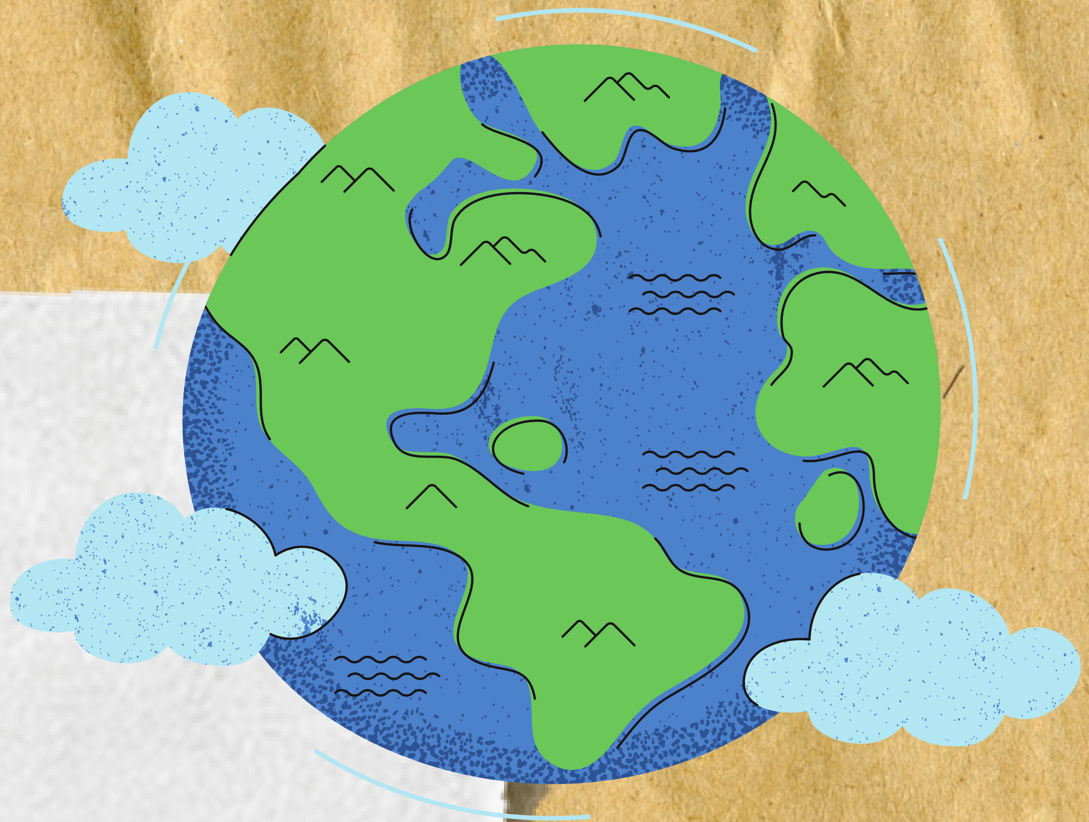
Decision

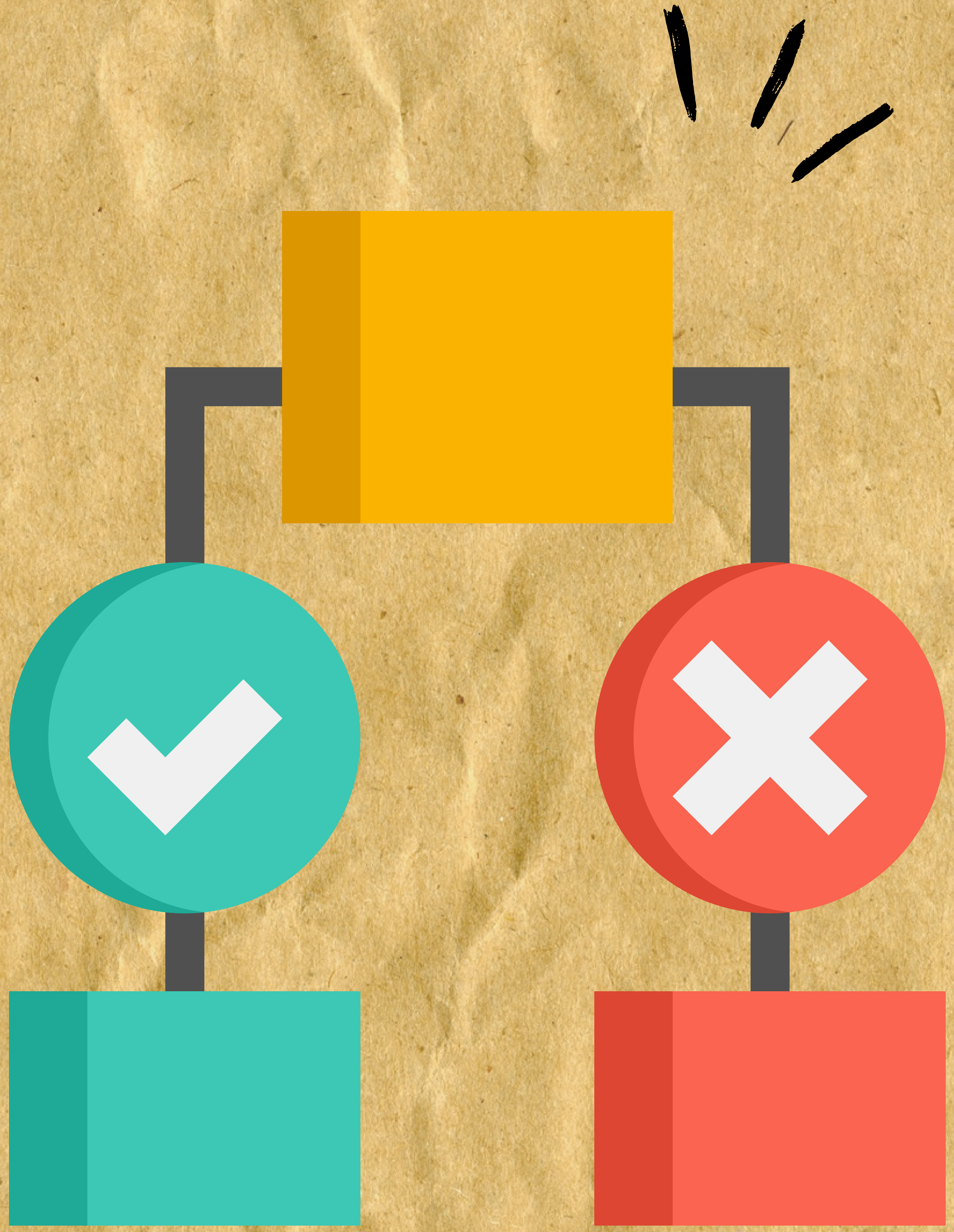
TREE



For classification

What is it?

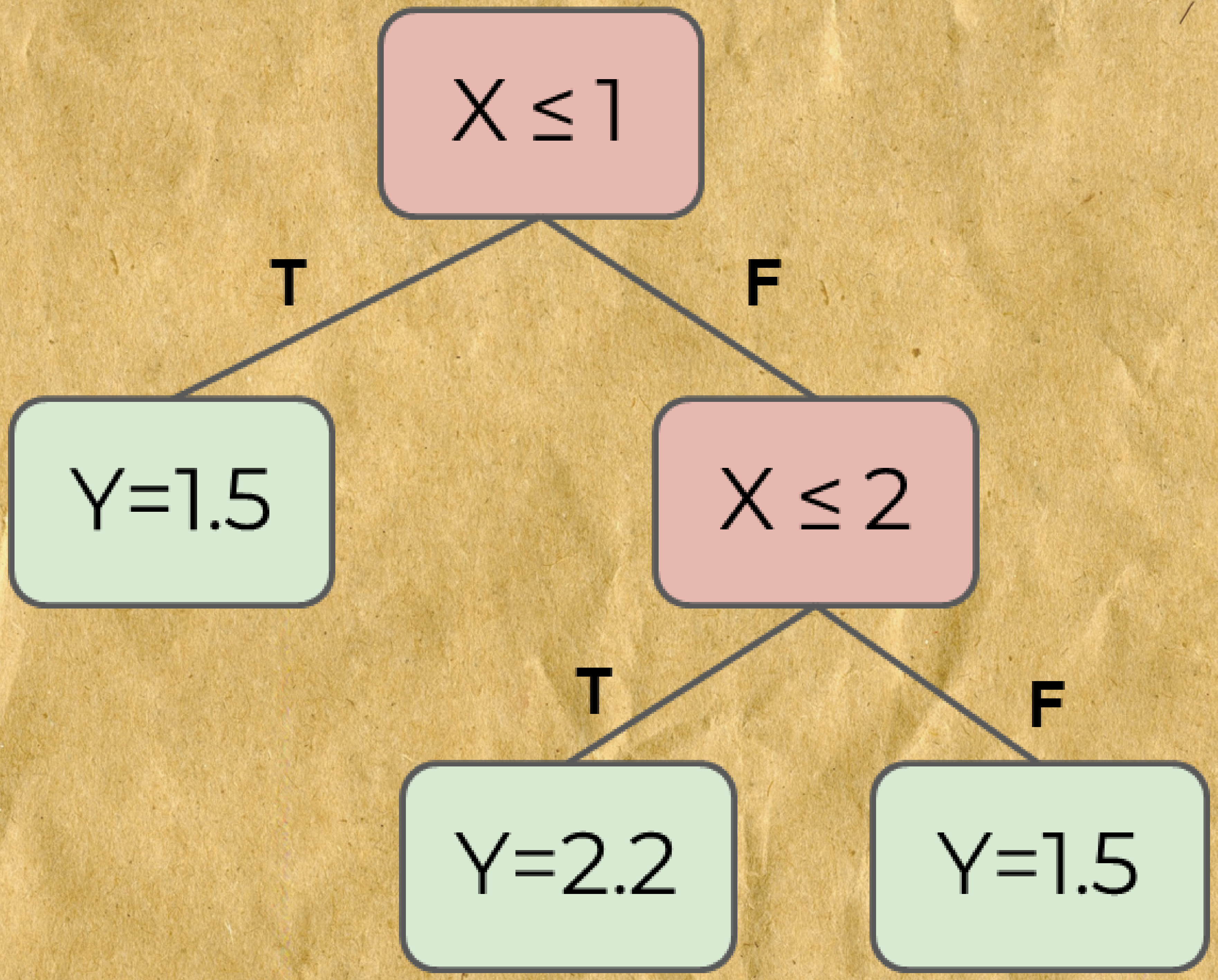
It is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome



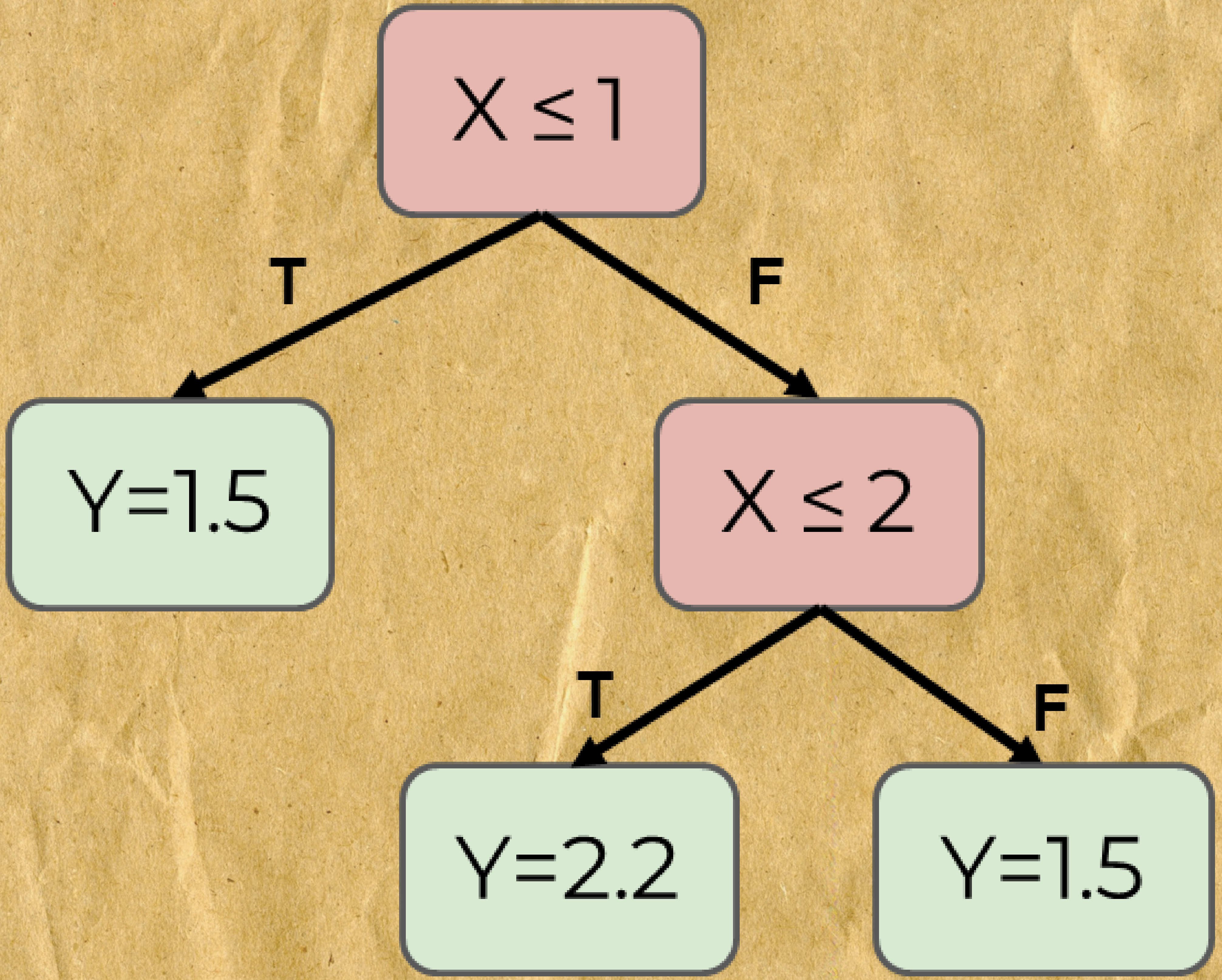
How is it constructed?

Classification and Regression Tree Algorithm (CART)

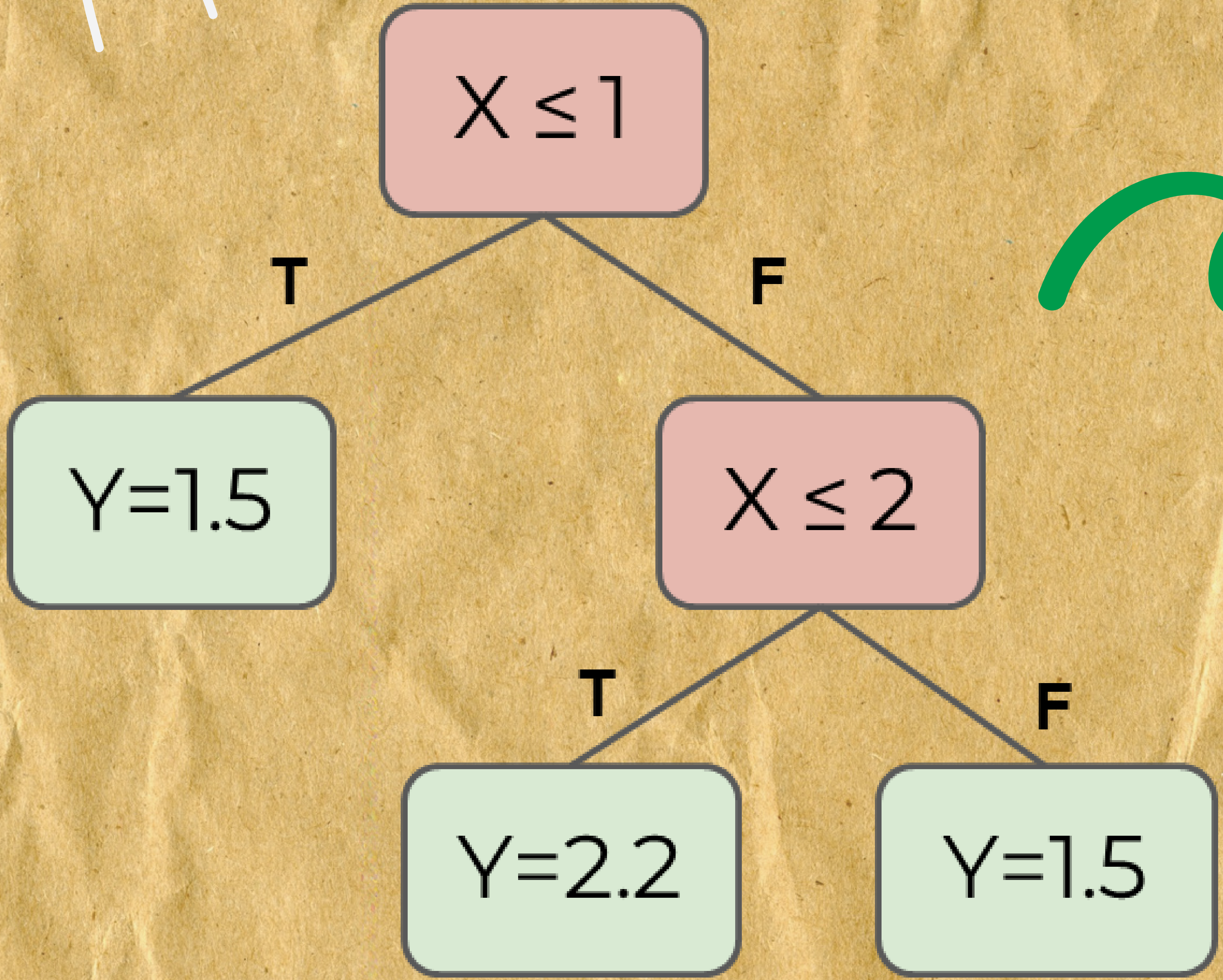
Asking Yes/No Questions

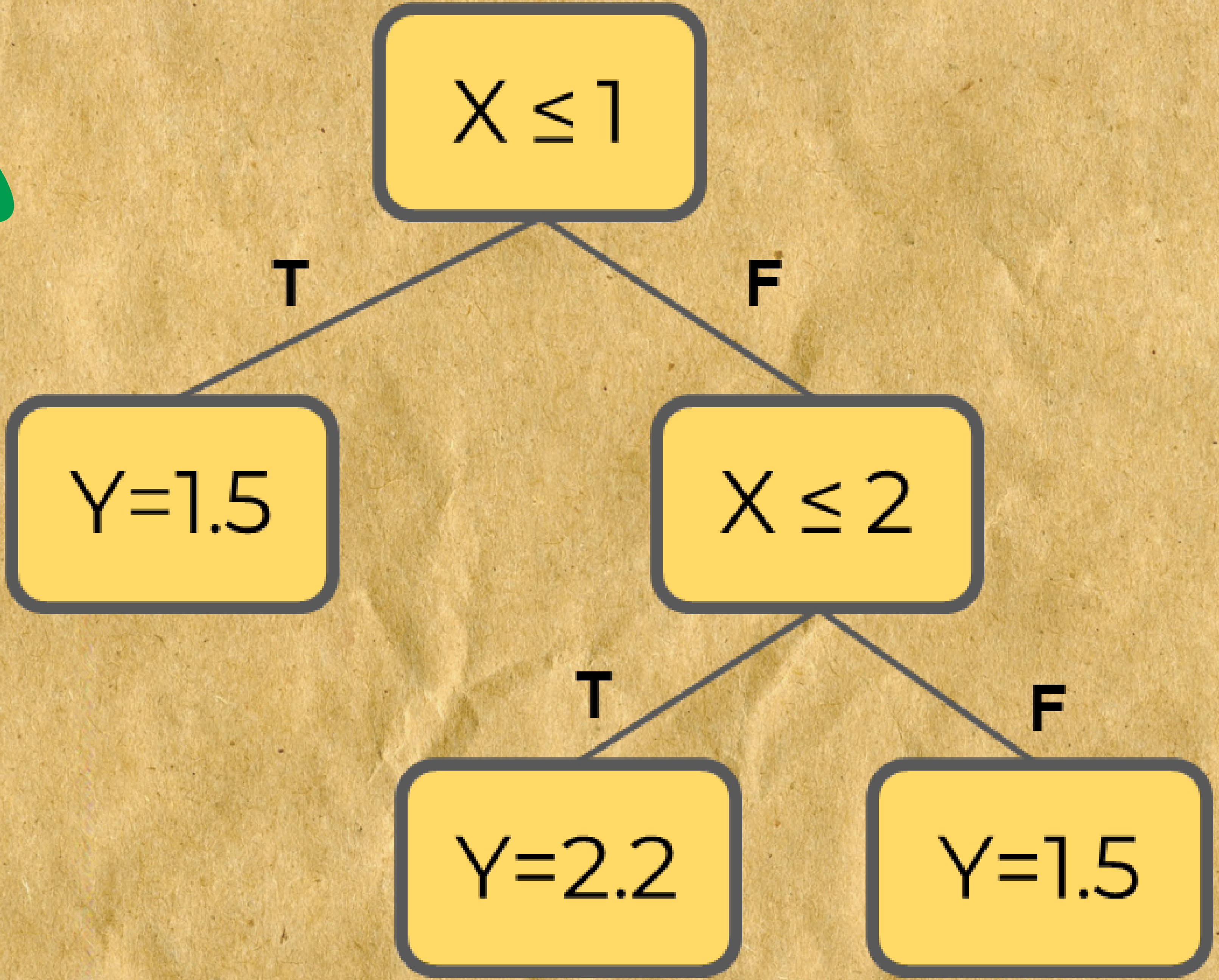


Splitting

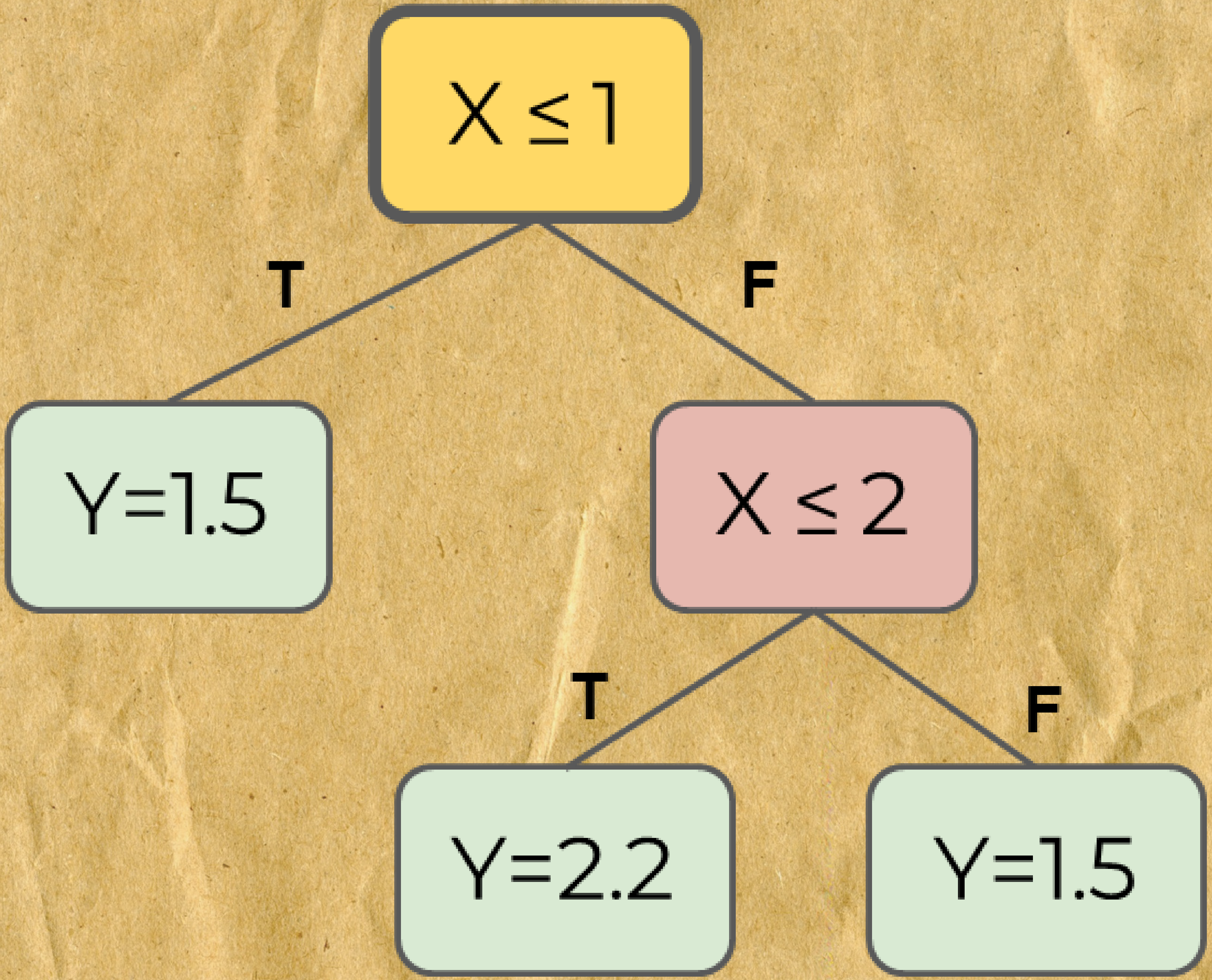


Nodes

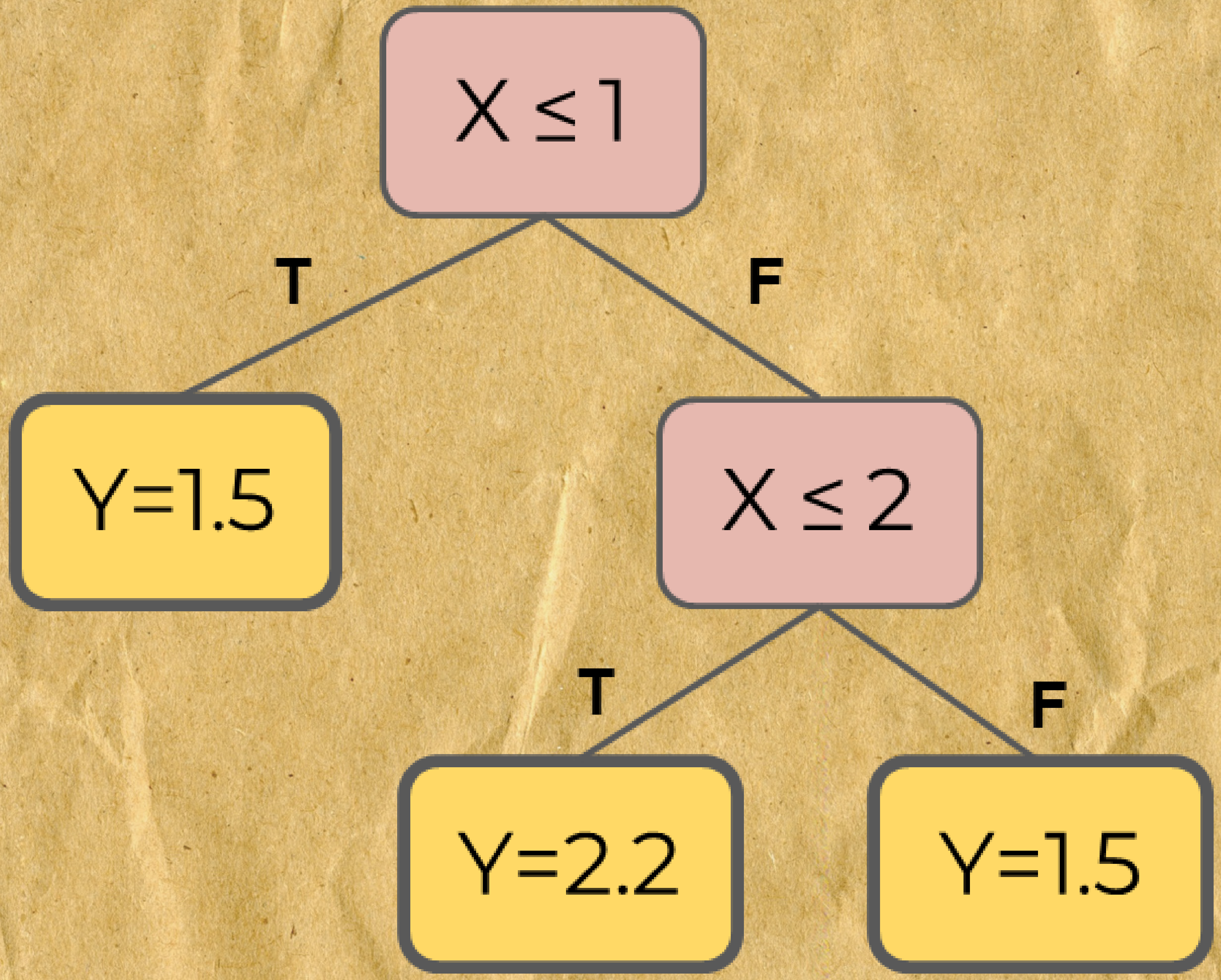




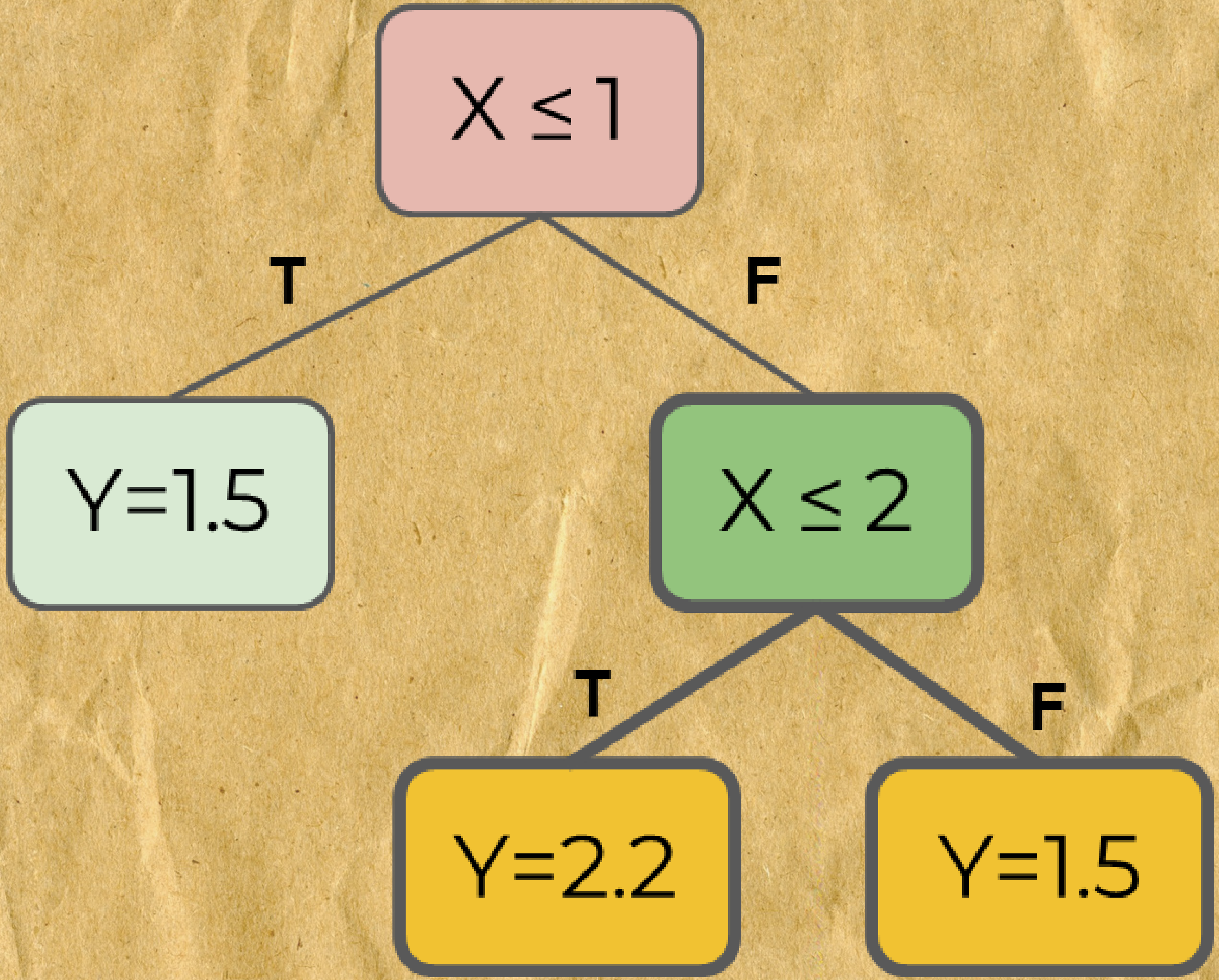
Root Node



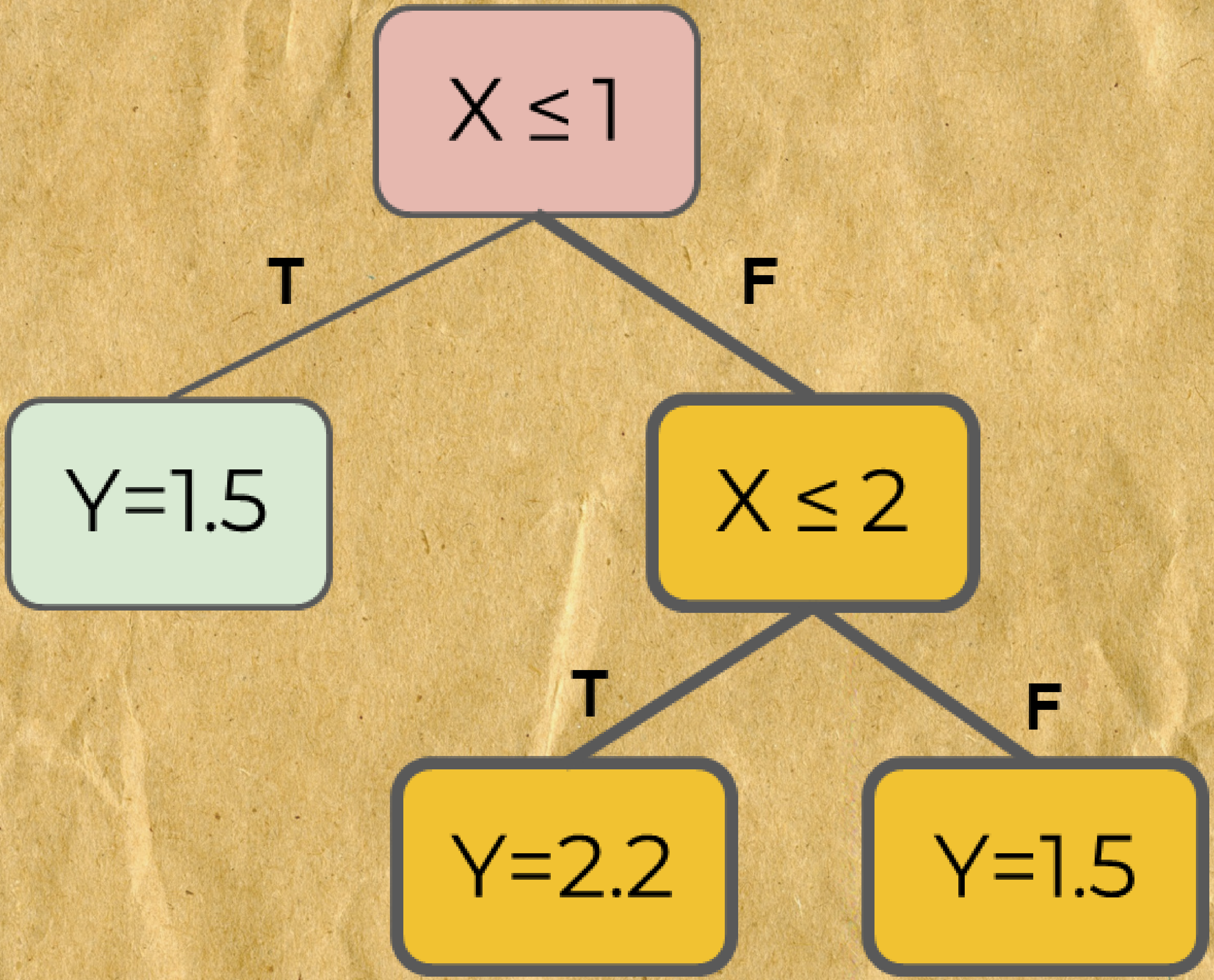
Leaf (Terminal) Nodes:



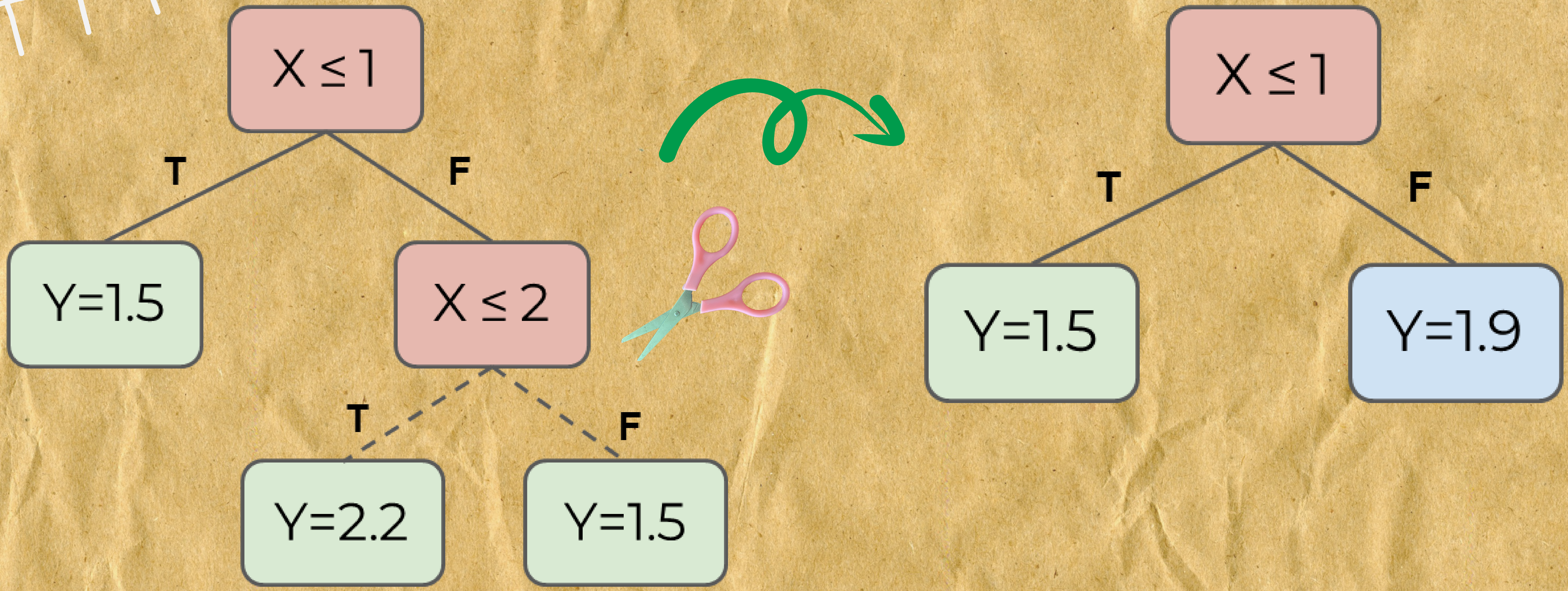
Parent & Children Nodes



Tree Branches (Sub Trees)



Pruning



Choose the best attribute

There are multiple ways to select the best attribute at each node, two methods, information gain (entropy) and Gini impurity, act as popular splitting criterion for decision tree models.

Entropy and Information Gain

Entropy is a concept that stems from information theory, which measures the impurity of the sample values. It is defined by the following formula, where:

$$\mathrm{Entropy}(S) = -\sum_{c \in C} p(c) \log_2 p(c)$$

- S represents the data set that entropy is calculated
- c represents the classes in set, S
- p(c) represents the proportion of data points that belong to class c to the number of total data points in set, S

The attribute with the highest information gain is preferred for splitting, as it effectively classifies training data based on the target classification. Information gain is typically calculated using the formula:

$$\text{Information Gain}(S,a) = \text{Entropy}(S) - \sum_{\text{vevcalues}(a)} \frac{|S_v|}{|S|} \text{Entropy}(S_v)$$

- a represents a specific attribute or class label
- Entropy(S) is the entropy of dataset, S
- |Sv|/ |S| represents the proportion of the values in Sv to the number of values in dataset, S
- Entropy(Sv) is the entropy of dataset, Sv

Gini Impurity

Gini impurity is the probability of incorrectly classifying random data point in the dataset if it were labeled based on the class distribution of the dataset. Similar to entropy, if set, S, is pure—i.e. belonging to one class) then, its impurity is zero. This is denoted by the following formula:

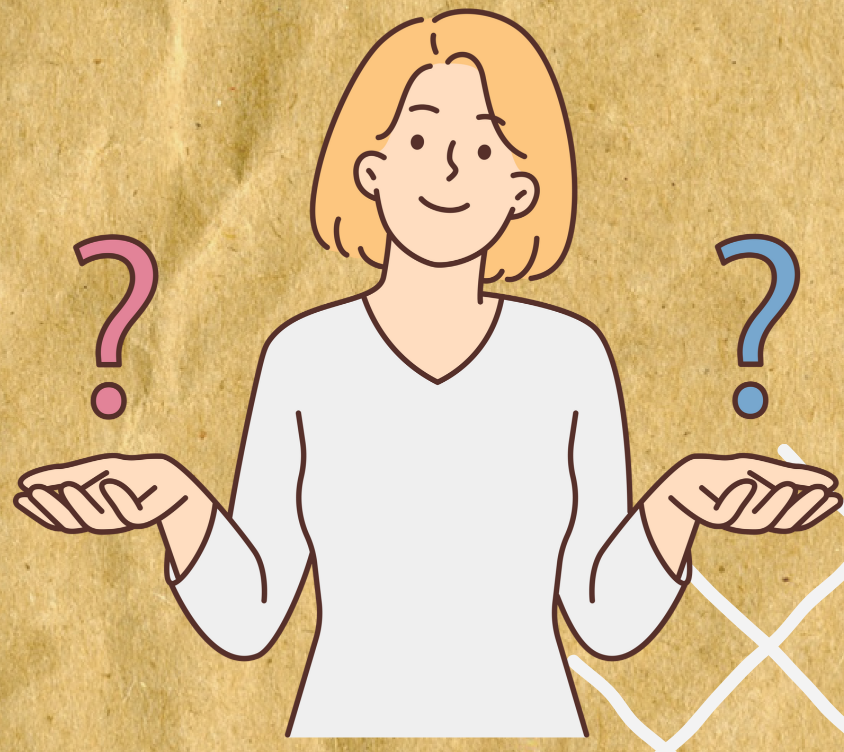
- For a set of classes C for a given dataset Q, pc is probability of class c.

$$\text{Gini Impurity} = 1 - \sum_i (p_i)^2$$

$$p_c = \frac{1}{N_Q} \sum_{x \in Q} \mathbb{1}(y_{class} = c) \qquad G(Q) = \sum_{c \in C} p_c(1 - p_c)$$

Advantages

- **Easy to interpret:** The hierarchical structure allows for a clear identification of important attributes, which might be less evident in other algorithms like neural networks.
- **More flexible:** Decision trees serve both classification and regression tasks, offering flexibility. They are insensitive to correlations between attributes, choosing only one feature to split on if variables are highly correlated.
- **Little to no data preparation required**: Decision trees are flexible and can handle various data types, including discrete or continuous values.



Disadvantages

Prone to overfitting:
Complex decision trees may overfit and struggle to generalize to new data. This can be addressed through pruning .

More costly:
The greedy search approach during construction makes decision trees more expensive to train compared to some other algorithms.

High variance estimators:
Small variations in data can lead to significantly different decision trees. While bagging (averaging estimates) can reduce variance, it may result in highly correlated predictors, limiting its effectiveness.