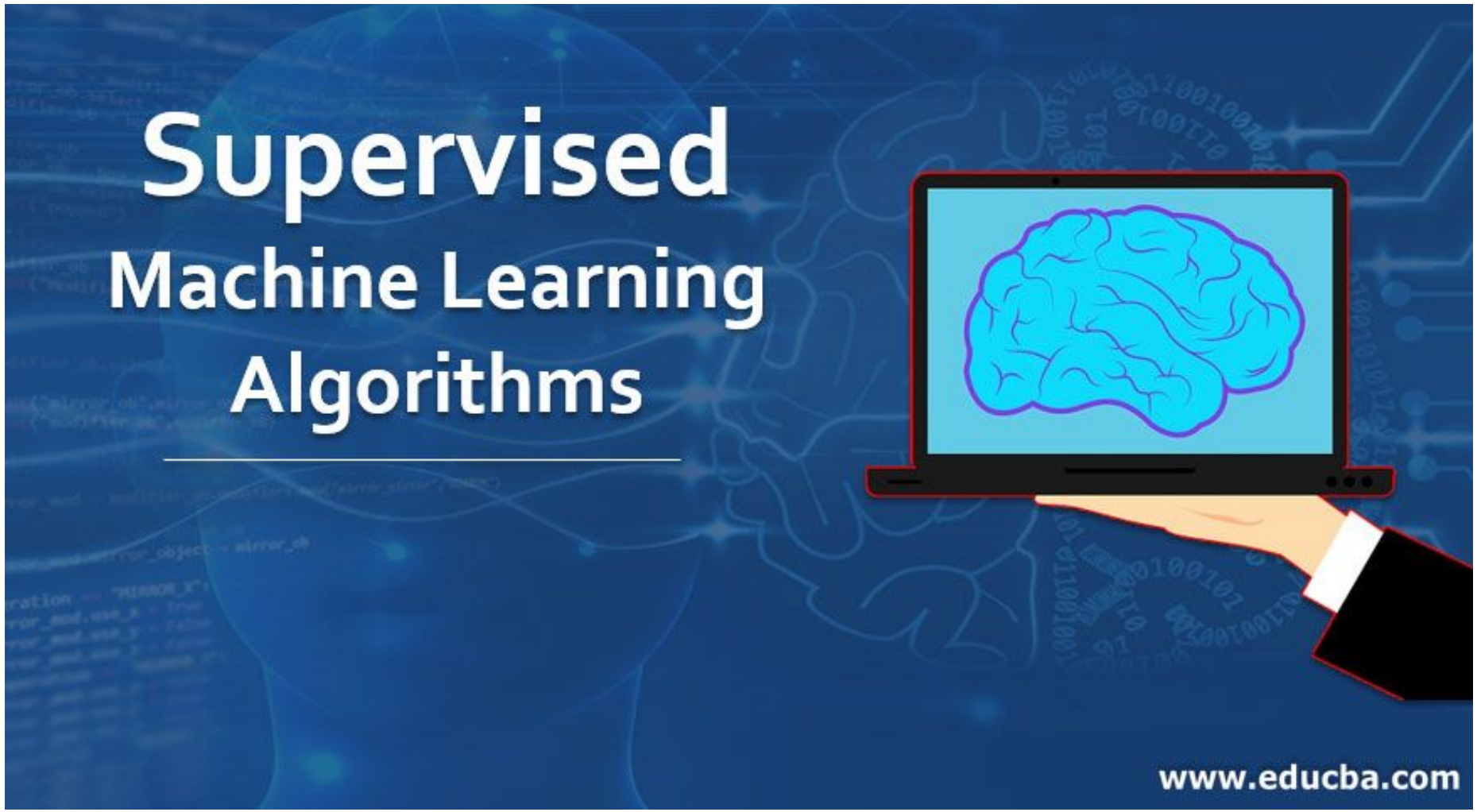

Supervised
Machine Learning
Algorithms
www.educba.com

Last lecture reminder

We learned about:

- SVM - Kernel trick computation
- SVM Classification - Python example
- SVM Classification - Linear Kernel
- SVM Classification - RBF Kernel
- SVM Classification - Polynomial Kernel
- SVM Classification - Grid Search
- SVR (Support Vector Regression) - SVM model for Continuous data

Decision Trees - Theory

Decision Trees algorithms → Decision Trees in Supervised Learning are algorithms that mimic the logical process of human decision-making, applied to predict an output given the provided inputs. In a decision tree, the internal nodes represent the features of a data set, the branches represent the decision rules, and each leaf node represents the outcome.

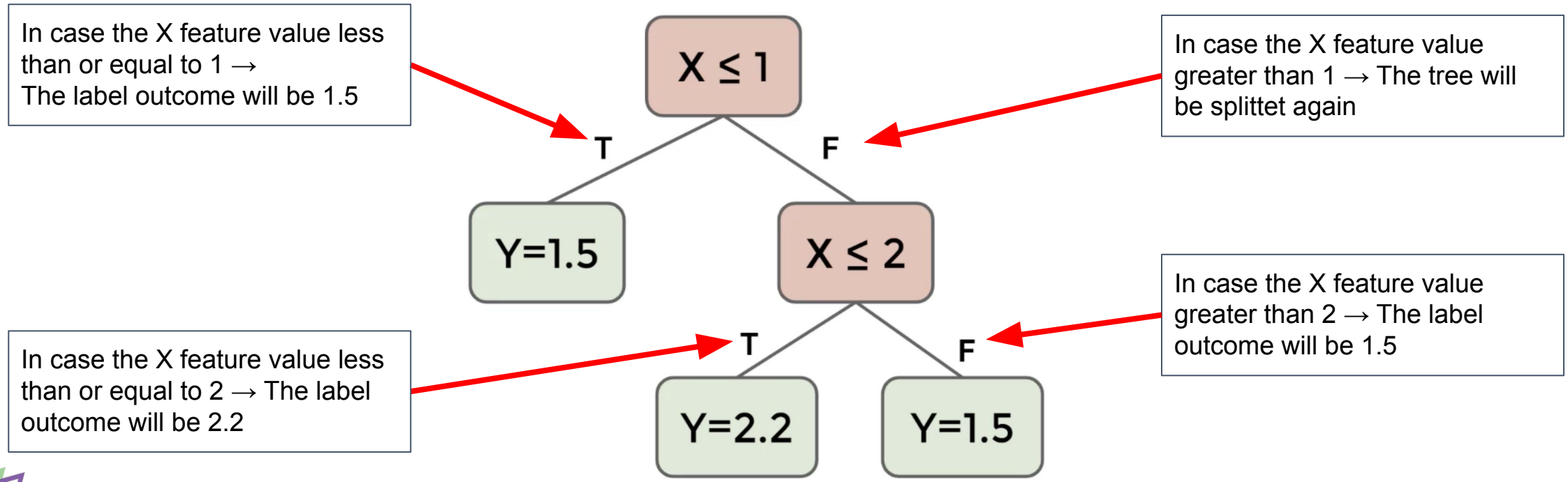
Conceptually, it can be considered as dividing a large set of conditions/data into smaller ones, which then further branch into even smaller sets, forming a tree-like structure.

The primary advantage of Decision Trees is their ease of interpretation and visualization, while their major disadvantage lies in their tendency to overfit data, low prediction accuracy and bias in handling imbalanced datasets.

Decision Tree - Terminology

Before starting to explore the different decision tree algorithms we first need to understand the decision tree basic terminology.

For that, let's take the simple following decision tree (X - feature value, Y - label value):

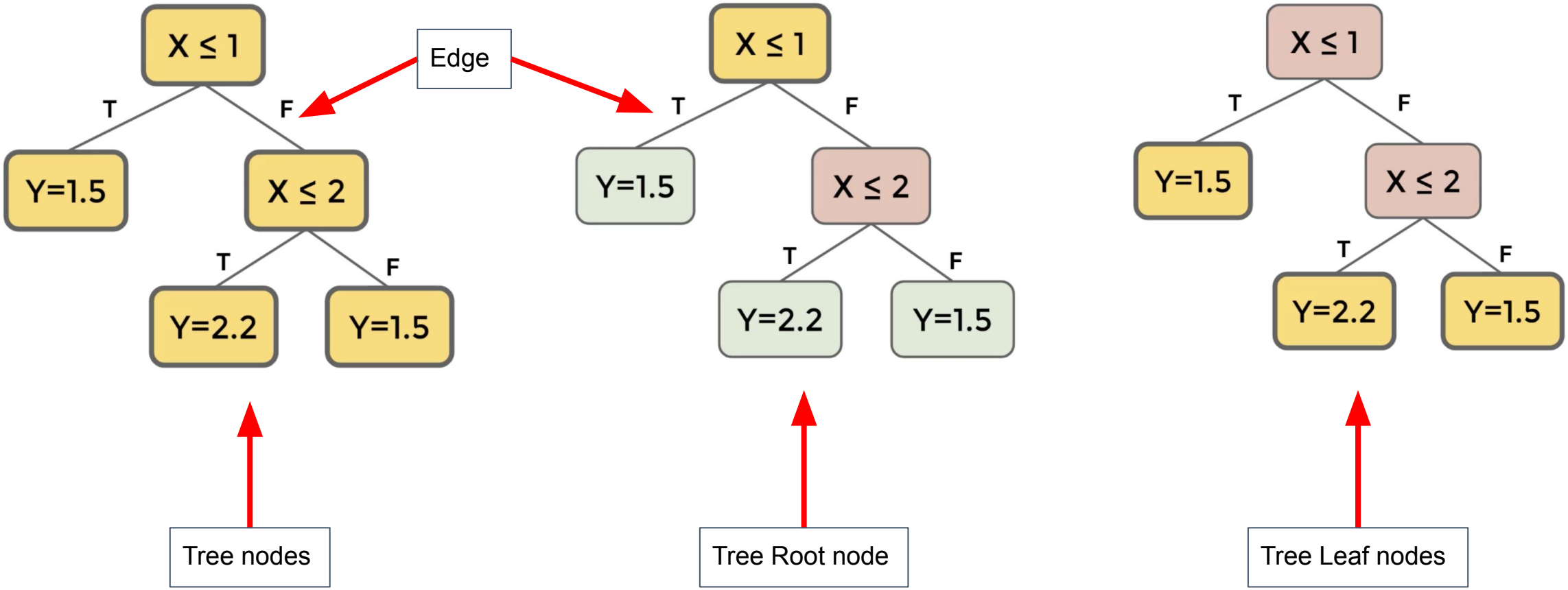


Decision Tree - Terminology

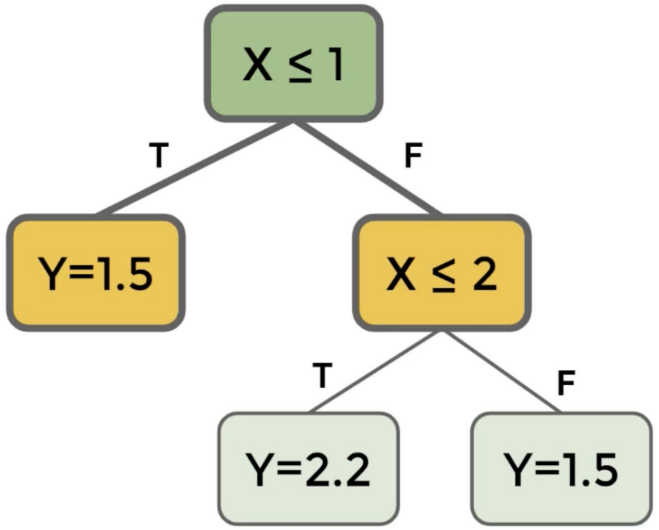
Let’s now understand some basic terminologies regarding decision trees and trees in general.

- **Node** → A node is an entity that contains a value or data and can be connected by links to other nodes.
- **Parent and child** → A node, which is divided into sub-nodes is called parent of sub-nodes whereas the sub-nodes are the children of a parent node.
- **Root** → The top node in a tree, with no parent meaning no nodes that are above him, the root is the first node in the tree and all the other nodes are it’s children.
- **Leaf** → A node which does not have any child node is called a leaf node.
- **Edge** → The connection between one node and another.
- **Height of a Node** → The length of the longest downward path to a leaf from that node.
- **Depth of a Node** → The length of the path to its root.
- **Sub-tree** → Subtree represents the descendants of a node.

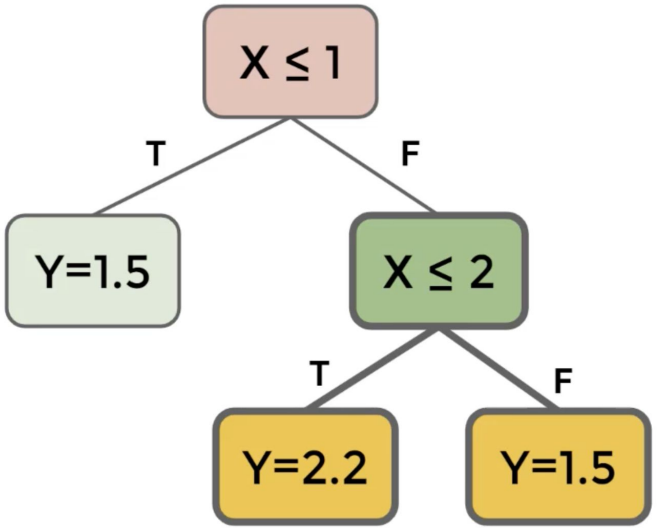
Decision Tree - Terminology



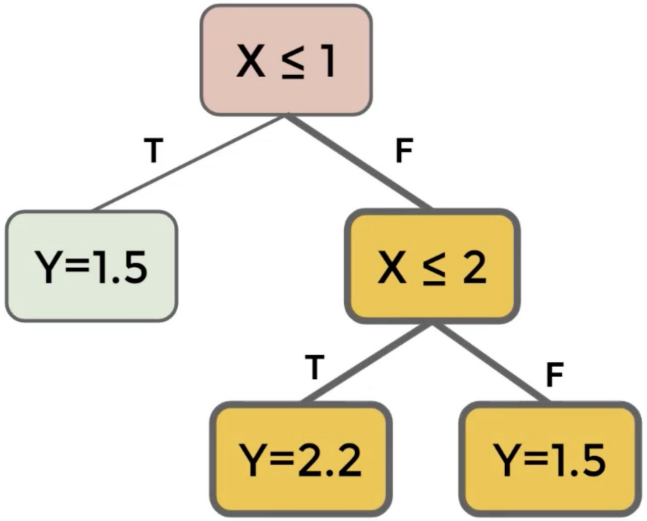
Decision Tree - Terminology



Parent node for
y=1.5 and x<= 2



Children nodes of
x<= 2 node

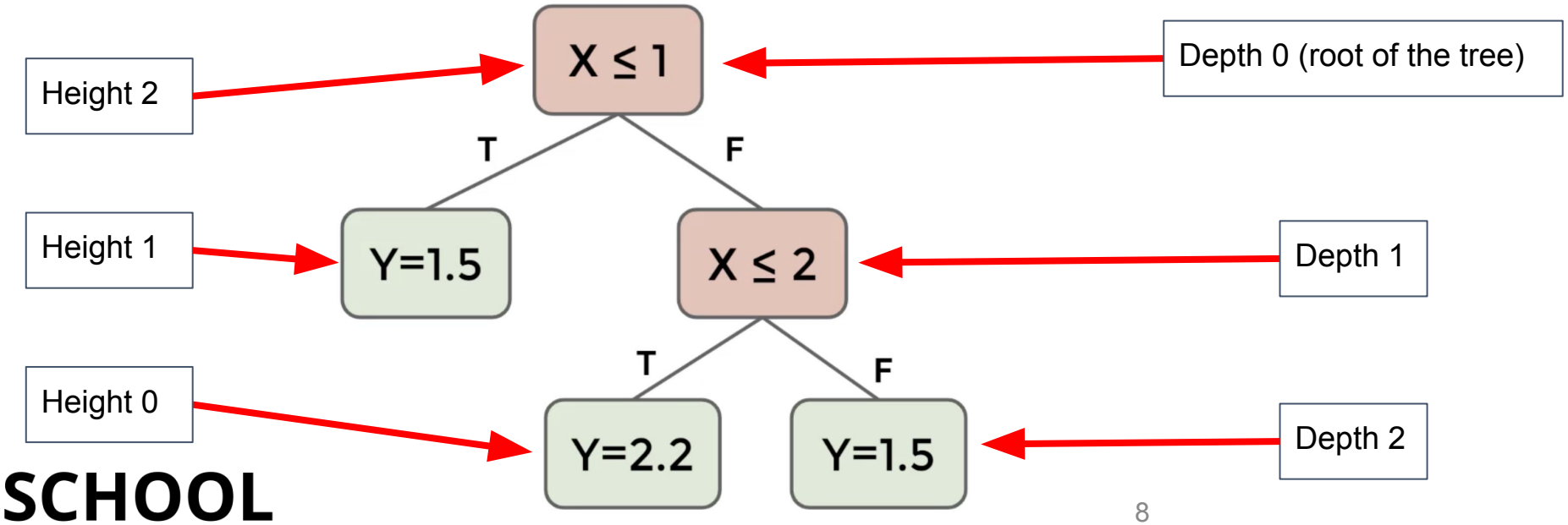


Subtree nodes

Decision Tree - Terminology

Tree depth → The depth of a tree could be seen as the depth of the deepest node in the tree, or the maximum level any node is on. The root node is the first level, its immediate child nodes are on the second level and so on.

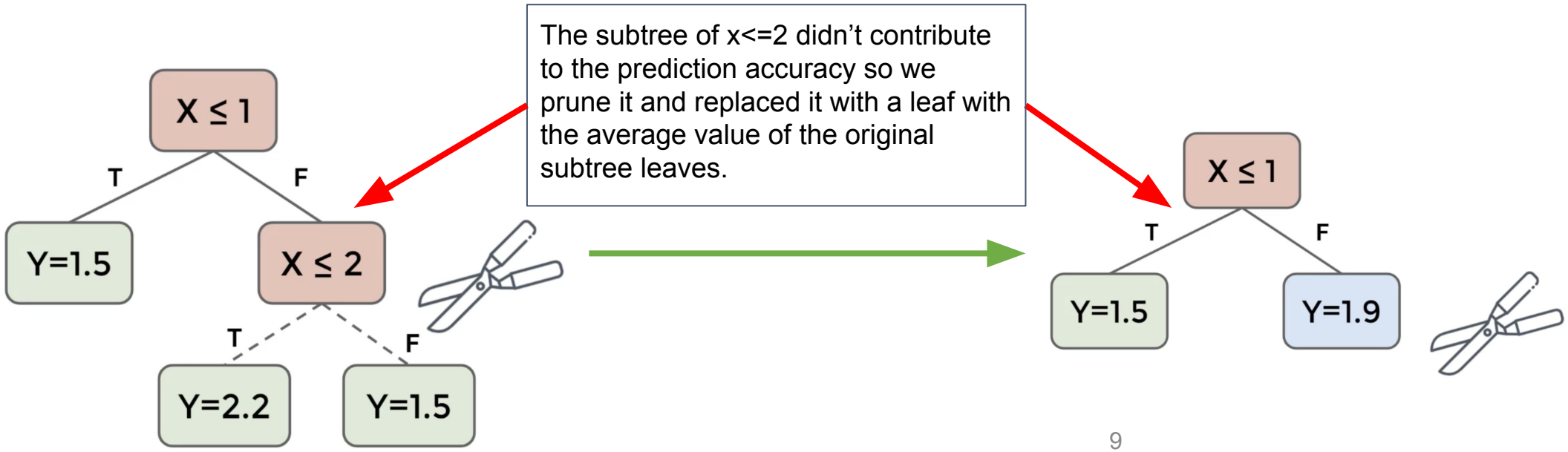
Tree height → The height of a tree is taken as the height of the root node, and is the length of the longest path from the root node to any leaf node in the tree.



Decision Tree - Terminology

Pruning → Pruning, in the context of machine learning and decision trees, is a technique used to reduce the size of trees by removing sections of the tree that do not provide power to classify instances. The goal of pruning is to improve the prediction accuracy of a decision tree model, or conversely to prevent overfitting.

When a subset of nodes (a sub-tree) doesn't significantly contribute to the prediction accuracy, it can be pruned (removed) to simplify the model.



Decision Tree - Gini Impurity

Until now we saw common terms in tree data structure but we still need to understand how can we convert our dataset in a decision tree.
In machine learning, when building decision trees, we need a metric or measurement to help the algorithm decide how to split the data at each node.

The most common information measurement that will help our algorithm decide how to build the decision tree from the provided data set is called **Gini Impurity**.

Gini Impurity → Gini Impurity is a metric used in decision trees algorithm for the task of decision making like deciding where to split the data. It measures the degree or probability of a specific variable being wrongly classified when it is randomly picked.

Decision Tree - Gini Impurity

Gini Impurity gives a measure of how often a randomly chosen element from a set would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset. The algorithm will try to minimize the Gini Impurity to make the best splits.

Gini Impurity formula for classification problems is the following

$$G(Q) = \sum_{c \in C} p_c(1 - p_c)$$

Q → Specific node in the decisıon tree

G(Q) → The Gini Impurity value of that Q node

Pc → The probability of belonging to a class C in Q node

1 - Pc → The probability of not belonging to a class C in Q node

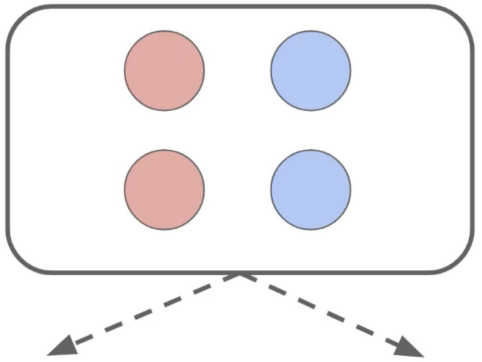
Σ → Sum of the probabilities of all classes in Q node

Decision Tree - Gini Impurity

For example → Let's take a simple classification problem with 2 classes and calculate the Gini Impurity of a specific node in the decision tree.

Let's say for example entire dataset has 4 points 2 of them classified as red and the other 2 points classified as blue.

$$G(Q) = \sum_{c \in C} p_c(1 - p_c)$$



Let's now calculate the G(Q) of that Q node:

Class red → Pc = Probability to get red = 2 / 4 = 0.5

1 - Pc = Probability of not get red = 1 - 0.5 = 0.5

Pc * (1 - Pc) = 0.5 * 0.5 = 0.25

Decision Tree - Gini Impurity

Class blue → Pc = Probability to get blue = 2 / 4 = 0.5

1 - Pc = Probability of not get blue = 1 - 0.5 = 0.5

Pc * (1 - Pc) = 0.5 * 0.5 = 0.25

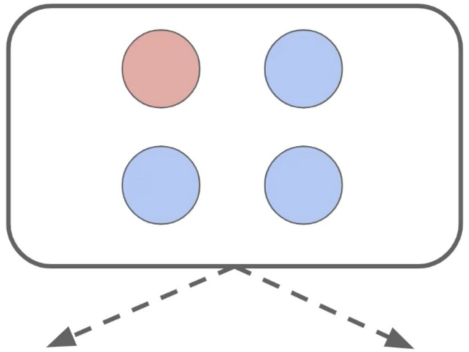
$G(Q) = \Sigma(Pc * (1 - Pc)) = 0.25 + 0.25 = 0.5$

So we got that the Gini Impurity value for the node is 0.5

Now, let's take another example which now the classes doesn't have the same amount of data points and calculate the Gini Impurity of the Q node.

Decision Tree - Gini Impurity

$$G(Q) = \sum_{c \in C} p_c(1 - p_c)$$



Let's now calculate the G(Q) of that Q node:

Class red → Pc = Probability to get red = 1 / 4 = 0.25

1 - Pc = Probability of not get red = 1 - 0.25 = 0.75

Pc * (1 - Pc) = 0.25 * 0.75 = 0.1875

Class blue → Pc = Probability to get blue = 3 / 4 = 0.75

1 - Pc = Probability of not get blue = 1 - 0.75 = 0.25

Pc * (1 - Pc) = 0.75 * 0.25 = 0.1875

Decision Tree - Gini Impurity

$G(Q) = \Sigma(Pc * (1 - Pc)) = 0.1875 + 0.1875 = 0.375$

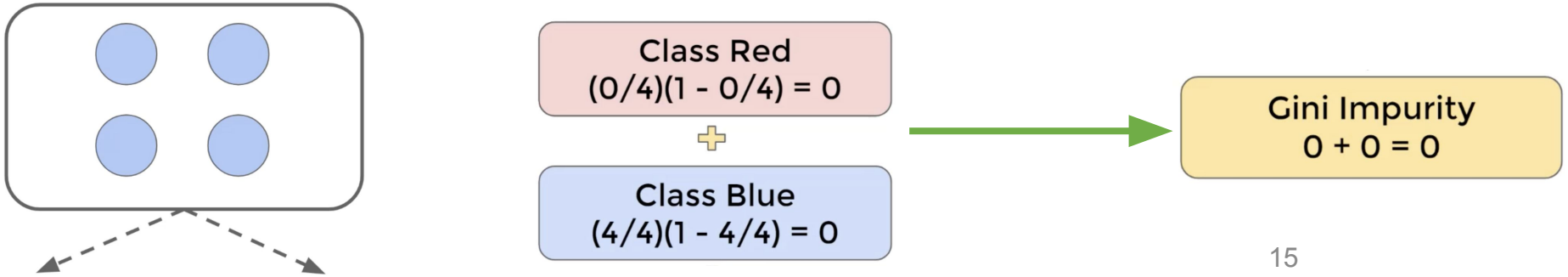
So we got that the Gini Impurity value for the node is 0.375

What we can see here is that the Gini Impurity value represent how “pure” our training dataset is.

When we have classes that have larger amount of data points then others our dataset becomes more “pure” so our Gini Impurity value decreases.

If we will take a look on a node with only blue data points (our data set is now the “purest” possible).

We will see that the Gini Impurity value of that node will be 0.

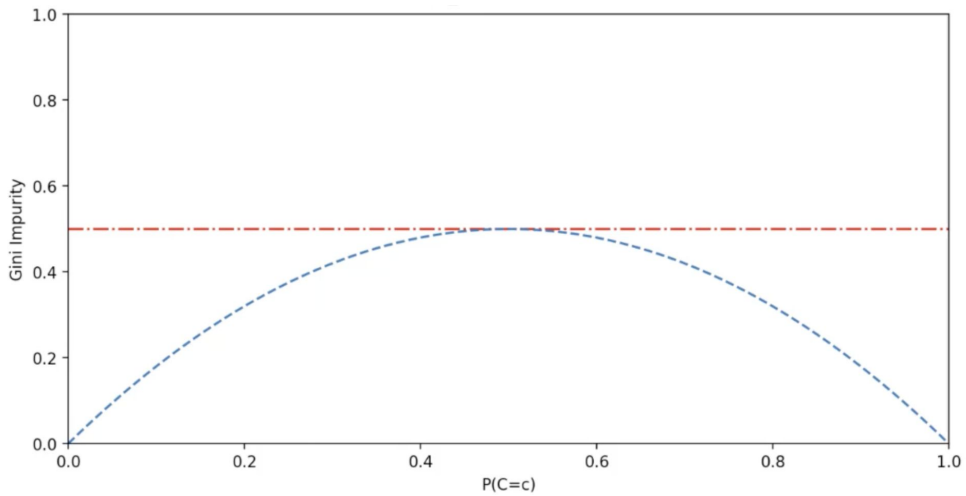


Decision Tree - Gini Impurity

When we had the most “pure” dataset, all the data points belong to one class and 0 data points belong to the other classes, we got Gini Impurity value of 0.
When we had the most “impure” dataset, each category had the same amount of points from the data set, we got Gini Impurity value of 0.5.

In conclusion, we can tell that for binary classification problem (problem with 2 categories) the Gini Impurity value can very between 0 - 0.5 while 0 indicate our data is the most “pure” and 0.5 indicate that our data is the most “impure”.

For binary classification problems:
Pc = 0.5 → G(Q) = 0.5 (max value)
Pc = 0 / 1 → G(Q) = 0 (min value)



Decision Tree - Gini Impurity Example

Let’s construct a simple decision tree from a given dataset and calculate it’s Gini Impurity:

For this example we will use a given dataset providing data about emails with URL link inside and if that email was a spam or not.

X - URL Link	Y-Spam
Yes	Yes
Yes	Yes
No	No
No	No
No	Yes
No	No
Yes	No

We want to predict according the given data if new emails will be spam or not.

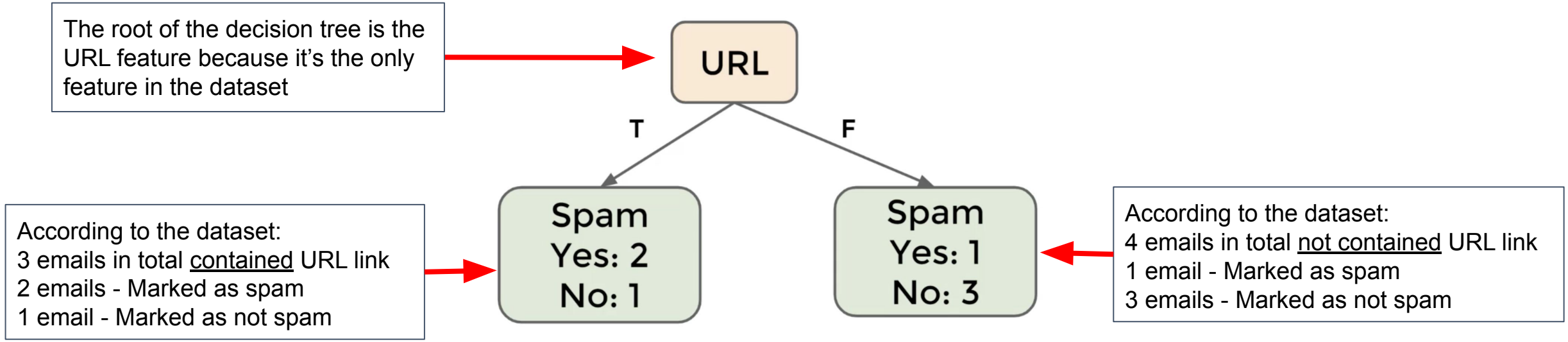
Decision Tree - Gini Impurity Example

According to the given data:

X (feature) → A binary categorical feature (has URL link or not)

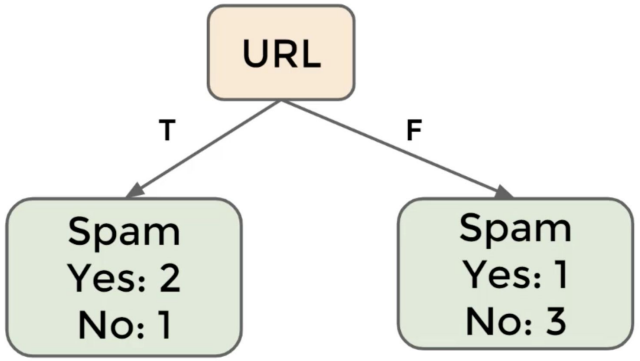
Y (Label) → A binary categorical label (marked as spam or not)

Now, let's construct a decision tree based on the provided dataset and calculate the Gini Impurity.



Decision Tree - Gini Impurity Example

Now we can calculate for each leaf node the Gini Impurity of that node. Once we calculated the Gini Impurity of all the leaf nodes in the tree we can also calculate the total Gini Impurity of the entire decision tree.



$$G(Q) = \sum_{c \in C} p_c(1 - p_c)$$

Gini Impurity calculation:

Left leaf node → (2/3)*(1-2/3) + (1/3)*(1-1/3) = 0.44

Right leaf node → (1/4)*(1-1/4) + (3/4)*(1-3/4) = 0.375

Total tree → (3/7)*0.44 + (4/7)*0.375 = 0.403 (explanation in the next slide)

Decision Tree - Gini Impurity Example

Calculating the Gini Impurity for the entire decision tree:

Once we have the Gini Impurity of each leaf node in our decision tree we can use it to calculate the weighted average of each leaf node.

What we did is taking the Gini Impurity of each leaf and multiply it by the percentage of data points that belong to that leaf node from the total data points in the dataset.

In our example →

Left node data points → 3 from 7

Left node Gini Impurity → 0.44

Right node data points → 4 from 7

Right node Gini Impurity → 0.375

Gini Impurity for the total tree → (3/7)*0.44 + (4/7)*0.375 = 0.403

Class Exercise - Gini Impurity

Instructions:

You are working as a data science in a car company and provided with the following dataset.

The dataset consists of the color of 10 cars and whether or not these cars are sports cars.

The available colors are Red (R) and Blue (B).

Your mission is to predict if a new car will be sport car or not.

- Construct a simple decision tree based on the dataset given
- Calculate for each leaf node the Gini Impurity value
- Calculate the entire tree Gini Impurity value

Class Exercise - Gini Impurity

Provided Data Set:

Car Number	Car Color	Is Sport Car
1	Red	Yes
2	Blue	No
3	Red	Yes
4	Blue	No
5	Red	Yes
6	Blue	Yes
7	Red	No
8	Blue	Yes
9	Red	No
10	Blue	No

Class Exercise Solution - Gini Impurity