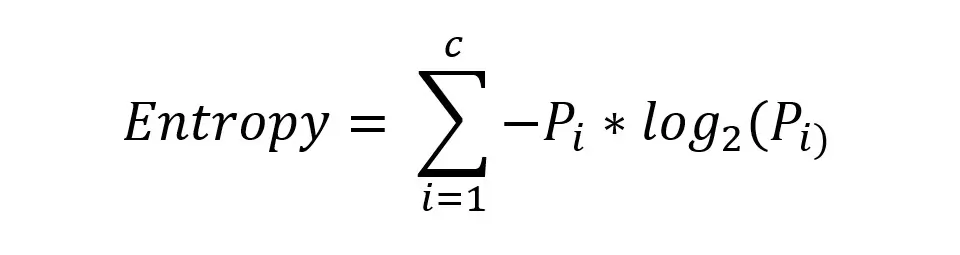
* **Types for Decision tree classifier:**

1. ID3

2. CART

* **To check a Pure split (leaf node):**



1.Entropy:

Min: 0, Max: 1

0.2



2. Geni Index/Impurity:

Min: 0, Max: 0.5

* **Information Gain (how the features are selected):**

Information Gain= Entropy (Parent)−Weighted Average Entropy (Children)

Entropy

Entropy is a measure of impurity or disorder in a set of data. In the context of decision trees, it is used to determine how a dataset can be split to create the most homogeneous branches.

Formula:

\[ \text{Entropy}(S) = -\sum\_{i=1}^{c} p\_i \log\_2(p\_i) \]

where \( p\_i \) is the proportion of the dataset belonging to class \( i \), and \( c \) is the number of classes.

Gini Index

The Gini Index, or Gini impurity, is another measure of impurity used in decision trees. It represents the probability of a randomly chosen element being incorrectly classified if it was randomly labeled according to the distribution of labels in the dataset.

Formula:

\[ \text{Gini}(S) = 1 - \sum\_{i=1}^{c} p\_i^2 \]

where \( p\_i \) is the proportion of the dataset belonging to class \( i \).

Information Gain

Information Gain measures the reduction in entropy or impurity from a dataset after a dataset is split on an attribute. It is used to decide which attribute to split the dataset on at each step in the tree.

Formula:

\[ \text{Information Gain}(S, A) = \text{Entropy}(S) - \sum\_{v \in \text{Values}(A)} \frac{|S\_v|}{|S|} \text{Entropy}(S\_v) \]

where:

- \( S \) is the original dataset.

- \( A \) is the attribute being split on.

- \( S\_v \) is the subset of \( S \) for which attribute \( A \) has value \( v \).

- \( |S\_v| \) is the number of elements in \( S\_v \).

- \( |S| \) is the number of elements in \( S \).

Key Points:

- Entropy and Gini Index are measures of impurity or disorder in a dataset.

- Information Gain quantifies the reduction in impurity by splitting the dataset based on an attribute.

- Decision Trees use these measures to decide the best attribute for splitting the dataset to form the most homogeneous branches.