

APRENDIZADO DE MÁQUINA PARA SAÚDE

# Avaliação de Desempenho: além da acurácia

---

Métricas, calibração, validação externa e equidade em modelos clínicos

APOSTILA TÉCNICA · ENCONTRO 05

Prof. Flávio Luiz Seixas  
Pós-graduação em Computação

## Sumário

|                                                    |   |
|----------------------------------------------------|---|
| Sumário .....                                      | 2 |
| Introdução .....                                   | 3 |
| 1. Matriz de confusão e medidas condicionais ..... | 3 |
| 2. Limiar, discriminação e desbalanceamento .....  | 3 |
| 3. Calibração e validação .....                    | 4 |
| 4. Equidade por subgrupos .....                    | 4 |
| Síntese do encontro.....                           | 4 |
| Referências .....                                  | 4 |

## Introdução

Avaliar um classificador clínico exige relacionar erros a prevalência, limiar e consequência. Acurácia agregada pode parecer alta mesmo quando o modelo não identifica o evento relevante, sobretudo em problemas desbalanceados. [1]

O encontro conecta matriz de confusão, sensibilidade, valor preditivo, curvas, calibração, validação e análise de subgrupos. A pergunta deixa de ser “qual número é maior?” e passa a ser “que propriedade foi medida e que decisão ela permite sustentar?”. [1]



Figura 1 — Matriz de confusão interpretada como consequências de classificação. Fonte: elaboração própria com base em [1].

## 1. Matriz de confusão e medidas condicionais

A matriz de confusão separa verdadeiros positivos (VP), falsos positivos (FP), verdadeiros negativos (VN) e falsos negativos (FN). A sensibilidade é  $VP/(VP+FN)$ : entre os casos com a condição, quantos são detectados. A especificidade é  $VN/(VN+FP)$ : entre os casos sem a condição, quantos recebem resultado negativo [1].

O valor preditivo positivo — equivalente à precisão no vocabulário de classificação — é  $VP/(VP+FP)$ . Diferentemente de sensibilidade e especificidade, valores preditivos dependem da prevalência na população. Portanto, um mesmo teste pode produzir proporções de positivos verdadeiros diferentes quando muda o contexto de uso [1].

### CONCEITO-CHAVE

A matriz de confusão separa verdadeiros positivos (VP), falsos positivos (FP), verdadeiros negativos (VN) e falsos negativos (FN).

## 2. Limiar, discriminação e desbalanceamento

Alterar o limiar desloca o equilíbrio entre falsos positivos e falsos negativos. A curva ROC resume sensibilidade contra taxa de falsos positivos ao longo dos limiares; a curva precisão–recall enfatiza o

comportamento da classe positiva. Nenhuma área sob curva escolhe sozinha o limiar clínico, porque custos e capacidade de intervenção permanecem externos à métrica [2].

A medida F1 combina precisão e sensibilidade por média harmônica. Ela é útil quando ambas importam, mas não incorpora diretamente especificidade, calibração ou utilidades clínicas. Métricas devem ser escolhidas a partir da tarefa e das consequências [5].

### 3. Calibração e validação

Discriminação ordena indivíduos por risco; calibração compara probabilidades previstas com frequências observadas. Quando probabilidades orientam decisão personalizada, calibração é central. Ela pode deteriorar com mudanças de população, incidência e prática [2].

Validação interna estima otimismo dentro do processo de desenvolvimento. Validação temporal e externa testa transportabilidade para outros períodos e instituições. Uma avaliação clínica completa avança do desempenho estatístico para integração, uso e impacto [3].

### 4. Equidade por subgrupos

Uma média global pode esconder diferenças entre grupos. Auditoria por subgrupo deve comparar cobertura dos dados, prevalência, calibração e taxas de erro relevantes. A presença de disparidade exige investigação de dados, formulação, modelo e uso, porque o viés pode surgir em todas essas fases [4].

### Síntese do encontro

|                                                                                                                                                                               |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Matriz de confusão e medidas condicionais: a matriz de confusão separa verdadeiros positivos (vp), falsos positivos (fp), verdadeiros negativos (vn) e falsos negativos (fn). |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

|                                                                                                                            |
|----------------------------------------------------------------------------------------------------------------------------|
| Limiar, discriminação e desbalanceamento: alterar o limiar desloca o equilíbrio entre falsos positivos e falsos negativos. |
|----------------------------------------------------------------------------------------------------------------------------|

|                                                                                                                                            |
|--------------------------------------------------------------------------------------------------------------------------------------------|
| Calibração e validação: discriminação ordena indivíduos por risco; calibração compara probabilidades previstas com frequências observadas. |
|--------------------------------------------------------------------------------------------------------------------------------------------|

|                                                                                 |
|---------------------------------------------------------------------------------|
| Equidade por subgrupos: uma média global pode esconder diferenças entre grupos. |
|---------------------------------------------------------------------------------|

### Referências

- [1] Owens, D. K.; Goldhaber-Fiebert, J. D.; Sox, H. C. Biomedical Decision Making: Probabilistic Clinical Reasoning. In: Biomedical Informatics. Springer, 2021.
- [2] Bellazzi, R.; Dagliati, A.; Nicora, G. Predicting Medical Outcomes. In: Intelligent Systems in Medicine and Health. Springer, 2022.
- [3] Shortliffe, E. H.; Sepúlveda, M.-J.; Patel, V. L. Framework for the Evaluation of Clinical AI Systems. In: Intelligent Systems in Medicine and Health. Springer, 2022.
- [4] Korngiebel, D. M.; Solomonides, A.; Goodman, K. W. Ethical and Policy Issues. In: Intelligent Systems in Medicine and Health. Springer, 2022.

- [5] Subramanian, D.; Cohen, T. A. Machine Learning Systems. In: Intelligent Systems in Medicine and Health. Springer, 2022.