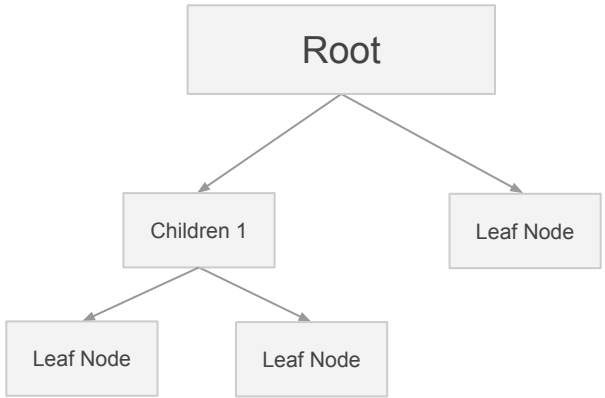
Understanding Decision Tree

Decision tree is a supervised learning algorithm used for both classification and regression tasks. It creates a model that predicts the value of a target variable by learning simple decision rules inferred from the training data.



Key Concepts of Decision Trees:

Tree Structure: A decision tree consists of nodes, branches, and leaves:

Nodes: Represent features or attributes in the dataset.

Branches: Connect nodes and represent decisions or rules.

Leaves: Terminal nodes that represent the outcome (class label or value) after applying all rules.

Decision Rules: Each internal node of the tree corresponds to a decision based on a feature, and each leaf node corresponds to a class label (in classification) or a numerical value (in regression).

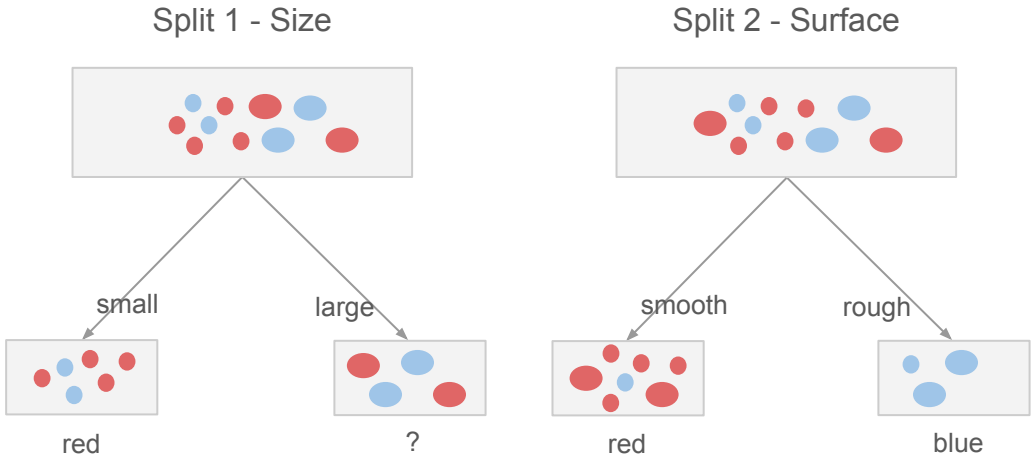
Splitting Criteria: Decision trees use various criteria (like Gini impurity for classification or variance reduction for regression) to split data at each node in a way that maximizes information gain or minimizes impurity.

Training: The algorithm recursively splits the data based on features that best separate the target variable. It selects splits that result in the purest subsets (homogeneous classes or reduced variance).

Prediction: To predict the target variable for new data, the algorithm navigates through the tree based on the learned decision rules until it reaches a leaf node, which provides the prediction.

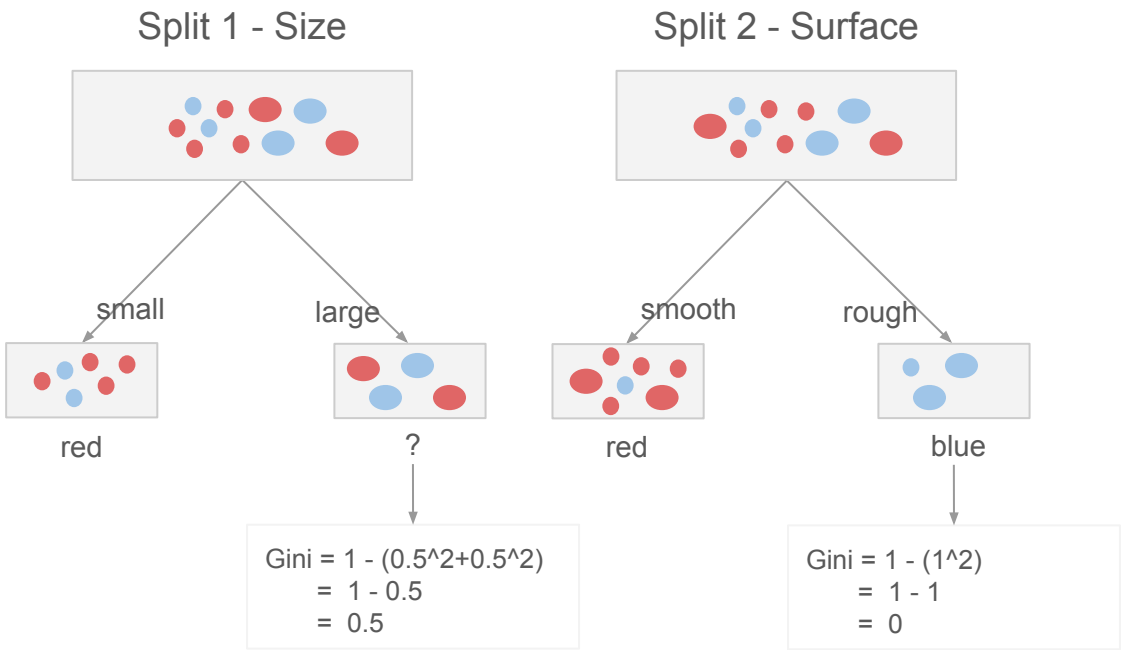
Decision Tree : Classification

Ball_id	Size	Surface	Ball_color
1	small	smooth	red
2	small	rough	blue
3	small	smooth	red
4	small	smooth	blue
5	small	smooth	red
6	small	smooth	red
7	large	smooth	red
8	large	rough	blue
9	large	smooth	red
10	large	rough	blue



- The internal nodes contains test conditions that separate instances based on their characteristics (feature values).
- The decision at a leaf node is made based on the majority class of the training instances that reach at the specific leaf node during the training phase.

Splitting Criterion



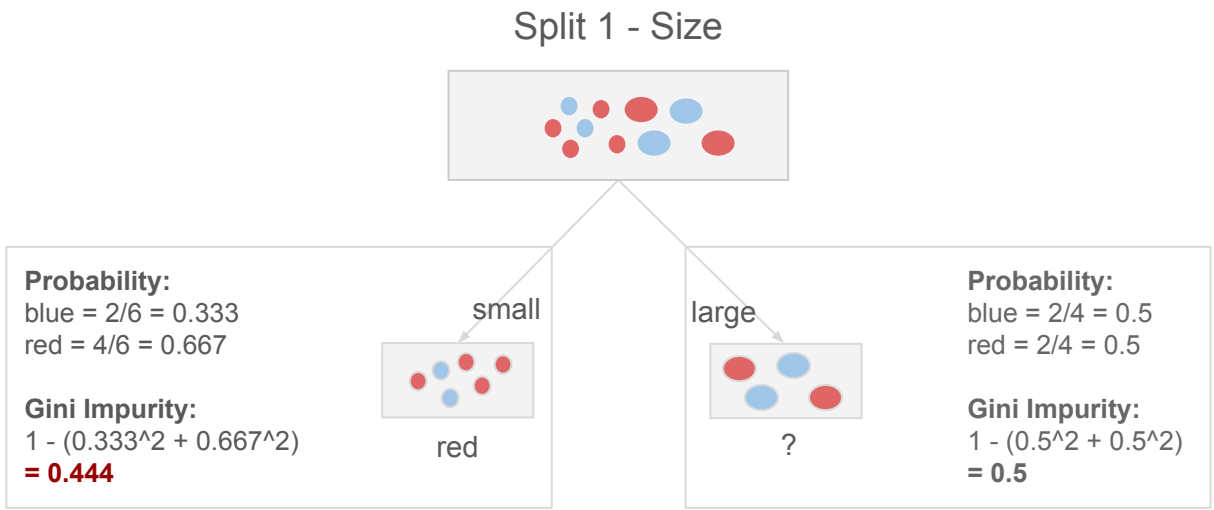
Gini Impurity: Gini impurity measures the degree of impurity or disorder in a set of examples. It is minimized when all instances belong to the same class (impurity is 0) and is maximized when the classes are equally distributed (impurity is 0.5).

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

Entropy: Entropy is a measure of uncertainty of a set of instances. The goal is to minimize entropy, and like Gini impurity, it is 0 when all instances belong to the same class and is higher when the classes are equally distributed.

$$Entropy = \sum_{i=1}^{C} -p_i * \log_2(p_i)$$

Weighted Gini Score



Probability:
blue = 2/6 = 0.333
red = 4/6 = 0.667

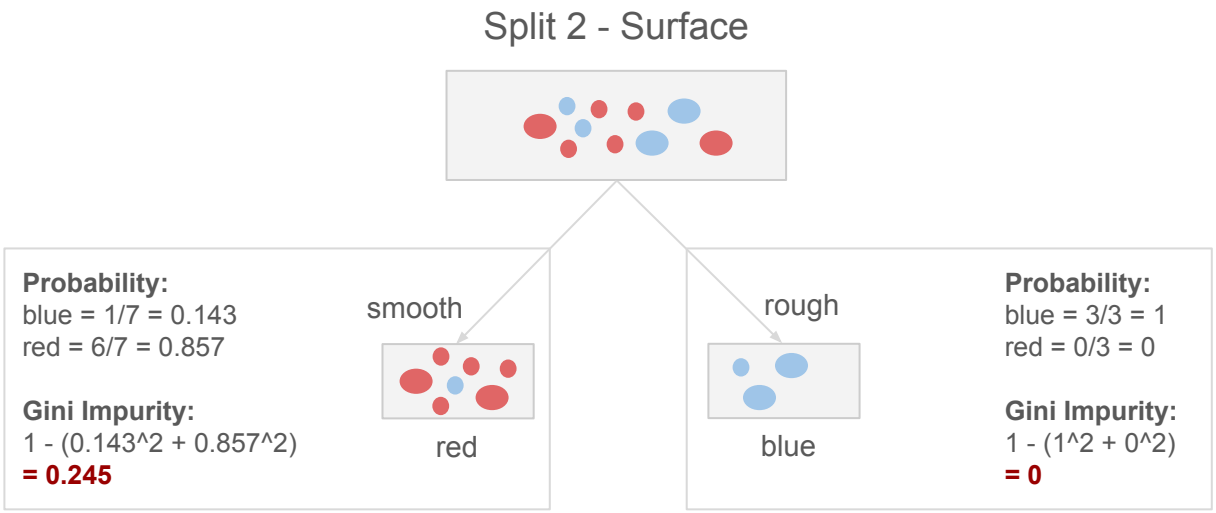
Gini Impurity:
1 - (0.333^2 + 0.667^2)
= 0.444

Probability:
blue = 2/4 = 0.5
red = 2/4 = 0.5

Gini Impurity:
1 - (0.5^2 + 0.5^2)
= 0.5

Weighted Gini Score for Split 1:
- Total items = 10
- Weights: **small** = 6/10 = 0.6, **large** = 4/10 = 0.4
- Weighted Gini score = 0.6 * 0.444 + 0.4 * 0.5 = 0.2664 + 0.2 = **0.4664**

Weighted Gini Score

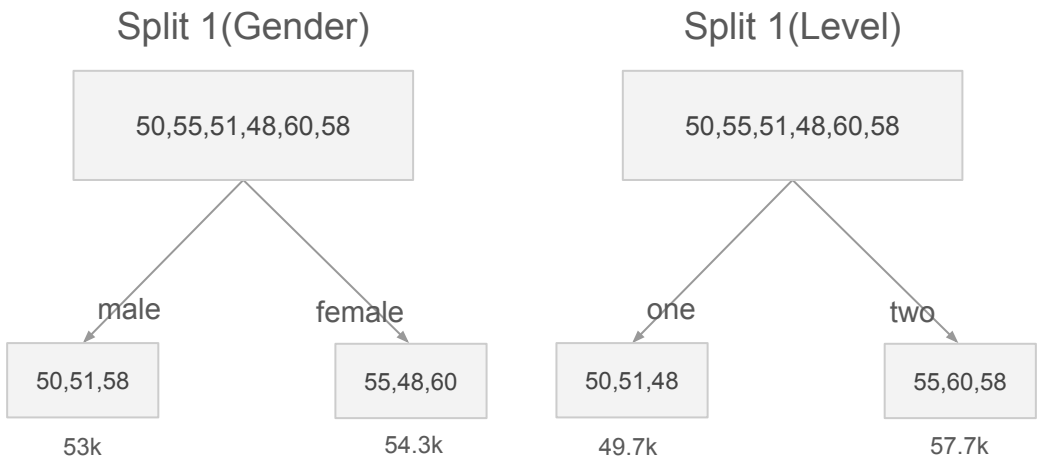


Weighted Gini Score for Split 2:

- Total items = 10
- Weights: **smooth** = 7/10 = 0.7, **rough** = 3/10 = 0.3
- Weighted Gini score = 0.7×0.245+0.3×0=0.1714+0=**0.1714**

Decision Tree : Regression

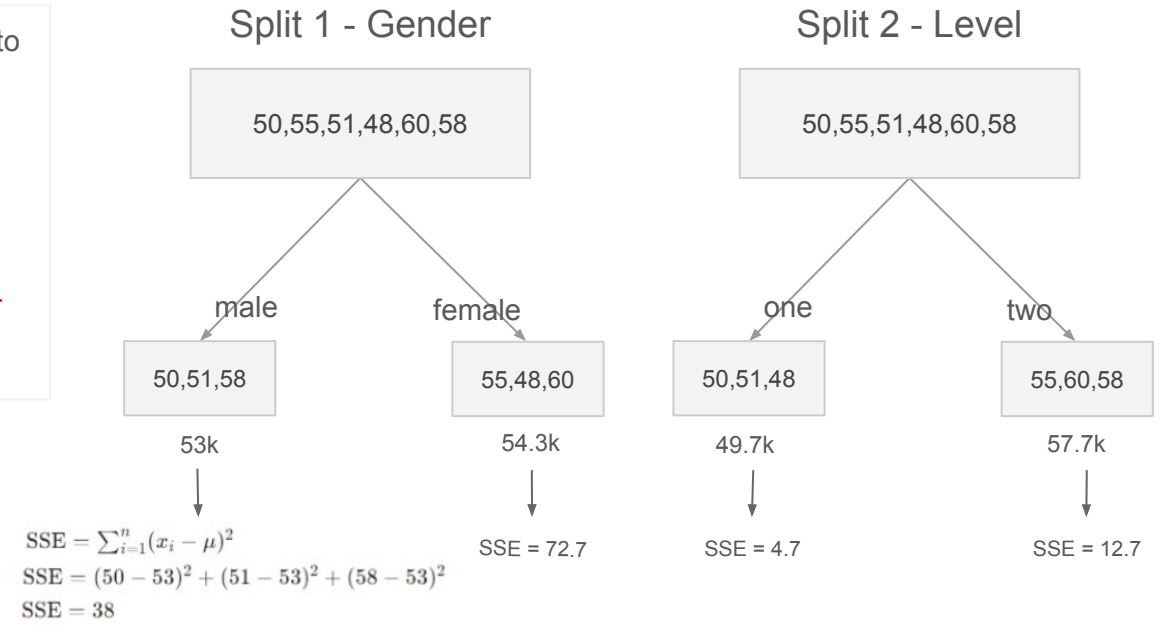
Emp_id	Gender	Level	Salary
1	male	one	50k
2	female	two	55k
3	male	one	51k
4	female	one	48k
5	female	two	60k
7	male	two	58k



The prediction at the leaf node is the mean (average) of the target variable values of the training instances that reach at the leaf node during the training phase.

Splitting Criterion

- In decision tree regression, the goal is to find splits in the data that result in subsets with minimal variability in the target variable.
- The splitting criterion is typically based on minimizing the sum of squared error (SSE).



$$\text{SSE} = \sum_{i=1}^{n}(x_i - \mu)^2$$
$$\text{SSE} = (50-53)^2 + (51-53)^2 + (58-53)^2$$
$$\text{SSE} = 38$$

Total SSE(Split 1) = 38+72.7 = 110.7

Total SSE(Split 2) = 4.7+12.7 = 17.33

Since the total SSE for split 2 is lower than the total SSE for split 1, split 2 will be chosen as the first split.

Strengths and Weaknesses of Decision Trees

Advantages of Decision Trees:

- Easy to understand and interpret, suitable for visual representation.
- Can capture nonlinear relationships between features and the target variable.
- Can handle both numerical and categorical data without requiring feature scaling.

Limitations of Decision Trees:

- Easily overfit noisy data, leading to poor generalization to unseen data.
- small variations in the data can result in different tree structures, making them unstable.
- In classification, may create biased trees if one class dominates the dataset.