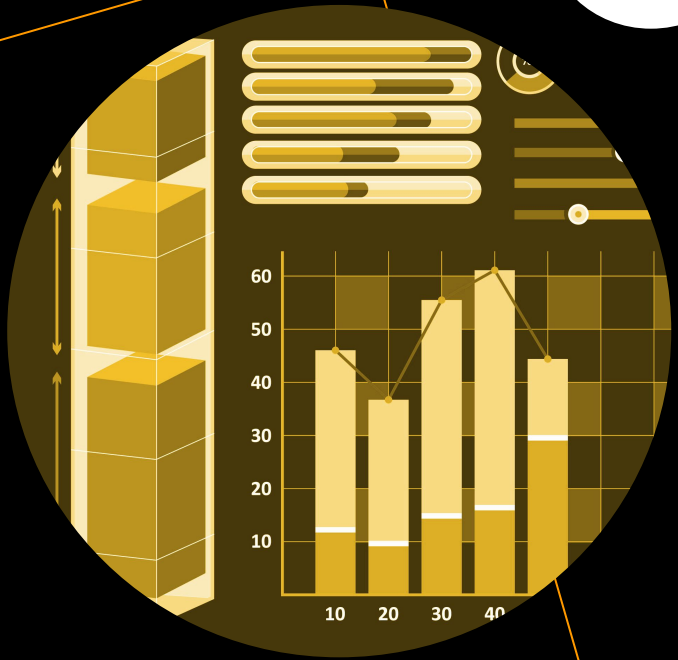
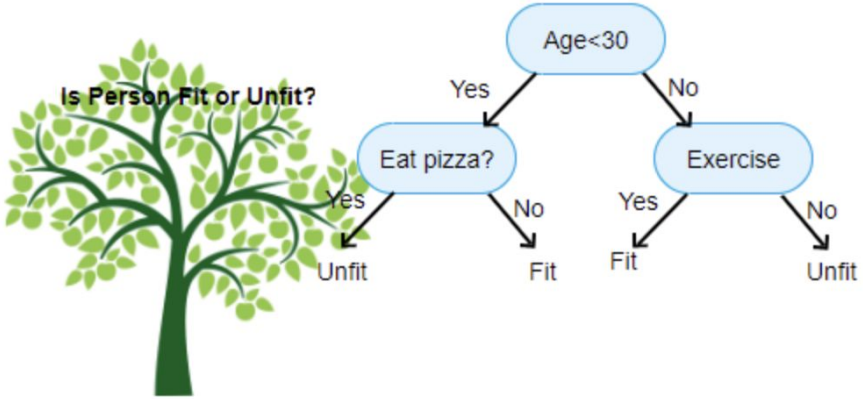
Decision Trees

Nhung Le



Decision Tree



- **Mục tiêu:**
  - Đi đến một kết luận dựa vào chia nhóm các chi tiết thông tin đã có.

- **Phương pháp:**
  - Đặt mục tiêu: vd. tối đa Information Gain
  - Với mỗi nhánh (vd. feature), đi qua các giá trị của nhánh trong dự liệu ban đầu
  - Chọn giá trị cut-off giúp đạt mục tiêu (ví dụ tối đa Information Gain)

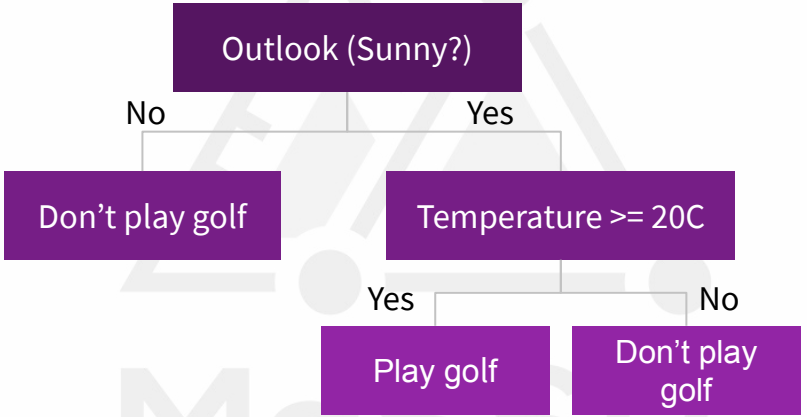
Nguồn:
https://www.aitimejournal.com/@akshay.chavan/a-comprehensive-guide-to-decision-tree-learning

Decision Tree

MaSSP

Bài toán: ***xét xem có nên đi chơi golf không***

1. Đặt mục tiêu (*objective*): tối đa ***Information Gain*** (***IG*** *- giá trị thông tin nhận được)*
2. Chọn nhánh *(feature)*: chọn *feature* mang lại nhiều giá trị thông tin nhất *(vd.* ***maximize information gain****)* thì sẽ đánh giá *feature* đấy trước.
3. Chọn ***cutoff value*** *(giá trị phân chia)* ở mỗi nhánh *(feature)*
   - **Categorical** *(giá trị phân loại):* vd. Thời tiết (mưa hay nắng). Nếu mưa -> không chơi golf, nếu nắng -> đánh giá tiếp nhiệt độ
   - **Numerical** *(giá trị số)*: Vd. Nhiệt độ, có các giá trị từ 15 đến 30 -> đi qua từng giá trị và xem tại giá trị nào khi phân chia nhóm sẽ mang lại **IG** lớn nhất để làm ***cutoff value***.



Note:
- Outlook: Thời tiết
- Temperature: Nhiệt độ

Nguồn: https://www.aitimejournal.com/@akshay.chavan/a-comprehensive-guide-to-decision-tree-learning

Decision Tree - Information Gain

Maximizing Information Gain = Minimizing Entropy

$$Gain(T, X) = Entropy(T) - Entropy(T, X)$$

G(PlayGolf, Outlook) = **E**(PlayGolf) – **E**(PlayGolf, Outlook)

= 0.940 – 0.693 = 0.247

Nguồn: information gain

Decision Tree - Entropy

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

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

Nguồn: entropy

Decision Tree - Entropy

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

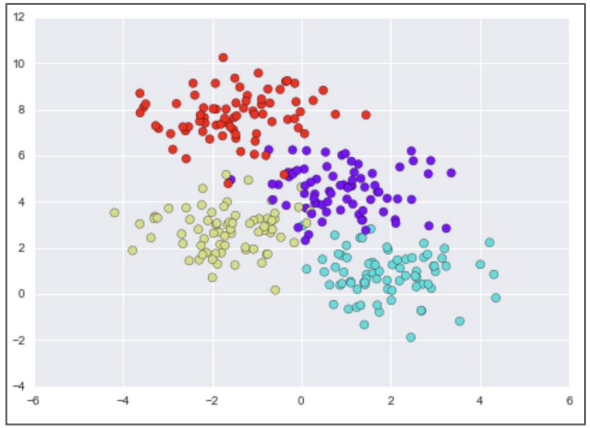
		Play Golf		
		Yes	No	
Outlook	Sunny	3	2	5
	Overcast	4	0	4
	Rainy	2	3	5
				14

$$\begin{aligned}\mathbb{E}(PlayGolf, Outlook) &= \mathbb{P}(Sunny) * \mathbb{E}(3,2) + \mathbb{P}(Overcast) * \mathbb{E}(4,0) + \mathbb{P}(Rainy) * \mathbb{E}(2,3) \\ &= (5/14)^{*}0.971 + (4/14) * 0.0 + (5/14) * 0.971 \\ &= 0.693\end{aligned}$$

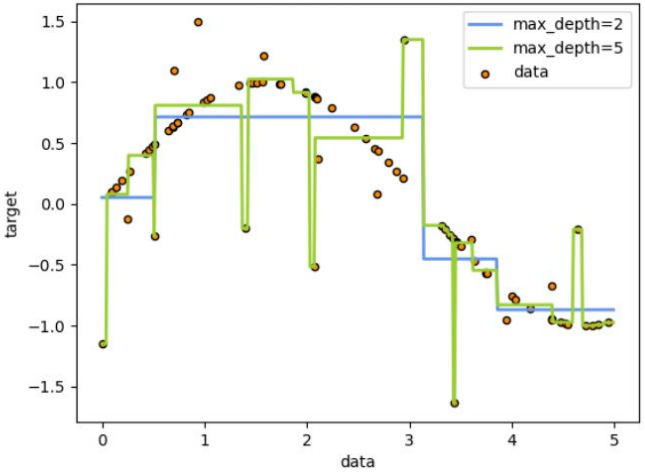
Note:
- Outlook: Thời tiết
- Overcast: Lượng mây

Nguồn: entropy

Tree Models - Ứng dụng



Decision Tree Classification



Decision Tree Regression

Tree models được sử dụng rộng rãi trong bài toán phân loại hay dự đoán khi ***không có linear relationship*** giữa X và Y

Ứng dụng:

- Dự đoán giá bất động sản
- Dự đoán giá cổ phiếu
- Nhận dạng hành vi ăn cắp thẻ tín dụng

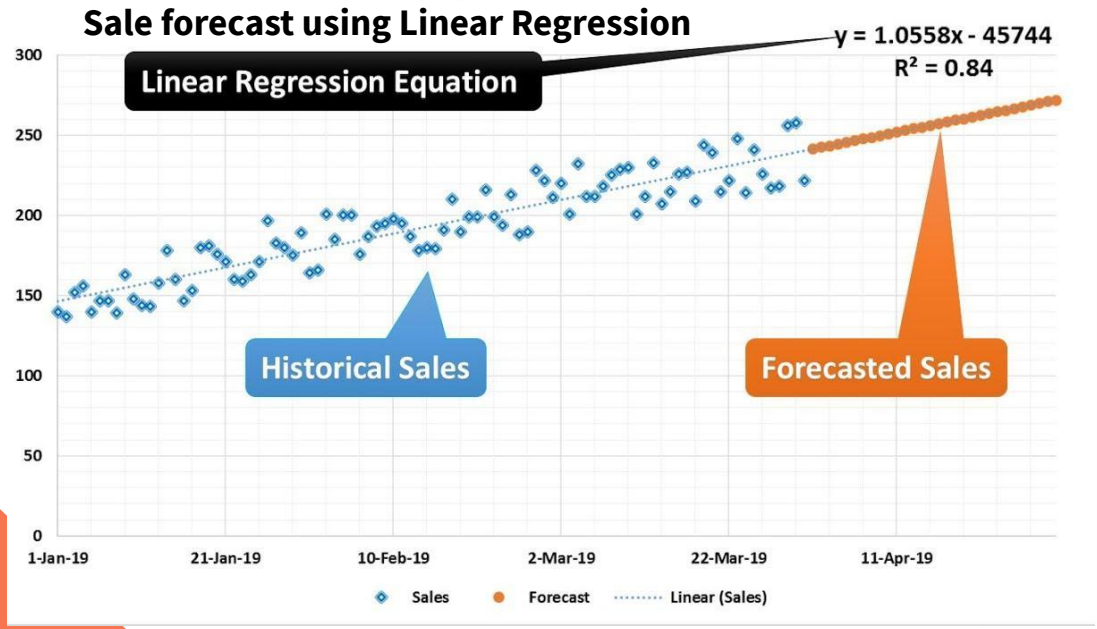
Nguồn: application

MaSSP

MaSSP

Appendix

Linear Regression - Ứng dụng



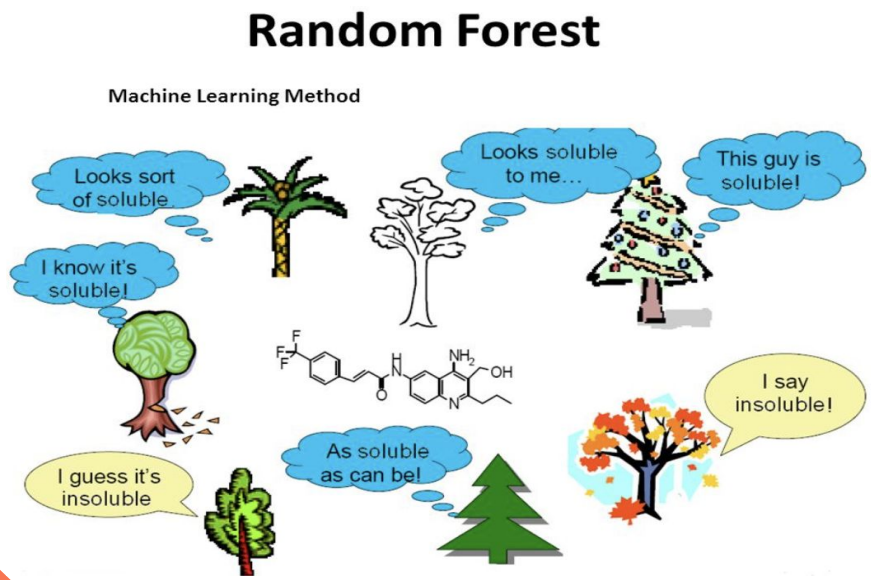
- Nghiên cứu thị trường
- Dự đoán doanh số bán hàng, tỉ suất người xem một show truyền hình
- Dự báo thời tiết

Nguồn: application

Linear Regression vs. Tree Models

	Linear Regression	Tree Models
PROs	- Simple - Computational efficiency - Not prone to overfitting	- Interpretability - Complicated relationship (rule based) - Trained on a small dataset
CONs	- Multiple constraints (e.g., Linearity assumption, independence of variables) - Less interpretable - Need large data set - Difficult to have a linear relationship in a large dataset	- Prone to overfitting - Affected by noise

Random Forest

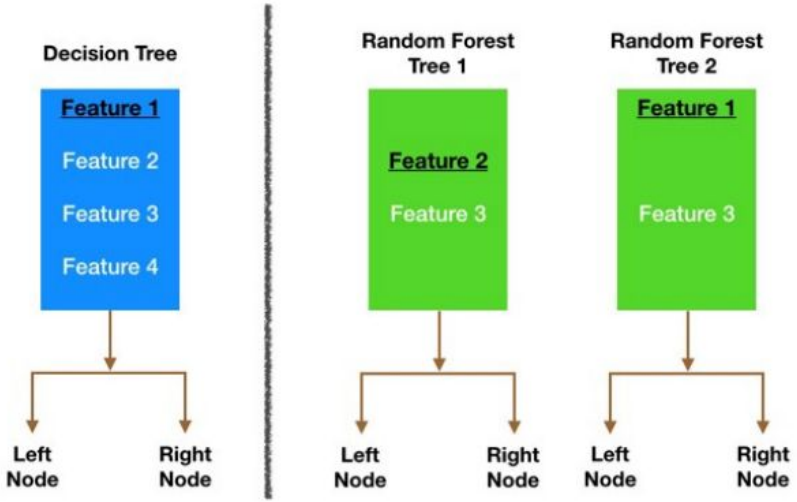


Trong rừng cây, mỗi cây sẽ đưa ra một quyết định.

Kết quả cuối cùng có thể tính bằng cách **1/ lấy giá trị trung bình** hoặc **2/ quyết định của số đông.**

Nguồn: random forest

Random Forest - Randomness



Node splitting in a random forest model is based on a random subset of features for each tree.

- **Random sampling:** *Mỗi cây trong Random Forest sẽ được huấn luyện (train) với* ***một nhóm data ngẫu nhiên***
- **Random feature selection:** *mỗi cây trong Random Forest sẽ sử dụng một* ***nhóm features (subset of features) khác nhau*** *để đảm bảo sự độc lập (independence)*

Nguồn: towarddatascien