Bài giảng môn học:

Kỹ nghệ tri thức và học máy (7080510)

CHƯƠNG 3:
HỌC CÓ GIÁM SÁT – PHẦN 1
(Supervised Learning)

Giảng viên: Đặng Văn Nam

Email: dangvannam@humg.edu.vn

Nội dung chương 3 – phần 2

3.2 Cây quyết định (Decision Tree)

a. Giới thiệu cây quyết định và thuật toán cây quyết định trong phân lớp.

b. Thực hành xây dựng mô hình phân lớp sử dụng thư viện Sklearn (Tập Iris – Tập Titanic)

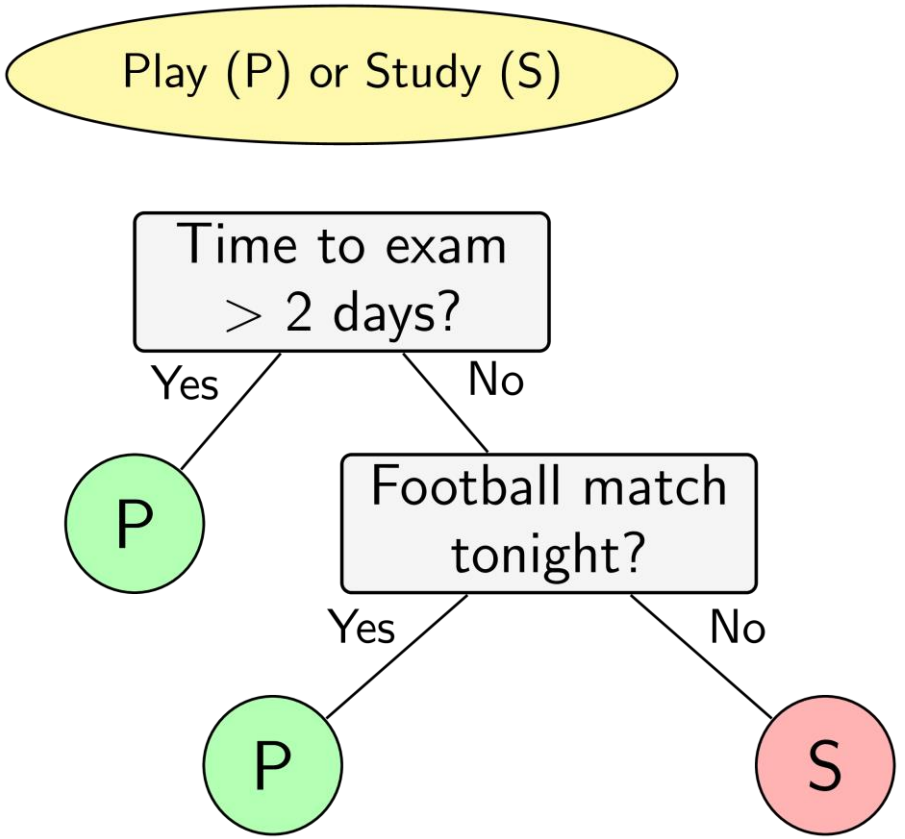
c. Bài thực hành số 7

3.2 Cây quyết định (Decision Tree)

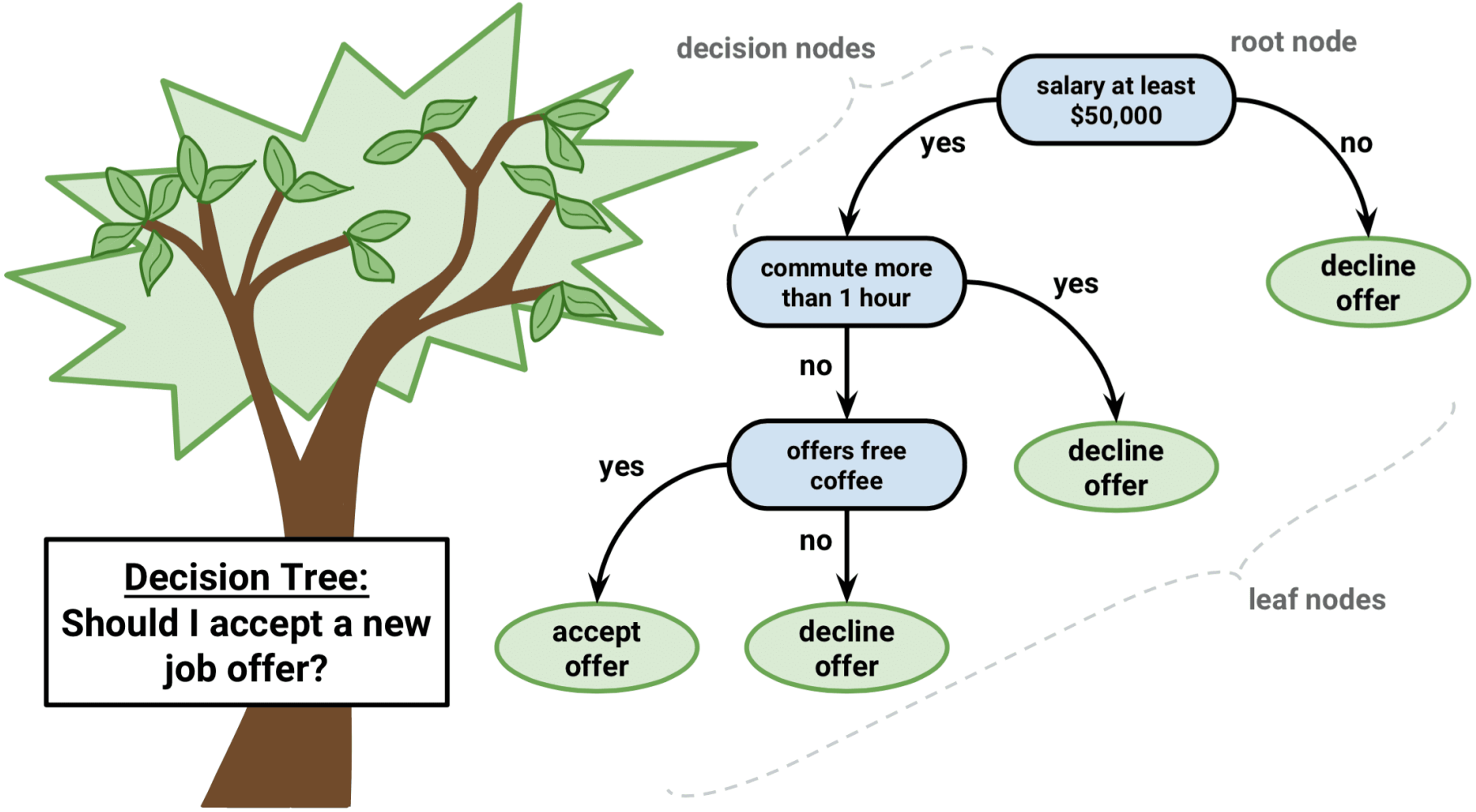
Giới thiệu cây quyết định:

- Việc quan sát, suy nghĩ và ra các quyết định của con người thường được bắt đầu từ các câu hỏi. Machine learning cũng có một mô hình ra quyết định dựa trên các câu hỏi. Mô hình này có tên là *cây quyết định* (*decision tree*).

- Decision tree là một mô hình học có giám sát, có thể được áp dụng vào cả hai bài toán classification và regression.

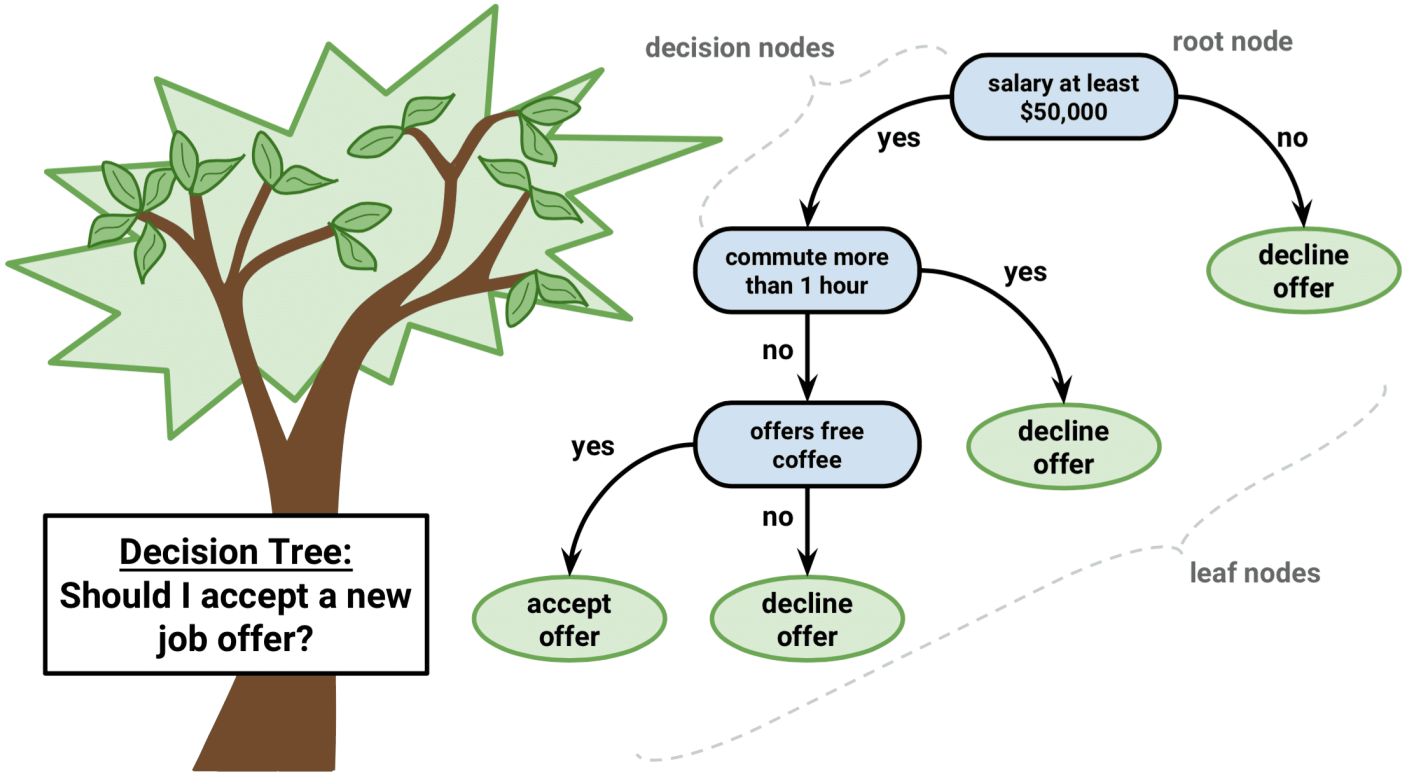


Cây quyết định:



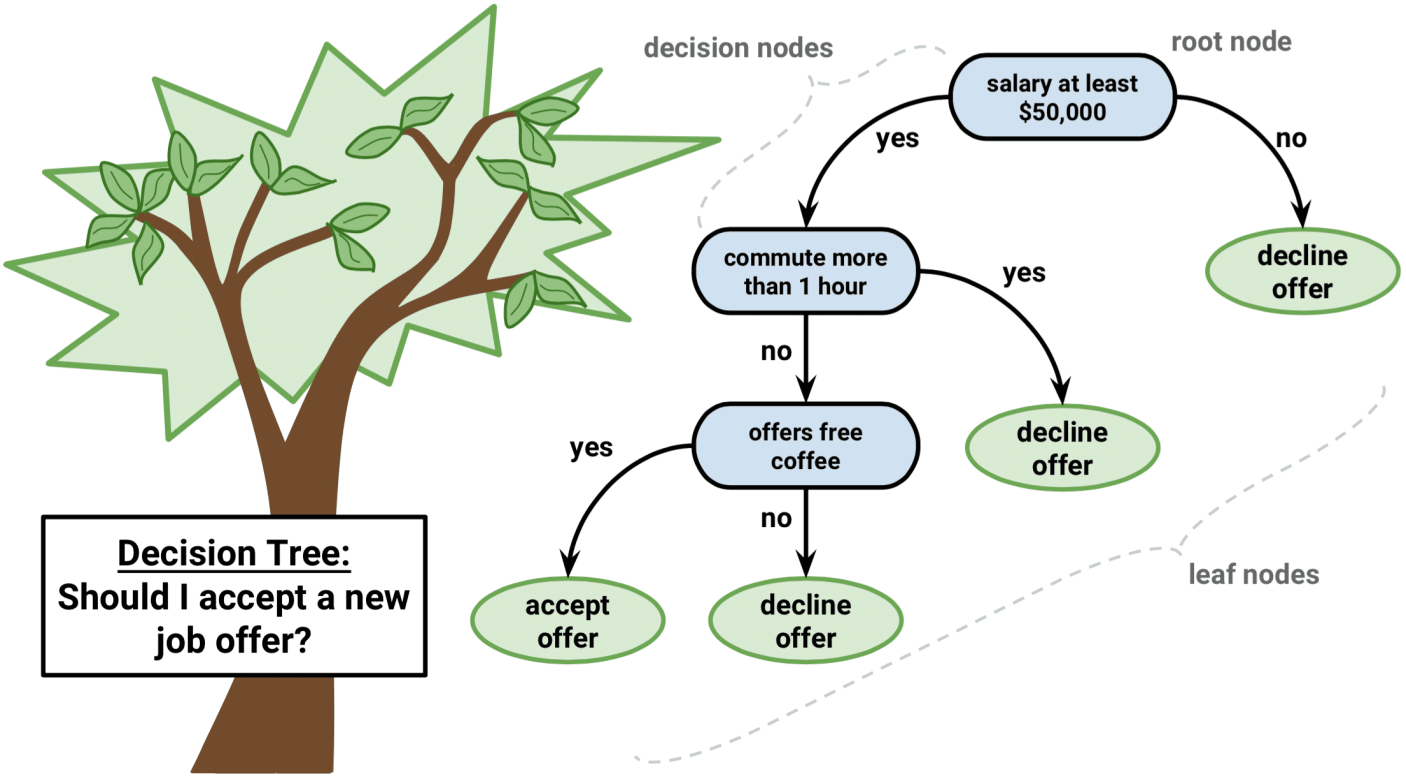
Việc xây dựng một decision tree trên dữ liệu huấn luyện cho trước là việc đi xác định các *câu hỏi* và *thứ tự của chúng*

Cây quyết định:



- Dùng cấu trúc cây để đưa ra một hàm phân lớp cần học (hàm mục tiêu có giá trị rời rạc)
- Một cây quyết định có thể được biểu diễn (diễn giải) bằng một tập các luật IF-THEN (dễ đọc và dễ hiểu)
- Được áp dụng thành công trong rất nhiều các bài toán ứng dụng thực tế

Cây quyết định:



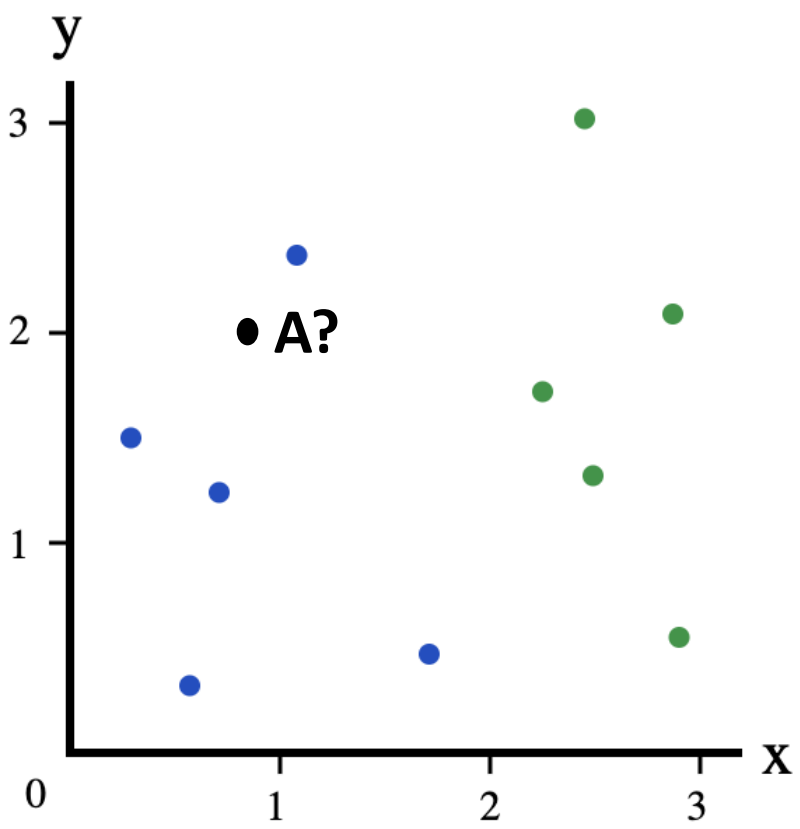
- **Root node (Nút gốc)**: Chứa toàn bộ dữ liệu mẫu. Dữ liệu này được chia thành các nhóm nhỏ hơn.
- **Decision node (Nút quyết định)**: Là các nút tiếp tục được phân chia.
- **Leaf node (Nút lá)**: Là các nút không được phân chia.

Thuật ngữ:

- **Splitting (Phân nhóm)**: Là quá trình chia các nhóm thành các nhóm nhỏ hơn.
- **Prunning (Tỉa cành)**: Loại bỏ một số nút phụ của cây.
- **Sub-Tree (nhánh)**: Là một bộ phận của cây.
- **Parent and Child node (Nút cha và nút con)**: Nút bị chia thành các nút phụ gọi là nút cha, các nút phụ của nút cha gọi là nút con.

Ví dụ cây quyết định:

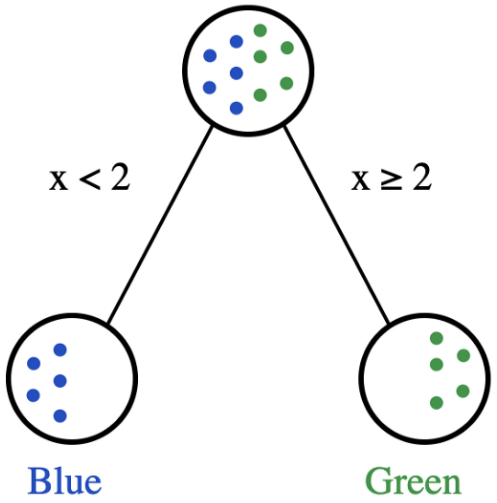
- Cho tập dữ liệu gồm 10 mẫu, thuộc 2 lớp:
  - Lớp blue: 5 mẫu
  - Lớp green: 5 mẫu

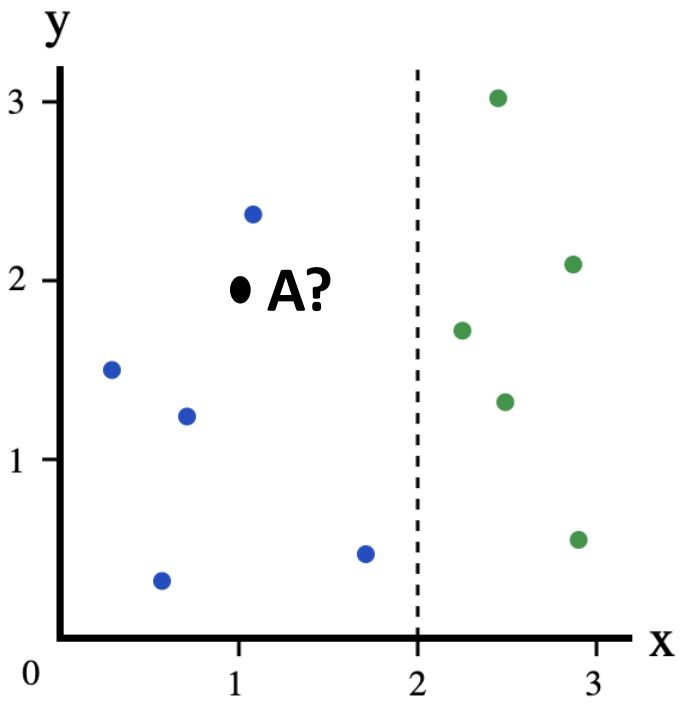


Có điểm dữ liệu mới với giá trị thuộc tính **A(x = 1, y = 2)** màu của điểm này nên là gì? (nên phân vào lớp nào?)

Ví dụ cây quyết định:

- Đây là cây quyết định đơn giản với một node phân loại kiểm tra xem x<2?

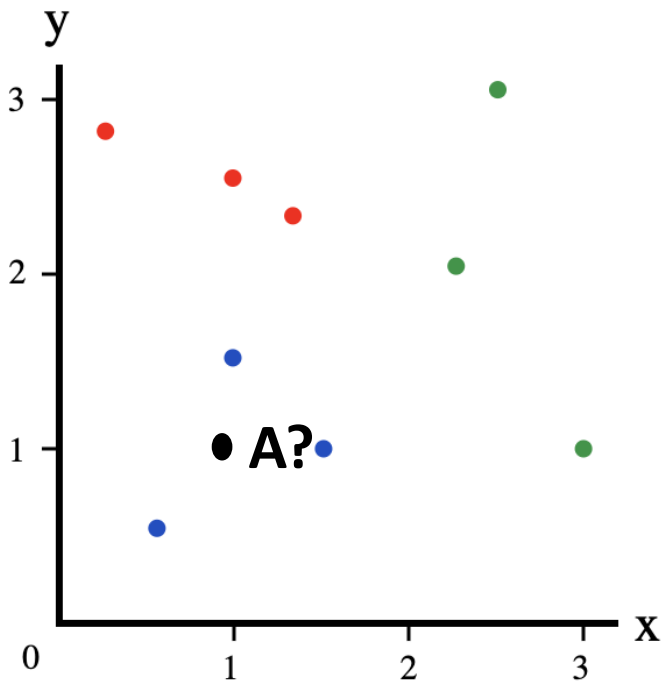




Nếu kiểm tra x < 2, chúng ta lấy nhánh trái và gán nhãn blue, nếu kiểm tra không đúng (x ≥ 2), chúng ta lấy nhánh phải và gán nhãn green.

Ví dụ cây quyết định:

- Ví dụ tập dữ liệu có 9 mẫu gồm 3 lớp:
  - Lớp Blue: 3 mẫu
  - Lớp Green: 3 mẫu
  - Lớp Red: 3 mẫu

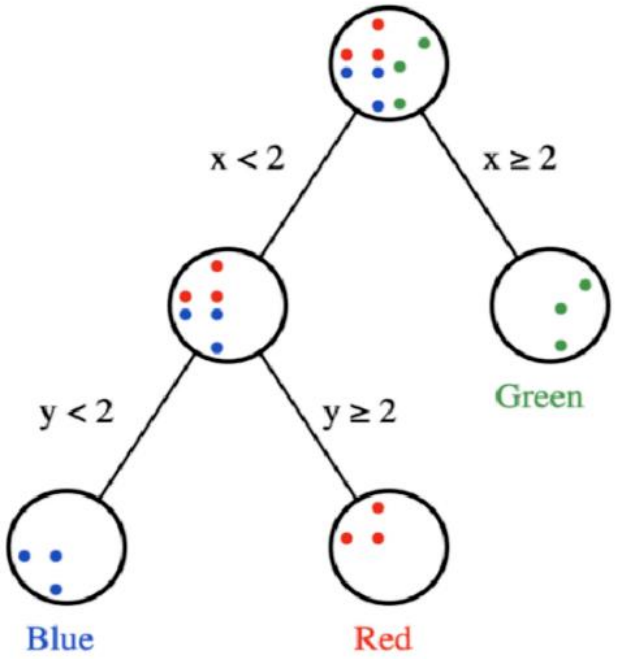


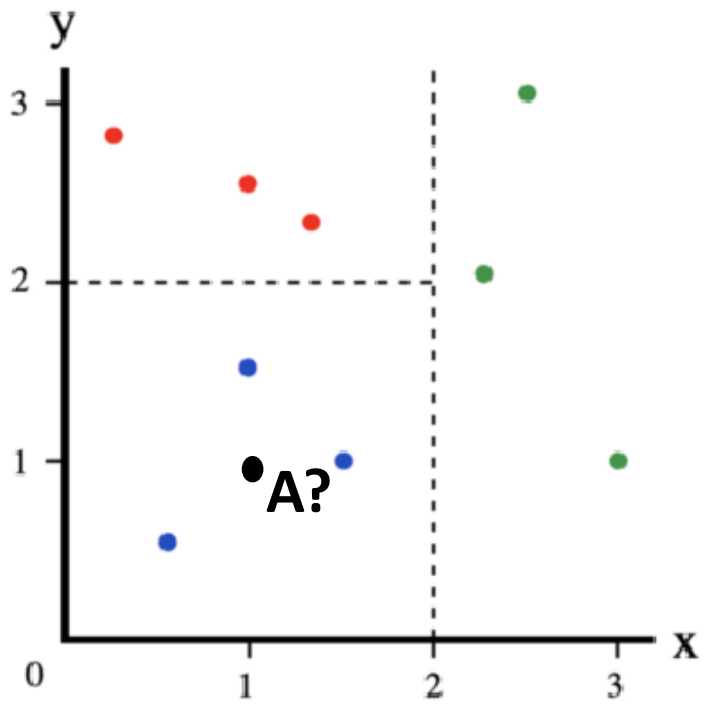
Cây quyết đĩnh cũ không hiệu quả, với mẫu dữ liệu mới A(x=1,y=1)

- Nếu x ≥ 2, chúng ta có thể vẫn tự tin phân loại vào Green
- Nếu x < 2, chúng ta không thể phân loại ngay vào Blue, nó cũng có thể vào Red.

Ví dụ cây quyết định:

- Chúng ta cần thêm node quyết định vào cây quyết định





Đó là ý tưởng chính của cây ra quyết định?

Các độ đo lựa chọn thuộc tính

- Một độ đo lựa chọn thuộc tính là một phương pháp tiên nghiệm (heuristic) để lựa chọn tiêu chí phân chia để phân tách tốt nhất phần dữ liệu D đã cho
- Một cách lý tưởng
  - Mỗi phần được chia ra nên thuần nhất
  - Mỗi phần thuần nhất là phần chứa các mẫu cùng thuộc một lớp
- Các độ đo phân chia thuộc tính (các luật phân chia)
  - Xác định các mẫu ở một node được phân chia thế nào
  - Đưa ra cách xếp hạng các thuộc tính
  - Thuộc tính với điểm cao nhất được lựa chọn
  - Xác định một điểm phân chia hoặc một tập con phân chia
- Các phương pháp
  - **Information gain (Entropy)**
  - **Gain ratio**
  - **Gini Index**

Các độ đo lựa chọn thuộc tính

Gini Index

$$I_G = 1 - \sum_{j=1}^{c} p_j^2$$

p_j: proportion of the samples that belongs to class c for a particular node

Entropy

$$I_H = -\sum_{j=1}^{c} p_j log_2(p_j)$$

p_j: proportion of the samples that belongs to class c for a particular node.

*This is the the definition of entropy for all non-empty classes ($p \neq 0$). The entropy is 0 if all samples at a node belong to the same class.

Thư viện Sklearn sử dụng 2 độ đo: Gini (Mặc định), Entropy

Ưu – Nhược điểm của Cây quyết định

- Cây quyết định có tốc độ học tương đối nhanh so với các phương pháp khác
- Đơn giản và dễ hiểu các luật phân loại trong cây ra quyết định
- Information Gain, Gain Ratio, và Gini Index là những phương pháp lựa chọn thuộc tính thông dụng nhất
- Cắt tỉa cây là cần thiết để loại bỏ những nhánh không tin cậy

- **Ưu điểm:**
  - Dễ hiểu: Cây biểu diễn trực quan
  - Hữu ích: Xác định được các biến quan trọng
  - Phi tham số: không cần giả định về phân phối
  - Không phục thuộc vào dữ liệu: Có thể áp dụng cả dữ liệu phân loại và liên tục
- **Nhược điểm:**
  - Dễ bị quá khớp (overfitting)

Một số vấn đề về xây dựng cây?

- Gốc của cây chứa tất cả dữ liệu, các node trung gian, các node lá.
- Các node được chia nhị phân:
  - Chọn một thuộc tính X_i
  - Chọn một điểm chia t_j
- Độ sâu của cây?
- Số mẫu trong mỗi node lá?
- Số node lá lớn nhất?

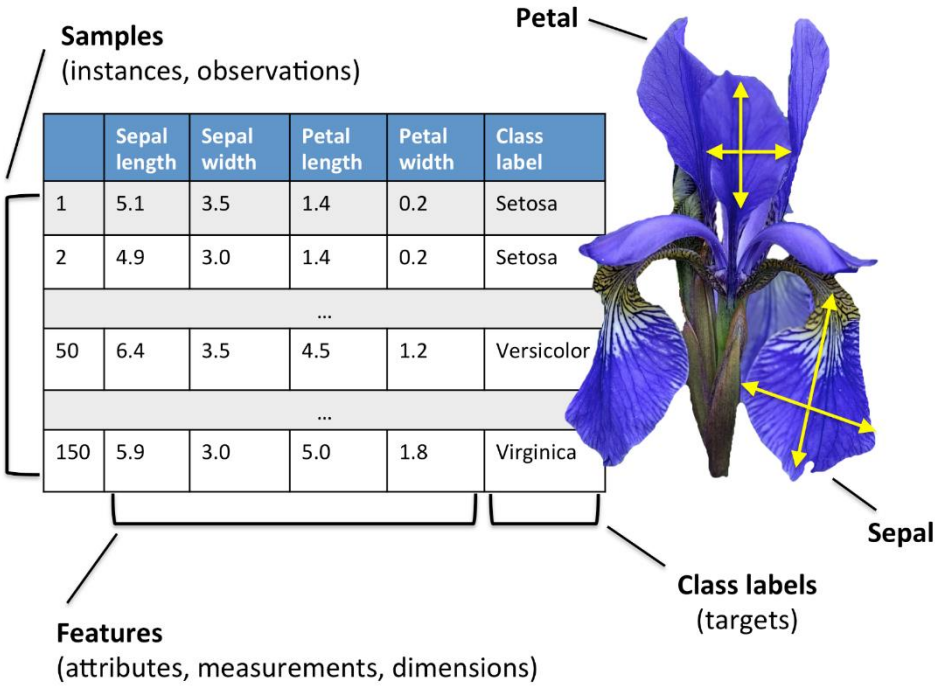
Ví dụ 1:
Phân lớp hoa lan

Ví dụ: Phân lớp hoa lan với Decision tree

- Tập dữ liệu bao gồm 150 mẫu về thông số chiều rộng, chiều dài của lá hóa và cánh hoa của 3 loại hoa Lan

IRIS DATASET

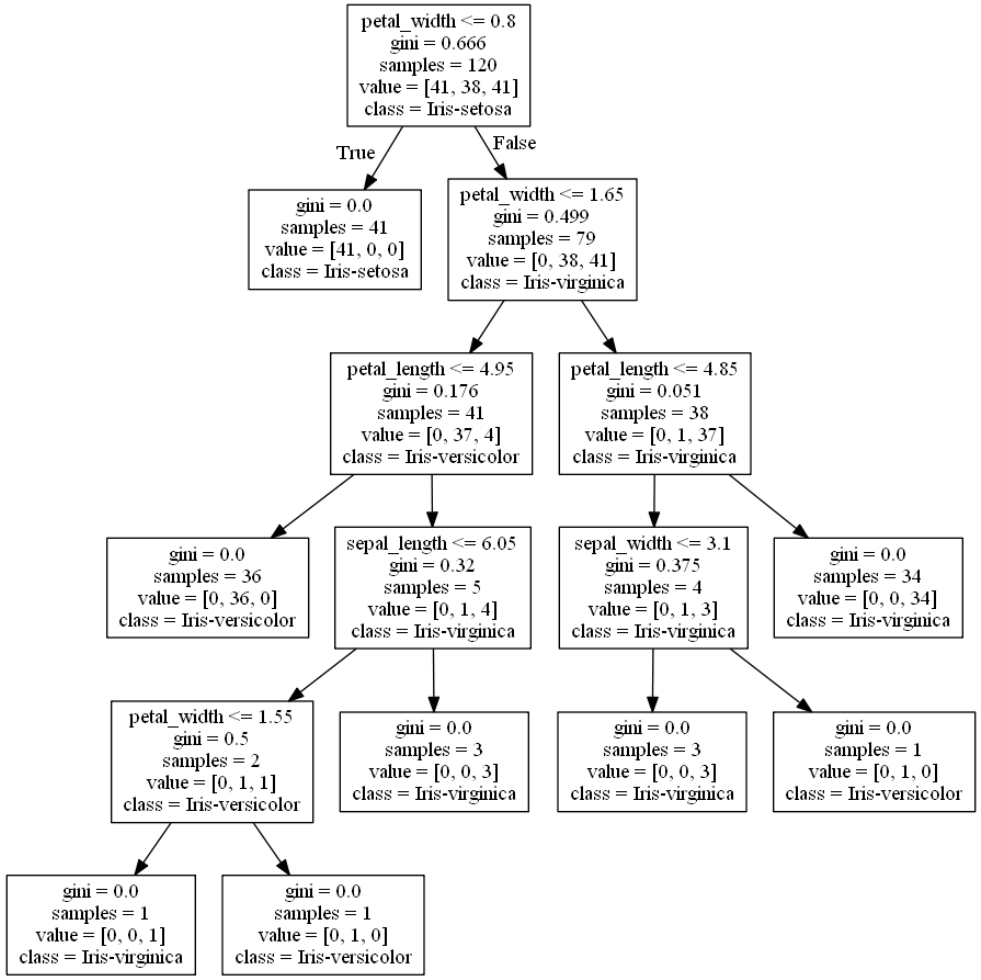
Classes	3
Samples per class	50
Samples total	150
Dimensionality	4
Features	real, positive



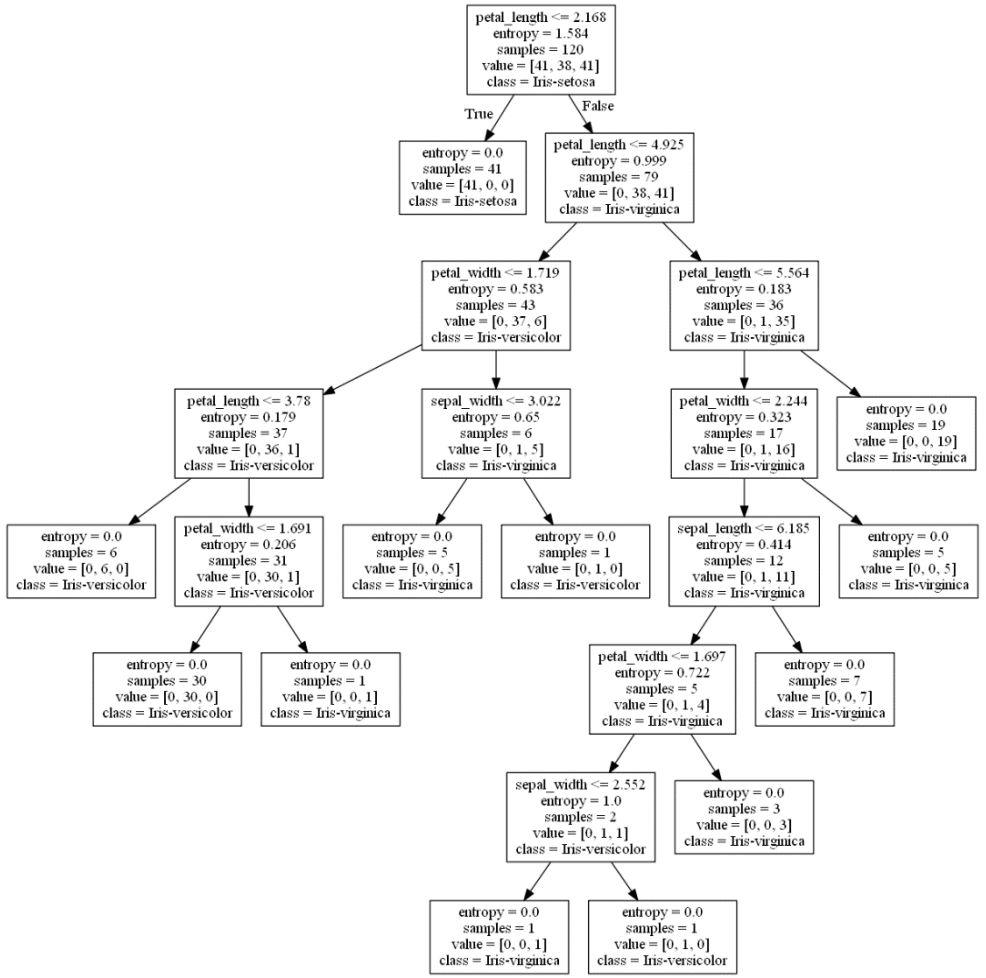
	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica

- **Tham khảo tiến trình thực hiện trong file code trên Jupyter Notebook**

Ví dụ: Phân lớp hoa lan với Decision tree



DecisionTreeClassifier(criterion= 'gini',
splitter='best',
random_state=0)



DecisionTreeClassifier(criterion= 'entropy',
splitter='random',
random_state=0)

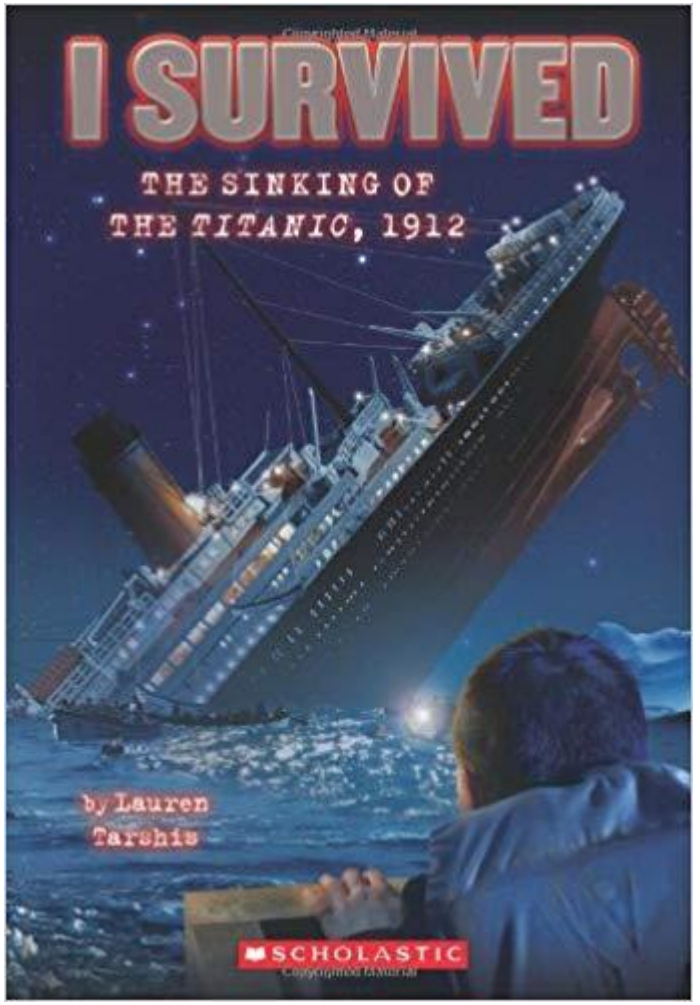
Ví dụ 2:
Bài toán Titanic

Ví dụ: Bài toán Titanic

- Xây dựng model học máy sử dụng Decision Tree dự đoán khả năng được cứu (1), Không được cứu (0) của hành khách trên tập dữ liệu đã được chuẩn bị ở chương 2:

	A	B	C	D	E	F	G
1	Survived	Pclass	Sex	Age	SibSp	Parch	Embarked
2	0	3	0	1	1	0	0
3	1	1	1	2	1	0	1
4	1	3	1	1	0	0	0
5	1	1	1	2	1	0	0
6	0	3	0	2	0	0	0
7	0	3	0	1	0	0	2
8	0	1	0	3	0	0	0
9	0	3	0	0	3	1	0
10	1	3	1	1	0	2	0
11	1	2	1	0	1	0	1
12	1	3	1	0	1	1	0
13	1	1	1	3	0	0	0
14	0	3	0	1	0	0	0
15	0	3	0	2	1	5	0

Data_Titanic_ok



- **Sinh viên làm trên Jupyter Notebook**

THỰC HÀNH 7

Yêu cầu 1:

- Sinh viên tìm hiểu về tập dữ liệu mẫu wine trong Dataset của Sklearn (xác định các features và label)

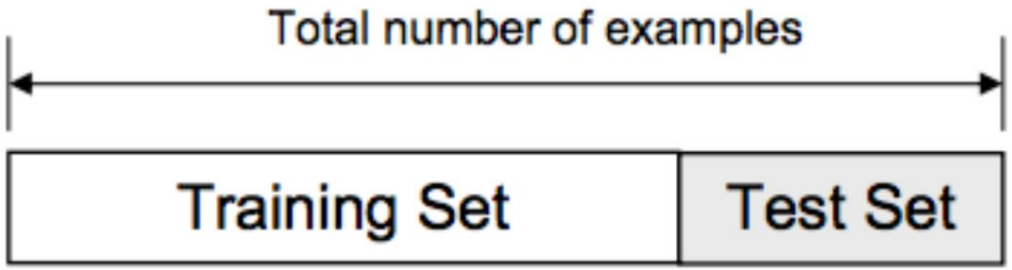
Number of Instances:	178 (50 in each of three classes)
Number of Attributes:	13 numeric, predictive attributes and the class
Attribute Information:	• Alcohol • Malic acid • Ash • Alcalinity of ash • Magnesium • Total phenols • Flavanoids • Nonflavanoid phenols • Proanthocyanins • Color intensity • Hue • OD280/OD315 of diluted wines • Proline

- **class:**
  - class_0
  - class_1
  - class_2

Classes	3
Samples per class	[59,71,48]
Samples total	178
Dimensionality	13
Features	real, positive

Yêu cầu 2:

- Tách tập dữ liệu data_wine thành 2 phần train – test theo tỷ lệ 75% - 25%



Yêu cầu 3:

- Sử dụng thuật toán Cây quyết định trong 2 trường hợp:
  - **Sử dụng độ đo Entropy**: Trực quan hóa cây quyết định thu được trên tập Huấn luyện, xác định thuộc tính quan trọng và vẽ biểu đồ; xác định độ chính xác của mô hình trên tập Test.
  - **Sử dụng độ đo Gini:** Trực quan hóa cây quyết định thu được trên tập Huấn luyện, xác định thuộc tính quan trọng và vẽ biểu đồ; xác định độ chính xác của mô hình trên tập Test.

Yêu cầu 4:

- Một mẫu rượu có các tham số như sau:

• Alcohol	: 12.7
• Malic acid	: 3.05
• Ash	: 1.88
• Alcalinity of ash	: 28.8
• Magnesium	: 101.1
• Total phenols	: 2.88
• Flavanoids	: 3.88
• Nonflavanoid phenols	: 0.44
• Proanthocyanins	: 2.88
• Color intensity	: 8.8
• Hue	: 1.48
• OD280/OD315 of diluted wines	: 3.88
• Proline	: 888

Sử dụng model huấn luyện được cho biết mẫu rượu này thuộc loại nào?


HANOI UNIVERSITY OF MINING AND GEOLOGY
KHOA
CÔNG NGHỆ THÔNG TIN
FACULTY OF INFORMATION TECHNOLOGY

Thank you!