

APRENDIZADO DE MÁQUINA PARA SAÚDE

Dados Clínicos: o que o modelo realmente vê?

Heterogeneidade, qualidade, missingness e vazamento de informação

APOSTILA TÉCNICA · ENCONTRO 02

Prof. Flávio Luiz Seixas
Pós-graduação em Computação

Sumário

Sumário	2
Introdução	3
1. O ecossistema de dados clínicos	3
2. Qualidade como adequação ao uso.....	4
3. Tempo, seleção e leakage	4
4. Governança e proveniência	4
Síntese do encontro.....	4
Referências	5

Introdução

Dados clínicos não são uma observação contínua e neutra do paciente. Eles resultam de encontros, instrumentos, práticas de documentação e decisões anteriores. Assim, a matriz entregue ao algoritmo contém tanto sinais biológicos quanto marcas do processo assistencial. [1]

O objetivo deste encontro é reconhecer tipos de dados, estrutura, temporalidade, mecanismos de ausência, vieses de seleção e vazamento. Essas propriedades determinam o que o modelo pode aprender e onde seu desempenho pode falhar. [1]

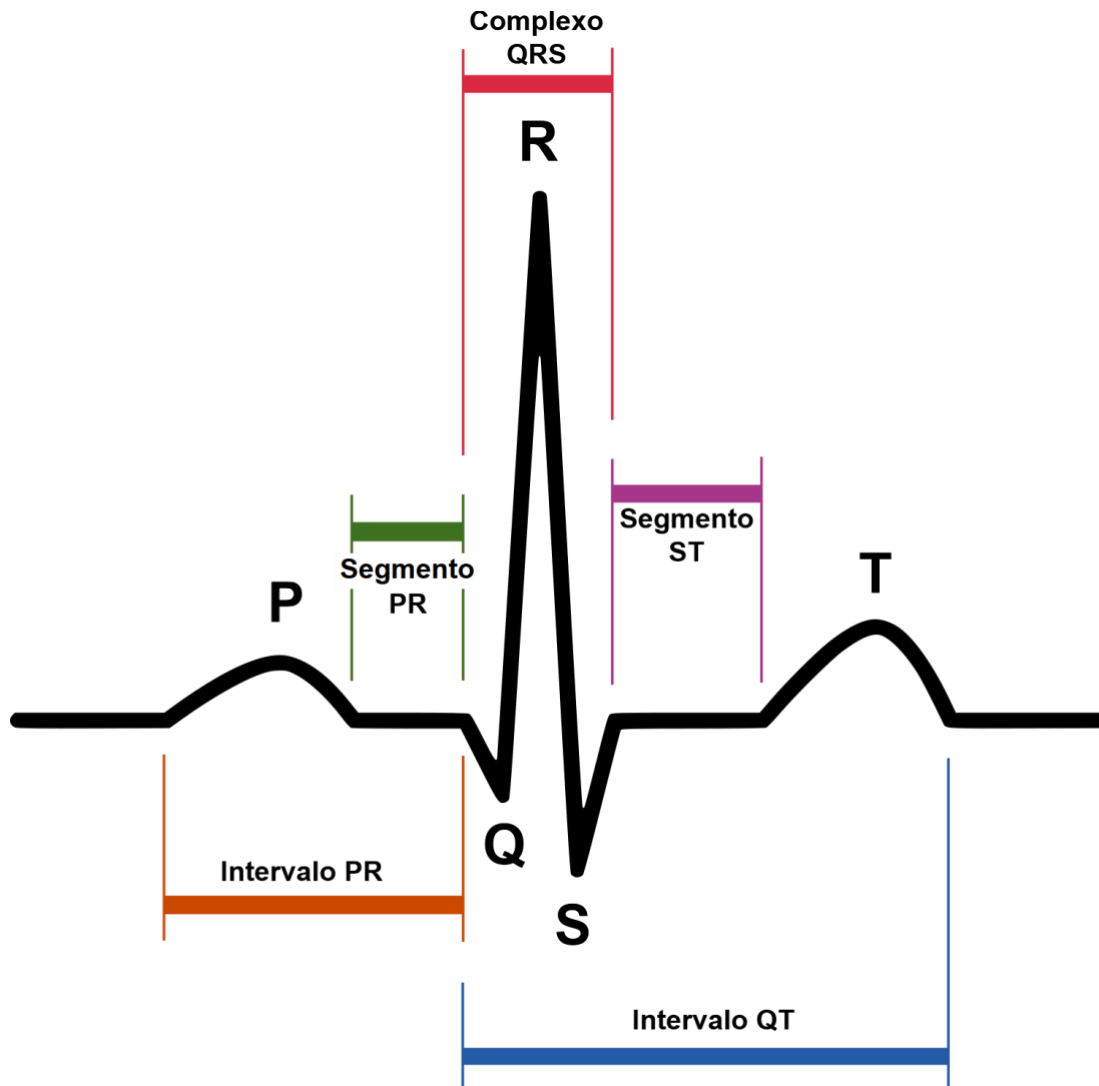


Figura 1 — Sinal eletrocardiográfico como exemplo de dado temporal. Fonte visual indicada no README do encontro; discussão baseada em [1] e [2].

1. O ecossistema de dados clínicos

O registro de saúde reúne observações produzidas por profissionais, pacientes e dispositivos. Dados estruturados facilitam consulta e agregação, enquanto narrativas preservam contexto e relações difíceis de antecipar em campos fixos. Nenhuma forma é automaticamente completa ou correta [1].

Imagens, sinais, dados moleculares e informações geradas pelo paciente têm estruturas e escalas distintas. Imagens digitais carregam pixels e metadados; sinais representam amostras no tempo; dados ômicos frequentemente combinam grande dimensionalidade e amostras menores. A integração exige preservar proveniência e significado [2].

CONCEITO-CHAVE

O registro de saúde reúne observações produzidas por profissionais, pacientes e dispositivos.

2. Qualidade como adequação ao uso

Qualidade deve ser julgada em relação à finalidade. Um código adequado para faturamento pode ser insuficiente para representar um fenótipo; um valor coletado no cuidado pode refletir seleção clínica e não uma medição planejada para pesquisa. Antes da modelagem, cada variável deve ser descrita por origem, unidade, tempo, transformação e motivo de coleta [1].

A ausência de dados pode ser completamente aleatória, depender de informações observadas ou relacionar-se a fatores não observados. Essas situações correspondem, respectivamente, a MCAR, MAR e MNAR. Excluir registros incompletos ou imputar valores sem considerar o mecanismo pode alterar a população analisada e o sinal disponível [3].

3. Tempo, seleção e leakage

Em predição clínica, o tempo define o que estaria disponível no momento do uso. Variáveis registradas depois do instante de decisão podem revelar o desfecho ou uma ação causada por ele. Esse vazamento produz estimativas otimistas que não se reproduzem quando o sistema é usado prospectivamente [4].

Também é preciso perguntar quem entra no conjunto de dados. A probabilidade de hospitalização, exame, retorno ou registro pode depender de acesso, gravidade e prática local. O hospital, o equipamento ou o protocolo podem tornar-se atalhos que o modelo aprende no lugar do fenômeno pretendido [4].

4. Governança e proveniência

A governança deve registrar quem pode acessar os dados, para qual finalidade e sob quais controles. Proveniência permite seguir a trajetória de uma variável desde a aquisição até a representação utilizada pelo modelo. Essa documentação é essencial para auditoria, reprodução e avaliação de mudança [5].

Síntese do encontro

O ecossistema de dados clínicos: o registro de saúde reúne observações produzidas por profissionais, pacientes e dispositivos.

Qualidade como adequação ao uso: qualidade deve ser julgada em relação à finalidade.

Tempo, seleção e leakage: em predição clínica, o tempo define o que estaria disponível no momento do uso.

Governança e proveniência: a governança deve registrar quem pode acessar os dados, para qual finalidade e sob quais controles.

Referências

- [1] Shortliffe, E. H.; Chiang, M. F. Biomedical Data: Their Acquisition, Storage, and Use. In: Biomedical Informatics. Springer, 2021.
- [2] Sim, I.; Sirota, M. Data and Computation: A Contemporary Landscape. In: Intelligent Systems in Medicine and Health. Springer, 2022.
- [3] Bellazzi, R.; Dagliati, A.; Nicora, G. Predicting Medical Outcomes. In: Intelligent Systems in Medicine and Health. Springer, 2022.
- [4] Subramanian, D.; Cohen, T. A. Machine Learning Systems. In: Intelligent Systems in Medicine and Health. Springer, 2022.
- [5] Goodman, K. W.; Miller, R. A. Ethics in Biomedical and Health Informatics: Users, Standards, and Outcomes. In: Biomedical Informatics. Springer, 2021.