

*Where is the Gini Index used ?*

Before diving into Gini Index it is essential to understand the area of Decision Tree, which is among the commonly used supervised machine learning algorithm for its simplicity in understanding. This algorithm supports predicting both classification and regression problems and Gini Index is used by Classification and Regression Tree (CART), one of the variants of Decision Tree algorithms.

*Why*  is Gini Index used?

Decision Tree supports making decisions by splitting the nodes of the tree into Root node, Decision node and Leaf node. To identify the best split the metric of Gini Index is used.

*How*  is Gini Index addressed?

Using the parameter criterion ‘Gini‘ should be passed as input. However in scikit-learn’s class constructors, Gini is passed as default value compared to other metrics such as ‘Entropy’.

*History*  of Gini Index?

Gini impurity is named after Italian mathematician [Corrado Gini](https://en.wikipedia.org/wiki/Corrado_Gini" \o "Corrado Gini), is a measure of how often a randomly chosen element from the set would be incorrectly labelled if it was randomly labelled according to the distribution of labels in the subset.

*Synonym*  Is Gini Index or Gini Impurity same?

It is to be noted that Gini Index and Gini Impurity are interchangeably used meaning the same.

Now coming to the most important question,

*What*  is Gini Index?

The metric used to perform the splitting of node when the target variable is *categorical*, is called as Gini Index. In Simple terms, if all the elements are linked with a single class then it is called pure.

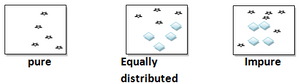
It ranges from 0-1

0 = all elements

1 = Randomly distributed

0.5 = equally distributed

It means an attribute with a lower Gini index should be preferred.



The formula is,



Gini is the probability of correctly labelling a randomly chosen element if it was randomly labelled according to the distribution of labels in the node.

The formula for Gini is:



Hence Gini Index is



Attribute with the lowest Gini score is the used for ROOT node. Gini Index is used by default and it is the preferred as it does not involve more computational intensive log mechanism

*Conclusion*

            Lower the Gini Impurity, higher is the homogeneity of the node. The Gini Impurity of a pure node is zero.Attribute with the lowest Gini score is the used for ROOT node. Gini Index is used by default and it is the preferred as it does not involve more computational intensive log mechanism.