DATA MINING

GROUP 8

NỘI DUNG

1. RANDOM FOREST ALGORITHM

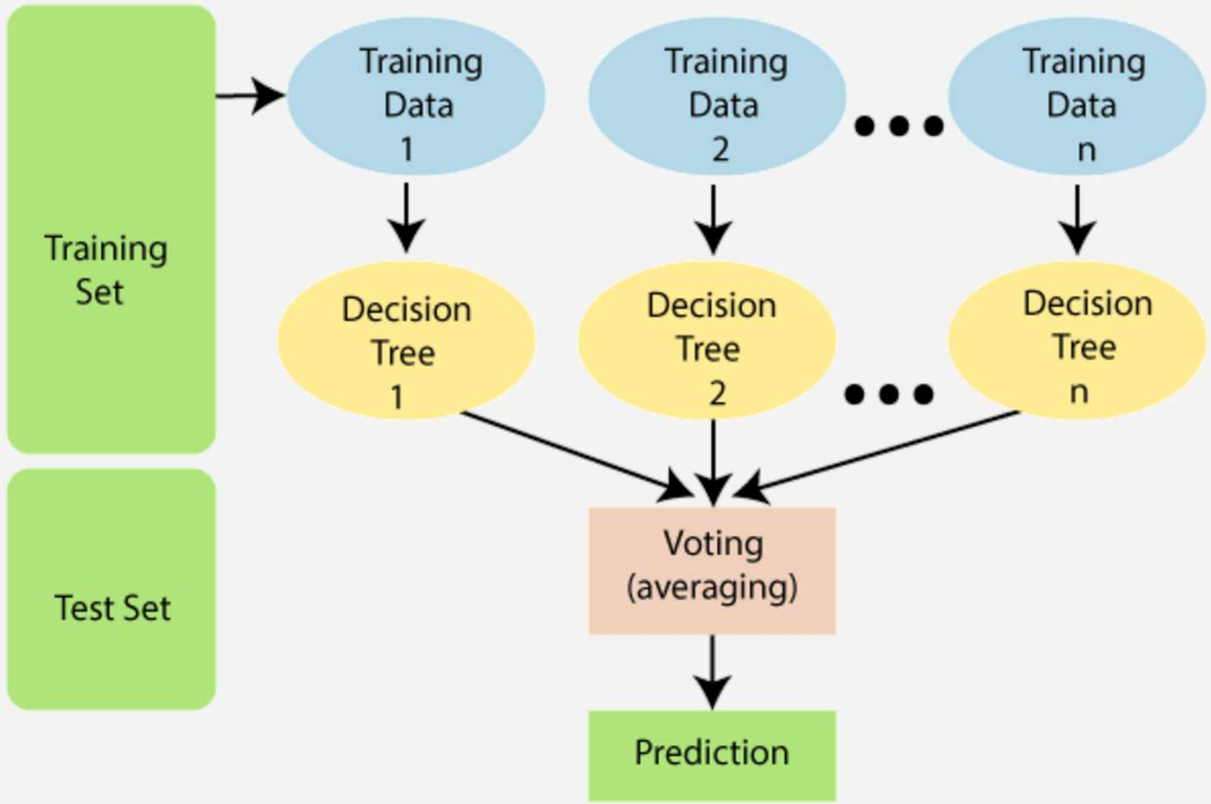
2. KNN ALGORITHM

RANDOM FOREST ALGORITHM

KHÁI NIỆM

- Là một thuật toán học máy được sử dụng cho các bài toán phân loại, hồi quy và các tác vụ khác liên quan đến dữ liệu được giám sát.

- Ở thuật toán này, sẽ xây dựng nhiều cây quyết định bằng thuật toán Decision Tree, và mỗi cây quyết định khác nhau.



Tiến Đạt

RANDOM FOREST ALGORITHM

ƯU ĐIỂM

- Có thể được sử dụng trong các bài toán phân loại và hồi quy
- Giải quyết vấn đề khớp quá mức
- Hoạt động tốt ngay cả khi dữ liệu chứa giá trị null/thiếu.
- Mỗi cây quyết định được tạo ra là độc lập với nhau
- Có tính ổn định cao
- Duy trì tính đa dạng
  Miễn nhiễm với lời nguyền của chiều không gian.

NHƯỢC ĐIỂM

- Random Forest rất phức tạp so với cây quyết định
- Toàn bộ quá trình tốn thời gian nhiều hơn so với các mô hình khác
- Khả năng dự đoán kém với dữ liệu ngoài phạm vi
- Dễ bị ảnh hưởng bởi dữ liệu nhiễu
- Chi phí tính toán cao

Tiến Đạt

ỨNG DỤNG

Dự đoán điểm tín dụng

Đánh giá rủi ro cho vay tiền

Marketing

Phân tích dữ liệu khách hàng, dự đoán xu hướng mua sắm

Thời tiết

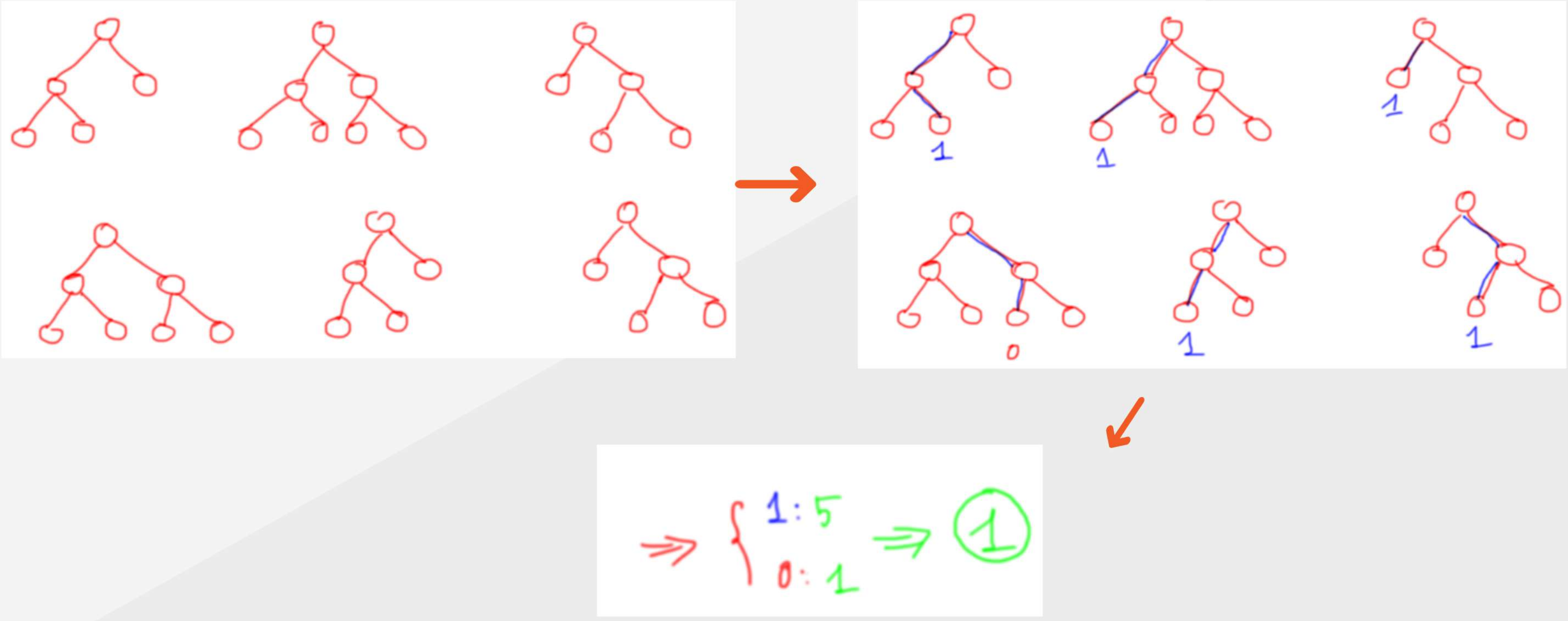
Dự báo thời tiết cho các vùng cụ thể

Y tế

Dự đoán nguy cơ bệnh lý, chẩn đoán bệnh

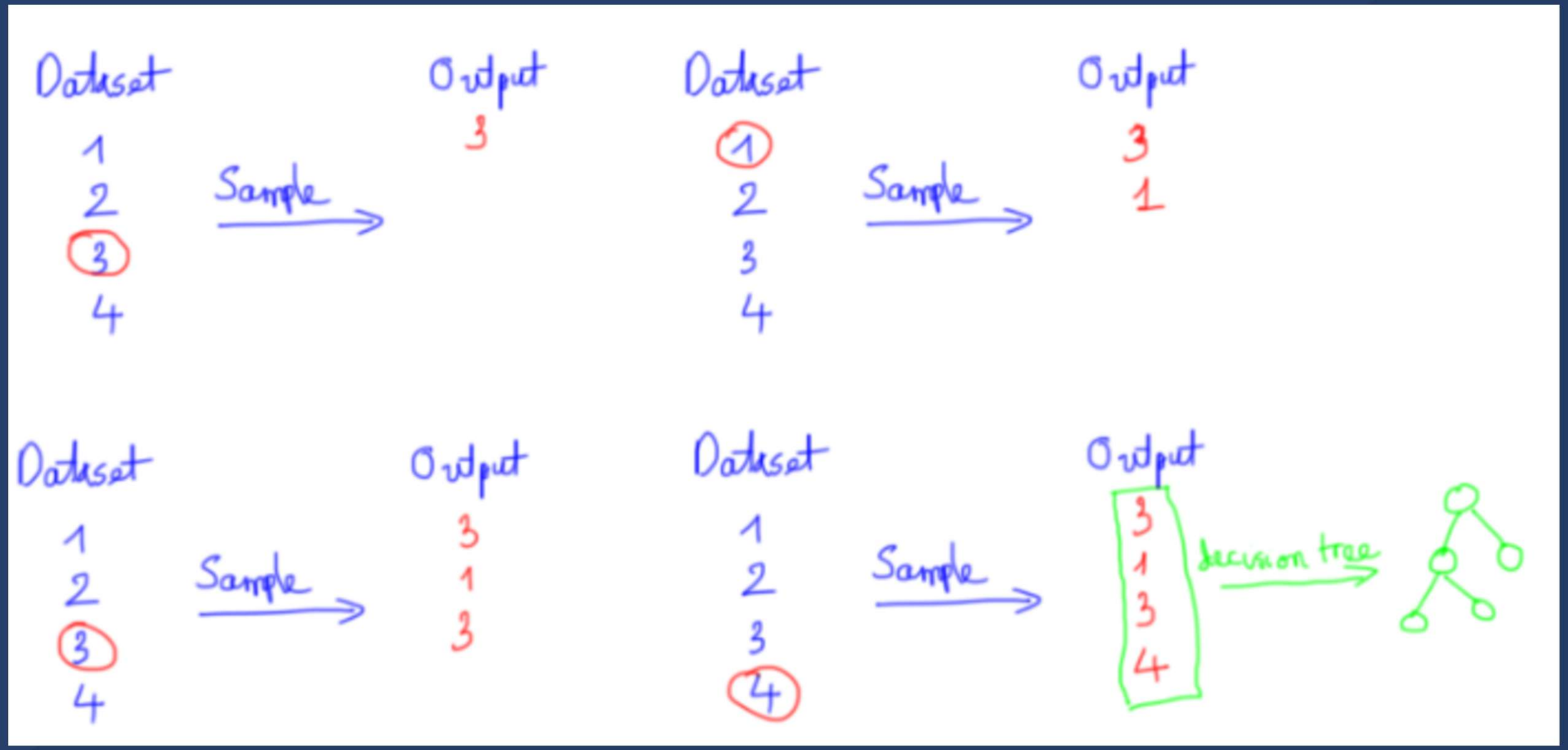
Tiến Đạt

CÁCH THỨC HOẠT ĐỘNG



Công Hiếu

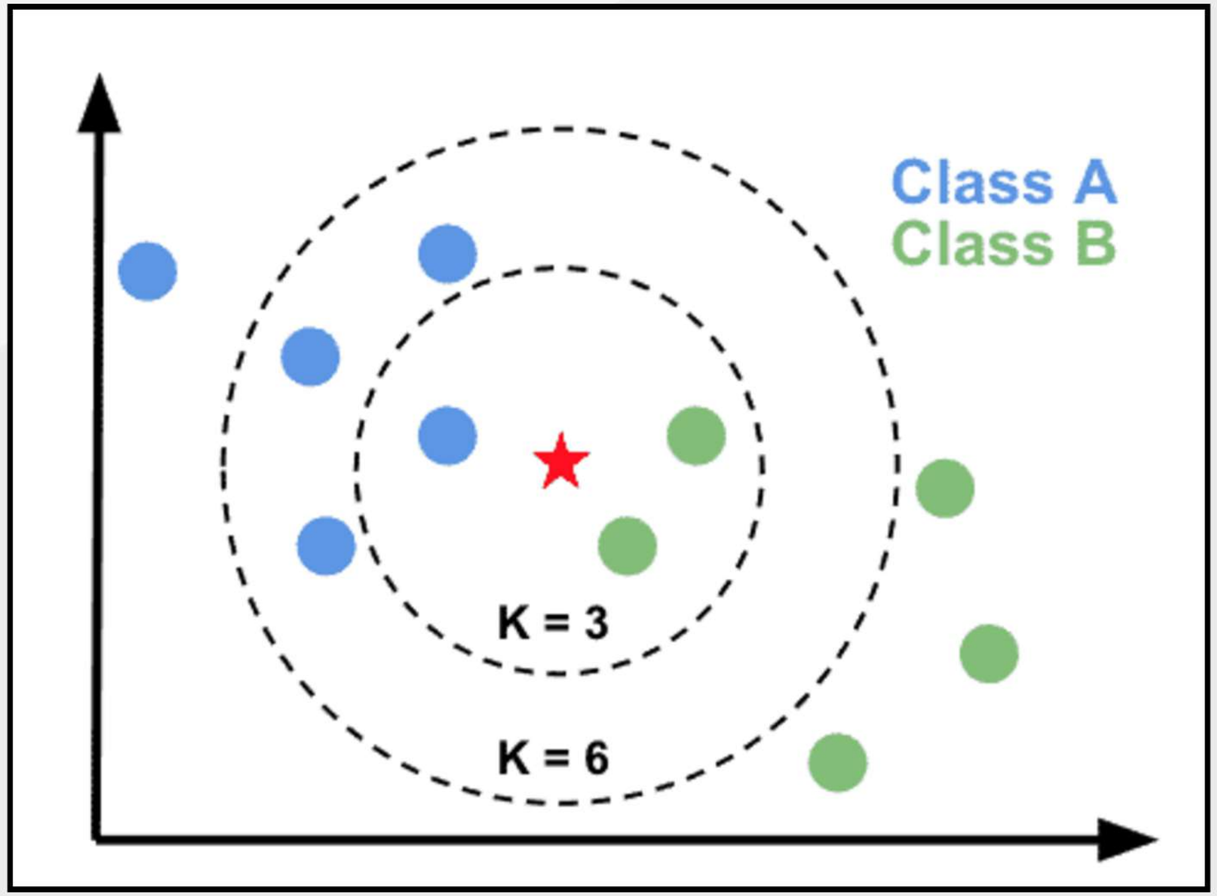
CÁC BƯỚC THỰC HIỆN



Công Hiếu

KNN

- Một trong những thuật toán có giám sát đơn giản nhất
- Thuật toán này được sử dụng rộng rãi trong nhiều lĩnh vực khác nhau từ phân loại, hồi quy cho đến hệ thống gợi ý.



Vũ Thị Hà

KNN

KHOẢNG CÁCH

Thuật toán KNN cho rằng những dữ liệu tương tự nhau sẽ tồn tại gần nhau trong một không gian, từ đó công việc của chúng ta là sẽ tìm k điểm gần với dữ liệu cần kiểm tra nhất.

Vũ Thị Hà

CÓ 3 CÁCH TÍNH KHOẢNG CÁCH

Euclidean $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$

Manhattan $\sum_{i=1}^{k}|x_i - y_i|$

Minkowski $\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q}$

→ $\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$

Vì dữ liệu trong một tập dữ liệu thường có độ đo khác nhau, dẫn đến sự chênh lệch về độ lớn.

→ **chuẩn hóa dữ liệu**

CÓ 2 CÁCH CHUẨN HÓA DỮ LIỆU

Min-max normalization

phương pháp này đưa các giá trị dữ liệu về khoảng giá trị nằm trong khoảng [0,1]

$$x_i' = \frac{x_i - \min(x)}{\max(x) - \min(x)}$$

Z-score normalization

phương pháp này được dùng để chuẩn hóa các giá trị dữ liệu sao cho chúng có trung bình bằng 0 và độ lệch chuẩn bằng 1.

$$x_i' = z = \frac{x_i - \bar{x}}{\sigma}$$

Vũ Thị Hà

VÍ DỤ

Ta có dữ liệu gồm: **Age, Loan, Default.**

Chọn **K = 3**, hãy dự đoán trường hợp **Age = 48, Loan = 142.000$** thì có khả năng vỡ nợ hay không?

Age	Loan	Default
25	$40.000	N
35	$60.000	N
45	$80.000	N
20	$20.000	N
35	$120.000	N
52	$18.000	N
23	$95.000	Y
40	$62.000	Y
60	$100.000	Y
48	$220.000	Y
33	$150.000	Y
48	$142.000	?

Default: Vỡ nợ

Vũ Thị Hà

VÍ DỤ

Bước 1:

Ta chuẩn hóa dữ liệu cột Age và Loan bằng Min-Max Normalization

*** Với Age (1) = 25:**

$$X_{25} = \frac{25-20}{60-20} = 0.125$$

*** Với Loan (1) = 40000:**

$$X_{\$40.000} = \frac{40000-18000}{220000-18000} = 0.109$$

Age	Loan	Default
0,125	0,109	N
0,375	0,208	N
0,625	0,307	N
0,000	0,010	N
0,375	0,505	N
0,800	0,000	N
0,075	0,381	Y
0,500	0,218	Y
1,000	0,406	Y
0,700	1,000	Y
0,325	0,653	Y
0,700	0,614	?

Tương tự bước trên ta được như bảng bên

Vũ Thị Hà

VÍ DỤ

Bước 2:

Tính khoảng cách, dùng công thức Euclidean

Euclidean $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$

***Distance (1)**

$$\text{Distance}(25,48)$$

$$=\sqrt{(0.7-0.125)^2+(0.614-0.109)^2}$$

$$=0.765$$

Age	Loan	Default	Distance
0,125	0,109	N	0,765
0,375	0,208	N	0,520
0,625	0,307	N	0,316
0,000	0,010	N	0,925
0,375	0,505	N	0,343
0,800	0,000	N	0,622
0,075	0,381	Y	0,667
0,500	0,218	Y	0,444
1,000	0,406	Y	0,365
0,700	1,000	Y	0,386
0,325	0,653	Y	0,377
0,700	0,614	?	

Vũ Thị Hà

VÍ DỤ

Bước 3:

Vì K = 3, chọn 3 giá trị nhỏ nhất trong cột Distance: 0.316, 0.343, 0.365

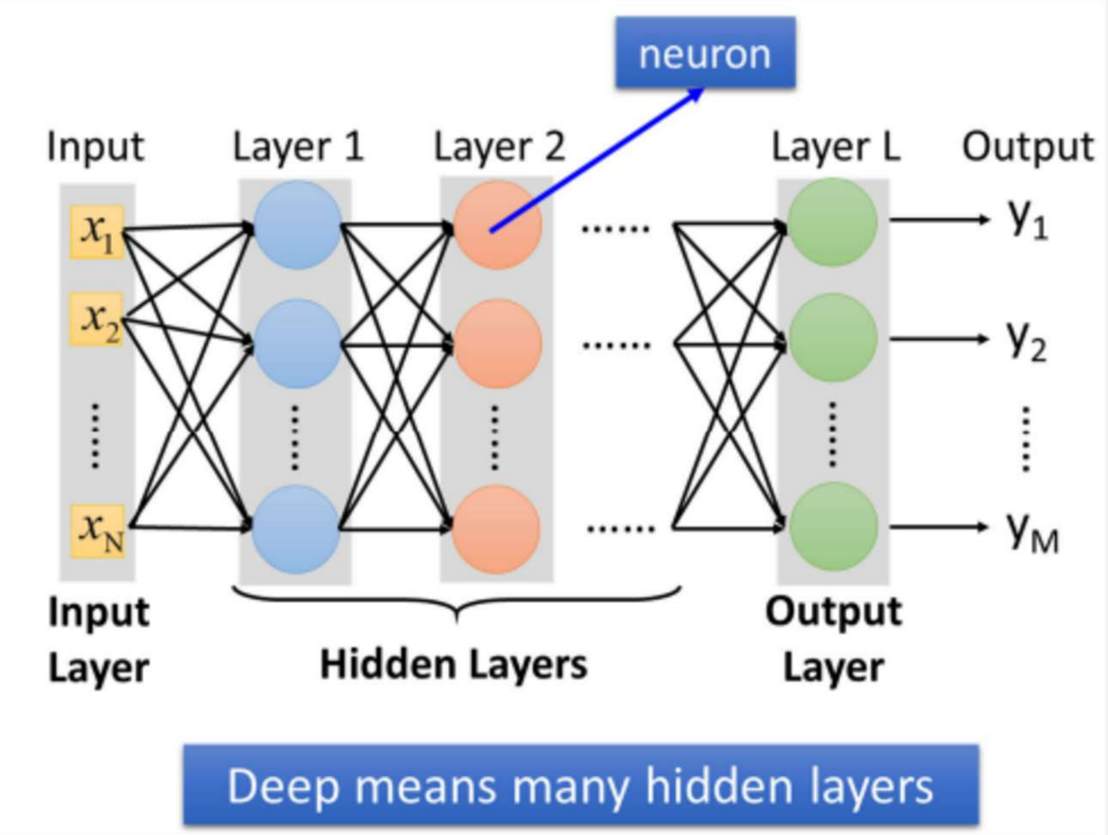
3 giá trị Distance này tương ứng với giá trị Default là: N, N, Y

Vì có 2/3 là giá trị N nên ta dự đoán với trường hợp

Age = 48, Loan = 142000 thì khả năng vỡ nợ là không (N).

Age	Loan	Default	Distance
0,125	0,109	N	0,765
0,375	0,208	N	0,520
0,625	0,307	N	0,316
0,000	0,010	N	0,925
0,375	0,505	N	0,343
0,800	0,000	N	0,622
0,075	0,381	Y	0,667
0,500	0,218	Y	0,444
1,000	0,406	Y	0,365
0,700	1,000	Y	0,386
0,325	0,653	Y	0,377
0,700	0,614	?	

NEURAL NETWORK



- **Input Layer:** là layer đầu tiên nó nhận dữ liệu đầu vào và chuyển tiếp chúng tới các hidden layer tiếp theo.
- **Hidden Layer:** Trích xuất các đặc trưng (features) từ dữ liệu đầu vào và chuyển tiếp đến các hidden layer tiếp theo hoặc output layer.
- **Output Layer:** là layer cuối cùng nó có nhiệm vụ tính toán và đưa ra kết quả dự đoán.

Phương Uyên

DEEP LEARNING

Một loại mạng thần kinh nhân tạo chứa nhiều lớp ẩn giữa lớp đầu vào và lớp đầu ra.

Các lớp ẩn này cho phép mạng nơ-ron học các biểu diễn phức tạp hơn của dữ liệu đầu vào bằng cách xử lý dữ liệu đó qua nhiều lớp biến đổi phi tuyến tính.

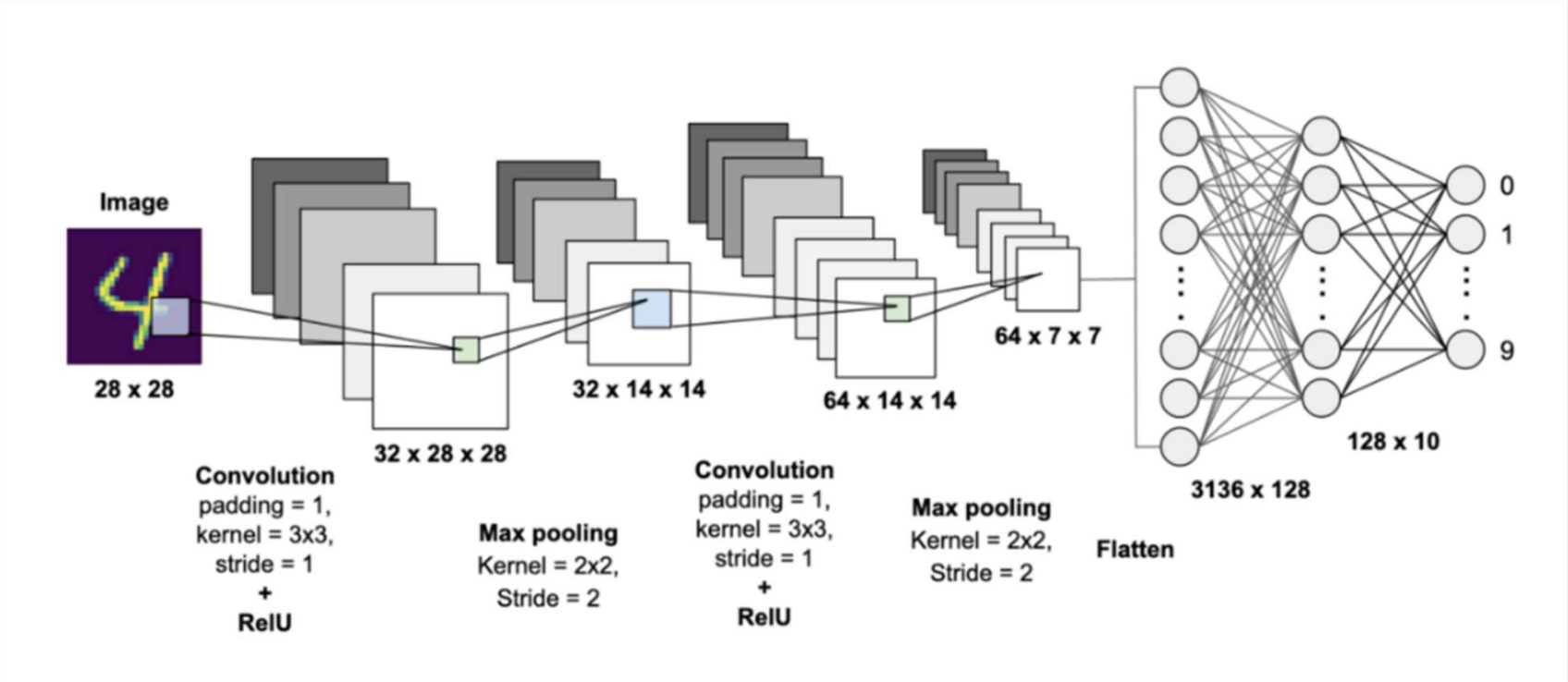
Các mô hình cơ bản

- **CNN**
- **RNN**
- **GRU**
- **LSTM**

Phương Uyên

CNN

Convolutional Neural Network (CNN) là một loại kiến trúc mạng neural được sử dụng rộng rãi trong xử lý ảnh, nhận dạng hình ảnh và công nghệ liên quan đến thị giác máy tính.



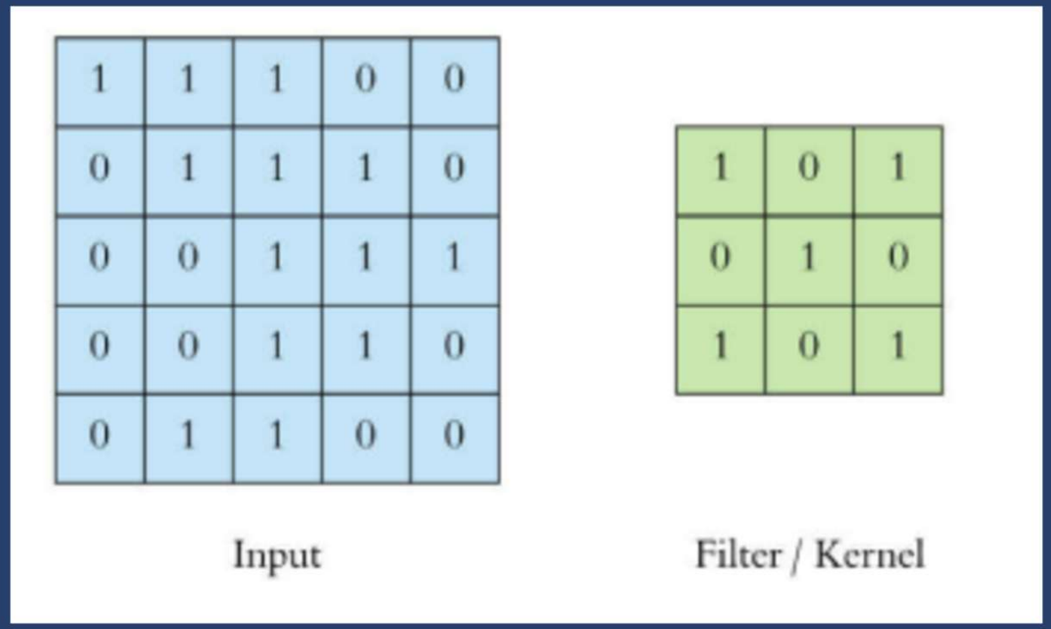
Gồm 3 lớp:

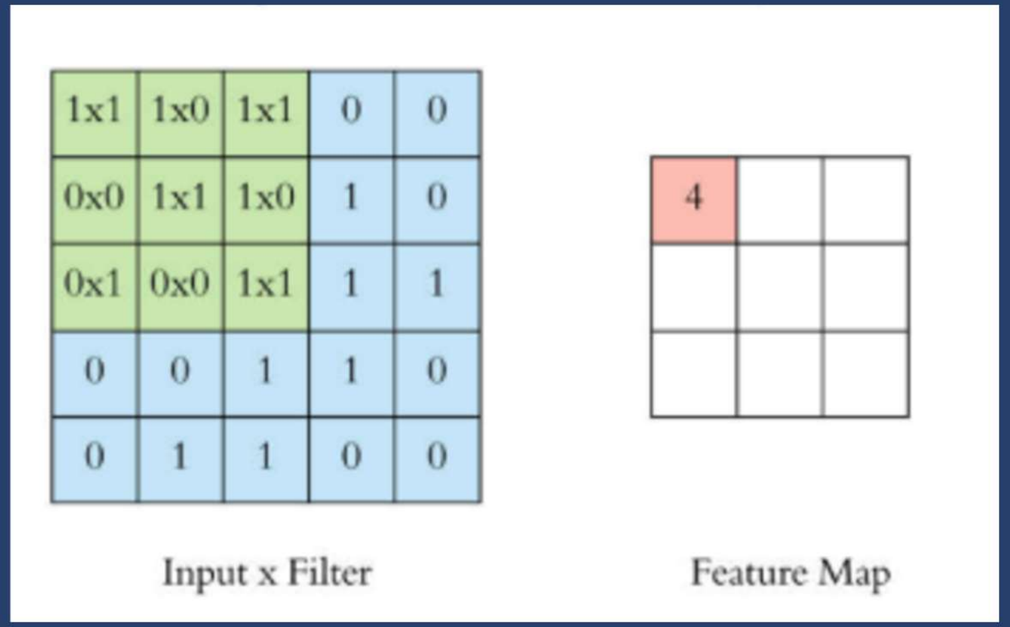
- **Lớp tích chập** (convolution)
- **Lớp gộp** (pooling)
- **Lớp kết nối đầy đủ** (full connected)

Phương Uyên

CONVOLUTIONAL LAYER

Quá trình tích chập được thực hiện bằng cách trượt bộ lọc qua đầu vào (input) của layer theo từng bước (stride) nhất định, và thực hiện phép nhân giữa bộ lọc và vùng đầu vào tương ứng (gọi là vùng nhân) và sau đó tổng hợp lại (summarize) để tạo ra đầu ra (output) của Conv layer.





Phương Uyên

THANK YOU FOR LISTENING!