Decision Tree

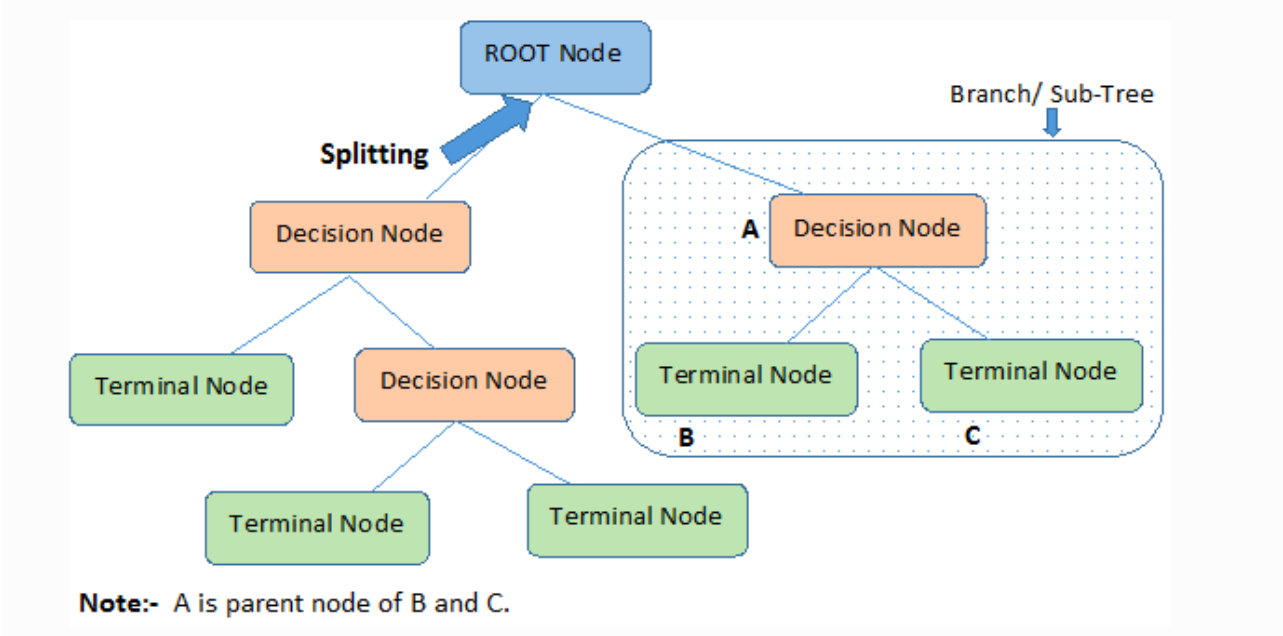
ماهي Decision Tree:
هي طريقة تعلم Supervised learning تستخدم لحل مشاكل classification و regression. والهدف منها إنشاء نموذج (model) يتنبأ بقيمة متغير (value of a target variable) من خلال تعلم قواعد قواعد الإقرار (decision rules) البسيطة التي يتم استنتاجها من data features.

مميزات Decision Tree:
- **سهل الفهم والتفسير. يمكن رسم Decision Tree على شكل رسوم بيانية.**
- **يتطلب القليل من إعداد البيانات.**
- **قادرة على التعامل مع البيانات numerical and categorical.**

عيوب Decision Tree:
- **لا يمكن استخدامه في البيانات الضخمة.**
- **غالبًا ما تتضمن شجرة القرار وقتًا أطول لتدريب النموذج.**

بناء Decision Tree:
تشبة هيكل الشجرة(tree structure) ، وقد تم بناء DT عن طريق تقسيم البيانات التدريبية(Training Data) بشكل متكرر إلى عينات أصغر وأصغر.

لدينا ثلاثة أنواع من العقد(nodes) في DT:
- **الـ Root node**
- **الـ Decision nodes**
- **الـ Leaf nodes**



طريقة عمل الـDecision Tree:

لابد ان نضع بعين الاعتبار ان DT عبارة rules, و بدايةً من root node يتم تحديدها بناءً على نتائج Attribute Selection Measure (ASM), ويتم تكرار عملية (ASM) حتى الى ان نصل الى Leaf node.

الإستراتيجيات للإنقسام في DT توثر بشكل كبير على دقة الشجرة (accuracy)

تختلف المعايير بالنسبة الى classification and regression trees.

ماهي Attribute Selection Measure (ASM) ؟

بإختصار هو يستخدم لتحديد معيار التقسيم الذي يقسم البيانات إلى افضل طريقة ممكنة.

له ثلاث انواع:

- **Gini index**

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

Gini Index

- **information Gain**
  - Entropy:

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i),$$

Entropy

الفرع(branch) الذي يحتوي على entropy من الصفر هو leaf node ويحتاج الفرع الذي يحتوي على entropy أكثر من الصفر إلى مزيد من الانقسام.

Play Golf	
Yes	No
9	5

Entropy(PlayGolf) = Entropy (5,9)
= Entropy (0.36, 0.64)
= - (0.36 $\log_2$ 0.36) - (0.64 $\log_2$ 0.64)
= 0.94

multiple attributes:

$$E(T,X) = \sum_{c \in X} P(c)E(c)$$

		Play Golf		
		Yes	No	
Outlook	Sunny	3	2	5
	Overcast	4	0	4
	Rainy	2	3	5
				14

E(PlayGolf, Outlook) = **P**(Sunny)*__E__(3,2) + **P**(Overcast)*__E__(4,0) + **P**(Rainy)*__E__(2,3)

= (5/14)*0.971 + (4/14)*0.0 + (5/14)*0.971

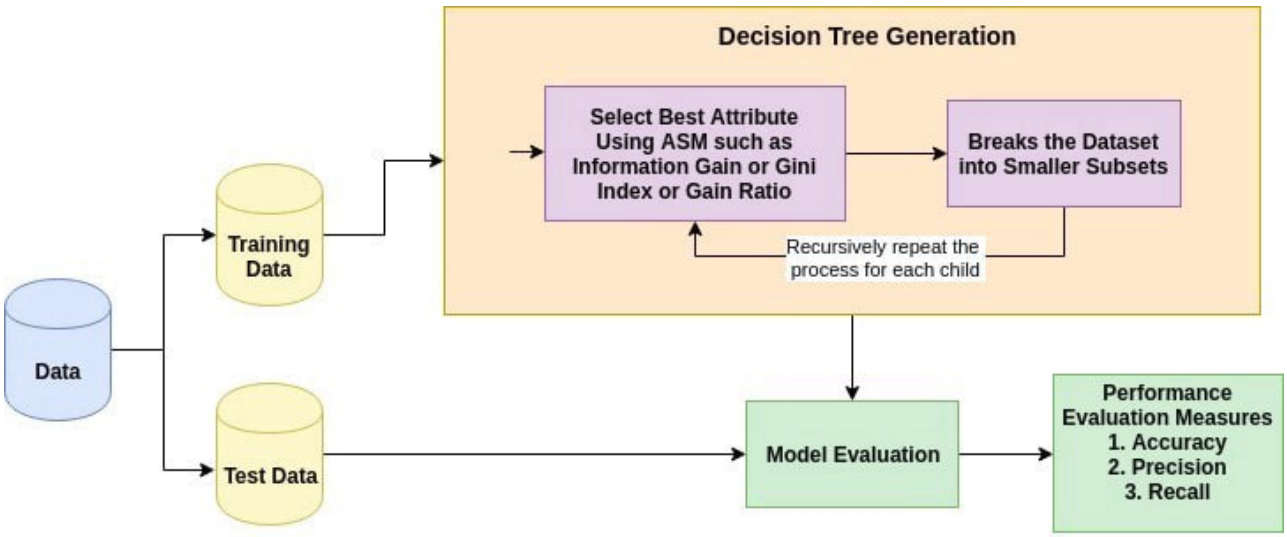
= 0.693

IG:

$$Information\ Gain = Entropy(before) - \sum_{j=1}^{K} Entropy(j, after)$$

- **Gain ratio**

$$Gain\ Ratio = \frac{Information\ Gain}{SplitInfo} = \frac{Entropy\ (before) - \sum_{j=1}^{K} Entropy(j, after)}{\sum_{j=1}^{K} w_j \log_2 w_j}$$



أنواع Decision Tree Algorathims:

- **DeacisionTreeClassification()**
- **DeacisionTreeRegression()**