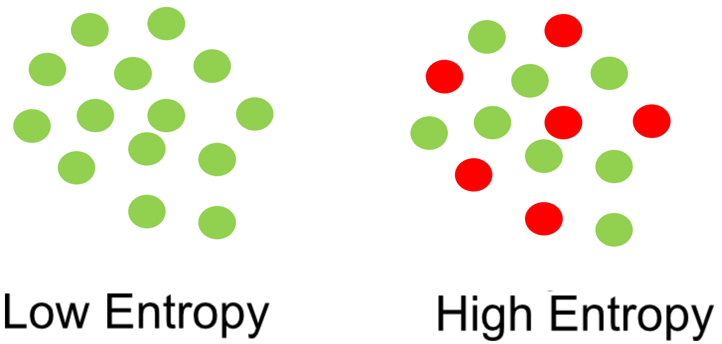
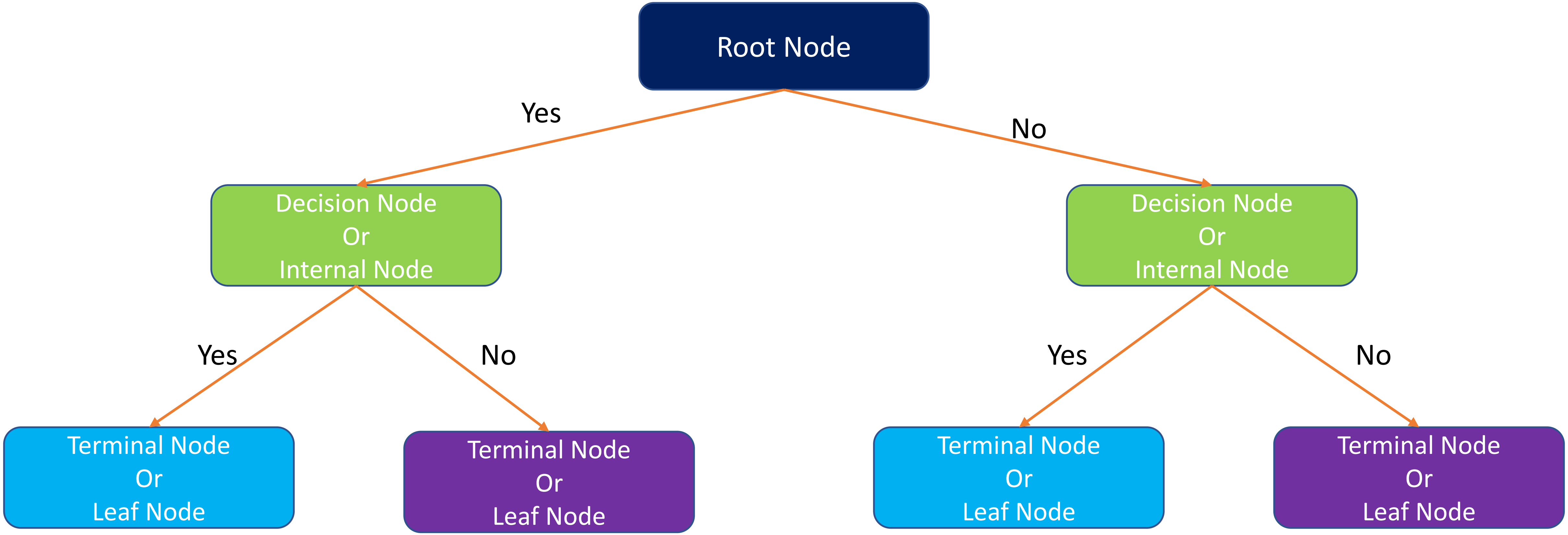
Siddhardhan

Entropy, Information Gain & Gini Impurity



Decision Tree - Terminologies

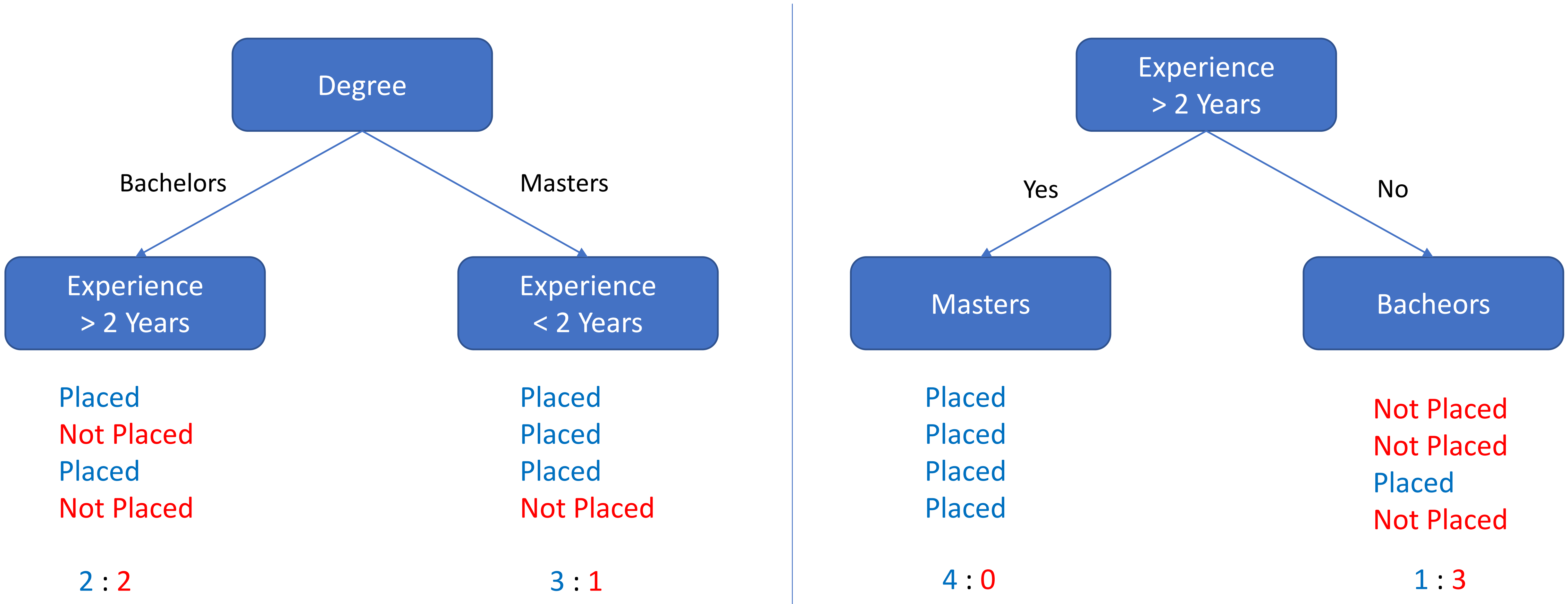


Decision Tree

Problem Statement: Build a Decision Tree to determine whether a person will **get a Job or not** based on their **Degree** & **Years of Experience**.

Degree	Experience in Years	Placed / Not Placed
Masters	2	Placed
Bachelors	0	Not Placed
Masters	3	Placed
Masters	1	Not Placed
Bachelors	2	Placed
Masters	3	Placed
Bachelors	0	Not Placed
Bachelors	1	Not Placed

Decision Tree



Degree

Bachelors

Masters

Experience
> 2 Years

Experience
< 2 Years

Placed
Not Placed
Placed
Not Placed

2 : 2

Placed
Placed
Placed
Not Placed

3 : 1

Experience
> 2 Years

Yes

No

Masters

Bacheors

Placed
Placed
Placed
Placed

4 : 0

Not Placed
Not Placed
Placed
Not Placed

1 : 3

Entropy: High
Information Gain: Low
Gini Impurity: High

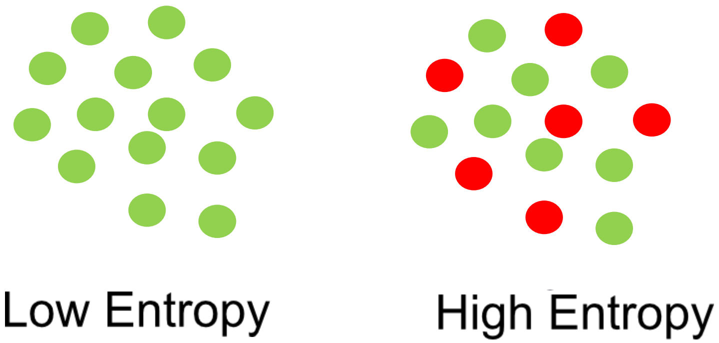
Entropy: Low
Information Gain: High
Gini Impurity: Low

Entropy

Entropy:

In Machine Learning, **Entropy** is the quantitative measure of the **randomness** of the information being processed.

A **high value of Entropy** means that the **randomness** in the system is **high** and thus making accurate predictions is tough.
A **low value of Entropy** means that the **randomness** in the system is **low** and thus making accurate predictions is easier.

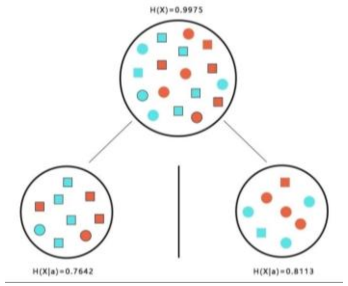


$$\textbf{Entropy} = \sum_{i=1}^{c} -p_i \log_2 p_i$$

c --> number of classes
p_i --> Probability of ith class

Information Gain

Information Gain is the measure of how much information a feature provides about a class. Low entropy leads to increased Information Gain and high entropy leads to low Information Gain.



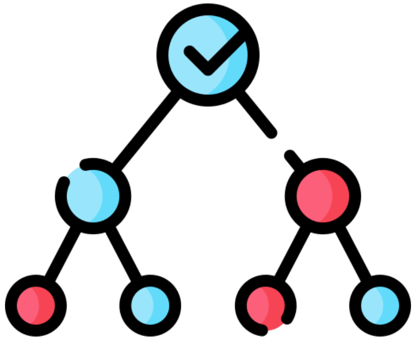
Information gain computes the difference between **entropy before split** and average entropy **after split** of the dataset based on a given feature.

Information gain (T, F) = $$\mathbf{Entropy}\ (\mathbf{T}) - \sum_{v \in F} \frac{|T_v|}{T} .\ \boldsymbol{Entropy}\ (\boldsymbol{T})$$

Gini Impurity

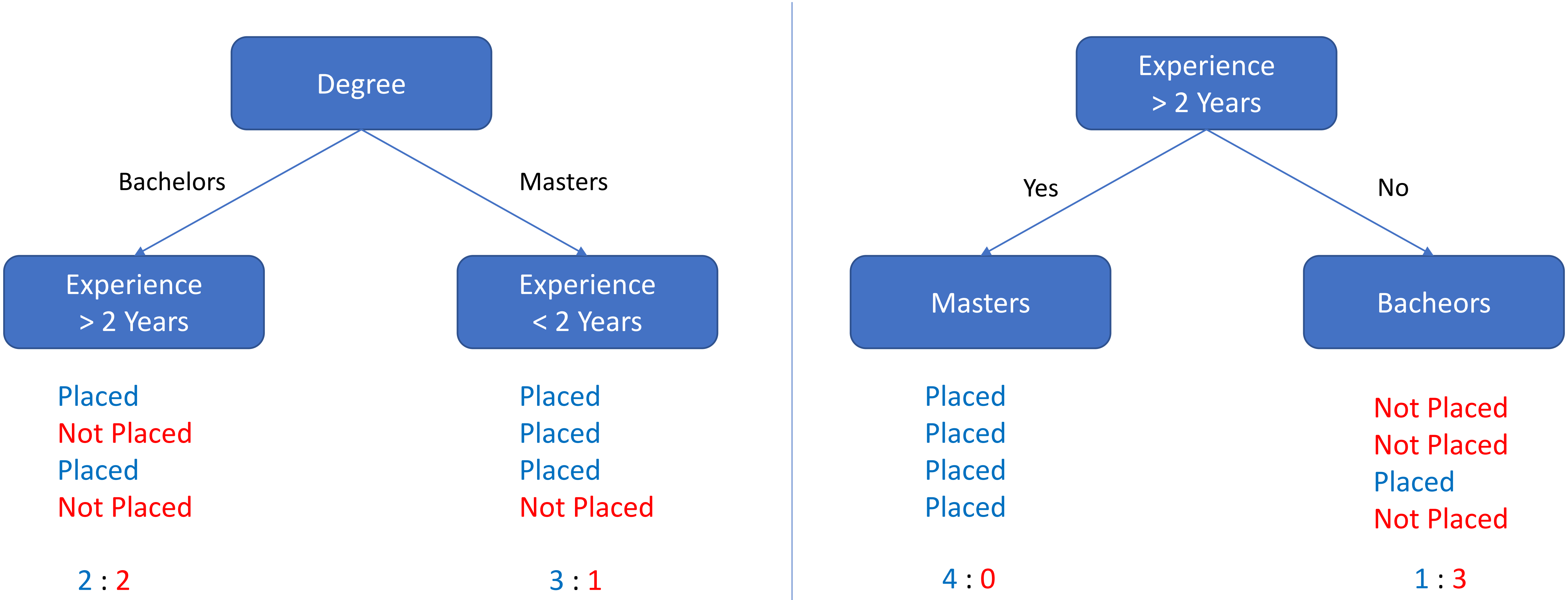
The split made in a Decision Tree is said to be pure if all the data points are accurately separated into different classes.

Gini Impurity measures the likelihood that a randomly selected data point would be incorrectly classified by a specific node.



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

Decision Tree



Degree

Bachelors

Masters

Experience
> 2 Years

Experience
< 2 Years

Placed
Not Placed
Placed
Not Placed

2 : 2

Placed
Placed
Placed
Not Placed

3 : 1

Experience
> 2 Years

Yes

No

Masters

Bacheors

Placed
Placed
Placed
Placed

4 : 0

Not Placed
Not Placed
Placed
Not Placed

1 : 3

Entropy: High
Information Gain: Low
Gini Impurity: High

Entropy: Low
Information Gain: High
Gini Impurity: Low