**Instruções:** Responda às questões de múltipla escolha assinalando apenas uma alternativa. Nas questões dissertativas, apresente o raciocínio passo a passo, incluindo fórmulas e cálculos intermediários quando solicitado. Mantenha a notação usada em aula: $y \in \{0, 1\}$ é o resultado real (1 = resultado positivo), $\hat{y} \in \{0, 1\}$ é a predição do modelo (1 = decisão favorável), $a \in \{0, 1\}$ é o atributo protegido (dois grupos) e $p_a = P(y{=}1 \mid a)$ é a taxa-base do grupo $a$. As taxas $TPR$, $FPR$ e $PPV$ são as definidas a partir da matriz de confusão por grupo. Use a vírgula como separador decimal.

**Questões de Múltipla Escolha.**

I) Um banco treina um classificador de concessão de crédito usando como rótulo de "bom pagador" o histórico de aprovações dos últimos vinte anos da própria instituição, período em que certos bairros raramente recebiam crédito. O modelo passa a negar crédito a moradores desses bairros mesmo quando suas finanças atuais são sólidas. Qual fonte de viés melhor descreve a origem do problema?

(a) Viés histórico, pois o rótulo reflete decisões passadas já enviesadas e não a capacidade real de pagamento.

(b) Viés de amostragem, pois o conjunto de treino contém menos exemplos do grupo afetado do que da maioria.

(c) Viés de medição, pois a renda foi registrada com um instrumento diferente para cada bairro.

(d) Viés de agregação, pois um único modelo foi ajustado a subpopulações com relações distintas entre atributos e rótulo.

II) Um sistema de triagem de currículos remove explicitamente o campo de gênero do vetor de entrada e, mesmo assim, continua aprovando homens em proporção muito maior. Investigando, descobre-se que o modelo usa fortemente o nome de clubes esportivos e de cursos extracurriculares listados pelos candidatos. Qual afirmação descreve corretamente a situação?

(a) A disparidade persiste porque os atributos restantes atuam como *proxies* do gênero, carregando a informação removida.

(b) A disparidade é impossível depois de remover o atributo, então a auditoria deve ter algum erro de medição.

(c) A remoção do atributo garante *fairness* individual, mas não *fairness* de grupo, que exige rebalanceamento.

(d) O modelo está calibrado por gênero, logo a diferença observada é estatisticamente esperada e aceitável.

III) Para o grupo $a$=0 obteve-se a matriz de confusão $TN = 120$, $FP = 80$, $FN = 30$, $TP = 70$ (linhas = resultado real, colunas = predição, positivo = decisão favorável). Qual é a taxa de verdadeiros positivos ($TPR$) desse grupo?

(a) 0,70

(b) 0,40

(c) 0,54

(d) 0,47

IV) Em uma plataforma de anúncios de vagas, 30% dos usuários do grupo $a$=0 e 30% dos usuários do grupo $a$=1 recebem o anúncio de uma vaga de tecnologia. Sabe-se ainda que a fração de pessoas realmente interessadas difere entre os grupos. Qual critério de *fairness* de grupo é satisfeito pelos números informados?

(a) Paridade demográfica, pois a taxa de decisões favoráveis é igual nos dois grupos.

(b) Igualdade de oportunidade, pois a taxa de verdadeiros positivos é igual nos dois grupos.

(c) Calibração por grupo, pois o valor preditivo positivo é igual nos dois grupos.

(d) Igualdade de chances, pois tanto $TPR$ quanto $FPR$ são iguais nos dois grupos.

V) Considere os dois grupos com as matrizes de confusão abaixo (positivo = decisão favorável).

| **Grupo** $a$=0 | $\hat{y}$=0 | $\hat{y}$=1 |
|---|---|---|
| $y$=0 | 150 | 50 |
| $y$=1 | 40 | 160 |

| **Grupo** $a$=1 | $\hat{y}$=0 | $\hat{y}$=1 |
|---|---|---|
| $y$=0 | 75 | 25 |
| $y$=1 | 30 | 70 |

Qual é o módulo da diferença entre os $TPR$ dos dois grupos (a lacuna de igualdade de oportunidade)?

(a) 0,10

(b) 0,05

(c) 0,25

(d) 0,15

VI) Usando as mesmas matrizes da questão anterior, deseja-se verificar se o critério de igualdade de chances (*equalized odds*) está satisfeito. Qual conclusão é correta?

(a) O critério é violado, pois embora os $FPR$ coincidam em 0,25, os $TPR$ diferem (0,80 contra 0,70).

(b) O critério é satisfeito, pois tanto os $TPR$ quanto os $FPR$ coincidem entre os grupos.

(c) O critério é violado, pois os $FPR$ diferem (0,25 contra 0,40), embora os $TPR$ coincidam.

(d) O critério é satisfeito, pois a taxa de decisões favoráveis é igual nos dois grupos.

VII) Um modelo de risco de reincidência tem, para o grupo $a$=1, $FPR = 0{,}45$ e $FNR = 0{,}28$. Sabendo que $TPR = 1 - FNR$, qual é o $TPR$ desse grupo?

(a) 0,72

(b) 0,55

(c) 0,45

(d) 0,28

VIII) Em um sistema de score de crédito, entre as pessoas do grupo $a$=0 classificadas como "baixo risco" ($\hat{y}$=1), 62% de fato pagaram o empréstimo; no grupo $a$=1, essa fração é de 61%. Já a fração de bons pagadores incorretamente marcados como alto risco difere bastante entre os grupos. Que critério está aproximadamente satisfeito e qual está em questão?

(a) Está satisfeita a calibração por grupo (igualdade de $PPV$); o que difere é a igualdade de $FNR$.

(b) Está satisfeita a igualdade de oportunidade (igualdade de $TPR$); o que difere é a calibração.

(c) Está satisfeita a paridade demográfica; o que difere é o valor preditivo positivo.

(d) Está satisfeita a igualdade de $FPR$; o que difere é o valor preditivo positivo.

IX) ☠️ Dois grupos têm taxas-base distintas, $p_0 = 0{,}30$ e $p_1 = 0{,}50$, e um classificador imperfeito é aplicado a ambos. A equipe ajusta o modelo para igualar o $PPV$ e o $FNR$ entre os grupos. Usando a identidade $FPR = \frac{p}{1-p} \cdot \frac{1-PPV}{PPV} \cdot (1 - FNR)$, o que necessariamente ocorre com os $FPR$?

(a) Os $FPR$ ficam diferentes, pois o fator $\frac{p}{1-p}$ cresce com $p$ e os demais fatores são iguais entre os grupos.

(b) Os $FPR$ ficam iguais, pois fixar $PPV$ e $FNR$ determina completamente o $FPR$ de cada grupo.

(c) Os $FPR$ ficam iguais, pois a identidade não depende da taxa-base quando o classificador é imperfeito.

(d) Os $FPR$ ficam diferentes apenas se o classificador for perfeito; com erro, eles coincidem.

X) Uma equipe recebe um modelo de linguagem de terceiros via API, sem acesso aos pesos nem ao pipeline de treino, e precisa reduzir associações tóxicas com um grupo religioso já na semana seguinte. Qual família de métodos de *debiasing* é a mais adequada à restrição?

(a) Métodos inferenciais, que atuam na inferência sem alterar parâmetros e funcionam sobre modelos de terceiros.

(b) Métodos distribucionais, que rebalanceiam o corpus e exigem retreinar o modelo do zero.

(c) Métodos de *one-step-training*, que adicionam um termo de regularização ao objetivo durante o treino.

(d) Métodos de *two-step-training*, que ajustam o modelo com uma segunda etapa de *fine-tuning*.

XI) ☠️ Uma empresa quer corrigir o viés de um modelo de fundação de bilhões de parâmetros que serve dezenas de aplicações. Retreinar do zero é proibitivo, e correções puramente na inferência têm se mostrado reversíveis sob *prompts* adversariais. Qual escolha de método é mais coerente com essas duas restrições simultâneas?

(a) *Two-step-training* via *fine-tuning* ou adapters sobre o modelo já treinado.

(b) Métodos distribucionais, rebalanceando todo o corpus de pré-treino.

(c) *One-step-training* com objetivo adversarial desde o início do pré-treino.

(d) Edição do espaço vetorial aplicada apenas no momento da inferência.

XII) Dois candidatos a crédito têm exatamente o mesmo histórico financeiro relevante e diferem apenas no atributo protegido $a$. O modelo aprova um e nega o outro. Considerando as duas famílias de critérios vistas em aula, qual noção de *fairness* é diretamente violada por esse caso isolado?

(a) *Fairness* individual, que exige tratamento similar para indivíduos similares na tarefa.

(b) Paridade demográfica, que exige taxas de aprovação iguais entre os grupos.

(c) Calibração por grupo, que exige que o escore tenha o mesmo significado nos grupos.

(d) Igualdade de oportunidade, que exige $TPR$ igual entre os grupos.

**Questões Dissertativas.**

I) (**Auditoria por grupo**) Um modelo de triagem para uma vaga foi avaliado em dois grupos. As matrizes de confusão (positivo = aprovado na triagem; linhas = resultado real, colunas = predição) são:

| **Grupo** $a$=0 | $\hat{y}$=0 | $\hat{y}$=1 |
|---|---|---|
| $y$=0 | 270 | 30 |
| $y$=1 | 40 | 160 |

| **Grupo** $a$=1 | $\hat{y}$=0 | $\hat{y}$=1 |
|---|---|---|
| $y$=0 | 240 | 60 |
| $y$=1 | 50 | 50 |

(a) Calcule, para cada grupo, a taxa de decisões favoráveis $P(\hat{y}{=}1 \mid a)$, o $TPR$ e o $FPR$. Mostre as contas.

(b) Decida se a paridade demográfica e a igualdade de oportunidade estão satisfeitas. Justifique com os números.

(c) A taxa-base $p_a = P(y{=}1 \mid a)$ é igual nos dois grupos? Explique como a diferença de taxas-base ajuda a entender por que a paridade demográfica e a igualdade de oportunidade podem discordar.

II) (**O teorema da impossibilidade**) Considere dois grupos com taxas-base $p_0 = 0{,}30$ e $p_1 = 0{,}60$ e um classificador imperfeito. Use a identidade

$$FPR = \frac{p}{1-p} \cdot \frac{1 - PPV}{PPV} \cdot (1 - FNR).$$

(a) Suponha que o modelo foi ajustado para ter $PPV = 0{,}60$ e $FNR = 0{,}30$ em **ambos** os grupos. Calcule o $FPR$ resultante de cada grupo e mostre que eles diferem.

(b) Explique, a partir do formato da identidade, por que é impossível satisfazer simultaneamente calibração (igualdade de $PPV$), igualdade de $FNR$ e igualdade de $FPR$ quando $p_0 \neq p_1$ e o classificador é imperfeito.

(c) Relacione esse resultado com a controvérsia COMPAS entre ProPublica e Northpointe: que critério cada lado mediu e por que ambos podiam estar "certos" ao mesmo tempo.

III) (**Escolha de critério e mitigação**) Para cada cenário abaixo, indique qual critério de *fairness* de grupo é o mais apropriado e justifique com base em quem paga o custo do erro mais grave. Em seguida, escolha uma família de métodos de *debiasing* (distribucional, *one-step*, *two-step* ou inferencial) e justifique a escolha pelas restrições de cada cenário.

(a) Triagem médica de uma doença grave porém tratável, em que deixar de detectar a doença é o desfecho mais danoso, com acesso total ao pipeline de treino e a dados clínicos rotulados.

(b) Distribuição de anúncios de vagas de emprego, em que o objetivo declarado é ampliar o acesso de grupos sub-representados às oportunidades, sobre um modelo de recomendação proprietário acessível apenas por API.

IV) ☠️ (**Limiar por grupo**) Um sistema usa um escore contínuo $s \in [0, 1]$ e decide $\hat{y}{=}1$ quando $s \geq \tau$. Com um limiar único $\tau = 0{,}5$ aplicado aos dois grupos, mediu-se $TPR_0 = 0{,}90$, $TPR_1 = 0{,}70$, ou seja, uma lacuna de igualdade de oportunidade de 0,20.

(a) Explique qualitativamente por que **baixar** o limiar $\tau_1$ aplicado ao grupo $a$=1 tende a aumentar o $TPR_1$ e, portanto, a reduzir a lacuna. O que acontece com o $FPR_1$ ao baixar esse limiar?

(b) Esse ajuste de limiar por grupo é um método de *debiasing* de qual família da taxonomia da aula? Cite uma vantagem e uma limitação dessa família.

(c) Suponha que, após igualar o $TPR$, os $FPR$ dos dois grupos passem a diferir. Conecte essa observação ao teorema da impossibilidade e explique por que não é contraditório "consertar" um critério e "quebrar" outro.

## Gabarito

### Múltipla Escolha.

I) **(a)** O problema está no rótulo: "bom pagador" foi definido a partir de decisões passadas já enviesadas (poucos bairros recebiam crédito), então o alvo aprendido não mede a capacidade real de pagamento, e sim o padrão histórico. Isso é viés histórico. Não é amostragem (não se trata de número de exemplos por grupo), nem medição (a renda não foi medida com instrumento distinto), nem agregação (não há um único modelo forçado sobre subpopulações com relações distintas; o defeito está na variável-alvo).

II) **(a)** Remover o atributo protegido (*fairness through unawareness*) não garante *fairness*: atributos correlacionados (clubes, cursos) funcionam como *proxies* e reconstroem a informação removida. A alternativa (b) está errada porque a disparidade é justamente esperada via *proxies*; (c) inverte o que a remoção garante, pois ela não garante nenhuma das duas noções; (d) calibração é outro critério e nada nos números informados a sustenta.

III) **(a)** $TPR = \frac{TP}{TP+FN} = \frac{70}{70+30} = \frac{70}{100} = 0{,}70$. A alternativa (d) traz $\frac{TP}{TP+FP} = \frac{70}{150} \approx 0{,}47$, que é o $PPV$, fácil de confundir com o $TPR$; (b) é $\frac{TP}{P_{\text{previstos}}}$ aproximado e (c) corresponde a outra razão da matriz, ambos cálculos com o denominador errado.

IV) **(a)** A paridade demográfica exige $P(\hat{y}{=}1 \mid a{=}0) = P(\hat{y}{=}1 \mid a{=}1)$. Como ambos os grupos recebem o anúncio a uma taxa de 30%, a paridade está satisfeita. As demais alternativas exigiriam condicionar em $y$ (oportunidade, chances) ou olhar $P(y{=}1 \mid \hat{y})$ (calibração); os dados informados não dizem nada sobre $y$, então não há como afirmar que esses critérios valem.

V) **(a)** $TPR_0 = \frac{160}{160+40} = \frac{160}{200} = 0{,}80$ e $TPR_1 = \frac{70}{70+30} = \frac{70}{100} = 0{,}70$. A lacuna é $|0{,}80 - 0{,}70| = 0{,}10$. O valor 0,25 em (c) é o $FPR$ comum aos dois grupos, medido no lugar do $TPR$; (b) e (d) resultam de usar as células erradas da matriz.

VI) **(a)** A igualdade de chances exige $TPR$ **e** $FPR$ iguais. Os $FPR$ coincidem: $FPR_0 = \frac{50}{50+150} = 0{,}25$ e $FPR_1 = \frac{25}{25+75} = 0{,}25$. Mas os $TPR$ diferem: 0,80 contra 0,70. Como basta uma das duas igualdades falhar, o critério está violado. A alternativa (b) afirma falsamente que as duas taxas coincidem; (c) inverte qual taxa difere; (d) confunde *equalized odds* com paridade demográfica.

VII) **(a)** $TPR = 1 - FNR = 1 - 0{,}28 = 0{,}72$. A alternativa (c) repete o $FPR$, e (d) repete o $FNR$, ambos confundindo qual taxa foi pedida.

VIII) **(a)** A igualdade de $PPV$ entre grupos é exatamente a versão binarizada da calibração ($P(y{=}1 \mid \hat{y}{=}1)$), aqui 62% contra 61%, aproximadamente satisfeita. A fração de bons pagadores marcados como alto risco que difere é o $FNR$ (entre $y{=}1$, os que recebem $\hat{y}{=}0$). As demais alternativas trocam o critério satisfeito ($TPR$, paridade, $FPR$), que não corresponde a $P(y{=}1 \mid \hat{y}{=}1)$.

IX) **(a)** Com $PPV$ e $FNR$ iguais nos dois grupos, na identidade só varia o fator $\frac{p}{1-p}$, que cresce com $p$. Como $p_1 > p_0$, o grupo $a{=}1$ tem $\frac{p}{1-p}$ maior e, portanto, $FPR$ maior; os $FPR$ não podem coincidir. (b) e (c) ignoram a dependência em $p$; (d) inverte o papel do classificador perfeito, que é justamente a exceção em que o teorema não se aplica.

X) **(a)** Sem acesso aos pesos nem ao treino e com prazo curto, só os métodos inferenciais se aplicam: atuam na inferência (*prompting*, edição de saída) sem alterar parâmetros e servem modelos de terceiros via API. Distribucionais (b) e *one-step* (c) exigem retreinar; *two-step* (d) exige acesso aos pesos para a segunda etapa de treino, indisponível aqui.

XI) **(a)** O *two-step-training* parte do modelo já treinado e adiciona uma etapa barata (*fine-tuning*, adapters), viável em foundation models (mesma economia do PEFT) e mais persistente que correções só na inferência. Distribucional (b) e *one-step* (c) implicam retreinar do zero, proibitivo na escala dada; a edição na inferência (d) é justamente o tipo de correção reversível que o enunciado quer evitar.

XII) **(a)** Dois indivíduos idênticos na tarefa recebem decisões opostas: isso viola diretamente a *fairness* individual (tratamento similar para similares). As noções de grupo (b, c, d) são estatísticas agregadas e podem estar satisfeitas mesmo com esse caso individual injusto, pois não olham pares de indivíduos.

**Dissertativas.**

I) (a) Grupo $a$=0 (total = 270 + 30 + 40 + 160 = 500): taxa favorável $P(\hat{y}{=}1 \mid a{=}0) = \frac{30+160}{500} = \frac{190}{500} = 0{,}38$; $TPR_0 = \frac{160}{160+40} = \frac{160}{200} = 0{,}80$; $FPR_0 = \frac{30}{30+270} = \frac{30}{300} = 0{,}10$. Grupo $a$=1 (total = 240 + 60 + 50 + 50 = 400): taxa favorável $P(\hat{y}{=}1 \mid a{=}1) = \frac{60+50}{400} = \frac{110}{400} = 0{,}275$; $TPR_1 = \frac{50}{50+50} = \frac{50}{100} = 0{,}50$; $FPR_1 = \frac{60}{60+240} = \frac{60}{300} = 0{,}20$.

(b) Paridade demográfica: exige taxas favoráveis iguais; 0,38 ≠ 0,275, logo **violada**. Igualdade de oportunidade: exige $TPR$ iguais; 0,80 ≠ 0,50 (lacuna de 0,30), logo **violada**.

(c) As taxas-base diferem: $p_0 = \frac{160+40}{500} = \frac{200}{500} = 0{,}40$ e $p_1 = \frac{50+50}{400} = \frac{100}{400} = 0{,}25$. Como a paridade demográfica olha $P(\hat{y}{=}1 \mid a)$ sem condicionar em $y$, ela mistura a taxa-base com o desempenho do modelo; já a igualdade de oportunidade condiciona em $y$=1. Com taxas-base diferentes, igualar uma das taxas em geral não iguala a outra, então os dois critérios podem (e costumam) discordar.

II) (a) Fator comum aos dois grupos: $\frac{1-PPV}{PPV} \cdot (1 - FNR) = \frac{1-0{,}60}{0{,}60} \cdot (1 - 0{,}30) = \frac{0{,}40}{0{,}60} \cdot 0{,}70 = 0{,}6667 \cdot 0{,}70 \approx 0{,}4667$. Grupo 0: $\frac{p_0}{1-p_0} = \frac{0{,}30}{0{,}70} \approx 0{,}4286$, então $FPR_0 \approx 0{,}4286 \cdot 0{,}4667 \approx 0{,}20$. Grupo 1: $\frac{p_1}{1-p_1} = \frac{0{,}60}{0{,}40} = 1{,}5$, então $FPR_1 \approx 1{,}5 \cdot 0{,}4667 \approx 0{,}70$. Os $FPR$ diferem fortemente (0,20 contra 0,70).

(b) Na identidade, $FPR$ é o produto de três fatores. Igualar $PPV$ e $FNR$ entre os grupos torna os dois últimos fatores idênticos, mas o primeiro, $\frac{p}{1-p}$, é estritamente crescente em $p$. Com $p_0 \neq p_1$, esse fator difere e arrasta o $FPR$ para valores distintos. Logo, fixar calibração e igualdade de $FNR$ **força** $FPR$ desiguais; não há classificador imperfeito que satisfaça os três ao mesmo tempo sob taxas-base diferentes.

(c) A Northpointe mediu calibração (igualdade de $PPV$, o escore significa o mesmo nos grupos) e a considerou aproximadamente satisfeita; a ProPublica mediu as taxas de erro (igualdade de $FPR$ e de $FNR$) e mostrou que estavam desbalanceadas. Pelo teorema, com as taxas-base de reincidência diferentes entre os grupos, nenhum classificador imperfeito poderia satisfazer calibração e igualdade de $FPR/FNR$ simultaneamente. Ambos os lados calcularam corretamente métricas legítimas e incompatíveis; a disputa era normativa, sobre qual noção de justiça deveria prevalecer.

III) (a) Critério: igualdade de oportunidade (ou igualdade de $FNR$). O erro mais grave é o falso negativo (deixar de detectar a doença), e quem paga é o paciente doente não identificado; igualar o $TPR$ entre grupos garante que doentes de qualquer grupo tenham a mesma chance de detecção. Família de *debiasing*: como há acesso total ao pipeline de treino e a rótulos clínicos, um método de *one-step-training* é adequado (por exemplo, um termo de regularização $\mathcal{L} = \mathcal{L}_{\text{tarefa}} + \lambda\, \mathcal{L}_{\text{fair}}$ que penaliza a disparidade de $TPR$), pois oferece controle

fino do trade-off via $\lambda$ e o custo de um treino completo é aceitável. Um método distribucional (rebalancear o conjunto clínico) também é defensável, já que ataca a causa nos dados.

(b) Critério: paridade demográfica. O objetivo declarado é redistribuir acesso a oportunidades, e o "erro" relevante é a exclusão de acesso de grupos sub-representados; igualar a taxa de exposição ao anúncio entre grupos atende a esse objetivo, e a paridade tem a vantagem de não exigir o rótulo verdadeiro. Família de *debiasing*: como o modelo é proprietário e só acessível por API (sem pesos nem treino), resta um método inferencial, por exemplo ajustar por pós-processamento a taxa de exibição por grupo. Vantagem: barato e aplicável a modelo de terceiros; limitação: a correção é superficial e não toca o viés latente do modelo.

IV) (a) Baixar $\tau_1$ faz com que mais indivíduos do grupo $a$=1 ultrapassem o limiar e recebam $\hat{y}$=1. Entre os verdadeiramente positivos ($y$=1), mais passam a ser corretamente aprovados, então o $TPR_1$ sobe e a lacuna em relação a $TPR_0$ diminui. O custo é que também mais negativos ($y$=0) ultrapassam o limiar, então o $FPR_1$ **aumenta** junto: o ajuste compra $TPR$ pagando em $FPR$.

(b) É um método inferencial (pós-processamento clássico de Hardt et al., ajuste de limiar por grupo), pois não altera nenhum parâmetro do modelo. Vantagem: barato, reversível e aplicável mesmo sem acesso aos pesos. Limitação: correção superficial, o viés permanece nos pesos/no escore e pode reaparecer; além disso, usar limiares diferentes por grupo é em si uma decisão normativa que precisa ser justificada.

(c) Igualar o $TPR$ entre grupos e observar que os $FPR$ passam a diferir é exatamente o que o teorema da impossibilidade prevê: com taxas-base distintas, não dá para igualar todos os critérios ao mesmo tempo. Consertar a igualdade de oportunidade (uma componente do *equalized odds*) força uma das outras quantidades a divergir. Não há contradição: a *fairness* não é booleana, é um trade-off, então melhorar um critério às custas de outro é o comportamento esperado, e a escolha de qual igualar deve ser decidida e documentada.