TS. NGUYỄN QUỐC HÙNG

- Mobile: 0912 251 253
- Email: hungngq@ueh.edu.vn
- Website: https://bit.ueh.edu.vn/nqhung/

KHOA HỌC DỮ LIỆU

- Mã học phần: **25D1INF50905948**
- Thời gian: **11/04/2025 - 16/05/2025**
- Hệ: ĐH, Chính quy
- Số lượng: 48 sinh viên
- Số tín chỉ: 2.00

PHÂN LỚP DỮ LIỆU - CLASSIFICATION

Thứ 6, thời gian: 12g45-17g05, Giảng đường: N1-303

PHÂN LỚP DỮ LIỆU

NỘI DUNG

- **Bài toán phân lớp dữ liệu**
  - Giới thiệu phân lớp dữ liệu
  - Các ứng dụng phân lớp dữ liệu trong kinh tế
- **Một số phương pháp phân lớp**
  - Hồi quy Logistic (Logistic Regression)
  - Cây quyết định (Decision Tree)
  - SVM (Support Vector Machine)
- **Các phương pháp đánh giá mô hình phân lớp**
  - Ma trận nhầm lẫn (Confusion matrix)
  - Tính chính xác (Accuracy)
  - ROC, AUC, Precision/Recall
  - Cross Validation: Holdout và K-fold cross validation
- **Minh họa bằng công cụ Orange**

NỘI DUNG

- **Bài toán phân lớp dữ liệu**
  - Giới thiệu phân lớp dữ liệu
  - Các ứng dụng phân lớp dữ liệu trong kinh tế
- **Một số phương pháp phân lớp**
  - Hồi quy Logistic (Logistic Regression)
  - Cây quyết định (Decision Tree)
  - SVM (Support Vector Machine)
- **Các phương pháp đánh giá mô hình phân lớp**
  - Ma trận nhầm lẫn (Confusion matrix)
  - Độ chính xác (Accuracy)
  - ROC, AUC, Precision/Recall
  - Cross Validation: Holdout và K-fold cross validation
- **Minh họa bằng công cụ Orange**

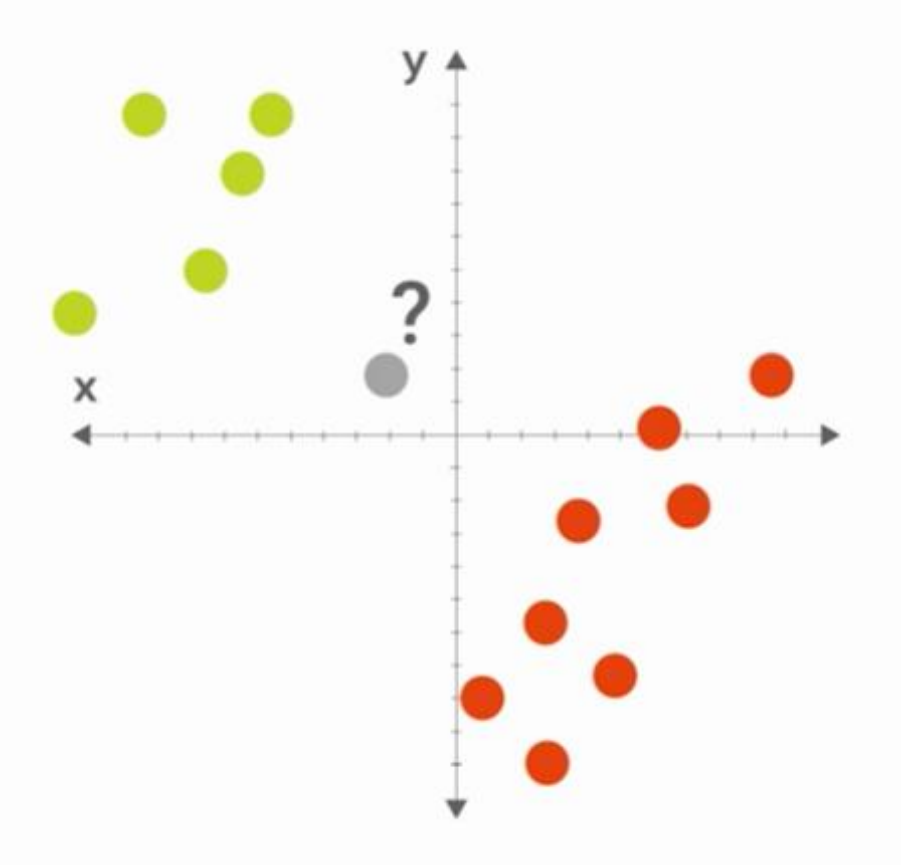
Giới thiệu phân lớp dữ liệu

Định nghĩa

Là quá trình phân một đối tượng dữ liệu vào **một** hay **nhiều** *lớp* (loại) đã cho trước nhờ một *mô hình phân lớp*.

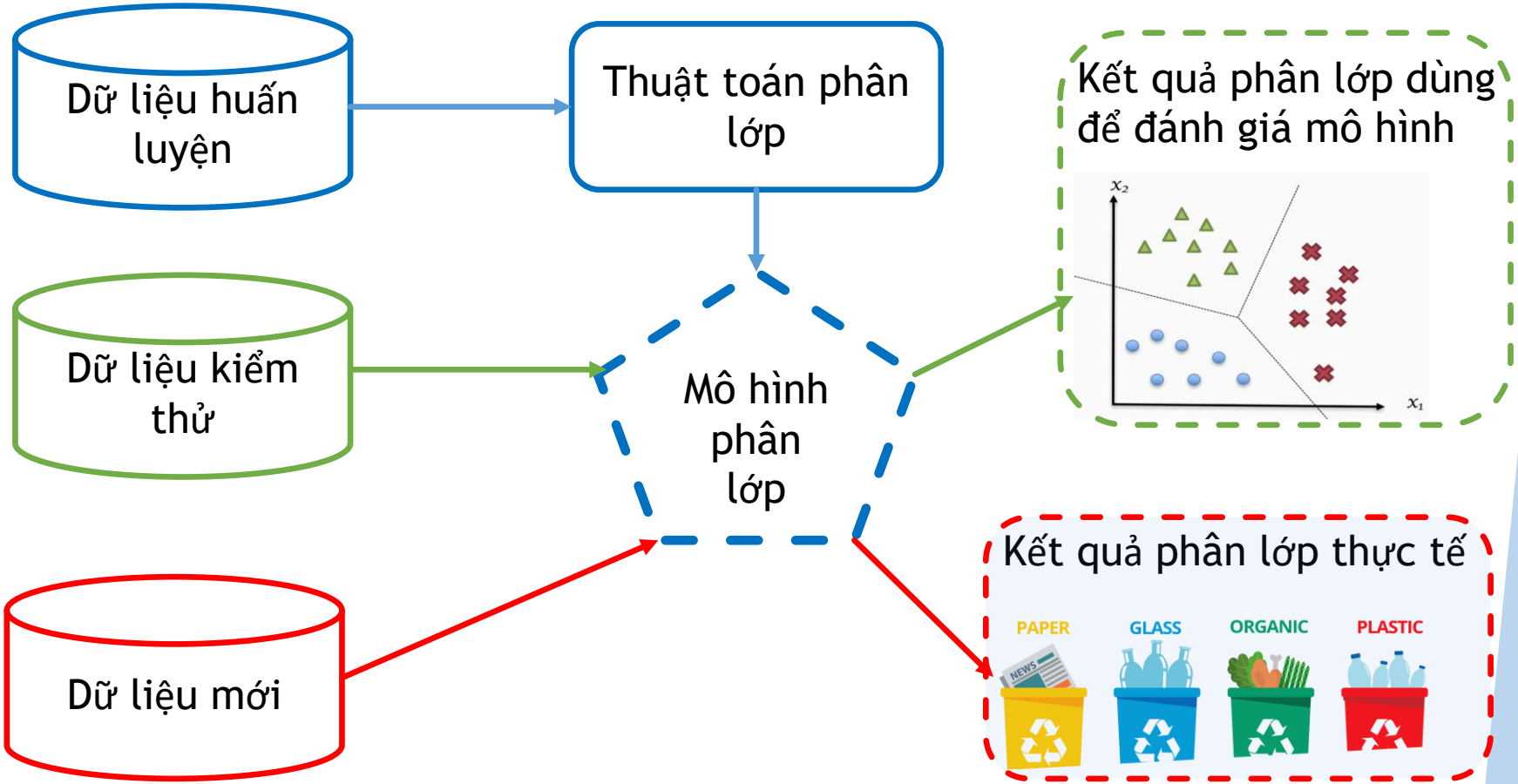
Mô hình này được xây dựng dựa trên một tập dữ liệu đã được gán nhãn trước đó (thuộc về lớp nào).

Quá trình gán nhãn (thuộc lớp nào) cho đối tượng dữ liệu chính là quá trình phân lớp dữ liệu.



Giới thiệu phân lớp dữ liệu

Quá trình phân lớp dữ liệu



Giới thiệu phân lớp dữ liệu

Quá trình phân lớp dữ liệu

Quá trình phân lớp dữ liệu gồm 2 bước chính:

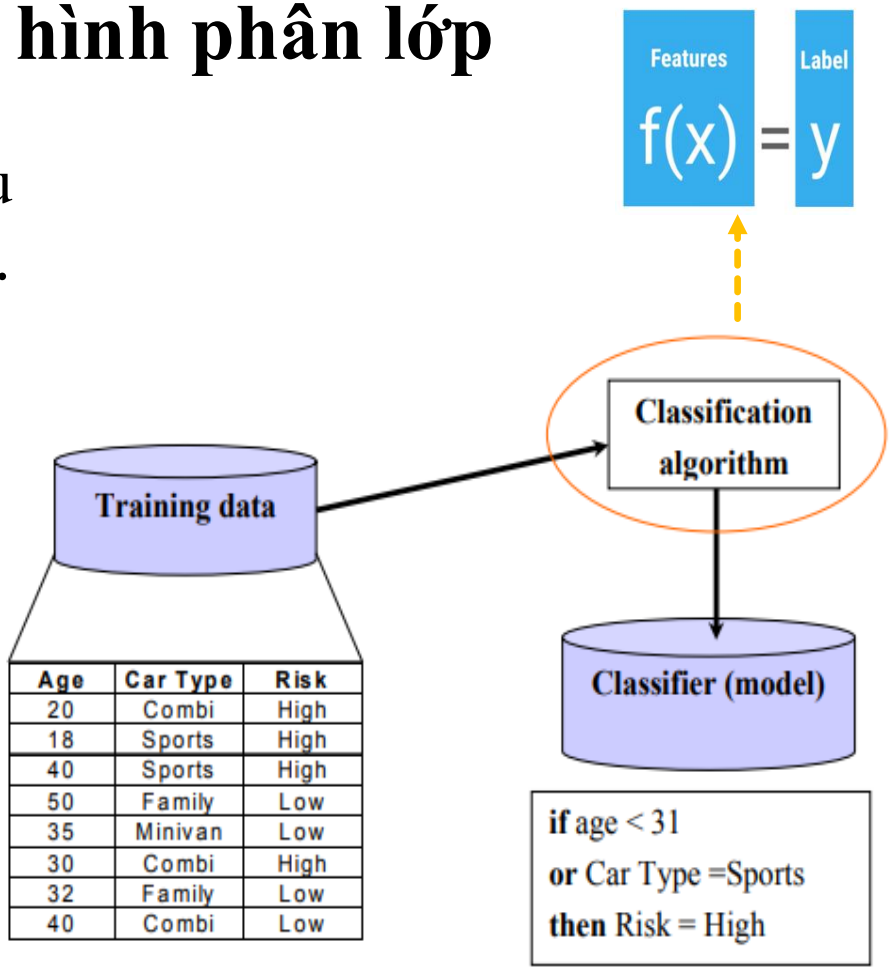
- **Bước 1**: Xây dựng mô hình (hay còn gọi là giai đoạn "học" hoặc "huấn luyện")

- **Bước 2**: Sử dụng mô hình chia thành 2 bước nhỏ.
  - *Bước 2.1*: Đánh giá mô hình (kiểm tra tính đúng đắn của mô hình)
  - *Bước 2.2*: Phân lớp dữ liệu mới

Giới thiệu phân lớp dữ liệu

Bước 1: Xây dựng mô hình phân lớp

- ✓ Dữ liệu đầu vào: là dữ liệu mẫu đã được gán nhãn và tiền xử lý.
- ✓ Các thuật toán phân lớp: cây quyết định, hàm số toán học, tập luật…
- ✓ Kết quả của bước này là **mô hình phân lớp** đã được huấn luyện (*trình phân lớp*)



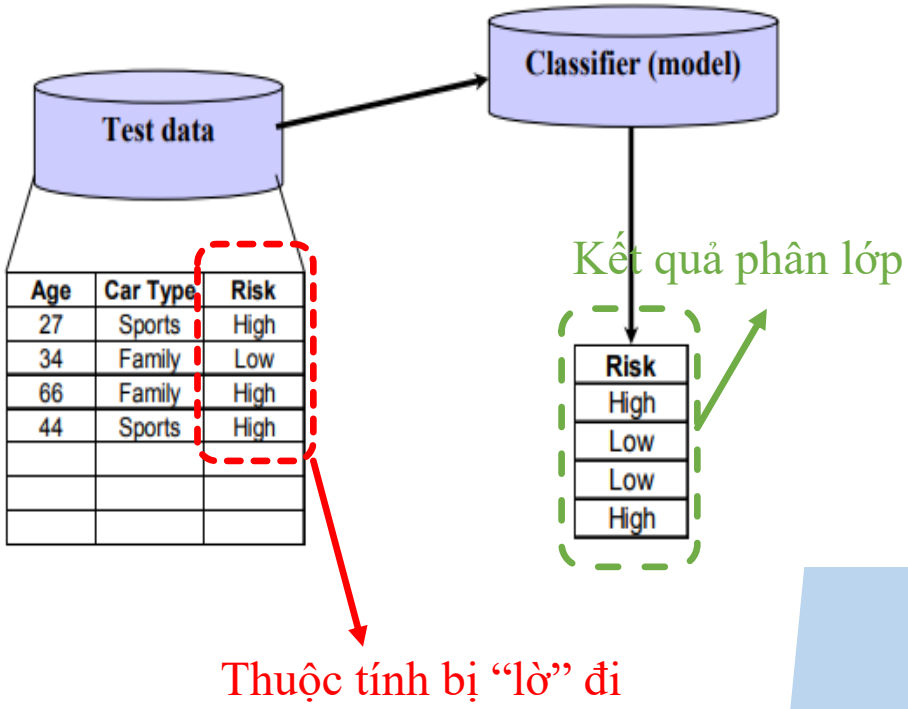
Age	Car Type	Risk
20	Combi	High
18	Sports	High
40	Sports	High
50	Family	Low
35	Minivan	Low
30	Combi	High
32	Family	Low
40	Combi	Low

Giới thiệu phân lớp dữ liệu

Bước 2.1: Đánh giá mô hình

- ✓ Dữ liệu đầu vào: là một tập dữ liệu mẫu khác đã được gán nhãn và tiền xử lý. Tuy nhiên lúc đưa vào mô hình phân lớp, ta "lờ" đi thuộc tính đã được gán nhãn.

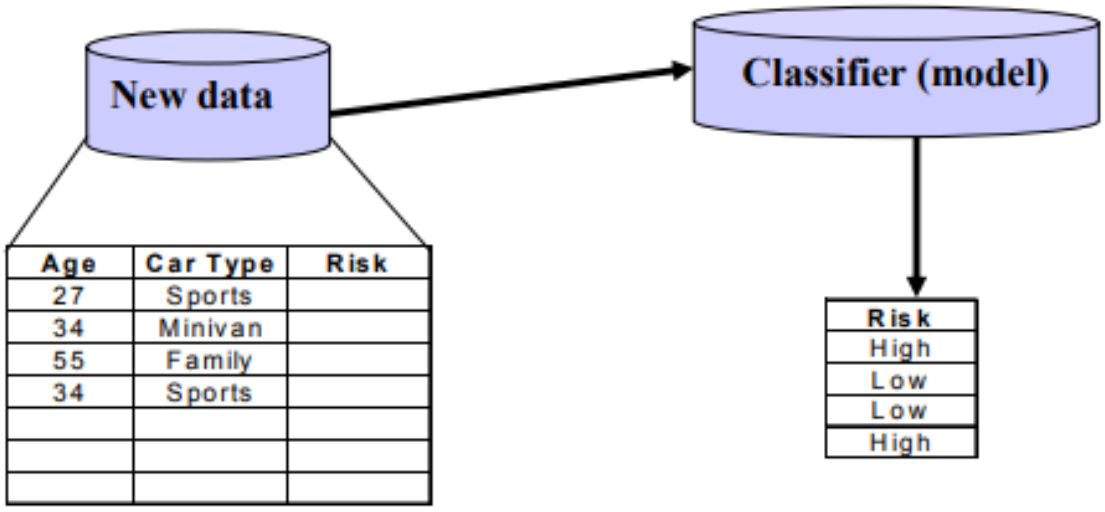
- ✓ Tính đúng đắn của mô hình sẽ được xác định bằng cách so sánh thuộc tính gán nhãn của dữ liệu đầu vào và kết quả phân lớp của mô hình.



Giới thiệu phân lớp dữ liệu

Bước 2.2: Phân lớp dữ liệu mới

- Dữ liệu đầu vào: là dữ liệu "khuyết" thuộc tính cần dự đoán lớp (nhãn)
- Mô hình sẽ tự động phân lớp (gán nhãn) cho các đối tượng dữ liệu này dựa vào những gì được huấn luyện ở bước 1



Giới thiệu phân lớp dữ liệu

Phân loại bài toán phân lớp

Nhiệm vụ của bài toán phân lớp là phân các đối tượng dữ liệu vào *n* lớp cho trước. Nếu:

- $n = 2$: **Phân lớp nhị phân**.
- $n > 2$: **Phân lớp đa lớp**.
- Mỗi đối tượng dữ liệu chỉ thuộc vào 1 lớp duy nhất: **Phân lớp đơn nhãn.**
- Một đối tượng dữ liệu có thể cùng lúc thuộc về nhiều lớp khác nhau: **Phân lớp đa nhãn**.

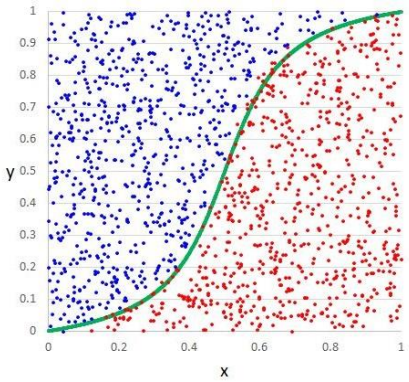
Các ứng dụng phân lớp dữ liệu trong kinh tế

- Tài chính ngân hàng
  - Dự báo giá chứng khoán
  - Xếp hạng tín dụng cá nhân và tổ chức
  - Đánh giá rủi ro tài chính
- Sales & Marketing
  - Dự báo doanh thu
  - Dự báo khách hàng trung thành
- Kinh tế học
  - Dự báo khủng hoảng kinh tế
  - Dự báo cung cầu

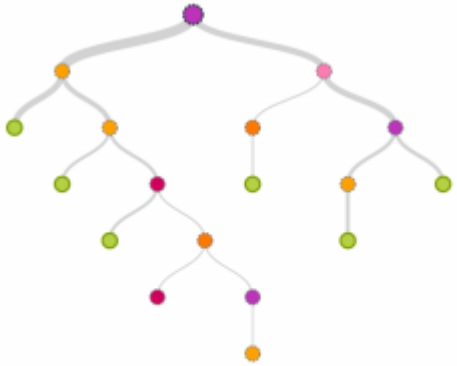
NỘI DUNG

- **Bài toán phân lớp dữ liệu**
  - Giới thiệu phân lớp dữ liệu
  - Các ứng dụng phân lớp dữ liệu trong kinh tế
- **Một số phương pháp phân lớp**
  - Hồi quy Logistic (Logistic Regression)
  - Cây quyết định (Decision Tree)
  - SVM (Support Vector Machine)
- **Các phương pháp đánh giá mô hình phân lớp**
  - Ma trận nhầm lẫn (Confusion matrix)
  - Độ chính xác (Accuracy)
  - ROC, AUC, Precision/Recall
  - Cross Validation: Holdout và K-fold cross validation
- **Minh họa bằng công cụ Orange**

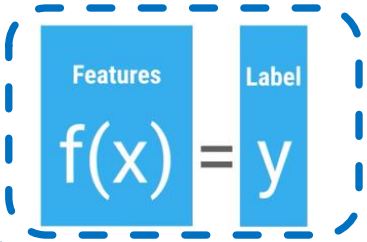
Một số phương pháp phân lớp

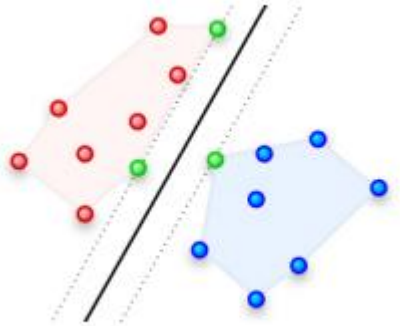


Logistic Regression

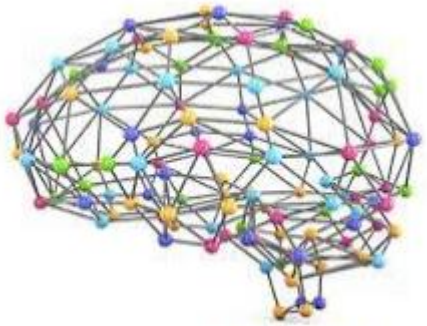


Decision Tree Induction
(ID3, C4.5,…)





Support Vector Machines



Neural Network

Một số phương pháp phân lớp

Hồi quy logistic (Logistic Regression)

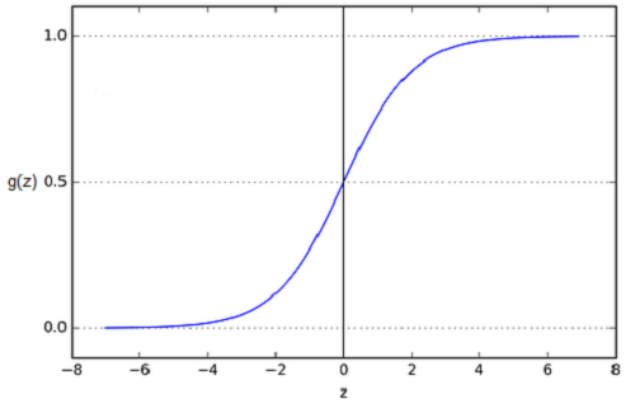
Định nghĩa: Là một mô hình xác suất dự đoán giá trị đầu ra rời rạc từ một tập các giá trị đầu vào (biểu diễn dưới dạng vector)

Mô tả: Đối với bài toán phân lớp:

Tập nhãn $\boldsymbol{y} = \{y_1, y_2, \dots, y_n\}$ *với* n *là số lớp*

Một đối tượng dữ liệu

$$\boldsymbol{x} = \{x_1, x_2, \dots, x_d\} \text{ với}$$



d là số thuộc tính của mỗi dòng dữ liệu và được biểu diễn dưới dạng vector

Hàm logistic $P(y = 1) = \frac{1}{1+e^{-(w_0+w_1x_1+w_2x_2+\cdots.+w_dx_d)}}$ *dự đoán đối tượng xem đối tượng x sở hữu các thuộc tính cụ thể sẽ thuộc vào lớp y nào.*

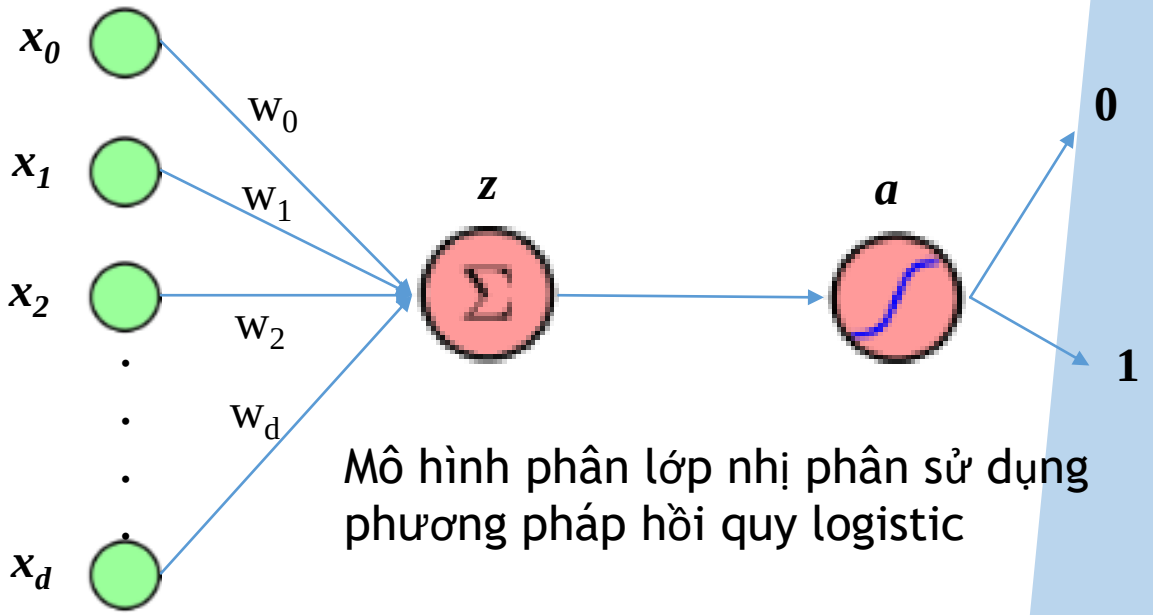
Một số phương pháp phân lớp

Hồi quy logistic (Logistic Regression)

Trong đó:
d là số lượng đặc trưng (thuộc tính) của dữ liệu,
w là trọng số, ban đầu sẽ được khởi tạo ngẫu nhiên, sau đó sẽ được điều chỉnh lại cho phù hợp

$$z = \sum_{d}^{i=0} w_i x_i$$

$$P(y) = sigmoid(z) = \frac{1}{1+e^{-z}}$$



Mô hình phân lớp nhị phân sử dụng phương pháp hồi quy logistic

Ví dụ đơn giản

- Xét bài toán phân lớp (nhị phân) phân lớp nguy cơ vỡ nợ trong tập khách hàng cá nhân để quyết định trong việc cho vay tiêu dùng:
  - Biến phân lớp (biến phụ thuộc): $y = 1$: vỡ nợ ; $y = 0$: không vỡ nợ.
  - Biến độc lập $x_1, x_2, \dots, x_d$ bao gồm: tuổi, học vấn, thu nhập, tài sản…
  - Hàm logistic (sigmoid): $P(y = 1) = \frac{1}{1+e^{-(w_0+w_1x_1+w_2x_2+\cdots+w_dx_d)}}$
  - Một ngưỡng t để phân lớp (nếu $P(y) \geq t$ thì phân vào lớp có thể vỡ nợ và ngược lại)
  - **Vấn đề**: cần tìm bộ hệ số (trọng số): $w_0, w_1, w_2 \dots w_d$ phù hợp để ước lượng

→ Bộ hệ số này sẽ được tính toán và điều chỉnh trong giai đoạn huấn luyện. Sau đó, sẽ được sử dụng trong quá trình đánh giá mô hình và phân lớp dữ liệu mới.

Một số phương pháp phân lớp

Cây quyết định (Decision Tree)

Khái niệm:

Trong lý thuyết quản trị, cây quyết định là đồ thị các quyết định cùng các kết quả khả dĩ đi kèm nhằm hỗ trợ quá trình ra quyết định.

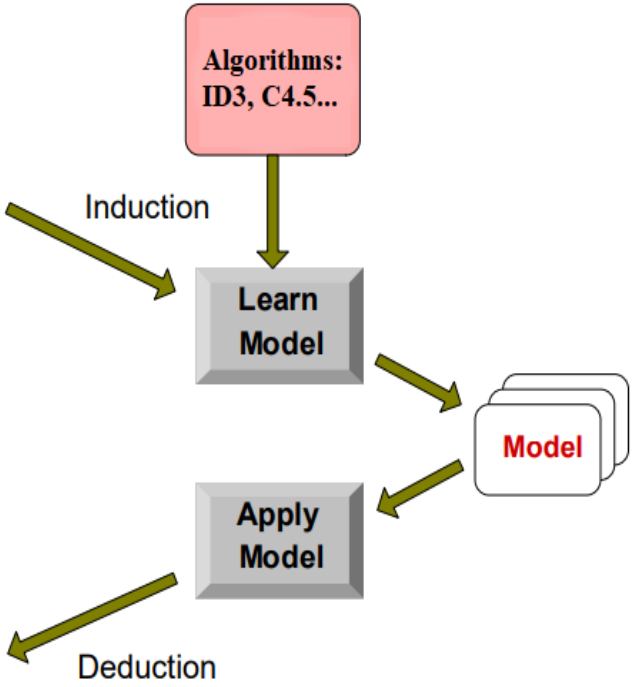
Trong lĩnh vực khai thác dữ liệu, cây quyết định là phương pháp nhằm mô tả, phân loại và tổng quát hóa tập dữ liệu cho trước.

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set



Một số phương pháp phân lớp

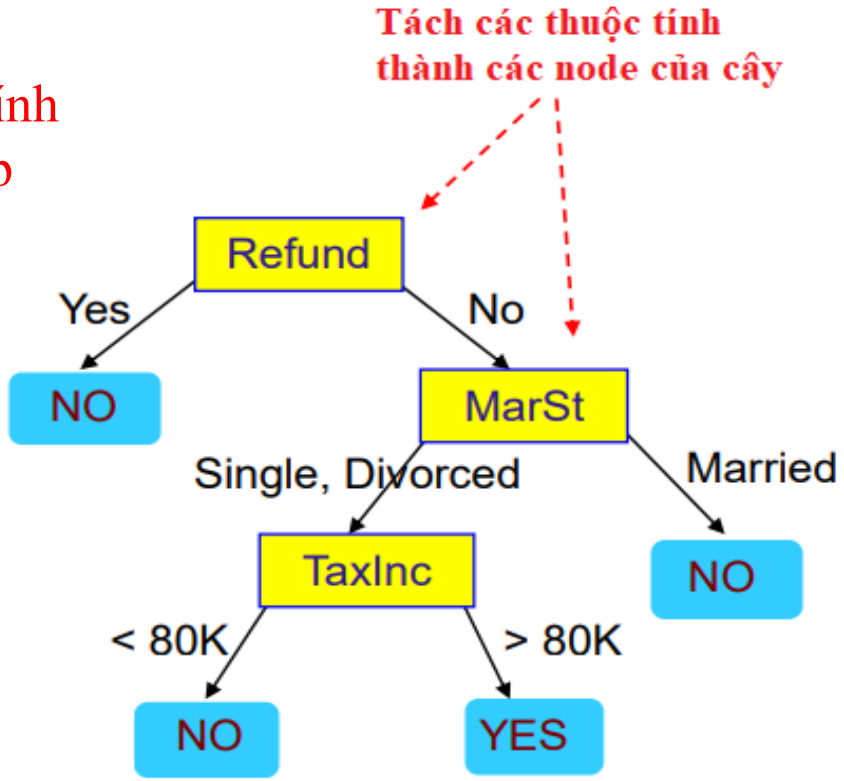
Cây quyết định (Decision Tree)

VÍ DỤ: XÂY DỰNG MÔ HÌNH CÂY QUYẾT ĐỊNH

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Dữ liệu huấn luyện

Thuộc tính phân lớp



Mô hình cây quyết định

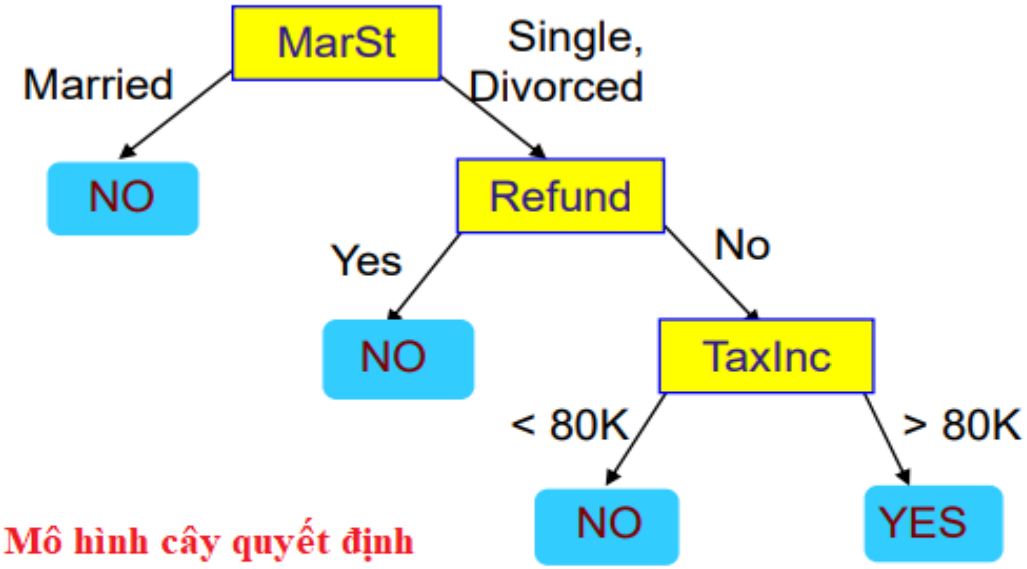
Một số phương pháp phân lớp

Cây quyết định (Decision Tree)

VÍ DỤ: XÂY DỰNG MÔ HÌNH CÂY QUYẾT ĐỊNH

Dữ liệu huấn luyện

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



Mô hình cây quyết định

Lưu ý:

- Một tập dữ liệu có thể được biểu diễn bởi nhiều cây quyết định tương ứng.

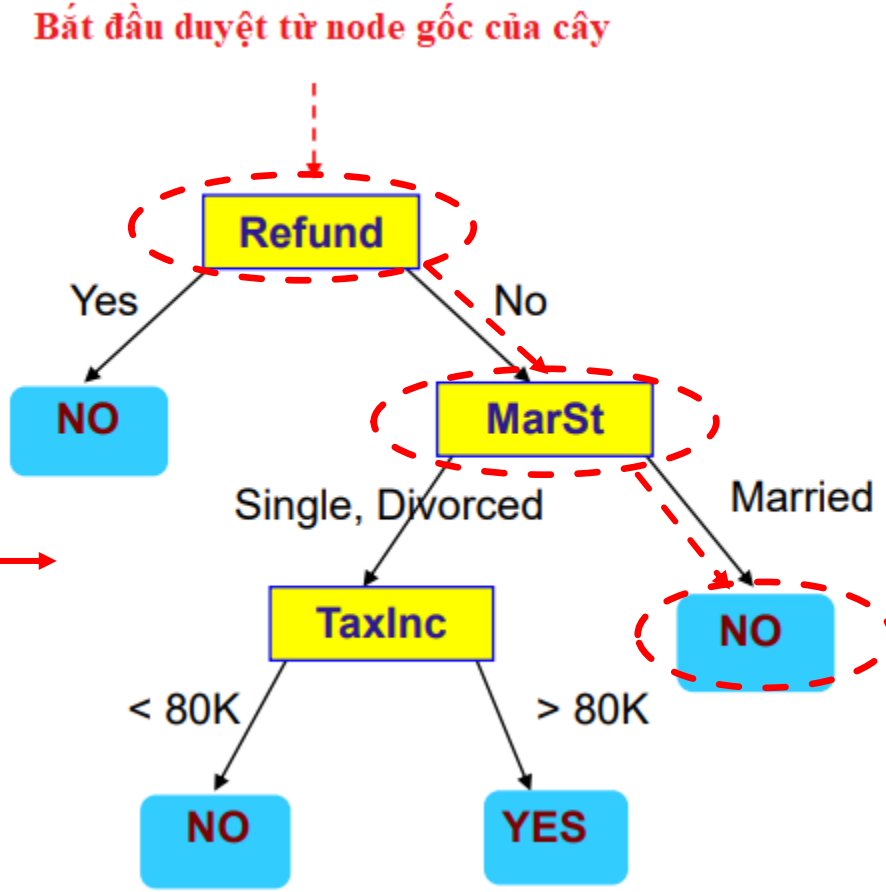
- Trong số đó, (theo nguyên lý Ockham's Razor) cây nào càng gọn thì càng tốt hơn

Cây quyết định (Decision Tree)

VÍ DỤ: PHÂN LỚP BẰNG MÔ HÌNH CÂY QUYẾT ĐỊNH

Dữ liệu đánh giá

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Kết quả: NO

Mô hình cây quyết định đã được xây dựng

Cây quyết định (Decision Tree)

Ưu khuyết điểm

Ưu điểm:

- Dễ hiểu
- Không đòi hỏi việc chuẩn hóa dữ liệu
- Có thể xử lý trên nhiều kiểu dữ liệu khác nhau.
- Xử lý tốt một lượng dữ liệu lớn trong thời gian ngắn

Khuyết điểm:

- Khó giải quyết trong tình huống dữ liệu phụ thuộc thời gian
- Chi phí xây dựng mô hình cao

Một số phương pháp phân lớp

SVM (Support Vector Machine)

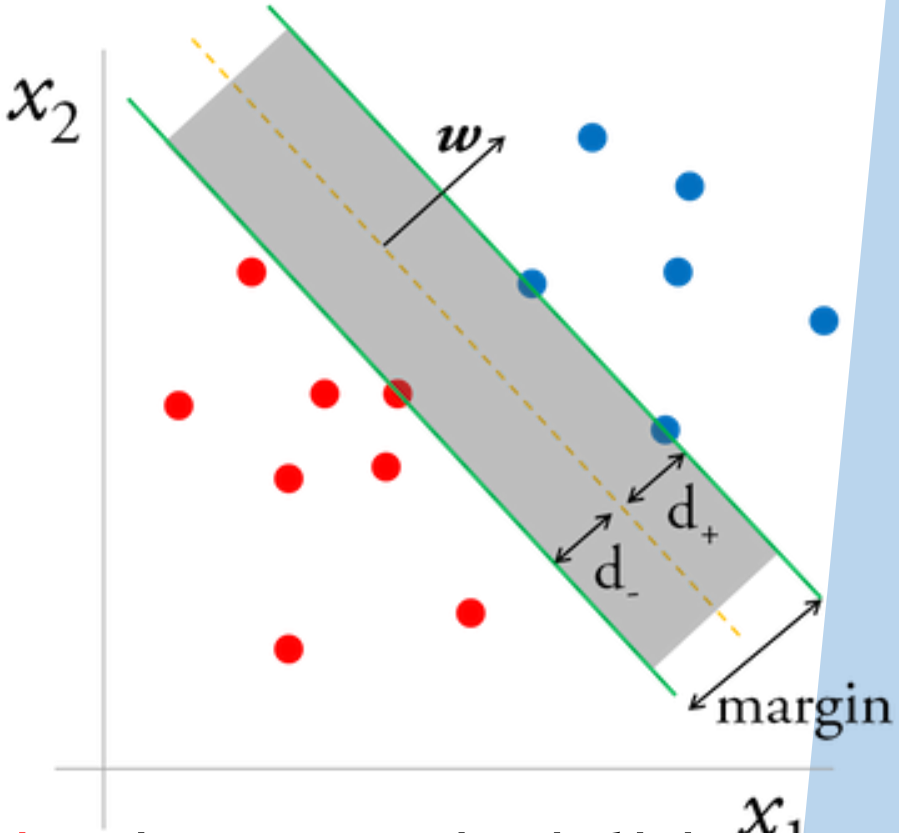
Giới thiệu:

- SVM là một thuật toán có giám sát, SVM nhận dữ liệu vào, xem chúng như những các vector trong không gian và phân loại chúng vào các lớp khác nhau bằng cách xây dựng một siêu phẳng trong không gian nhiều chiều làm mặt phân cách các lớp dữ liệu.
- Để tối ưu kết quả phân lớp thì phải xác định siêu phẳng (hyperplane) có khoảng cách đến các điểm dữ liệu (margin) của tất cả các lớp xa nhất có thể.
- SVM có nhiều biến thể phù hợp với các bài toán phân loại khác nhau.

SVM (Support Vector Machine)

Một số khái niệm:

- *Margin*: là khoảng cách giữa siêu phẳng (trong trường hợp không gian 2 chiều là đường thẳng) đến 2 điểm dữ liệu gần nhất tương ứng với 2 phân lớp.



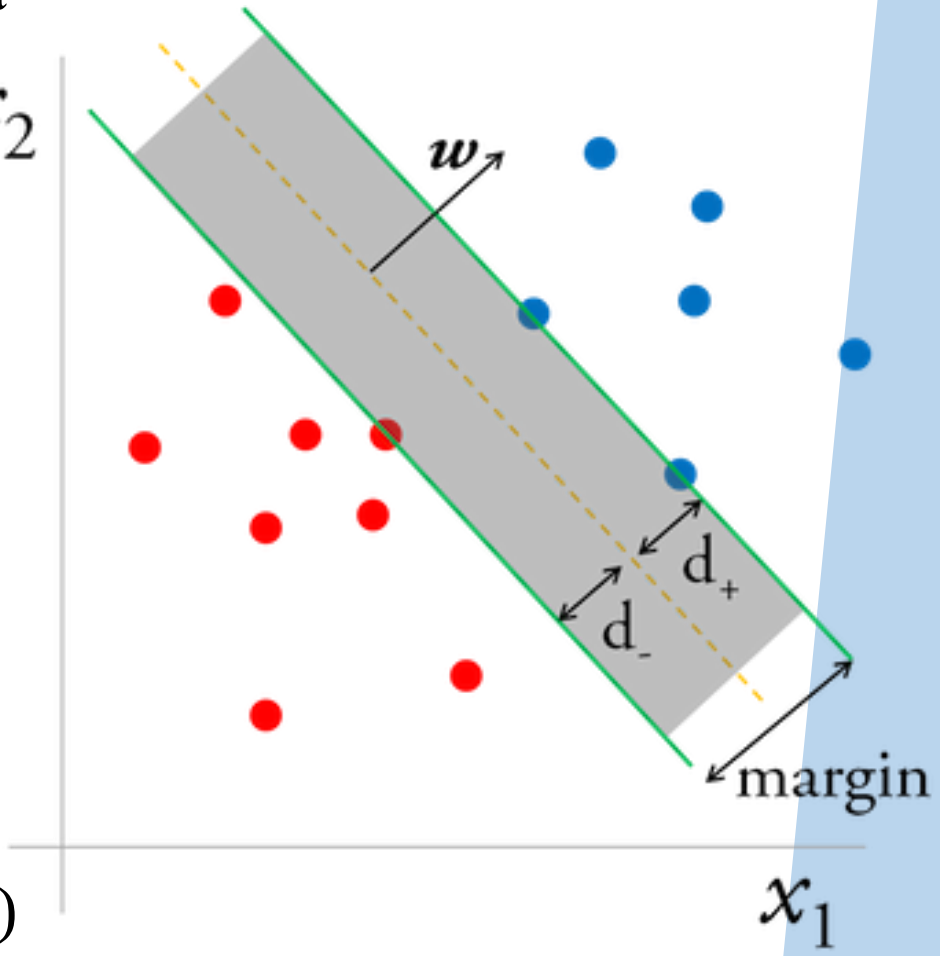
SVM cố gắng tối ưu bằng cách tối đa hóa giá trị margin này, từ đó tìm ra siêu phẳng đẹp nhất để phân 2 lớp dữ liệu. Nhờ vậy, SVM có thể giảm thiểu việc phân lớp sai (misclassification) đối với điểm dữ liệu mới đưa vào.

SVM (Support Vector Machine)

➢ Support Vectors:

Bài toán của chúng ta trở thành tìm ra 2 đường biên của 2 lớp dữ liệu sao cho khoảng cách giữa 2 đường này là lớn nhất. Siêu phẳng cách đều 2 biên đó chính là siêu phẳng cần tìm.

Các điểm xanh, đỏ nằm trên 2 đường biên (màu xanh lá) được gọi là các support vector, vì chúng có nhiệm vụ hỗ trợ để tìm ra siêu phẳng (màu cam)



SVM (Support Vector Machine)

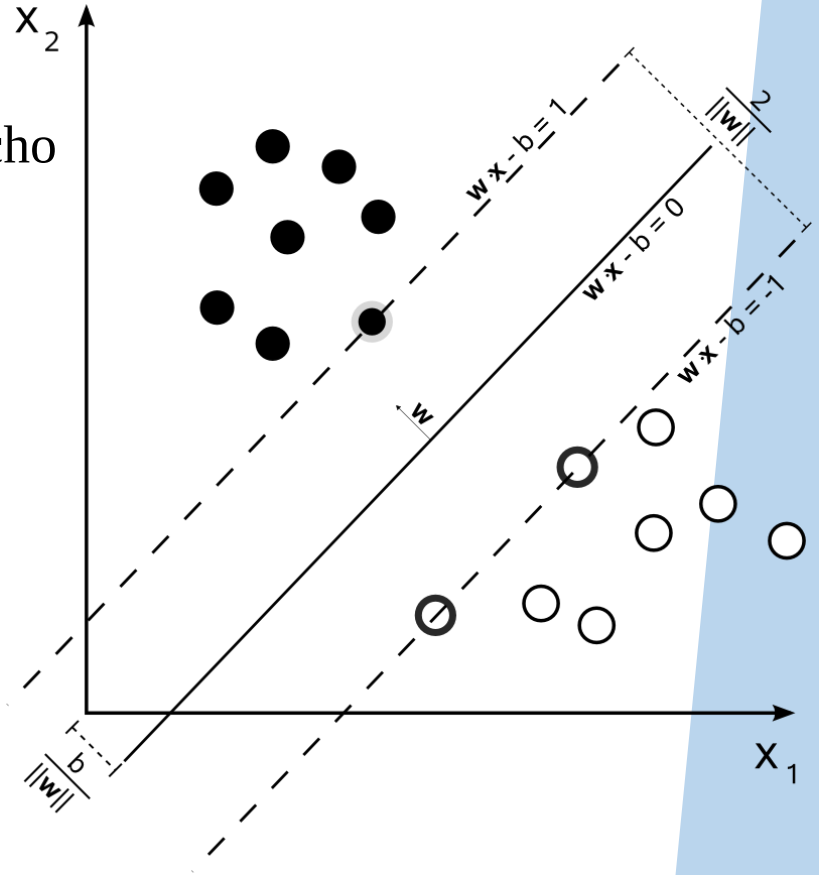
VÍ DỤ: Bài toán trong không gian hai chiều

- ✓ y_i: là các lớp chứa các điểm dữ liệu x_i. Ở ví dụ này y mang giá trị 1 và -1 (có 2 lớp)
- ✓ x_i : là một vector thực nhiều chiều đại diện cho một đối tượng dữ liệu cụ thể.
- ✓ Giả sử 2 đường thẳng song song đi qua các support vector của 2 lớp dữ liệu lần lượt là: $w_1x_1 + w_2x_2 + b = 1$

$$w_1x_1 + w_2x_2 + b = -1$$

Viết gọn lại là:

$$wx - b = 1 \text{ và } wx - b = -1.$$



Trong đó:

- w: là một vector pháp tuyến
- Khoảng cách giữa hai đường thẳng chính là margin $= 2/\sqrt{{w_1}^2 + {w_2}^2}$
- Khi đó đường thẳng phân cách cần tìm là: $w_1x_1 + w_2x_2 + b = 0$ hay $wx - b = 0$

SVM (Support Vector Machine)

Tổng quát hóa trong không gian nhiều chiều

- Số chiều của không gian bài toán (còn gọi là không gian đặc trưng) tương ứng với số lương thuộc tính (đặc trưng) của một đối tượng dữ liệu.
- Phương trình biểu diễn siêu phẳng cần tìm (hyperlane) trong không gian đa chiều là: $w^T x + b = 0$ và giá trị margin = $\frac{2|w^T x+b|}{\|w\|} = \frac{2}{\|w\|}$
- Mục tiêu của SVM là cần tìm giá trị margin cực đại đồng nghĩa với việc $\|w\|$ đạt cực tiểu với điều kiện:

$$y_n(\mathbf{w}^T \mathbf{x}_n + b) \geq 1, \forall n = 1, 2, \ldots, N$$

- Hàm mục tiêu cần tối ưu là một norm nên là một hàm lồi => bài toán quy hoạch toàn phương (Quadratic Programing)

SVM (Support Vector Machine)

Các biến thể của SVM

Loại SVM	Tính chất
Hard Margin SVM	Hai lớp cần phân lớp là có thể phân chia tuyến tính (linearly seperable)
Soft Margin SVM	Hai lớp cần phân lớp là "gần" phân chia tuyến tính (almost linear seperable)
Multi-class SVM	Phân lớp đa lớp (biên giữa các lớp là tuyến tính)
Kernel SVM	Dữ liệu là phi tuyến

Các biến thể của SVM

- Ưu điểm:
  - Tiết kiệm bộ nhớ (do quá trình test chỉ cần so điểm dữ liệu mới với mặt siêu phẳng tìm được mà không cần tính toán lại)
  - Linh hoạt: vừa có thể phân lớp tuyến tính và phi tuyến (sử dụng các kernel khác nhau)
  - Xử lý được trong không gian nhiều chiều
- Khuyết điểm:
  - Trong trường hợp số chiều dữ liệu lớn hơn số dòng dữ liệu thì SVM cho kết quả không tốt.
  - Chưa thể hiện tính xác suất trong phân lớp.

NỘI DUNG

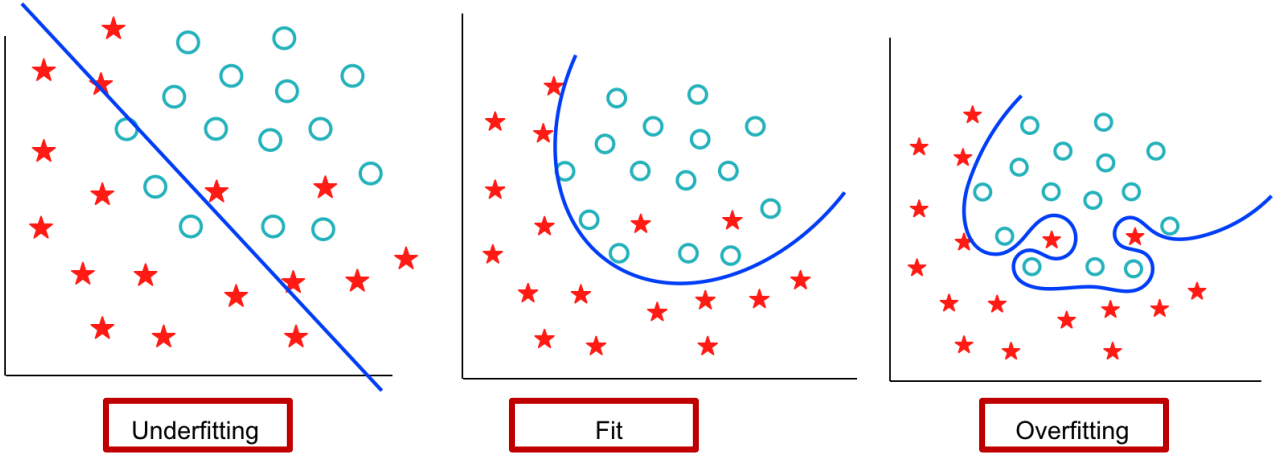
- **Bài toán phân lớp dữ liệu**
  - Giới thiệu phân lớp dữ liệu
  - Các ứng dụng phân lớp dữ liệu trong kinh tế
- **Một số phương pháp phân lớp**
  - Hồi quy Logistic (Logistic Regression)
  - Cây quyết định (Decision Tree)
  - SVM (Support Vector Machine)
- **Các phương pháp đánh giá mô hình phân lớp**
  - Ma trận nhầm lẫn (Confusion matrix)
  - Tính chính xác (Accuracy)
  - ROC, AUC, Precision/Recall
  - Cross Validation: Holdout và K-fold cross validation
- **Minh họa bằng công cụ Orange**

Các phương pháp đánh giá mô hình phân lớp

Khái niệm:

Là các phương pháp nhằm kiểm tra tính hiệu quả của mô hình phân lớp trên dữ liệu có đặc thù cụ thể, từ đó quyết định có sử dụng mô hình đó hay không.

Một mô hình lý tưởng là một mô hình không quá đơn giản, không quá phức tạp và không quá nhạy cảm với nhiễu (tránh underfitting và overfitting).



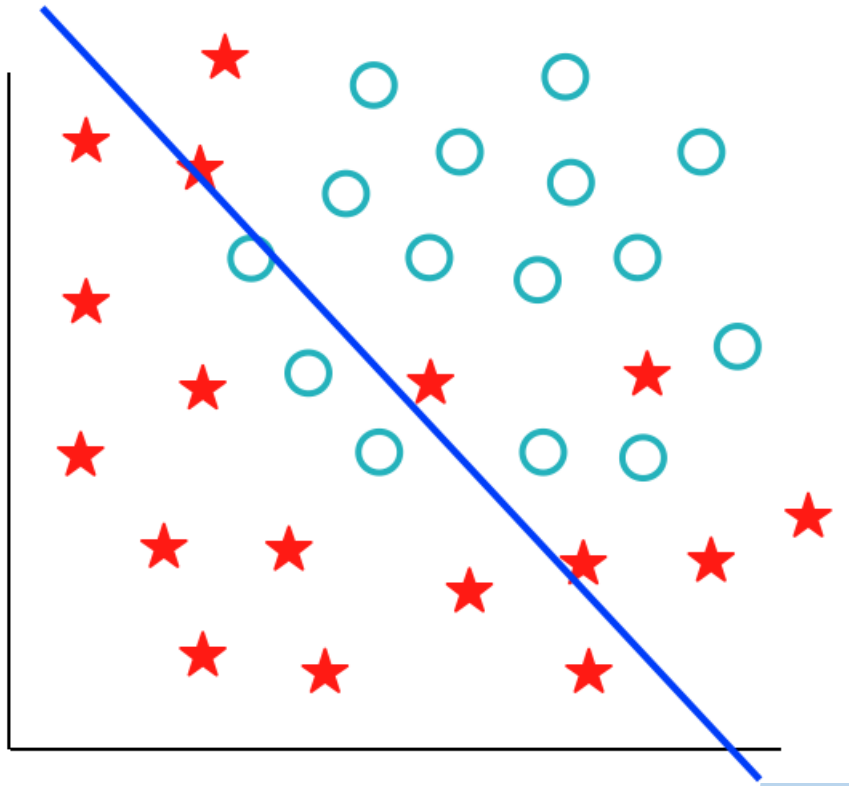
Các phương pháp đánh giá mô hình phân lớp

Underfitting (chưa khớp):

Mô hình được coi là chưa khớp nếu nó chưa được chưa phù hợp với tập dữ liệu huấn luyện và cả các mẫu mới khi dự đoán.

Nguyên nhân có thể là do mô hình chưa đủ độ phức tạp cần thiết để bao quát được tập dữ liệu.

Tồn tại nhiều điểm dữ liệu mà mô mình không phân loại được đúng dẫn đến độ chính xác mô hình thấp.

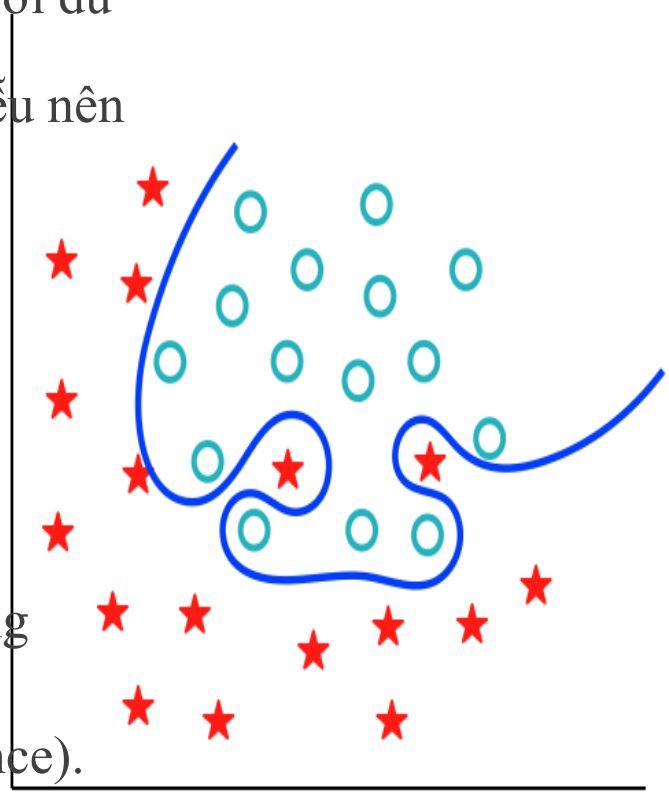


Các phương pháp đánh giá mô hình phân lớp

Overfitting (quá khớp):

Overfitting là hiện tượng mô hình tìm được *quá khớp* với dữ liệu huấn luyện. Điều này dẫn đến việc dự đoán cả nhiễu nên mô hình không còn tốt khi phân lớp trên dữ liệu mới.

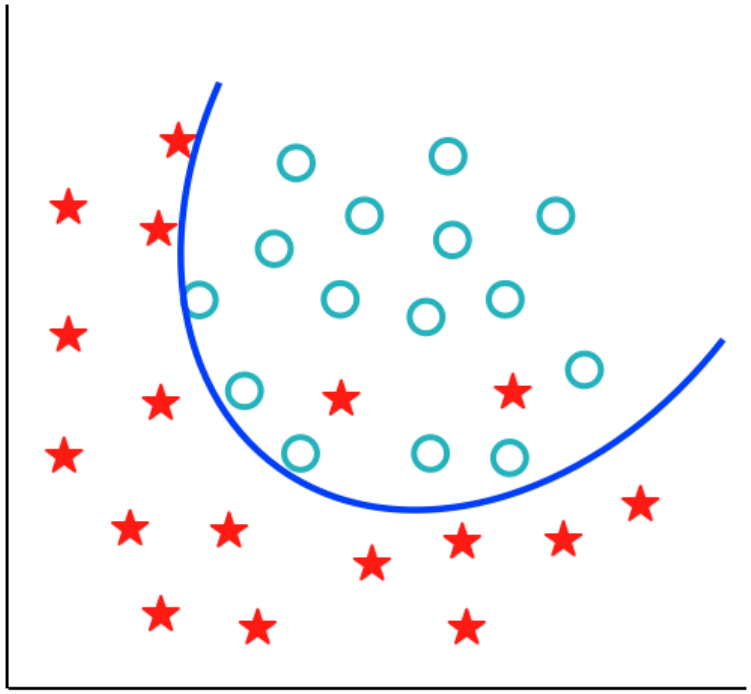
Quá khớp xảy ra khi lượng dữ liệu huấn luyện quá nhỏ trong khi độ phức tạp của mô hình quá cao nên mặc dù độ chính xác cao nhưng không thể mô tả được xu hướng tổng quát của dữ liệu mới (còn được gọi là High Variance).



Các phương pháp đánh giá mô hình phân lớp

Good fitting: Là trường hợp mô hình cho ra kết quả hợp lý với cả tập dữ liệu huấn luyện và các giá trị mới, tức mang tính tổng quát.

Ngoài thực tế mô hình tốt là mô hình cho kết quả hợp lý một cách chấp nhận được trên dữ liệu mẫu lẫn dữ liệu mới.



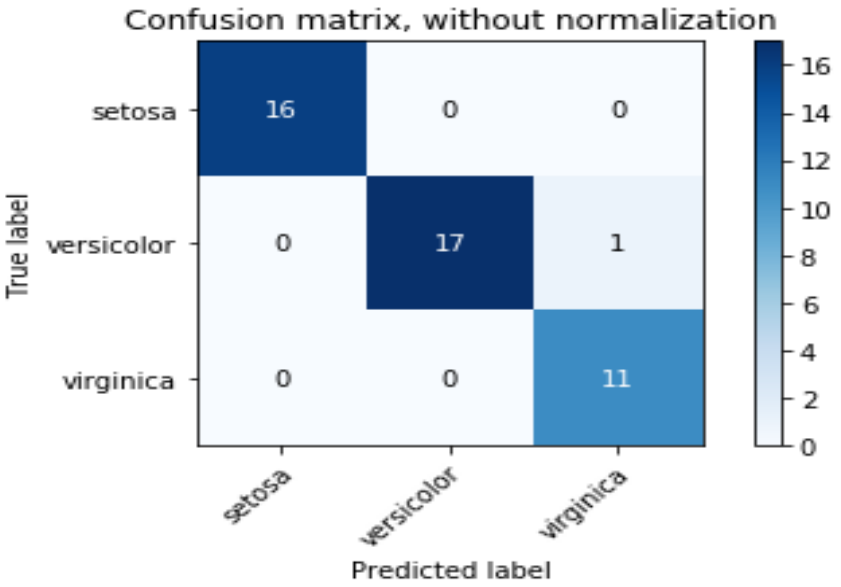
Trong tất cả các giả thiết có thể giải thích được một hiện tượng, ta nên chọn giả thiết đơn giản nhất (Occam's razor)

=> Do đó, trong tất cả các model "**đúng**", chọn model **đơn giản** nhất.

Ma trận nhầm lẫn (Confusion Matrix)

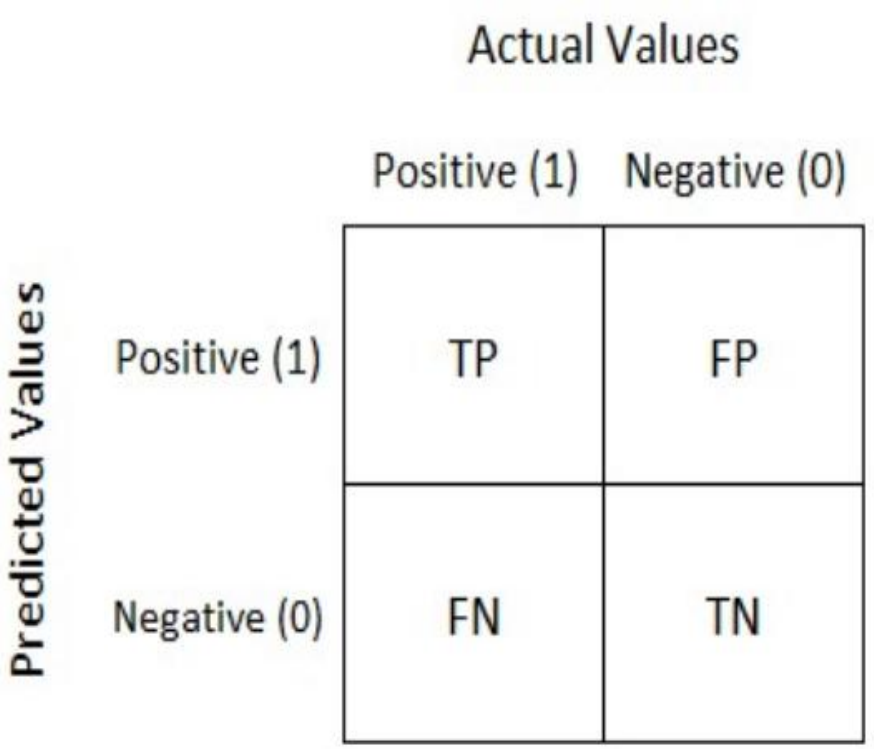
- Ma trận nhầm lẫn: là ma trận chỉ ra có bao nhiêu điểm dữ liệu thực sự thuộc vào một lớp cụ thể, và được dự đoán là rơi vào lớp nào.
- Confusion matrix là có kích thước *k x k* với *k* là số lượng lớp của dữ liệu.

VD: Ở bộ dữ liệu Iris có 3 nhãn dữ liệu là: Sentosa, versicolor và virginica và có ma trận nhầm lẫn được vẽ bằng python như hình bên dưới.
Giá trị tại ô *(i;j)* cho biết số lượng mẫu *i* bị phân vào lớp *j*.



Ví dụ về ma trận nhầm lẫn

- Bài toán chuẩn đoán ung thư ta có 2 lớp: lớp bị ung thư được chuẩn đoán Positive và lớp không bị ung thư được chuẩn đoán là Negative:
- **TP (True Positive)**: Số lượng dự đoán chính xác. Là khi mô hình dự đoán đúng một người bị ung thư.
- **TN (True Negative)**: Số lượng dự đoán chính xác một cách gián tiếp. Là khi mô hình dự đoán đúng một người không bị ung thư, tức là việc không chọn trường hợp bị ung thư là chính xác.
- **FP (False Positive - Type 1 Error)**: Số lượng các dự đoán sai lệch. Là khi mô hình dự đoán một người bị ung thư nhưng người đó hoàn toàn khỏe mạnh.
- **FN (False Negative - Type 2 Error)**: Số lượng các dự đoán sai lệch một cách gián tiếp. Là khi mô hình dự đoán một người không bị ung thư nhưng người đó bị ung thư, tức là việc không chọn trường hợp bị ung thư là sai.



Accuracy (tính chính xác)

- Là tỷ lệ số mẫu được phân lớp đúng trong toàn bộ tập dữ liệu.

$acc = {}^{(TP+TN)}/_{n} => Error\ rate = 1 - acc$ *là độ lỗi của mô hình*

- **Accuracy** chỉ cho chúng ta biết được tỷ lệ dữ liệu được phân loại đúng mà không chỉ ra được cụ thể mỗi loại được phân loại như thế nào, lớp nào được phân loại đúng nhiều nhất, và dữ liệu thuộc lớp nào thường bị phân loại nhầm vào lớp khác.

		Actual Values: Positive (1)	Actual Values: Negative (0)
Predicted Values	Positive (1)	TP	FP
Predicted Values	Negative (0)	FN	TN

Precision, Recall, F_1- score

- Precision (độ chính xác): cho biết trong số *m* mẫu được phân vào lớp *i* thì có tỷ lệ bao nhiêu mẫu có đúng (tránh nhầm lẫn với tính chính xác accuracy)

$$precision = {TP}/{(TP + FP)}$$

- Recall (độ phủ) còn gọi là độ phủ hay độ nhạy (sensitivity) hay TPR (True Positive Rate)

$$recall = {TP}/{(TP + FN)}$$

- F_1-score: giá trị trung bình điều hòa (harmonic mean) của hai độ đo *Precision* và *Recall*.

$$F_1 = 2\frac{(precision \times recall)}{(precision + recall)}$$

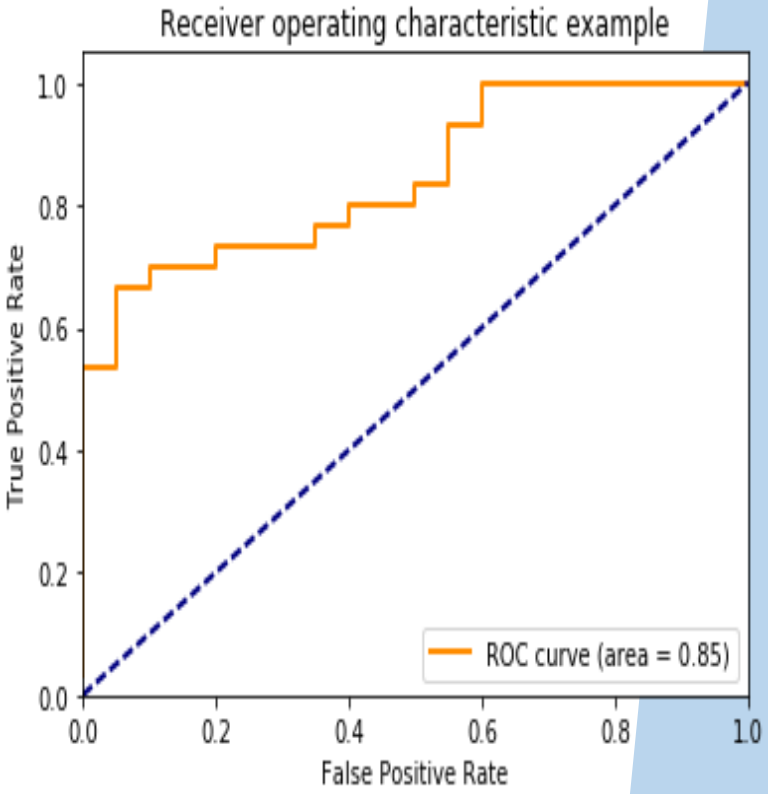
F_1 có giá trị gần với giá trị nào nhỏ hơn giữa 2 giá trị Precision và Recall.
F_1 sẽ có giá trị lớn nếu cả 2 giá trị Precision và Recall đều lớn.

ROC và AUC

- ROC (Receiver Operating Characteristic)

Là một đồ thị được sử dụng khá phổ biến trong đánh giá các mô hình phân loại nhị phân. Đường cong này được tạo ra bằng cách biểu diễn tỷ lệ dự báo true positive rate (TPR) dựa trên tỷ lệ dự báo false positive rate (FPR) tại các ngưỡng khác nhau.

Một mô hình hiệu quả khi có FPR thấp và TPR cao, hay ROC càng tiệm cận với điểm (0;1) trong đồ thị thì mô hình càng hiệu quả.



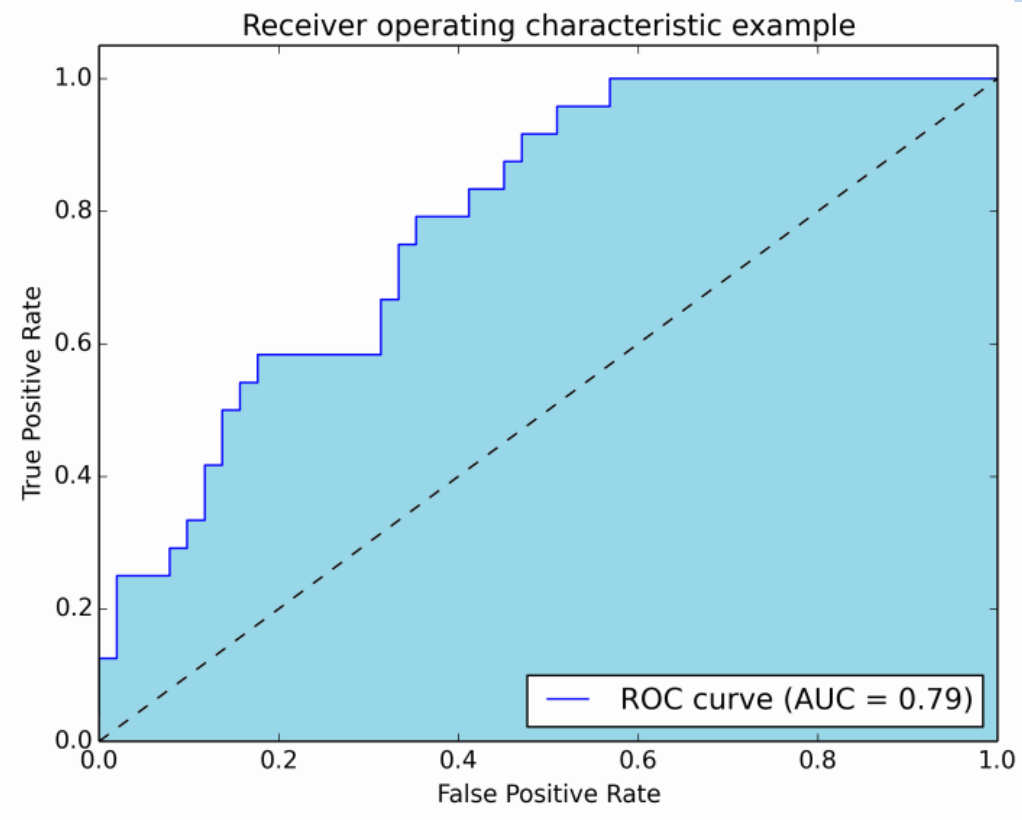
ROC và AUC

- AUC (Area Under the Curve)

Là diện tích nằm dưới đường cong ROC.

Giá trị này là một số dương nhỏ hơn hoặc bằng 1.

Giá trị này càng lớn thì mô hình càng tốt.

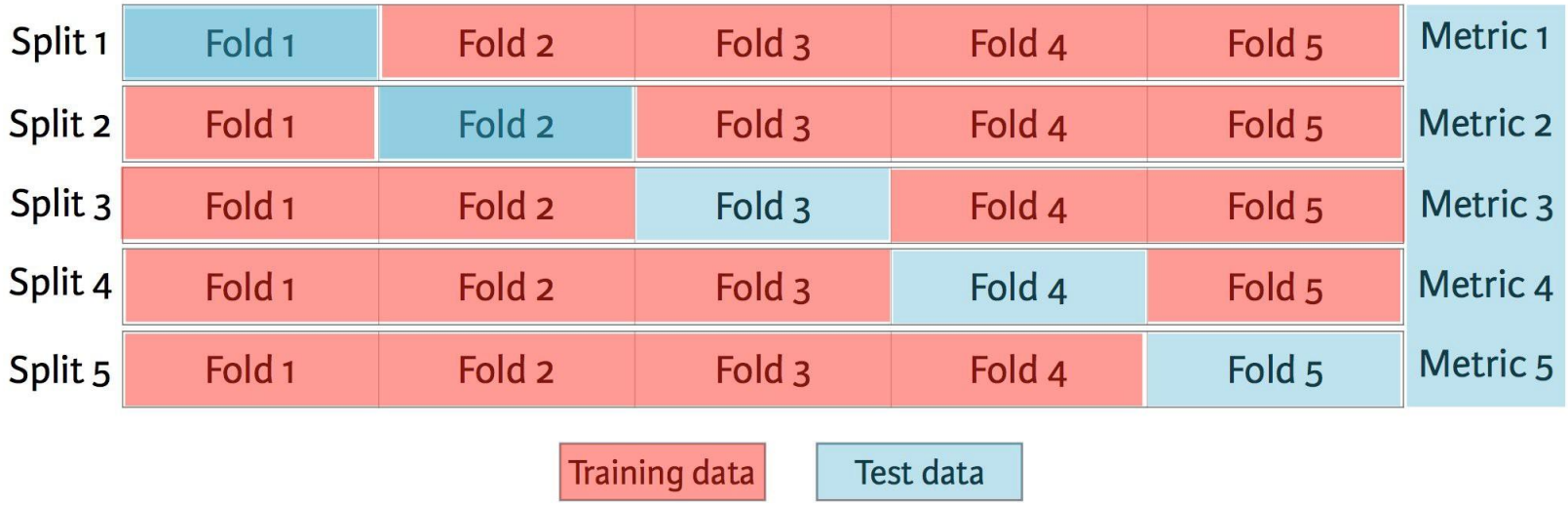


Phương pháp phân chia dữ liệu Hold-out

- Phương pháp Hold-out phân chia tập dữ liệu ban đầu thành 2 tập độc lập theo 1 tỷ lệ nhất định. Ví dụ, tập huấn luyện (training set) chiếm 70%, tập thử nghiệm (testing set) chiếm 30%.
- Phương pháp này thích hợp cho các tập dữ liệu nhỏ. Tuy nhiên, các mẫu có thể không đại diện cho toàn bộ dữ liệu (thiếu lớp trong tập thử nghiệm).
- Có thể cải tiến bằng cách dùng phương pháp lấy mẫu sao cho mỗi lớp được phân bố đều trong cả 2 tập dữ liệu huấn luyện và đánh giá. Hoặc lấy mẫu ngẫu nhiên : thực hiện holdout k lần và độ chính xác acc(M) = trung bình cộng k giá trị chính xác.

K-fold cross validation

- Phương pháp này phân chia dữ liệu thành k tập con có cùng kích thước (gọi là các ***fold***).
- Một trong các ***fold*** được sử dụng làm tập dữ liệu đánh giá và phần còn lại được sử dụng làm tập huấn luyện.
- Quá trình lặp lại cho đến khi tất cả các fold đều đã được dùng làm tập dữ liệu đánh giá.
- Xét ví dụ: 5-fold cross validation



Hold-out vs. K-fold cross validation

- Phương pháp K-fold thường được sử dụng nhiều hơn do mô hình sẽ được huấn luyện và đánh giá trên nhiều phần dữ liệu khác nhau. Từ đó tang độ tin cậy cho các độ đo đánh giá của mô hình.
- Phương pháp Hold-out thường cho hiệu quả tốt trên các tập dữ liệu lớn. Tuy nhiên, ở các tập dữ liệu nhỏ hoặc vừa phải, hiệu quả của mô hình sử dụng phương pháp này phụ thuộc nhiều vào cách chia cũng như tỷ lệ chia dữ liệu.

NỘI DUNG

- **Bài toán phân lớp dữ liệu**
  - Giới thiệu phân lớp dữ liệu
  - Các ứng dụng phân lớp dữ liệu trong kinh tế
- **Một số phương pháp phân lớp**
  - Hồi quy Logistic (Logistic Regression)
  - Cây quyết định (Decision Tree)
  - SVM (Support Vector Machine)
- **Các phương pháp đánh giá mô hình phân lớp**
  - Ma trận nhầm lẫn (Confusion matrix)
  - Độ chính xác (Accuracy)
  - ROC, AUC, Precision/Recall
  - Cross Validation: Holdout và K-fold cross validation
- **Minh họa bằng công cụ Orange**

Demo bằng công cụ Orange

- Cho tập dữ liệu bán hàng (Sales_Data.xlsx)

KhachHang	GioiTinh	Tuoi	ThuNhap	TrinhDo	ChucVu	SoLanMua	GiaTri	LoaiKhachHang
1	Nam	29	15	DH	Nhan vien	3	30	Thuong
2	Nu	22	10	CD	Nhan vien	5	50	VIP
3	Nam	31	22	DH	Nhan vien	1	10	Thuong
4	Nu	23	9	CD	Nhan vien	4	20	VIP
5	Nam	30	21	DH	Nhan vien	3	10	Thuong
6	Nam	40	25	CH	Quan ly	6	50	VIP
7	Nam	50	22	DH	Nhan vien	2	24	VIP
8	Nam	46	24	DH	Quan ly	1	2	Thuong
9	Nam	34	20	CD	Quan ly	1	3	Thuong
10	Nu	46	29	CH	Quan ly	2	4	Thuong
11	Nu	37	30	CH	Quan ly	4	6	Thuong
12	Nu	38	22	DH	Quan ly	5	16	Thuong
13	Nam	39	17	CD	Nhan vien	3	24	Thuong
14	Nam	29	15	DH	Nhan vien	3	30	Thuong
15	Nu	22	10	CD	Nhan vien	5	50	VIP
16	Nam	31	22	DH	Quan ly	1	10	Thuong

Demo bằng công cụ Orange

- Bài toán: cần dự báo phân loại khách hàng (Forecast_Data.xlsx)

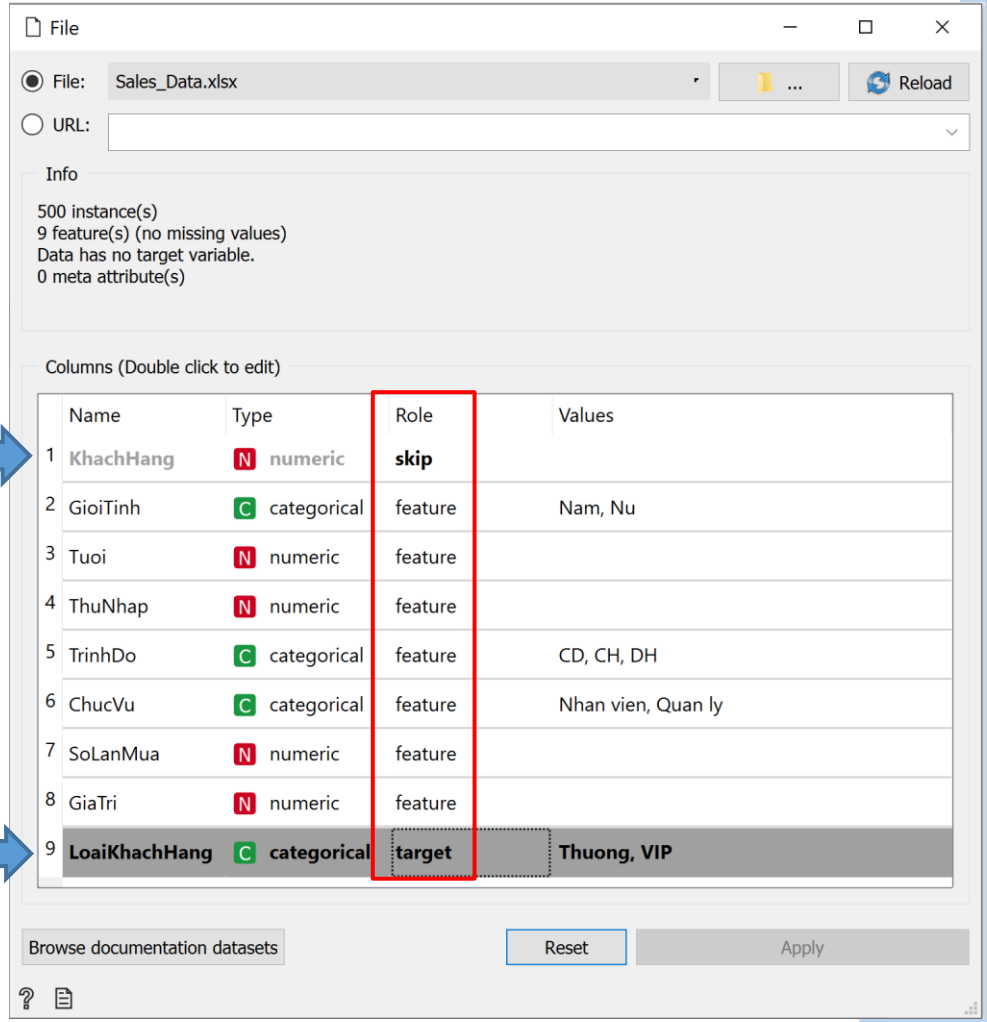
KhachHang	GioiTinh	Tuoi	ThuNhap	TrinhDo	ChucVu	SoLanMua	GiaTri	LoaiKhachHang
1001	Nu	29	15	DH	Nhan vien	2	30	?
1002	Nam	22	10	CD	Nhan vien	5	50	?
1003	Nam	21	22	DH	Nhan vien	1	10	?
1004	Nam	23	9	CD	Nhan vien	4	20	?
1005	Nu	30	21	DH	Nhan vien	3	10	?
1006	Nu	50	25	CH	Quan ly	6	50	?
1007	Nu	41	22	DH	Nhan vien	2	24	?
1008	Nam	28	24	DH	Quan ly	2	7	?
1009	Nam	34	20	CD	Quan ly	1	8	?
1010	Nam	41	29	CH	Quan ly	2	13	?

Demo bằng công cụ Orange

- Chọn nguồn dữ liệu và xác định mô hình dự báo

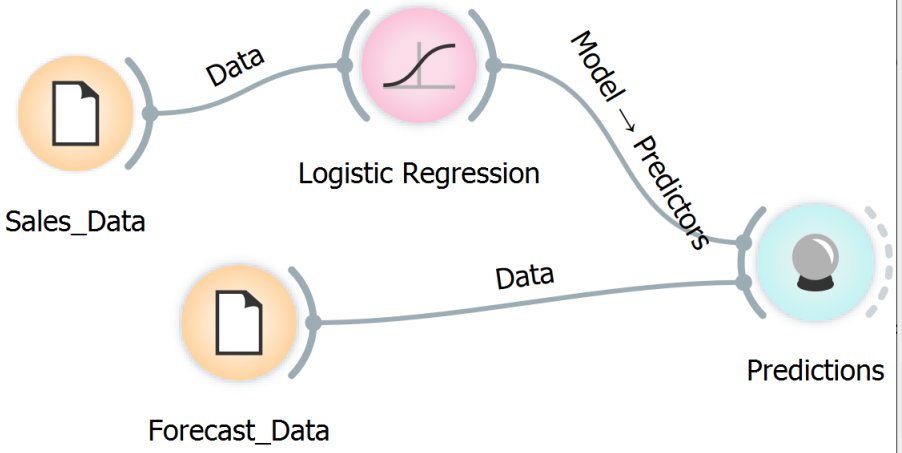
Biến không tham gia vào mô hình

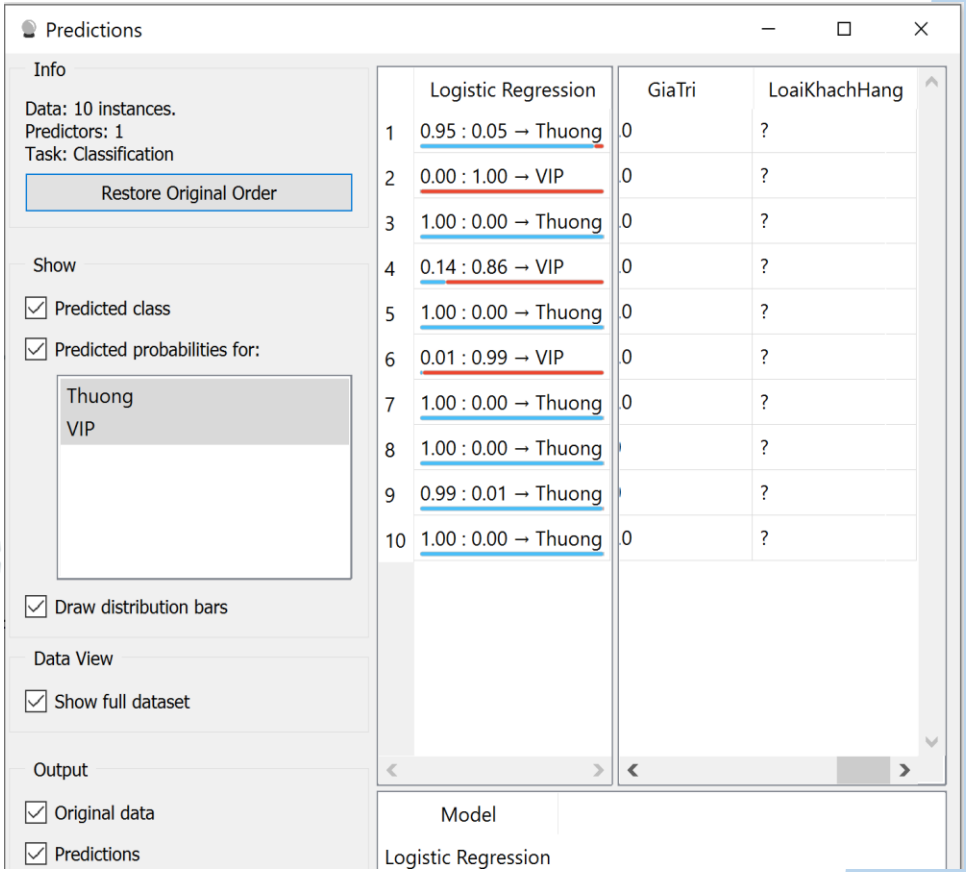
Biến phụ thuộc



Demo bằng công cụ Orange

- Dự báo bằng 1 thuật toán cụ thể (VD: LR)

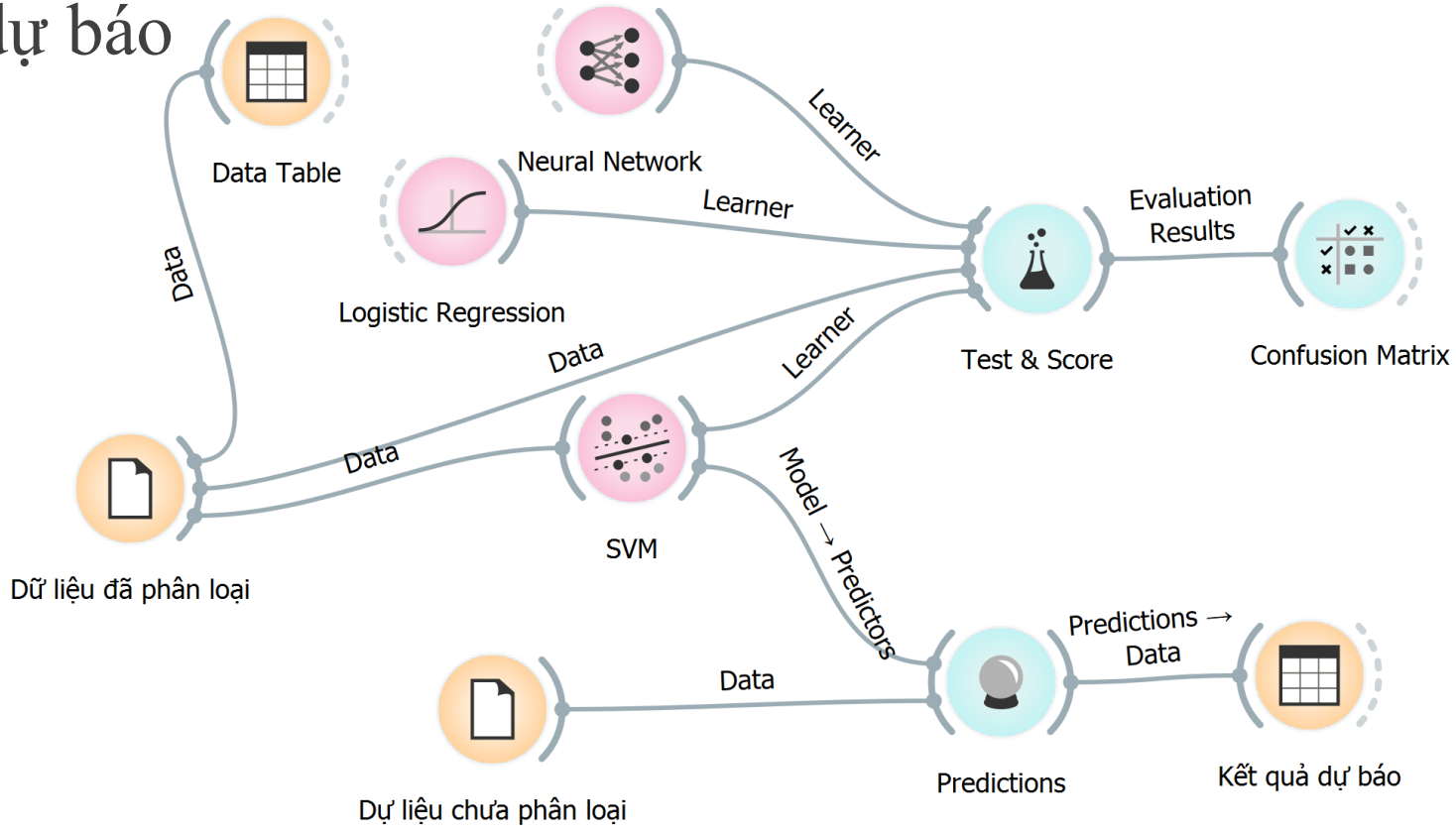




Kết quả

Demo bằng công cụ Orange

- So sánh các thuật toán, rồi dùng thuật toán tốt nhất để dự báo

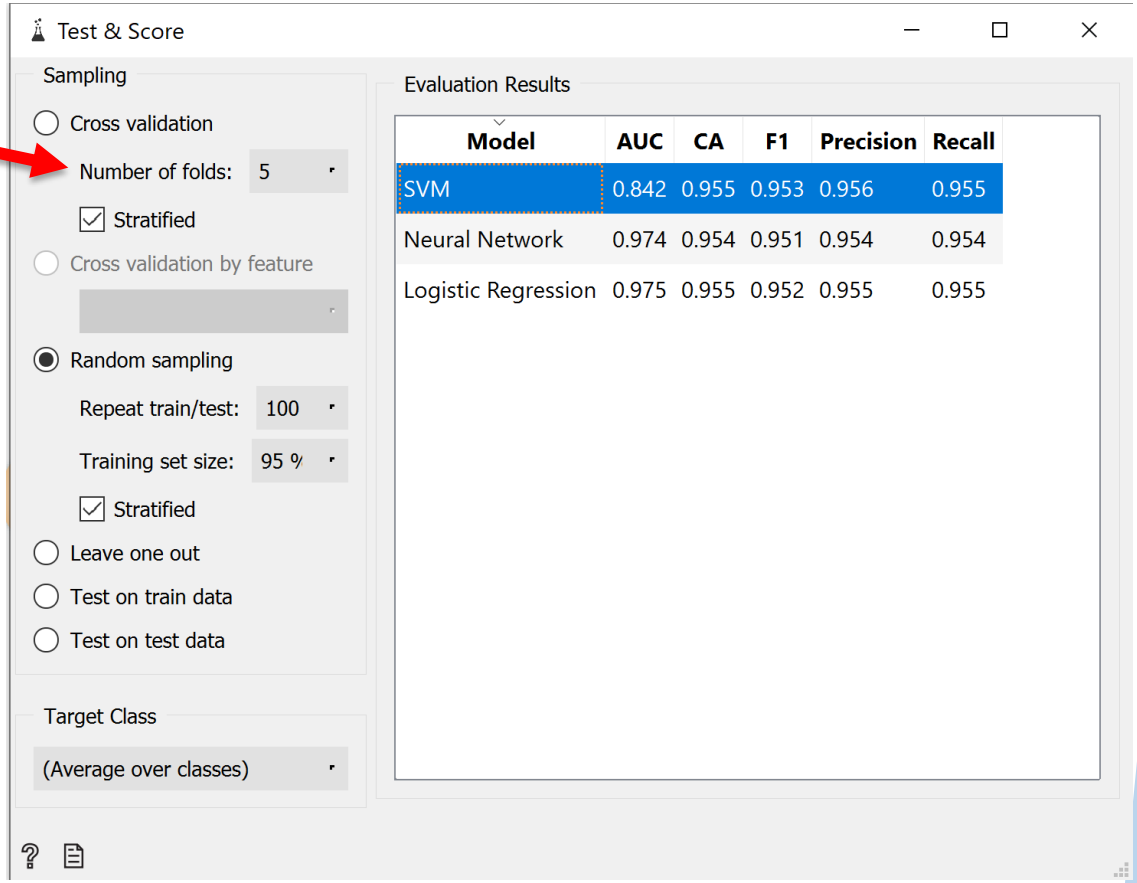


Demo bằng công cụ Orange

- Kết quả huấn luyện với 3 thuật toán

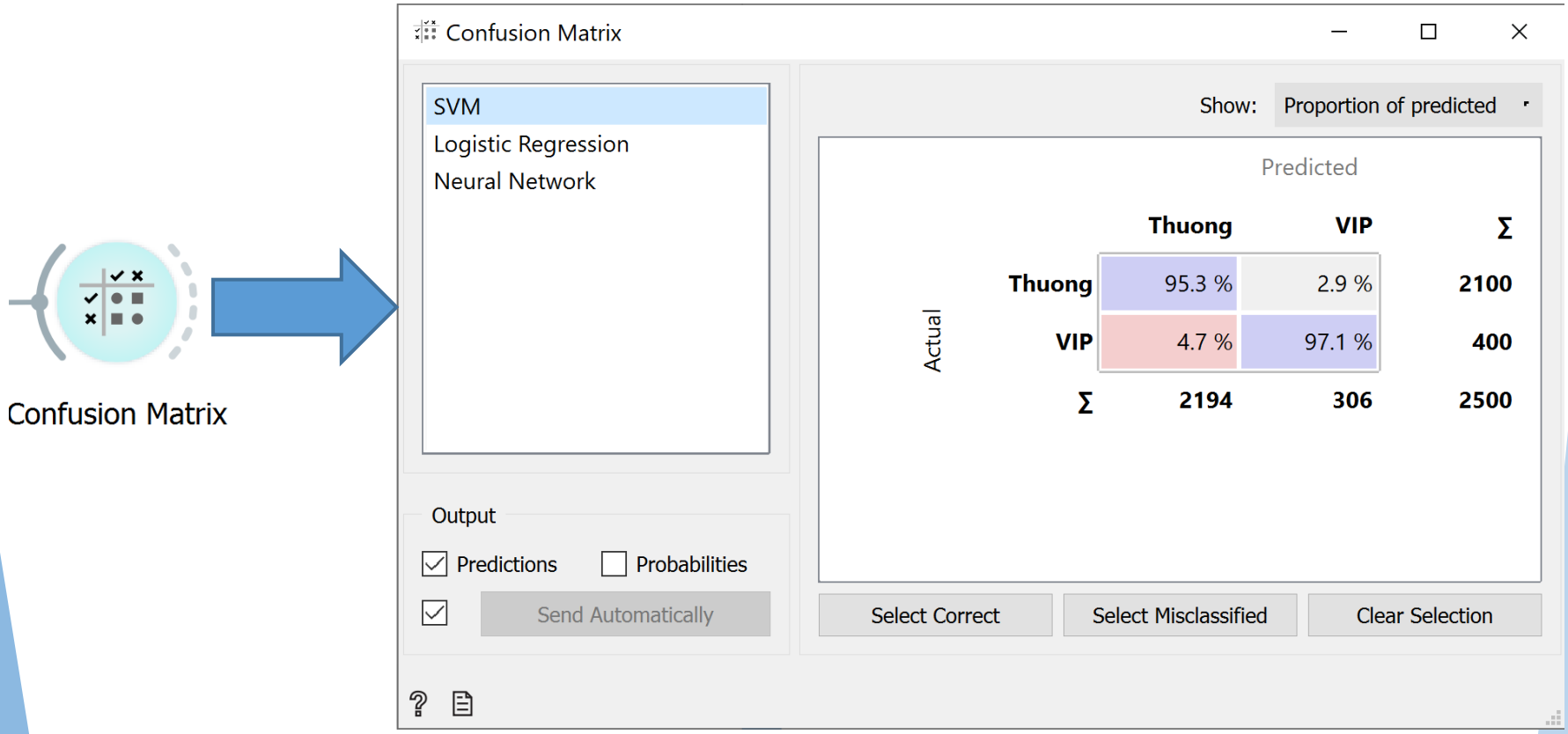
K-fold với k=5

AUC: Area Under the Curve
CA: Accuracy
F1, Precision, Recall



Demo bằng công cụ Orange

- Xem kết quả Ma trận nhầm lẫn



Demo bằng công cụ Orange

- Kết quả dự báo bằng phương pháp SVM

Predictions