TÍNH TOÁN CÁC CHỈ SỐ ĐÁNH GIÁ CHO BÀI TOÁN PHÂN LOẠI NHỊ PHÂN NHẰM SO SÁNH HIỆU SUẤT GIỮA CÁC MÔ HÌNH VÀ LỰA CHỌN MÔ HÌNH PHÙ HỢP NHẤT

I. Mở đầu

Hãy hình dung một phiên chợ xưa, nơi người ta mang thóc đổi gà, gà đổi cá. Mọi giao dịch đều dựa vào cảm tính, khiến việc so sánh giá trị trở nên mơ hồ: một con gà đáng giá bao nhiêu thóc, hay cá quý hơn gà? Chỉ đến khi vàng được dùng làm chuẩn quy đổi chung, việc trao đổi mới trở nên rõ ràng và minh bạch.

Hình 1: Khung cảnh chợ xưa: người dân trao đổi hàng hóa theo kiểu hàng đổi hàng, trước khi vàng trở thành thước đo chung cho giá trị.

Tương tự, trong học máy, nếu thiếu các evaluation metrics (tạm dịch: chỉ số đánh giá), ta cũng không biết mô hình nào thực sự tốt hơn. Những chỉ số này chính là "vàng" giúp so sánh và lựa chọn mô hình một cách khách quan.

Chúng ta có thể chưa quá quen thuộc với các khái niệm, vì vậy bảng chú thích dưới đây sẽ giúp tóm tắt các thuật ngữ chính được dùng xuyên suốt bài.

Bảng chú thích thuật ngữ

Thuật ngữ	Giải thích
Binary Classification(phân loại nhị phân)	Bài toán phân loại dữ liệu đầu vào vào đúng một trong hai lớp khác nhau.
Multi-Class Classification (phân loại đa lớp)	Mỗi mẫu chỉ thuộc về duy nhất một lớp, nhưng tổng số lớp lớn hơn hai.
Multi-Label Classification (phân loại đa nhãn)	Một mẫu có thể đồng thời mang nhiều nhãn.
Confusion Matrix(ma trận nhầm lẫn)	Bảng 2×2 tóm tắt kết quả dự đoán so với nhãn thật, gồm TP, TN, FP, FN.
True Positive (TP)	Mẫu Positive thật được dự đoán đúng.
True Negative (TN)	Mẫu Negative thật được dự đoán đúng.
False Positive (FP)	Mẫu Negative bị nhầm thành Positive (báo động giả).
False Negative (FN)	Mẫu Positive bị nhầm thành Negative (bỏ sót).
Accuracy (độ chính xác)	Tỉ lệ dự đoán đúng trên toàn bộ dữ liệu.
Precision (độ chuẩn xác)	Trong các mẫu dự đoán là Positive, có bao nhiêu mẫu thật sự Positive.
Recall / Sensitivity / TPI (độ phủ)	Trong các mẫu Positive thật, mô hình phát hiện đúng bao nhiêu.
Specificity / TNR(độ đặc hiệu)	Trong các mẫu Negative thật, mô hình dự đoán đúng bao nhiêu.
F1-score	Trung bình điều hòa giữa Precision và Recall, cân bằng cả hai loại lỗi FP và FN.
Balanced Accuracy	Trung bình giữa Recall và Specificity, hữu ích khi dữ liệu mất cân bằng.
MCC (Matthews Correlation Coefficient)	Chỉ số tương quan tổng hợp dùng cả TP, TN, FP, FN.
ROC Curve	Đường cong biểu diễn quan hệ giữa TPR và FPR khi thay đổi ngưỡng phân loại.
AUC (Area Under Curve)	Diện tích dưới đường ROC.

MỤC LỤC

II. Một số evaluation metrics có trong Binary Classification

II.1. Tổng quan về classification

Trước khi chúng ta đi sâu vào cách đo lường hiệu suất, chúng ta sẽ ôn lại khái niệm **Classification** (tạm dịch: phân loại) là gì và các dạng bài toán phổ biến của nó.

Classification là một trong những nhiệm vụ cốt lõi của supervised learning (tạm dịch: học có giám sát). Nhiệm vụ này hướng đến việc xây dựng những mô hình có khả năng ánh xạ dữ liệu đầu vào sang một label (tạm dịch: nhãn) hoặc một category (tạm dịch: loại) rời rạc đã được xác định trước.

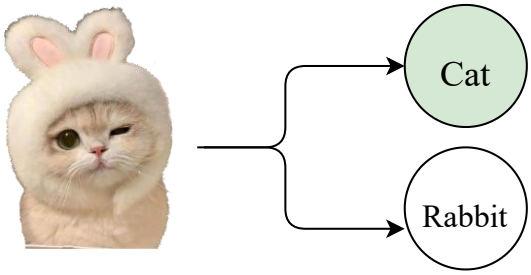
- Supervised Learning
  - Regression
  - Classification
    - Binary Classification
    - Multi-Class Classification
    - Multi-Label Classification
    - ... ...

Hình 2: Một ví dụ về Binary Classification

Điểm cốt lõi của Classification nằm ở khả năng khái quát hóa rằng mô hình không chỉ dừng lại ở việc ghi nhớ dữ liệu huấn luyện, mà còn có thể đưa ra dự đoán chính xác cho những dữ liệu hoàn toàn mới. Chính nhờ năng lực này, các hệ thống Classification đã trở thành nền tảng của nhiều ứng dụng trí tuệ nhân tạo hiện đại, từ nhận dạng hình ảnh và xử lý ngôn ngữ tự nhiên, đến phát hiện gian lận và hỗ trợ ra quyết định trong các lĩnh vực như y tế, tài chính hay an ninh mạng. Trong thực tế, các hệ thống Classification thường được triển khai dưới ba dạng bài toán chính:

Dạng bài toán đầu tiên và cũng cơ bản nhất là **Binary Classification** (tạm dịch: phân loại nhị phân), trong đó nhiệm vụ của mô hình là phân loại mỗi mẫu dữ liệu đầu vào vào một trong hai class (tạm dịch: lớp) không trùng nhau.

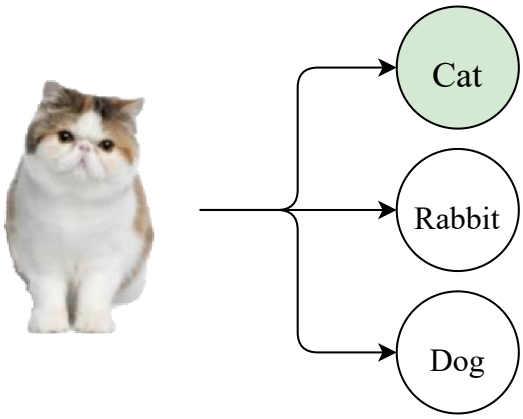
Mỗi dự đoán chỉ có thể rơi vào đúng một trong hai khả năng, chẳng hạn như *có/không*, *đúng/sai*, *spam/không spam*.



Hình 3: Minh họa về Binary Classification.

Multi-Class Classification (tạm dịch: phân loại đa lớp) là dạng bài toán trong đó số lượng lớp lớn hơn hai, nhưng mỗi mẫu dữ liệu đầu vào vẫn chỉ có thể thuộc về duy nhất một lớp. Các lớp này cũng tách biệt và không trùng nhau.

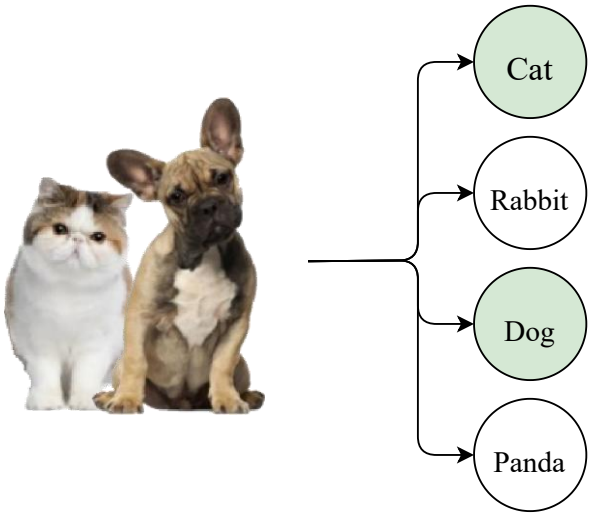
Mô hình phải chọn ra đúng một lớp trong nhiều khả năng có sẵn, chẳng hạn như *chữ số từ 0 đến 9*, *thể loại bài báo như Thể thao, Chính trị, Công nghệ*, hay *loài động vật như chó, mèo, chim*.



Hình 4: Minh họa về Multi-Class Classification.

Multi-Label Classification (tạm dịch: phân loại đa nhãn) là dạng bài toán phức tạp hơn, trong đó mỗi mẫu dữ liệu đầu vào có thể đồng thời thuộc về nhiều lớp. Khác với hai dạng trước, các nhãn trong trường hợp này có thể cùng tồn tại song song.

Một đối tượng có thể được gán nhiều nhãn cùng lúc, chẳng hạn như *một bộ phim vừa thuộc Hành động, Phiêu lưu và Khoa học viễn tưởng*, *một bệnh nhân cùng lúc mắc tiểu đường và cao huyết áp*, hoặc *một bài báo đề cập đến cả Kinh tế lẫn Pháp luật*.



Hình 5: Minh họa về Multi-Label Classification.

Việc phân biệt rõ ba dạng bài toán phân loại giúp chúng ta lựa chọn cách tiếp cận phù hợp và, đồng thời, xác định được phương pháp đánh giá hiệu suất của mô hình một cách hợp lý.

Để tiện theo dõi, bảng dưới đây tổng hợp một số ký hiệu sử dụng xuyên suốt bài viết này:

Bảng Ký hiệu Toán học

Ký hiệu	Ý nghĩa
N	Tổng số mẫu.
$x_i \in \mathbb{R}^n$	Vector đặc trưng của mẫu thứ i.
$y_i \in \{0, 1\}$	Nhãn thực (1: Positive, 0: Negative).
$\hat{p}_i \in [0, 1]$	Xác suất dự đoán $P(y_i = 1 \mid x_i)$.
$\tau \in [0, 1]$	Ngưỡng phân loại: $\hat{y}_i = \mathbf{1}[\hat{p}_i \geq \tau]$.
$\hat{y}_i \in \{0, 1\}$	Nhãn dự đoán (sau khi áp ngưỡng).
TP	Mẫu Positive thật được dự đoán đúng.
TN	Mẫu Negative thật được dự đoán đúng.
FP	Mẫu Negative bị nhầm thành Positive (báo động giả).
FN	Mẫu Positive bị nhầm thành Negative (bỏ sót).
Accuracy	$\dfrac{TP + TN}{TP + TN + FP + FN}$.
Precision	$\dfrac{TP}{TP + FP}$.
Recall	$\dfrac{TP}{TP + FN}$.
Specificity	$\dfrac{TN}{TN + FP}$.
$F1 - score$	$2 \cdot \dfrac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$.
Balanced Accuracy	Balanced Accuracy $= \dfrac{\text{Recall} + \text{Specificity}}{2}$.
MCC	$\dfrac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$.
AUC_{ROC}	Diện tích dưới đường ROC.

II.2. Tổng quát về các evaluation metrics

II.2.1. Phát biểu bài toán

Trong bài toán Binary Classification, ta có tập dữ liệu huấn luyện $D = \{(x_i, y_i)\}_{i=1}^{N}$, trong đó $x_i \in \mathrm{R}^n$ là vector đặc trưng và $y_i \in \{0,1\}$ là nhãn (với 1 biểu thị lớp *Positive*, 0 biểu thị lớp *Negative*).

Khi dự đoán, mô hình có thể trả về:

(1) Nhãn rời rạc $\hat{y}_i \in \{0,1\}$

(2) Xác suất $\hat{p}_i = P(y_i = 1 \mid x_i) \in [0,1]$.

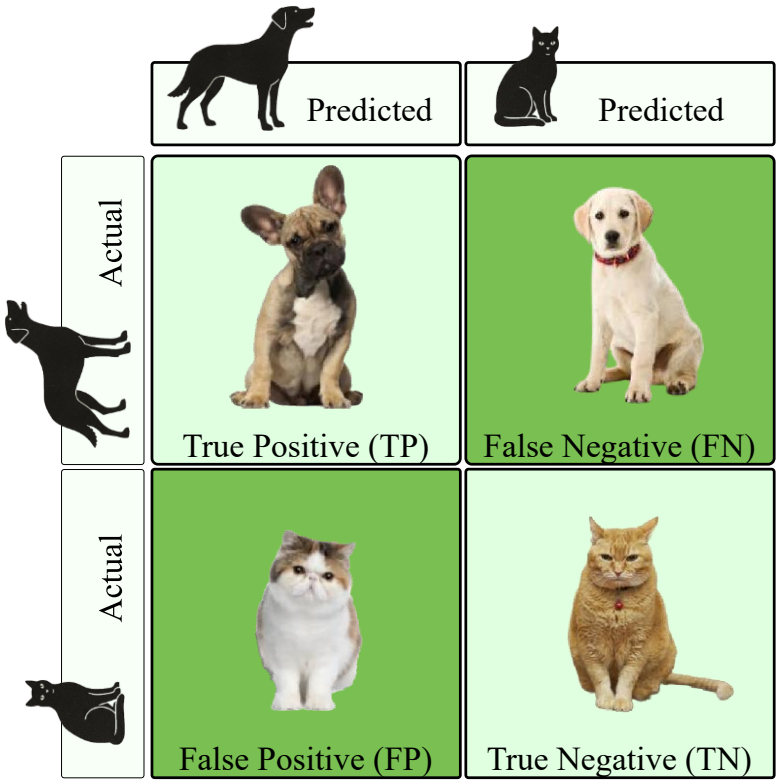
Trường hợp mô hình xuất ra xác suất $\hat{p}_i$, ta cần chuyển đổi thành nhãn rời rạc bằng cách chọn một **ngưỡng** (threshold) $\tau \in [0,1]$:

$$\hat{y}_i = \begin{cases} 1 & \text{nếu } \hat{p}_i \geq \tau, \\ 0 & \text{nếu } \hat{p}_i < \tau. \end{cases}$$

Thông thường, ngưỡng mặc định là $\tau = 0.5$.

Để đánh giá hiệu quả dự đoán của mô hình, công cụ nền tảng được sử dụng là **Confusion Matrix** (tạm dịch: ma trận nhầm lẫn). Từ cấu trúc này, nhiều evaluation metrics được phát triển nhằm khắc phục hạn chế của những chỉ số đơn lẻ và phản ánh đầy đủ hơn hiệu năng mô hình.

Trong các phần tiếp theo, ta sẽ minh hoạ cách áp dụng các chỉ số này thông qua ví dụ phân loại hình ảnh *chó* và *mèo*.



Hình 6: Confusion Matrix cho bài toán phân loại hình ảnh chó mèo.

Xét cụ thể trong bài toán phân loại chó và mèo, **Confusion Matrix** được minh hoạ như Hình 6 và bao gồm bốn thành phần chính:

- **True Positive (TP)**: số mẫu chó thật sự được mô hình dự đoán đúng là chó;
- **True Negative (TN)**: số mẫu mèo thật sự được dự đoán đúng là mèo;
- **False Positive (FP)**: số mẫu mèo nhưng bị nhầm thành chó;
- **False Negative (FN)**: số mẫu chó nhưng bị nhầm thành mèo.

Dựa trên bốn thành phần của Confusion Matrix, ta có thể xây dựng nhiều chỉ số đánh giá khác nhau nhằm phản ánh các khía cạnh riêng biệt về hiệu quả của mô hình, trong đó phổ biến nhất là *Accuracy*, *Precision*, *Recall* và *F1-score*. Khi mở rộng sang bài toán Multi-Class Classification, Confusion Matrix không còn giới hạn ở dạng 2×2 mà trở thành một ma trận vuông, với mỗi hàng và mỗi cột tương ứng với một lớp cụ thể, từ đó cho phép phân tích chi tiết hiệu năng của mô hình trên từng lớp.

II.2.2. Accuracy

Từ Confusion Matrix trên, ta có thể định nghĩa chỉ số đơn giản và phổ biến nhất là **Accuracy** (tạm dịch: độ chính xác). Chỉ số này đo tỉ lệ tổng số mẫu được mô hình dự đoán đúng (cả chó và mèo) trên toàn bộ dữ liệu, được tính theo công thức:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Trong đó:

- *Accuracy* là độ chính xác;
- *TP* là số mẫu chó thật sự được mô hình dự đoán đúng là chó;
- *TN* là số mẫu mèo thật sự được dự đoán đúng là mèo;
- *FP* số mẫu mèo nhưng bị nhầm thành chó;
- *FN* số mẫu chó nhưng bị nhầm thành mèo.

Nói cách khác, Accuracy cho biết trong toàn bộ các bức ảnh chó mèo cần phân loại, mô hình đã dự đoán đúng bao nhiêu phần trăm.

Lưu ý: Mặc dù dễ hiểu và trực quan, Accuracy lại có hạn chế lớn trong trường hợp dữ liệu mất cân bằng.

Chính vì vậy, cần đến các chỉ số khác để bổ sung cho Accuracy, giúp đánh giá mô hình toàn diện hơn trong nhiều bối cảnh.

II.2.3. Precision

Sau khi đã xem xét Accuracy trong bài toán phân loại chó mèo, ta tiếp tục đến với một chỉ số khác tập trung vào chất lượng của các dự đoán Positive, đó là **Precision** (tạm dịch: độ chuẩn xác).

Trong ngữ cảnh này, Precision cho biết trong số tất cả các bức ảnh mà mô hình dự đoán là *chó*, có bao nhiêu bức thực sự đúng là chó. Chỉ số này được tính theo công thức:

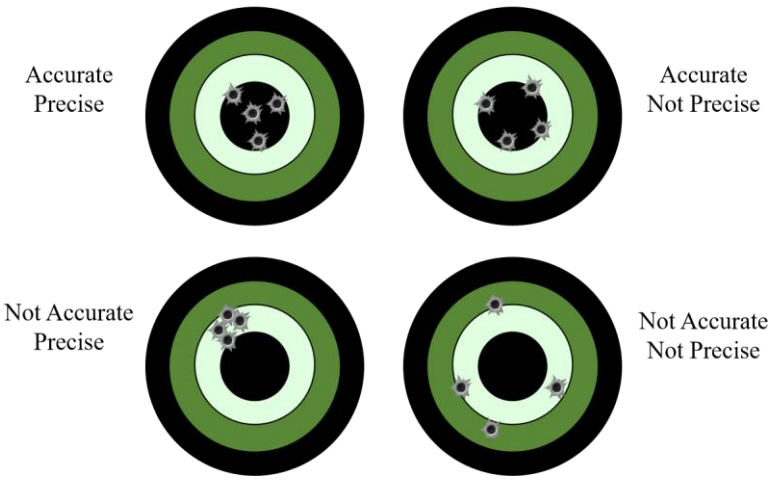
$$Precision = \frac{TP}{TP + FP}$$

Trong đó:

- *Precision* là độ chuẩn xác;
- *TP* là số mẫu chó thật sự được mô hình dự đoán đúng là chó;
- *FP* số mẫu mèo nhưng bị nhầm thành chó.

Điều này có thể hiểu như sau:

- Precision cao → mô hình hiếm khi nhầm mèo thành chó, hầu hết các dự đoán chó đều chính xác.
- Precision thấp → mô hình thường xuyên gán nhãn nhầm mèo thành chó, gây nhiều báo động giả.



Hình 7: Mối quan hệ giữa Accuracy và Precision.

Lưu ý: Accuracy và Precision đều phản ánh mức độ chính xác của dự đoán, nhưng ở hai khía cạnh khác nhau, được minh họa ở Hình 7. **Accuracy** đo lường độ gần đúng trung bình so với giá trị thực (các dự đoán gần "tâm"), trong khi **Precision** nhấn mạnh vào độ tập trung và ổn định của các dự đoán (các dự đoán nằm sát nhau, dù có thể lệch khỏi tâm).

.

II.2.4. Recall

Bên cạnh Precision, một chỉ số quan trọng khác trong bài toán phân loại chó mèo là **Recall** (tạm dịch: độ phủ) (còn được gọi là **Sensitivity** hoặc **True Positive Rate**). Nếu Precision đánh giá độ "chắc chắn" của những dự đoán chó, thì Recall lại đo lường khả năng *không bỏ sót* các bức ảnh chó thật sự.

Nói cách khác, Recall cho biết trong toàn bộ các ảnh chó, mô hình đã nhận diện đúng được bao nhiêu. Chỉ số này được tính theo công thức:

$$Recall = \frac{TP}{TP + FN}$$

Trong đó:

- *Recall* là độ phủ;
- *TP* là số mẫu chó thật sự được mô hình dự đoán đúng là chó;
- *FN* số mẫu chó nhưng bị nhầm thành mèo.

Điều này có thể hiểu như:

- Recall cao → mô hình phát hiện hầu hết các ảnh chó, rất ít bị bỏ sót.
- Recall thấp → mô hình bỏ qua nhiều ảnh chó, làm giảm độ tin cậy trong những ứng dụng quan trọng.

II.2.5. Trade-off giữa Precision và Recall

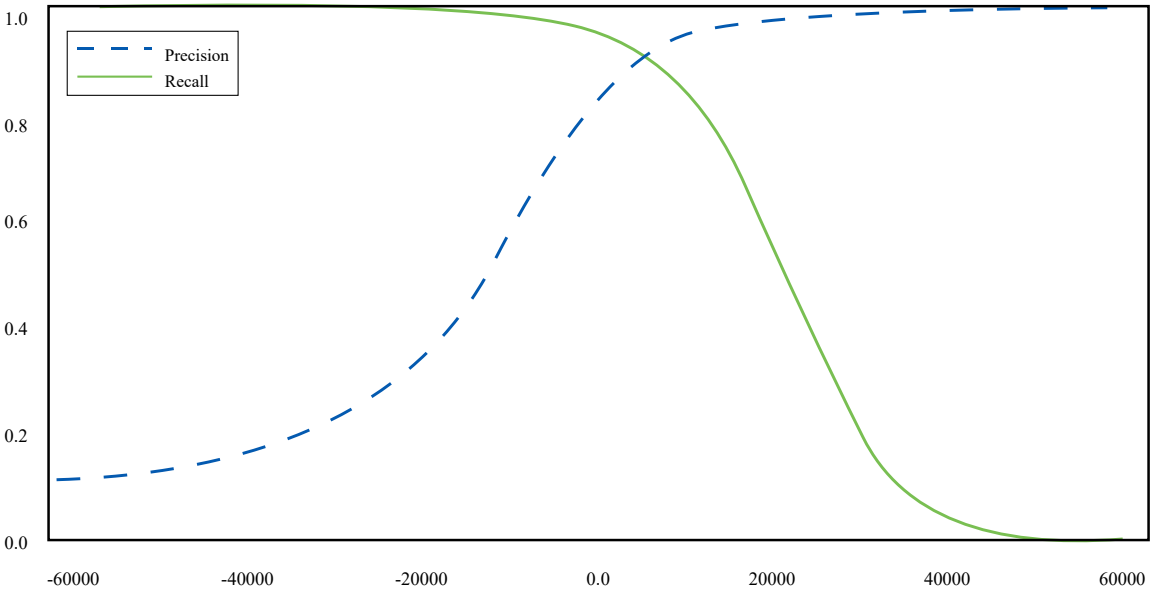
Từ hai chỉ số Precision và Recall trong bài toán phân loại chó mèo, ta có thể thấy rằng chúng phản ánh hai góc nhìn khác nhau: Precision nhấn mạnh vào độ chắc chắn của những dự đoán "chó", còn Recall lại nhấn mạnh vào việc không bỏ sót các ảnh chó thật. Vấn đề là hai chỉ số này hiếm khi đạt giá trị cao cùng lúc, mà thường tồn tại mối quan hệ trade-off (tạm dịch: đánh đổi).

Nguyên nhân xuất phát từ ngưỡng phân loại mà ta lựa chọn. Ví dụ:

- Nếu đặt ngưỡng cao (ví dụ 0.9), mô hình chỉ gán nhãn "chó" khi rất chắc chắn. Khi đó Precision cao (hầu hết các ảnh được dự đoán là chó đều thật sự là chó), nhưng Recall thấp vì nhiều ảnh chó bị bỏ sót.

- Ngược lại, nếu đặt ngưỡng thấp (ví dụ 0.3), mô hình trở nên "dễ" hơn khi gán nhãn "chó". Khi đó Recall cao (hầu hết các ảnh chó đều được phát hiện), nhưng Precision thấp vì nhiều ảnh mèo bị nhầm thành chó.

- Ngưỡng



Hình 8: Mối quan hệ trade-off giữa Precision và Recall theo ngưỡng.

Như vậy, Precision và Recall giống như hai mặt của một chiếc cân: khi mặt này nâng lên thì mặt kia hạ xuống. Do đó, trong thực tế không thể chỉ tối ưu một phía mà cần có sự cân bằng, tùy thuộc vào mục tiêu của ứng dụng. Với những tình huống đòi hỏi cả hai yếu tố, ta thường dùng các chỉ số tổng hợp như **F1-score** để đánh giá toàn diện hơn.

II.2.6. Specificity

Song song với Recall, một chỉ số khác cũng quan trọng trong bài toán phân loại chó mèo là **Specificity** (tạm dịch: độ đặc hiệu) (còn được gọi là **True Negative Rate**). Nếu Recall đo lường khả năng mô hình nhận diện hết

các ảnh chó thật sự, thì Specificity phản ánh mức độ chính xác trong việc phát hiện mèo, tức là không gán nhầm mèo thành chó.

Có thể hiểu rằng, Specificity cho biết trong toàn bộ các ảnh mèo, mô hình đã dự đoán đúng được bao nhiêu. Chỉ số này được tính theo công thức:

$$Specificity = \frac{TN}{TN + FP}$$

Trong đó:

- *Specificity* là độ đặc hiệu;
- *TN* là số mẫu mèo thật sự được dự đoán đúng là mèo;
- *FP* số mẫu mèo nhưng bị nhầm thành chó.

Ý nghĩa trong thực tế:

• Specificity cao → mô hình hầu như không nhầm mèo thành chó, đảm bảo độ tin cậy cho lớp Negative.

• Specificity thấp → mô hình thường xuyên gán nhầm mèo thành chó, tạo ra nhiều báo động giả.

II.2.7. F1-score

Sau khi đã xem xét Precision và Recall trong bài toán phân loại chó mèo, ta dễ dàng nhận thấy rằng hai chỉ số này thường mâu thuẫn với nhau.

Để dung hòa hai yếu tố này, ta sử dụng **F1-score** (tạm dịch: điểm F1), một chỉ số được định nghĩa dựa trên trung bình điều hòa của Precision và Recall:

$$F1 - score = 2 . \frac{Precision . Recall}{Precision + Recall}$$

Trong đó:

- $F1 - score$ là chỉ số cân bằng giữa Precision và Recall;

- *Precision* là tỉ lệ số ảnh được dự đoán là chó và thực sự đúng là chó;
- *Recall* là tỉ lệ số ảnh chó thật sự được mô hình nhận diện đúng.

F1-score đặc biệt hữu ích khi dữ liệu bị mất cân bằng, chẳng hạn số lượng ảnh mèo nhiều hơn chó. Khi đó, F1-score giúp cân nhắc đồng thời cả hai loại lỗi: bỏ sót chó thật (FN) và báo động giả khi nhầm mèo thành chó (FP).

Nói cách khác, ta vừa muốn mô hình nhận diện được hầu hết các bức ảnh chó (Recall cao), vừa muốn đảm bảo rằng những ảnh được dự đoán là chó thực sự đúng là chó (Precision cao). F1-score chính là thước đo phù hợp để cân bằng giữa hai yêu cầu này trong bài toán phân loại chó-mèo.

Lưu ý: Đặc điểm của F1-score là chỉ cao khi cả Precision và Recall đều cao; nếu một trong hai quá thấp thì F1-score cũng giảm theo.

II.2.8. Balanced Accuracy

Trong trường hợp dữ liệu phân loại chó mèo bị mất cân bằng, ví dụ số lượng ảnh mèo nhiều hơn hẳn số lượng ảnh chó, chỉ số **Accuracy** thông thường có thể trở nên thiếu công bằng. Khi đó, **Balanced Accuracy** (tạm dịch: độ chính xác cân bằng) được sử dụng để điều chỉnh bằng cách lấy trung bình giữa Recall và Specificity. Chỉ số này được tính theo công thức:

$$Balaced\ Accuracy = \frac{Recall + Specificity}{2}$$

Trong đó:

- *Balanced Accuracy* là độ chính xác cân bằng;
- *Recall* đo tỉ lệ số ảnh chó được mô hình nhận diện đúng;
- *Specificity* đo tỉ lệ số ảnh mèo được mô hình nhận diện đúng.

Nhờ cách tính trung bình này, Balanced Accuracy phản ánh công bằng hiệu quả của mô hình trên cả hai lớp chó và mèo, thay vì chỉ bị chi phối bởi lớp chiếm đa số.

II.2.9. MCC (Matthews Correlation Coefficient)

Một chỉ số tổng hợp khác mạnh mẽ hơn là **MCC** (Matthews Correlation Coefficient), còn được gọi là hệ số tương quan Matthews. MCC không chỉ xem xét TP và TN, mà còn tính đến cả FP và FN, nhờ đó cung cấp cái nhìn cân bằng hơn về chất lượng mô hình. Chỉ số này được tính theo công thức:

$$MCC = \frac{TP.TN - FP.FN}{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}$$

Trong đó:

- *MCC* là hệ số tương quan Matthews;
- *TP* là số mẫu chó thật sự được mô hình dự đoán đúng là chó;
- *TN* là số mẫu mèo thật sự được dự đoán đúng là mèo;
- *FP* số mẫu mèo nhưng bị nhầm thành chó;
- *FN* số mẫu chó nhưng bị nhầm thành mèo.

Ý nghĩa của MCC:

- Giá trị dao động từ −1 đến 1.
- *MCC* = 1: mô hình phân loại hoàn hảo (nhận diện đúng tất cả chó và mèo).
- *MCC* = 0: mô hình không tốt hơn đoán ngẫu nhiên.
- *MCC* = −1: mô hình dự đoán hoàn toàn ngược lại so với thực tế.

Chỉ số MCC đặc biệt hữu ích khi dữ liệu mất cân bằng, vì nó đánh giá hiệu quả tổng thể dựa trên cả bốn giá trị của Confusion Matrix, tránh thiên lệch về lớp chiếm ưu thế.

. **Lưu ý:** Tất cả các chỉ số như Accuracy, Precision, Recall, F1-score hay thậm chí Balanced Accuracy và MCC đều được tính toán dựa trên một ngưỡng cố định để phân loại.

tuy nhiên, trong thực tế, mô hình không chỉ xuất ra nhãn rời rạc mà còn đưa ra xác suất $\hat{p}_i$. Nếu ta thay đổi ngưỡng phân loại, ví dụ từ 0.5 sang 0.7

trong bài toán chó-mèo, số ảnh được gán nhãn là chó sẽ thay đổi, kéo theo toàn bộ các chỉ số trên cũng thay đổi theo.

II.2.10. ROC Curve và AUC

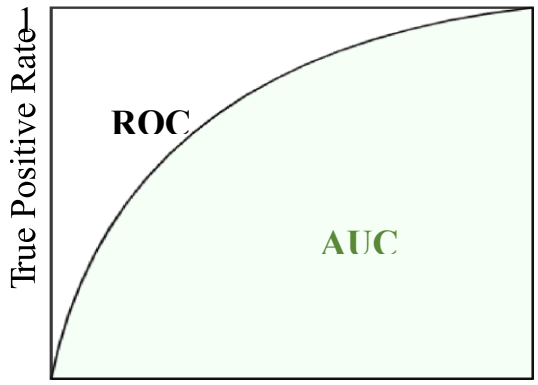
Để có thể so sánh mô hình một cách khách quan mà không phụ thuộc vào việc chọn ngưỡng, ta sử dụng **ROC Curve** và **AUC**. Đây là những công cụ đánh giá trực quan và tổng quát hơn, dựa trên việc quét toàn bộ các giá trị ngưỡng thay vì cố định tại một điểm duy nhất.

Trước hết, ta định nghĩa hai chỉ số quan trọng:

$$TPR = \frac{TP}{TP + FN}, \qquad FPR = \frac{FP}{FP + TN}.$$

Trong đó:

• *TPR* (True Positive Rate) hay Recall: tỉ lệ ảnh chó thật được nhận diện đúng là chó; • *FPR* (False Positive Rate): tỉ lệ ảnh mèo bị nhầm thành chó.



0 False Positive Rate 1

Hình 9: Đường cong ROC và phần diện tích AUC.

ROC Curve là đồ thị biểu diễn mối quan hệ giữa *TPR* và *FPR* khi thay đổi ngưỡng phân loại. Một mô hình tốt sẽ có đường ROC nằm gần góc trên bên trái, thể hiện vừa nhận diện đúng nhiều chó (TPR cao), vừa hạn chế nhầm mèo thành chó (FPR thấp).

AUC (Area Under Curve) là diện tích dưới đường cong ROC:

- AUC = 1.0 → mô hình phân loại hoàn hảo (phân biệt chó và mèo chính xác tuyệt đối).
- AUC = 0.5 → mô hình không tốt hơn đoán ngẫu nhiên.

III.Thực hành

III.1. Pháp biểu bài toán

Giả sử Doanh nghiệp AI đang thử nghiệm ba mô hình học máy để phân loại xe đã qua sử dụng thành hai lớp: **Lớp 0 (Chất lượng tốt)** và **Lớp 1 (Rủi ro cao)**. Mục tiêu là so sánh hiệu suất giữa các mô hình theo đầy đủ thước đo đã trình bày (Accuracy, Precision, Recall, Specificity, F1, Balanced Accuracy, MCC), từ đó chọn phương án phù hợp với mục tiêu kinh doanh.

Bảng 1: Bộ dữ liệu kiểm định xe và dự đoán của các mô hình.

ID Xe	Năm SX	Km đã đi	Nhãn thực tế	Model A	Model B	Model C
1	2022	12000	Tốt	Tốt	Tốt	Tốt
2	2018	140000	Rủi ro	Rủi ro	Rủi ro	Rủi ro
3	2021	35000	Tốt	Tốt	Tốt	Rủi ro
4	2020	55000	Tốt	Rủi ro	Tốt	Tốt
5	2017	190000	Rủi ro	Tốt	Tốt	Rủi ro
6	2022	20000	Tốt	Tốt	Tốt	Rủi ro
7	2019	80000	Tốt	Tốt	Tốt	Tốt

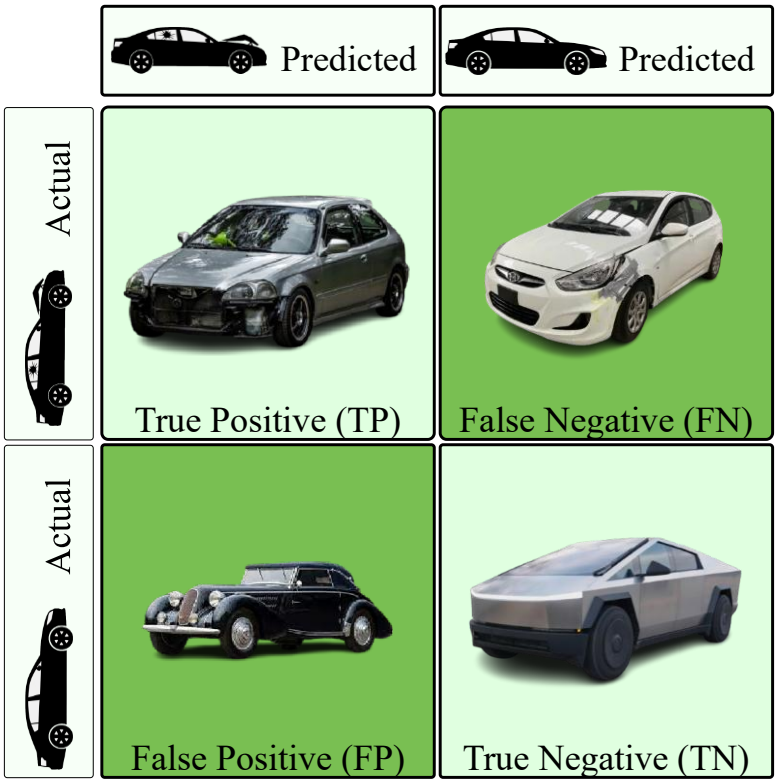
III.1.1. Phân tích sơ bộ bài toán Rủi ro kinh doanh:

- Sai lầm nghiêm trọng nhất là *bỏ sót* (False Negative) - xe thực sự **rủi ro** nhưng bị mô hình dự đoán thành **tốt**.
- Hậu quả: chi phí bảo hành hoặc sửa chữa cao hơn, đồng thời làm giảm uy tín và niềm tin của khách hàng.
- Trong thực tế, doanh nghiệp thường ưu tiên mô hình có **Recall** cao để phát hiện được càng nhiều xe rủi ro càng tốt.

- Việc chấp nhận một số báo động nhầm (False Positive) thường được xem là ít nguy hại hơn so với bỏ sót xe rủi ro.

III.2. Lời giải

Bước 1: Định nghĩa Confusion Matrix



Hình 10: Confusion Matrix cho bài toán phân loại xe.

Trong đó:

- **TP (True Positive):** số xe *rủi ro* được mô hình dự đoán đúng là rủi ro;
- **TN (True Negative):** số xe *tốt* được mô hình dự đoán đúng là tốt;
- **FP (False Positive):** số xe *tốt* nhưng bị mô hình dự đoán nhầm là rủi ro (báo động giả);
- **FN (False Negative):** số xe *rủi ro* nhưng bị mô hình dự đoán nhầm là tốt (bỏ sót).

Đối chiếu dự đoán với nhãn thực từ Bảng 1, ta thu được:

Mô hình	TP	FN	FP	TN
Model A	1	1	1	4

Model B	1	1	0	5
Model C	2	0	2	3

Bước 2: Tính đầy đủ các chỉ số

Với mỗi mô hình, ta lần lượt tính toàn bộ các chỉ số đã giới thiệu ở phần trước:

(1) Accuracy

Accuracy đo tỷ lệ dự đoán đúng trên toàn bộ dữ liệu. Công thức:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}.$$

Áp dụng cho từng mô hình, ta được:

$$Accuracy_A = \frac{TP_A + TN_A}{TP_A + TN_A + FP_A + FN_A} = \frac{1+4}{1+4+1+1} = \frac{5}{7} \approx 71.4\%.$$

$$Accuracy_B = \frac{TP_B + TN_B}{TP_B + TN_B + FP_B + FN_B} = \frac{1+5}{1+5+0+1} = \frac{6}{7} \approx 85.7\%.$$

$$Accuracy_C = \frac{TP_C + TN_C}{TP_C + TN_C + FP_C + FN_C} = \frac{2+3}{2+3+2+0} = \frac{5}{7} \approx 71.4\%.$$

(2) Precision

Precision đo tỷ lệ số xe thực sự rủi ro trong toàn bộ số xe mà mô hình dự đoán là rủi ro. Công thức:

$$Precision = \frac{TP}{TP + FP}.$$

Với từng mô hình, ta có:

$$Precision_A = \frac{1}{1+1} = \frac{1}{2} = 50\%.$$

$$Precision_B = \frac{1}{1+0} = 1.0 = 100\%.$$

$$Precision_C = \frac{2}{2+2} = \frac{2}{4} = 50\%.$$

(3) Recall

Recall đo tỷ lệ số xe rủi ro được phát hiện đúng trong toàn bộ số xe rủi ro thực tế. Công thức:

$$Recall = \frac{TP}{TP + FN}.$$

Với từng mô hình, ta có:

$$Recall_A = \frac{1}{1+1} = \frac{1}{2}$$

$$= 50\%.\ Recall_B =$$

$$\frac{1}{1+1} = \frac{1}{2} = 50\%.$$

$$Recall_C = \frac{2}{2+0} = 1.0 = 100\%.$$

(4) Specificity

Specificity đo tỷ lệ số xe tốt được nhận diện đúng trong toàn bộ số xe tốt thực tế. Công thức:

$$Specificity = \frac{TN}{TN + FP}.$$

Với từng mô hình, ta có:

$$Specificity_A = \frac{4}{4+1} = \frac{4}{5} = 80\%.$$

$$Specificity_B = \frac{5}{5+0} = 1.0 = 100\%.$$

$$Specificity_C = \frac{3}{3+2} = \frac{3}{5} = 60\%.$$

(5) F1-score

F1-score là trung bình điều hoà giữa Precision và Recall, giúp cân bằng giữa độ chính xác và khả năng phát hiện. Công thức:

$$F1 - score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}.$$

Với từng mô hình, ta có:

$$F1 - score_A = \frac{2 \cdot 0.5 \cdot 0.5}{0.5 + 0.5} = \frac{0.5}{1} = 50\%.$$

$$F1 - score_B = \frac{2 \cdot 1.0 \cdot 0.5}{1.0 + 0.5} = \frac{1.0}{1.5} \approx 66.7\%.$$

$$F1 - score_C = \frac{2 \cdot 0.5 \cdot 1.0}{0.5 + 1.0} = \frac{1.0}{1.5} \approx 66.7\%.$$

(6) Balanced Accuracy

Balanced Accuracy được định nghĩa là trung bình cộng giữa Recall và Specificity, giúp đánh giá công bằng trong trường hợp dữ liệu mất cân bằng. Công thức:

$$Balanced\ Accuracy = \frac{Recall + Specificity}{2}.$$

Với từng mô hình, ta có:

$$BalancedAcc_A = \frac{50\% + 80\%}{2} = 65\%.$$

$$BalancedAcc_B = \frac{50\% + 100\%}{2} = 75\%.$$

$$BalancedAcc_C = \frac{100\% + 60\%}{2} = 80\%.$$

(7) MCC (Matthews Correlation Coefficient)

MCC là hệ số tương quan Matthews, xem xét đồng thời TP, TN, FP, FN để đưa ra thước đo cân bằng toàn diện. Công thức:

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}.$$

Với từng mô hình, ta có:

$$MCC_A = \frac{1 \cdot 4 - 1 \cdot 1}{\sqrt{(1+1)(1+1)(4+1)(4+1)}} = \frac{3}{\sqrt{100}} = 0.300.$$

$$MCC_B = \frac{1 \cdot 5 - 0 \cdot 1}{\sqrt{(1+0)(1+1)(5+0)(5+1)}} = \frac{5}{\sqrt{60}} \approx 0.645$$

$$MCC_C = \frac{2 \cdot 3 - 2 \cdot 0}{\sqrt{(2+2)(2+0)(3+2)(3+0)}} = \frac{6}{\sqrt{120}} \approx 0.548$$

Sau khi tính toán chi tiết cho từng chỉ số trên ba mô hình, ta tổng hợp kết quả vào bảng dưới đây.

Bảng 2: Bảng chỉ số đầy đủ cho từng mô hình (đơn vị: %). MCC hiển thị

Mô hình	Accuracy	Precision	Recall	Specificity	F1	Balanced Acc.	MCC
Model A	71.4	50.0	50.0	80.0	50.0	65.0	0.300
Model B	**85.7**	**100.0**	50.0	**100.0**	66.7	75.0	**0.645**
Model C	71.4	50.0	**100.0**	60.0	**66.7**	**80.0**	0.548

Lưu ý: Trong dữ liệu này ta chỉ có nhãn rời rạc (không có xác suất dự đoán nên không thể tính AUC-ROC

Bước 3: Phân tích kết quả & Kết luận

Quan sát Bảng trên, có thể thấy sự khác biệt rõ rệt giữa ba mô hình:

- **Model B:** Là mô hình có nhiều chỉ số vượt trội nhất. Với *Accuracy* cao nhất (85.7%), *Precision* tuyệt đối (100%), *Specificity* tối đa (100%) và hệ số *MCC* lớn nhất (0.645), mô hình B gần như không tạo ra báo động giả. Khi mô hình này gắn cờ một xe là *rủi ro*, thì dự đoán đó hầu như chắc chắn chính xác. Đây là lựa chọn phù hợp nếu mục tiêu kinh doanh là **tối ưu độ tin cậy cảnh báo và hiệu quả tổng thể**.
- **Model C:** Nổi bật ở *Recall* (100%) và *Balanced Accuracy* cao nhất (80%). Điều này có nghĩa là mô hình C không bỏ sót bất kỳ xe rủi ro nào, song lại phải đánh đổi bằng số lượng báo động giả (FP) cao hơn. Mô hình này phù hợp trong bối cảnh doanh nghiệp đặt ưu tiên vào **an toàn và uy tín**, chấp nhận kiểm tra bổ sung để xử lý các trường hợp gắn cờ nhầm.

- **Model A:** Các chỉ số đều ở mức trung bình, với *Precision* và *Recall* chỉ đạt 50%, đồng thời MCC thấp nhất (0.300). Điều này cho thấy mô hình này thiếu ổn định và không có ưu điểm nổi bật so với hai mô hình còn lại.

  Kết luận: Như vậy, quyết định chọn mô hình phụ thuộc vào ưu tiên chiến lược của doanh nghiệp:

  - Model C: thích hợp khi ưu tiên an toàn tuyệt đối (không bỏ sót xe rủi ro).
  - Model B: là lựa chọn tối ưu nếu doanh nghiệp cần độ tin cậy cảnh báo cao và hiệu quả tổng thể.

TÀI LIỆU THAM KHẢO

[1] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[2] T. Fawcett, "An introduction to roc analysis", *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

[3] D. M. W. Powers, "Evaluation: From precision, recall and f-measure to roc, informedness, markedness & correlation", *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37– 63, 2011.

[4] H. He et al., "Learning from imbalanced data", *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.

[5] D. Chicco et al., "The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation", *PLOS ONE*, vol. 15, no. 6, e0233589, 2020.

[6] G. Tsoumakas et al., "Multi-label classification: An overview", in *International Journal of Data Warehousing and Mining*, 2007, pp. 1–13.