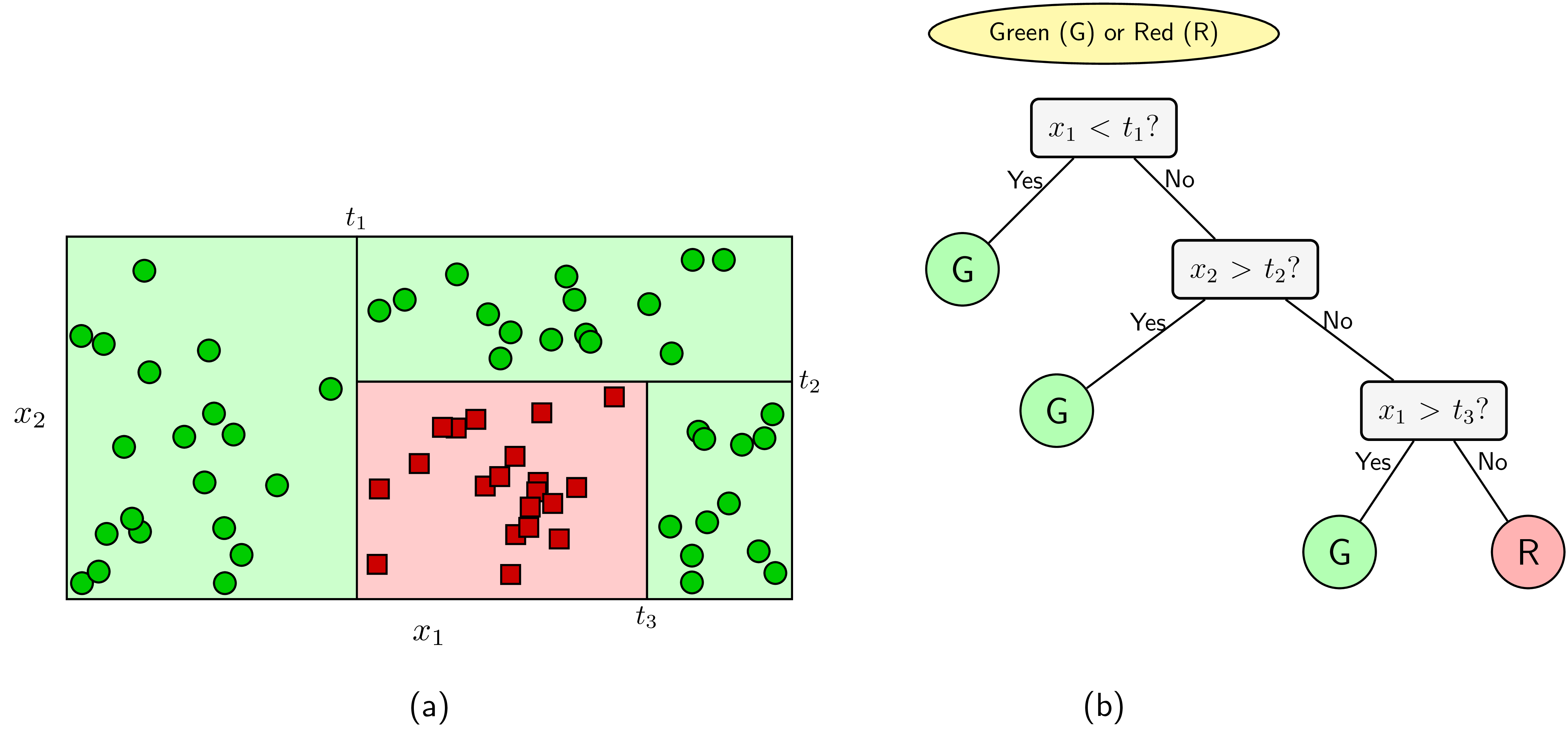


Decision node: is supposed to be feature of the data point

Branch: represent the output of decision

Leaf node: represent the final output

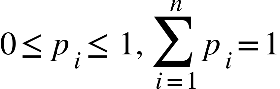
**2.Example:**

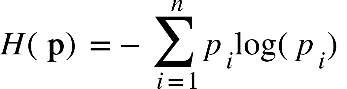


As you can see in this image. Our target is that finding the best border to classify 2 classes. The simple operation of decision tree algoritm: If data points get the value {"mathml":"<math style=\"font-family:stix;font-size:16px;\" xmlns=\"http://www.w3.org/1998/Math/MathML\"><mstyle mathsize=\"16px\"><msub><mi>x</mi><mn>1</mn></msub><mo>&lt;</mo><msub><mi>t</mi><mn>1</mn></msub></mstyle></math>","origin":"MathType for Microsoft Add-in"} , they will belong to Green class and the remaining will belong to Red class, if data points get the value {"mathml":"<math style=\"font-family:stix;font-size:16px;\" xmlns=\"http://www.w3.org/1998/Math/MathML\"><mstyle mathsize=\"16px\"><msub><mi>x</mi><mn>2</mn></msub><mo>&gt;</mo><msub><mi>t</mi><mn>2</mn></msub></mstyle></math>","origin":"MathType for Microsoft Add-in"} , they will belong to Green class and the remaining will belong to Red class,if data points get the value {"mathml":"<math style=\"font-family:stix;font-size:16px;\" xmlns=\"http://www.w3.org/1998/Math/MathML\"><mstyle mathsize=\"16px\"><msub><mi>x</mi><mn>1</mn></msub><mo>&gt;</mo><msub><mi>t</mi><mn>3</mn></msub></mstyle></math>","origin":"MathType for Microsoft Add-in"} , they will belong to Green class and the remaining will belong to Red class. We can do the decision step until finding the best border.

**3.Algorithm:**

Entropy function:

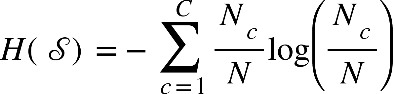
The probability distribution of a random variable {"mathml":"<math style=\"font-family:stix;font-size:16px;\" xmlns=\"http://www.w3.org/1998/Math/MathML\"><mstyle mathsize=\"16px\"><mi>x</mi></mstyle></math>","origin":"MathType for Microsoft Add-in"} can get values {"mathml":"<math style=\"font-family:stix;font-size:16px;\" xmlns=\"http://www.w3.org/1998/Math/MathML\"><mstyle mathsize=\"16px\"><msub><mi>x</mi><mn>1</mn></msub><mo>,</mo><msub><mi>x</mi><mn>2</mn></msub><mo>,</mo><mo>&#x2026;</mo><mo>,</mo><msub><mi>x</mi><mi>n</mi></msub></mstyle></math>","origin":"MathType for Microsoft Add-in"} . So the probability {"mathml":"<math style=\"font-family:stix;font-size:16px;\" xmlns=\"http://www.w3.org/1998/Math/MathML\"><mstyle mathsize=\"16px\"><mi>x</mi></mstyle></math>","origin":"MathType for Microsoft Add-in"} get the value {"mathml":"<math style=\"font-family:stix;font-size:16px;\" xmlns=\"http://www.w3.org/1998/Math/MathML\"><mstyle mathsize=\"16px\"><msub><mi>x</mi><mi>i</mi></msub></mstyle></math>","origin":"MathType for Microsoft Add-in"} is {"mathml":"<math xmlns=\"http://www.w3.org/1998/Math/MathML\" style=\"font-family:stix;font-size:16px;\"><msub><mi>p</mi><mi>i</mi></msub><mo>=</mo><mi>p</mi><mo>(</mo><mi>x</mi><mo>=</mo><msub><mi>x</mi><mi>i</mi></msub><mo>)</mo></math>","origin":"MathType for Microsoft Add-in"}.. This probability is assigned as {"mathml":"<math xmlns=\"http://www.w3.org/1998/Math/MathML\" style=\"font-family:stix;font-size:16px;\"><mi mathvariant=\"bold\">p</mi><mo>=</mo><mo>(</mo><msub><mi>p</mi><mn>1</mn></msub><mo>,</mo><msub><mi>p</mi><mn>2</mn></msub><mo>,</mo><mo>&#x2026;</mo><mo>,</mo><msub><mi>p</mi><mi>n</mi></msub><mo>)</mo></math>","origin":"MathType for Microsoft Add-in"}. So the Entropy of the distribution {"mathml":"<math style=\"font-family:stix;font-size:16px;\" xmlns=\"http://www.w3.org/1998/Math/MathML\"><mstyle mathsize=\"16px\"><mi mathvariant=\"bold\">p</mi></mstyle></math>","origin":"MathType for Microsoft Add-in"} is defined as:



ID3 algorithm:

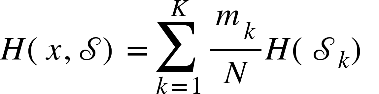
Problem with {"mathml":"<math style=\"font-family:stix;font-size:16px;\" xmlns=\"http://www.w3.org/1998/Math/MathML\"><mstyle mathsize=\"16px\"><mi>C</mi></mstyle></math>","origin":"MathType for Microsoft Add-in"} classes, suppose we are considering a decision-node that make a decision on the set {"mathml":"<math style=\"font-family:stix;font-size:16px;\" xmlns=\"http://www.w3.org/1998/Math/MathML\"><mstyle mathsize=\"16px\"><mi>S</mi></mstyle></math>","origin":"MathType for Microsoft Add-in"} of data points have the size equal {"mathml":"<math style=\"font-family:stix;font-size:16px;\" xmlns=\"http://www.w3.org/1998/Math/MathML\"><mstyle mathsize=\"16px\"><mo>|</mo><mi mathvariant=\"script\">S</mi><mo>|</mo><mo>=</mo><mi>N</mi></mstyle></math>","origin":"MathType for Microsoft Add-in"} . We also suppose that {"mathml":"<math style=\"font-family:stix;font-size:16px;\" xmlns=\"http://www.w3.org/1998/Math/MathML\"><mstyle mathsize=\"16px\"><mi>N</mi></mstyle></math>","origin":"MathType for Microsoft Add-in"} data points have {"mathml":"<math xmlns=\"http://www.w3.org/1998/Math/MathML\" style=\"font-family:stix;font-size:16px;\"><msub><mi>N</mi><mi>c</mi></msub><mo>,</mo><mi>c</mi><mo>=</mo><mn>1</mn><mo>,</mo><mn>2</mn><mo>,</mo><mo>&#x2026;</mo><mo>,</mo><mi>C</mi></math>","origin":"MathType for Microsoft Add-in"} belong to class {"mathml":"<math xmlns=\"http://www.w3.org/1998/Math/MathML\" style=\"font-family:stix;font-size:16px;\"><mi>c</mi></math>","origin":"MathType for Microsoft Add-in"}.

So the Entropy of this decision node is fomulated as:



We suppose the decision node {"mathml":"<math style=\"font-family:stix;font-size:16px;\" xmlns=\"http://www.w3.org/1998/Math/MathML\"><mstyle mathsize=\"16px\"><mi>x</mi></mstyle></math>","origin":"MathType for Microsoft Add-in"} (feature {"mathml":"<math style=\"font-family:stix;font-size:16px;\" xmlns=\"http://www.w3.org/1998/Math/MathML\"><mstyle mathsize=\"16px\"><mi>x</mi></mstyle></math>","origin":"MathType for Microsoft Add-in"}). The set {"mathml":"<math style=\"font-family:stix;font-size:16px;\" xmlns=\"http://www.w3.org/1998/Math/MathML\"><mstyle mathsize=\"16px\"><mi>S</mi></mstyle></math>","origin":"MathType for Microsoft Add-in"} of data points have {"mathml":"<math xmlns=\"http://www.w3.org/1998/Math/MathML\" style=\"font-family:stix;font-size:16px;\"><mi>K</mi></math>","origin":"MathType for Microsoft Add-in"} child node, and the number of data points of each child node are {"mathml":"<math xmlns=\"http://www.w3.org/1998/Math/MathML\" style=\"font-family:stix;font-size:16px;\"><msub><mi>m</mi><mn>1</mn></msub><mo>,</mo><msub><mi>m</mi><mn>2</mn></msub><mo>,</mo><mo>&#x2026;</mo><mo>,</mo><msub><mi>m</mi><mi>K</mi></msub></math>","origin":"MathType for Microsoft Add-in"},respectively.

So the Entropy of the decision node {"mathml":"<math xmlns=\"http://www.w3.org/1998/Math/MathML\" style=\"font-family:stix;font-size:16px;\"><mi>x</mi></math>","origin":"MathType for Microsoft Add-in"} is formulated as:



Next, we can define information gain based on feature x:

{"mathml":"<math style=\"font-family:stix;font-size:16px;\" xmlns=\"http://www.w3.org/1998/Math/MathML\"><mstyle mathsize=\"16px\"><mi>G</mi><mo>(</mo><mi>x</mi><mo>,</mo><mi mathvariant=\"script\">S</mi><mo>)</mo><mo>=</mo><mi>H</mi><mo>(</mo><mi mathvariant=\"script\">S</mi><mo>)</mo><mo>-</mo><mi>H</mi><mo>(</mo><mi>x</mi><mo>,</mo><mi mathvariant=\"script\">S</mi><mo>)</mo></mstyle></math>","origin":"MathType for Microsoft Add-in"}

Information gain is supposed to be a tool to build the structure of tree.

I seems to be theoretical, we need a practical example to understand these formulas.

We get an classification example here:

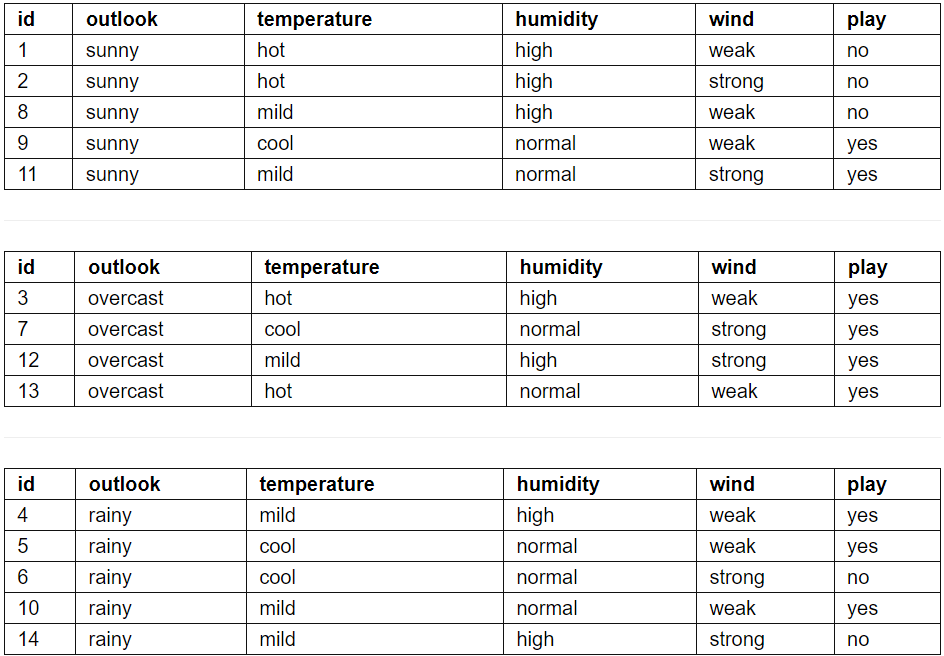
We need to predict the value of ‘play’ (yes/no) , based on 4 features ‘outlook’ , ’temperature’ , ’humidity’ ,wind’

Ảnh có chứa văn bản, số, ảnh chụp màn hình, Phông chữ

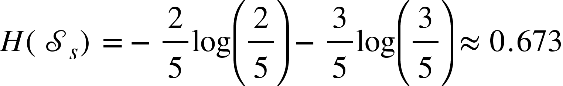
Mô tả được tạo tự động

Next, we split the table into 3 types of ‘outlook’ feature (‘sunny’ ,’overcast’, ‘rainy’):

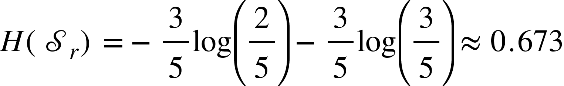
We get the set of {"mathml":"<math style=\"font-family:stix;font-size:16px;\" xmlns=\"http://www.w3.org/1998/Math/MathML\"><mstyle mathsize=\"16px\"><msub><mi>S</mi><mi>s</mi></msub><mo>,</mo><msub><mi>S</mi><mi>o</mi></msub><mo>,</mo><msub><mi>S</mi><mi>r</mi></msub></mstyle></math>","origin":"MathType for Microsoft Add-in"} represent 3 ‘child node’ of ‘outlook’ feature. {"mathml":"<math style=\"font-family:stix;font-size:16px;\" xmlns=\"http://www.w3.org/1998/Math/MathML\"><mstyle mathsize=\"16px\"><msub><mi>m</mi><mi>s</mi></msub><mo>,</mo><msub><mi>m</mi><mi>o</mi></msub><mo>,</mo><msub><mi>m</mi><mi>r</mi></msub></mstyle></math>","origin":"MathType for Microsoft Add-in"} are the number of sunny, overcast, rainy , respectively.

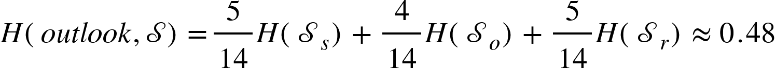


We can easily compute the Entropy of outook:



{"mathml":"<math xmlns=\"http://www.w3.org/1998/Math/MathML\" style=\"font-family:stix;font-size:16px;\"><mi>H</mi><mo>(</mo><msub><mi mathvariant=\"script\">S</mi><mi>o</mi></msub><mo>)</mo><mo>=</mo><mn>0</mn></math>","origin":"MathType for Microsoft Add-in"}



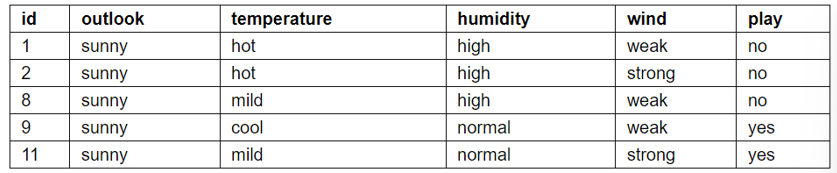


We also repeatedly compute Entropy of temperature, humidity, wind:

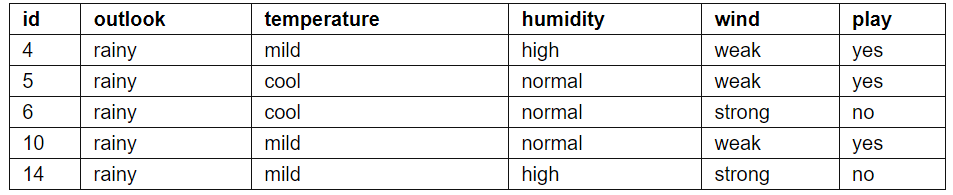
{"mathml":"<math style=\"font-family:stix;font-size:16px;\" xmlns=\"http://www.w3.org/1998/Math/MathML\"><mstyle mathsize=\"16px\"><mi>H</mi><mo>(</mo><mi>t</mi><mi>e</mi><mi>m</mi><mi>p</mi><mi>e</mi><mi>r</mi><mi>a</mi><mi>t</mi><mi>u</mi><mi>r</mi><mi>e</mi><mo>,</mo><mi mathvariant=\"script\">S</mi><mo>)</mo><mo>&#x2248;</mo><mn>0</mn><mo>.</mo><mn>631</mn><mo>&#xA0;</mo><mo>,</mo><mo>&#xA0;</mo><mi>H</mi><mo>(</mo><mi>h</mi><mi>u</mi><mi>m</mi><mi>i</mi><mi>d</mi><mi>i</mi><mi>t</mi><mi>y</mi><mo>,</mo><mi mathvariant=\"script\">S</mi><mo>)</mo><mo>&#x2248;</mo><mn>0</mn><mo>.</mo><mn>547</mn><mo>,</mo><mo>&#xA0;&#xA0;</mo><mi>H</mi><mo>(</mo><mi>w</mi><mi>i</mi><mi>n</mi><mi>d</mi><mo>,</mo><mi mathvariant=\"script\">S</mi><mo>)</mo><mo>&#x2248;</mo><mn>0</mn><mo>.</mo><mn>618</mn></mstyle></math>","origin":"MathType for Microsoft Add-in"}

We realized that the Entropy of outlook is the highest, so we choose ‘outlook’ feature as the first decision node or root node.

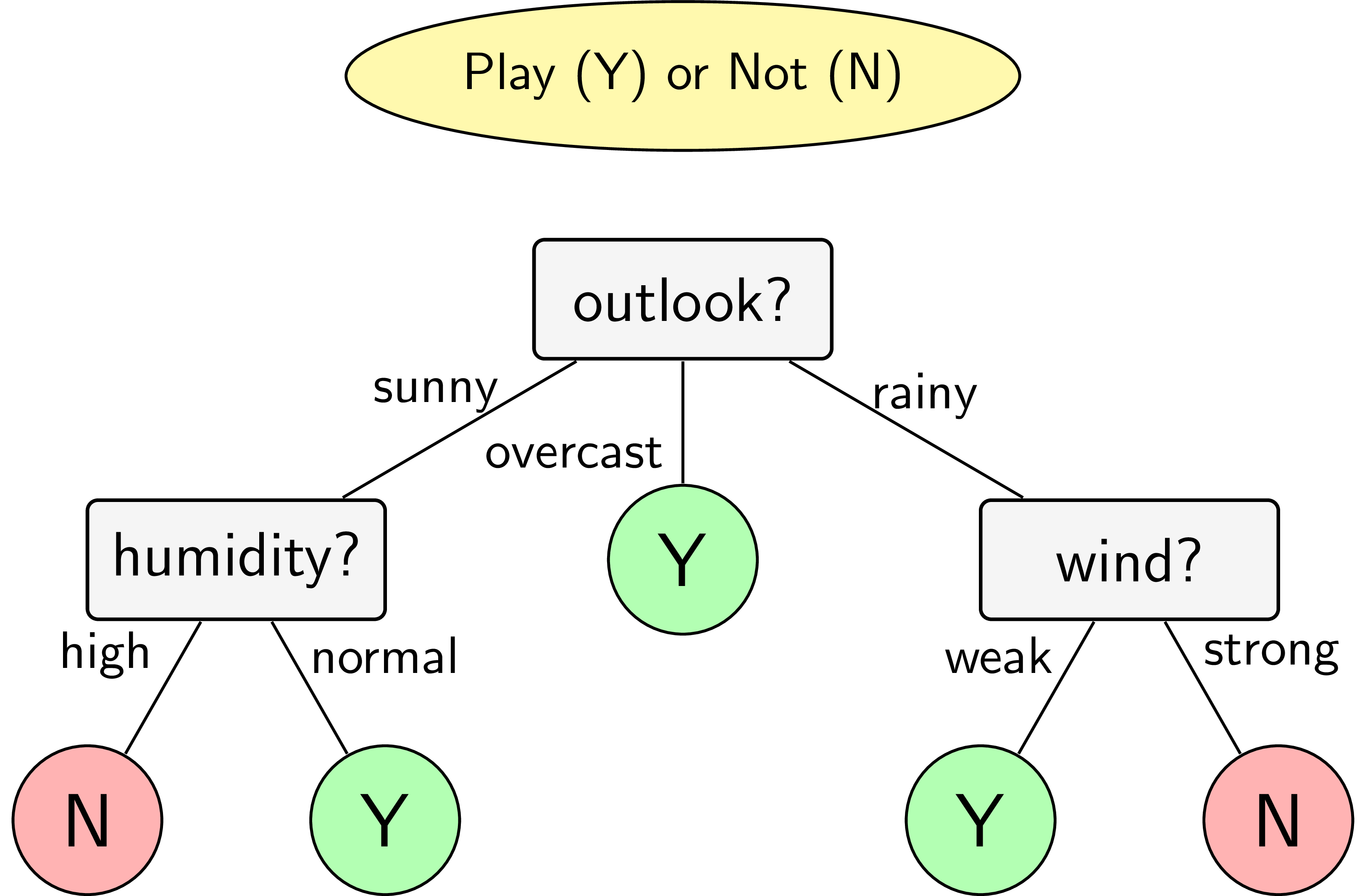
In ‘sunny’ branch, the ‘play’ is no only when humidity is high , the ‘play’ is yes only when humidity is normal. The ‘humidity’ feature will be a decision node of ‘sunny’ branch.



We also get ‘wind’ feature be decision node of ‘rainy’ branch based on below table:



Finally, we get the decision tree for this problem here:



General steps of Decision Tree algorithm:

Step-1: Begin the tree with the root node, says S, which contains the complete dataset.

Step-2: Find the best attribute in the dataset

Step-3: Divide the S into subsets that contains possible values for the best attributes.

Step-4: Generate the decision tree node, which contains the best attribute.

Step-5: Recursively make new decision trees using the subsets of the dataset created in step -3. Continue this process until a stage is reached where you cannot further classify the nodes and called the final node as a leaf node.

If we repeatly do decision step (step-3) on an excessive way. It can lead to overfitting. So it’s necessary for us to propose a condition to stop the decision step below:

* If a node have entropy equal 0, it means all data points belong to one class
* If a node have the number of data points is smaller than a fixed threshold
* Limit the tree depth that can reduce the complexity of tree
* If leaf node is higher than a fixed threshold
* If the decision step don’t make the entropy substantly decreased , it means the information gain is smaller than a fixed threshold.

**4. Advantages and Disadvantages:**

Advantages:

* It is simple to understand as it follows the same process which a human follow while making any decision in real-life.
* It can be very useful for solving decision-related problems.
* It helps to think about all the possible outcomes for a problem.
* There is less requirement of data cleaning compared to other algorithms.

Disadvantages:

* The decision tree contains lots of layers, which makes it complex.
* It may have an overfitting issue, which can be resolved using the Random Forest algorithm.
* For more class labels, the computational complexity of the decision tree may increase.