خلاصه‌ای جمع و جور از متریک‌های سنجش:

🧮 معیارهای اصلی ارزیابی مدل‌های دسته‌بندی (با پیش فرض همون ماتریس ۲×۲)

همه بر اساس چهار عدد بدیهی ماتریس درهم‌ریختگی (Confusion Matrix) تعریف می‌شوند:

- TP: تعداد نمونه‌های مثبت واقعی که مثبت پیش‌بینی شده‌اند.
- FP: تعداد نمونه‌های منفی واقعی که مثبت پیش‌بینی شده‌اند.
- FN: تعداد نمونه‌های مثبت واقعی که منفی پیش‌بینی شده‌اند.
- TN: تعداد نمونه‌های منفی واقعی که منفی پیش‌بینی شده‌اند.

---

۱) Accuracy (دقت کل)

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

نسبت کل پیش‌بینی‌های درست به کل نمونه‌ها.

---

۲) Precision (دقت / صحت)

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

از بین پیش‌بینی‌های مثبت، چند درصد واقعاً مثبت بوده‌اند.

---

۳) Recall / Sensitivity / True Positive Rate (TPR)

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

از بین همه‌ی مثبت‌های واقعی، چند درصدش را پیدا کردیم.

---

۴) Specificity / True Negative Rate (TNR)

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

از بین همه‌ی منفی‌های واقعی، چند درصدش را به درستی منفی تشخیص دادیم.
مکمل برای کلاس منفی Recall

---

۵) False Positive Rate (FPR) / Fall-out

$$FPR = \frac{FP}{FP + TN} = 1 - Specificity$$

نسبت نمونه‌های منفی که اشتباهاً مثبت اعلام شده‌اند.
(هرچه کمتر باشد، بهتر است.)

---

۶) False Negative Rate (FNR) / Miss Rate

$$FNR = \frac{FN}{FN + TP} = 1 - Recall$$

نسبت نمونه‌های مثبت که اشتباهاً منفی اعلام شده‌اند.
(هرچه کمتر باشد، بهتر است.)

---

۷) F1-Score

میانگین هارمونیک Precision و Recall:

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

زمانی که بخواهیم تعادل بین Precision و Recall را با یک عدد نشان دهیم.
(بین ۰ و ۱، هر چه نزدیک‌تر به ۱ بهتر است.)

---

۸) Matthews Correlation Coefficient (MCC)

یک معیار متعادل که هر چهار خانه ماتریس را در نظر می‌گیرد و برای مجموعه‌های نامتوازن عالی است:

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

در بازه‌ی $[-1,1]$:

- 1 = پیش‌بینی کامل
- 0 = پیش‌بینی تصادفی
- -1 = پیش‌بینی کاملاً برعکس

---

۹) Negative Predictive Value (NPV)

$$NPV = \frac{TN}{TN + FN}$$

از بین پیش‌بینی‌های منفی، چند درصد واقعاً منفی بوده‌اند.
(شبیه Precision ولی برای کلاس منفی)

---

۱۰) Area Under the ROC Curve (AUC-ROC)

این یک منحنی است و برای فهم کلی:
رابطه‌ی بین Recall (TPR) و FPR را در آستانه‌های مختلف نشان می‌دهد .AUC سطح زیر این منحنی است (بین ۰ و ۱).
۱ = عالی، ۰.۵ = تصادفی.

برای محاسبه‌ی دقیق نیاز به احتمال‌های خروجی مدل دارد، نه فقط ماتریس ۲×۲، ولی از همان ستون‌ها می‌توان نسخه‌ی ساده‌شده‌ای را متصور شد.

حالا بریم در یک مثال بررسی‌شون کنیم 👇

یک تست پزشکی روی ۱ میلیون نفر

فرض می‌کنیم آمار واقعی به این صورت می‌باشد:

- کل جمعیت: 1,000,000
- واقعیت: 100 نفر بیمار (مثبت واقعی) و 999,900 نفر سالم (منفی واقعی)
- پیش‌بینی مدل: 20 نفر را بیمار اعلام کرده (Positive) و بقیه را سالم (Negative)
- جزئیات بیشتر: از آن 20 نفر، 15 نفر واقعاً بیمارند، 5 نفر سالمند که اشتباهاً بیمار گفته شده‌اند.

پس ماتریس درهم‌ریختگی:

	پیش‌بینی: منفی (سالم)	پیش‌بینی: مثبت (بیمار)
واقعاً مثبت (بیمار)	FN = 100 - 15 = 85	TP = 15
واقعاً منفی (سالم)	TN = 999,900 - 5 = 999,895	FP = 5

محاسبه‌ی معیارها:

Accuracy

$$\frac{TP + TN}{N} = \frac{15 + 999{,}895}{1{,}000{,}000} = \frac{999{,}910}{1{,}000{,}000} = 0.99991 \ (99.991\%)$$

Precision

$$\frac{TP}{TP + FP} = \frac{15}{15 + 5} = \frac{15}{20} = 0.75\ (75\%)$$

Recall (Sensitivity)

$$\frac{TP}{TP + FN} = \frac{15}{15 + 85} = \frac{15}{100} = 0.15\ (15\%)$$

Specificity

$$\frac{TN}{TN + FP} = \frac{999{,}895}{999{,}895 + 5} = \frac{999{,}895}{999{,}900} \approx 0.999995\ (99.9995\%)$$

False Positive Rate (FPR)

$$\frac{FP}{FP + TN} = \frac{5}{999{,}900} \approx 0.000005\ (0.0005\%)$$

False Negative Rate (FNR)

$$\frac{FN}{FN + TP} = \frac{85}{100} = 0.85\ (85\%)$$

F1-Score

$$2 \cdot \frac{0.75 \cdot 0.15}{0.75 + 0.15} = 2 \cdot \frac{0.1125}{0.90} = 0.25$$

MCC

$$\frac{15 \cdot 999{,}895 - 5 \cdot 85}{\sqrt{(20)(100)(999{,}900)(999{,}980)}} = \frac{14{,}998{,}425 - 425}{\sqrt{20 \cdot 100 \cdot 999{,}900 \cdot 999{,}980}}$$

مخرج خیلی بزرگ است ← مقدار MCC بسیار کوچک، نزدیک به 0.037 (نشان‌دهنده‌ی همبستگی ضعیف است)

NPV

$$\frac{TN}{TN+FN} = \frac{999{,}895}{999{,}895+85} = \frac{999{,}895}{999{,}980} \approx 0.999915\ (99.9915\%)$$

تحلیل و تفسیرات:

- Accuracy ٪۹۹.۹۹۱ است؛ در نگاه اول مدل عالی به نظر می‌رسد. اما این فریبنده است، چون کلاس سالم بسیار بزرگ است.
- Precision ٪۷۵ یعنی از هر ۴ نفری که مدل بیمار اعلام کرد، ۳ نفر واقعاً بیمار بودند، یک نفر سالم بوده که بد نیست.
- Recall فقط ٪۱۵ است!!! یعنی از ۱۰۰ بیمار واقعی، فقط ۱۵ نفر را پیدا کرده، ۸۵ نفر را از دست داده است که این فاجعه است.
- Specificity تقریباً کامل (٪۹۹.۹۹۹۵): تقریباً هیچ سالمی را اشتباهی بیمار نخوانده و این عالیه.
- FPR بسیار پایین (٪۰.۰۰۰۵) یعنی آلارم‌های اشتباهی بسیار کم.
- FNR بسیار بالا (٪۸۵) یعنی اکثر بیماران نادیده گرفته شده‌اند.
- F1-Score ۰.۲۵ که نشان‌دهنده‌ی توازن بد Precision و Recall است. مدل یا Recall را فدای Precision کرده، یا برعکس؛ اما F1 در نهایت ضعیف است.
- MCC نزدیک به صفر (۰.۰۳۷) یعنی مدل تقریباً تصادفی عمل می‌کند، با وجود Accuracy بالا! این مهر تاییدی است بر اینکه Accuracy در مجموعه‌های نامتوازن گمراه‌کننده است.
- NPV بسیار بالا (٪۹۹.۹۹۱۵) یعنی اگر مدل بگوید سالمه، تقریبا مطمئنی که سالم هست. اما این به خاطر فراوانی سالم‌هاست.

نتیجه: این مدل عملا فقط ۱۵٪ بیماران را پیدا کرده و اصلا مناسب یک سیستم غربالگری نیست!! تمام معیارهای غیر از Accuracy این ضعف را فریاد می‌زنند. پس برای ارزیابی واقعی یک مدل باید ترکیب معیارها (حداقل Recall ، Precision، Specificity و F1-score را دید، نه فقط Accuracy)