Escola Politécnica da Universidade de São Paulo

Avaliação de Discriminação em Aprendizagem de Máquina usando Técnicas de Interpretabilidade

Mestranda: Juliana Cesaro
Orientador: Fabio Gagliardi Cozman

Agenda

- Motivação
- Objetivo da pesquisa
- Avaliação de discriminação e remoção de viés
- Interpretabilidade
- Resultados
- Conclusão

RESEARCH ARTICLE

ECONOMICS

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

Ziad Obermeyer[1,2]*, Brian Powers[3], Christine Vogeli[4], Sendhil Mullainathan[5]*†

Health systems rely on commercial prediction algorithms to identify and help patients with complex health needs. We show that a widely used algorithm, typical of this industry-wide approach and affecting millions of patients, exhibits significant racial bias: At a given risk score, Black patients are considerably sicker than White patients, as evidenced by signs of uncontrolled illnesses. Remedying this disparity would increase the percentage of Black patients receiving additional help from 17.7 to 46.5%. The bias arises because the algorithm predicts health care costs rather than illness, but unequal access to care means that we spend less money caring for Black patients than for White patients. Thus, despite health care cost appearing to be an effective proxy for health by some measures of predictive accuracy, large racial biases arise. We suggest that the choice of convenient, seemingly effective proxies for ground truth can be an important source of algorithmic bias in many contexts.

Racial bias in a medical algorithm favors white patients over sicker ...

A widely used algorithm that flags patients for extra medical care is biased against black patients, a study found.

washingtonpost.com

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

WASHINGTONPOST.COM

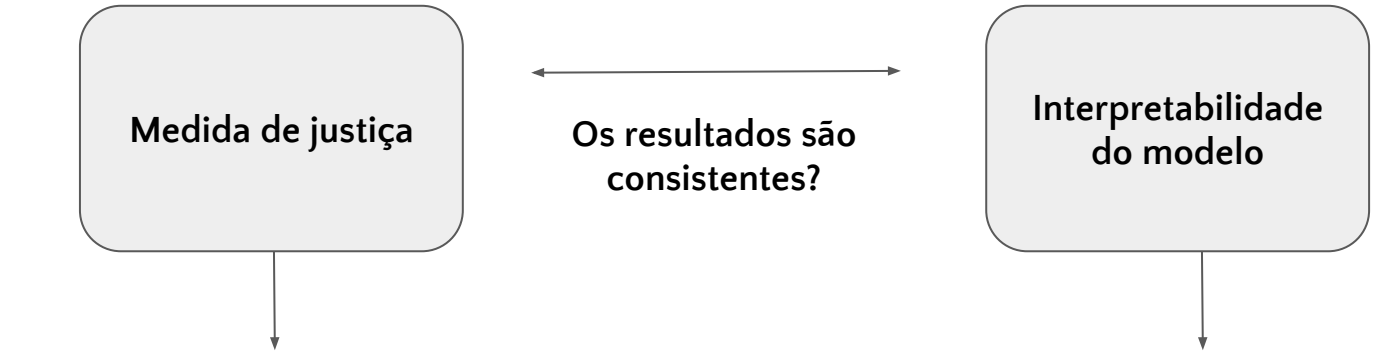
A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear.

BBC News Brasil @bbcbrasil · Oct 31, 2016

Sistema de algoritmo que determina pena de condenados cria polêmica nos EUA bbc.in/2f9pasa

O que deve ser feito para evitar a construção de modelos de ML discriminatórios?

Considerações durante o desenvolvimento do modelo para evitar discriminação

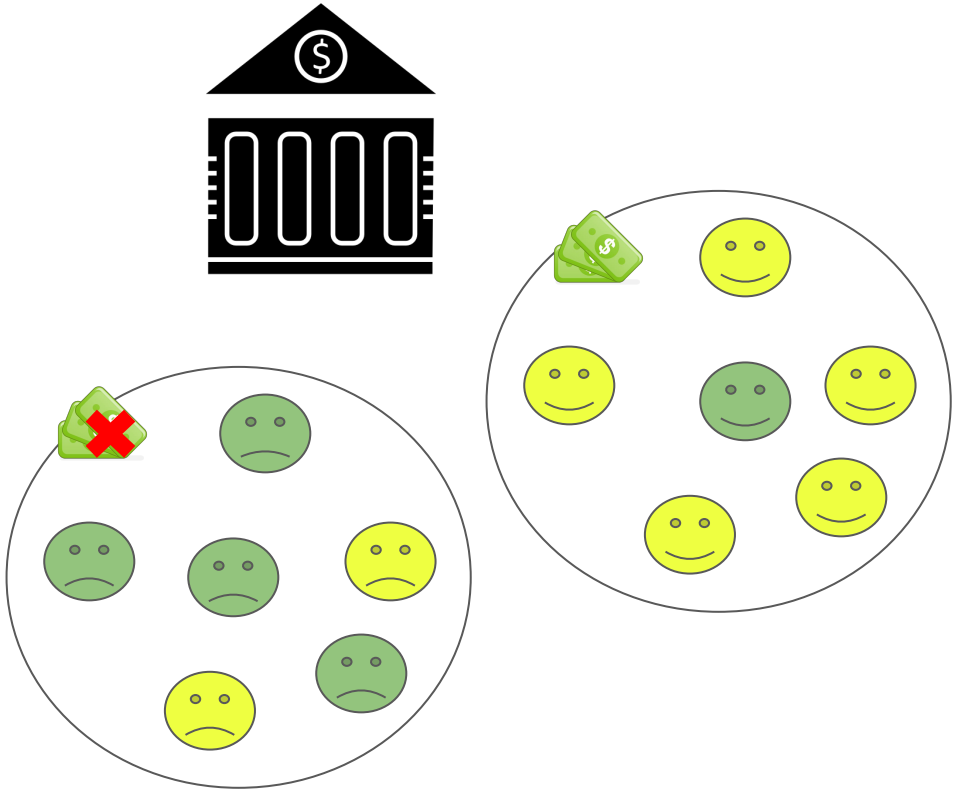


Definição de justiça das Institutas de Justiniano

“A constante e perpétua vontade de dar a cada um o que é seu”

Conceitos

- **Grupo privilegiado:** pessoas de cor amarela
- **Grupo desprivilegiado:** pessoas de cor verde
- **Variável sensível:** cor
- **Viés:** erro sistemático que faz com que o grupo privilegiado tenha vantagem sistemática



Definições de justiça

Justiça entre grupos

- Paridade estatística:

$$P(\hat{Y} = 1|A = 0) = P(\hat{Y} = 1|A = 1)$$

- Igualdade de oportunidade

$$P(\hat{Y} = 1|A = 0, Y = 1) = P(\hat{Y} = 1|A = 1, Y = 1)$$

Justiça entre grupos e indivíduos

- Índice de entropia generalizado:

$$\frac{1}{n\alpha(\alpha-1)}\sum_{i=1}^{n}\left[\left(\frac{b_i}{\mu}\right)^{\alpha}-1\right]$$

$$b_i = \hat{y}_i - y_i + 1. \qquad \mu_g = \frac{1}{|g|}\sum_{i\in g} b_i.$$

Justiça entre indivíduos

- Consistência:

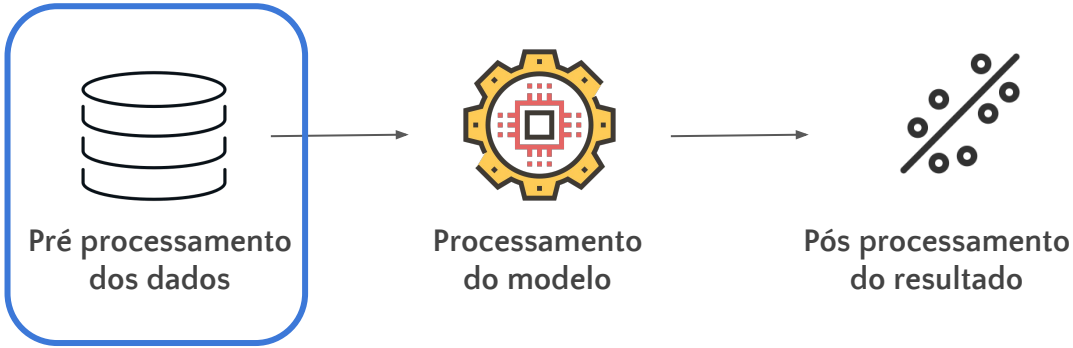
$$1 - \frac{1}{N}\sum_{n=1}^{N}\left|\hat{y}_n - \frac{1}{k}\sum_{j \epsilon kNN(X')}\hat{y}_j\right|$$

Abordagens causais

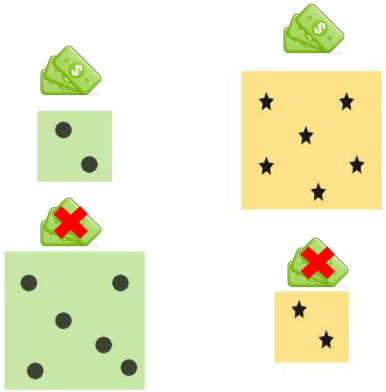
- Justiça contrafactual:

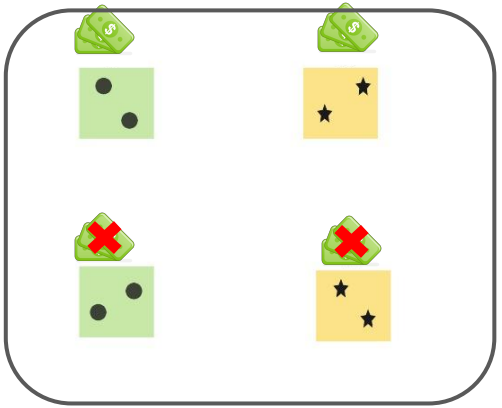
$$P(\hat{Y}_{A\leftarrow a'}|X = x, A = a) = P(\hat{Y}_{A\leftarrow a}|X = x, A = a)$$

Técnica para remoção de viés

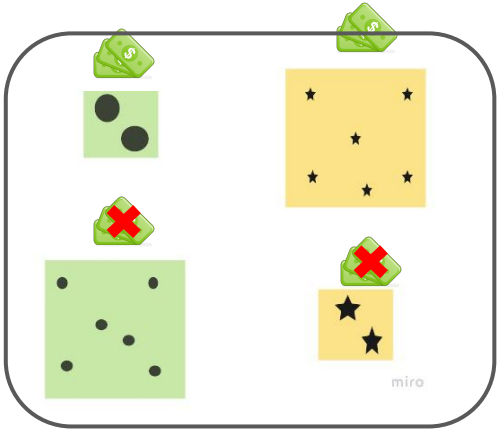


Técnica para remoção de viés: pré processamento



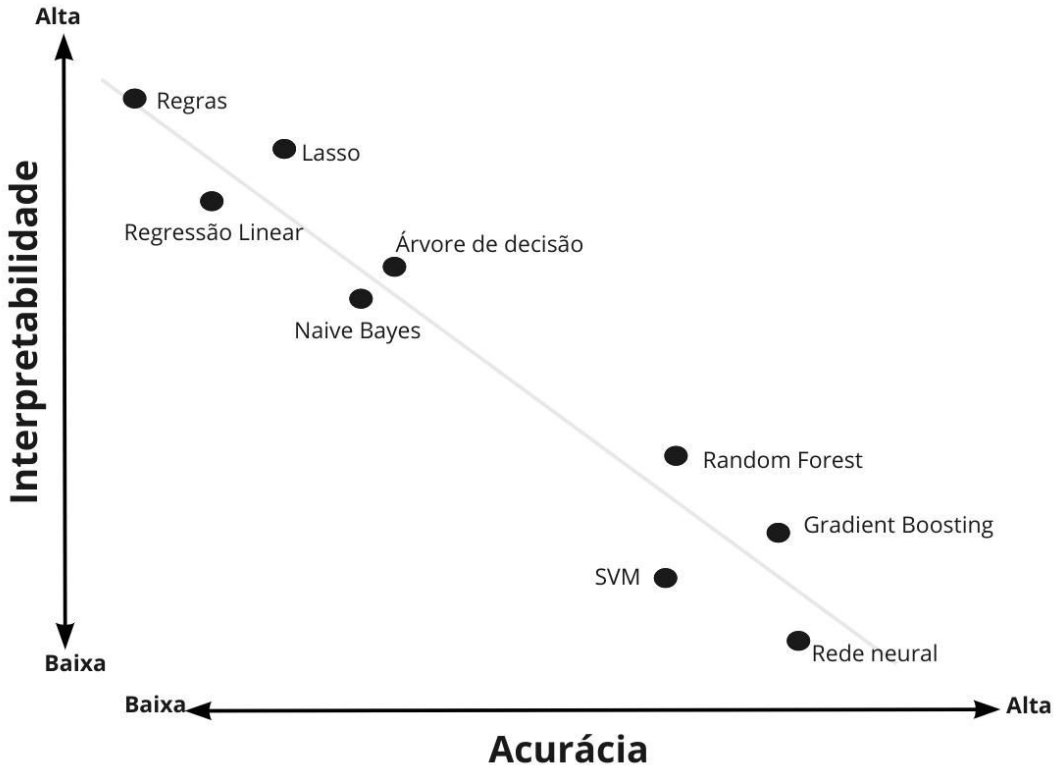


Subamostragem parametrizada



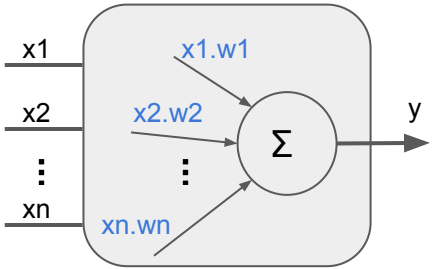
Reponderação

Interpretabilidade do modelo

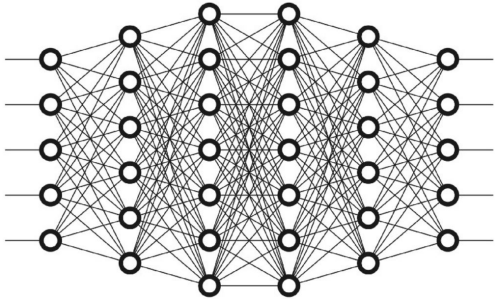


Interpretabilidade do modelo

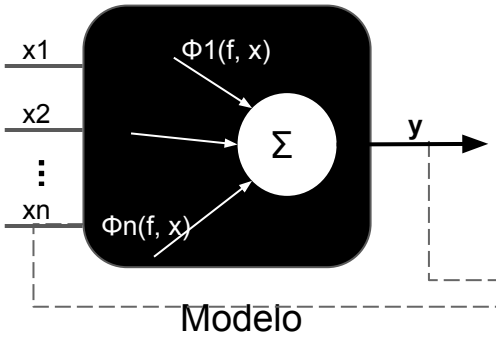
Modelo Linear



Rede neural

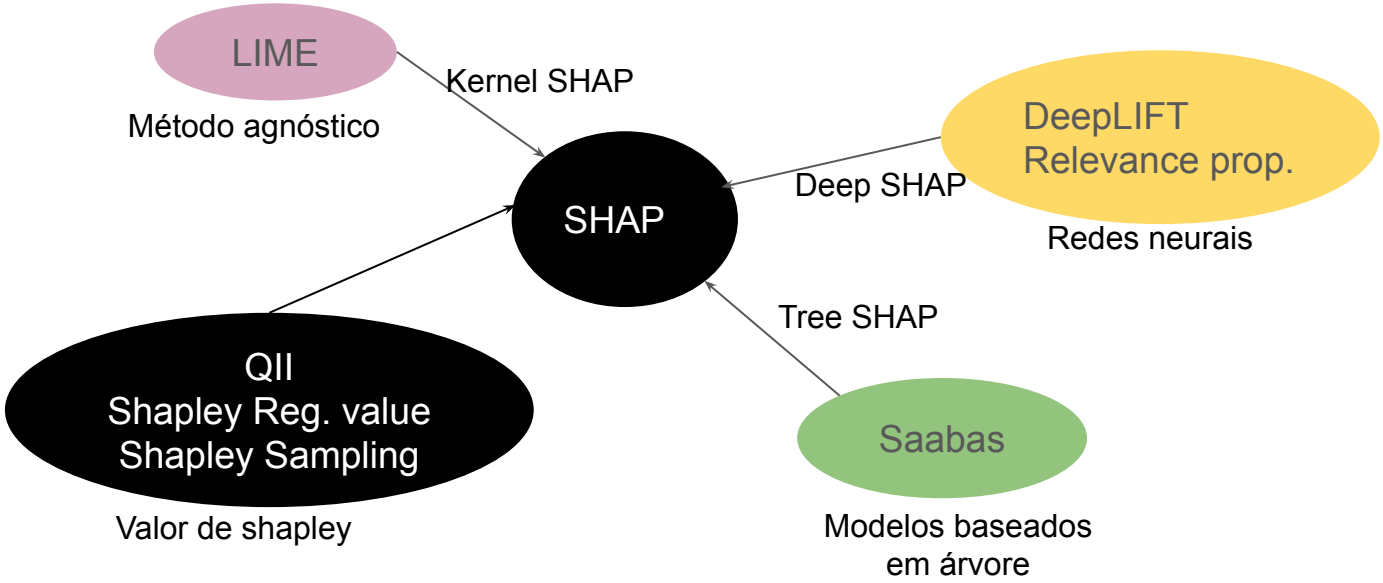


Métodos de atribuição aditiva da feature



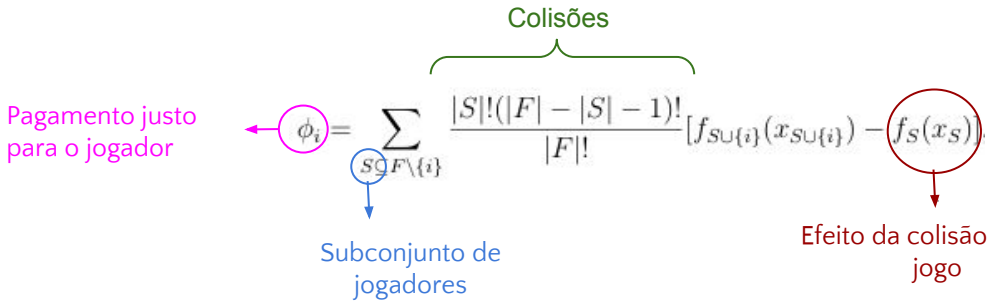
Resultado do **SHAP**

Métodos de atribuição aditiva da feature



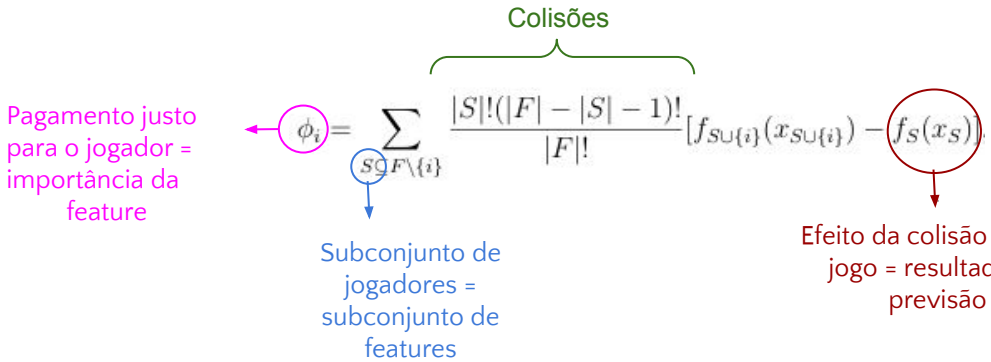
Valor de Shapley

Prevê o efeito de cada jogador considerando as diferentes colisões que ele pode participar no jogo, e com isso calcular qual seria o pagamento justo de cada jogador.



Valor de Shapley

Prevê o efeito de cada jogador considerando as diferentes colisões que ele pode participar no jogo, e com isso calcular qual seria o pagamento justo de cada jogador.



Valor de Shapley

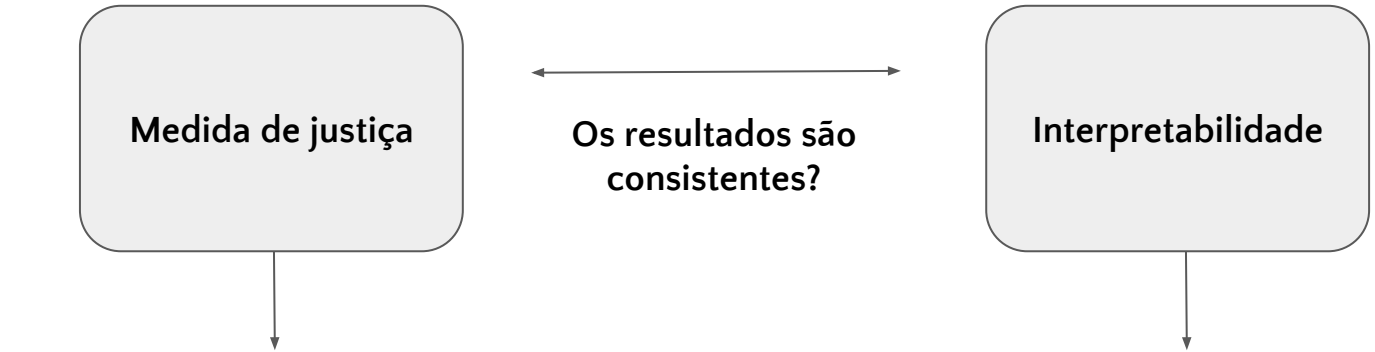
Propriedades garantidas:

- **Acuracidade local:** resultado da explicação é igual ao do modelo

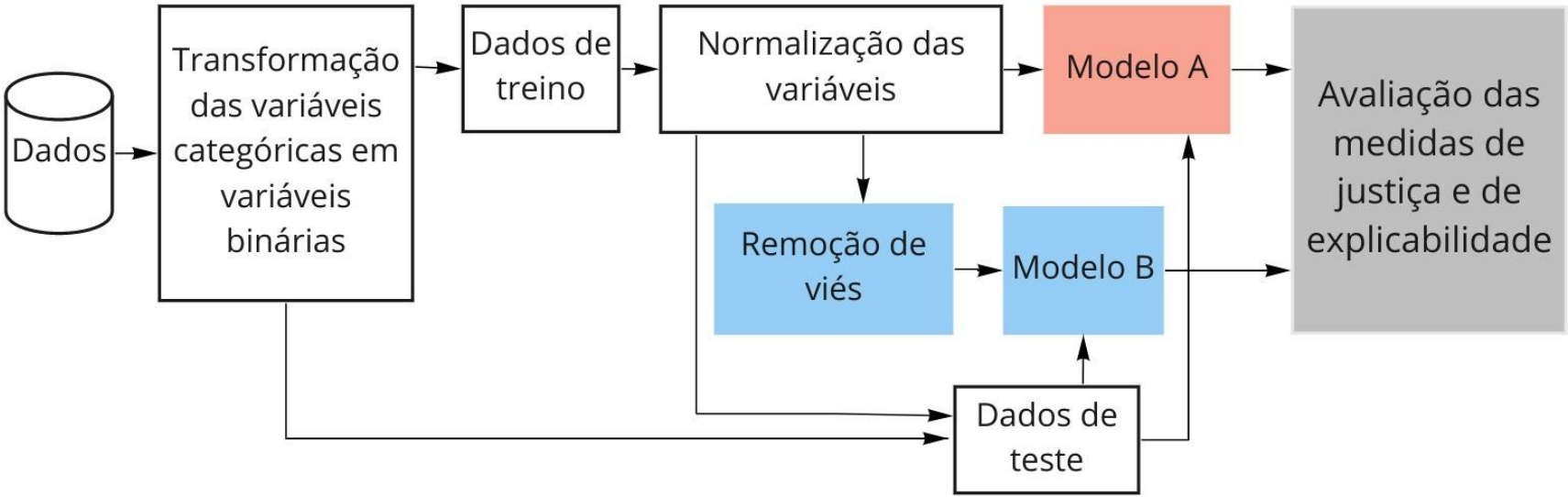
$$f(h_x(z')) = g(z') = \phi_0 + \sum_{i=0}^{M} \phi_i z'_i.$$

- **Consistência:** se a influência de uma feature no modelo fica maior ou igual, a importância gerada não deve diminuir
- **Omissão:** se uma feature não altera o resultado do modelo sua importância deve ser nula

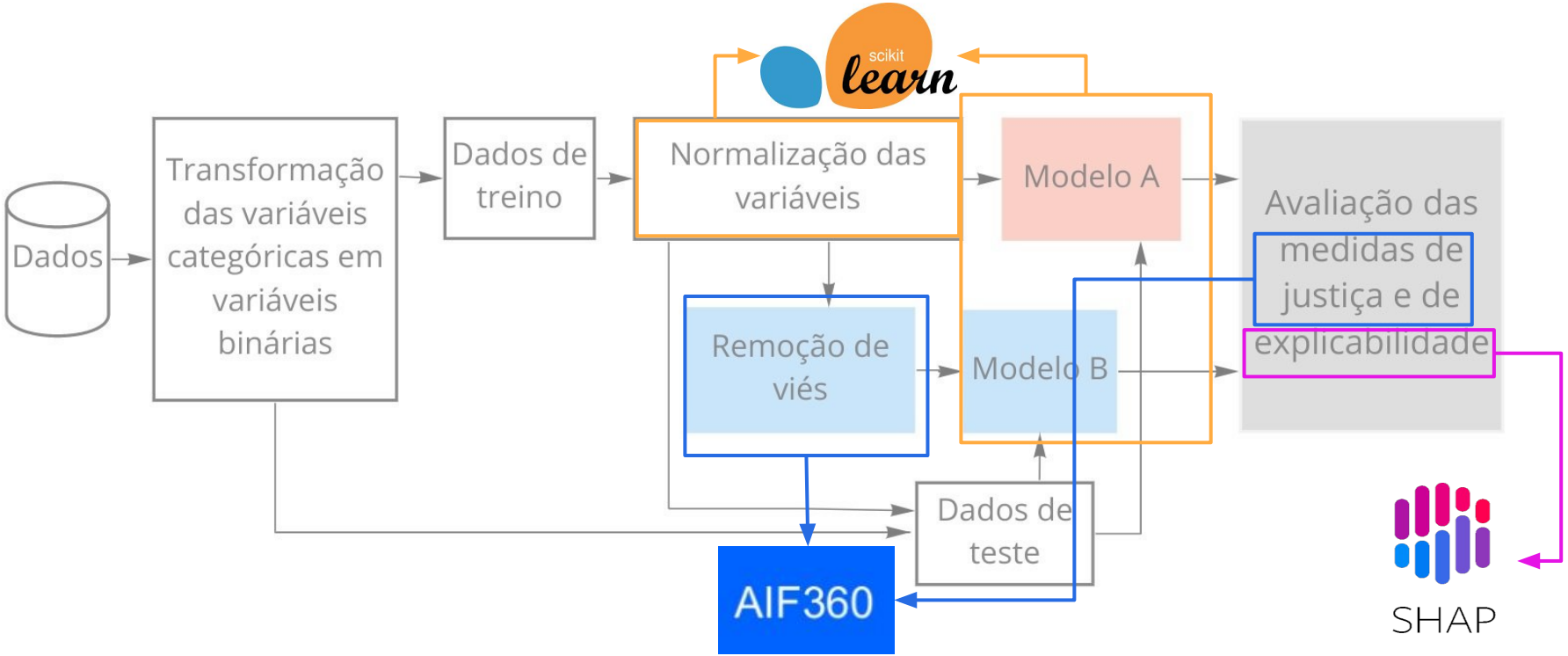
Considerações durante o desenvolvimento do modelo para evitar discriminação



Considerações durante a modelagem para evitar discriminação



Considerações durante a modelagem para evitar discriminação



Modelos

Regressão logística

Interpretabilidade:
- Kernel SHAP
- Linear SHAP

Random Forest

Interpretabilidade:
- Kernel SHAP
- Tree SHAP

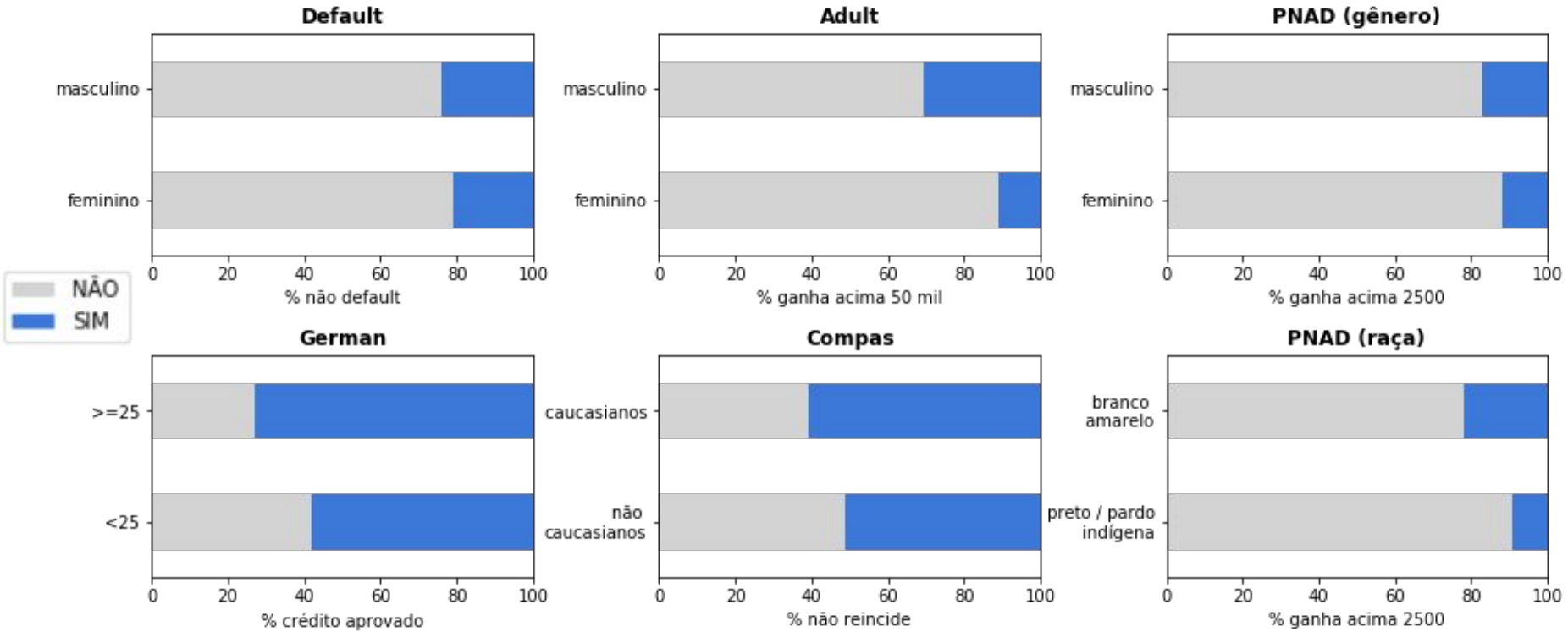
Gradient Boosting

Interpretabilidade:
- Kernel SHAP
- Tree SHAP

SVM

Interpretabilidade:
- Kernel SHAP

Datasets


Default
masculino
feminino
0
20
40
60
80
100
% não default
Adult
masculino
feminino
0
20
40
60
80
100
% ganha acima 50 mil
PNAD (gênero)
masculino
feminino
0
20
40
60
80
100
% ganha acima 2500
NÃO
SIM
German
>=25
<25
0
20
40
60
80
100
% crédito aprovado
Compas
caucasianos
não caucasianos
0
20
40
60
80
100
% não reincide
PNAD (raça)
branco amarelo
preto / pardo indígena
0
20
40
60
80
100
% ganha acima 2500

Avaliação global do resultado com SHAP

Importância do atributo:

$$\frac{1}{N}\sum_{j=1}^{N} \mid \phi_i^{(j)} \mid$$

Importância da feature sensível no modelo

Disparidade do SHAP:

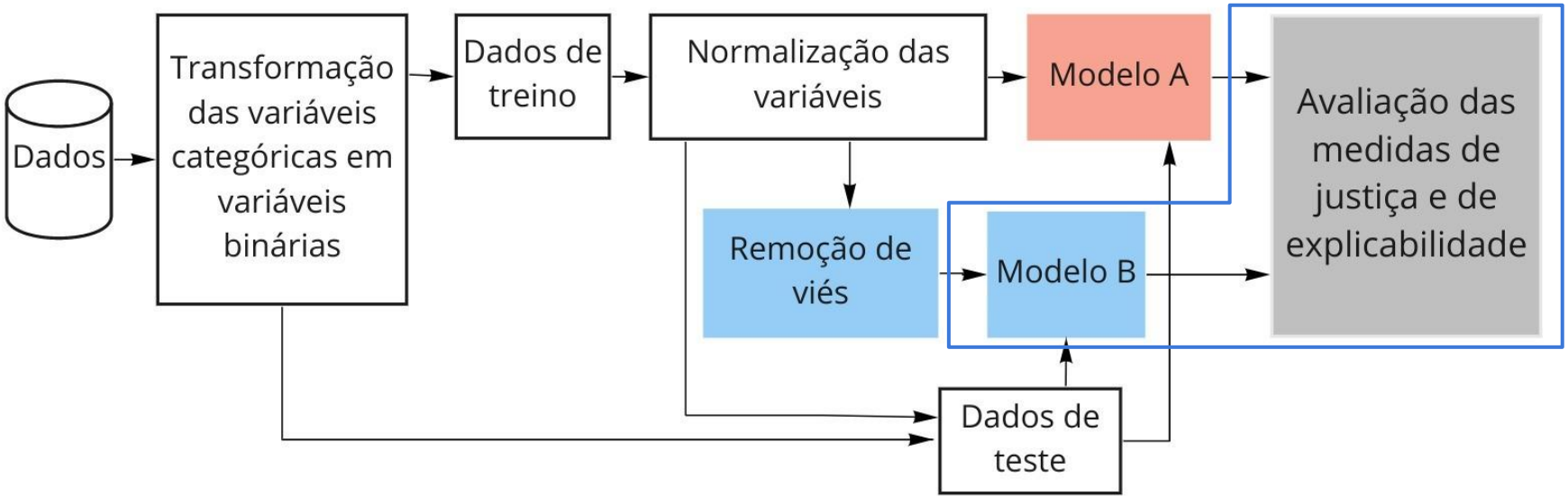
$$\frac{1}{N_k}\sum_{k=1}^{N_k} \phi_i^{(k)} - \frac{1}{N_l}\sum_{l=1}^{N_l} \phi_i^{(l)}$$

Diferença da importância entre grupos desprivilegiado e privilegiado

Avaliação do resultado por cenário

		Igualdade entre os grupos	Favorecimento do grupo privilegiado	Favorecimento do grupo desprivilegiado
$P(\hat{Y}=1\|A=0)-P(\hat{Y}=1\|A=1)$	Paridade estatística	0	> 0	< 0
$P(\hat{Y}=1\|A=0,Y=1)-P(\hat{Y}=1\|A=1,Y=1)$	Igualdade de oportunidade	0	> 0	< 0
$P(\hat{Y}_{A\leftarrow 0}\|X=x,A=1)-P(\hat{Y}_{A\leftarrow 1}\|X=x,A=1)$	Contrafactual	0	> 0	< 0
$\frac{1}{N}\sum_{n=1}^{N}\left\|\hat{y}_n-\frac{1}{k}\sum_{j\epsilon kNN(X')}\hat{y}_j\right\|$	1 - consistência	0	≠ 0	≠ 0
$\frac{1}{n\alpha(\alpha-1)}\sum_{i=1}^{n}\left[\left(\frac{b_i}{\mu}\right)^{\alpha}-1\right]$	Índice de entropia gen.	0	≠ 0	≠ 0
$\frac{1}{N_k}\sum_{k=1}^{N_k}\phi_i^{(k)}-\frac{1}{N_l}\sum_{l=1}^{N_l}\phi_i^{(l)}$	Disparidade do SHAP	0	> 0	< 0
$\frac{1}{N}\sum_{j=1}^{N}\|\phi_i^{(j)}\|$	Importância da feature	0	≠ 0	≠ 0

Avaliação do resultado por cenário



Resultado por cenário

Comparação com Disparidade do SHAP

	Igualdade entre os grupos		Favorecimento do grupo privilegiado		Favorecimento do grupo desprivilegiado	
	Condizente	Contradizente	Condizente	Contradizente	Condizente	Contradizente
Paridade estatística	2	7	30	72	23	10
Igualdade de oportunidade	4	4	26	54	34	22
Contrafactual	**32**	7	**35**	5	**61**	4

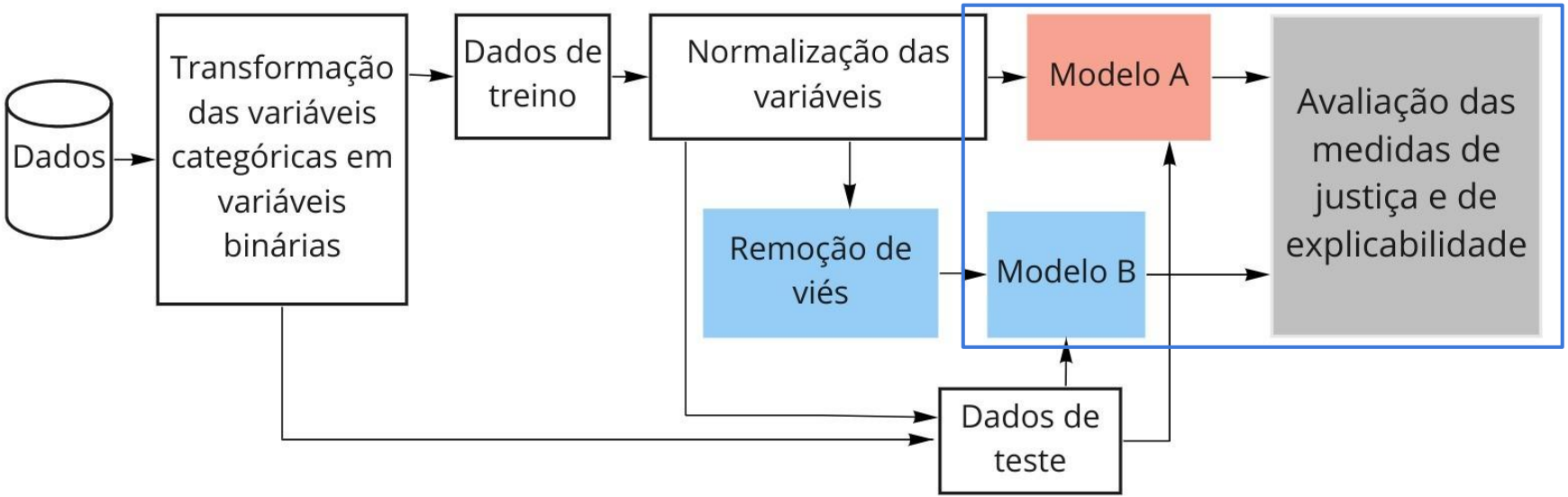
	Igualdade entre os grupos		Desigualdade entre os grupos	
	Condizente	Contradizente	Condizente	Contradizente
1 - consistência	0	0	103	41
Índice de entropia gen.	0	0	103	41

Comparação com importância do atributo

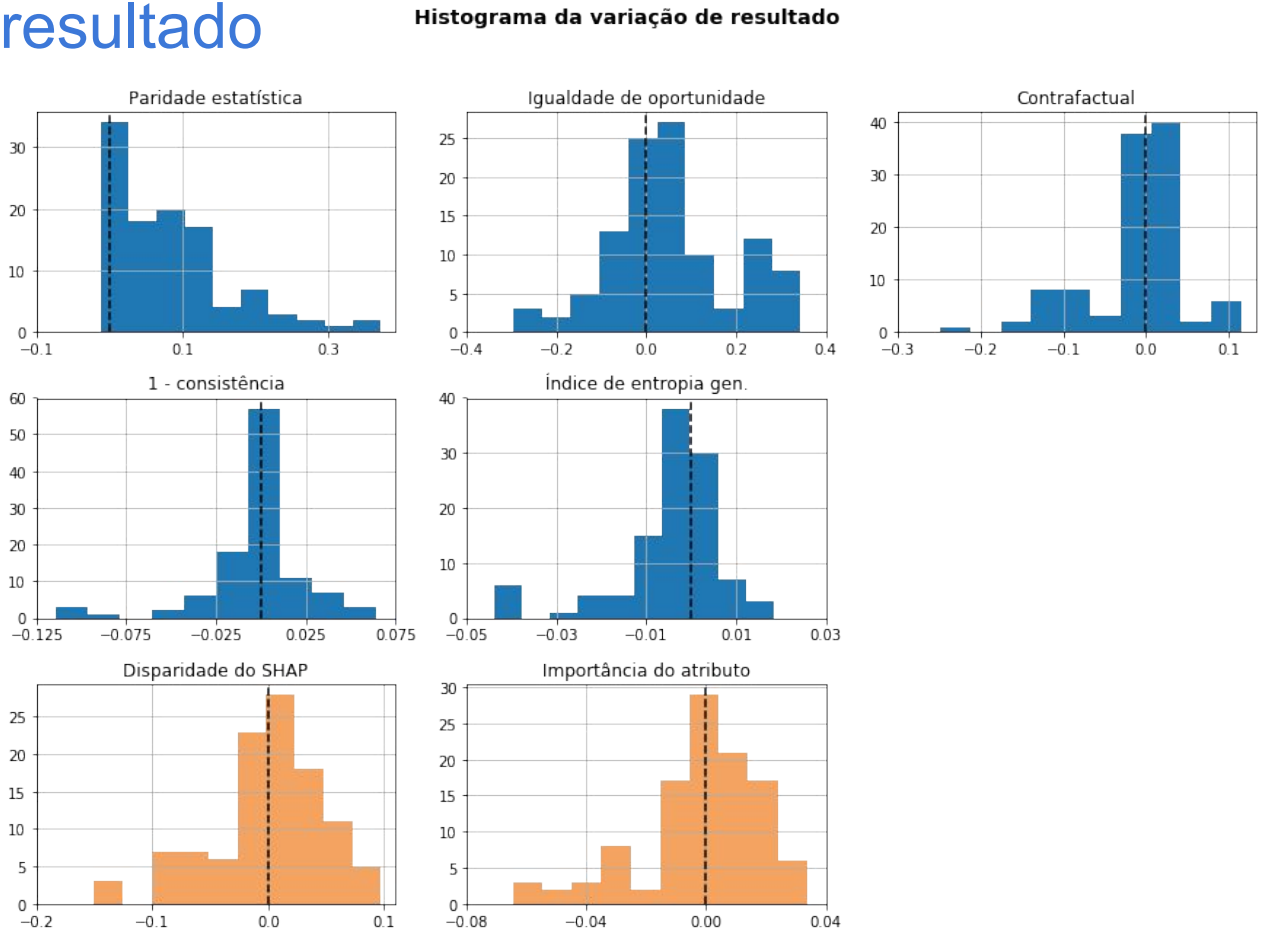
	Igualdade entre os grupos		Desigualdade entre os grupos	
	Condizente	Contradizente	Condizente	Contradizente
Paridade estatística	3	6	81	54
Igualdade de oportunidade	6	2	85	51
Contrafactual	**39**	0	**87**	18
1 - consistência	0	0	87	57
Índice de entropia gen.	0	0	87	57

Avaliação da variação de resultado

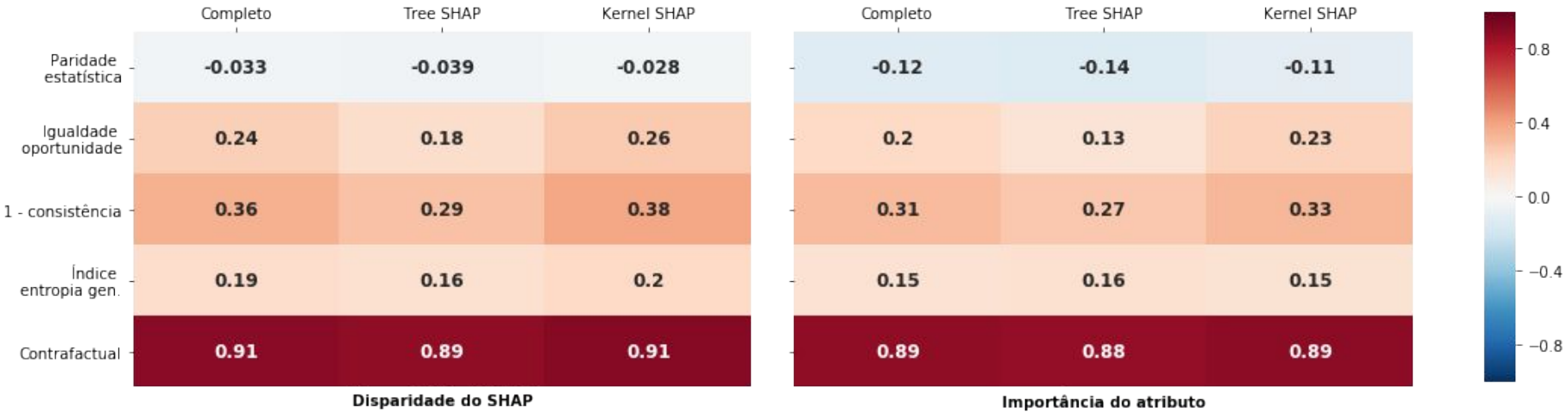
$$|medida_i(modelo_A)| - |medida_i(modelo_{B_j})|$$



Variação de resultado



Variação de resultado



Conclusão

- Justiça contrafactual apresenta grande consenso com resultados do SHAP
- Paridade estatística apresenta maior divergência com resultado do SHAP

- Resultados obtidos com medidas de justiça podem não ser consistente com resultados de interpretabilidade

Obrigada!