ACH2016 - Inteligência Artificial
Aula 08 - Desempenho e Viés

Valdinei Freire da Silva

valdinei.freire@usp.br - Bloco A1 100-O

Russell e Norvig, Capítulo 18
Artigos sobre Viés

Tarefa de Aprendizado Supervisionado

Dado um conjunto de treinamento com N exemplos de pares entrada-saída

$$(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N),$$

onde cada y_i foi gerado por uma função f desconhecida, isto é, $y_i = f(x_i)$.

Descubra uma função h que aproxima a verdadeira função f.

x é a entrada e y é a saída.

x e y pode ser qualquer valor, números ou categorias, x usualmente é um vetor de valores (atributos).

Melhor Hipótese

Genericamente pode-se pensar em uma função de perda:

$$L(h, \mathcal{E})$$

que avalia a qualidade da hipótese h aplicada na população $\mathcal{E}$.

A hipótese ótima é dada por:

$$h^* = \arg\min_{h \in \mathcal{H}} L(h, \mathcal{E}).$$

Empiricamente, para um conjunto de exemplos E, temos:

$$\hat{h}^* = \arg\min_{h \in \mathcal{H}} L(h, E).$$

Exemplo de função de perda (*Loss Function*):

- Acurácia: taxa de exemplos que são classificados corretamente.
- Verosimilhança: apenas para hipóteses probabilísticas.

Função de Perda Proxy

Objetivo de **inferência probabilística**: escolher o modelo probabilístico que melhor explica a população (ou dados observados).

Objetivo da **classificação**: escolher o modelo que melhor classifica.

Verosimilhança: mensuração probabilística que privilegia modelos que melhor explica a população.

Acurácia: mensuração que privilegia modelos que melhor classifica.

Considere um modelo probabilístico binomial h', então um modelo de classificação binária h pode ser construído da seguinte forma:

$$h(x) = \begin{cases} 0, & \text{se } h'(x) < T \\ 1, & \text{caso contrário} \end{cases}$$

T é um limiar (*threshold*) arbitrário.

Matriz de Confusão[1]

		Predição do Modelo	
		Negativo	**Positivo**
Valores Corretos	**Negativo** (N)	Acerto (TN)	Erro (FP)
	Positivo (P)	Erro (FN)	Acerto (TP)

$$accuracy = \frac{TP + TN}{TP + FN + FP + TN} = \frac{TP + TN}{P + N} = \Pr(Y = \hat{Y})$$

$$recall = \frac{TP}{TP + FN} = \frac{TP}{P} = \Pr(\hat{Y} = 1 | Y = 1)$$

$$precision = \frac{TP}{TP + FP} = \Pr(Y = 1 | \hat{Y} = 1)$$

[1] TP = true positive; FP = false positive; FN = false negative; TN = true negative; P = positivos na população; N = negativos na população

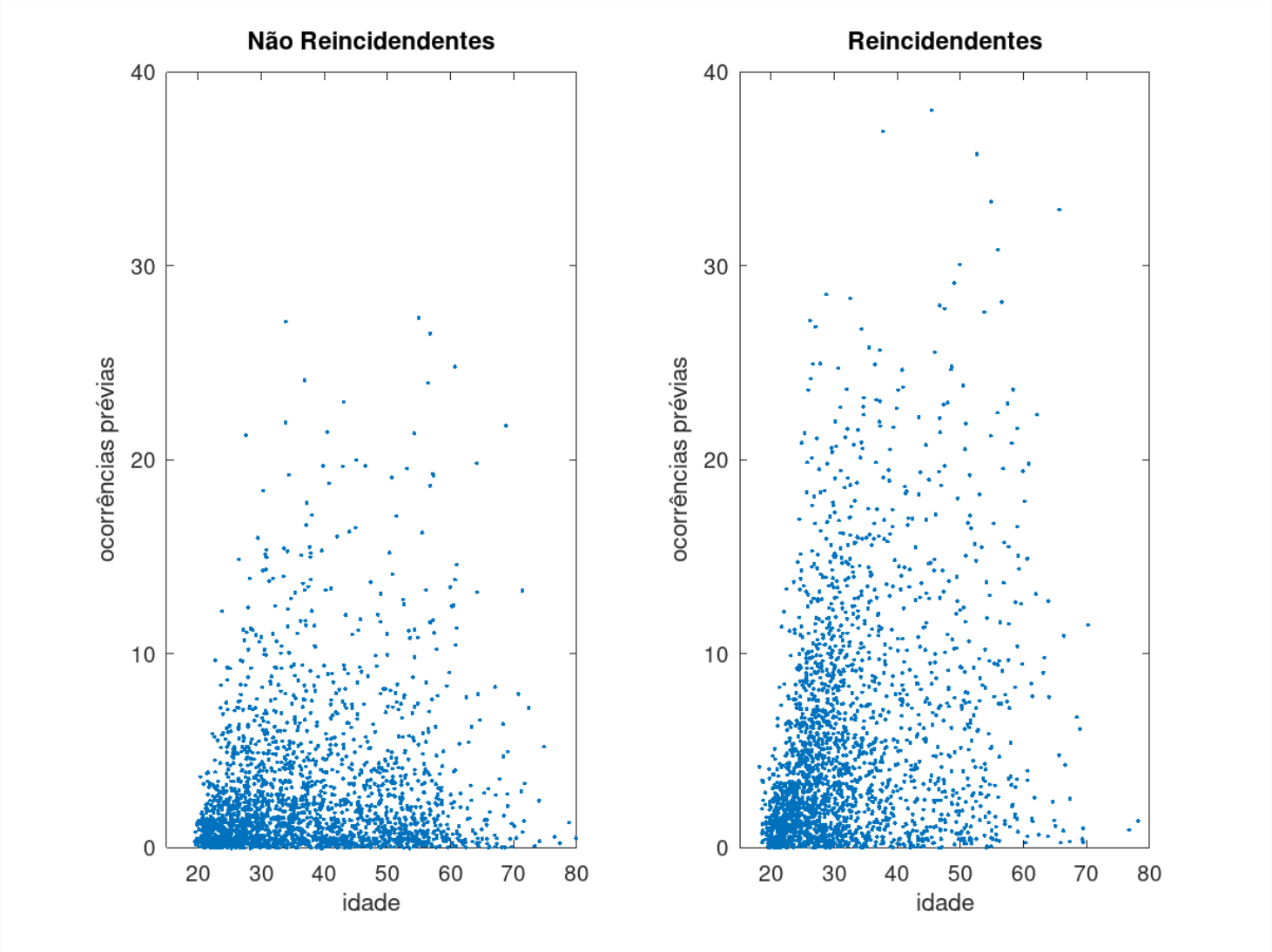
Exemplo: Risco de Reincidência

entrada x						saída y
idade	gênero	raça	⋯	grau da infr.	ocorr. prévias	reinc. 2 anos
41	homem	afro-americana	⋯	contravenção	0	falsa
47	mulher	caucasiana	⋯	contravenção	1	verdadeira
23	homem	caucasiana	⋯	crime	5	verdadeira
...	...	...	...	...	...	...

Quem tem interesse na predição?

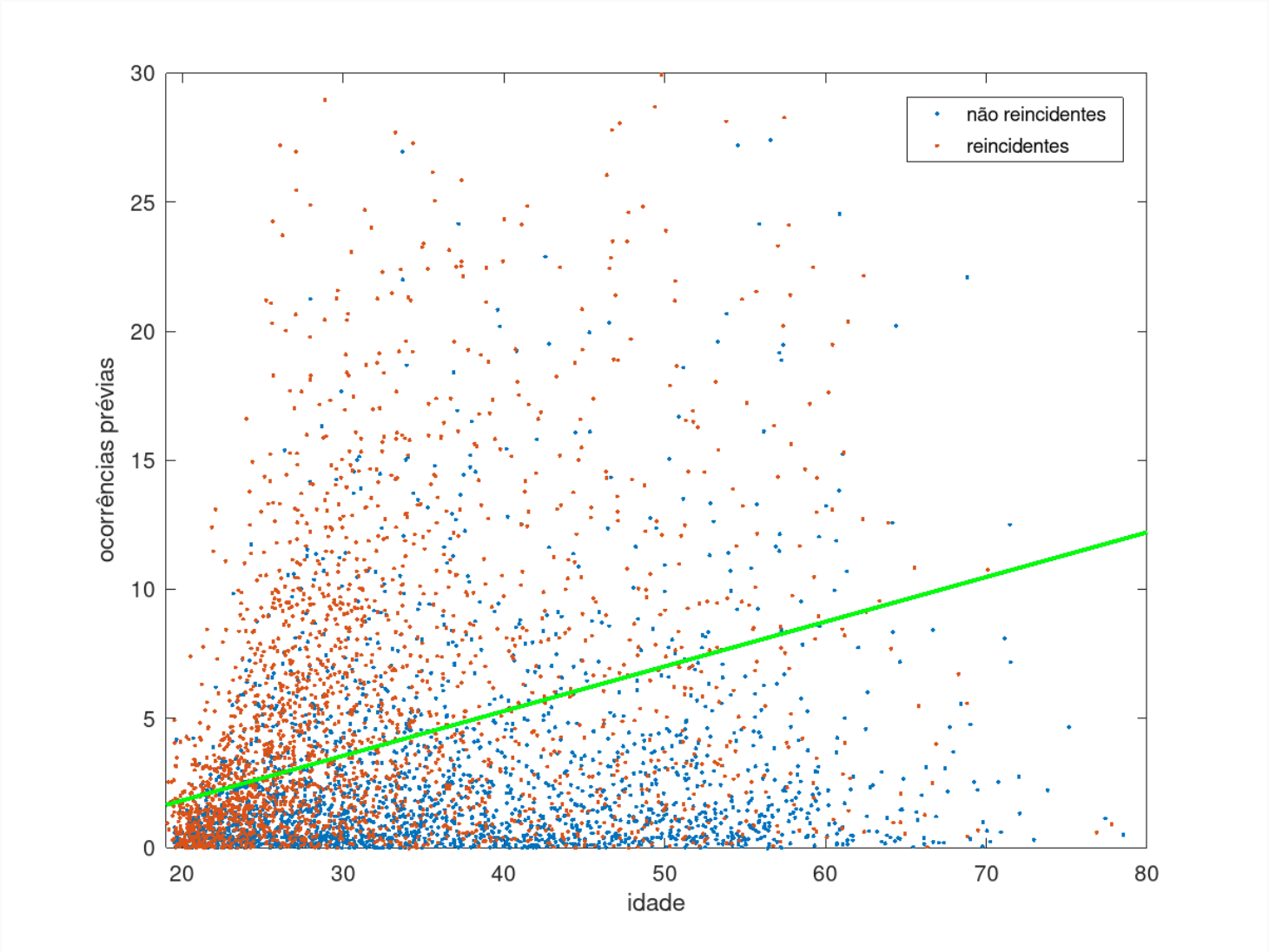
- juiz: uma predição sobre reincidência pode ser utilizada para decisões sobre liberdade condicional ou fiança
- réu: precisão (precision), todos não-reincidentes recebem predição correta
- vítima futura: revocação (recall), todos reincidentes recebem predição correta

Dados ProPublica



Regressão Logística

Atributos: idade e ocorrências prévias



Erros

Prevalência de casos positivos: 0.47.

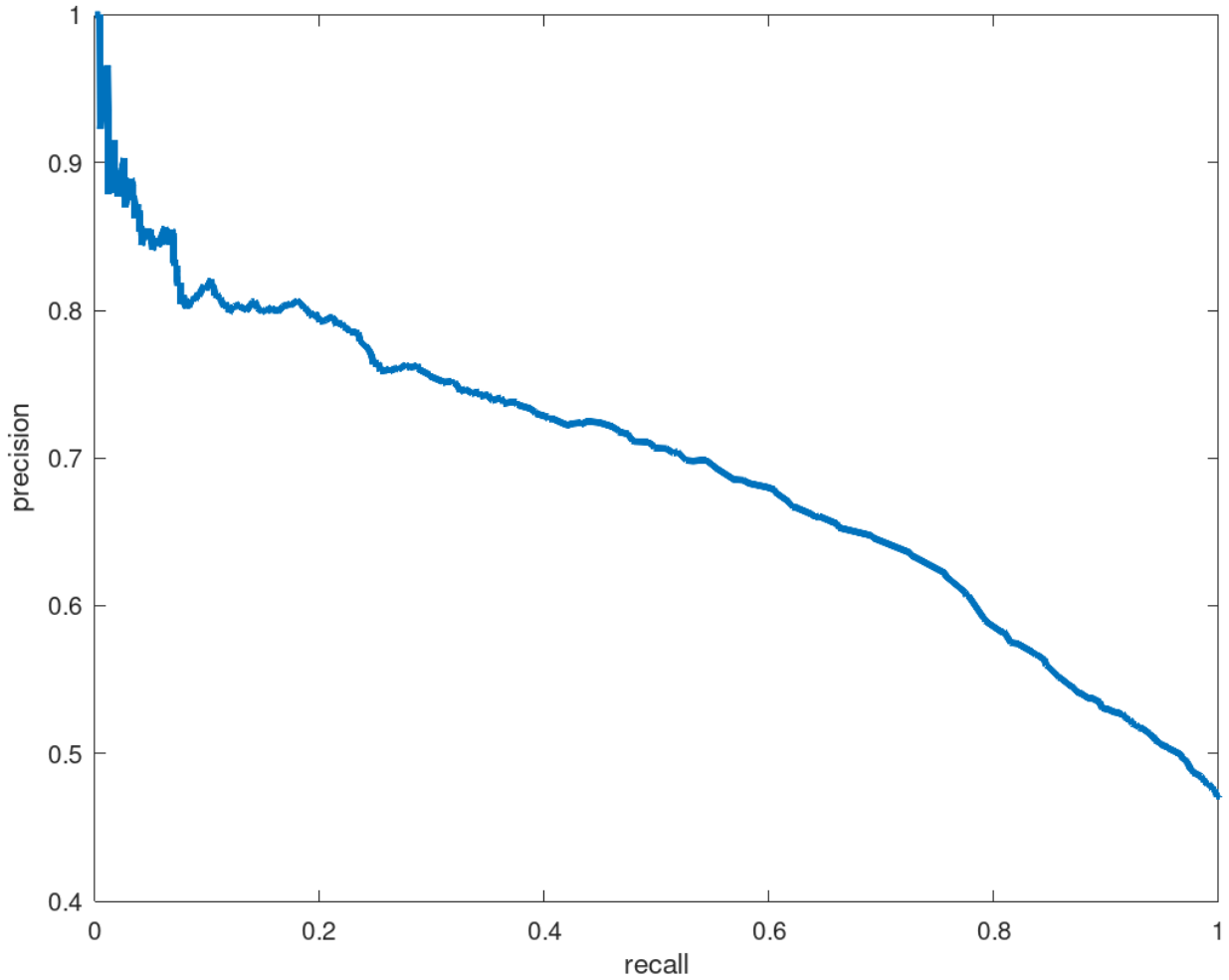
Resultados com $T = 0.5$		**Predição do Modelo**	
		Negativo	**Positivo**
Valores	(N = 2795)	(TN = 2356)	(FP = 439)
Corretos	(P = 2483)	(FN = 1343)	(TP = 1140)

- Acurácia (accuracy): 0.662
- Revocação (recall): 0.459
- Precisão (precision): 0.722

Qual compromisso entre Revocação e Precisão?

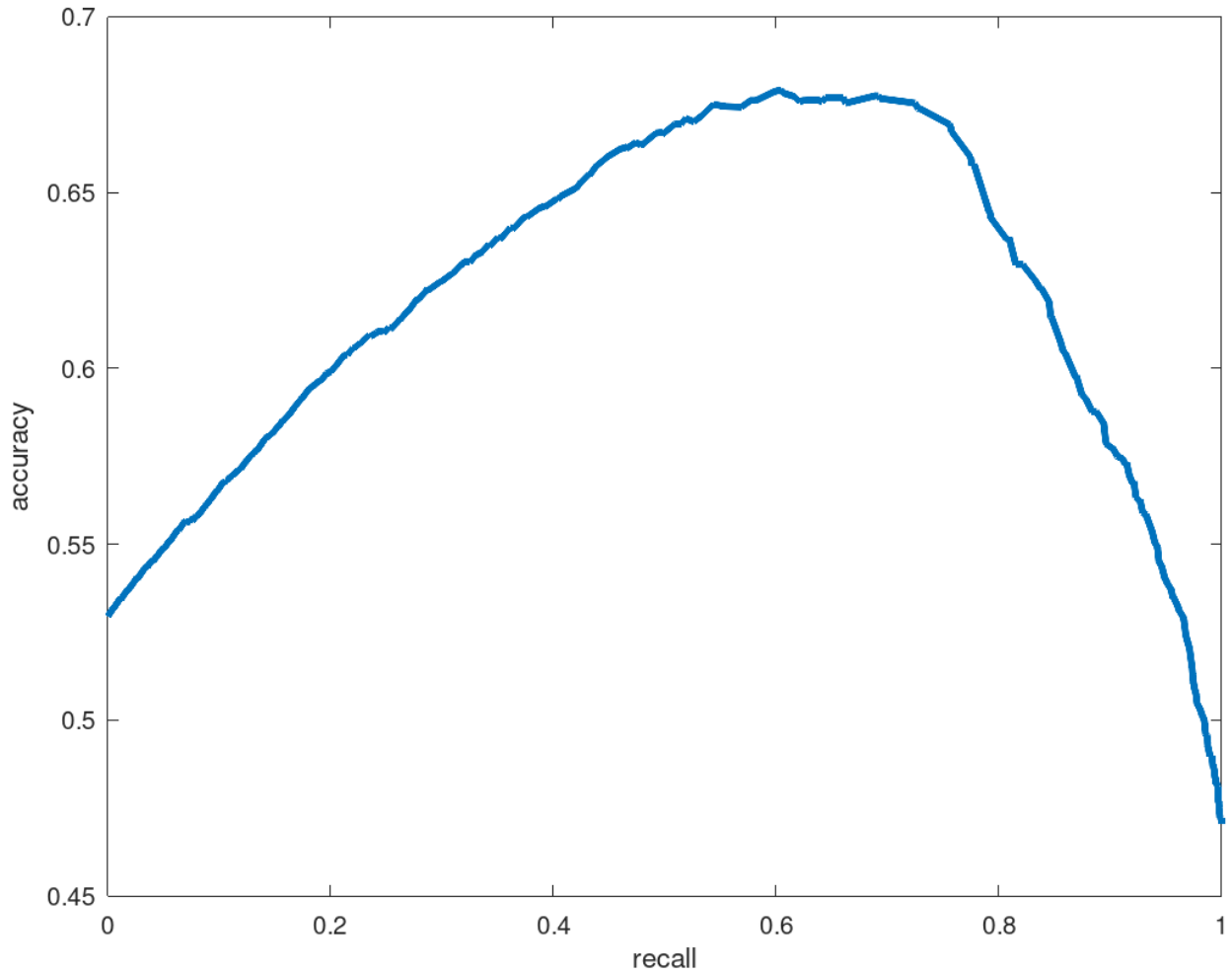
Curva Precisão-Revocação[2]

Variando o valor de T (*threshold*), obtém-se diferentes compromissos.



[2]Fawcett. An introduction to ROC analysis.

Curva Acurácia-Revocação



Outras Medidas[3]

		Predicted condition			
	Total population $= P + N$	**Positive (PP)**	**Negative (PN)**	Informedness, bookmaker informedness (BM) $= TPR + TNR - 1$	Prevalence threshold (PT) $= \frac{\sqrt{TPR \times FPR} - FPR}{TPR - FPR}$
Actual condition	**Positive (P)**	**True positive (TP),** **hit**	**False negative (FN),** **type II error, miss,** **underestimation**	True positive rate (TPR), recall, sensitivity (SEN), probability of detection, hit rate, power $= \frac{TP}{P} = 1 - FNR$	False negative rate (FNR), miss rate $= \frac{FN}{P} = 1 - TPR$
	Negative (N)	**False positive (FP),** **type I error, false alarm,** **overestimation**	**True negative (TN),** **correct rejection**	False positive rate (FPR), probability of false alarm, fall-out $= \frac{FP}{N} = 1 - TNR$	True negative rate (TNR), specificity (SPC), selectivity $= \frac{TN}{N} = 1 - FPR$
	Prevalence $= \frac{P}{P + N}$	Positive predictive value (PPV), precision $= \frac{TP}{PP} = 1 - FDR$	False omission rate (FOR) $= \frac{FN}{PN} = 1 - NPV$	Positive likelihood ratio (LR+) $= \frac{TPR}{FPR}$	Negative likelihood ratio (LR−) $= \frac{FNR}{TNR}$
	Accuracy (ACC) $= \frac{TP + TN}{P + N}$	False discovery rate (FDR) $= \frac{FP}{PP} = 1 - PPV$	Negative predictive value (NPV) $= \frac{TN}{PN} = 1 - FOR$	Markedness (MK), deltaP (Δp) $= PPV + NPV - 1$	Diagnostic odds ratio (DOR) $= \frac{LR+}{LR-}$
	Balanced accuracy (BA) $= \frac{TPR + TNR}{2}$	F_1 score $= \frac{2\,PPV \times TPR}{PPV + TPR} = \frac{2\,TP}{2\,TP + FP + FN}$	Fowlkes–Mallows index (FM) $= \sqrt{PPV \times TPR}$	Matthews correlation coefficient (MCC) $= \sqrt{TPR \times TNR \times PPV \times NPV} - \sqrt{FNR \times FPR \times FOR \times FDR}$	Threat score (TS), critical success index (CSI), Jaccard index $= \frac{TP}{TP + FN + FP}$

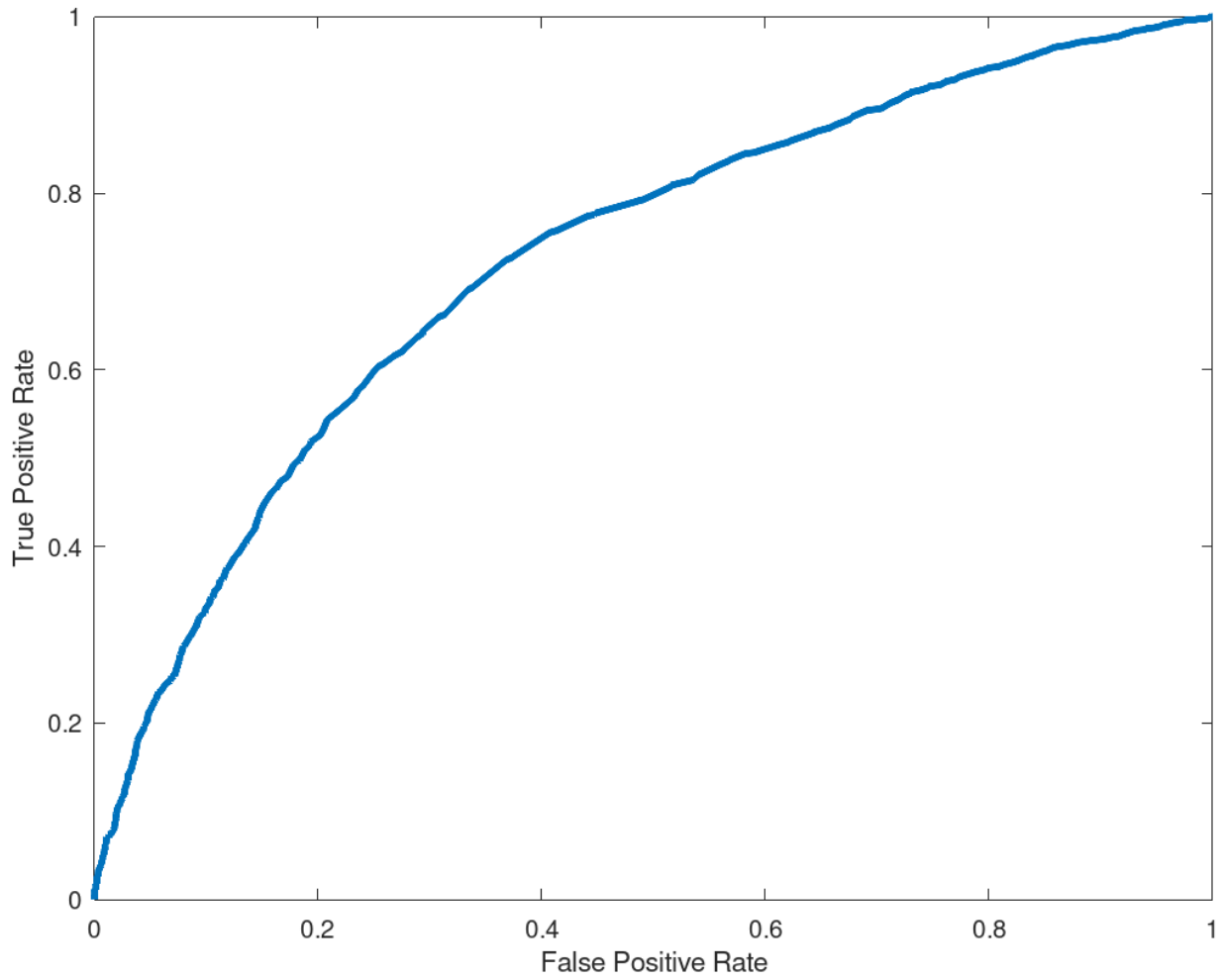
Sources: [20][21][22][23][24][25][26][27] view • talk • edit

[3] https://en.wikipedia.org/wiki/Confusion_matrix

Receiver Operating Characteristic (ROC)

AUC = Area Under the Curve

Invariante à prevalência de casos positivos.



Caso COMPAS: Risco de Reincidência

- COMPAS
  - Correctional Offender Management Profiling for Alternative Sanctions
  - Gerenciamento de Perfis Correcionais de Infratores para Sanções Alternativas
- ProPublica
  - Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks.
  - Viés de Máquina: Há software usado em todo o país para prever futuros criminosos. E é tendencioso contra os negros.

entrada x						saída y'	saída y
idade	gênero	raça	···	grau da infr.	ocorr. prévias	COMPAS	reinc. 2 anos
41	homem	afro-americana	···	contravenção	0	alto	falsa
47	mulher	caucasiana	···	contravenção	1	baixo	verdadeira
23	homem	caucasiana	···	crime	5	alto	verdadeira
...	...	...	...	...	...	...	...

Resultados ProPublica

	Todos Réus		Réus Pretos		Réus Brancos	
	baixo	alto	baixo	alto	baixo	alto
não reincidiu	2681	1282	990	805	1139	349
reincidiu	1216	2035	532	1369	461	505
	Acuracy=0.654 Recall=0.626 Precision=0.614 FPR=0.324 FNR=0.374		Acuracy=0.638 Recall=0.720 Precision=0.630 FPR=0.449 FNR=0.280		Acuracy=0.670 Recall=0.523 Precision=0.591 FPR=0.235 FNR=0.477	

Conclusão: 50% das pessoas brancas que reincidiram, receberiam risco baixo.

Viés (Bias) e Racismo

Atributo Sensível: os atributos X contém ou implicitamente codifica algum atributo sensível (raça, gênero, orientação sexual, etc.) de uma entidade (pessoa).

Viés em Estatística: $Bias[\hat{Y}|X = x] = E[\hat{Y}|X = x] - E[Y|X = x]$.

Viés Sociais: decisões sistematicamente erradas em favor ou contra determinados grupos determinados pelo atributo sensível.

Conceitos:

- A é o atributo sensível: preto ou branco?
- Y é a variável alvo: a pessoa tem uma doença que precisa ser tratada?
- $\widehat{Y}$ é a decisão dada pelo classificador: recomenda o tratamento

Viés e Racismo

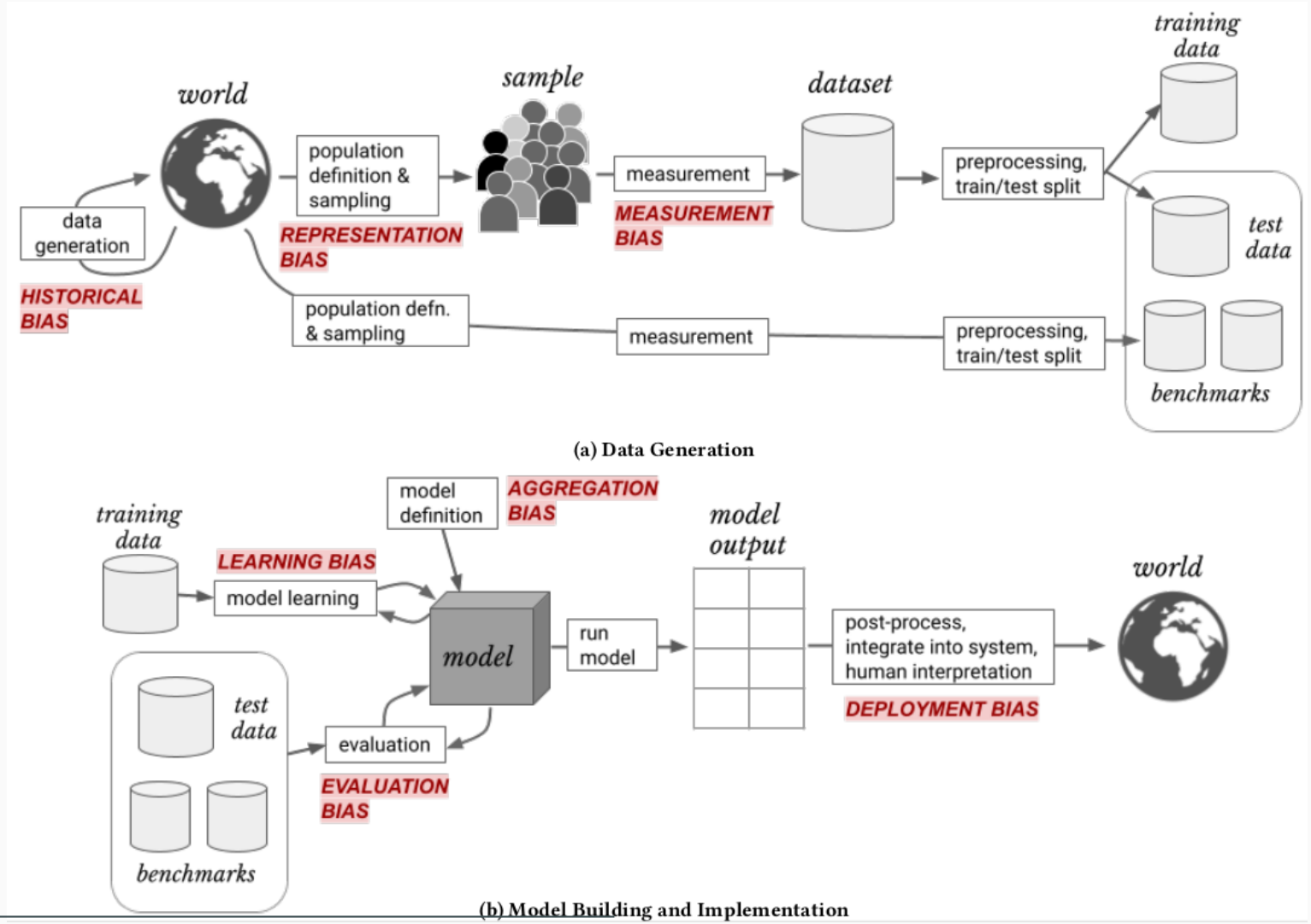
Conceitos:

- A é o atributo sensível
- Y é a variável alvo
- $\widehat{Y}$ é a decisão dada pelo classificador

Justiça (**Fairness**):

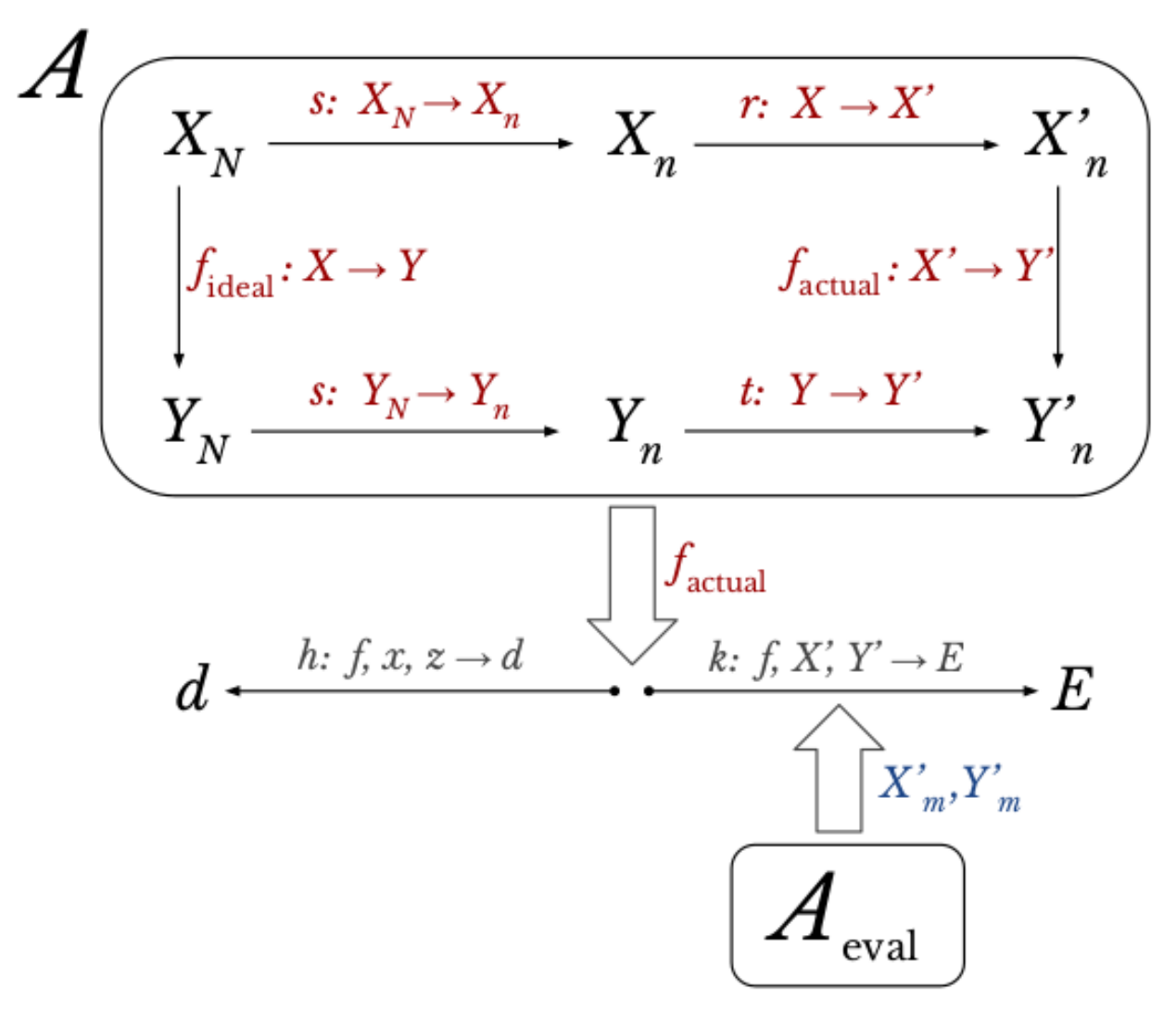
- **Cotas**: $\Pr(\widehat{Y}|A = \textit{negra}) = \Pr(\widehat{Y}|A = \textit{branca})$
  As pessoas que recebem o tratamento para uma doença são proporcionais à ocorrência do grupo na população
- **Recall:** $\Pr(\widehat{Y}|A = \textit{negra}, Y) = \Pr(\widehat{Y}|A = \textit{branca}, Y)$
  As pessoas que recebem o tratamento são proporcionais à ocorrência do grupo na população que tem a doença
- **Precision:** $\Pr(Y|A = \textit{negra}, \widehat{Y}) = \Pr(Y|A = \textit{branca}, \widehat{Y})$
  As pessoas que tem a doença são proporcionais à ocorrência do grupo na população que recebeu tratamento

Fontes de Viés[4]



[4]Suresh e Guttag. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle, 2021.

Fontes de Viés



Dados ProPublica - Coleta de Dados

Característica dos Dados:

- 60.16% pretos (apenas 15% na população entre brancos e pretos)
- Reincidência de pretos: 52.32%
- Reincidência de brancos: 39.01%

Fontes de Viés:

- Histórica
- Representação
- Mensuração
- Avaliação

Estudos de Casos

Black is to Criminal as Caucasian is to Police: Detecting and Removing Multiclass Bias in Word Embeddings

Thomas Manzini, Yao Chong Lim, Yulia Tsvetkov, Alan W Black

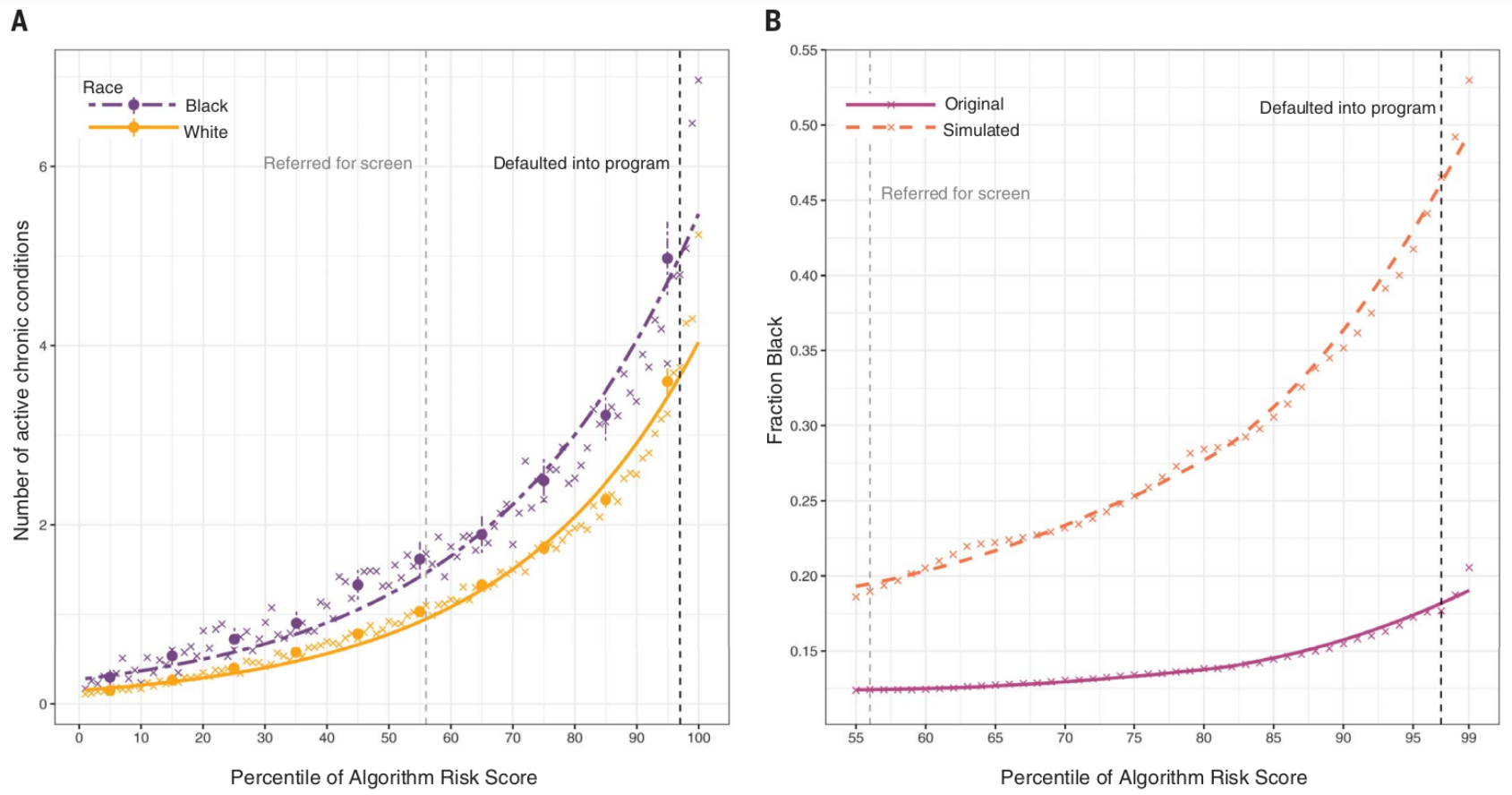
Gender Biased Analogies	
man → doctor	woman → nurse
woman → receptionist	man → supervisor
woman → secretary	man → principal
Racially Biased Analogies	
black → criminal	caucasian → police
asian → doctor	caucasian → dad
caucasian → leader	black → led
Religiously Biased Analogies	
muslim → terrorist	christian → civilians
jewish → philanthropist	christian → stooge
christian → unemployed	jewish → pensioners

Table 1: Examples of gender, racial, and religious biases in analogies generated from word embeddings trained on the Reddit data from users from the USA.

Estudos de Casos

Dissecting racial bias in an algorithm used to manage the health of populations

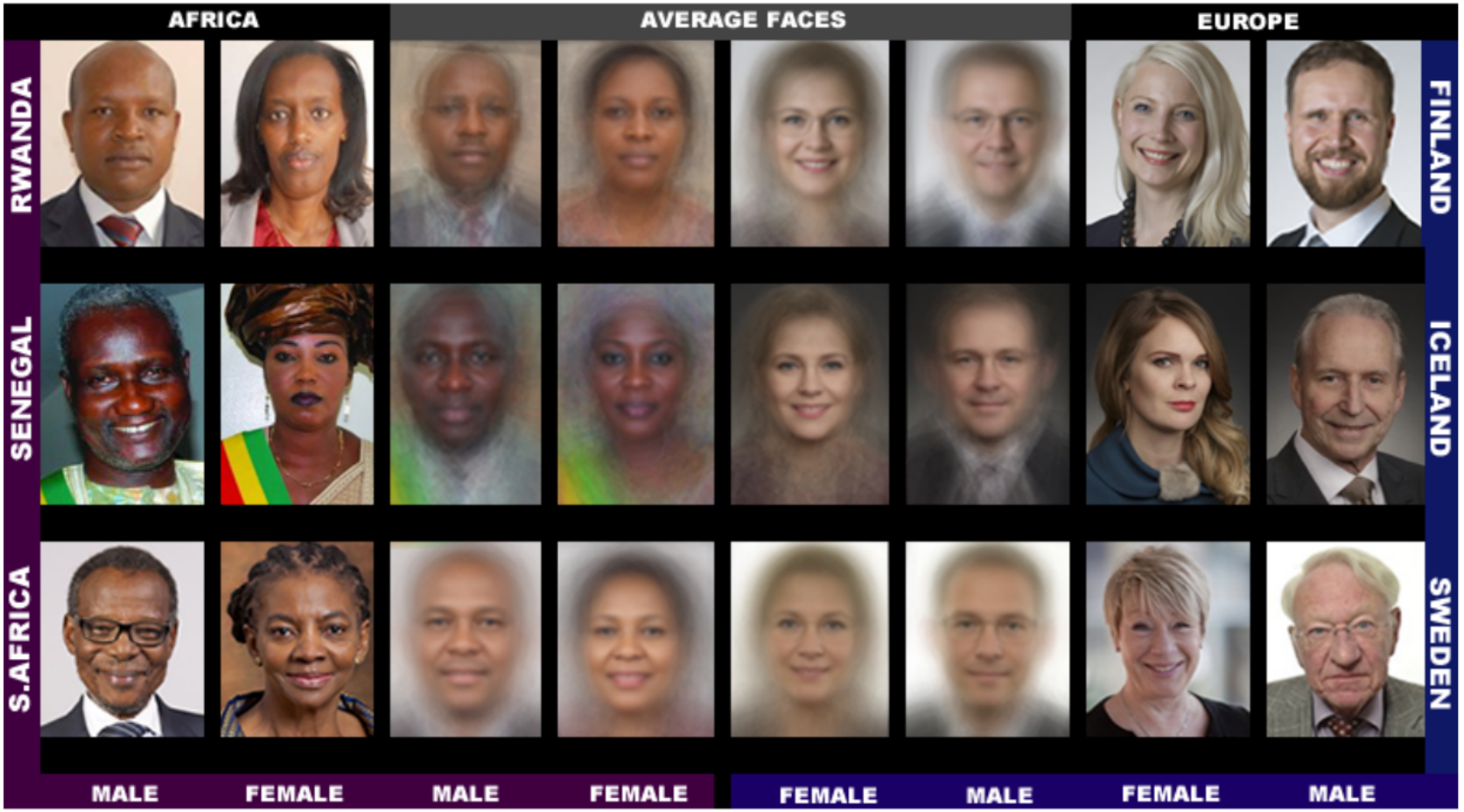
Ziad Obermeyer, Brian Powers, Christine Vogeli, Sendhil Mullainathan



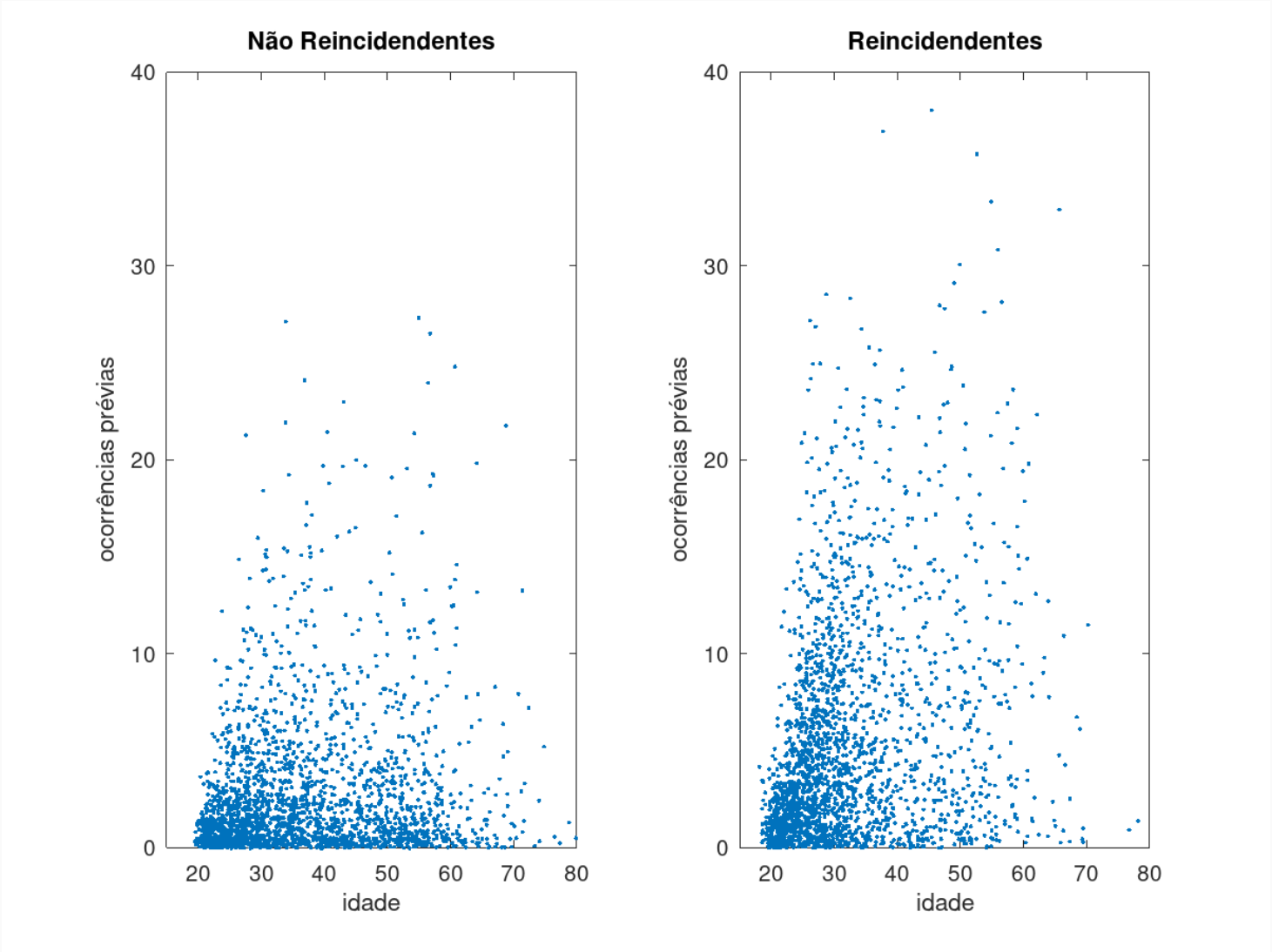
Estudos de Casos

Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification

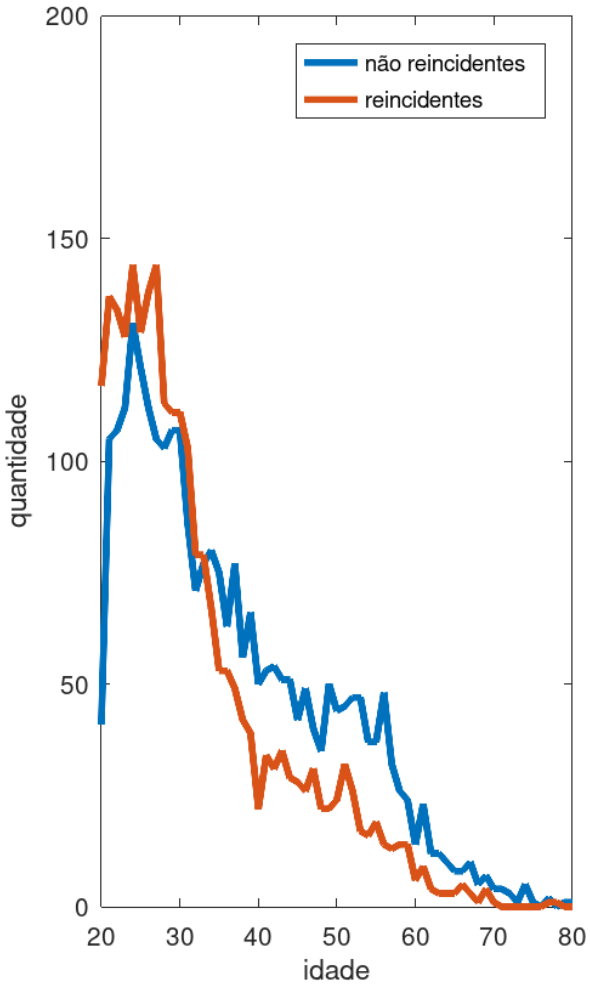
Joy Buolamwini, Timnit Gebru

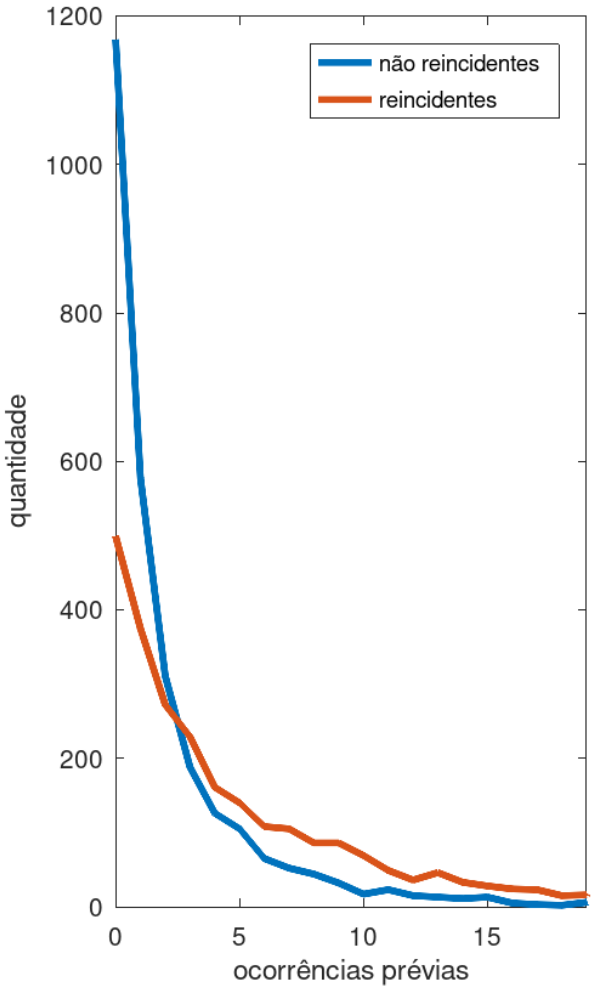


Dados ProPublica - Aprendizado de Máquina

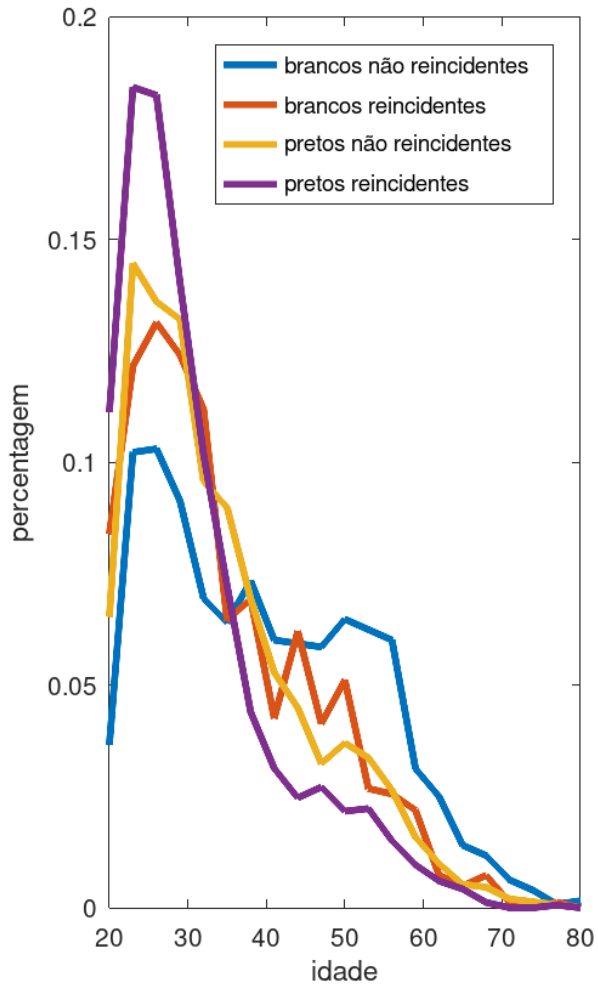


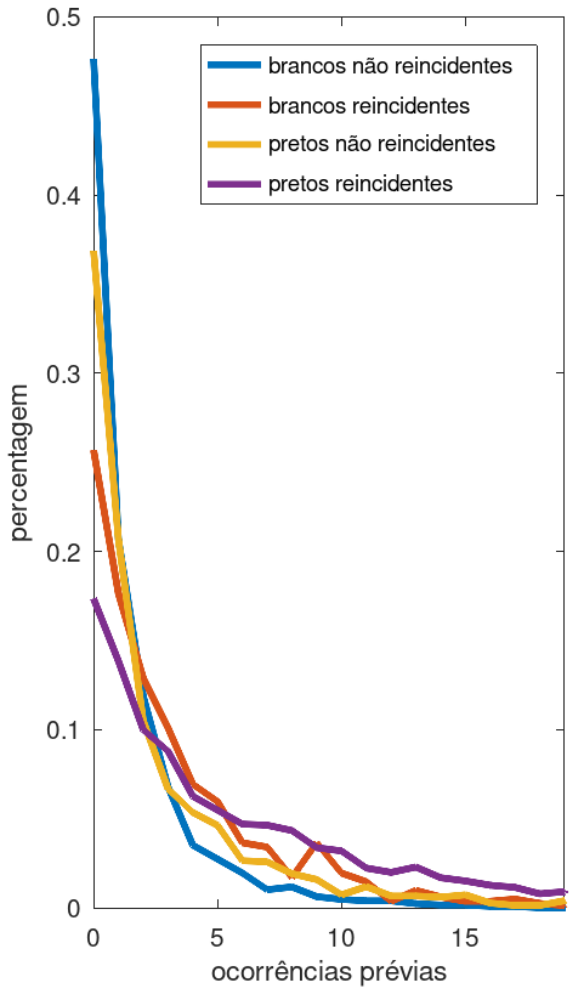
Dados ProPublica - Aprendizado de Máquina



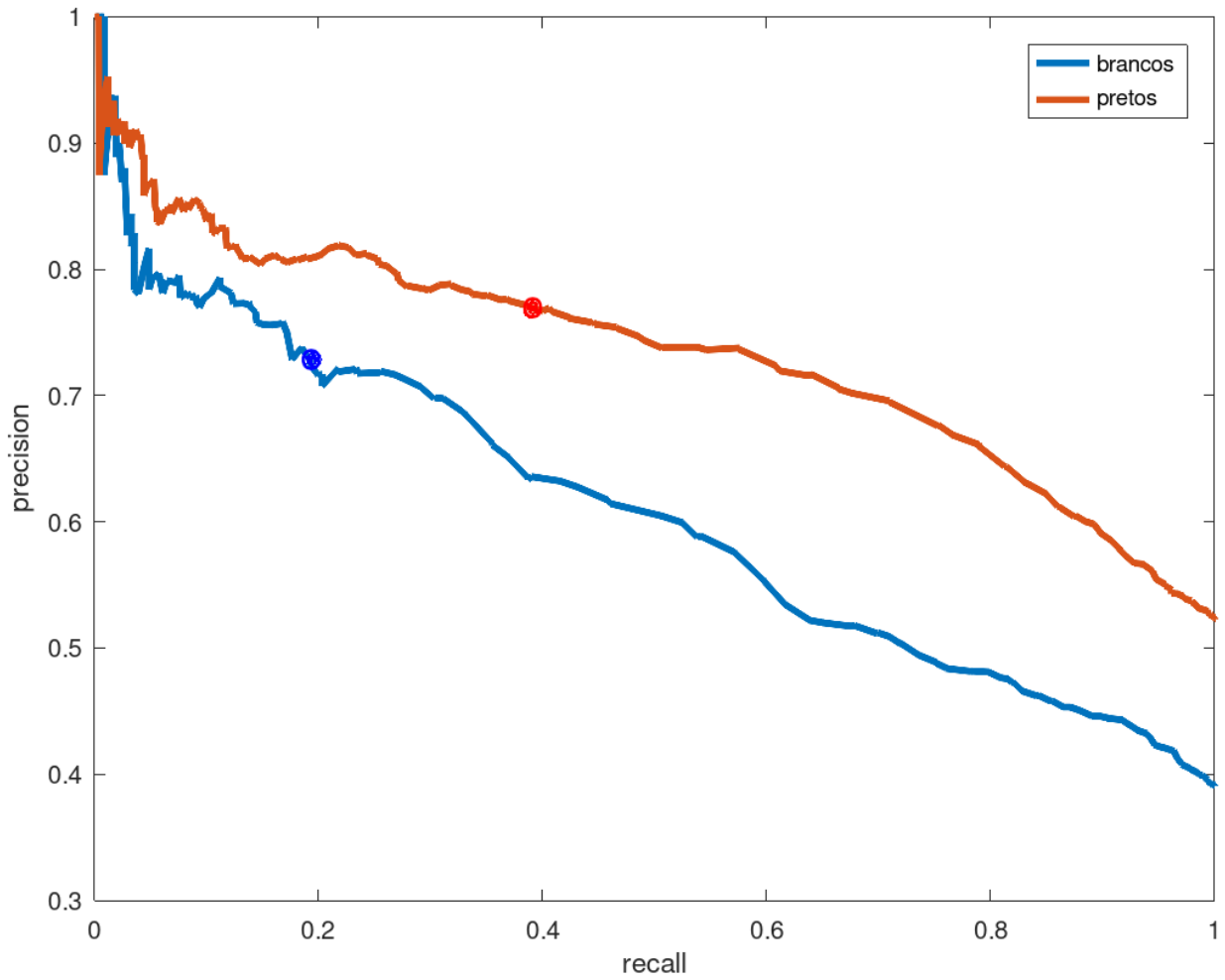


Dados ProPublica - Viés de Agregação

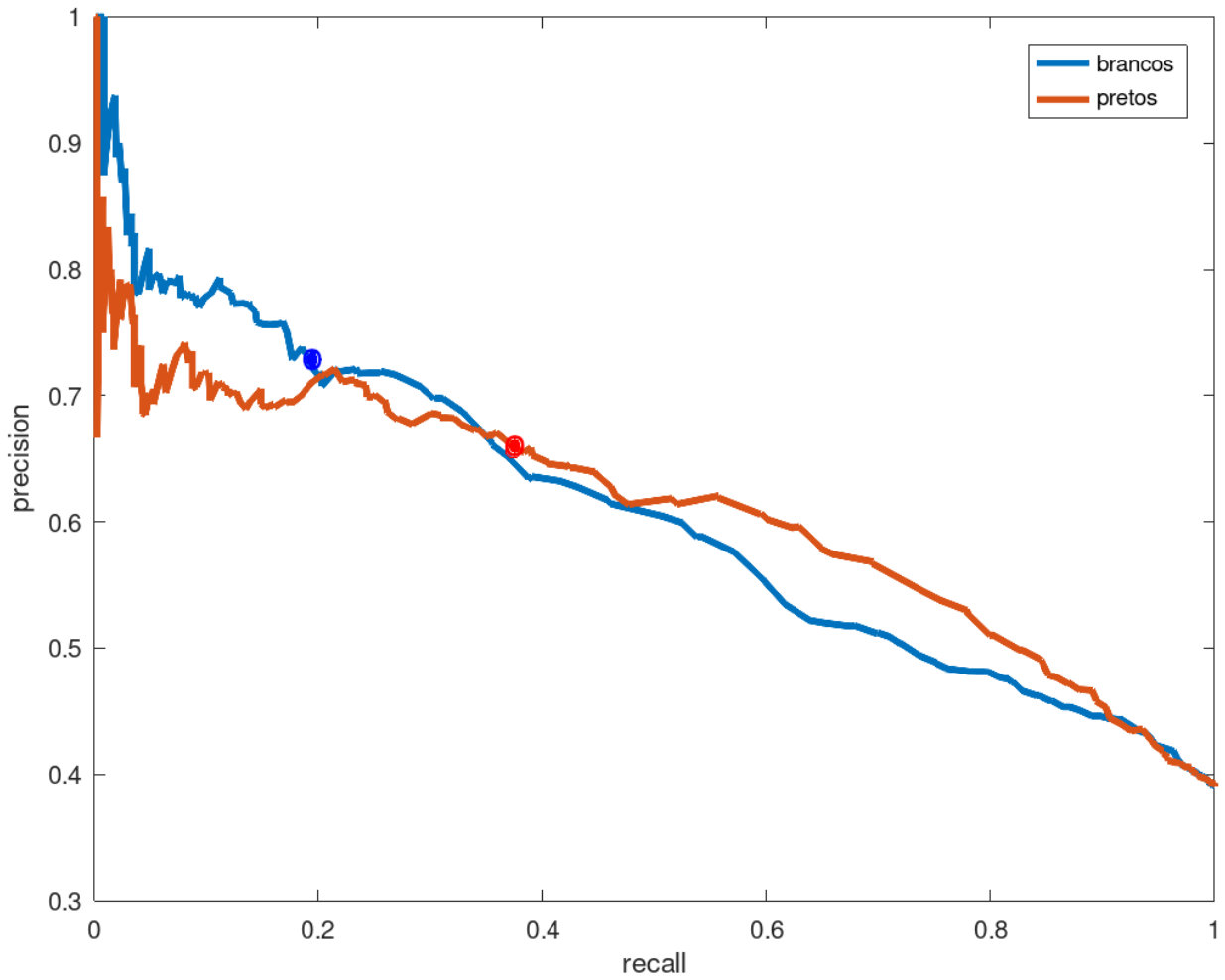




Dados ProPublica - Viés de Aprendizado



Dados ProPublica - Balanceamento de Dados



Conclusão

- Aprendizado de Máquina **reconhece padrões em amostras** e **generaliza** para entidades **desconhecidas**.
- **Erros são inerentes** a qualquer método de aprendizado de máquina.
- Como garantir que generalizações sejam feitas da forma correta e pela razão correta?
- Como garantir que decisões de alto-risco que moldam as chances na vida são justas e corretas?
- **Quem** está disposto para aceitar erros de generalização e em **qual** sistema?
- Quais **medidas de desempenho** mudaria a disposição para aceitar generalização?
- E se for necessário utilizar **regras diferentes** para **grupos diferentes**?
- Como garantir que **sistemas** adotem essas medidas?
- Como garantir que não houve **erros procedurais** nas decisões realizadas?

Conclusão

"Is our goal to faithfully reflect the data? Or do we have an obligation to question the data, and to design our systems to conform to some notion of equitable behavior, regardless of whether or not that's supported by the data currently available to us?"[5]

"Transparent algorithms provide defendants and the public with imperative information about tools used for safety and justice, allowing a wider audience to participate in the discussion of fairness. . . . . We argue that it is not fair that life-changing decisions are made with an error-prone system, without entitlement to a clear, verifiable, explanation."[6]

---

[5] Barocas, Hardt, e Narayanan. Fairness and Machine Learning: Limitations and Opportunities.

[6] Rudin, Wang e Coker. The age of secrecy and unfairness in recidivism prediction.