Tuần 12: Giả thuyết thống kê và mô hình dự báo (tiếp theo)

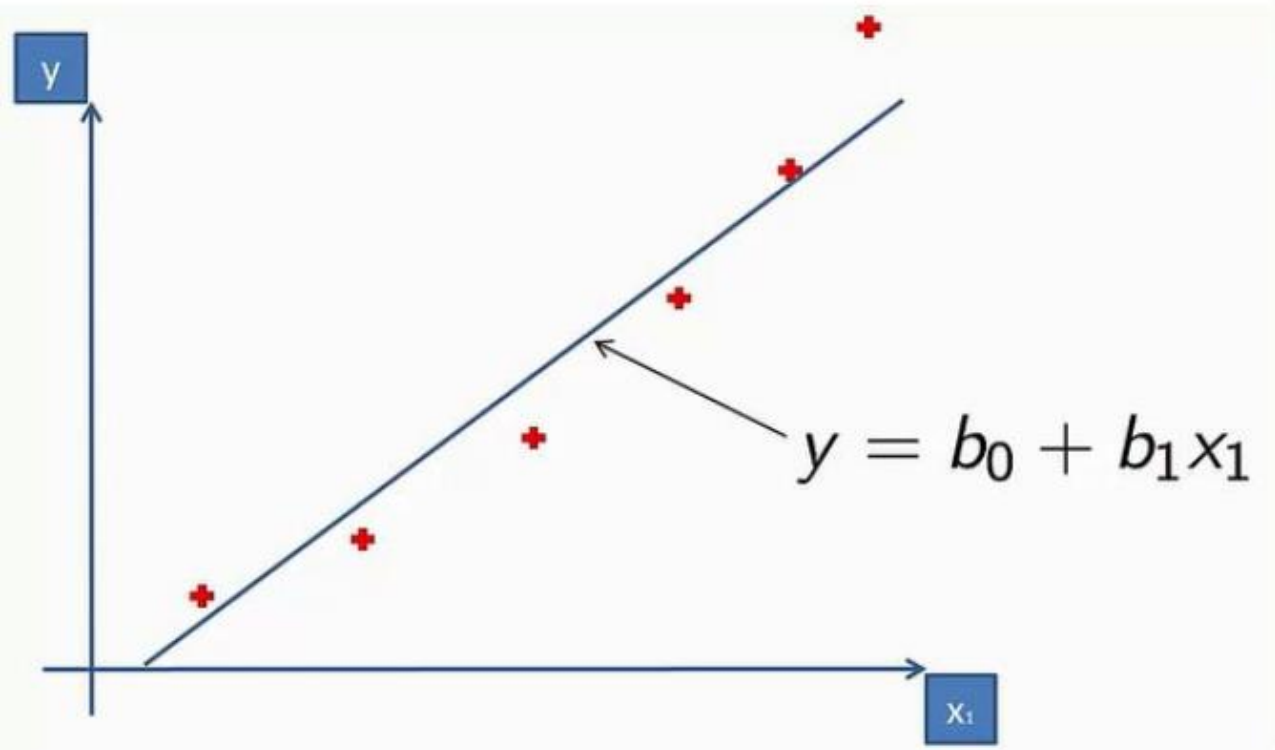
Giảng viên: ThS. Lê Minh Tân

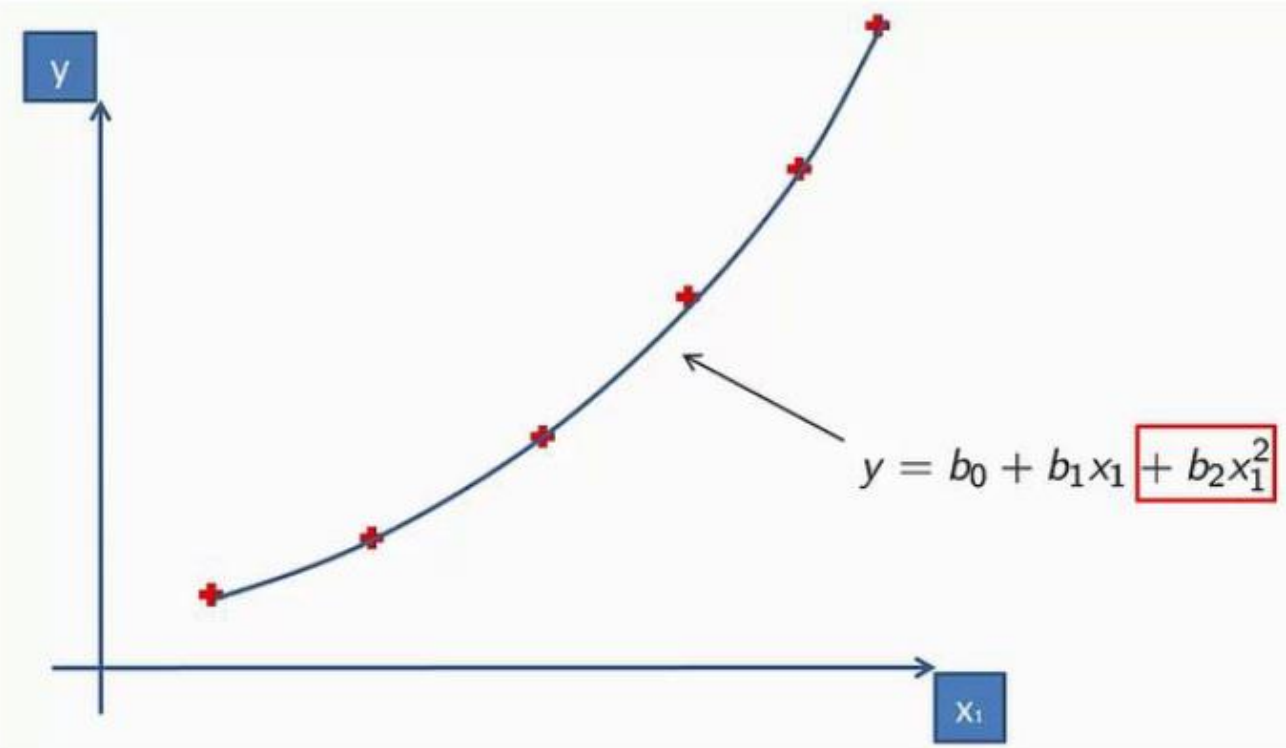
Nội dung

I. Nhóm mô hình hồi quy
II. Nhóm mô hình CART
III. Thước đo đánh giá mô hình dự báo

I. Nhóm mô hình hồi quy

- Ngoài hồi quy tuyến tính, ta có:
  - Hồi quy đa thức (polynomial regression)
  - Hồi quy logistic


y
x₁
$y = b_0 + b_1 x_1$


y
x₁
$y = b_0 + b_1 x_1 + b_2 x_1^2$

$$y = \beta_1 x + \beta_2 x^2 + \cdots + \beta_n x^n + \beta_0 + \varepsilon$$

$$\beta = (X'^T X')^{-1} X'^T y$$

$$y = \beta_1 x + \beta_2 x^2 + \cdots + \beta_n x^n + \beta_0 + \varepsilon$$

$$\beta = (X'^T X')^{-1} X'^T y$$

$x_{0,0}$	$x_{0,1}$	…	$x_{0,n-1}$
$x_{1,0}$	$x_{1,1}$	…	$x_{1,n-1}$
…	…	…	…
$x_{m-1,0}$	$x_{m-1,1}$	…	$x_{m-1,n-1}$

$$y = \beta_1 x + \beta_2 x^2 + \cdots + \beta_n x^n + \beta_0 + \varepsilon$$

$$\beta = (X'^T X')^{-1} X'^T y$$

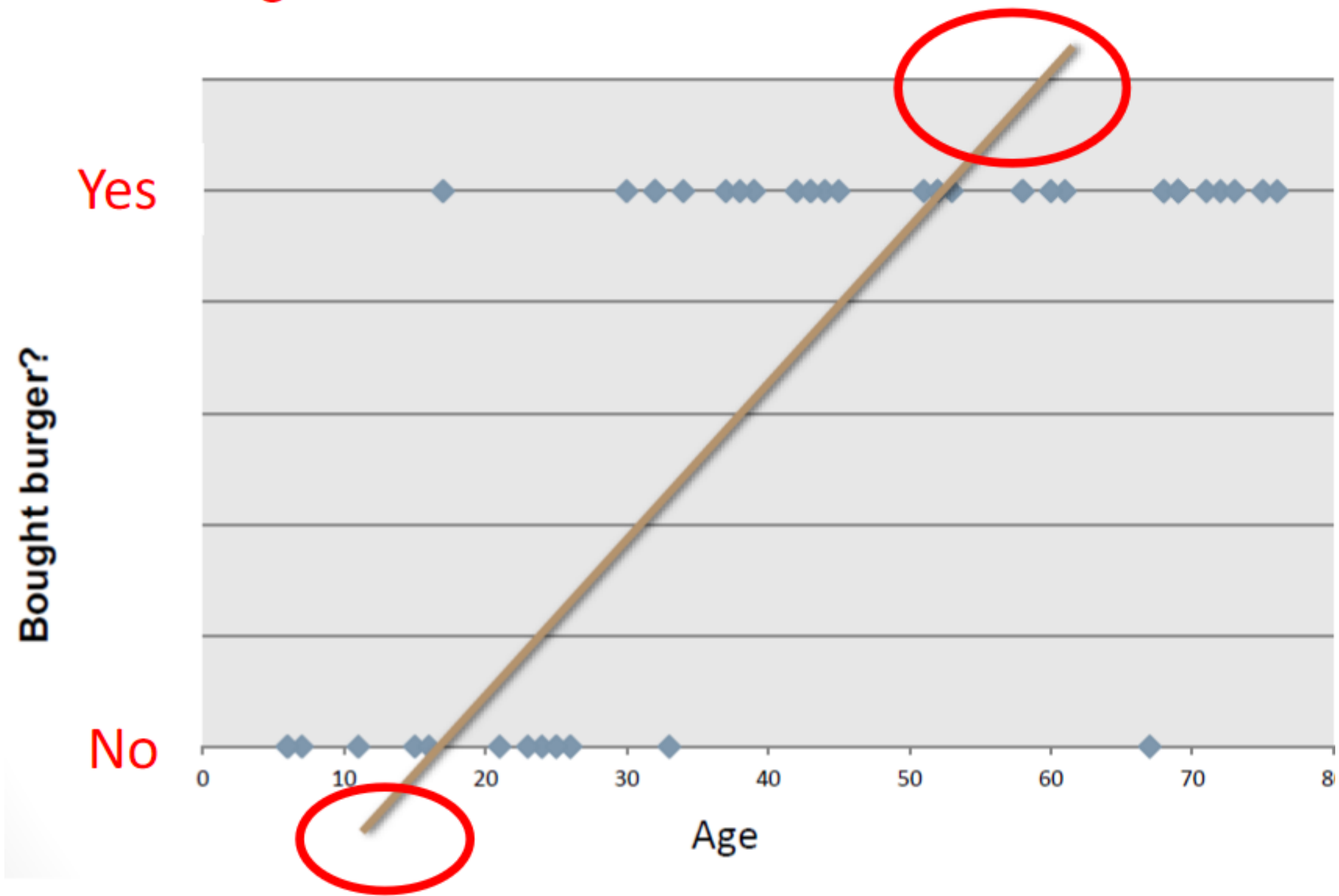
$x_{0,0}^2$	$x_{0,1}^2$	…	$x_{0,n-1}^2$
$x_{1,0}^2$	$x_{1,1}^2$	…	$x_{1,n-1}^2$
…	…	…	…
$x_{m-1,0}^2$	$x_{m-1,1}^2$	…	$x_{m-1,n-1}^2$

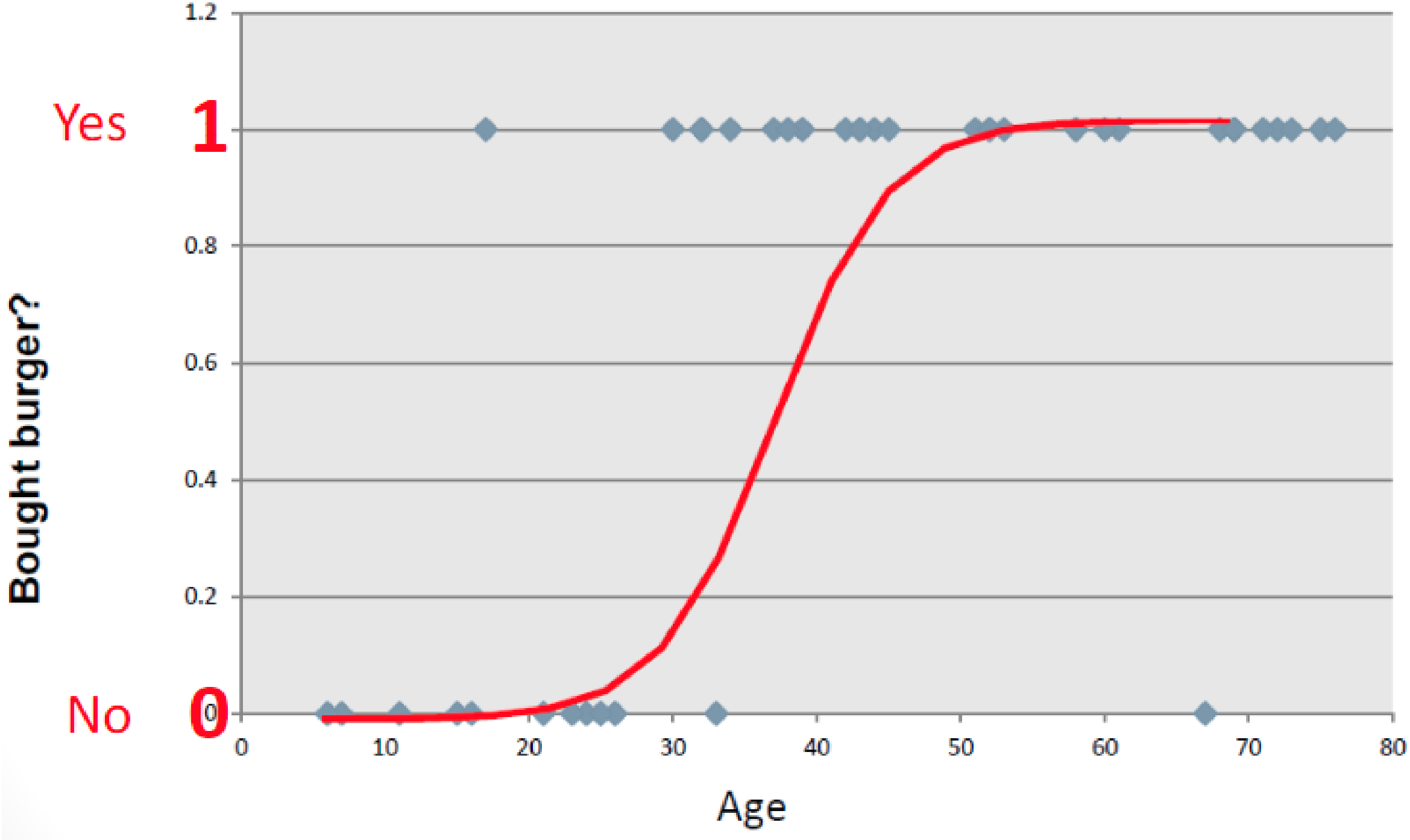
Tật mắt: Cận/Không cận?

Bán hàng: Mua / không mua?

Marketing: Click / không click?

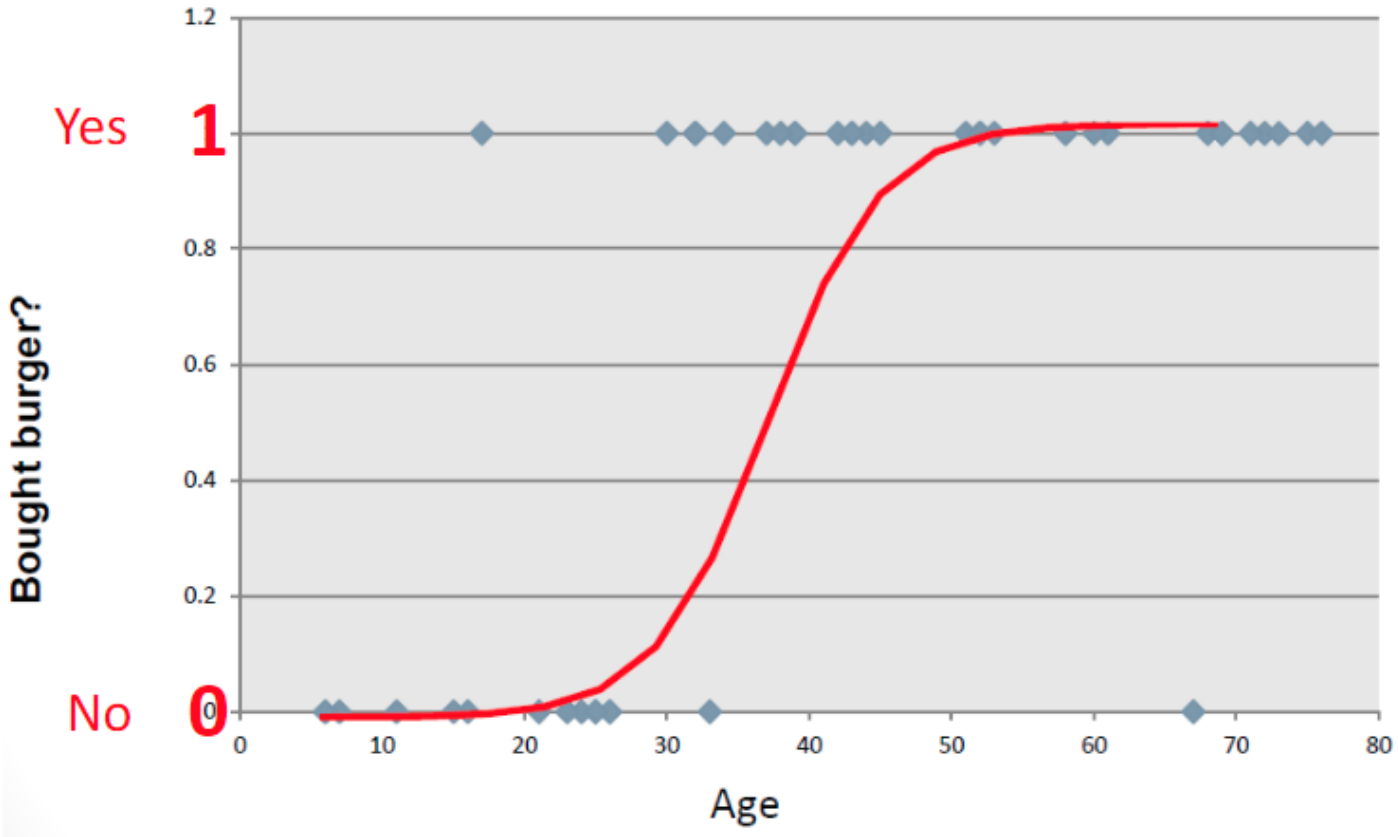
Kết quả môn học: Qua môn / rớt?


Yes
No
Bought burger?
0
10
20
30
40
50
60
70
Age


Yes
1
No
0
Bought burger?
1.2
0.8
0.6
0.4
0.2
0
0
10
20
30
40
50
60
70
80
Age

Hàm logistic:

$$f(x) = \frac{L}{1 + e^{-k(x-x_0)}} = p(x)$$



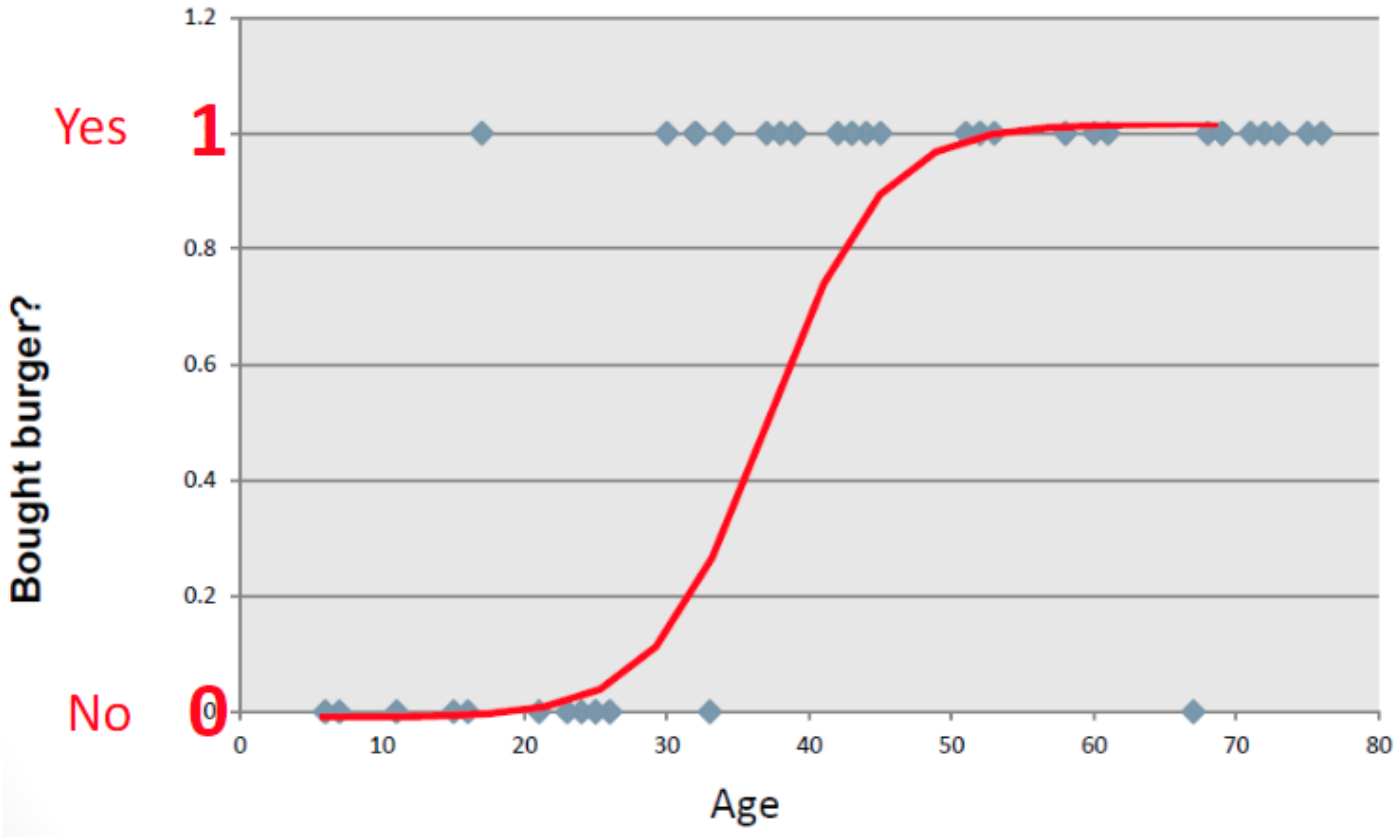
https://en.wikipedia.org/wiki/Logistic_function

Hàm logistic:

$$f(x) = \frac{L}{1 + e^{-k(x-x_0)}} = p(x)$$

Xác suất có điều kiện:

$$p(x) = \frac{1}{1 + e^{-kx+kx_0}}$$



https://en.wikipedia.org/wiki/Logistic_function

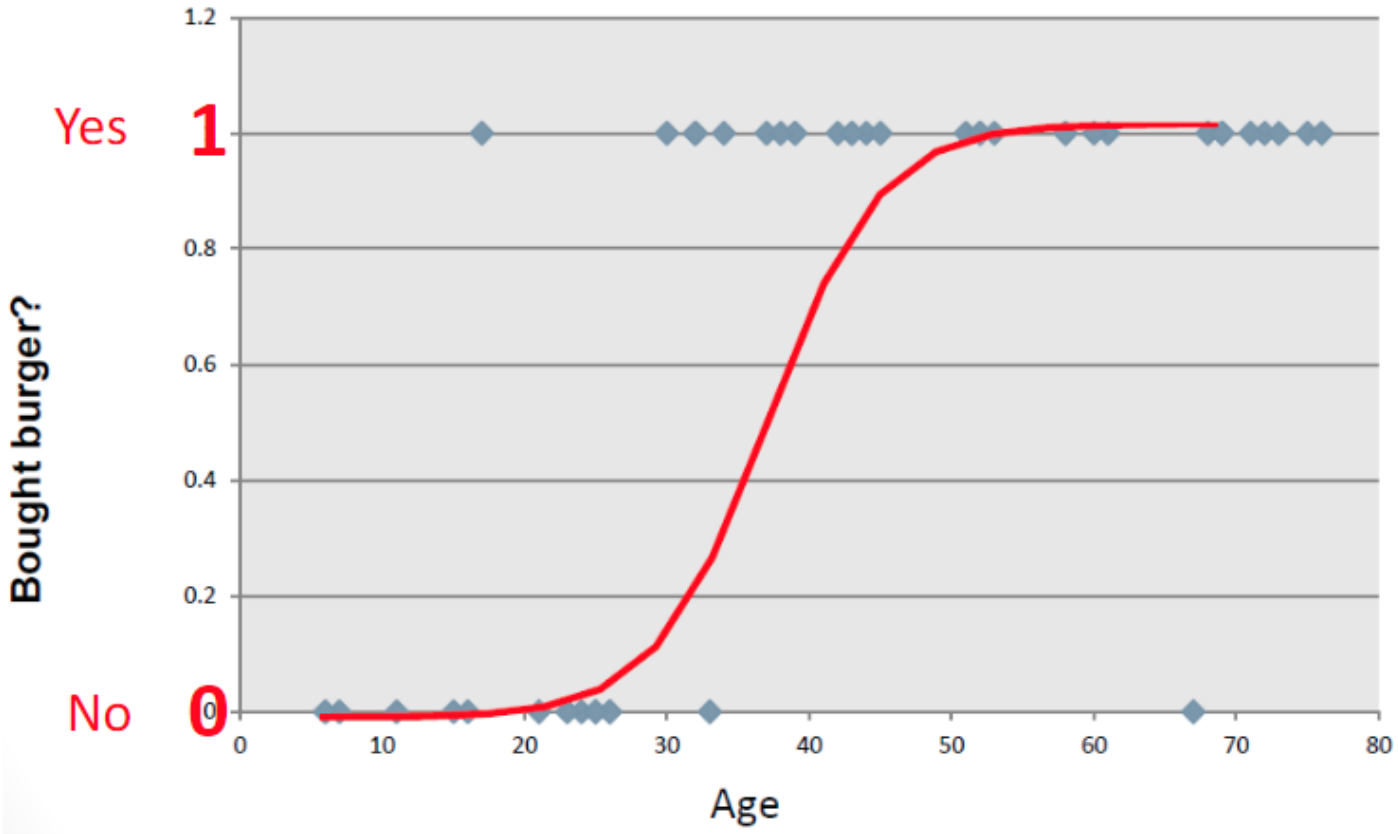
Hàm logistic:

$$f(x) = \frac{L}{1 + e^{-k(x-x_0)}} = p(x)$$

Xác suất có điều kiện:

$$p(x) = \frac{1}{1 + e^{-kx+kx_0}}$$

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



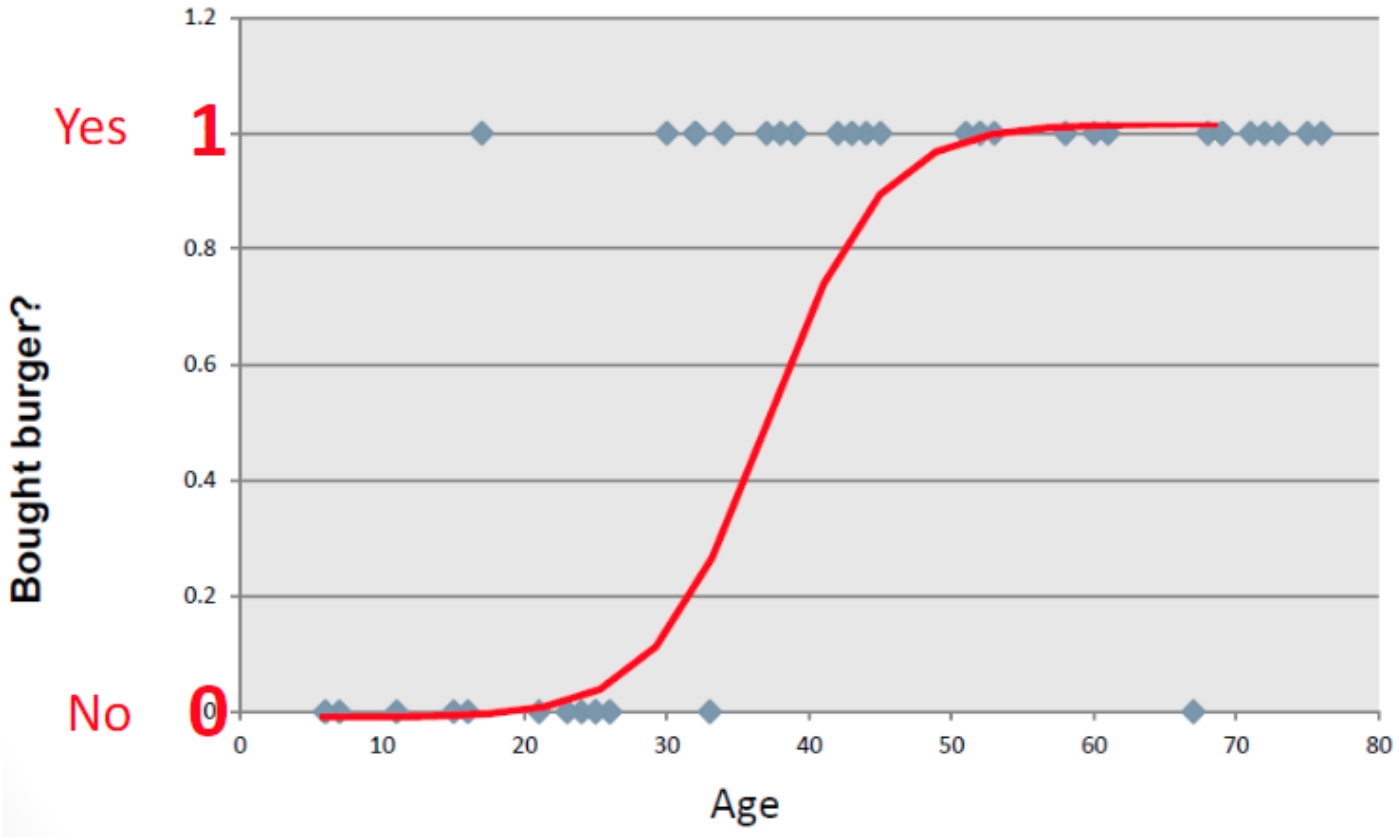
https://en.wikipedia.org/wiki/Logistic_function

Hàm logistic:

$$f(x) = \frac{L}{1 + e^{-k(x-x_0)}} = p(x)$$

Xác suất có điều kiện:

$$p(y|x) = \frac{1}{1 + e^{-y}}$$



https://en.wikipedia.org/wiki/Logistic_function

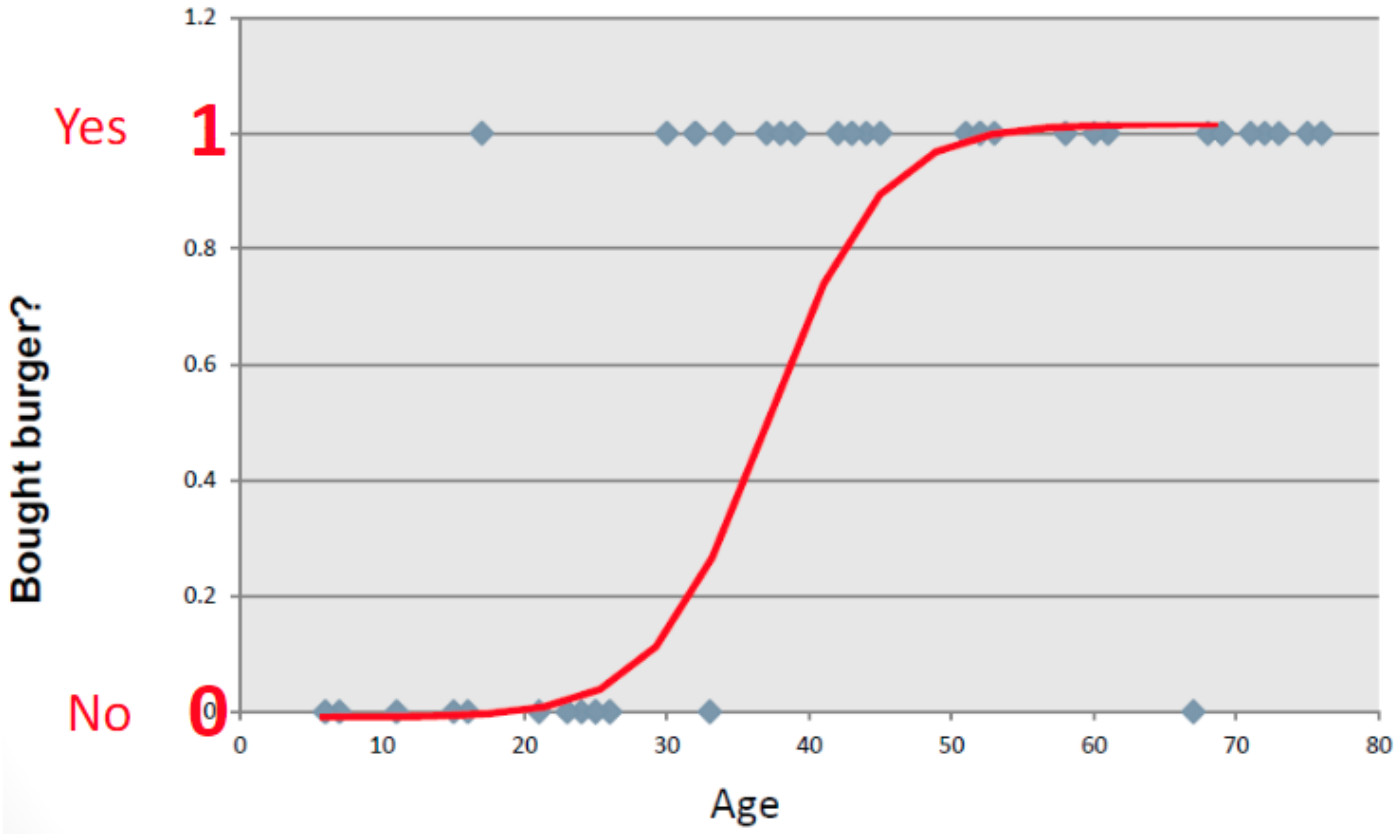
Hàm logistic:

$$f(x) = \frac{L}{1 + e^{-k(x-x_0)}} = p(x)$$

Xác suất có điều kiện:

$$p(y|x) = \frac{1}{1 + e^{-y}}$$

$$e^{-y} = \frac{1}{p} - 1 = \frac{1-p}{p}$$



https://en.wikipedia.org/wiki/Logistic_function

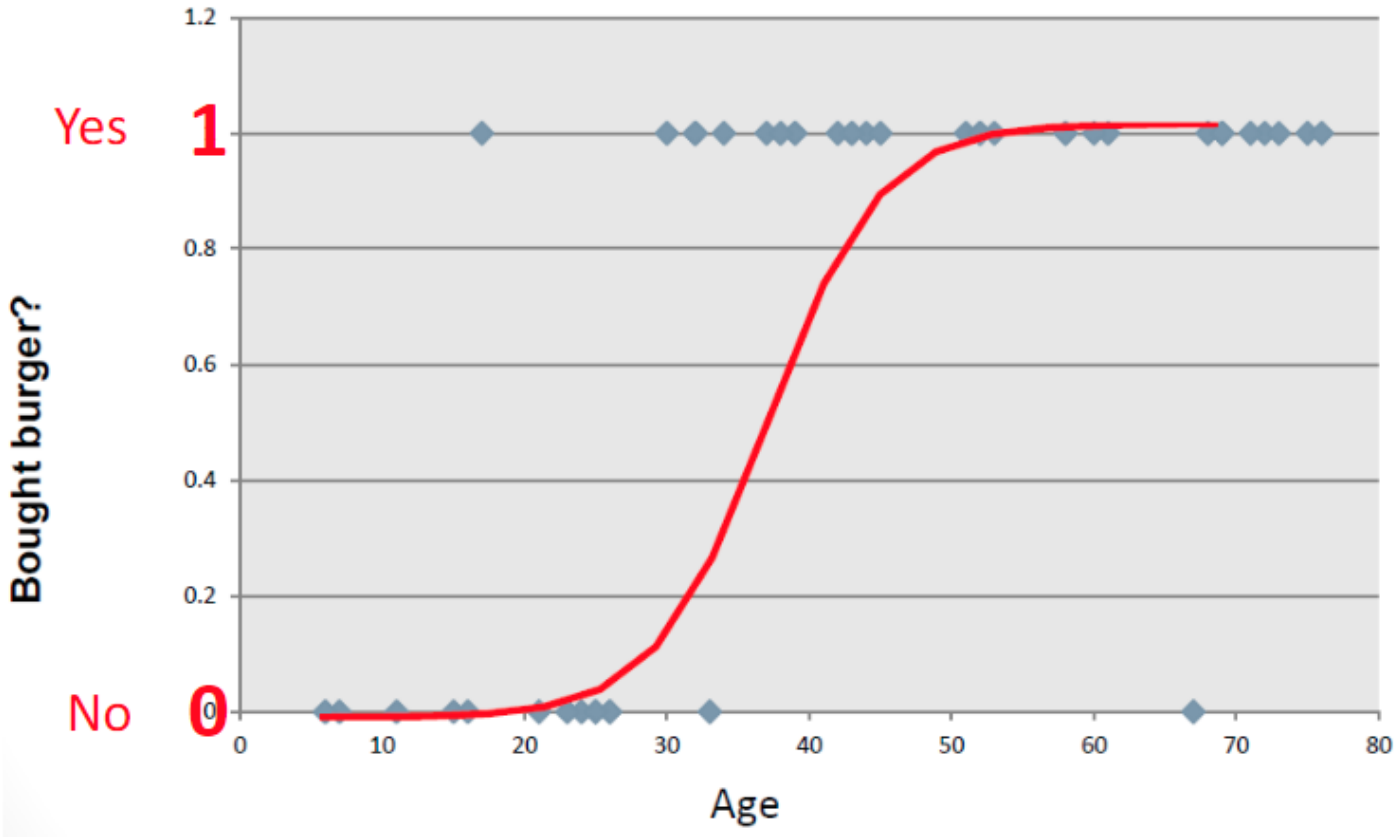
Xác suất có điều kiện:

$$p(y|x) = \frac{1}{1 + e^{-y}}$$

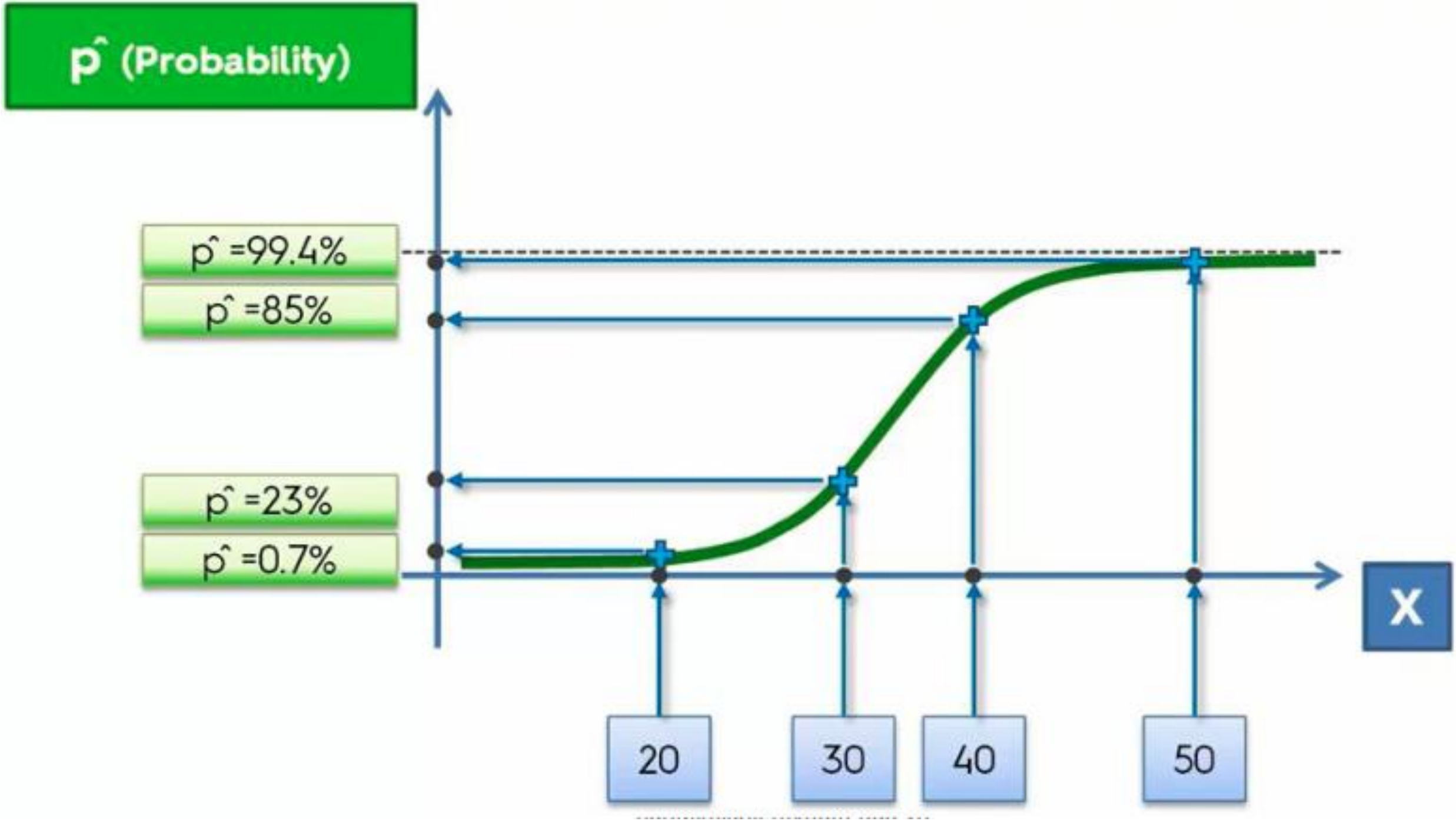
$$e^{-y} = \frac{1}{p} - 1 = \frac{1 - p}{p}$$

$$e^{y} = \frac{p}{1 - p}$$

$$y = \ln\left(\frac{p}{1 - p}\right)$$

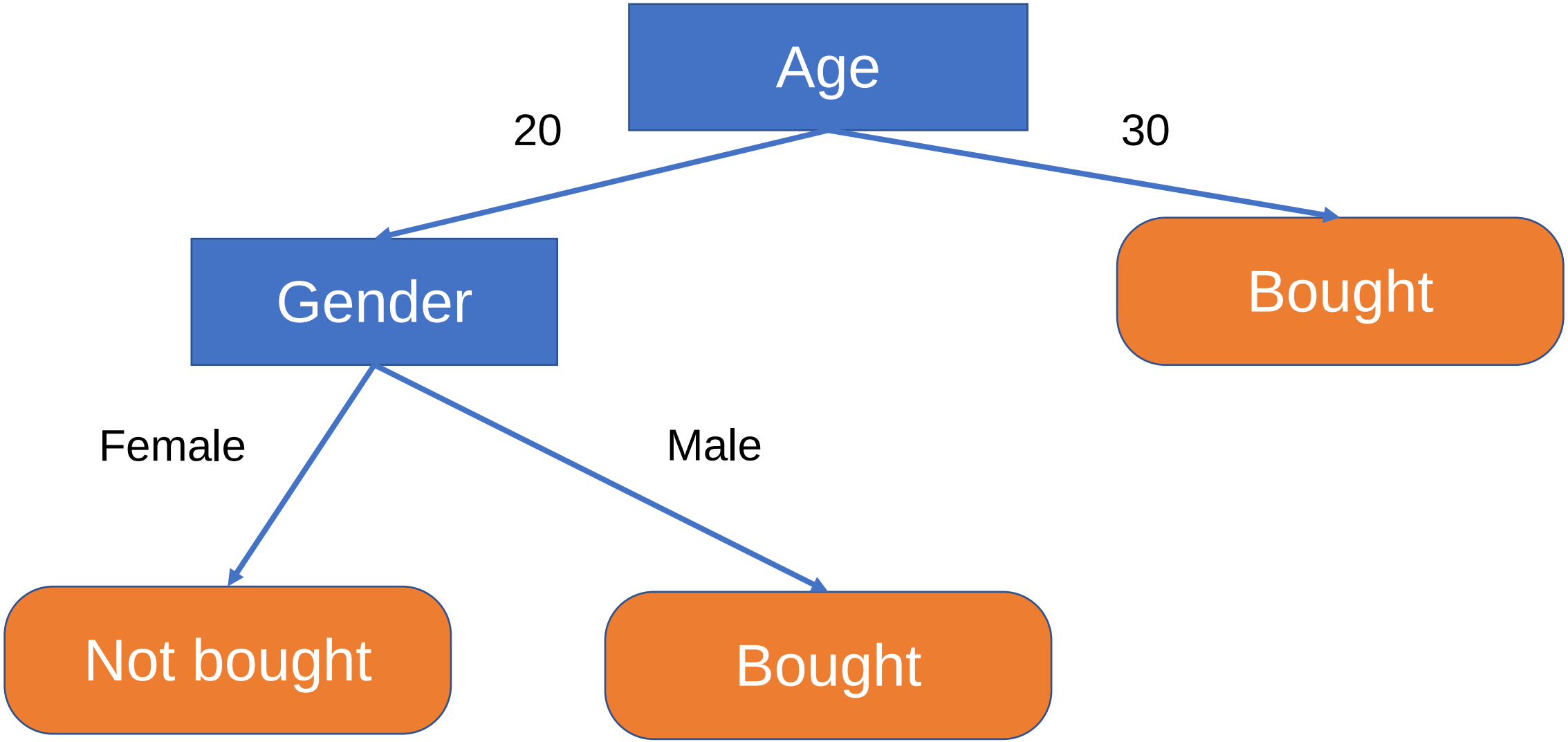


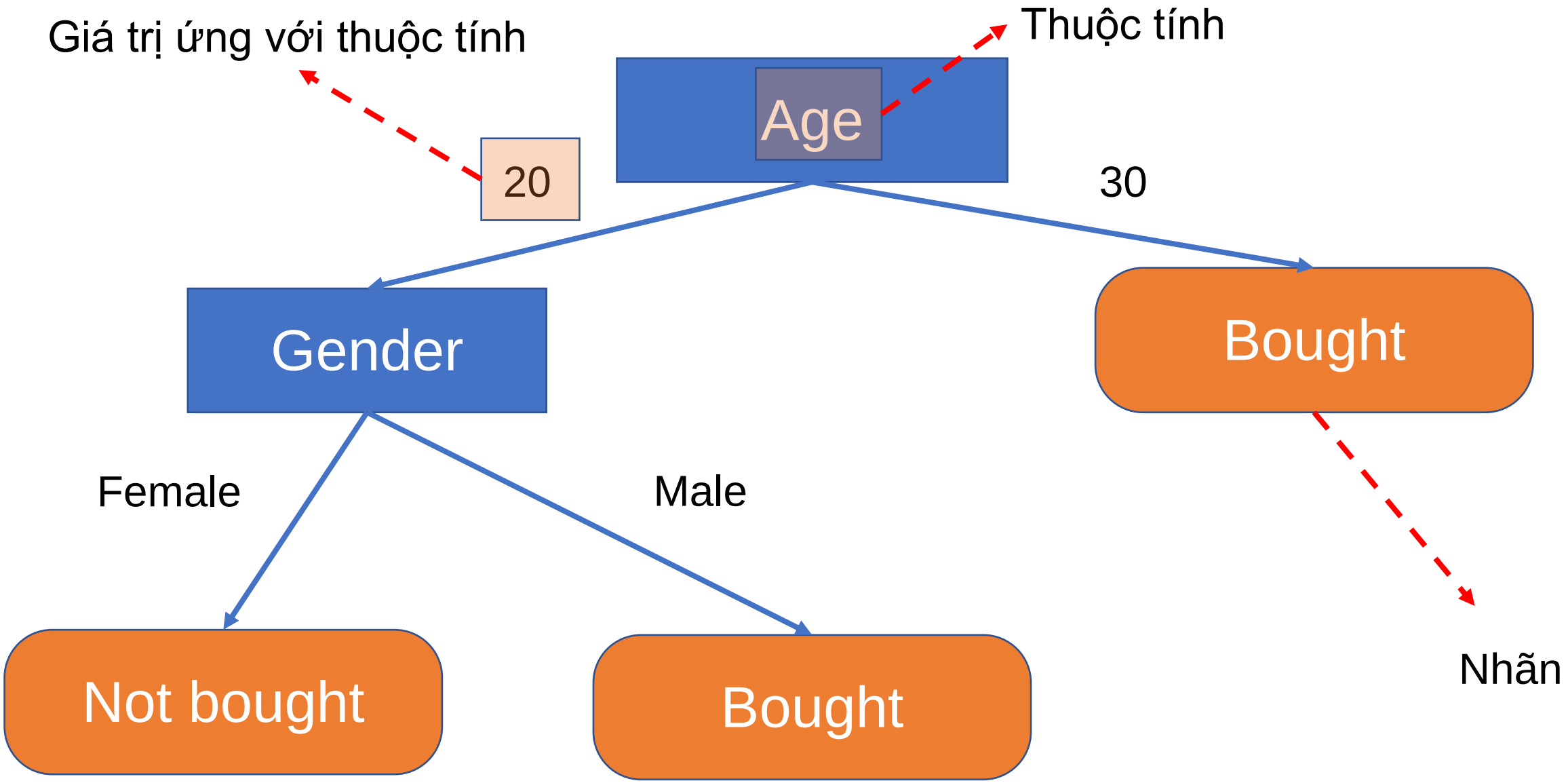
https://en.wikipedia.org/wiki/Logistic_function

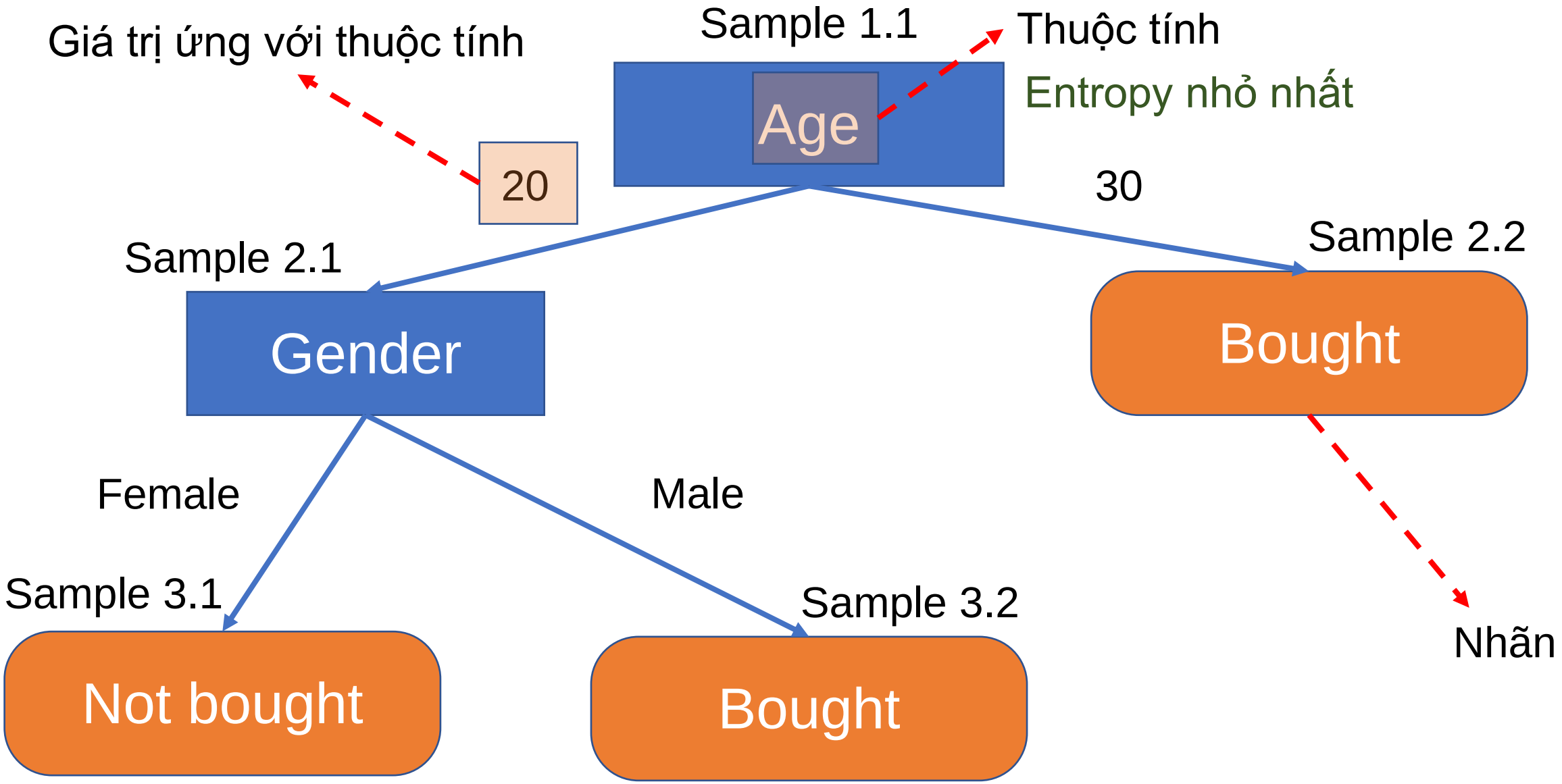

p̂ (Probability)
p̂ =99.4%
p̂ =85%
p̂ =23%
p̂ =0.7%
X
20
30
40
50

II. Nhóm mô hình CART

- **Phân loại** (classification) là hành động gán nhãn với x.
- **Nhãn** (label) là các giá trị rời rạc.
- Các thuật toán xây dựng cây để phân loại hay hồi quy gọi là **cây quyết định** và thuộc nhóm **CART** (classification and regression tree).


Age
20
30
Gender
Bought
Female
Male
Not bought
Bought


Giá trị ứng với thuộc tính
Thuộc tính
Age
20
30
Gender
Bought
Female
Male
Not bought
Bought
Nhãn


Giá trị ứng với thuộc tính
Sample 1.1
Thuộc tính
Entropy nhỏ nhất
Age
20
30
Sample 2.1
Gender
Sample 2.2
Bought
Female
Male
Sample 3.1
Sample 3.2
Not bought
Bought
Nhãn

```
def id3(x, x_attrs, node):
    node ← khởi tạo nút mới
    Nếu toàn bộ hàng trong x đều chỉ thuộc 1 nhãn label:
        Thêm nhãn label vào node
        Trả về node
    Nếu x rỗng:
        Thêm nhãn trội vào node
        Trả về node
    best_attr ← Thuộc tính có entropy nhỏ nhất
    Thêm best_attr vào node
    …
```

…

Với mỗi value thuộc tập giá trị độc nhất của best_attr:

Thêm value vào nút con mới của node

matchRows ← Các hàng mà giá trị best_attr chứa value

Nếu matchRows rỗng:

nextNode ← Nhãn trội

Ngược lại:

newY ← Danh sách các thuộc tính y mà không có best_attr

nextNode ← id3(matchRows, newY, label, node)

Trả về node

III. Thước đo đánh giá mô hình dự báo

Phân loại: Confusion matrix (ma trận hỗn loạn)

	Positive (PP)	Negative (PN)
Positive (P)	True positive	False negative
Negative (N)	False positive	True negative

III. Thước đo đánh giá mô hình dự báo

Phân loại: Confusion matrix (ma trận hỗn loạn)

	Positive (PP)	Negative (PN)
Positive (P)	True positive	False negative
Negative (N)	False positive	True negative

Phân loại: Confusion matrix (ma trận hỗn loạn)

		Dự đoán	
		Mua	Không mua
Sự thật	Mua	20	15
	Không mua	10	25

- Ordinary Least Square (OLS)
  - *Bình phương tối thiểu*
- Mean Squared Error (MSE)
  - *Sai số toàn phương trung bình*
- Root Mean Squared Error (RMSE)
  - *Sai số tiêu chuẩn*
- Mean Absolute Error (MAE)
  - *Sai số tuyệt đối trung bình*