Recap – Decision Tree

Rina BUOY

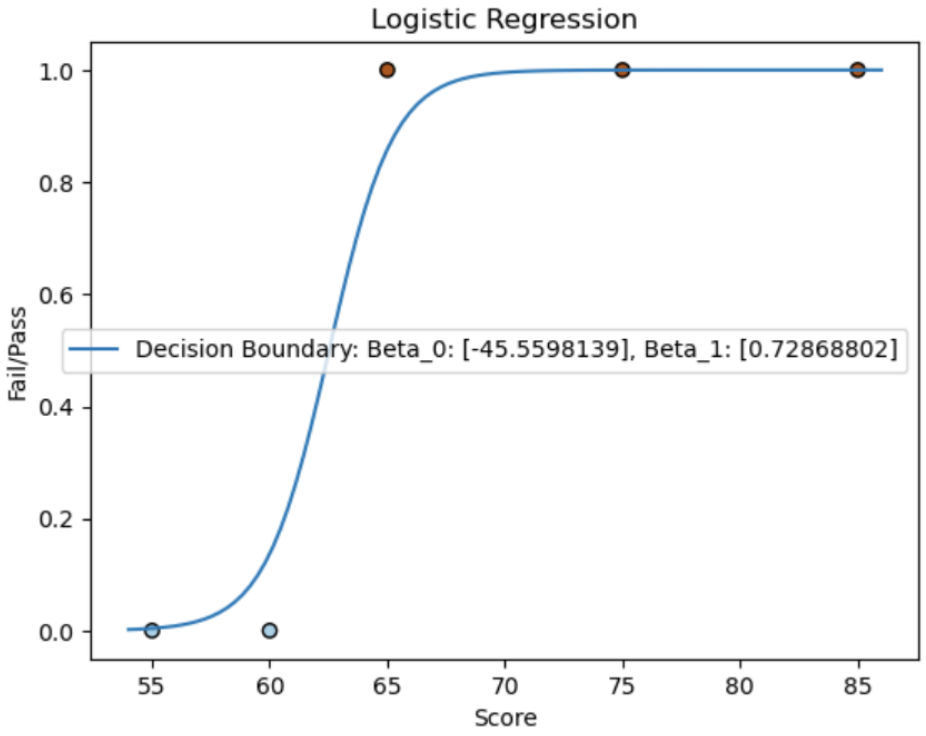
AMERICAN UNIVERSITY
OF PHNOM PENH

STUDY LOCALLY. LIVE GLOBALLY.

Parametric vs. Non-parametric Models?

Exam Score	Pass/Fail
65	Pass
75	Pass
55	Fail
85	Pass
60	Fail

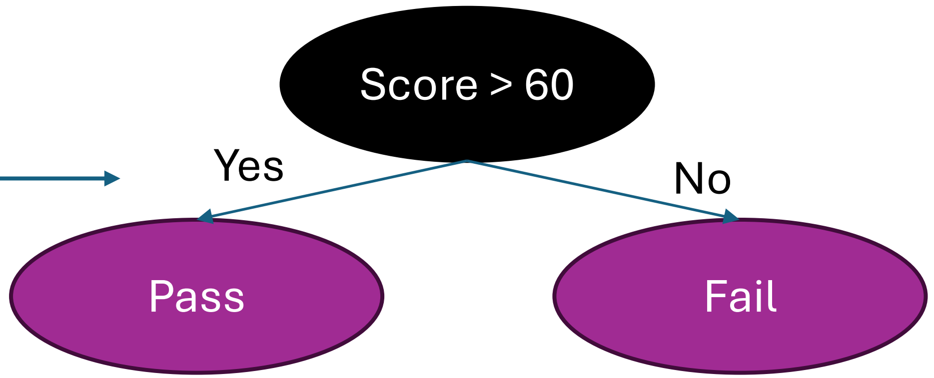
$$Y = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$



What are the parameters ?

Parametric vs. Non-parametric Models?

Exam Score	Pass/Fail
65	Pass
75	Pass
55	Fail
85	Pass
60	Fail



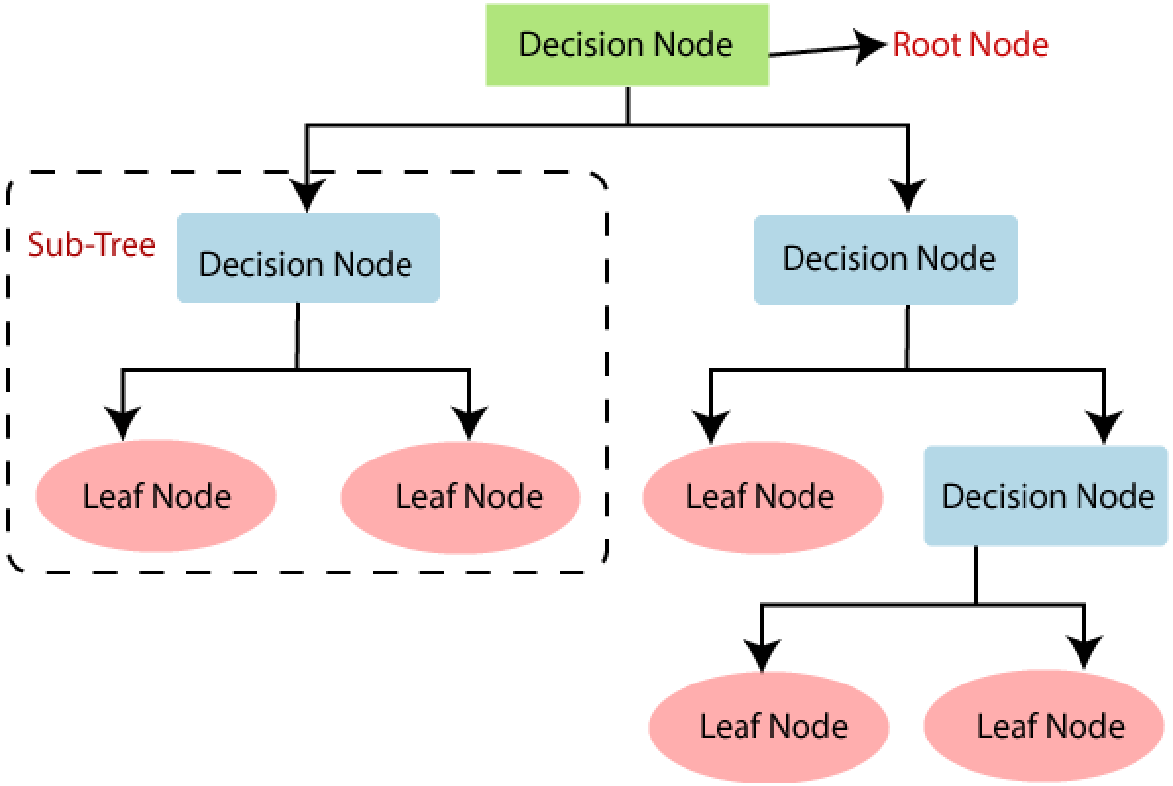
65 – Pass
75 – Pass
85 - Pass

60 – Fail
55 – Fail

What are the parameters ?

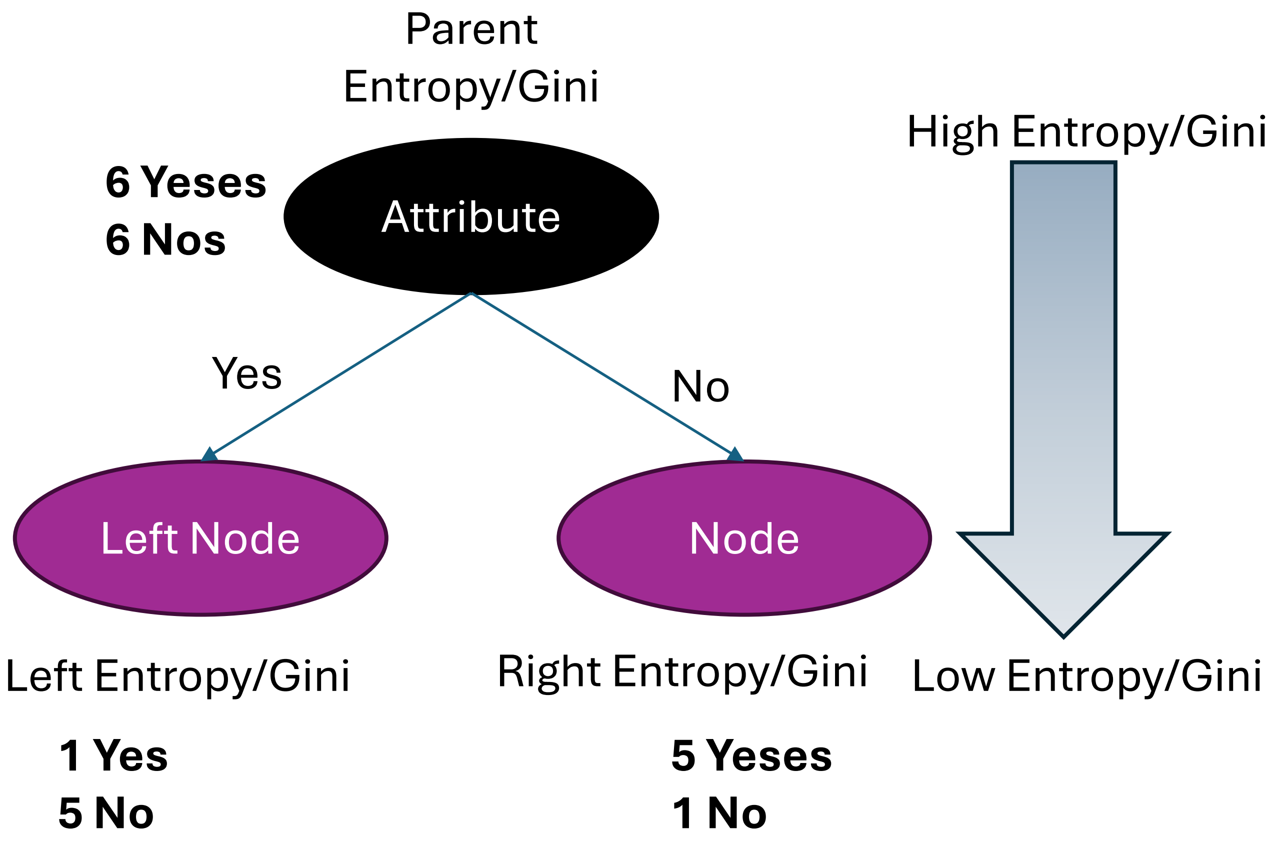
Multiple Features

Outlook	Temperature	Humidity	Windy	Play Tennis?
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No



A LG model will require multiple parameters, while a DT requires multiple branches.

Identify the best attributes for subtrees



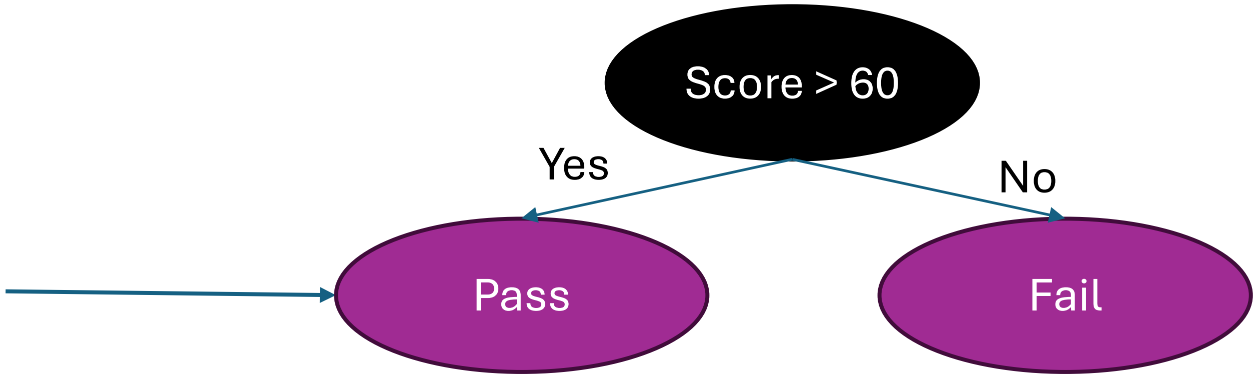
IG = Entropy(parent) – Entropy(children)

Attribute	IG
Attribute-1	…
…	…
Attribute-n	…

Pick the attribute with the highest IG.

Continuous attribute case

Exam Score	Pass/Fail
65	Pass
75	Pass
55	Fail
85	Pass
60	Fail



65 – Pass
75 – Pass
85 - Pass

60 – Fail
55 – Fail

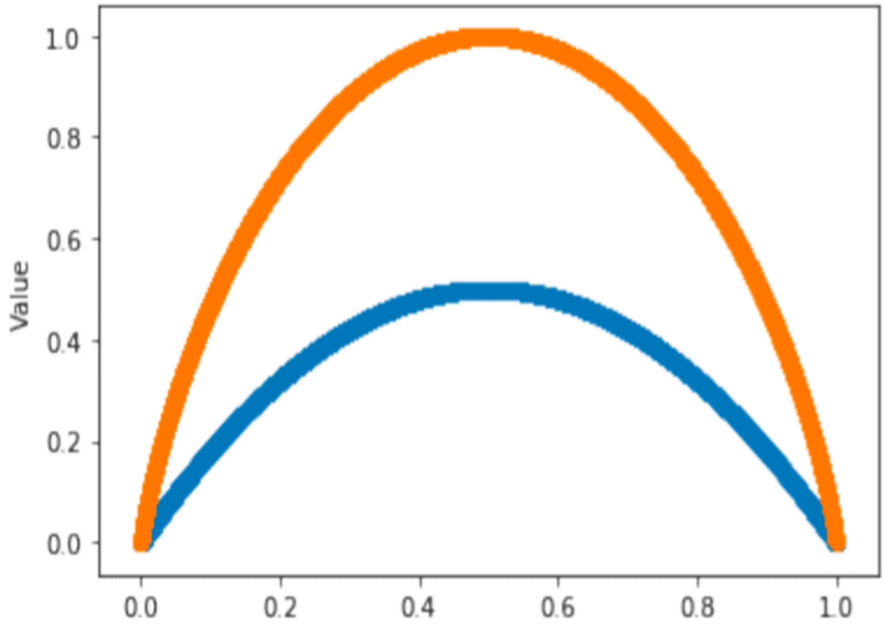
Split Value	60	…	85
IG	…		…

Pick the split point with the highest IG.

Entropy vs. Gini

$$Gini = 1 - \sum_j p_j^2$$

$$Entropy = -\sum_j p_j \log_2 p_j$$



Entropy (orange) vs. Gini (blue) for a binary distribution (i.e., Bernoulli)

Decision Tree Variants

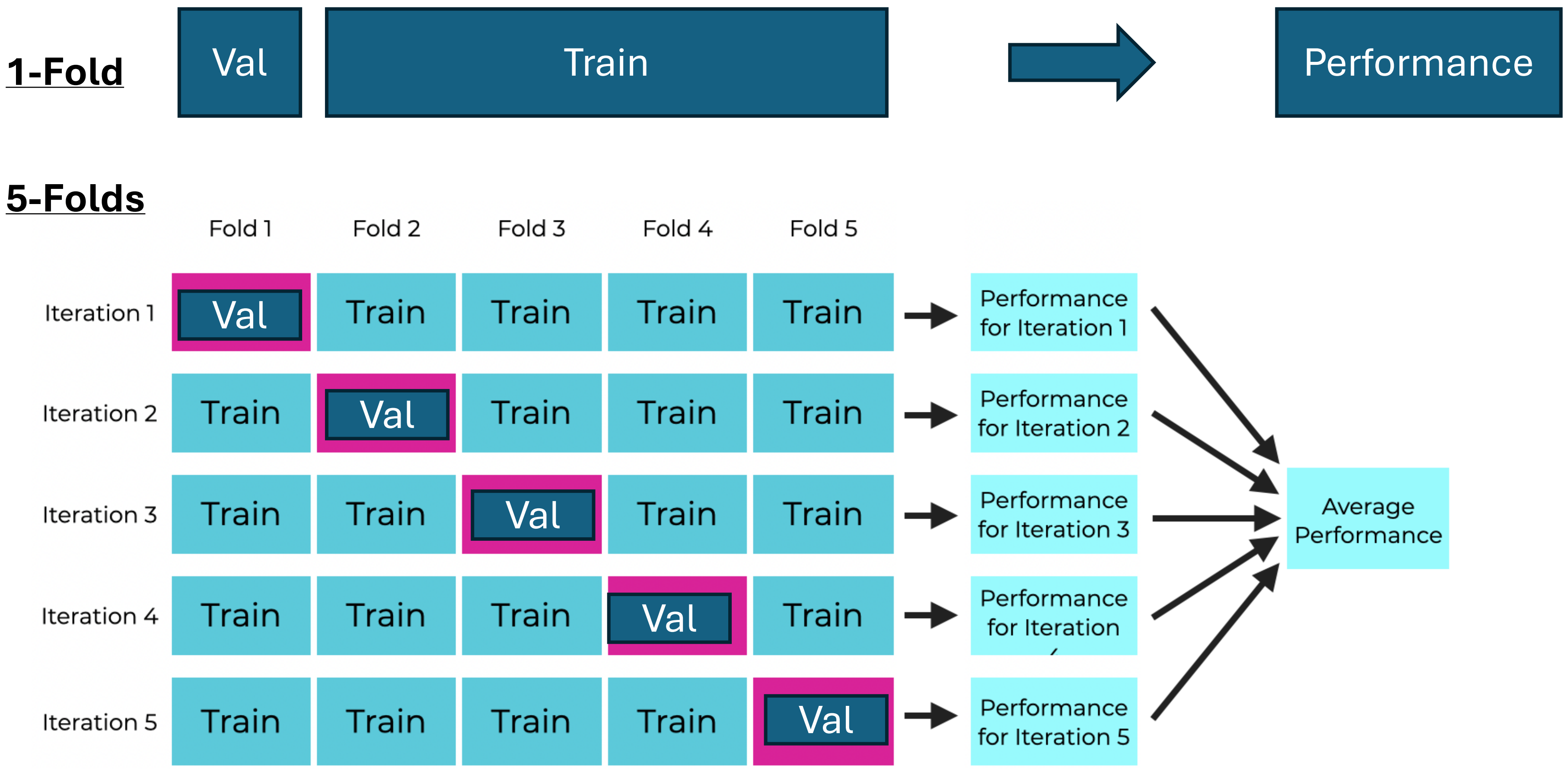
1. **CART (Classification and Regression Trees)**: CART is a versatile decision tree algorithm that can be used for both classification and regression tasks. It works by recursively partitioning the feature space into regions, aiming to minimize impurity (Gini index or entropy) in each split.
2. **ID3 (Iterative Dichotomiser 3)**: ID3 is one of the earliest decision tree algorithms primarily designed for classification tasks. It builds the tree by iteratively selecting the best attribute to split on based on information gain.
3. **C4.5**: C4.5 is an extension of ID3 that addresses some of its limitations. It handles both discrete and continuous attributes, deals with missing values, and incorporates pruning to reduce overfitting.

Hyper-parameters

1. **Criterion**: The function used to measure the quality of a split. Common criteria are "gini" for the Gini impurity and "entropy" for information gain.
2. **Max Depth**: The maximum depth of the decision tree. It limits the number of levels in the tree and helps prevent overfitting.
3. **Min Samples Split**: The minimum number of samples required to split an internal node. It controls the creation of new nodes by requiring a minimum amount of data in each split.
4. **Min Samples Leaf**: The minimum number of samples required to be at a leaf node. It controls the size of terminal nodes and prevents the model from creating nodes with very few samples.
5. **Max Features**: The number of features to consider when looking for the best split. It helps to reduce overfitting and speed up the training process by randomly selecting a subset of features at each split.

- Etc.

Reminder - Cross Validation



Hyper-parameter Search

- Grid search exhaustively searches through all combinations of hyperparameters, while randomized search samples hyperparameter combinations randomly.
- Grid search is more thorough and guarantees finding the optimal combination within the specified grid, but it can be computationally expensive.
- Randomized search is more efficient in terms of computational resources and can handle a larger hyperparameter space, but it may not guarantee finding the optimal combination.
- Grid search is preferred when computational resources are not a constraint and the hyperparameter space is small. Randomized search is preferred for large hyperparameter spaces or when computational resources are limited.