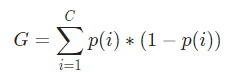
# Decision Tree

Decision Tree is one of the most commonly used, practical approaches for supervised learning. It can be used to solve both Regression and Classification tasks with the latter being put more into practical application.

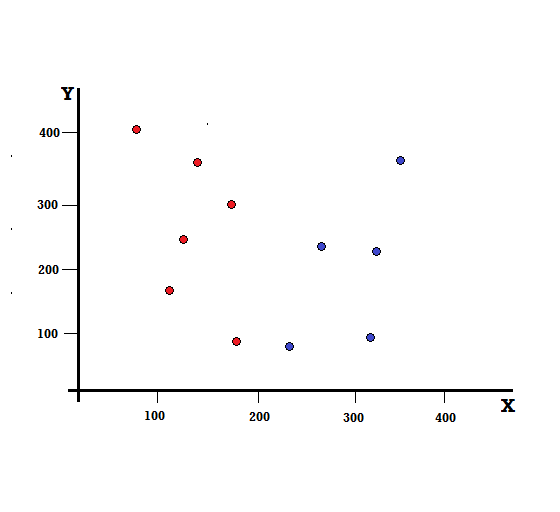
The regression tree is used when the predicted outcome is a real number and the classification tree is used to predict the class to which the data belongs. These two terms are collectively called as Classification and Regression Trees (CART).

The Gini Index or Gini Impurity is calculated by subtracting the sum of the squared probabilities of each class from one. It favours mostly the larger partitions and are very simple to implement. In simple terms, it calculates the probability of a certain randomly selected feature that was classified incorrectly.

Mathematically, The Gini Index is represented by



Let us understand the calculation of the Gini Index with a simple example. In this, we have a total of 10 data points with two variables, the reds and the blues. The X and Y axes are numbered with spaces of 100 between each term. From the given example, we shall calculate the Gini Index and the Gini Gain.



In the above example, we have C=2 and p(1) = p(2) = 0.5, Hence the Gini Index can be calculated as,

*G =p(1) \* (1−p(1)) + p(2) \* (1−p(2))*

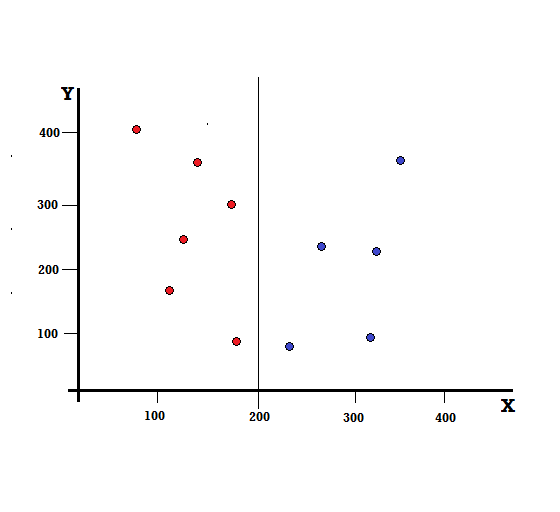
*=0.5 \* (1−0.5) + 0.5 \* (1−0.5)*

*=0.5*

Where 0.5 is the total probability of classifying a data point imperfectly and hence is exactly 50%.

Now, let us calculate the Gini Impurity for both the perfect and imperfect split that we performed earlier,

## Perfect Split



The left branch has only reds and hence its Gini Impurity is,

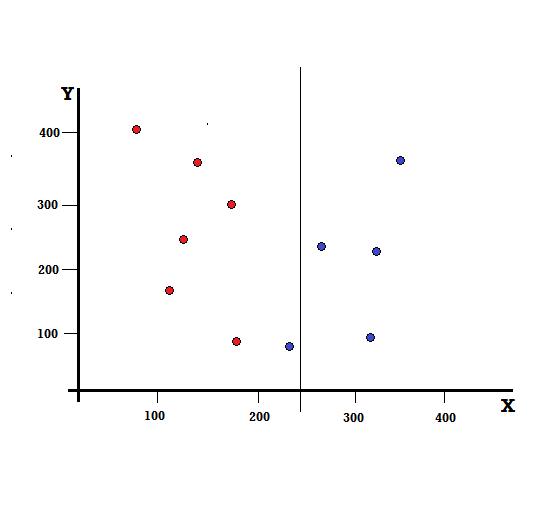
*G(left) =1 \* (1−1) + 0 \* (1−0) = 0*

The right branch also has only blues and hence its Gini Impurity is also given by,

*G(right) =1 \* (1−1) + 0 \* (1−0) = 0*

From the quick calculation, we see that both the left and right branches of our perfect split have probabilities of 0 and hence is indeed perfect. A Gini Impurity of 0 is the lowest and the best possible impurity for any data set.

## Imperfect Split



In this case, the left branch has 5 reds and 1 blue. Its Gini Impurity can be given by,

*G(left) =1/6 \* (1−1/6) + 5/6 \* (1−5/6) = 0.278*

The right branch has all blues and hence as calculated above its Gini Impurity is given by,

*G(right) =1 \* (1−1) + 0 \* (1−0) = 0*

Now that we have the Gini Impurities of the imperfect split, in order to evaluate the quality or extent of the split, we will give a specific weight to the impurity of each branch with the number of elements it has.

*(0.6\*0.278) + (0.4\*0) = 0.167*

Now that we have calculated the Gini Index, we shall calculate the value of another parameter, Gini Gain and analyse its application in Decision Trees. The amount of impurity removed with this split is calculated by deducting the above value with the Gini Index for the entire dataset (0.5)

*0.5 – 0.167 = 0.333*

This value calculated is called as the “**Gini Gain**”. In simple terms, **Higher Gini Gain = Better Split**.

Hence, in a Decision Tree algorithm, the best split is obtained by maximizing the Gini Gain, which is calculated in the above manner with each iteration.

After calculating the Gini Gain for each attribute in the data set, the class, sklearn.tree.DecisionTreeClassifier will choose the largest Gini Gain as the Root Node. When a branch with Gini of 0 is encountered it becomes the leaf node and the other branches with Gini more than 0 need further splitting. These nodes are grown recursively till all of them are classified.

## Pruning

The performance of a tree can be further increased by **pruning**. It involves **removing the branches that make use of features having low importance**. This way, we reduce the complexity of tree, and thus increasing its predictive power by reducing overfitting.

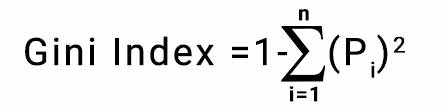
## Advantages of CART

* Simple to understand, interpret, visualize.
* Decision trees implicitly perform variable screening or feature selection.
* Can handle both numerical and categorical data. Can also handle multi-output problems.
* Decision trees require relatively little effort from users for data preparation.
* Nonlinear relationships between parameters do not affect tree performance.

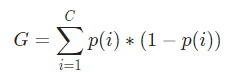
## **Disadvantages of CART**

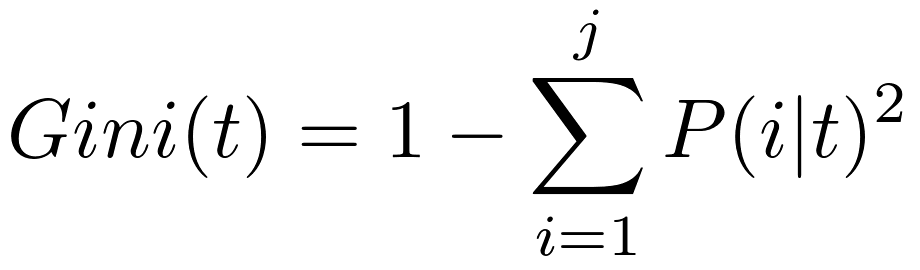
* Decision-tree learners can create over-complex trees that do not generalize the data well. This is called overfitting.
* Decision trees can be unstable because small variations in the data might result in a completely different tree being generated. This is called [variance](https://medium.com/towards-data-science/balancing-bias-and-variance-to-control-errors-in-machine-learning-16ced95724db), which needs to be lowered by methods like bagging and [**boosting**](https://towardsdatascience.com/boosting-the-accuracy-of-your-machine-learning-models-f878d6a2d185).
* Greedy algorithms cannot guarantee to return the globally optimal decision tree. This can be mitigated by training multiple trees, where the features and samples are randomly sampled with replacement.
* Decision tree learners create [biased](https://medium.com/towards-data-science/balancing-bias-and-variance-to-control-errors-in-machine-learning-16ced95724db) trees if some classes dominate. It is therefore recommended to balance the data set prior to fitting with the decision tree.

The Gini Index is determined by deducting the sum of squared of probabilities of each class from one, mathematically, Gini Index can be expressed as:



Are they same mathematically ?





## Gini Index vs Information Gain

Take a look below for the getting discrepancy between Gini Index and Information Gain,

1. The Gini Index **facilitates the bigger distributions** so easy to implement whereas the Information Gain **favors lesser distributions** having small count with multiple specific values.
2. The method of the Gini Index is**used by CART algorithms**, in contrast to it, Information Gain is**used in**[**ID3, C4.5 algorithms**](https://medium.com/datadriveninvestor/tree-algorithms-id3-c4-5-c5-0-and-cart-413387342164).
3. Gini index **operates on the categorical target variables** in terms of “success” or “failure” and **performs only binary split**, in opposite to that Information Gain **computes the difference between entropy** **before and after the split** and indicates the impurity in classes of elements.