What is Entropy?

- Entropy is the degree of uncertainty, impurity or disorder of a random variable, or a measure of purity. It characterizes the impurity of an arbitrary class of examples.

- *Entropy is the measurement of impurities or randomness in the data points.*

- Here, if all elements belong to a single class, then it is termed as “Pure”, and if not then the distribution is named as “Impurity”.

- **It is computed between 0 and 1**, however, heavily relying on the number of groups or classes present in the data set it can be more than 1 while depicting the same significance i.e. extreme level of disorder.

- In more simple terms, if a dataset contains homogeneous subsets of observations, then no impurity or randomness is there in the dataset, and **if all the observations belong to one class, the entropy of that dataset becomes zero.**

What is Information Gain?

- The concept of entropy plays an important role in measuring the information gain. However, “Information gain is based on the information theory”.

- Information gain is used for determining the best features/attributes that render maximum information about a class. It follows the concept of entropy while aiming at decreasing the level of entropy, beginning from the root node to the leaf nodes.

- Information gain computes the difference between entropy before and after split and specifies the impurity in class elements.

Information Gain = Entropy before splitting - Entropy after splitting

Given a probability distribution such that

P = (p1 , p2 ,.....pn),

and where (p_i) is the probability of a data point in the subset of Di of a dataset D,

Therefore, Entropy is defined as the

$$Entropy\ (P) = \ -\sum_{i=1}^{n} p_i \log_2 (p_i)$$

Generally, it is not preferred as it involves ‘log’ function that results in the computational complexity. Moreover;

1. Information gain is non-negative.
2. Information Gain is symmetric such that switching of the split variable and target variable, the same amount of information gain is obtained. (Source)
3. Information gain determines the reduction of the uncertainty after splitting the dataset on a particular feature such that if the value of information gain increases, that feature is most useful for classification.
4. The feature having the highest value of information gain is accounted for as the best feature to be chosen for split.

What is Gain Ratio?

Proposed by **John Ross Quinlan**, Gain Ratio or Uncertainty Coefficient is used to normalize the information gain of an attribute against how much entropy that attribute has. Formula of gini ratio is given by

Gain Ratio=Information Gain/Entropy

From the above formula, it can be stated that if entropy is very small, then the gain ratio will be high and vice versa.
Be selected as splitting criterion, Quinlan proposed following procedure,

1. First, determine the information gain of all the attributes, and then compute the average information gain.
2. Second, calculate the gain ratio of all the attributes whose calculated information gain is larger or equal to the computed average information gain, and then pick the attribute of higher gain ratio to split.

What is Gini Index?

- The gini index, or gini coefficient, or gini impurity computes the degree of probability of a specific variable that is wrongly being classified when chosen randomly and a variation of gini coefficient.

- It works on categorical variables, provides outcomes either be "successful" or "failure" and hence conducts binary splitting only.

- **The degree of gini index varies from 0 to 1,**

- **Where 0 depicts that all the elements be allied to a certain class, or only one class exists there.**
- **The gini index of value as 1 signifies that all the elements are randomly zdistributed across various classes, and**
- **A value of 0.5 denotes the elements are uniformly distributed into some classes.**

- It was proposed by Leo Breiman in 1984 as an impurity measure for decision tree learning and is given by the equation/formula;

$$Gini\,(P) = \sum_{i=1}^{n} p_i\,(1 - p_i\,) = 1 - \sum_{i=1}^{n} (p_i)^2$$

- where $P=(p_1 , p_2 ,.....p_n)$, and p_i is the probability of an object that is being classified to a particular class.

- Also, an attribute/feature with least gini index is preferred as root node while making a decision tree.

Calculation

- 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.

- The Gini Index varies between 0 and 1, where 0 represents purity of the classification and 1 denotes random distribution of elements among various classes. A Gini Index of 0.5 shows that there is equal distribution of elements across some classes.

Mathematically, The Gini Index is represented by

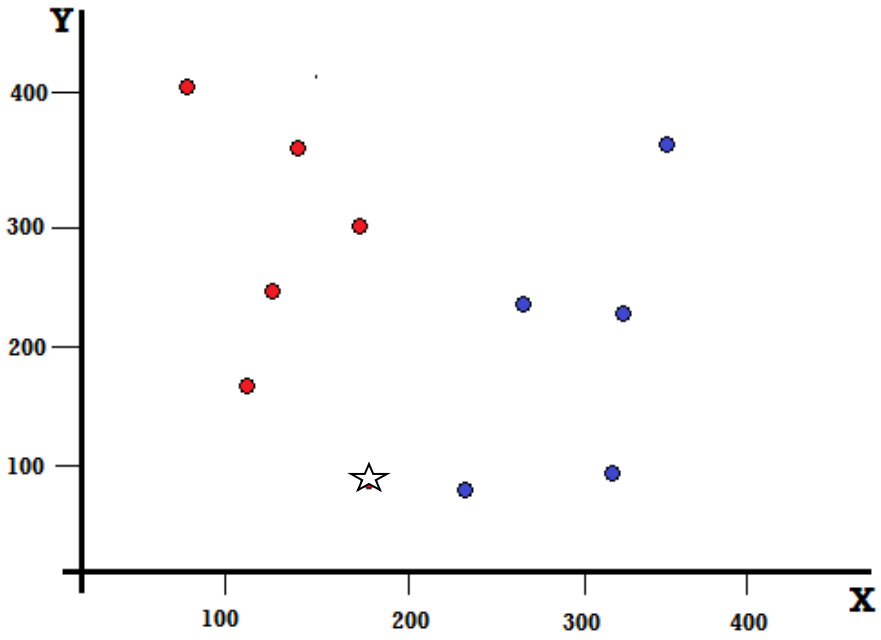
$$G = \sum_{i=1}^{C} p(i) * (1 - p(i))$$

- The Gini Index works on categorical variables and gives the results in terms of “success” or “failure” and hence performs only binary split.

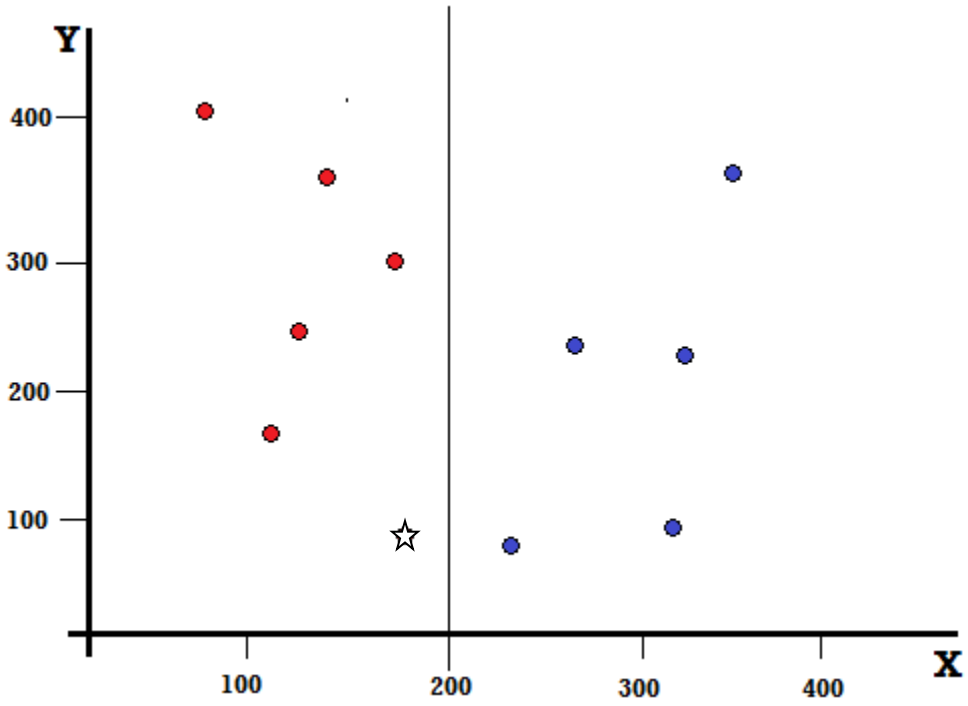
- From the Gini Index, the value of another parameter named Gini Gain is calculated whose value is maximised with each iteration by the Decision Tree to get the perfect CART(Classification and regression tree)

Example

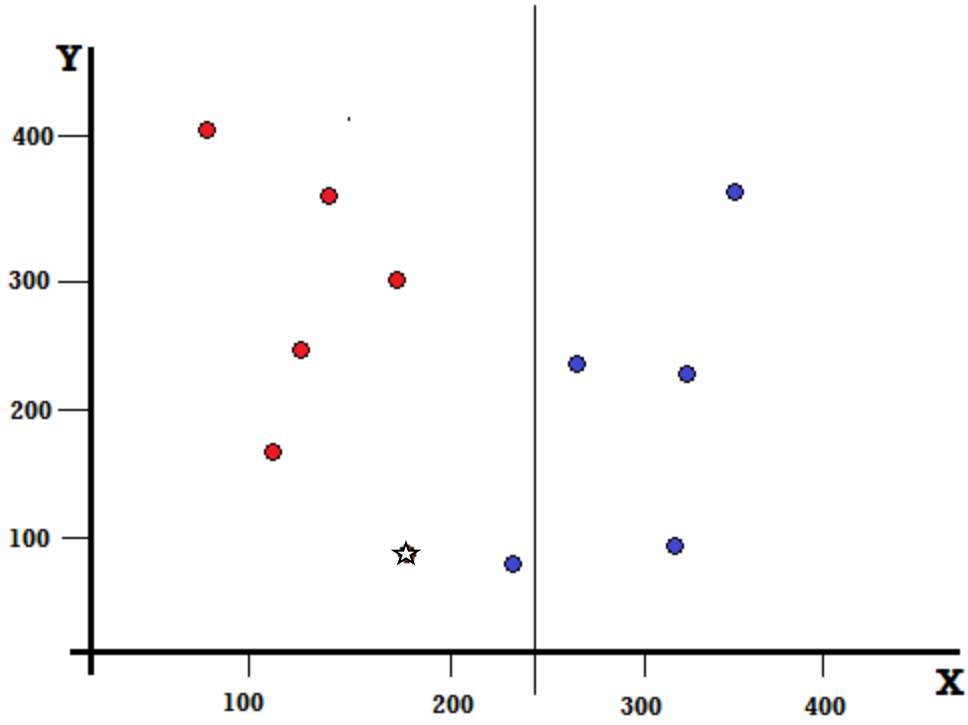
- Let us understand the calculation of the Gini Index with a simple example. In this, we have a total of 10 data points with two variables, the reds and the blues. The X and Y axes are numbered with spaces of 100 between each term. From the given example, we shall calculate the Gini Index and the Gini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



- For a decision tree, we need to split the dataset into two branches. Consider the following data points with 5 Reds and 5 Blues marked on the X-Y plane. Suppose we make a binary split at X=200, then we will have a perfect split as shown below.



- It is seen that the split is correctly performed and we are left with two branches each with 5 reds (left branch) and 5 blues (right branch).But what will be the outcome if we make the split at X=250?



We are left with two branches, the left branch consisting of 5 reds and 1 blue, while the right branch consists of 4 blues. The following is referred to as an imperfect split. In training the Decision Tree model, to quantify the amount of imperfectness of the split, we can use the Gini Index.

Basic Mechanism

- To calculate the Gini Impurity, let us first understand it's basic mechanism.

  - First, we shall randomly pick up any data point from the dataset
  - Then, we will classify it randomly according to the class distribution in the given dataset. In our dataset, we shall give a data point chosen with a probability of 5/10 for red and 5/10 for blue as there are five data points of each colour and hence the probability.

- Now, in order to calculate the Gini Index, the formula is given by

$$G = \sum_{i=1}^{C} p(i) * (1 - p(i))$$

Where, C is the total number of classes and p(*i*) is the probability of picking the data point with the class *i.*

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%.

Steps to calculate the highest information gain on a data set

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Entropy of the whole data set

14 records, 9 are “yes”

$$-\left(\frac{9}{14}\log_2\frac{9}{14}+\frac{5}{14}\log_2\frac{5}{14}\right)\approx 0.94$$

Expected new entropy for each attribute

outlook

The outlook attribute contains 3 distinct values:

- overcast: 4 records, 4 are “yes”

$$-\left(\frac{4}{4}\log_2\frac{4}{4}\right)=0$$

- rainy: 5 records, 3 are “yes”

$$-\left(\frac{3}{5}\log_2\frac{3}{5}+\frac{2}{5}\log_2\frac{2}{5}\right)\approx 0.97$$

- sunny: 5 records, 2 are “yes”

$$-\left(\frac{2}{5}\log_2\frac{2}{5}+\frac{3}{5}\log_2\frac{3}{5}\right)\approx 0.97$$

Expected new entropy:

$$-\left(\frac{4}{14}\times 0+\frac{5}{14}\times 0.97+\frac{5}{14}\times 0.97\right)\approx 0.69$$

temperature

Distinct value	**Yes records**	**Entropy**
cool	4 records, 3 are “yes”	0.81
hot	4 records, 2 are “yes”	1.0
mild	6 records, 4 are “yes”	0.92

Expected new entropy:

$$-\left(\frac{4}{14}\times 0.81+\frac{4}{14}\times 1.0+\frac{6}{14}\times 0.92\right)\approx 0.91$$

humidity

Discrete values

Distinct value	Yes records	Entropy
normal	7 records, 6 are “yes”	0.59
high	7 records, 2 are “yes”	0.86

Expected new entropy:

$$-\left(\frac{7}{14}\times 0.81+\frac{7}{14}\times 0.86\right)\approx 0.72$$

Continues values

Consider every possible binary partition; choose the partition with the highest gain

windy

Distinct value	Yes records	Entropy
weak	8 records, 6 are “yes”	0.81
strong	6 records, 3 are “yes”	0.97

Expected new entropy:

$$-\left(\frac{8}{14}\times 0.81+\frac{6}{14}\times 0.97\right)\approx 0.87$$

Gain

Attribute	Information Gain
outlook	0.94 - 0.69 = 0.25
temperature	0.94 - 0.91 = 0.03
humidity	0.94 - 0.72 = 0.22
windy	0.94 - 0.87 = 0.07

The highest information gain is with the outlook attribute