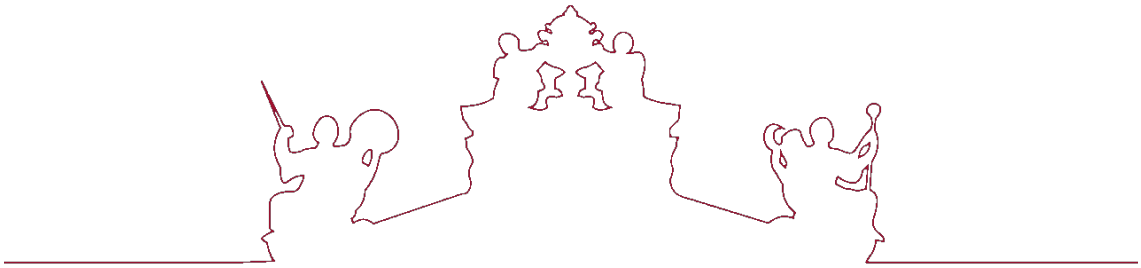

Um revisitar da regressão logística: da modelação ao impacto na tomada de decisão

(Sumário pormenorizado de lição no âmbito das Provas de Agregação)

Paulo de Jesus Infante dos Santos

Universidade de Évora
Escola de Ciências e Tecnologia
Departamento de Matemática

Outubro de 2025

Sumário pormenorizado

O sumário pormenorizado da lição com o título “Um revisitar da regressão logística: da modelação ao impacto na tomada de decisão” é apresentado à Universidade de Évora no âmbito das Provas de Agregação em Matemática ao abrigo da alínea c) do artigo 5.º do Decreto-Lei n.º 239/2007, de 19 de junho, com a atualização introduzida pelo Decreto-Lei n.º 64/2023, de 31 de julho.

Évora, 21 de outubro de 2025

Assinado por: **Paulo de Jesus Infante dos Santos**
Num. de Identificação: 09628833
Data: 2025.10.21 15:07:36 +0100

Paulo de Jesus Infante dos Santos

Preâmbulo

Esta sessão assume deliberadamente um carácter de lição-seminário. Mais do que expor técnicas, pretende problematizar decisões de modelação e mostrar como escolhas metodológicas críticas, como interações, validação, calibração, definição de pontos de corte e respeito pelo desenho amostral, transformam a análise em tomadas de decisões eficazes e informadas. A opção por este formato resulta do perfil da turma (formações diversas e objetivos profissionais distintos) e da própria natureza da unidade curricular *Análise de Dados Categóricos*, onde se insere.

Esta unidade curricular (UC) privilegia a aplicação prática dos métodos estatísticos a dados reais, com atenção à reprodutibilidade e à comunicação adequada a diferentes públicos. Esta abordagem é estruturante, procurando transformar o domínio técnico em competência aplicada: documentar processos, justificar escolhas, diagnosticar pressupostos e saber comunicar resultados. A regressão logística, tema central desta lição, constitui o núcleo metodológico da UC: é a técnica de referência para respostas binárias, largamente utilizada nas mais diversas áreas do conhecimento, e o ponto de partida para aprofundar a análise através de extensões e boas práticas. É também uma das ferramentas estatísticas que o autor desta lição mais tem utilizado para produzir conhecimento que apoia a tomada de decisão.

A pertinência do tema "Um Revisitar da Regressão Logística: da modelação ao impacto na tomada de decisão" decorre da necessidade de ir além do ajustamento técnico, abordando escolhas de modelação que têm impacto direto na interpretação e na ação. Os tópicos abordados são decisivos para transformar resultados estatísticos em recomendações fiáveis e equitativas em domínios como a saúde pública, a sinistralidade rodoviária ou a gestão desportiva municipal.

Esta lição-seminário reflete o ADN da unidade curricular: rigor científico, aplicabilidade prática e comunicação clara orientada à decisão. O conceito de revisitar sinaliza o voltar aos fundamentos para os levar mais longe, tornando explícito o que muda quando passamos da técnica ao apoio efetivo à tomada de decisão. E, neste caso particular, quando trabalhamos com bases reais de saúde e atividade física, e quando as recomendações têm impacto direto em pessoas, serviços e territórios, o objetivo fundamental é garantir que a evidência científica se traduz em políticas públicas consistentes, equitativas e orientadas para o bem-estar coletivo.

Índice

1. Introdução 5
2. Objetivos 6
3. Roteiro da Lição 6
4. Súmula da Lição 7
    4.1. Interações em regressão logística: importância para a interpretação e apoio à tomada de decisão 7
    4.2. Validação e calibração em regressão logística: garantir probabilidades fiáveis para o apoio à tomada de decisão 9
    4.3. Ponto de corte na regressão logística: implicações para a classificação e apoio à tomada de decisão 11
    4.4. Amostras complexas e pesos de calibração: o que muda, porquê e as implicações no apoio à tomada de decisão 13
    4.5. O que a perspetiva de *machine learning* pode acrescentar ao apoio à tomada de decisão 14
5. Bibliografia Recomendada 16

1. Introdução

De acordo com a organização teórica da UC *Análise de Dados Categóricos*, a lição-seminário proposta insere-se no seu núcleo metodológico: os modelos lineares generalizados (GLM) aplicados a respostas categóricas. Em particular, centra-se na regressão logística binária, matéria central do programa e fundamento para extensões a modelos ordinais e multinomiais. A compreensão desta lição apoia-se em conteúdos trabalhados em momentos anteriores: a introdução aos GLM, a estimação via máxima verosimilhança, os diagnósticos de bondade de ajustamento e de desempenho preditivo e a interpretação de coeficientes. Estes elementos constituem a base necessária para que os estudantes possam acompanhar uma abordagem mais avançada, que problematiza aspetos frequentemente negligenciados, mas fundamentais para a validade e utilidade prática de um modelo.

A regressão logística foi escolhida por duas razões complementares. Por um lado, pela sua ubiquidade, pois é talvez a técnica mais usada na análise de dados categóricos em saúde pública, ciências sociais, economia, educação ou desporto, sendo um elo comum entre diferentes percursos académicos da turma. Por outro lado, pela sua plasticidade pedagógica, pois permite consolidar fundamentos estatísticos e, ao mesmo tempo, mostrar como escolhas metodológicas aparentemente técnicas, tais como considerar ou não interações, definir ou não um ponto de corte (limiar) de classificação, respeitar ou ignorar pesos de amostragem, têm consequências reais na interpretação e na decisão. Revisitar a logística traduz-se, assim, na expansão do olhar: a transição de modelos ajustados em 'modo automático' para modelos que resultam de uma consciencialização crítica e que asseguram a sua relevância aplicada.

O tema ganha ainda mais pertinência no quadro da filosofia pedagógica da UC. Esta não se limita à transmissão de técnicas, mas procura desenvolver competência transversal em análise de dados*:* capacidade de ler resultados estatísticos de forma crítica, de dialogar com outras unidades curriculares e de comunicar conclusões de modo transparente para públicos não especialistas. A lição-seminário reforça esta dimensão, ao aproximar estatística, aplicação prática e tomada de decisão em contextos de intervenção, como os cuidados de enfermagem (na predição de risco em saúde) ou a gestão de programas de atividade física municipal.

Finalmente, importa sublinhar a heterogeneidade da turma, formada por estudantes dos mestrados em Matemática, Modelação Estatística e Análise de Dados, Inteligência Artificial e Ciência de Dados e Epidemiologia. Essa diversidade é simultaneamente um desafio e uma oportunidade: exige uma exposição clara, capaz de articular teoria e prática, mas também enriquece a discussão, pois cada estudante pode reconhecer nos exemplos tratados a relevância da modelação estatística para o seu domínio específico. É nesta confluência que a lição encontra o seu sentido mais profundo: constituir-se como um exercício de elevado rigor metodológico que se converte em competência transversal, fundamental para o futuro especialista transformar a análise de dados em apoio estratégico à tomada de decisão.

2. Objetivos

Esta lição-seminário tem como objetivo revisitar a regressão logística, não para repetir conteúdos, mas para aprofundar aspetos críticos que condicionam a validade dos resultados e a utilidade prática da análise. Pretende-se que os estudantes:

i) Consolidem fundamentos da regressão logística como base para uma análise crítica mais avançada.
ii) Reconheçam o impacto de escolhas metodológicas cruciais (inclusão de interações, definição de pontos de corte, validação e calibração, incorporação de pesos em amostras complexas) na interpretação e no apoio à decisão.
iii) Desenvolvam sensibilidade crítica para distinguir entre desempenho "na amostra" e utilidade real em contextos aplicados.
iv) Reforcem competências de literacia alargada, capazes de ligar rigor estatístico a comunicação clara e a decisões informadas.

Mais do que apresentar novos métodos, esta lição procura formar uma atitude analítica: compreender que cada escolha de modelação pode alterar as conclusões possíveis e, por consequência, as decisões que delas derivam.

3. Roteiro da Lição

Com base nos objetivos definidos, a lição-seminário organiza-se de forma progressiva: começamos por revisitar a regressão logística no contexto da unidade curricular, avançamos para decisões de modelação que alteram interpretações e recomendações (interações, validação e calibração), passamos à transformação de probabilidades em ação (pontos de corte e critérios operacionais) e discutimos o que muda quando os dados provêm de amostras complexas (pesos e inferência populacional). Serão sempre incluídos exemplos com dados reais para fomentar discussão. Encerramos com uma nota complementar sobre o que abordagens de *machine learning* podem acrescentar, não para substituir a logística, mas para enriquecer a decisão.

1. Revisitar a regressão logística.
   1.1. A logística como técnica central para respostas binárias no núcleo da *Análise de Dados Categóricos*.
   1.2. Do ajustamento à decisão: ir além dos efeitos principais e integrar escolhas que mudam recomendações.

2. Interações em regressão logística: importância para a interpretação e apoio à tomada de decisão.
   2.1. Quando efeitos "médios" escondem efeitos condicionais.
   2.2. Variáveis aparentemente secundárias que se tornam determinantes em interação.
   2.3. Implicações para a decisão: recomendações que mudam ao incorporar interações.

3. Validação e calibração em regressão logística: garantir probabilidades fiáveis para o apoio à tomada de decisão.
   3.1. Divisão treino–teste e limitações em amostras pequenas.

3.2. Validação cruzada e *bootstrap*: estabilidade e correção do otimismo.
3.3. Calibração — intercepto, declive e curva: consequências práticas de não calibrar.

4. Ponto de corte na regressão logística: implicações para a classificação e apoio à tomada de decisão
4.1. Ponto de corte fixo (0,5) vs. critérios orientados por objetivo (Youden; sensibilidade mínima).
4.2. Ponto de corte por capacidade (*Top-K*): alinhar decisão com recursos disponíveis.
4.3. Equidade e eficiência: como os diferentes critérios de limiar moldam a priorização das ações.

5. Amostras complexas e pesos de calibração: o que muda, porquê e as implicações no apoio à tomada de decisão
5.1. Estratificação, conglomerados e calibração: o que está em causa.
5.2. Modelar e avaliar com pesos: coeficientes, incerteza e métricas populacionais.
5.3. Decidir para a amostra vs. decidir para a população: implicações práticas.

6. O que a perspetiva de *machine learning* pode acrescentar ao apoio à tomada de decisão.
6.1. Menção dos Modelos de Random Forest (RF) e Extreme Gradient Boosting (XGBoost) para captura de padrões não lineares, interações implícitas e contributos individuais via Shapley Additive Explanations (SHAP).
6.2. Leitura integrada e implicações práticas: reforço do papel das infraestruturas, segmentação de políticas e prestação de contas transparente.

4. Súmula da Lição

Segue-se agora o desenvolvimento de cada tópico em maior detalhe. Iniciaremos pela análise das interações em regressão logística, dado que este constitui um ponto nevrálgico para compreender como variáveis aparentemente pouco relevantes ou mesmo não significativas podem assumir papel determinante quando consideradas em interação com outras. Do mesmo modo, permite evidenciar como as conclusões que sustentam a tomada de decisão podem ser distintas se analisarmos as variáveis isoladamente ou em interação. Esta escolha não é apenas metodológica, sendo também pedagógica, pois demonstra de forma clara aos estudantes que conclusões simplificadas podem ocultar padrões relevantes, conduzindo a decisões desajustadas em diferentes contextos.

4.1. Interações em regressão logística: importância para a interpretação e apoio à tomada de decisão

A regressão logística é amplamente utilizada na análise de dados categóricos nas mais diversas áreas. É comum que as análises se restrinjam aos efeitos principais, pressupondo que a influência de cada preditor se mantém constante em todos os indivíduos ou contextos. Contudo, a realidade empírica mostra que muitos efeitos são condicionais, isto é, dependem do nível de outros fatores. Ignorar estas interações pode conduzir a

conclusões enviesadas, mascarando padrões reais e comprometendo a eficácia das intervenções.

Com a abordagem deste tópico pretende-se consolidar o que foi referido em lições anteriores acerca da modelação estatística de dados categóricos binários. Especificamente, procura-se:

i) mostrar que variáveis não significativas nos efeitos principais podem revelar-se determinantes quando consideradas em interação;

ii) demonstrar como as conclusões sobre os determinantes de um evento mudam quando se ignoram ou incorporam interações significativas;

iii) discutir as implicações práticas para a tomada de decisão.

Com recurso a dados reais de um inquérito municipal à atividade física em Évora de 2012, modela-se a prática de caminhadas entre praticantes de atividade física. São comparados dois modelos: um modelo composto apenas por efeitos principais e outro que integra interações estatisticamente significativas e contextualmente fundamentadas (*sexo* × *tempo de prática*; *uso da ecopista* × i*dade*).

No modelo sem interações, a variável *uso de ecopista* não é estatisticamente significativa, embora revele indícios de poder ter efeito na prática de atividade física, pois é marginalmente significativa. Num processo de modelação automatizada (p. ex., *stepwise*), ou mesmo seguindo os passos usuais ensinados na fase inicial da unidade curricular de Análise de Dados Categóricos, esta variável seria excluída e considerada irrelevante. No entanto, ao incluir a interação com a idade, torna-se evidente que o efeito do uso da ecopista depende da faixa etária: irrelevante em jovens, mas claramente positivo e crescente a partir da meia-idade.

De forma análoga, a interação entre sexo e duração das sessões revela um padrão claro: em sessões curtas, não há diferenças relevantes entre sexos; em sessões de maior duração, as mulheres apresentam maior propensão para optar por caminhadas. Nos homens, pelo contrário, sessões mais curtas associam-se a maior probabilidade de caminhar, enquanto nas mulheres o tempo de prática não altera significativamente essa opção. Além disso, entre não utilizadores de ecopista a idade, por si só, não é determinante; já entre utilizadores, a idade reforça progressivamente a propensão para caminhar, evidenciando um efeito ambiental que cresce com a idade.

Deste modo, o exemplo mostra que uma variável "insignificante" no efeito principal pode, em interação, revelar padrões cruciais; e que modelos que ignoram interações produzem conclusões simplistas, incapazes de refletir a heterogeneidade real. Se ignorarmos as interações, a recomendação tenderia para políticas uniformes, indiferentes a sexo, idade, padrões de prática e acesso/uso de ecopista, com risco de desperdício de recursos e manutenção de desigualdades. Integrando as interações, a decisão muda de natureza e torna-se possível de operacionalizar:

i) Sessões curtas dirigidas a homens como porta de entrada (adesão inicial), articulando horários e locais de conveniência.
ii) Nas mulheres, não penalizar sessões mais longas (não reduzir duração por receio de "quebra" na adesão), oferecendo rotas/horários compatíveis.

iii) Ecopistas como alavanca a partir da meia-idade: priorizar manutenção e ligações seguras nos corredores com maior presença de adultos mais velhos; garantir conforto (sombra, bancos, iluminação) e continuidade dos traçados.
iv) Para não utilizadores de ecopista, trabalhar barreiras de acesso e de conhecimento (sinalética, mapas simples, ações de conhecimento), porque sem uso não há efeito e, nesse caso, a idade isolada pouco acrescenta.
v) Comunicação segmentada: mensagens e canais diferenciados por sexo, idade e padrão de prática (p. ex., grupos de caminhada curtos para homens iniciantes; rotas mais longas para mulheres).
vi) Monitorização por segmento e por acesso à ecopista (utiliza vs. não utiliza), para avaliar se as recomendações estão a reduzir as disparidades esperadas.

A diferença entre estes dois cenários é relevante: num, prevalecem políticas generalistas, pouco eficazes e desajustadas; no outro, implementam-se intervenções direcionadas, eficientes e socialmente mais equitativas, onde a ecopista deixa de ser uma "variável marginal" e passa a instrumento tático cuja efetividade depende de quem a usa e em que fase da vida se encontra. Consequentemente, este exemplo permite não ilustrar apenas a importância das interações na análise estatística aplicada, mas mostrar que a omissão deste aspeto metodológico pode traduzir-se em más políticas públicas, ao passo que a sua incorporação abre caminho para decisões baseadas em evidência, segmentadas e com impacto real.

Antes de fechar este tópico, reforçamos três aspetos já trabalhados em aulas anteriores: i) mesmo que não seja estatisticamente significativa, uma variável deve permanecer no modelo quando a experiência e o conhecimento sobre o fenómeno assim o justificarem, sendo muitas vezes necessário recorrer à colaboração interdisciplinar, dado que o estatístico pode não dominar todos os aspetos contextuais envolvidos; ii) nem todas as interações estatisticamente significativas devem ser incluídas no modelo, sendo importante avaliar se fazem sentido à luz do problema em estudo; iii) a inclusão de múltiplas interações, sobretudo envolvendo as mesmas variáveis, pode complicar desnecessariamente a comunicação; por isso, privilegia-se parcimónia, mantendo apenas as interações que apresentam maior robustez estatística e relevância prática.

4.2. Validação e calibração em regressão logística: garantir probabilidades fiáveis para o apoio à tomada de decisão

Antes de transformar probabilidades em decisões, é indispensável assegurar duas propriedades: validação (desempenho fora da amostra de ajustamento) e calibração (as probabilidades preditas corresponderem, em média, às frequências observadas). Um modelo pode apresentar um excelente desempenho na amostra de treino e falhar na prática (sobreajustamento); ou pode discriminar bem quem tem ou não tem o evento de interesse e, ainda assim, produzir probabilidades não calibradas, induzindo decisões erradas.

Em amostras moderadas a grandes, uma estratégia robusta é dividir os dados em conjunto de treino e conjunto de teste (80/20 ou 70/30). Todo o ciclo de modelação (seleção de

variáveis, interações, pré-processamento) faz-se exclusivamente no treino; o teste permanece "intocado" até ao fim, para estimar o desempenho fora da amostra. Em contextos temporais, a divisão deve ser temporal (treino: período inicial; teste: período posterior), evitando fugas de informação. Em qualquer cenário, a divisão deve ser estratificada pelo evento para preservar a prevalência.

Quando a amostra é pequena ou existe forte desequilíbrio entre classes (poucos casos na categoria minoritária), a simples divisão 80/20 ou 70/30 desperdiça informação para treino e torna a avaliação no teste instável. Nesses cenários, é preferível recorrer à validação cruzada (por exemplo, dividir a amostra em 10 subconjuntos e alternar entre treino e teste, repetindo o processo para reduzir variância) ou ao bootstrap (reamostrar com reposição para estimar o otimismo e corrigir a sobreavaliação do desempenho). A ideia central é usar todas as observações de forma rotativa para treinar e validar, obtendo métricas mais estáveis e menos dependentes de uma única partição.

Em contextos de pequeno número de observações informativas ou forte desequilíbrio entre classes, aumenta o risco de sobreajustamento e de separação ou quase-separação, situações em que uma combinação de preditores distingue perfeitamente (ou quase perfeitamente) as classes, levando a coeficientes muito grandes e instáveis. Nestes casos, recomenda-se parcimónia (limitar o número de preditores), penalização (ridge, lasso) ou correções como o método de Firth (verosimilhança penalizada), que estabilizam as estimativas e evitam resultados espúrios.

A calibração deve ser verificada explicitamente em três dimensões: (i) *calibration-in-the-large* (calibração na média), verificando que o intercepto do modelo recalibrado é ≈ 0, garantindo que a probabilidade média do evento ocorrer não é sobrestimada nem subestimada; (ii) *calibration slope* (declive ≈ 1), que assegura que as previsões não são excessivamente extremas (sobreajustamento) nem demasiado conservadoras (subajustamento); e (iii) curva de calibração (probabilidades previstas *versus* observadas). Em complemento, o *Brier score* fornece uma medida global que combina discriminação e calibração, permitindo comparar diferentes modelos de forma agregada. Um modelo com área abaixo da curva ROC (AUC) elevada, mas má calibração pode induzir em erro: "30% de risco" não é 30 em 100 pacientes, mas pode, por exemplo, ser 25 ou 35. Se necessário, procede-se a recalibração (ajuste de intercepto/declive) ou a calibração pós-ajuste (p. ex., isotónica), sempre com validação interna.

Por fim, como ponte para o tópico seguinte (ponto de corte) refere-se que em desbalanceamentos elevados é prudente introduzir um terceiro bloco 60/20/20 (treino/validação/teste). O treino ajusta o modelo; a validação escolhe pontos de corte e afina decisões (p. ex., maximizar sensibilidade mínima desejada, equilibrar *precisão–recall*), e o teste estima o desempenho final. Quando a amostra não comporta 60/20/20, recorre-se a validação cruzada aninhada, onde se usam *folds* externos para avaliar o desempenho e internos para determinar o ponto de corte, sempre com estratificação e sem "olhar" para o teste ao decidir o ponto de corte. Qualquer reamostragem para lidar com desbalanceamento (p. ex., SMOTE – sobreamostragem sintética para equilibrar o conjunto de dados) faz-se apenas no treino, mantendo validação e teste intactos.

Para apresentar este tópico considera-se o modelo preditivo para o risco de desenvolvimento de feridas complexas em contexto de cuidados de saúde. Mostra-se que a divisão treino/teste e a validação cruzada apresentam desempenho consistente fora da amostra e métricas estáveis, enquanto o *bootstrap* confirma otimismo residual de pequena magnitude. A verificação da calibração revela intercepto próximo de zero e declive próximo de um, sinal de que as probabilidades previstas são, em média, fiáveis. Emerge, ainda, um ponto didático importante: categorias pouco frequentes (p. ex., ferida em membro superior) induzem quase-separação, com coeficientes inflacionados e probabilidades extremas; a regressão logística com penalização de *Firth* mostra-se apropriada para estabilizar a inferência sem perda de utilidade prática.

Faz-se notar que para este caso, se não validássemos e calibrássemos, correríamos o risco de confiar num modelo que não se replica noutras unidades, levando a atrasos no escalonamento de cuidados (falsos negativos) ou a sobrecarga de equipas (falsos positivos). Ignorando a calibração podemos distorcer prioridades de vigilância. Se não identificarmos e corrigirmos a quase-separação, certos coeficientes parecem "determinísticos", induzindo regras clínicas artificiais (p. ex., "local X implica complexidade") que não se sustentam e prejudicam a comunicação entre equipas. Conclui-se que validar, calibrar e, quando necessário, aplicar penalização é condição para transformar probabilidades em decisões seguras e exequíveis: priorizar quem precisa, evitar atrasos e alocar recursos de forma realista.

4.3. Ponto de corte na regressão logística: implicações para a classificação e apoio à tomada de decisão

A regressão logística traduz relações entre variáveis em probabilidades de ocorrência de um evento. Contudo, a tomada de decisão raramente se esgota na leitura dessas probabilidades. Na prática, é necessário definir um ponto de corte (*threshold*) que transforme previsões contínuas em decisões binárias de ocorrência desse evento e que conduz a agir ou não agir, classificar como provável ou improvável, incluir ou não incluir num determinado grupo. Essa escolha, mais do que um detalhe técnico, traduz prioridades, recursos e critérios de equidade, e é por isso determinante na ligação entre o modelo e a ação. Embora frequentemente se utilize o valor 0,50 por convenção (por defeito em vários *softwares*), esse ponto de corte raramente é o mais adequado. A escolha deve refletir não apenas critérios estatísticos, mas também os objetivos e os recursos disponíveis.

Neste contexto, várias estratégias são possíveis, entre as quais destacamos:

i) Ponto de corte padrão (0,50): intuitivo e de fácil comunicação, mas pode não maximizar nem a sensibilidade nem a especificidade.

ii) Ponto de corte de Youden: procura o equilíbrio entre sensibilidade e especificidade, maximizando a capacidade discriminativa global.

iii) Ponto de corte pelo critério sensibilidade mínima/desejada**:** para cenários onde a prioridade máxima é detetar (quase) todos os positivos (e.g., doenças de alto contágio/risco), mesmo à custa de mais falsos positivos;

iv) Ponto de corte por capacidade (top-K) - seleciona, de forma explícita, os casos com maior probabilidade até perfazer a fração da população que pode ser atendida com os recursos disponíveis. Este critério alinha a decisão com a capacidade real de resposta, tornando transparente quem é priorizado e porquê, e evitando que o limiar resulte apenas de convenções ou arbitrariedade.

Aplicando esta abordagem ao modelo validado e calibrado, a comparação entre os quatro pontos de corte (incluindo aqui o usual 0,5) mostra padrões em grande parte convergentes, reforçando robustez, mas com diferenças notáveis no top-20%:

i) com o ponto de corte por defeito, obtém-se um compromisso equilibrado, mas não otimizado;
ii) com o critério de Youden, a sensibilidade sobe ligeiramente e a especificidade desce um pouco, mantendo desempenho global semelhante;
iii) com sensibilidade mínima pré-definida, a política torna-se clínico-centrada, assegurando deteção de (quase) todos os casos relevantes;
iv) com a estratégia top-K, maximiza-se a precisão e reduz-se a carga da equipa, à custa de sensibilidade inferior, útil quando a capacidade é a restrição dominante.

Pensando num contexto de enfermagem domiciliária, a escolha do ponto de corte tem consequências diretas na equidade e na afetação de recursos. Se a prioridade é não falhar casos graves (prevenção de feridas complexas), um ponto de corte definido por sensibilidade mínima aumenta o número de visitas preventivas e reduz falsos negativos, aceitando mais falsos positivos. Se a limitação principal é a capacidade semanal, um critério top-K (p. ex., priorizar os X% com maior risco) torna explícito quem é visto primeiro e garante transparência na lista de espera. Em ambos os casos, é essencial registar e fundamentar o ponto de corte escolhido, garantindo que a decisão clínica possa ser verificada e seja equitativa.

Observa-se, ainda, que neste exemplo concreto três dos quatro critérios conduziram a resultados muito semelhantes. À primeira vista, tal pode parecer redundante ou pouco útil. Na realidade, constitui uma confirmação de robustez. Mostra que, neste conjunto de dados, diferentes perspetivas de decisão convergem, reforçando a confiança no modelo. A lição pedagógica é clara: não se trata de procurar pontos de cortes diferentes “a qualquer custo”, mas de documentar e comparar alternativas. Quando coincidem, temos evidência adicional de consistência; quando se verificam diferenças, tornam-se evidentes as concessões cruciais para a tomada de decisão. Em saúde, falsos negativos podem ter custos clínicos elevados; em gestão de recursos, falsos positivos podem ser mais críticos, pelo que o ponto de corte deve equilibrar estes riscos conforme a prioridade.

4.4. Amostras complexas e pesos de calibração: o que muda, porquê e as implicações no apoio à tomada de decisão

A maioria dos inquéritos em saúde, educação e desporto não assenta em amostras aleatórias simples, mas em desenhos amostrais complexos, que combinam estratificação, conglomerados, probabilidades desiguais de seleção e ajustes à não resposta. Para corrigir os desvios que daí resultam, recorremos a pesos: (i) de desenho, quando refletem a probabilidade de seleção de cada unidade; (ii) de não resposta, quando compensam ausências sistemáticas de determinados grupos; e (iii) calibrados, quando ajustam a amostra a marginais populacionais conhecidos, como a distribuição por sexo e idade.

Na prática, cada respondente passa a representar não apenas a si próprio, mas também um número variável de indivíduos da população (o peso estatístico). Ignorar este mecanismo equivale a assumir, de forma implícita, que a amostra é automaticamente representativa, o que conduz a consequências previsíveis: erros-padrão subestimados, mais associações "estatisticamente significativas" e conclusões que não se confirmam à escala populacional.

Neste tópico consideramos o inquérito municipal à atividade física em Évora de 2025, pois permite ilustrar claramente esta diferença. Ajustam-se dois modelos logísticos: um incorporando os pesos calibrados, outro tratando a amostra como se fosse aleatória simples. No primeiro, emergem determinantes estáveis e plausíveis à escala da população, como a idade, sexo, infraestruturas e participação em eventos, com incerteza realista associada às estimativas. No segundo, surgem associações artificiais, reflexo da sobrerrepresentação de alguns grupos de respondentes. Aparentemente, ambos os modelos podem exibir métricas globais semelhantes, como a AUC ou a exatidão, mas na verdade descrevem realidades distintas, pois um retrata o concelho no seu conjunto enquanto o outro avalia o desempenho apenas nos inquiridos.

Esta distinção revela toda a sua importância quando se passa da modelação para a decisão prática. Um ponto de corte ou um critério de priorização top-K baseados no modelo não ponderado produzem listas que refletem apenas a amostra; com ponderação, a mesma regra exprime-se em termos populacionais, transmitindo quem deve ser priorizado no município. As implicações são substanciais. Na afetação de recursos, por exemplo, o modelo ingénuo tende a destacar equipamentos pontuais frequentados por grupos muito presentes no inquérito, sugerindo concentrar investimentos em locais visíveis apenas para os respondentes. O modelo ponderado permite confirmar o papel estrutural das ecopistas e ciclovias como determinantes da prática física à escala concelhia, reorientando as prioridades para a malha infraestrutural. Esta ponderação não elimina por completo o enviesamento potencial da autosseleção, em que pessoas mais ativas tendem a responder mais. Contudo, corrige os desequilíbrios observáveis entre grupos etários e de sexo, aproximando a amostra da composição real da população adulta. Na prática, o modelo ponderado dá voz também a quem respondeu menos, frequentemente os menos ativos, oferecendo uma leitura mais representativa e mais justa da realidade concelhia.

Por fim, fazemos notar que até os indicadores de monitorização são afetados. Séries temporais baseadas apenas na amostra podem sugerir progressos rápidos, quando na

realidade medem apenas quem respondeu; ao incorporar pesos, as metas anuais tornam-se comparáveis entre freguesias e ao longo do tempo, garantindo uma avaliação justa.
Com pesos, a evidência orienta para linhas de ação que sustentam infraestruturas contínuas (ecopistas/ciclovias) como pilares, segmentação por idade e sexo para adequar oferta e comunicação, e políticas de literacia de equipamentos para reduzir barreiras de acesso. Assim, os pesos transformam o inquérito numa ferramenta de planeamento justa, representativa e sustentável.
Do ponto de vista pedagógico, esta comparação é uma lição de inferência aplicada: a regressão logística não se esgota em ajustar coeficientes; exige respeitar o alvo da decisão. Ignorar pesos é decidir para quem respondeu; incorporar pesos é decidir para quem vive no concelho. É isso que confere equidade e legitimidade às políticas.

4.5. O que a perspetiva de *machine learning* pode acrescentar ao apoio à tomada de decisão

A regressão logística continua a ser a técnica de referência na análise de respostas binárias, pela sua transparência, interpretabilidade e capacidade de quantificar efeitos em razões de chances, tornando-se insubstituível sempre que a prioridade é comunicar resultados a decisores ou fundamentar políticas. No entanto, os avanços em *machine learning* (ML) permitem ampliar esta perspetiva, captando padrões não lineares, interações implícitas e efeitos limiares que a especificação paramétrica dificilmente revela de forma completa. Em complemento à regressão logística os modelos de Random Forest (RF) e Extreme Gradient Boosting (XGBoost) oferecem uma perspetiva adicional: captam não linearidades e interações de alta ordem sem necessidade de as pré-especificar, e permitem "abrir" o modelo com decomposições Shapley Additive Explanations (SHAP).
Recorrendo novamente ao inquérito municipal à atividade física em Évora de 2025, mostra-se que três modelos (logística, RF e XGB) convergem em pontos estruturais: o papel central das infraestruturas abertas (ecopistas, ciclovias), a relevância de hábitos de lazer ativos/socializadores (como passear). Em contraste, comportamentos de lazer marcadamente sedentários surgem associados a menor probabilidade de prática. As explicações SHAP tornam tangível esta leitura, mostrando, caso a caso, como cada fator impulsiona a probabilidade para cima ou para baixo, e evidenciando com nitidez a força infraestrutural das ecopistas e ciclovias, a modulação da idade e os padrões de tempos livres ativos.
O modelo RF privilegia a sensibilidade, que neste caso se traduz em menos falhas na deteção de praticantes, enquanto o modelo XGB privilegia a precisão, que neste caso se traduz em menos esforços desperdiçados em perfis não praticantes. Esta diferença não é um detalhe técnico. Traduz-se em escolhas de política: abrangência quando a prioridade é mobilizar e não deixar praticantes de fora; racionalização quando o foco é concentrar a intervenção onde a probabilidade de prática é realmente elevada. A regressão logística, por seu turno, confirma a direção dos efeitos, fornece intervalos de confiança e clarifica o que é estrutural face ao que é circunstancial. No caso real analisado, permite validar que

infraestruturas abertas e hábitos de lazer ativos ou socializadores estão associados a maior probabilidade de prática; que a idade reduz progressivamente essa probabilidade, enquanto o sexo masculino continua associado a níveis mais elevados; e que eventos-âncora funcionam como reforço, sobretudo em públicos já predispostos.

Os modelos de *machine learning*, em contraste, não especificam interações nem formas funcionais a priori, mas aprendem-nas automaticamente a partir dos dados. Isto permite detetar combinações de variáveis e efeitos limiares que a modelação paramétrica dificilmente captaria com a mesma agilidade. Assim, os algoritmos de RF e XGBoost complementam a regressão logística: enriquecem a leitura ao revelar padrões complexos de associação, enquanto a logística fornece a estrutura inferencial necessária para comunicar resultados, justificar investimentos e traduzir evidência em recomendações políticas ou clínicas.

Consolida-se com as implicações práticas que são diretas neste caso. Para a gestão municipal, a leitura integrada aponta em consolidar ecopistas/ciclovias como espinha dorsal da mobilidade ativa, reforçar circuitos e ligações seguras nos bairros com menor prática, usar eventos-âncora como portas de entrada para hábitos regulares (grupos de caminhada/corrida) e segmentar por idade e género a oferta e os horários. Em paralelo, as explicações SHAP facilitam a prestação de contas e a transparência, pois tornam verificável por que razão um indivíduo é priorizado, algo particularmente útil quando se aplicam pontos de corte (sensibilidade mínima) ou critérios por capacidade (top-K).

Aproveita-se este caso para sublinhar a coerência com o que foi observado noutros momentos da lição onde se considerou o inquérito municipal à atividade Física de 2012. As diferenças aparentes refletem contextos que se alteraram, como oferta de infraestruturas, perfis demográficos e padrões de lazer, não uma contradição metodológica. Em 2025, o papel das infraestruturas consolida-se, não sendo apenas um fator de contexto, mas um determinante transversal do comportamento ativo, captado pelos modelos de ML e confirmado pela regressão logística como efeito consistente e estatisticamente robusto.

Ao longo desta lição percorremos um fio lógico: das interações, que revelam efeitos condicionais, à validação e calibração, que asseguram probabilidades fiáveis; da escolha criteriosa do ponto de corte, que traduz modelos em decisões operacionais, à incorporação de amostras complexas, que garantem legitimidade populacional; culminando no diálogo com modelos de ML, que enriquecem a perspetiva preditiva e ampliam a explicabilidade.

Encerrando esta lição, fica clara a importância de revisitar a regressão logística para além da sua formulação elementar. Os tópicos abordados são etapas que, quando ignoradas, comprometem não só a qualidade estatística do modelo, mas sobretudo a relevância das decisões dele decorrentes. Ao articular exemplos concretos de saúde pública e de atividade física, procura-se evidenciar que a estatística aplicada não é um exercício abstrato, mas uma ferramenta concreta de apoio a políticas e práticas com impacto direto na comunidade.

Este percurso também mostra, com dados municipais de Évora e estudos de prevalência de feridas do Alentejo, como a regressão logística, enriquecida por técnicas de validação,

calibração, definição explícita de pontos de corte e integração com ML, se converte numa ferramenta de apoio à decisão, capaz de orientar recursos de forma justa e eficaz.

Naturalmente, esta não é uma abordagem exaustiva. Questões como o desbalanceamento severo da variável resposta, os desafios das amostras pequenas, a visualização aprofundada dos efeitos ou o ajustamento de modelos logísticos em contextos de *big data* constituem temas igualmente centrais e desafiantes. As pontes entre a regressão logística e os modelos de ML, já exploradas neste sumário, constituem hoje um campo de investigação particularmente ativo, sobretudo no domínio da explicabilidade de modelos complexos. O desenvolvimento de ferramentas como as aplicadas em estudos de sinistralidade rodoviária mostra que o rigor estatístico da logística e o desempenho preditivo dos ML são complementares, e não mutuamente exclusivos.

Estas temáticas exigem tempo próprio para serem exploradas. Ficam, assim, como pontos de continuidade, para futuras lições e seminários de investigação.

5. Bibliografia Recomendada

Agresti, A. (2013). *Categorical data analysis* (3rd ed.). Wiley.

Bauman, A. E., Reis, R. S., Sallis, J. F., Wells, J. C., Loos, R. J. F., & Martin, B. W. (2012). Correlates of physical activity: Why are some people physically active and others not? *The Lancet, 380*(9838), 258–271. https://doi.org/10.1016/S0140-6736(12)60735-1

Carreiras, A., Infante, P., & Afonso, A. (2019). Avaliação do efeito do desenho de amostragem em modelos de regressão logística. In H. Bacelar Nicolau, F. Sousa, C. Marcelo, A. S. Ferreira, P. Infante, & A. Figueiredo (Eds.), *Classificação e análise de dados: Métodos e aplicações III (CLADMAp III)* (pp. 77–86). INE. http://hdl.handle.net/10174/25648

Dobson, A. J., & Barnett, A. (2018). *An introduction to generalized linear models* (4th ed.). CRC Press.

Heeringa, S. G., West, B. T., & Berglund, P. A. (2017). *Applied survey data analysis* (2nd ed.). Chapman & Hall/CRC.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). Wiley.

Infante, P., Jacinto, G., Afonso, A., Rego, L., Nogueira, V., Quaresma, P., Saias, J., Santos, D., Nogueira, P., Silva, M., Pisco Costa, R., Góis, P., & Manuel, P. R. (2022). Comparison of statistical and machine-learning models on road traffic accident severity classification. *Computers, 11*(5), 80. https://doi.org/10.3390/computers11050080

Infante, P., Jacinto, G., Santos, D., Nogueira, P. M., Afonso, A., Quaresma, P., Silva, M. G., Nogueira, V. B., Rego, L., & Saias, J. (2023). Prediction of Road Traffic Accidents on a Road in Portugal: A Multidisciplinary Approach Using Artificial Intelligence, Statistics, and Geographic Information Systems. *Information*, 14 (4). https://doi.org/10.3390/info14040238

Lumley, T. (2010). *Complex surveys: A guide to analysis using R*. Wiley.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in neural information processing systems* (Vol. 30, pp. 4765–4774). Curran Associates.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). Chapman & Hall.

Molnar, C. (2022). *Interpretable machine learning: A guide for making black box models explainable* (2nd ed.). Independently published.

Pepe, M. S. (2003). *The statistical evaluation of medical tests for classification and prediction*. Oxford University Press.

Steyerberg, E. W. (2019). *Clinical prediction models: A practical approach to development, validation, and updating* (2nd ed.). Springer. https://doi.org/10.1007/978-3-030-16399-0

Szklo, M., & Nieto, F. J. (2019). *Epidemiology: Beyond the basics* (4th ed.). Jones & Bartlett Learning.

Turkman, M. A. A., & Silva, G. L. (2000). *Modelos lineares generalizados*. Edições SPE.

Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., & Steyerberg, E. W. (2019). Calibration: The Achilles heel of predictive analytics. *BMC Medicine, 17*, 230. https://doi.org/10.1186/s12916-019-1466-7

Zweig, M. H., & Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. *Clinical Chemistry, 39*(4), 561–577. https://doi.org/10.1093/clinchem/39.4.561