JULIANA CESARO

AVALIAÇÃO DE DISCRIMINAÇÃO EM APRENDIZAGEM DE MÁQUINA USANDO TÉCNICAS DE INTERPRETABILIDADE

São Paulo
2020

JULIANA CESARO

AVALIAÇÃO DE DISCRIMINAÇÃO EM APRENDIZAGEM DE MÁQUINA USANDO TÉCNICAS DE INTERPRETABILIDADE

Versão revisada

Dissertação apresentada à Escola Politécnica da Universidade de São Paulo para obtenção do Título de Mestre em Ciências.

Área de Concentração:
Engenharia da Computação

Orientador:
Fabio Gagliardi Cozman

São Paulo
2020

Catalogação-na-publicação

Cesaro, Juliana
Avaliação de Discriminação em Aprendizagem de Máquina usando Técnicas de Interpretabilidade / J. Cesaro, F. Cozman -- São Paulo, 2020.
76 p.

Dissertação (Mestrado) - Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia de Computação e Sistemas Digitais.

1.inteligência artificial 2.discriminação 3.teoria dos jogos I.Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia de Computação e Sistemas Digitais II.t. III.Cozman, Fabio

AGRADECIMENTOS

Primeiro gostaria de agradecer ao meu orientador Fabio Cozman por ter aceitado conduzir o meu trabalho de pesquisa. Obrigada pela compreensão e paciência durante todos esses anos, e por sempre ter conseguido me dar respostas rápidas mesmo com sua intensa rotina acadêmica.

Aos membros da banca, Hugo Neri e Marisa Vasconcelos, por aceitarem participar desta dissertação, e pelos pertinentes apontamentos e sugestões que contribuíram para a melhoria da pesquisa.

Ao Luis Moneda pela valiosa ideia de usar interpretabilidade para avaliar viés em modelo. Além de ter sido o pontapé inicial para essa pesquisa, foi um tema que me interessou muito e serviu como motivação para concluir a pesquisa.

Ao Allan Dieguez pelo incentivo no mestrado e pelas contribuições no artigo.

A minha mãe Terezinha Mós, a quem devo grande parte de minhas conquistas, e ao meu pai Decio Cesaro, que não pode participar deste momento tão especial da minha vida, mas a quem também devo muito pelas minhas realizações.

Agradecimento profundo e especial ao meu marido Rafael Vital, que além de todo o apoio e amor dado durante esses anos difíceis do mestrado, inúmeras vezes contribui de forma significativa com observações na pesquisa.

Por fim, agradeço a todos aqueles que contribuíram, direta ou indiretamente, para a conclusão deste trabalho.

RESUMO

Preconceitos presentes na sociedade podem criar vieses em modelos aprendidos a partir de dados. Para avaliar a existência de viés, alguns pesquisadores propõem o uso de definições de "justiça", enquanto outros usam técnicas de interpretabilidade. Porém, parece não existir nenhum estudo que compara as medidas de justiça (através de várias definições de justiça) e os resultados de interpretabilidade (através de várias noções de interpretabilidade). Nesse trabalho foi proposto metodologias para examinar e comparar essas técnicas. A ideia é avaliar como as medidas de justiça e o resultado de interpretabilidade variam em um modelo com viés e em outro sem viés. O foco foi no uso do SHAP (SHapley Additive exPlanations) como técnica de interpretabilidade, que usa conceito da teoria dos jogos cooperativos para calcular a contribuição de cada atributo em uma previsão gerada pelo modelo; foi apresentado resultados com alguns conjuntos de dados propensos a injustiça. Com os experimentos foi identificado qual a medida de justiça tem relação alta e baixa com o resultado do SHAP, o que auxiliaria a decidir quando é recomendável usar o SHAP como técnica de interpretabilidade ou quando é melhor usar outra técnica.

Palavras-Chave – Interpretabilidade, discriminação, SHAP, valor de Shapley, importância de atributo.

ABSTRACT

Prejudices present in society and in data can introduce biases in a model. In order to evaluate the presence of bias in machine learning, some proposals use fairness measures, while others use interpretability techniques. However, there seems to be no study that compares fairness measures (across various definitions of fairness) and interpretability results (across various notions of interpretability). In this work, we propose ways to evaluate and compare such notions. The idea is to evaluate how fairness measures and interpretability results vary in a model with bias and another one without bias. We focus in particular on SHAP (SHapley Additive exPlanations) as the interpretability technique, which uses cooperative game theory concepts to calculate each feature contribution in a forecast generated by the model; we present results for a number of unfairness-prone datasets. The experiments allow us to identify the fairness measures with high and low connection with SHAP results, which would help decide when it is recommended to use SHAP as an interpretability technique or when it is better to use another technique.

Keywords – Interpretability, fairness, SHAP, Shapley, feature importance.

SUMÁRIO

1 INTRODUÇÃO

O uso de aprendizado de máquina vem se expandindo para uma vasta gama de aplicações, incluindo decisões de crédito, justiça criminal e seleção de candidatos em um processo seletivo. Muitas dessas aplicações envolvem questões éticas, onde as principais preocupações são [41]:

- Discriminação e viés: modelos de aprendizado de máquina podem reproduzir e amplificar padrões de discriminação existentes na sociedade. Um exemplo disso foi o algoritmo desenvolvido pela Amazon de seleção de candidatos para vagas na área de tecnologia que favorecia homens na seleção. O algoritmo ficou enviesado porque foi treinado com dados históricos de seleção de candidatos onde havia predominância na contratação masculina para a área, conforme relato de Jeffrey Dastin em [22], onde constam os detalhes e dados.

- Previsões inexplicáveis ou injustificáveis: algoritmos que fornecem explicação aumentam a confiança do usuário em relação ao resultado gerado. Além disso, permitem que o afetado pela decisão do algoritmo recorra caso se sinta injustiçado. Um caso polêmico foi a condenação de Eric Loomis a seis anos de prisão auxiliada pelo resultado de um algoritmo chamado COMPAS, que calcula o risco da pessoa voltar a reincidir. Loomis foi condenado porque estava dentro de um carro em que um dos passageiros disparou tiros durante uma perseguição policial, porém Loomis nega ter sido o responsável pelos disparos. Quando Loomis foi apreendido teve que responder a um questionário, que junto com seu histórico foi usado pelo algoritmo do COMPAS para determinar a sentença. O resultado foi que ele teria alta probabilidade de reincidir, por isso foi associado a uma das sentenças mais graves. Loomis discordou desta sentença, porém não teve como recorrer porque o algoritmo não era capaz de gerar explicação do resultado gerado. Este caso foi relatado por Liu et al. na Ref. [43].

- Previsões inseguras: o resultado de um algoritmo pode ser inseguro tanto por pro-

blemas de design na construção do sistema, ou porque o algoritmo foi treinado em um ambiente diferente ou não previsto do utilizado na prática. Esse foi o caso de um chatbot chamado Tay, construído pela Microsoft para interagir com jovens entre 18 e 24 anos, que em menos de 24 horas após o lançamento começou a dar respostas racistas, sexistas e de genocídio. O objetivo do chatbot Tay era analisar e aprender com mensagens de tweets enviadas para ele, porém a Microsoft não previu o envio intencional de mensagens inapropriadas e ofensivas. Este caso foi relatado por Jane Wakefield na Ref. [64], onde constam os detalhes do caso.

- Invasão de privacidade: em muitas aplicações dados pessoais são extraídos sem o consentimento apropriado do titular. Um modelo que utiliza esses dados pode infringir o poder de escolha do usuário em ter suas decisões moldadas por influência de determinada tecnologia.

Estas preocupações com comportamento ético tem motivado a discussão crescente em entidades governamentais, na academia, e na indústria. Entidades governamentais de diversos países têm discutido a IA (inteligência artificial) e seus impactos éticos na sociedade, com destaque na atuação do parlamento Europeu com o lançamento de um guia sobre o assunto [49].

Na literatura, diversos artigos e livros sobre ética em IA têm sido publicados [20,36,41]. Muitas conferências recentes na área de aprendizado de máquina têm dado destaque ao tema, como a NeurIPS (Neural Information Processing Systems) de 2019 [57], e em outras foi o tema principal, com destaque para a FAccT (Fairness, Accountability, and Transparency) que acontece desde 2018.

Na indústria, casos de modelos com viés como os citados anteriormente vem causando grande aversão do público. Isso tem feito muitas empresas se preocuparem mais com os possíveis impactos éticos dos projetos de IA; além dessa mobilização interna, diversas empresas têm atuado externamente. Uma das formas de mobilização têm sido com a organização de eventos e divulgação sobre o assunto; nessa frente podemos citar a associação "Partnership on AI" [53] que conta com a colaboração de diversas empresas, como Google, Amazon, Facebook, IBM e Microsoft.

Além disso, grandes empresas têm disponibilizado diversas ferramentas abertas ao público para auxiliar o desenvolvimento de um projeto de IA ético. Uma delas é a interface interativa da Google "What if tool" [65], que permite gerar diversos gráficos para interpretação das variáveis do modelo e avaliação de viés. Outra é um conjunto de três bibliotecas disponibilizados pela IBM: "AIF-360" [10] para identificação e mitigação de

viés no modelo, “AIX-360” [6] para a explicabilidade do modelo, e a “ART” [52] para testar o modelo contra ataques adversários auxiliando dessa forma a construção de um modelo mais robusto e seguro.

Para o desenvolvimento de um sistema de IA ético, os temas mais abordados são explicabilidade, transparência e justiça das decisões geradas. Sendo que quando o sistema impacta diretamente a vida de pessoas há maior preocupação com discriminação de certos grupos, principalmente se dentre os grupos afetados existir algum que historicamente foi discriminado ou injustiçado.

Em situações onde há preocupação com discriminação, na etapa de modelagem é necessário escolher um tipo de definição de justiça para avaliar se o resultado do modelo é justo para os diversos grupos presentes. Além disso, é necessário investir na interpretabilidade do comportamento do modelo e das influências dos atributos.

O aumento da interpretabilidade dá mais confiança no modelo tanto para quem o implementa, como para o usuário que é afetado pela sua decisão com o fornecimento de uma explicação para o resultado gerado. Para a escolha da definição de justiça não existe uma regra, ela é definida de acordo com o escopo do projeto. Enquanto que para a escolha da metodologia de explicabilidade o ideal é usar a melhor metodologia existente para a aplicação.

Com o uso de modelos menos complexos que são mais facilmente interpretáveis, como regressão linear e árvore de decisão, é possível gerar diretamente explicação do resultado [51]. Por isso, normalmente é mais recomendável iniciar a modelagem com esses modelos mais simples que além de serem diretamente interpretáveis, costumam ser mais rápidos de serem implementados e serviriam como baseline para os modelos mais complexos.

Entretanto, em contexto onde há grande volume de dados e variáveis, modelos mais complexos costumam ser melhores. Por exemplo, para problemas de classificação de imagem costuma ser obtido melhor acurácia com o uso de rede neural profunda com camadas convolucionais, que é um modelo altamente não interpretável [51]. Nesse contexto que é melhor usar um modelo mais complexo e é necessário a explicabilidade do seu resultado, é possível usar técnicas de interpretabilidade externa ao modelo para explicar o seu resultado.

Existem diversas técnicas para explicar o resultado de um modelo de aprendizado de máquina [9,44,48,56]. Algumas são técnicas globais que visam entender o modelo como um todo, outras são técnicas locais usadas para explicar previsões específicas. O foco aqui será em técnicas locais, pois o interesse é conseguir explicar um resultado específico gerado, por

exemplo, para alguém que se sentiu injustiçado pela decisão do modelo. Nessa dissertação o foco será no uso do SHAP, que atualmente é uma técnica local de interpretabilidade considerada referência. O SHAP usa conceitos da teoria dos jogos cooperativos para calcular a contribuição de cada atributo em uma previsão gerada pelo modelo.

Apesar de existirem diversas técnicas de interpretabilidade, ainda não existe nenhum formalismo para a avaliação da explicação gerada e definição de requisitos que essas explicações devem atender [29]. Assim como não existe nenhum benchmark para avaliar e comparar técnicas de interpretabilidade. Tudo isso dificulta a escolha da técnica de interpretabilidade para ser usada.

Além disso, a explicação gerada com certa técnica de interpretabilidade poderia contradizer o resultado de outras métricas usadas para avaliar o modelo. Por exemplo, em um problema em que existe preocupação com discriminação, ainda não sabemos se o resultado obtido com medidas de justiça refletem o mesmo resultado da técnica de interpretabilidade. Apesar de haverem estudos que usam técnicas de interpretabilidade para avaliar se o modelo é discriminatório [2,61,63], parece não existir nenhum estudo comparativo para avaliar a consistência entre resultados gerados pelas técnicas de interpretabilidade e pelas métricas de justiça.

O objetivo dessa dissertação é propor um framework para comparar técnicas de interpretabilidade e de justiça. As contribuições dessa dissertação são:

- proposta de um framework que permita avaliar se as medidas obtidas com técnicas de interpretabilidade refletem medidas de justiça, baseado na comparação entre os resultado quando o modelo é treinado com dados enviesados e dados com redução de viés;
- aplicação do framework em um estudo com cinco conjuntos de dados enviesados e quatro tipos de modelos.

Uma parte dos resultados dessa dissertação apareceu no artigo [18], onde foi usado o framework proposto para comparar as medidas de justiça e do SHAP em um cenário mais restrito em relação a variedade de conjunto de dados, modelos, técnicas de mitigação de viés e medidas de justiça.

O Capítulo 2 contém definições de justiça que foram feitas ao longo da história, e as definições matemáticas que vem sendo propostas para avaliar se um modelo é justo. Além disso, nele é descrito algumas metodologias de mitigação de viés do modelo. No Capítulo 3 são apresentadas algumas técnicas de interpretabilidade, com destaque na descrição do

SHAP. O Capítulo 4 contém a especificação do funcionamento do framework proposto para comparar as medidas de justiça e do SHAP, e o detalhamento das métricas que serão utilizadas nessa comparação. No Capítulo 5 é feita a caracterização da metodologia que será utilizada para a implementação do framework proposto, e a descrição dos conjuntos de dados que serão utilizados para avaliar os resultados obtidos. No Capítulo 6 as conclusões finais são retratadas.

2 DISCRIMINAÇÃO E JUSTIÇA EM APRENDIZAGEM DE MÁQUINA

O conceito de justiça pode ter diferentes significados em diferentes contextos. Uma definição que pode ser considerada base [50] vem da lei romana das Institutas de Justiniano do século VI d.C., que define justiça como "a constante e perpétua vontade de dar a cada um o que é seu". Nessa definição "constante e perpétua" requer imparcialidade e consistência na aplicação da lei: dois casos relativamente similares deveriam ser tratados da mesma forma.

A necessidade de justiça surge em reivindicações, pedindo por exemplo por oportunidade ou recursos, e para resolver tal reivindicação a justiça determina o que cada pessoa tem direito a ter, que na definição de Justiniano é retratado por "dar a cada um o que é seu". Essas reivindicações por justiça podem ser feitas por um indivíduo ou por um grupo.

Na literatura existem duas visões distintas de justiça, uma conservadora que defende que justiça acontece quando a lei é aplicada corretamente, enquanto outra pedindo por reforma nas normas para um regime de justiça ideal de distribuição, que requer que as pessoas sejam igualmente beneficiadas [50]. John Rawls é um dos defensores da justiça ideal, e na sua obra *Uma Teoria da Justiça* de 1971 [55] aborda o problema da distribuição dos direitos, apresentando dois princípios fundamentais de justiça: o da liberdade e o da igualdade democrática.

No princípio da liberdade, Rawls defende que todos devem ter o mesmo direito para a liberdade básica igual. Rawls defende que as desigualdades econômicas devem ser consideradas de modo a beneficiar os menos favorecidos da sociedade, e estar vinculadas a posições e cargos acessíveis a todos em condições de igualdade de oportunidades.

Esta preocupação com justiça têm atraído cada vez mais atenção da área de aprendizagem de máquina, pois sem os devidos cuidados na construção do modelo, resultados podem prejudicar certos grupos. Um das causas disso vem de preconceitos presente na

sociedade e nos dados, que leva a construção de modelos "enviesados". Um caso famoso foi do algoritmo COMPAS, que tem como objetivo a previsão de reincidência criminal. Ao fazer um estudo da assertividade do algoritmo, foi verificado que a taxa de erro para negros era em torno de duas vezes maior do que a dos brancos [4].

Esse tipo de discriminação não é permitido pela Constituição Federal, pois constituem como objetivos fundamentais da República Federativa do Brasil: "promover o bem de todos, sem preconceitos de origem, raça, sexo, cor, idade e quaisquer outras formas de discriminação". Dessa forma, para evitar preconceito, decisões geradas por modelos de aprendizado de máquina não deveriam depender dessas características [14].

Uma tentativa para evitar discriminação nos resultados dos modelos é a proibição do uso de dados sensíveis de uma pessoa prevista na LGPD (Lei Geral de Proteção de Dados) [14]. Sendo que no Artigo 5º parágrafo II os seguintes dados são definidos como sensíveis: "dado pessoal sobre origem racial ou étnica, convicção religiosa, opinião política, filiação a sindicato ou a organização de caráter religioso, filosófico ou político, dado referente à saúde ou à vida sexual, dado genético ou biométrico, quando vinculado a uma pessoa natural". Existem alguma exceções na regra, como para o caso de procedimentos realizados por profissionais da saúde, ou quando a pessoa afetada pela decisão do algoritmo explicitamente autoriza o uso do dado sensível para a aplicação de uso, conforme descrito no Artigo 7º da LGPD.

Entretanto, apenas a não utilização da variável sensível pode não ser suficiente para a construção de um algoritmo justo. Pois nos dados podem existir variáveis que são dependentes da variável sensível removida [54], em alguns casos uma série de variáveis com correlação fraca com a variável sensível, quando usadas em conjunto podem ser suficiente para fornecer relevante informação ao modelo da variável sensível. Esse efeito é retratado no trabalho de Pamela et al. [15], onde apenas com a informação de texto e imagens feitas por usuários do Pinterest foi possível prever com alta acurácia o gênero dos usuários.

Pela dificuldade em avaliar se um algoritmo é justo, diversas definições matemáticas foram propostas. Porém, apesar do estudo crescente na área, ainda não há um consenso sobre as definições que vêm sendo propostas. Verma e Rubin [62] fazem uma revisão da literatura, e mostram que existem mais de vinte definições de justiça definidas nos últimos tempos, sendo possível obter classificações contradizente de justiça de acordo com a definição usada. Hutchinson e Mitchell [31] fazem um estudo das definições quantitativas de justiça usadas nos últimos 50 anos, e como elas se relacionam com as definições atuais

usadas em problemas de aprendizado de máquina. Na filosofia política, o tópico de discriminação é de grande interesse, por isso existe um esforço significativo em seu formalismo. Aproveitando disso, o estudo abordado por Binns em [12] traz conceitos de justiça da filosofia política para aplicar no contexto de aprendizado de máquina.

2.1 Definições de justiça

As definições de justiça podem ser divididas em dois principais tipos: justiça entre grupos e justiça entre indivíduos. Essas definições quantificam a relação entre grupos "privilegiado" e "desprivilegiado" para determinado atributo sensível. Nessa dissertação iremos focar em problemas em que a variável sensível contem apenas os grupos privilegiado e desprivilegiado.

Representaremos a variável sensível por um atributo binário A no modelo, atribuindo valor 0 ao grupo desprivilegiado e o valor 1 ao grupo privilegiado. Além disso, focaremos em problemas de classificação binária, em que a variável alvo Y pode assumir valores $\{0, 1\}$, sendo que o valor 1 indica a classe desejável, como um bom score de crédito. Representaremos o valor previsto da variável alvo por $\hat{Y}$.

2.1.1 Justiça entre grupos

Justiça entre grupos é obtida quando o grupo privilegiado e o desprivilegiado são tratados da mesma forma. Um conceito comumente usado é o de *paridade estatística* (statistical parity) ou *paridade demográfica* (demographic parity) [17, 71], que determina que um classificador é justo se sujeitos do grupo privilegiado e desprivilegiado têm probabilidades iguais de receberem uma determinada classificação. Essa definição é representada pela formulação abaixo:

$$P(\hat{Y} = 1|A = 0) = P(\hat{Y} = 1|A = 1). \tag{2.1}$$

Uma definição equivalente porém menos rigorosa, é a de *efeito da disparidade* (disparate impact) [26], em que um classificador é considerado injusto quando:

$$\frac{P(\hat{Y} = 1|A = 0)}{P(\hat{Y} = 1|A = 1)} \leq \tau. \tag{2.2}$$

Conforme essa definição, um classificador é injusto quando a razão entre as probabilidade de previsão da classe desejada entre o grupo desprivilegiado e o grupo privilegiado

é menor que τ, a ser definido de acordo com o problema. O caso de uso citado no artigo que propõe essa medida [26], é o atendimento à recomendação da EEOC (US Equal Employment Opportunity Commission), em que as empresas devem considerar uma taxa de seleção de grupos minoritários de pelo menos 80% da taxa do grupo majoritário. Por isso o valor adotado para τ é de 0.8.

Ao tentar tornar o modelo justo usando as definições de paridade estatística e de efeito da disparidade, pode ser introduzido um erro considerável no resultado. Pois ambas as definições não levam em consideração o valor real da variável alvo Y. A fim de amenizar esse efeito, outra definição propõe em além de comparar a previsão gerada para os diferentes grupos da variável sensível A, fazer essa comparação considerando também o valor de Y [30]. A definição de *probabilidade equalizada* (equalized odds), descrita pela Equação (2.3), usa esse conceito:

$$P(\hat{Y} = 1|A = 0, Y = y) = P(\hat{Y} = 1|A = 1, Y = y), \tag{2.3}$$

onde y pode ter os valores $\{0, 1\}$. Para o caso em que $y = 1$ a condição requer a igualdade da taxa de verdadeiros positivos, e para o caso em que $y = 0$ requer a igualdade da taxa de falso positivos.

A definição de probabilidade equalizada ao requerer igualdade quando $y = 0$ e $y = 1$, pode punir modelos que performam bem na maioria dos casos [30]. A fim de evitar isso, uma possibilidade de definição menos rígida é requerer igualdade apenas para o caso da classe desejável, que definimos como $y = 1$. Essa definição é chamada de *igualdade de oportunidade* (equality of opportunity):

$$P(\hat{Y} = 1|A = 0, Y = 1) = P(\hat{Y} = 1|A = 1, Y = 1). \tag{2.4}$$

2.1.2 Justiça entre indivíduos

Metodologias que buscam equalizar a relação entre grupos podem aumentar a injustiça entre indivíduos. Considerando o exemplo de seleção de candidatos para uma vaga de emprego, ao tentar aproximar a relação entre grupos, candidatos menos qualificados para a vaga podem ser selecionados. Para evitar isso, o ideal é usar definições de justiça que levam em consideração características em comum dos indivíduos.

Uma das primeiras definições com esse intuito foi a de *justiça através do conhecimento*

(fairness through awareness) [25], em que um classificador é considerado justo se:

$$D(M(X_1), M(X_2)) \leq d(X_1, X_2), \tag{2.5}$$

sendo X_i os indivíduos que vão ser classificados, M uma função que mapeia indivíduos para distribuição de probabilidade prevista ($M : X \rightarrow \Delta(\hat{Y})$). d e D são métricas para medir similaridade entre indivíduos e da distribuição prevista respectivamente. A métrica d é definida de acordo com a necessidade do problema. Ao mesmo tempo que isso dá maior flexibilidade, também envolve maior dificuldade de formulação da métrica.

A fim de eliminar a necessidade de definição de uma métrica de similaridade entre indivíduos, Zemel et al. [68] propõe uma medida de justiça chamada *consistência* (consistency), que através de padrões nos dados classifica indivíduos semelhantes. Consistência é definida pela Equação (2.6); a métrica compara a previsão do modelo para um indivíduo n com a média das previsões dos k indivíduos mais semelhante, obtidos através do k-nearest neighbors, $kNN(X)$:

$$consistência = 1 - \frac{1}{N}\sum_{n=1}^{N}\left|\hat{y}_n - \frac{1}{k}\sum_{j \epsilon kNN(X')}\hat{y}_j\right|. \tag{2.6}$$

Uma observação é que caso a variável sensível seja utilizada como entrada do modelo, é necessário removê-la para o cálculo do kNN. Por isso denotamos X' sendo X com remoção da variável sensível.

2.1.3 Justiça entre grupos e indivíduos

Anteriormente descrevemos definições que avaliam justiça entre grupos e justiça entre indivíduos. Apesar destes conceitos comumente serem definidos de forma separada, em diversos casos existe a necessidade de garantir simultaneamente ambos os tipos de justiça [60]. Nesse tipo de situação pode ser complexo construir um modelo que garanta um balanço ótimo entre mais de uma medida de justiça e a acurácia. Uma solução mais efetiva nesse caso é utilizar algum tipo de medida que avalie simultaneamente ambos os tipos de justiça.

Uma definição com esse intuito é a de *índice de entropia generalizado* (generalized entropy index) [60], que usa conceitos de índices de desigualdade que já foram extensivamente estudados em Economia e em Ciências Sociais para avaliar o quão desigual é a distribuição de renda entre grupos e indivíduos em uma população.

Para o uso desta definição deve ser definida uma função de benefício b_i para avaliar se um indivíduo i é beneficiado pelo modelo, que depende do contexto de aplicação. Speicher et al. [60] sugere o uso da Equação (2.7) para o cálculo da função de benefício b_i em problemas de classificação binário. Com essa equação o maior valor ocorreria quando $b_i = 2$ com falsos positivos (indivíduos erroneamente beneficiados pelo resultado modelo), valores moderados aconteceriam com $b_i = 1$ com verdadeiros positivos e verdadeiros negativos (indivíduos que receberam a classificação correta), e o menor valor ocorreria com $b_i = 0$ com os falsos negativos (indivíduos erroneamente prejudicados pelo resultado modelo):

$$b_i = \hat{y}_i - y_i + 1. \tag{2.7}$$

Com a definição da função de benefício que avalia o efeito da justiça entre indivíduos é possível obter o benefício para um conjunto de indivíduos pertencentes a certo grupo, e dessa forma calcular o efeito da justiça entre grupos. Este calculo é feito pela Equação (2.8), através da média dos benefícios recebidos pelos indivíduos de um grupo g:

$$\mu_g = \frac{1}{|g|} \sum_{i \in g} b_i. \tag{2.8}$$

Utilizando o resultado da função de benefício é possível avaliar a justiça global de um modelo com índices de desigualdade, sendo proposto o uso do índice de entropia generalizado definido pela Equação (2.9):

$$\varepsilon^{\alpha}(b_1, b_2, ..., b_n) = \frac{1}{n\alpha(\alpha - 1)} \sum_{i=1}^{n} \left[\left(\frac{b_i}{\mu} \right)^{\alpha} - 1 \right], \tag{2.9}$$

sendo α uma constante diferente de 0 e 1. Com esta definição quanto menor o resultado obtido mais justo é considerado o modelo.

Alguns índices de desigualdade, como o índice de entropia generalizado, garantem a propriedade da decomposição de subgrupos, que permite reescrever o seu cálculo pela soma de uma componente que captura a relação dentro do grupo e uma componente que captura a relação entre grupos. Para o contexto de avaliação da justiça de um modelo, essa é uma propriedade importante garantida, para permitir a avaliação separada da influência de cada componente. Usando a propriedade da decomposição de subgrupos podemos reescrever a Equação (2.9) como a Equação (2.10):

$$\begin{aligned} \varepsilon^{\alpha}(b_1, b_2, ..., b_n) &= \sum_{g=1}^{|G|} \frac{n_g}{n} \left(\frac{\mu_g}{\mu} \right) \varepsilon^{\alpha}(b^g) + \sum_{g=1}^{|G|} \frac{n_g}{n\alpha(\alpha - 1)} \left[\left(\frac{\mu_g}{\mu} \right)^{\alpha} - 1 \right] \\ &= \varepsilon_w^{\alpha}(b) + \varepsilon_{\beta}^{\alpha}(b), \end{aligned} \tag{2.10}$$

sendo que $\varepsilon_w^\alpha(b)$ captura a relação dentro do grupo e $\varepsilon_\beta^\alpha(b)$ captura a relação entre grupos.

O índice de entropia generalizado pode ser usado em problemas em que existe mais de uma variável sensível, por exemplo raça e gênero, nesse casos existem diversos subgrupos formado pela intersecção das variáveis sensíveis, como de mulheres brancas. Conforme demonstrado por Speicher et al. [60] quanto maior o número desses subgrupos maior a contribuição da componente entre grupos ($\varepsilon_\beta^\alpha(b)$).

2.1.4 Abordagens causais

Outra abordagem são as baseadas em relações causais [28, 37, 38], onde um tipo de definição adotada é a de *justiça contrafactual* (counterfactual fairness) [38] descrito pela Equação (2.11). Ela indica que uma previsão $\hat{Y}$ é justa quando ao mudar o grupo da variável sensível o valor da previsão continua o mesmo:

$$P(\hat{Y}_{A\leftarrow a'}|X = x, A = a) = P(\hat{Y}_{A\leftarrow a}|X = x, A = a), \tag{2.11}$$

sendo que $\hat{Y}_{A\leftarrow a'}$ representa a intervenção aplicada, onde o grupo da variável sensível A foi mudado de a para a'.

Para o caso mais comum em que a variável sensível não é usada no modelo, para avaliar se um modelo é justo é necessário representar graficamente a relação entre as variáveis e verificar se a variável prevista não depende das variáveis descendentes da variável sensível. Enquanto que para o caso em que a variável sensível é utilizada no modelo essa verificação é mais simples, sendo necessário trocar o valor da variável sensível no momento da previsão do modelo já treinado para analisar se a previsão varia.

2.2 Mitigação de viés

O termos viés pode assumir diferentes significados, e é importante esclarecer que nessa dissertação foi utilizado para se referir a discriminação presente no resultado do modelo para certo grupo. As técnicas de amenização de viés de um modelo podem ser divididas em três categorias: pré-processamento dos dados antes do treinamento do modelo [34, 68]; alteração da função de otimização do modelo [35]; e pós-processamento do resultado do modelo [30].

Técnicas de pré-processamento tem como vantagem serem agnósticas ao modelo utilizado, porém não garantem uma boa acurácia do modelo por não terem acesso ao valor

previsto. Técnicas que alteram a função otimização do modelo buscam encontrar uma relação ótima entre a acurácia do modelo e as medidas de justiça, porém a desvantagem é que essas técnicas são específicas para certos modelos e envolvem sua modificação, o que dependendo da situação pode não ser possível. As técnicas de pós-processamento também são agnósticas ao modelo, costumam gerar bons resultados para as métricas de justiça, porém trazem como desvantagem a necessidade do uso da variável sensível após a previsão ser gerada, o que pode não ser possível.

Nessa dissertação serão usadas dois tipos de técnica de pré-processamento dos dados: *reponderação* (reweighing) e *subamostragem parametrizada* (parametrised sampling). Ambas as técnicas têm como vantagem não mudarem o valor das variáveis de entrada do modelo, como o método proposto por Feldman et al. [26], e não limitarem a aplicação a modelos específicos e que envolvem a sua modificação, como o método abordado por Zhang et al. [69].

2.2.1 Reponderação

A técnica de reponderação [34] associa pesos aos dados de entrada do modelo. Para compensar o viés nos dados, pesos maiores são atribuídos aos indivíduos pertencentes ao grupo desprivilegiado. O valor do peso atribuído a cada instância X é calculada pela razão entre a probabilidade esperada P_{esp} e a probabilidade observada P_{obs} para a variável alvo Y e a variável sensível A:

$$W(X) = \frac{P_{esp}(A = a \wedge Y = y)}{P_{obs}(A = a \wedge Y = y)}. \tag{2.12}$$

No caso ideal de um conjunto de dados D sem viés, as variáveis Y e A são independentes. Dessa forma, a probabilidade esperada é:

$$\begin{aligned} P_{esp}(A = a \wedge Y = y) &= P(A = a) \times P(Y = y) \\ &= \frac{|\{X \in D | X(A) = a\}|}{|D|} \times \frac{|\{X \in D | X(Y) = y\}|}{|D|}. \end{aligned} \tag{2.13}$$

Porém em um conjunto de dados enviesado Y e A são dependentes, por isso a probabilidade observada é:

$$P_{obs}(A = a \wedge Y = y) = \frac{|\{X \in D | X(A) = a, X(Y) = y\}|}{|D|}. \tag{2.14}$$

2.2.2 Amostragem parametrizada

A técnica de amostragem parametrizada [67] tem como objetivo amenizar o viés de um modelo através da reamostragem do dado de treino. Para a reamostragem é definido um parâmetro d, que pode assumir valores entre -1 e 1. A flexibilidade na escolha do parâmetro d permite otimizar um classificador para atingir uma definição de justiça específica.

O dado de treino do modelo é definido como D, que contém uma variável sensível A com grupo privilegiado F e grupo desprivilegiado U, sendo a taxa de positivos PR calculada pela razão entre o total de amostras com variável alvo positiva e o total de amostras.

Em um conjunto de dados enviesado a taxa de positivos PR do grupo privilegiado é maior que a do grupo desprivilegiado, conforme representado na Figura 1A. O objetivo da amostragem é definir uma função f^+ para corrigir a taxa de positivos do grupo privilegiado, e uma função u^+ para corrigir a taxa de positivos do grupo desprivilegiado. Essas funções devem atender aos seguintes critérios:

- $d = 1$: dados de treino D sem modificação;
- $d = 0$: grupos F e U com mesma taxa de positivos PR;
- $d = -1$: grupos F e U com taxa de positivos PR completamente reversa.

A função mais simples para atender esses critérios é uma quadrática polinomial com $u^+(d) = f^+(-d)$, definida na Equação (2.15):

$$\begin{aligned} f^+(d) &= ad^2 + bd + c, \\ u^+(d) &= ad^2 - bd + c, \end{aligned} \tag{2.15}$$

sendo:

$$a = \frac{PR(F) + PR(U)}{2} - PR(D), \quad b = \frac{PR(F) - PR(U)}{2}, \quad c = PR(D).$$

Para diferentes técnicas de amostragem descritas no trabalho de Zelaya et al. [67] é usada como base a mesma Expressão (2.15). Conforme mostrado no trabalho, aplicando essas técnicas de amostragem em diferentes conjuntos de dados foram obtidos resultados de acurácia e de medida de justiça muito semelhantes. Por isso nessa dissertação o foco será apenas na técnica de subamostragem. Nela o grupo privilegiado com variável alvo

positiva e o grupo desprivilegiado com variável alvo negativa são aleatoriamente subamostrados.

O Pseudo Código 1 descreve o processo de subamostragem. A Figura 1 representa em A um conjunto de dados enviesado, e em B a transformação que acontece no dado com a subamostragem para o caso em que $d = 0$.

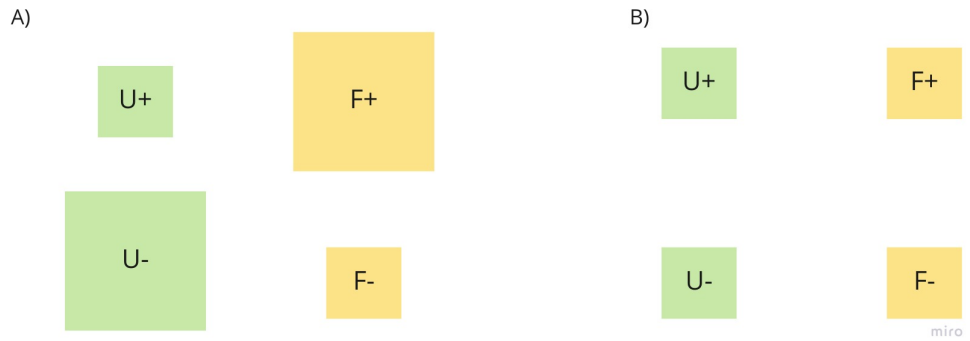


Figura 1: A) Conjunto de dados enviesado onde a quantidade de amostras com variável alvo positivo, $+$, é maior no grupo privilegiado F do que no grupo desprivilegiado U. B) Transformação no tamanho das amostras com o processo de subamostragem que acontece com o grupo privilegiado com variável alvo positiva, $F+$, e com o grupo desprivilegiado com variável alvo negativa, $U-$.

Algorithm 1: Subamostragem parametrizada

Entradas:

$X \leftarrow$ Conjunto de dados de treino

$d \in [-1, 1] \leftarrow$ parâmetro de correção do viés nos dados

Resultado: $X\_corr \leftarrow$ Conjunto de dados de treino corrigido

XF_pos = $X(A = 1 \ \& \ Y = 1)$

XF_neg = $X(A = 1 \ \& \ Y = 0)$

XU_pos = $X(A = 0 \ \& \ Y = 1)$

XU_neg = $X(A = 0 \ \& \ Y = 0)$

PR_F = tamanho[XF_pos] / tamanho[XF_pos + XF_neg]

PR_U = tamanho[XU_pos] / tamanho[XU_pos + XU_neg]

PR_D = tamanho[$X(Y = 1)$] / tamanho[X]

a = ((PR_F + PR_U) / 2) − PR_D

b = (PR_F − PR_U) / 2

c = PR_T

$fpr = a * d * *2 + b * d + c$

$upr = a * d * *2 - b * d + c$

$fpos = fpr \ / \ (1 - fpr)$

$uneg = (1 - upr) \ / \ upr$

$fpossize = fpos * tamanho$(XF_neg)

$unegsize = uneg * tamanho$(XU_pos)

XF_pos = subamostragem(XF_pos, $tam = fpossize$)

XU_neg = subamostragem(XU_neg, $tam = unegsize$)

X_corr = concat(XF_pos, XF_neg, XU_pos, XU_neg)

3 INTERPRETABILIDADE

Interpretabilidade é importante para quem desenvolve um modelo, pois permite um melhor entendimento de como o modelo se comportaria em diversos cenários, e auxilia na criação e seleção dos atributos do modelo. Além disso, interpretabilidade é importante para o usuário final afetado pelo resultado do modelo, para dar mais confiança em relação ao resultado gerado e permitir que o usuário recorra caso se sinta prejudicado.

Alguns modelos são naturalmente interpretáveis, como exemplo uma Regressão Linear que contém um único peso por atributo, e sua previsão é gerada através da multiplicação destes pesos pelos valores do atributo, conforme representado na Figura 2. Por isso, em um modelo linear é fácil entender quais são os atributos que mais contribuem para aumentar ou diminuir a previsão.

Entretanto, em aplicações com grande volume de dados e variáveis podem ser obtidos resultados consideravelmente melhores com o uso de modelos menos interpretáveis, como com conjunto de árvores e redes neurais profundas. O gráfico da Figura 3 sugere uma relação entre interpretabilidade e acurácia de alguns modelos, este gráfico não foi obtido de forma científica, mas da uma noção intuitiva da relação entre interpretabilidade e acurácia para o caso de modelos treinados com dados tabulares com grande quantidade de amostras e variáveis.

Em situações onde são usados modelos que não são naturalmente interpretáveis e

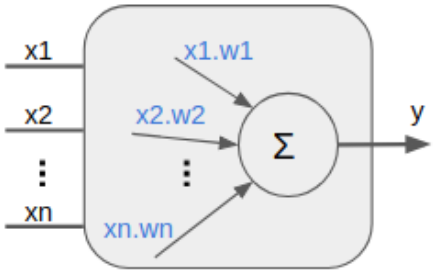


Figura 2: Representação da previsão gerada por um modelo linear, que é obtida pela somatória dos valores de seus atributos multiplicado pelos pesos do modelo.

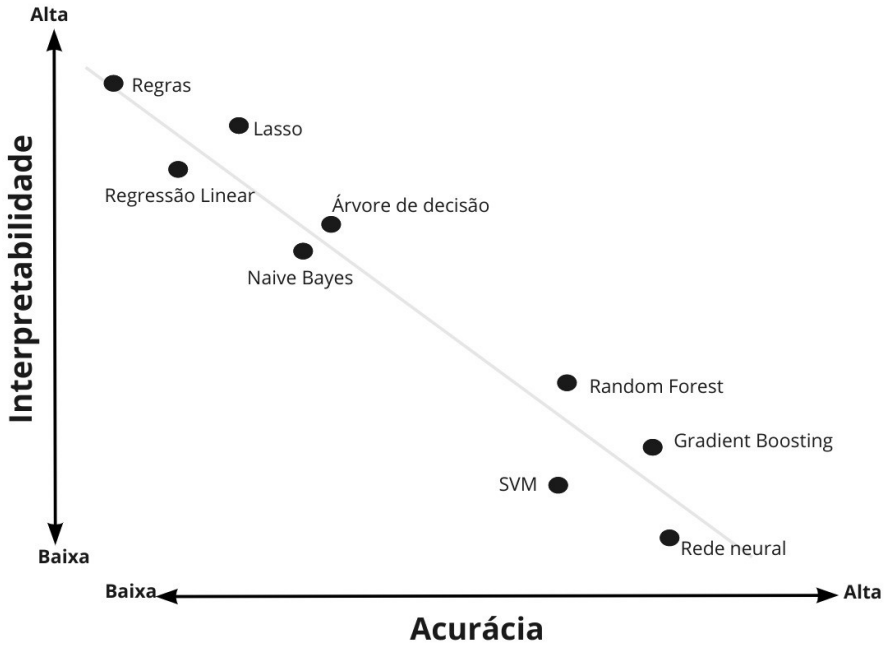


Figura 3: Interpretabilidade vs. acurácia de modelos.

existe a necessidade de explicação do resultado gerado, é necessário o uso de técnicas de interpretabilidade externa ao modelo para explicar o seu resultado. Existem diversas técnicas de interpretabilidade, que podem ser classificadas conforme diferentes critérios. Uma possibilidade é classificar de acordo com o escopo da interpretabilidade, e dividir em duas principais categorias: técnicas globais que visam entender o funcionamento do modelo como um todo [9, 44], e técnicas locais usadas para interpretar uma previsão específica [48, 56].

Outra possibilidade é dividir em dois grupos de acordo com a dependência do modelo. O primeiro é específico para certo tipo de modelo, e a sua implementação depende de considerável entendimento do funcionamento do modelo [7, 21, 70]. O segundo tipo é agnóstico ao modelo, por isso é usado a mesma metodologia de interpretabilidade para qualquer tipo de modelo [27, 56], e a explicação é calculada a partir dos valores de entrada e saída do modelo.

Nessa dissertação o foco foi em metodologias locais de interpretabilidade, as duas metodologias locais mais conhecidas e utilizadas são o LIME (Locally Interpretable Model-agnostic Explanations) [56] e o SHAP (SHapley Additive exPlanations) [48]. Sendo escolhido utilizar aqui o SHAP, pois ele garante algumas propriedades desejadas que serão descritas posteriormente.

SHAP possui diversas variações; uma delas é agnóstica ao modelo, e outras são es-

pecíficas para certos modelos para otimizar o cálculo. SHAP usa conceitos da teoria cooperativa dos jogos de métodos baseados no cálculo valor de Shapley [23, 42, 72]. Porém o cálculo do valor de Shapley é muito custoso, por isso o SHAP também usa conceitos de outros metodologias de interpretabilidade, como do LIME e da abordagem de Saabas [58], para aproximar o resultado. Nesta seção descreveremos as variações do SHAP usadas nesta pesquisa.

3.1 Metodologias de atribuição aditiva do atributo

O SHAP e suas variantes têm em comum a mesma Expressão (3.1) de atribuição aditiva de atributos, que aproxima linearmente o resultado de um modelo:

$$g(z') = \phi_0 + \sum_{i=0}^{M} \phi_i z'_i, \tag{3.1}$$

sendo g a função aditiva e M o número de atributos do modelo. A variável z' pode assumir valores $\{0, 1\}$. A variável ϕ_i representa o efeito do atributo no modelo f a ser explicado. Com isso é obtido um peso por atributo que é usado para explicar o seu resultado, conforme representado na Figura 4.

Em métodos locais como o SHAP, a explicação é gerada para uma amostra específica x. O valor original de x pode ser difícil para o usuário interpretar, por isso é comumente feita uma transformação para valores binários x', que indicam a presença ou ausência do atributo na explicação. Para obter os valores originais da amostra x é definida uma função $h_x(x') = x$. Métodos locais tentam garantir que $g(z') \approx f(h_x(z'))$ para a amostra que deseja ser explicada ($z' = x'$).

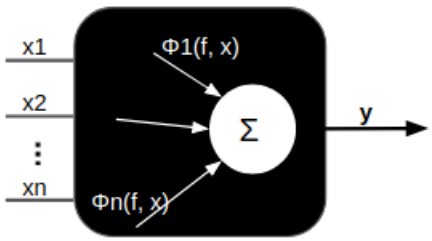


Figura 4: Representação do método de atribuição aditiva do atributo para aproximar linearmente a previsão de um modelo, atribuindo um peso para cada atributo.

3.2 LIME

A metodologia proposta por Ribeiro et al. do LIME [56] fornece explicação de um modelo f para a previsão gerada para uma amostra x aproximando localmente um classificador g por uma função linear. Na Figura 5 estão representadas as etapas envolvidas no treinamento do classificador local g.

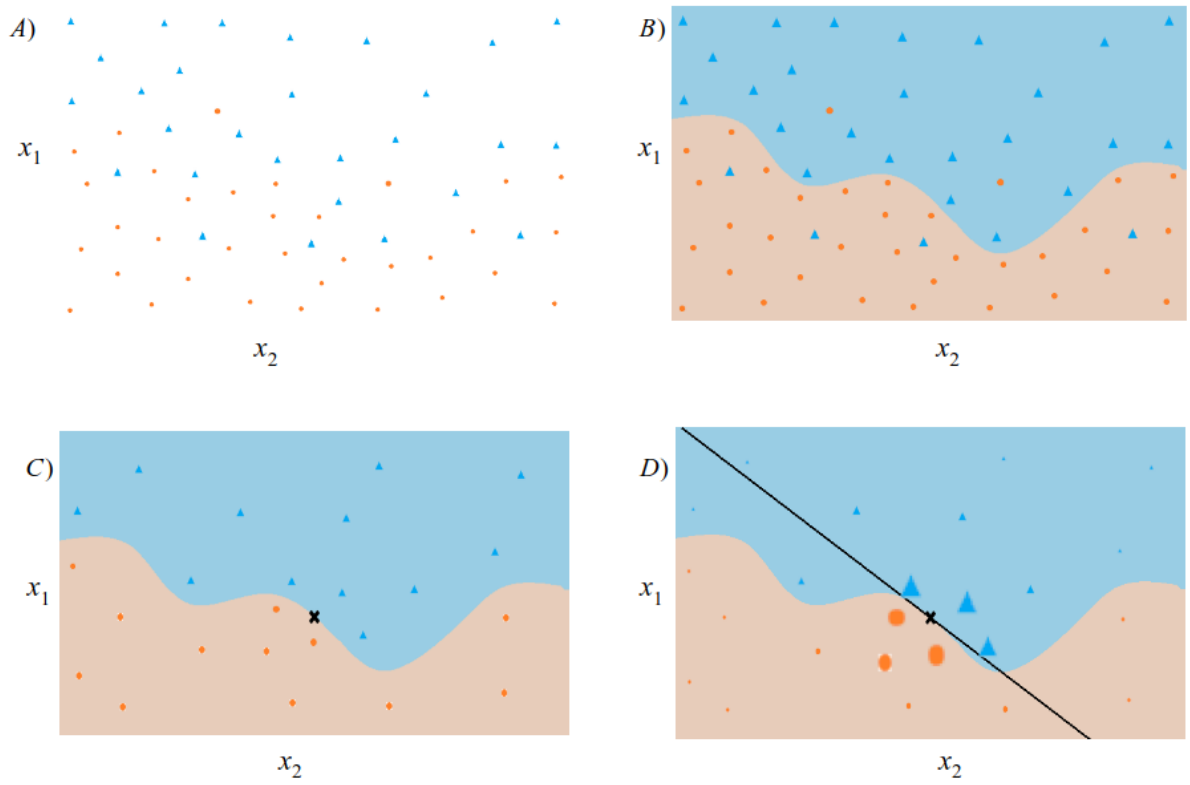


Figura 5: Treinamento do LIME com dados tabulares. A) Amostra de dados X com dois atributos (x_1 e x_2) e 2 classes possíveis (círculo laranja e triângulo azul). B) Resultado do treinamento de um modelo f com dados X. C) O símbolo preto representa uma amostra x que deseja gerar explicação, em torno dela são geradas amostras z. D) Ponderação de z de acordo com a distância de x (representado pelo tamanho dos símbolos) e treinamento de um modelo linear g que aproxima localmente f.

Na Figura 5A está representado um conjunto de dados X com dois atributos (x_1 e x_2), e duas classificações possíveis (círculos laranja e triângulos azul). Um modelo f foi treinado com X, gerando a classificação na Figura 5B representada pelas cores de fundo azul e laranja. O símbolo preto X na Figura 5C mostra uma amostra x em relação à qual se deseja obter explicação para a classificação gerada por f. Para isso na Figura 5C foram criadas amostras sintéticas z ao redor de x, que são classificadas pelo modelo f. As amostras z são ponderadas por um kernel $\pi_{x'}$ de acordo com a proximidade de x, dessa forma amostras mais próximas de x recebem pesos maiores e as mais distantes

pesos menores. O resultado da ponderação está representado em 5D pelo tamanho do símbolo. Usando as amostras z, a ponderação $\pi_{x'}$ e a classificação $f(z)$ é treinado um modelo linear g para explicar o resultado da amostra x.

A explicação gerada pelo modelo linear g é descrita pela Equação (3.1), em que ϕ_i representa o impacto de cada atributo no modelo f. Para o cálculo dos parâmetros ϕ_i, LIME minimiza a seguinte função objetivo:

$$\xi = \arg\min_{g \in \mathcal{G}} \mathcal{L}(f, g, \pi_{x'}) + \Omega(g), \tag{3.2}$$

sendo $\mathcal{L}$ a função de custo, que mede o erro da aproximação local de f por g. $\mathcal{G}$ é a classe de modelos lineares, e $\Omega(g)$ é uma medida de complexidade da explicação gerada por g. A função de custo $\mathcal{L}$ é medida pelo erro quadrático médio, ponderado pelo kernel $\pi_{x'}$:

$$\mathcal{L}(f, g, \pi_{x'}) = \sum_{z' \in Z} (f(h_x(z')) - g(z'))^2 \pi_{x'}(z'), \tag{3.3}$$

sendo $h_x(x') = x$ uma função que mapeia valores binários no seu espaço original. A função $\pi_{x'}$ mede a proximidade entre z' e x', e o tipo de kernel deve ser escolhido de forma arbitrária de acordo com o problema. No LIME é sugerido o uso de um kernel exponencial:

$$\pi_{x'}(z') = exp\left(\frac{-D(x', z')^2}{\sigma^2}\right), \tag{3.4}$$

sendo D uma função de distância com largura σ definida de acordo com a aplicação, exemplo distância L2 para imagens.

Uma das maiores dificuldades no LIME é definir a função do kernel, pois o resultado pode variar muito de acordo com a forma como esta função é definida [3, 51]. No artigo que apresenta o LIME não é feita nenhuma definição formal de como escolher o kernel ótimo.

Por exemplo, para o caso de um kernel exponencial é necessário determinar o valor de σ, sendo que que esse valor determina a taxa de decaimento da função exponencial, e quanto maior ele for maior será a taxa de decaimento da função, tornando o kernel mais estreito. Dependendo da aplicação pode ser melhor um kernel mais estreito, onde as amostras devem estar muito próximas de x para influenciar o modelo g treinado localmente. Enquanto que em outros casos é melhor um kernel mais largo, que considera também amostras mais distantes.

No LIME o processo de criação das amostras z varia se o atributo for numérico ou categórico. Se o atributo for numérico, os valores são gerados usando uma distribuição

Algorithm 2: Amostragem de dados no LIME

Entradas:
x ← Amostra que será explicada
X ← Conjunto de dados real, usado para treinar o modelo f
N ← Tamanho do conjunto de dados Z que será criado
posicao_feature_categorica ← lista com índices das colunas categóricas de X
Resultado: Z ← Conjunto de dados criado
Iteração sobre as colunas **X**
for X_j *in* X **do**
  if j *in posicao_feature_categorica* **then**
    $v_j \leftarrow$ valores distintos de X_j
    $f_j \leftarrow$ distribuição de X_j
    $Z_j \leftarrow$ random_int(valores possíveis=v_j, probabilidade associada = f_j)
  else
    $\sigma_j \leftarrow$ desvio de X_j
    $\mu_j \leftarrow$ média de X_j
    $Z_j \leftarrow$ random_normal(μ =μ_j, σ =σ_j)
  end if
end for

normal, com média e desvio igual ao do atributo original. Se o atributo for categórico, são gerados os mesmo valores existentes no atributo com base na sua frequência de ocorrência. O Pseudo Código 2 representa o processo de amostragem para gerar as amostras z. Este processo de amostragem têm alguns problemas, como de considerar os atributos independentes e com distribuição normal. Isso nem sempre é verdade, e pode gerar problemas na explicação gerada pelo modelo linear g [13, 39].

3.3 Abordagem de Saabas

Saabas [58] propõe uma metodologia de explicação local para modelos baseado em árvore. A explicação é feita considerando a contribuição de cada atributo no caminho percorrido na árvore para gerar o resultado.

Considerando o caso mais simples de uma árvore de decisão f, o resultado gerado para uma amostra x pode ser descrito pela Equação (3.1) de atribuição aditiva. Nela ϕ_0 é á média dos valores da variável alvo do dado de treino, ϕ_i é a contribuição de cada atributo no caminho percorrido e M é o número de atributos.

Iremos descrever como é feito o cálculo da contribuição de cada atributo com o exemplo mostrado na Figura 6. Nela é mostrado um caso fictício de um modelo de árvore de decisão f para aproximar os dados reais descritos por F, que contém os atributos v_1, v_2 e v_3. As

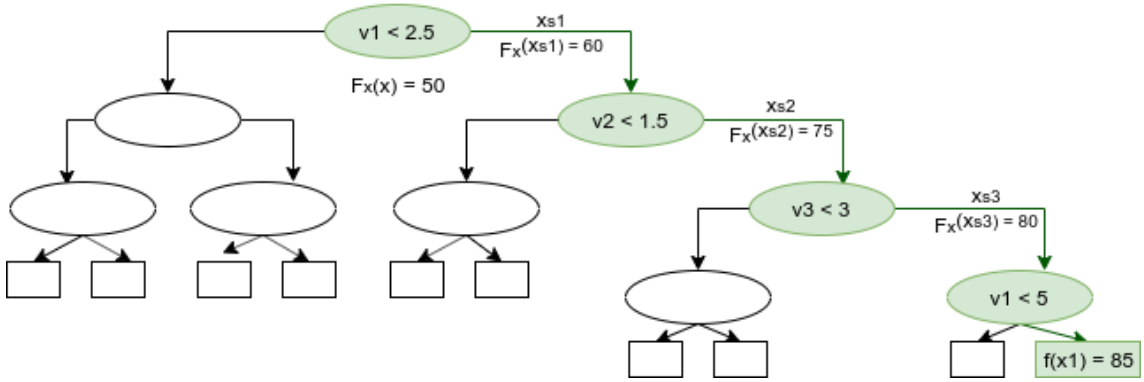


Figura 6: Exemplo fictício de modelo de árvore de decisão, onde a parte em verde enfatiza o caminho percorrido para gerar explicação do resultado para amostra x_1.

elipses são os nós da árvore, e os retângulos são as folhas. Para gerar o resultado para uma amostra x_1 foi percorrido o caminho em verde. Temos que X indica os dados usados no treino da árvore f, e $F_X(z) = E(F(z)|z \in X_{S_i})$ é uma função que calcula o valor médio das variáveis alvo para certo subconjunto X_{S_i} do dado de treino. Nesse caso, usando a Equação (3.1) a explicação é:

$$f(x) = 50 + (60 - 50) + (75 - 60) + (80 - 75) + (85 - 80) = 85,$$
$$f(x) = 50 + (10 + 5) \text{ [importância de } x_1\text{]} + 15 \text{ [importância de } x_2\text{]}.$$

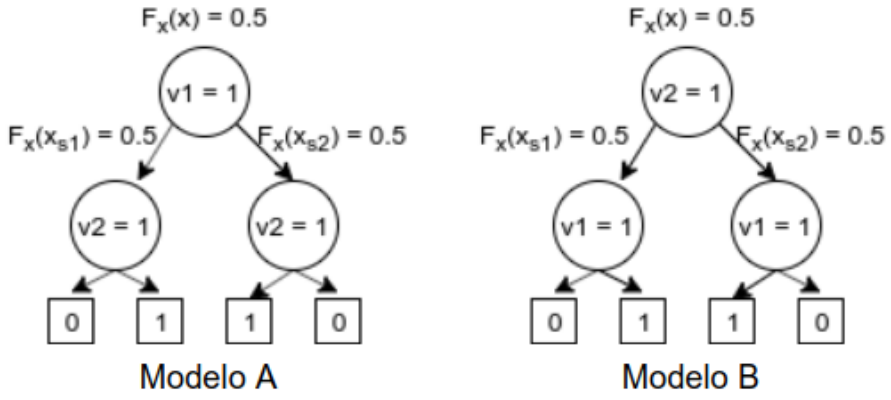
A Equação (3.1) é específica para o caso de uma árvore de decisão. Para modelos que são conjunto de árvores, como Random Forest, e explicação é gerada pela Equação (3.5):

$$f(x) = \frac{1}{J}\sum_{j=1}^{J}\phi_{0j} + \sum_{i=1}^{M}\left(\frac{1}{J}\sum_{j=1}^{J}\phi_{ij}\right), \tag{3.5}$$

sendo J o número de árvores.

A abordagem proposta por Saabas tem como vantagem garantir a acurácia local, em que o resultado gerado pela explicação é igual ao resultado do modelo, porém essa metodologia não garante consistência. Isso pode ser visto pelo exemplo mostrado na Figura 7 de uma árvore XOR com atributos v_1 e v_2. Nela a diferença entre os modelos A e B é que v_1 e v_2 foram trocados de posição.

Considerando que nos dois modelos haja quantidades iguais de pontos em cada folha, para garantir consistência no resultado da explicação a alteração na posição dos atributos não deveria alterar a sua importância, porém vemos que não é isso que acontece. No modelo A vemos que o atributo v_1 sempre têm importância 0 e a importância v_2 é sempre diferente de 0. Porém quando a ordem é trocada no modelo B esse cenário se inverte, o atributo v_2 passa a ter importância 0 e a importância de v_1 é sempre diferente de 0.



			Modelo A		Modelo B	
v1	v2	v1 ⊕ v2	ϕ(v1)	ϕ(v2)	ϕ(v1)	ϕ(v2)
1	1	0	0	-0.5	-0.5	0
1	0	1	0	0.5	0.5	0
0	1	1	0	0.5	0.5	0
0	0	0	0	-0.5	-0.5	0

Figura 7: Dois modelos de árvore de decisão XOR, onde a única diferença entre eles é a ordem dos atributo v_1 e v_2. A tabela mostra a importância $\phi(v_1)$ e $\phi(v_2)$ calculadas pela metodologia do Saabas. Pelo resultado vemos que a alteração da ordem alterou a importância dos atributos, o que mostra inconsistência na explicação gerada.

3.4 Valor de Shapley

Na teoria dos jogos, são considerados jogos cooperativos aqueles em que acontece disputa entre grupos de jogadores (colisões). Sendo o objetivo prever 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. O valor de Shapley [59] é uma das solução para a teoria dos jogos cooperativa. De acordo com o valor de Shapley, o pagamento de um jogador i em um jogo com M jogadores é calculado pela Equação (3.6):

$$\phi_i = \sum_{S \subseteq M \setminus \{i\}} \frac{|S|!(M - |S| - 1)!}{M!} (v(S \cup \{i\}) - v(S))), \tag{3.6}$$

sendo $v(S)$ uma função que calcula o efeito da colisão em um jogo com um subconjunto S de jogadores.

Na Equação (3.6) o efeito marginal do jogador i em cada colisão S é calculado através da diferença entre $v(S \cup \{i\}) - v(S)$. Para o cálculo do pagamento do jogador é feito uma

média ponderada do seu efeito marginal entre todas as colisões possíveis. Dentre os subconjuntos possíveis também é considerado o subconjunto vazio $S = \emptyset$. Para exemplificar, a seguir está sendo calculado qual seria o pagamento de cada jogador em um jogo com 3 jogadores.

$$\begin{aligned}
\phi_1 &= \frac{1}{3}\Big(v(\{1,2,3\}) - v(\{2,3\})\Big) + \frac{1}{6}\Big(v(\{1,2\}) - v(\{2\})\Big) + \\
&\quad \frac{1}{6}\Big(v(\{1,3\}) - v(\{3\})\Big) + \frac{1}{3}\Big(v(\{1\}) - v(\{\emptyset\})\Big), \\
\phi_2 &= \frac{1}{3}\Big(v(\{1,2,3\}) - v(\{1,3\})\Big) + \frac{1}{6}\Big(v(\{1,2\}) - v(\{1\})\Big) + \\
&\quad \frac{1}{6}\Big(v(\{2,3\}) - v(\{3\})\Big) + \frac{1}{3}\Big(v(\{2\}) - v(\{\emptyset\})\Big), \\
\phi_3 &= \frac{1}{3}\Big(v(\{1,2,3\}) - v(\{1,2\})\Big) + \frac{1}{6}\Big(v(\{1,3\}) - v(\{1\})\Big) + \\
&\quad \frac{1}{6}\Big(v(\{2,3\}) - v(\{2\})\Big) + \frac{1}{3}\Big(v(\{3\}) - v(\{\emptyset\})\Big).
\end{aligned}$$

Usando a definição da Equação (3.6), o valor de Shapley diz como distribuir de forma justa o pagamento entre jogadores, e garante as seguintes propriedades:

Eficiência: todo o ganho é distribuído entre os jogadores:

$$\sum_{i=1}^{N} \phi_i = v(N).$$

Simetria: se dois jogadores contribuem igualmente em todas as colisões possíveis $v(S \cup \{i\}) = v(S \cup \{j\})$, então seus valores de Shapley são iguais: $\phi_i = \phi_j$

Linearidade: ϕ_i é uma função linear; ou seja, satisfaz:

- Aditividade: $\phi_i(v + w) = \phi_i(v) + \phi_i(w)$;
- Homogeneidade: $\phi_i(av) = a\phi_i(v)$.

Jogador sem influência: um jogador que não influencia no resultador do jogo recebe pagamento zero. Isso ocorre quando $v(S \cup \{i\}) = v(S)$ para todas as colisões S possíveis.

Em 1985 Young demonstrou que o valor de Shapley é a única solução que satisfaz essas propriedades [66]. Alguns métodos usam conceitos do valor de Shapley para estimar a importância dos atributos do modelo [23, 42, 72]. Nesse contexto, o resultado de uma previsão é o jogo, os atributos do modelo são os jogadores e a importância de cada atributo é o pagamento distribuído de forma justa.

Um dos primeiros métodos a usar o valor de Shapley para o cálculo da importância do atributo foi proposto por Lipovetsky [42], usado para o caso de um modelo linear com

presença de multicolinearidade. O impacto de cada atributo i é calculado pela Equação (3.7), em que para a sua estimativa são treinados diversos modelos para cada colisão possível, dessa forma é estimado o impacto de cada colisão marginal através da diferença entre as previsões $f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)$:

$$\phi_i = \sum_{S\subseteq F\setminus\{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!}[f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)], \tag{3.7}$$

sendo F o conjunto de todas os atributos.

Como o cálculo da Equação (3.7) é muito custoso pois exige treinar diversos modelos, foram propostas diversas aproximações desse cálculo. Uma das propostas foi aplicar a amostragem dos dados de certo atributo para simular o efeito da sua remoção no modelo, sem ser necessário dessa forma retreinar um modelo. Isso foi feito em *Shapley sampling value* [72] e *quantitative inpute influence* [23].

3.5 SHAP

SHAP [48], proposto por Lundberg et al., é outra metodologia que usa conceitos do valor de Shapley para calcular a importância dos atributos do modelo. Como descrito previamente o cálculo exato do valor de Shapley pela Equação (3.7) é muito custoso, por isso Lundberg et al. propõe metodologias de aproximação. Dentre essas metodologias

Uma dessas metodologias é o *Kernel SHAP*, que usa conceitos do LIME para gerar a explicação. Essa metodologia tem como vantagem ser agnóstica ao tipo de modelo, porém é possível obter ainda maior eficiência no cálculo quando a explicação é restrita para certos tipos de modelo. Isso foi feito por exemplo com o *Tree SHAP*, que é específico para modelos baseados em árvore, o *Deep SHAP*, que é específico para modelos de Deep Learning, e o *Linear SHAP*, que é específico para modelos lineares.

Na Figura 8 é mostrado algumas das variantes do SHAP, com especificação das metodologias que cada variante herda conceitos e o escopo de aplicação de cada variante. O conceito base comum a todas estas variantes é do valor de Shapley, sendo retratado na figura pela elipse preta, onde estão descritas as metodologias que usam esse mesmo conceito.

A metodologia do SHAP gera explicação para uma amostra x treinada com um modelo f, aproximando localmente por uma função linear g. Essa definição segue a mesma formulação de atribuição aditiva da Equação (3.1). As amostras $z' \in \{0, 1\}$ nessa for-

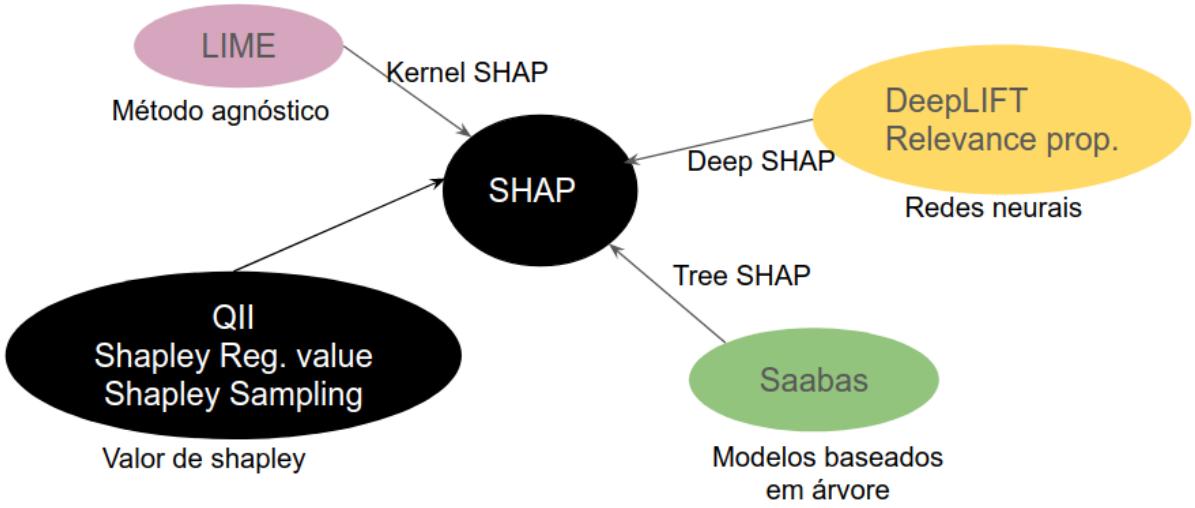


Figura 8: Representação de algumas variantes do SHAP, com identificação do seu escopo de aplicação e de algumas metodologias que cada variante usa conceitos. A elipse preta com descrição valor de Shapley representa o conceito base do SHAP, sendo especificado nela metodologias que previamente usaram esse mesmo conceito.

mulação representam as possíveis colisões entre os atributo, sendo que 1 indica presença do atributo e 0 ausência. O efeito de cada atributo ϕ_i nesse formulação é chamado de valor de SHAP.

O valor de Shapley é a única solução que garante eficiência, simetria, linearidade e jogador sem influência. SHAP também satisfaz essas propriedades, e conforme descrito no seu artigo [48], delas podem ser derivadas outras propriedade interessantes para a explicação gerada:

Acuracidade local: O resultado da explicação local gerada pelo modelo linear g para uma amostra x é igual ao resultado gerado pelo modelo f:

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

Essa propriedade é equivalente a propriedade da eficiência do valor de Shapley quando todos os $z'_i = 1$.

Consistência: Se um modelo f muda para um modelo f' de tal forma que a influência de um atributo i fica maior ou igual, a importância ϕ_i calculada não deve diminuir. Dessa forma, para qualquer dois modelos f e f', se:

$$f'_x(z') - f'_x(z'_1) \geq f_x(z') - f_x(z'_1).$$

Para qualquer entrada z' com valores $\{0, 1\}$, a importância ϕ_i calculada para a amostra x deve ser $\phi_i(f', x) \geq \phi_i(f, x)$.

Omissão: equivalente a propriedade já descrita do "Jogador sem influência", em que se $f_x(S \cup i) = f_x(S)$, para qualquer subconjunto S, então o atributo i não deve ter impacto na explicação, por isso o $\phi_i = 0$. Na prática, a metodologia do SHAP apenas considera um atributo completamente sem influência se os valores desse atributo forem constantes em todo o conjunto de dados [45].

Métodos que não usam conceitos do valor de Shapley para o cálculo da importância do atributo não garantem todas as propriedades descritas acima. O LIME por exemplo garante apenas a propriedade da omissão.

3.5.1 Kernel SHAP

A metodologia do Kernel SHAP usa conceitos do LIME para o cálculo dos parâmetros ϕ_i de importância dos atributos; por isso o processo de cálculo destes parâmetros é semelhante ao representado na Figura 5. Nesse processo, para obter explicação do resultado de um modelo f para uma amostra x é feita a amostragem de dados em torno de x, criando novas amostras z que são ponderadas por um kernel $\phi_{x'}$. Usando as amostras z e o kernel $\pi_{x'}$ é treinado um modelo linear g para explicar o resultado de x. Os parâmetros do modelo linear são definidos através da minimização da função objetivo 3.3. As principais diferenças nesse processo do Kernel SHAP em relação ao LIME estão na metodologia de amostragem dos dados e na definição do kernel.

No LIME, a definição do kernel $\pi_{x'}$ e do termo de regularização Ω da função custo 3.3 é livre, sendo definida de acordo com o problema. Dependendo da forma como o kernel é escolhido, a explicação gerada por g pode variar muito [51]. Esse problema é resolvido no Kernel SHAP, em que ele mostra que para ser obtido um resultado compatível com valor de Shapley, $\pi_{x'}$ e Ω devem ser definidos da seguinte forma:

$$\Omega(g) = 0,$$
$$\pi_{x'}(z') = \frac{M-1}{(M \text{ choose } |z'|)|z'|(M-|z'|)},$$

sendo $|z'|$ o número de elementos diferentes de zero.

Para a amostragem dos dados, inicialmente é criada uma matriz Z' com valores binários representado as colisões entre os atributos. Esta matriz Z' é usada como base para criação de Z. Quando $z'_{ij} = 1$ é atribuído a z_{ij} o mesmo valor do atributo j da amostra x que deseja ser explicado, enquanto que quando $z'_{ij} = 0$ é atribuído um valor aleatório dentro os valores possíveis do atributo j. Esse processo de amostragem não

Algorithm 3: Amostragem de dados no Kernel SHAP

Entradas:
$\boldsymbol{x_{1\times F}} \leftarrow$ Amostra que será explicada com F atributos
$\boldsymbol{X_{N\times F}} \leftarrow$ Conjunto de dados real, usado para treinar modelo f
$\boldsymbol{N_{max}} \leftarrow$ Tamanho máximo do conjunto de dados X' que será criado
Resultado: $Z \leftarrow$ Conjunto de dados criado
$M \leftarrow$ Número de atributos não omissos
$Ns \leftarrow max(2^M, N_{max})$
Cria matriz com valores $\{0, 1\}$ representado as possíveis colisões de atributos:
$Z'_{Ns\times F} \leftarrow random\_binario(num\_linhas = N, num\_cols = F)$
Iteração sobre as colunas **X**
for *i in range(1, Ns)* **do**
  for *j in range(1, F)* **do**
    if $Z'[i, j] == 1$ **then**
      $Z[i, j] \leftarrow x[1, j]$
    else
      $Z[i, j] \leftarrow random\_choose(valores\_possiveis{=}X_j)$
    end if
  end for
end for

leva em consideração a dependência entre os atributos, sendo que a explicação gerada com o Kernel SHAP pode ser problemática caso exista uma alta dependência entre os atributos [1]. O Psudeo Código 3 descreve o processo de amostragem do Kernel SHAP.

Lundberg et al. compara em seu trabalho [48] o Kernel SHAP, LIME e Shapley sampling value para calcular a importância de um atributo. O valor correto da importância do atributo foi obtido pelo valor do Shapley sem aproximação. Nesta comparação ele mostra que o LIME diverge consideravelmente do resultado correto, e o Shapley sampling value precisa de um número consideravelmente maior de iterações em relação ao Kernel SHAP para aproximar corretamente e diminuir a incerteza.

3.5.2 Linear SHAP

Para modelos lineares, se for assumida independência entre os atributos, o cálculo da importância do atributo ϕ_i é exato. Para o caso de um modelo linear $f(x) = \sum_{j=1}^{M} w_j x_j + b$, os valores de SHAP são:

$$\phi_0(f, x) = b,$$
$$\phi_i(f, x) = w_j(x_j - E[x_j]).$$

Esse resultado foi mostrado anteriormente por Štrumbelj e Kononenko [72]. Para o caso que considera dependência entre os atributos o cálculo de ϕ_i é mais complexo. Em geral assume-se que o conjunto de dados X tem uma distribuição gaussiana multivariável com média μ e covariância Σ, e o valor de SHAP é então obtido pela equação abaixo:

$$\phi_i = w\left[\frac{1}{M!}\sum_R([Q_{S_i^{\bar{R}}\cup i} - R_{S_i^R\cup i}] - [Q_{S_i^{\bar{R}}} - R_{S_i^R}])\right]\mu,$$
$$+w\left[\frac{1}{M!}\sum_R([Q_{S_i^R\cup i} + R_{S_i^R\cup i}] - [Q_{S_i^R} + R_{S_i^R}])\right]x.$$

sendo $R_S = P_{\bar{S}}^T P_{\bar{S}} \Sigma P_S^T (P_S \Sigma P_S^T)^{-1} P_S$ e $Q_S = P_S^T P_S$, onde P_S é a matriz de projeção que seleciona um subset S. A demonstração do cálculo de ϕ_i é apresentada no repositório do projeto do SHAP [46].

3.5.3 Tree SHAP

A maioria das metodologias para explicar modelos baseados em árvore presentes na literatura são metodologias globais, sendo que as três principais métricas usadas são:

- Ganho: Calcula a contribuição do atributo em reduzir o erro ou impuridade devido as divisões presentes na árvore [16].
- Contagem de divisões: Contagem do número de vezes que o atributo é usado como divisão da árvore [19].
- Permutação: Quantifica a variação do erro do modelo ao permutar aleatoriamente os valores de um atributo [8].

Uma das poucas abordagens locais existentes para explicar modelos de árvore é a proposta por Saabas [58]. Essa metodologia possui semelhanças com o SHAP, pois assim como ele propõe o uso de uma função de atribuição aditiva do atributo, Equação (3.1), para explicar a previsão. Porém diferentemente do SHAP que calcula a importância do atributo considerando o efeito da sua inclusão em todas as combinações possíveis com outros atributos, a metodologia de Saabas considera apenas o caminho percorrido na árvore para gerar a previsão. Por isso, a metodologia de Saabas não garante consistência. No caso de um modelo que é um conjunto de árvores com todas as combinações possíveis de atributos o valor obtido pela abordagem de Saabas convergiria para o valor obtido com o SHAP e seria obtido consistência [47].

Conforme descrito anteriormente, o cálculo do valor exato de Shapley pela Equação (3.7) é muito custoso. Para modelos baseado em árvore, para estimar $f_S(x_s) = E[f(x)|x_s]$ não é necessário retreinar o modelo, pode ser estimado de maneira semelhante à abordagem de Saabas, que calcula o valor médio do subconjunto de treino que passa por certo nó. Mesmo assim a complexidade do cálculo continua elevada, sendo da ordem de $O(TL2^M)$, sendo T o número de árvores, M o número de atributos e L o número máximo de folhas em cada árvore.

Lundberg et al. propõe com o Tree SHAP [47] uma metodologia otimizada de calcular o valor exato de Shapley, reduzindo a complexidade para ordem polinomial $O(TLD^2)$, sendo D a profundidade máxima da árvore. No artigo que descreve a metodologia é apresentado o pseudo código do cálculo, a ideia para a otimização é armazenar a proporção e tamanho de todos os subconjuntos de atributos que geram o resultado para cada folha, e assim evitar fazer diversas varreduras na árvore.

Essa metodologia permite gerar explicação apenas para a saída direta do modelo, porém dependendo da aplicação pode ser necessário a explicação do resultado de uma transformação não linear da saída do modelo [47]. Como não é simples adaptar o cálculo do Tree SHAP para gerar explicação para uma transformação não linear da saída do modelo, foi proposta uma variação do Tree SHAP esses casos chamada que não considera dependência entre os atributos de acordo com regras de inferência causal [32]. O pseudo código desta metodologia também apresentada por Lundberg et al. em seu artigo [47].

3.5.4 Interpretação global do resultado

Além do SHAP propor diversas metodologias para explicar localmente o resultado da previsão do modelo, também se propõe metodologias de agregação da explicação para avaliar o modelo de forma global.

Uma das propostas é calcular a importância global do atributo pela Equação (3.8). Esse tipo de resultado é semelhante ao fornecido por metodologias como a de ganho [16] para modelos baseado em árvores, que foi descrita anteriormente. Uma das formas de avaliar o resultado da importância do atributo é com o gráfico de barras mostrado na Figura 9A. Porém essa forma de visualização não mostra a distribuição dos valores obtidos para a explicação do atributo, e como esses valores se relacionam com seu valor real.

Como o SHAP calcula a importância do atributo por previsão, com ele é possível mostrar esse tipo de relação. Que é a proposta do gráfico chamado *summary plot*, mostrado

na Figura 9B. Nele os atributos são ordenadas de acordo com o resultado da Equação (3.8), os valores de SHAP ϕ_i obtidos por atributos são representados por círculos, e são ordenados e plotados horizontalmente. Cada valor de ϕ_i é colorido de acordo com o valor do atributo, sendo associado a cor azul aos valores menores e vermelho aos maiores.

$$\frac{1}{N}\sum_{j=1}^{N} \mid \phi_i^{(j)} \mid . \tag{3.8}$$

Os gráficos da Figura 9 mostram a importância do atributo de uma Random Forest treinada com o conjunto de dados do Adult. As variáveis categóricas desse conjunto de dados foram transformadas em atributos binários, para representar por exemplo a categoria *Masters* da variável *Education* foi criado o atributo *Education_Masters*, sendo associado valor 1 para o caso em que a categoria de *Education* é *Masters* e 0 para os outros casos.

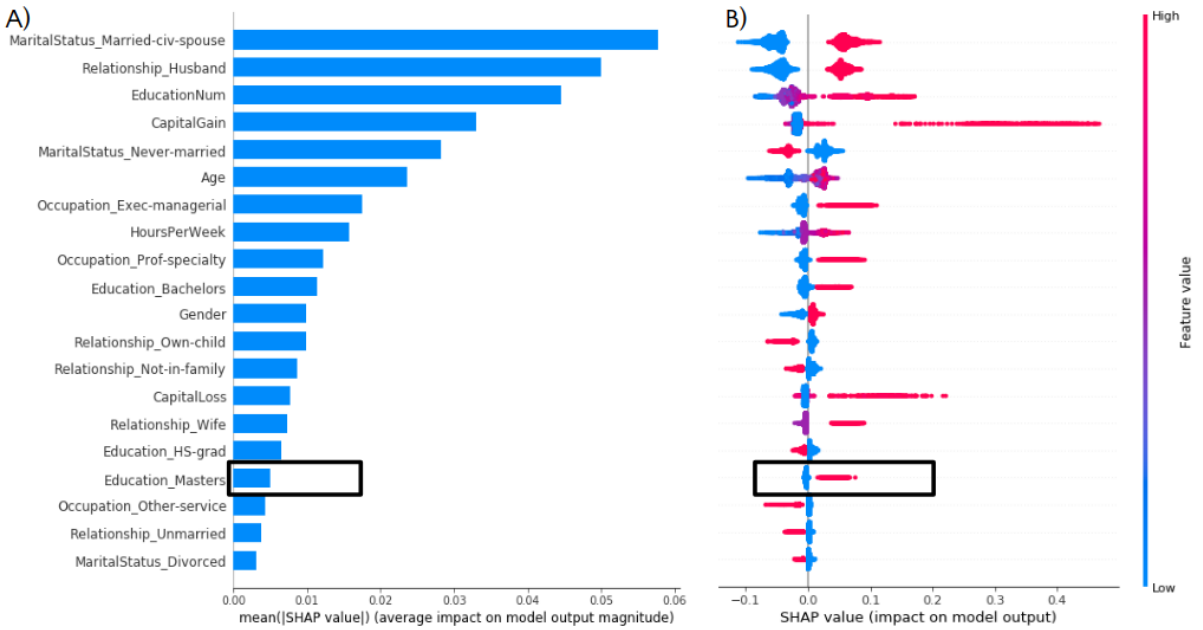


Figura 9: Diferentes formas de visualização para mostrar a importância global dos atributos através dos valores do SHAP, para explicar o resultado de uma Random Forest treinado com o conjunto de dados do Adult. **A**) Gráfico de barras com um único valor por atributo para representar sua importância no modelo. **B**) Gráfico com todos os valores de SHAP da importância do atributo. Com isso é possível visualizar a distribuição do valores, além disso esse gráfico mostra a relação do valor de SHAP com o valor do atributo.

Pelo resultado da Figura 9A, o atributo *Education_Masters* parece ter pouca importância no modelo, porém pela Figura 9B vemos que apenas quando a categoria de *Education* não é *Masters* que de fato a importância é praticamente nula. Enquanto

que nos casos em que a pessoa tem *Masters* há um efeito de aumentar consideravelmente o resultado da previsão. Como em aproximadamente 95% das amostras o atributo $Education\_Masters = 0$, que é associado a um ϕ_i aproximadamente nulo, o resultado do impacto global desse atributo pela Equação (3.8) é baixo.

Outro tipo de gráfico do SHAP interessante é o de Partial Dependence Plot (PDP) mostrado na Figura 10, que permite entender com mais detalhe como um atributo específico afeta o resultado do modelo. Ele permite avaliar a relação entre os valores do atributo (eixo x) e o valores do SHAP (eixo y). Além disso, os pontos nos gráficos são coloridos de acordo com os valores do atributo mais relacionado com escala de cor representada no lado esquerdo do gráfico.

O gráfico de PDP da Figura 10 foi gerado para o atributo de idade com um modelo de Gradient Boosting treinado com o conjunto de dados do PNAD. A idade foi normalizada para usar no modelo, isso explica o seu valor variar de -2 a 3. O conjunto de dados do PNAD contém dados da população brasileira, e o objetivo do modelo construído foi prever se a renda de certa pessoa era superior a 2500 reais.

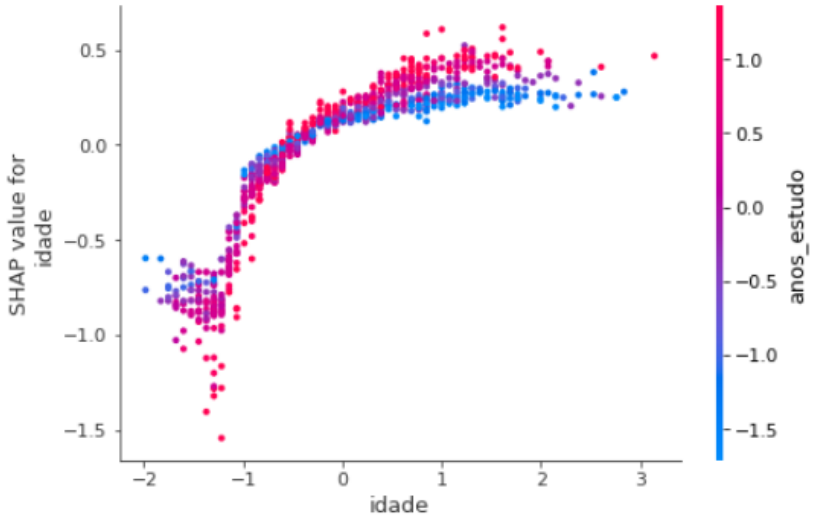


Figura 10: Gráfico de PDP para a variável *idade* de um Gradient Boosting treinado com o conjunto de dados do PNAD. Com esse gráfico vemos a relação entre os valores do atributo, no eixo x, e os valores de SHAP, no eixo y.

Avaliando o gráfico vemos que para *idade* até -1, os valores de SHAP são negativos, com isso nessa faixa de valores *idade* influencia para diminuir a previsão. Além disso, vemos que de -1 até 1 quanto mais aumenta a *idade* mais o atributo impacta para aumentar o valor previsto, e à partir da *idade* 1 o valor de SHAP fica constante. Outra relação importante mostrada no gráfico é que até certa idade, anos de estudos não impactam para aumentar o valor previsto. Porém a partir de certa idade, o aumento dos anos de

estudos tem impacto em aumentar a previsão.

4 PROPOSTA

Esta seção contém a descrição do framework proposto para comparar as medidas de justiça e os resultados de interpretabilidade obtidos com o SHAP, e a especificação das métricas que serão utilizadas nessa comparação. Uma das dificuldades da comparação é que o SHAP gera um resultado local por amostra, enquanto que as medidas de justiça capturam o comportamento global do modelo. Pela maior facilidade de comparação entre as medidas utilizando um único resultado por modelo, nesta pesquisa serão propostas formas para avaliar o resultado do SHAP globalmente.

4.1 Medidas avaliadas

Lundberg et al. propõe o uso da Equação (3.8) para avaliar a importância global do atributo através dos valores de SHAP. Será utilizada esta mesma equação para avaliar a importância do atributo sensível no modelo.

Porém a avaliação do quanto certo atributo é mais importante do que os outros costuma ser mais útil do que um valor numérico para quantificar sua importância, pensando nisso será calculada a posição no ranking do atributo sensível. Para obter esse ranking será calculada a importância global de cada atributo do modelo com a Equação (3.8) e ordenado os valores obtidos em ordem decrescente, e com isso será obtida a posição do atributo sensível. Além disso, nessa dissertação é proposto uma terceira métrica com nome de *disparidade do SHAP*.

Com isso, serão utilizados três medidas para avaliar o resultado do SHAP: importância do atributo sensível, posição no ranking do atributo sensível e disparidade do SHAP. Essas medidas estão descritas na Tabela 1, nela $\phi_i^{(k)}$ são os valores de SHAP para o grupo desprivilegiado e $\phi_i^{(l)}$ os valores de SHAP para o grupo privilegiado.

As medidas do SHAP serão comparadas com as seguintes definições de justiça: a diferença da *paridade estatística*, a diferença da *igualdade de oportunidade*, a *consistência*,

Medida do SHAP	**Fórmula**
Disparidade do SHAP	$\frac{1}{N_k}\sum_{k=1}^{N_k}\phi_i^{(k)} - \frac{1}{N_l}\sum_{l=1}^{N_l}\phi_i^{(l)}$
Importância da feature	$\frac{1}{N}\sum_{j=1}^{N} \mid \phi_i^{(j)} \mid$

Tabela 1: Equações das medidas do SHAP que utilizaremos nesse trabalho.

a diferença da *justiça contrafactual* e o *índice de entropia generalizado*. Por simplificação, a diferença da paridade estatística, a diferença da igualdade de oportunidade e a diferença da justiça contrafactual serão chamadas de paridade estatística, igualdade de oportunidade, e contrafactual. A Tabela 2 contém as medidas de justiça com as fórmulas utilizadas para calculá-las.

Medida de justiça	**Fórmula**
Paridade estatística	$P(\hat{Y}=1\|A=0) - P(\hat{Y}=1\|A=1)$
Igualdade de oportunidade	$P(\hat{Y}=1\|A=0,Y=1) - P(\hat{Y}=1\|A=1,Y=1)$
1 - consistência	$\frac{1}{N}\sum_{n=1}^{N}\left\|\hat{y}_n - \frac{1}{k}\sum_{j\epsilon kNN(X')}\hat{y}_j\right\|$
Contrafactual	$P(\hat{Y}_{A\leftarrow 0}\|X=x,A=1) - P(\hat{Y}_{A\leftarrow 1}\|X=x,A=1)$
Índice de entropia gen.	$\frac{1}{n\alpha(\alpha-1)}\sum_{i=1}^{n}\left[\left(\frac{b_i}{\mu}\right)^{\alpha} - 1\right]$

Tabela 2: Fórmulas das medidas de justiça que utilizaremos nesse trabalho.

Nas equações da Tabela 2, A é a variável sensível, sendo o valor 1 atribuído ao grupo privilegiado e 0 ao grupo desprivilegiado. $\hat{Y}$ é a variável prevista pelo modelo e Y é a variável alvo do modelo, sendo $Y = 1$ o valor da classe desejada, como aprovação de crédito. Todos os modelos construídos foram restritos a um problema de classificação binário em que variável sensível também é binária, porém a técnica do SHAP poderia ser aplicada com qualquer tipo de classificador e qualquer tipo de variável sensível.

4.2 Framework

A Figura 11 resume o processo que será implementado para comparar as medidas de justiça com as medidas do SHAP. Inicialmente as variáveis categóricas serão transformadas em binárias, depois os dados serão divididos em 80% para treino e 20% para teste, e será realizada a normalização dos atributos. Após isso, dois processos paralelos serão realizados: (i) treinamento de um modelo diretamente com os dados que serão normalizados (representado pela cor vermelha); (ii) aplicação de uma etapa adicional de mitigação do viés dos dados para treinar outro modelo (representado pela cor azul). Finalmente, as medidas de justiça e as obtidas com o SHAP, que serão geradas pelos modelos A e B respectivamente, serão comparadas.

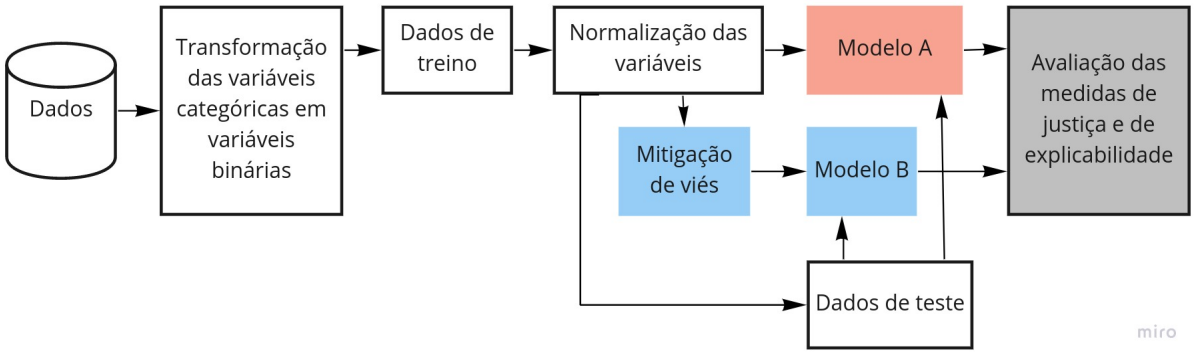


Figura 11: Framework para comparação de medidas de justiça com as medidas de explicabilidade obtidas com o SHAP. Nele é comparado a variação de resultado de um modelo enviesado, em vermelho, com o de um modelo com viés removido, em azul.

Para a etapa de mitigação do viés os dois tipos de técnicas que são detalhados no Capítulo 2 serão utilizados: reponderação e subamostragem parametrizada. Na subamostragem parametrizada é necessário definir o valor do parâmetro d, que pode assumir valores entre -1 e 1, essa flexibilidade do parâmetro permite escolher um tipo de definição de justiça para o modelo atingir. Quanto menor o valor de d maior a mudança no comportamento do modelo, sendo que o valor 1 não causa nenhuma alteração no modelo, nesse trabalho foi escolhido os valores 0 e -1 para d. Com isso, serão obtidos três modelos distintos das técnicas de mitigação de viés, que serão comparados com o resultado do modelo em que não foi aplicado nenhum tipo de técnica de mitigação de viés.

4.3 Resultados esperados

Em todos os conjuntos de dados selecionados nessa pesquisa existe uma diferença perceptível de favorecimento do grupo privilegiado em relação ao valor previsto, que

pode ser verificado pela Tabela 5. Com isso espera-se que o modelo A, onde não foi aplicado nenhum tipo de técnica de mitigação de viés, favoreça com suas previsões o grupo privilegiado. Porém, com a aplicação da técnica de mitigação de viés são esperados três cenários:

- **Igualdade entre os grupos**: cenário ideal esperado onde a técnica de fato removeu de viés, pois os grupos privilegiados e desprivilegiados são igualmente favorecidos pelo resultado do modelo.
- **Favorecimento do grupo privilegiado**: nesse cenário a técnica redução do viés não foi efetiva, onde apesar de ter diminuído a desigualdade entre os grupos, o grupo desprivilegiado continua sendo desfavorecido.
- **Favorecimento do grupo desprivilegiado**: nesse cenário houve inversão de favorecimento entre grupos, onde o grupo desprivilegiado passou a ser o favorecido pelo resultado do modelo.

A Tabela 3 mostra os valores de cada medida de justiça (em branco) e medidas do SHAP (em amarelo) nos três cenários descritos anteriormente. A medida de posição no ranking do atributo sensível não foi inclusa nessa tabela, pois não é esperado um valor específico com ela para cada cenário. Essa medida de posição será utilizada de forma auxiliar as outras medidas do SHAP, tendo como principal objetivo mostrar o quão significativa foi variação da posição no ranking do atributo sensível nos modelos com e sem mitigação de viés.

	Igualdade entre os grupos	Favorecimento do grupo privilegiado	Favorecimento do grupo desprivilegiado
Paridade estatística	0	> 0	< 0
Igualdade de oportunidade	0	> 0	< 0
Contrafactual	0	> 0	< 0
1 - consistência	0	≠ 0	≠ 0
Índice de entropia gen.	0	≠ 0	≠ 0
Disparidade do SHAP	0	> 0	< 0
Importância do atributo	0	≠ 0	≠ 0

Tabela 3: Descrição dos valores esperados com as medidas de justiça (em branco) e medidas do SHAP (em amarelo) nos cenários de igualdade e desigualdade entre grupos.

Pelo resultado da pesquisa de Verma et al. [62] que compara o resultado de mais de vinte definições de justiça, já é esperado que a classificação dos cenários poderão ser

diferentes de acordo com a definição de justiça usada. Porém, o objetivo nessa dissertação é avaliar se para alguma dessas medidas de justiça existe um maior consenso em relação ao resultado do SHAP.

5 EXPERIMENTOS, DADOS E RESULTADOS

Este capítulo contém a descrição dos aspectos da implementação do framework proposto e dos conjunto de dados utilizados, e a avaliação dos resultados obtidos.

5.1 Experimentos

Para testar o framework descrito no Capítulo 4 serão utilizados os modelos de Regressão Logística, Random Forest, Gradient Boostaing e Support Vector Machine (SVM). Para os experimentos será usada a biblioteca do scikit-learn[1], com cinco conjuntos de dados enviesados. Todos os testes serão feitos usando os mesmos hiperparâmetros nos modelos.

Para gerar os resultados com o SHAP será usado o código aberto disponível para esta metodologia.[2] A biblioteca AIF-360[3] será utilizada para aplicar a reponderação e calcular as medidas de justiça. Parte do código aberto[4] será usado para implementação da subamostragem parametrizada. A Figura 12 mostra em quais etapas do framework proposta para comparação das medidas de justiça e do SHAP essas bibliotecas serão utilizadas. Todos os testes e as bases utilizadas nesse experimento ficarão disponíveis em um repositório do github.[5]

Conforme descrito anteriormente, existem diversas variações de técnicas do SHAP para explicar o resultado do modelo. A única que é agnóstica ao modelo é o *Kernel SHAP*, por isso essa técnica será aplicada em todos os modelos. Para alguns modelos que serão utilizados também é possível aplicar técnicas do SHAP que são específicas para certos modelos, estas técnicas também serão utilizadas. A Tabela 4 contém as técnicas que serão usadas por modelo.

[1] http://scikit-learn.org
[2] https://github.com/slundberg/shap
[3] https://aif360.mybluemix.net
[4] https://github.com/vladoxNCL/fairCorrect/blob/master/Fairness-Multi.ipynb
[5] https://github.com/cesarojuliana/feature_importance_fairness_pt2

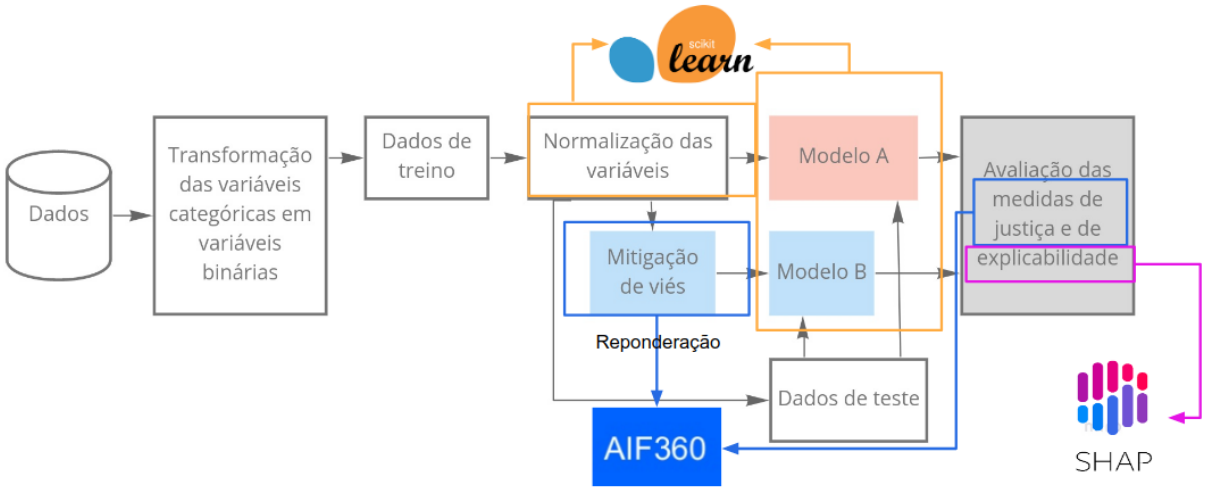


Figura 12: Framework proposto para comparação das medidas de justiça e do SHAP, com especificação das bibliotecas do Python que foram utilizadas em cada etapa.

Modelo	**Técnica**
Regressão Logística	Linear e Kernel SHAP
Random Forest	Tree e Kernel SHAP
Gradient Boostaing	Tree e Kernel SHAP
SVM	Kernel SHAP

Tabela 4: Modelos utilizados nesse projeto e técnicas do SHAP que serão aplicadas para interpretar o seu resultado.

Para juntar resultados de diferentes metodologias do SHAP é importante garantir que todos os valores do SHAP estejam na mesma unidade. Seria esperado que todas as metodologias do SHAP calculassem os valores na mesma unidade, porém não é isso que acontece. Todas as metodologias usam a mesma Equação (3.1) de atribuição aditiva do atributo para aproximar o valor previsto pelo modelo, e com isso é obtido o valor de SHAP. Porém, a diferença entre essas metodologias é que cada uma faz a aproximação da previsão do modelo em escalas diferentes.

Por exemplo, para o caso de um modelo de classificação algumas metodologias fazem a aproximação da probabilidade prevista pelo modelo, enquanto que outras fazem a aproximação do logaritmo da chance de sucesso previsto (log odds ou logit). Em algumas variantes do SHAP não é possível escolher a unidade da previsão do modelo para fazer a aproximação, e nesses casos não é simples adaptar o cálculo para permitir esse tipo de escolha.

Para padronizar será juntado apenas os resultados das técnicas de SHAP que fazem a aproximação da probabilidade prevista pelo modelo. Com o Linear SHAP não é possível aproximar a probabilidade prevista, e é sempre feita a aproximação do log odds previsto;

por isso não será juntado os resultados do Linear SHAP com os outros resultados obtidos para a maioria das análises.

A avaliação do Linear SHAP será feita em uma seção específica para isso, onde será comparado a consistência do seu resultado com o do Kernel SHAP para o modelo de Regressão Logística. As duas variações do Linear SHAP serão utilizadas, a que considera dependência entre os atributos e a que aplica intervenção nos dados. No caso do Tree SHAP apenas a variação que aplica intervenção nos dados será utilizada, pois ela é a única que permite definir para em todos os casos aproximar a probabilidade prevista pelo modelo.

5.2 Dados

Para treinar os modelos serão usados os conjuntos de dados do Adult, German, Default, COMPAS e PNAD. Com exceção do PNAD, o outros são conhecidas por terem viés, e por isso comumente usados para avaliar viés em modelo. O conjunto de dados do COMPAS foi obtido do ProPublica [4], o conjunto de dados do PNAD foi obtido do Centro de Estudos e Metrópole da USP [40] e os conjuntos de dados do Adult, German e Default foram obtidos do repositório da UCI [24].

A Tabela 5 traz um resumo por conjunto de dados de quais são as variáveis alvo e sensível, com descrição dos seus grupos privilegiado e desprivilegiado. A Figura 13 mostra por grupo o percentual de casos que a variável alvo tem valor 1 (em azul) e valor 0 (em cinza). Nela é perceptível que em todos os conjuntos de dados é maior o percentual de casos em que a variável alvo tem valor 1 do grupo privilegiado em relação ao grupo desprivilegiado. Como foi determinado o valor 1 para a classe desejável da variável alvo, esse maior percentual de casos do grupo privilegiado evidencia a situação de privilégio do grupo.

O conjunto de dados do Adult é o resultado de um censo realizado em 1994 nos Estados Unidos, em que o objetivo é predizer se o salário anual de uma pessoa é superior ou inferior a 50 mil dólares. O gênero será usado como variável sensível conforme feito por Kamishima et al. [35]. O conjunto de dados do Default contêm informação de clientes de cartão de crédito de Taiwan em 2005, sendo objetivo prever se o cliente irá pagar a fatura do cartão. O gênero será considerado como variável sensível conforme feito por Berk et al. [11].

O conjunto de dados do German contêm informação bancária de clientes, sendo o

Dataset	Variável alvo =1 (Y=1)	Variável sensível	Grupo	
			Privilegiado	Desprivilegiado
Adult	Salário anual acima 50 mil dólares	Gênero	Masculino	Feminino
Compas	Não reincidir em 2 anos	Raça	Caucasianos	Não caucasianos
German	Baixo risco de crédito	Idade	Acima de 25 anos	Abaixo de 25 anos
Default	Pagamento da fatura do cartão	Gênero	Masculino	Feminino
PNAD gênero	Salário mensal acima 2500 reais	Gênero	Masculino	Feminino
PNAD raça	Salário mensal acima 2500 reais	Raça	Branco e amarelo	Preto, pardo e indígena

Tabela 5: Descrição dos conjuntos de dados utilizados nesse trabalho.

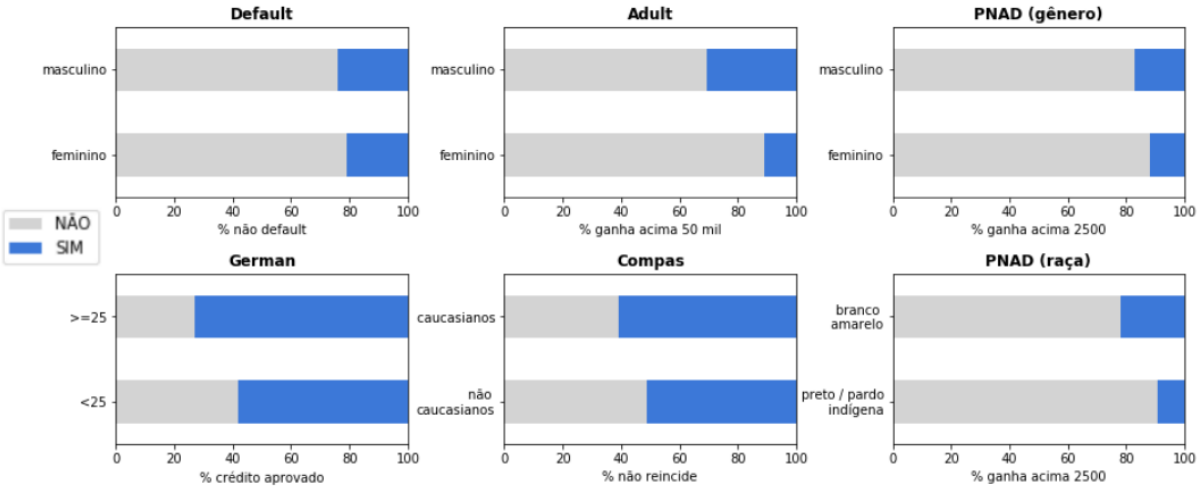


Figura 13: Percentual de amostras para os grupos privilegiados e desprivilegiados que a variável alvo tem valor 1 (SIM, em azul) e valor 0 (NÃO, em cinza).

objetivo classificar o risco de crédito como bom ou ruim para esses clientes. A variável sensível considerada será idade, conforme feito por Kamiran e Calders [33]. O conjunto de dados do COMPAS contém informação de réus criminais de Broward County, na Florida, sendo o objetivo prever a probabilidade do réu reincidir em um período de dois anos. A raça será usada como variável sensível, reclassificando essa variável em dois grupos, caucasianos e não caucasianos.

Nos dados do COMPAS serão usados os mesmos filtros que foram usados por Angwin et al. [5], e serão selecionadas as seguintes variáveis: *race*, *age*, *c_charge_degree*, *v_score_text*, *sex*, *priors_count*, *is_recid*, *days_b_screening_arrest*, *v_decile_score*, *two_year_recid*. No conjunto de dados do Adult será excluída a variável *fnlwgt*, pois essa variável não tem

relação com a variável prevista. No conjunto de dados do Default será excluída a variável *id* pelo mesmo motivo.

O conjunto de dados do PNAD (Pesquisa Nacional por Amostra de Domicílios Contínua)[6] é o resultado de uma pesquisa realizada nos domicílios brasileiros para avaliar a força de trabalho e desenvolvimento socioeconômico do país. Essa pesquisa acontece anualmente, e nesse trabalho será utilizado o resultado de 2015. Com base em uma análise feita nele é perceptível uma relevante disparidade da renda mensal com relação ao gênero e a raça, mostrado na Figura 14. O resultado completo da análise está disponível online.[7]

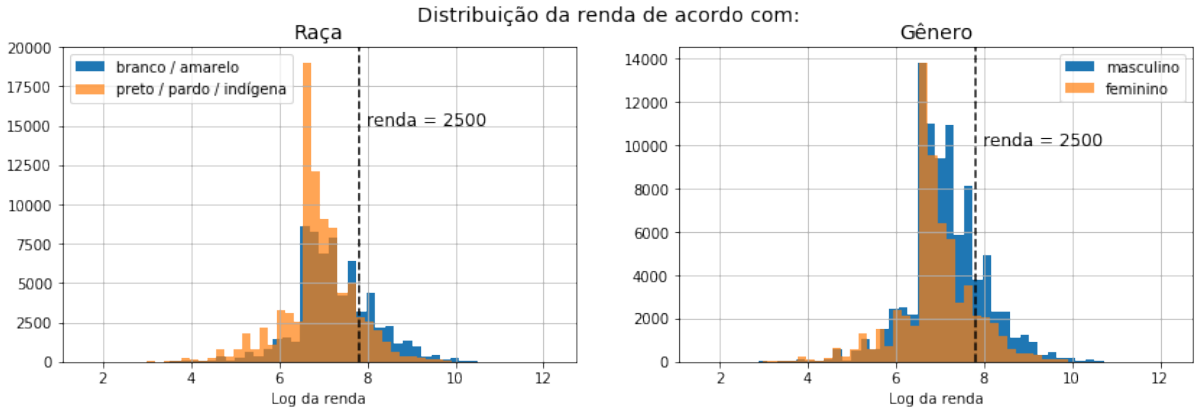


Figura 14: Distribuição da renda mensal por gênero e raça de acordo resultado do PNAD de 2015.

Visto a relevante desigualdade de renda existente em relação a raça e ao gênero, e para manter o padrão dos outros dados que contém uma única variável sensível binária, dois tipos de modelo serão construídos: um modelo com a variável sensível raça que não tem a variável gênero, e outro modelo com a variável sensível gênero que não tem a variável raça. Ambos os modelos terão como objetivo prever a renda da pessoa.

Além disso, a variável contínua de renda será transformada em uma variável binária categórica. Para isso é necessário definir um limiar para dividir a renda em duas categorias de forma a atender a dois critérios: (i) deve ser um valor alto o suficiente para diferenciar a renda de acordo com a variável sensível; (ii) não deve ser tão alto, para evitar que o conjunto de dados fique muito desbalançado. Considerando esses critérios, por meio de inspeções visuais foi selecionado como limiar renda de R$ 2500. O resultado das Figuras 14 e 15 mostra que com esse valor é perceptível a diferença de renda entre os grupos, além disso o conjunto de dados não ficou muito desbalanceado.

[6] https://www.ibge.gov.br/estatisticas/sociais/educacao/9173-pesquisa-nacional-por-amostra-de-domicilios-continua-trimestral.html?=t=o-que-e

[7] https://github.com/cesarojuliana/feature_importance_fairness_pt2/blob/master/EDA - PNAD.ipynb

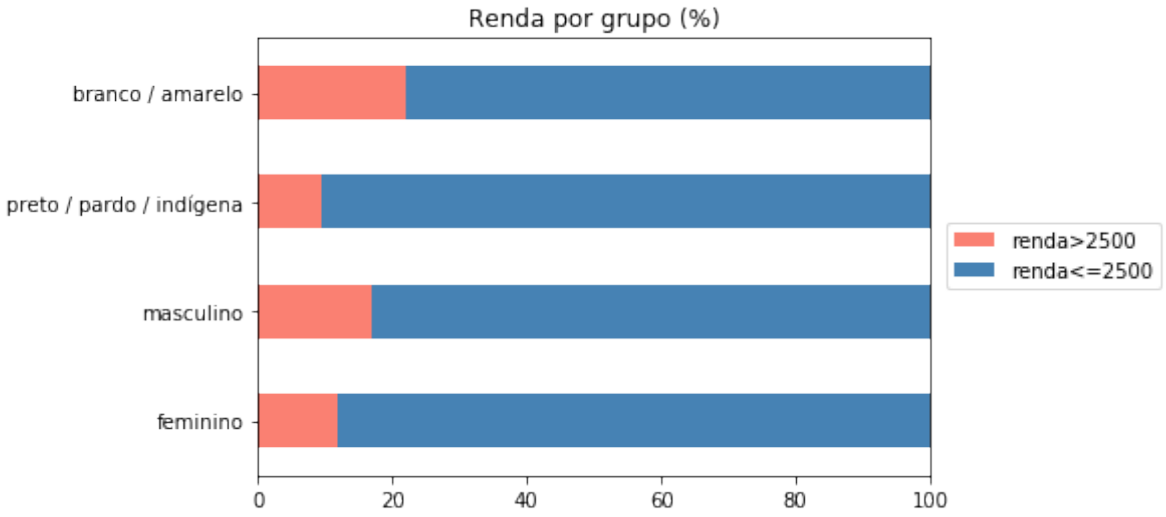


Figura 15: Percentual de pessoas que ganham acima de R$ 2500, em vermelho, e abaixo de R$ 2500, em azul, de acordo com gênero e raça.

O conjunto de dados do PNAD de 2015 contém informação de aproximadamente 357 mil entrevistados, para uso nessa pesquisa foram selecionados os casos em que havia registro da renda com valor maior que zero, o que reduziu o número de registros para aproximadamente 153 mil. O conjunto de dados contém 436 variáveis, sendo que diversas possuem muitos valores nulos, foram selecionadas as variáveis com poucos valores nulos e que por conhecimento de negócio são mais diretamente relacionadas a renda.

Com isso as seguintes variáveis foram selecionadas: $V0302$ (gênero), $V8005$ (idade), $V0404$ (raça), $V4011$ (estado civil), $V4803$ (anos estudo), $V4707$ (faixa de horas de trabalho semanal), $V4706$ (posição no trabalho), $V4809$ (grupamento de atividade principal do empreendimento do trabalho), $V4810$ (grupamento ocupacionais do trabalho) e $V4718$ (renda).

Usando as variáveis selecionadas para caracterizar os entrevistados, foi avaliado se a diferença de renda é causada exclusivamente por diferença nas características descritas por essas variáveis. Ou se majoritariamente entrevistados do grupo privilegiado com características muito semelhantes as do grupo desprivilegiado tem renda superior.

Para verificar se esse efeito acontece será feito um teste simples. Será treinado um modelo com as variáveis selecionadas, e será gerado previsões para amostras do grupo privilegiado. Depois será alterado artificialmente o valor da variável sensível dessas amostras tornando-as pertencentes ao grupo desprivilegiado, e com estas amostrar será gerado novas previsões. Em situação de justiça, pela definição de justiça contrafactual, a previsão não deveria alterar, porém caso haja favorecimento do grupo privilegiado será verificado

diminuição de casos que a renda prevista foi acima de R$ 2500.

Esse teste será realizado para o caso com gênero como variável sensível e para o caso com raça como variável sensível. Para não restringir a análise a um tipo de modelo, a avaliação será feita com dois tipos: Gradient Boosting e Regressão Logística. Em ambos os casos será utilizado o limiar de 0.5 em probabilidade posterior para classificar em 0 ou 1.

A distribuição de probabilidade prevista pelos modelos com a variável sensível original (sem alteração) e com variável sensível invertida são mostrados nas Figuras 16 e 17. A Figura 16 mostra a distribuição de probabilidade dos modelos com variável sensível gênero, e na Figura 17 a distribuição de probabilidade com a variável sensível raça. Em ambos os casos verifica-se deslocamento da distribuição de probabilidade, com consequente diminuição de casos previstos de renda acima de R$ 2500 ao inverter o valor da variável sensível.

Avaliando o modelo de Gradient Boosting com variável sensível gênero, foram previstas renda acima de R$ 2500 para as amostras com gênero masculino em 72 casos, porém ao artificialmente mudar apenas o gênero nessas amostras para feminino houve diminuição para 43 casos com renda acima de 2500, o que representa uma diminuição de aproximadamente 40% dos casos.

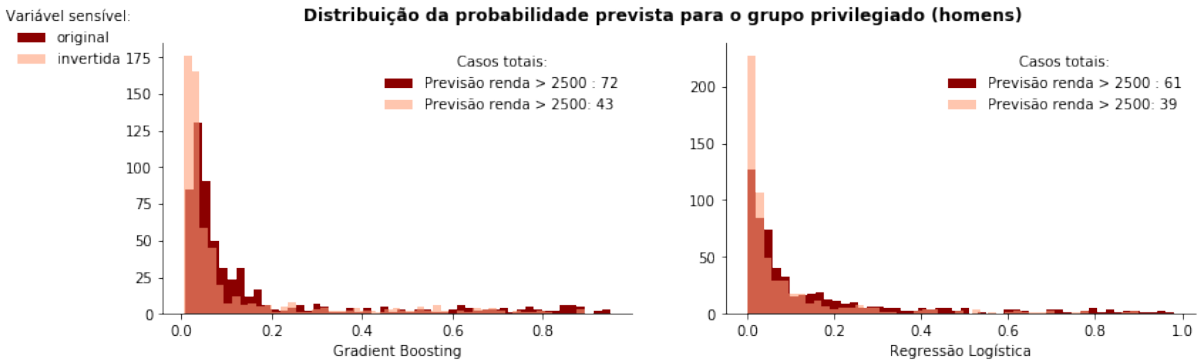


Figura 16: Distribuição da probabilidade prevista para entrevistados do gênero masculino, de modelos treinados com o conjunto de dados do PNAD para prever se a renda é acima de R$2500. Em marrom está representado a distribuição de probabilidade com o valor original da variável gênero, e em salmão a previsão para as mesmas amostras com o valor da variável gênero alterada.

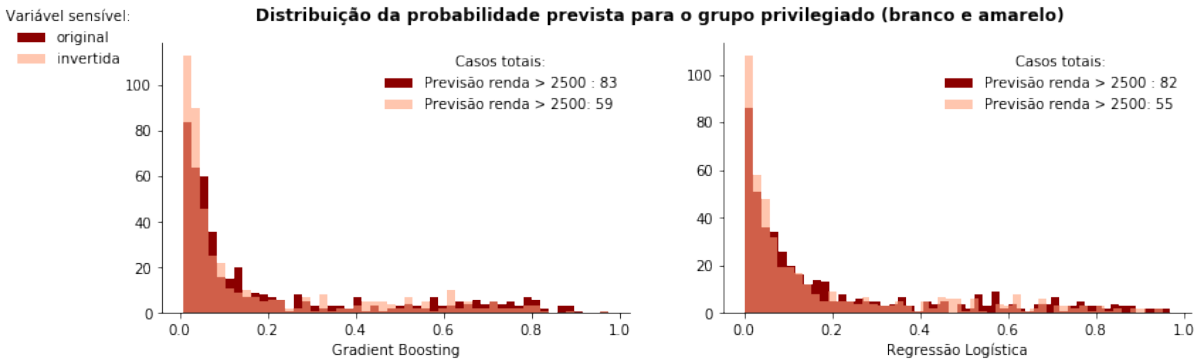


Figura 17: Distribuição da probabilidade prevista de modelos treinados com o conjunto de dados do PNAD para as amostras com raça branca e amarela. Em marrom está representado a distribuição de probabilidade com o valor original da variável raça, e em salmão a previsão para as mesmas amostras com o valor da variável raça alterada.

5.3 Comparação das medidas de justiça e do SHAP

Nesta seção os resultados obtidos com as medidas de justiça e as do SHAP serão comparados. Sendo comparado o cenário de justiça obtidos com cada medida. Além disso, será calculado a variação de valores de cada medida nos modelos com e sem mitigação de viés, para verificar se existe semelhança nessa variação entre as medidas.

5.3.1 Comparação do cenário de justiça

Anteriormente 3 cenários de justiça foram descritos: *igualdade entre grupos*, *favorecimento do grupo privilegiado* e *favorecimento do grupo desprivilegiado*. Para cada cenário foi especificado na Tabela 3 os valores esperados com as medidas de justiça e as do SHAP. Já é esperado que a classificação do cenário poderá ser diferente de acordo com a medida usada, porém será avaliado se para alguma das medidas de justiça existe maior consenso em relação ao cenário classificado com as medidas do SHAP.

A comparação de cenário será feita usando os resultados dos modelos com mitigação de viés, pois com estes modelos é esperado qualquer um dos três cenários, enquanto que nos modelos sem mitigação de viés é esperado na maioria dos casos o cenário de favorecimento do grupo privilegiado. Isso porque em todos os conjuntos de dados usados nesse projeto o grupo privilegiado é consideravelmente mais favorecido (conforme mostrado na Tabela 5). Por ser tão significativa a diferença, imagina-se que sem a aplicação de técnica para reduzir o viés do modelo a maioria das medidas de justiça mostrem esse privilégio.

Com algumas das medidas é possível avaliar apenas se existe desigualdade entre os grupos, sem conseguir diferenciar qual o grupo que está sendo favorecido. Isto acontece

com as medidas de *consistência*, *índice de entropia generalizada* e *disparidade do SHAP*. Conforme retratado na Tabela 3, essas medidas possuem o mesmo valor para os cenários de favorecimento do grupo privilegiado e favorecimento do grupo desprivilegiado, o que evidencia que não é possível diferenciar esses cenários. Por isso, a avaliação dessas medidas será feita considerando apenas dois cenários: igualdade entre os grupos e desigualdade entre os grupos.

Para os casos em que um cenário acontece com um valor exato da medida, consideramos uma tolerância de $\pm 10^{-3}$ do valor para considerar como pertencente ao cenário. Fizemos isso pois sabemos que no cálculo das medidas são feitas diversas aproximações computacionais. Conforme exposto na Tabela 3, o único cenário com valor exato é o de igualdade entre os grupos que acontece quando a medida é igual a zero.

A Tabela 6 mostra a comparação por cenário entre cada medida de justiça com a disparidade do SHAP. Para cada medida foi calculado o percentual de resultados por cenário, dividindo entre os casos condizentes e contradizentes com o cenário da disparidade do SHAP. A Tabela 7 apresenta esse mesmo tipo de avaliação por cenário, comparando as medidas de justiça com a importância do atributo.

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	1.4%	4.9%	20.8%	50.0%	16.0%	6.9%
Igualdade de oportunidade	2.8%	2.8%	18.1%	37.5%	23.6%	15.3%
Contrafactual	**22.2%**	4.9%	**24.3%**	3.5%	**42.4%**	2.8%

	Igualdade entre os grupos		Desigualdade entre os grupos	
	Condizente	Contradizente	Condizente	Contradizente
1 - consistência	0.0%	0.0%	71.5%	28.5%
Índice de entropia gen.	0.0%	0.0%	71.5%	28.5%

Tabela 6: Análise do percentual de casos por cenário de justiça em que as medidas de justiça são condizentes e contradizentes com a disparidade do SHAP.

A Tabela 8 expõe o percentual de casos em que o cenário da medida de justiça foi condizente com o da disparidade do SHAP, e a Tabela 9 o percentual de casos em que o cenário da medida de justiça foi condizente com a importância do atributo. Nessas tabelas os resultados estão separados por metodologia do SHAP, sendo que em completo está presente a avaliação com todos os resultados.

Avaliando os resultados verifica-se que para todos os cenários e todas as metodologias do SHAP a medida contrafactual foi a mais condizente com a disparidade do SHAP e com a importância do atributo.

Comparação com importância do atributo

	Igualdade entre os grupos		Desigualdade entre os grupos	
	Condizente	Contradizente	Condizente	Contradizente
Paridade estatística	2.1%	4.2%	56.3%	37.5%
Igualdade de oportunidade	4.2%	1.4%	59.0%	35.4%
Contrafactual	**27.1%**	0.0%	**60.4%**	12.5%
1 - consistência	0.0%	0.0%	60.4%	39.6%
Índice de entropia gen.	0.0%	0.0%	60.4%	39.6%

Tabela 7: Avaliação entre os cenário de justiça do percentual de casos em que as medidas de justiça e a importância do atributo sensível são condizentes e contradizentes.

% Condiz com Disparidade do SHAP

	Completo	Tree SHAP	Kernel SHAP
Paridade estatística	38%	38%	38%
Igualdade de oportunidade	44%	43%	37%
Contrafactual	**89%**	**78%**	**91%**
1 - consistência	72%	62%	74%
Índice de entropia gen.	72%	62%	74%

Tabela 8: Avaliação separada por técnica do SHAP do percentual de casos em que as medidas de justiça e a disparidade do SHAP são condizentes. Completo apresenta a comparação com todos resultados gerados.

% Condiz com importância do atributo

	Completo	Tree SHAP	Kernel SHAP
Paridade estatística	58%	46%	59%
Igualdade de oportunidade	63%	54%	63%
Contrafactual	**88%**	**81%**	**87%**
1 - consistência	60%	49%	61%
Índice de entropia gen.	60%	49%	61%

Tabela 9: Cálculo do percentual de casos em que as medidas de justiça condiz com importância do atributo. Avaliação feita de forma separada por técnica do SHAP, sendo que completo compara todos resultados.

5.3.2 Comparação da variação dos resultados

Nesta seção será avaliado a variação de resultado entre os modelos com e sem mitigação de viés para verificar se o modelo ficou mais justo com a mitigação de viés. A variação de resultado foi calculada pela equação abaixo:

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

sendo que $medida_i$ é a medida de justiça ou do SHAP, $modelo_A$ é o modelo sem mitigação de viés, e $modelo_{B_j}$ é o modelo que foi aplicado a técnica j de mitigação de viés.

Variação de resultado positiva indica que a mitigação de viés tornou o modelo mais justo, enquanto que negativa indica que o modelo ficou mais injusto. Está sendo calculado a diferença do módulo das medidas para permitir sempre a mesma da interpretação do resultado, de que quanto menor ele for mais justo ficou o modelo com redução de viés.

A Figura 18 mostra o histograma da variação de resultado das medidas, sendo que os histogramas em azul são das medidas de justiça, e em laranja das medidas do SHAP. Os resultados nos gráficos mostram que apenas a medida de paridade estatística considerou que quase todos os modelo ficaram mais justo com a mitigação de viés, enquanto que de acordo com as outras medidas, em aproximadamente metade dos casos os modelos ficaram mais justos com a mitigação de viés e a outra metade dos casos ficaram mais injustos.

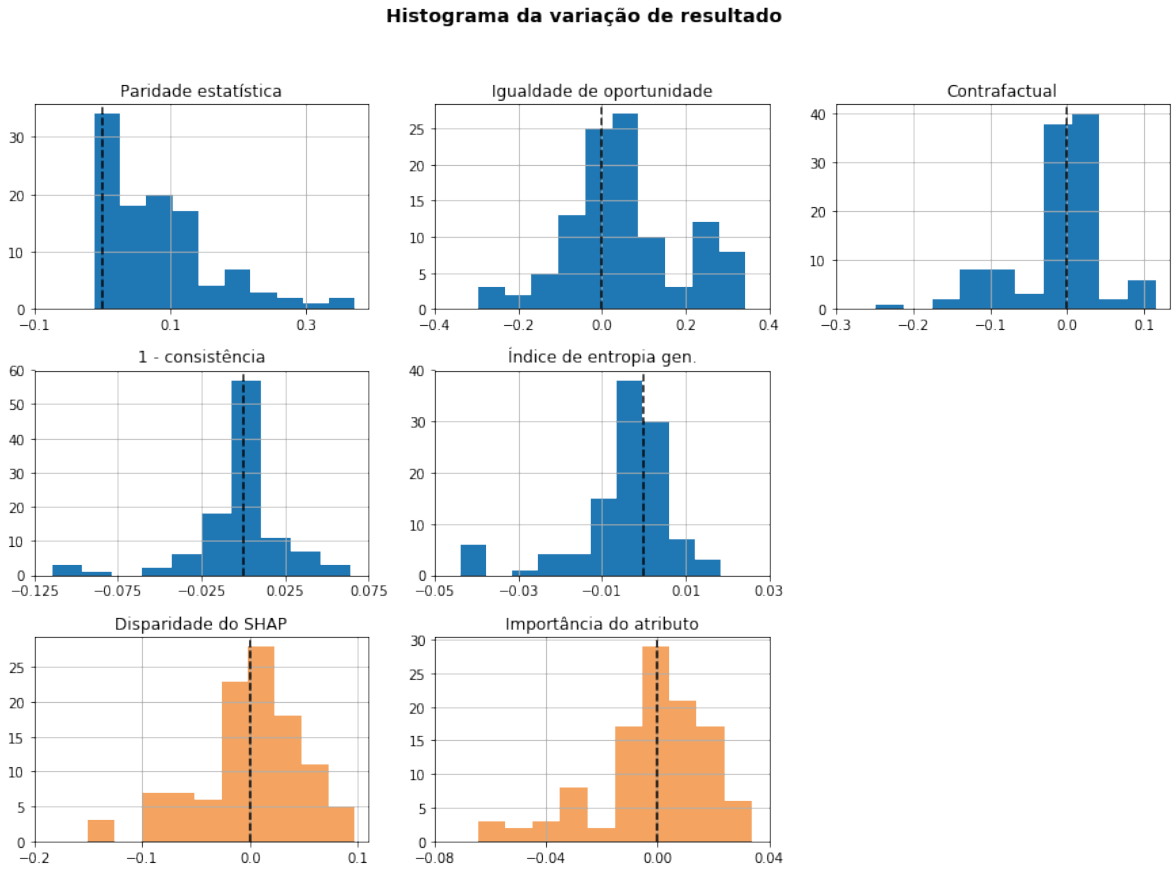


Figura 18: Histograma da variação de resultados por medida de justiça, em azul, e medida do SHAP, em laranja.

A Figura 19 mostra heatmaps com a correlação calculada entre a variação de resultado das medidas. No heatmap da esquerda foi calculado a correlação entre a variação das medidas de justiça e da disparidade do SHAP, e no heatmap da direita foi calculada a

correlação entre a variação das medidas de justiça e da importância do atributo. Nos dois gráficos foram feitas análise separada de acordo com a metodologia do SHAP utilizada, sendo que completo mostra a correlação com todos os resultados.

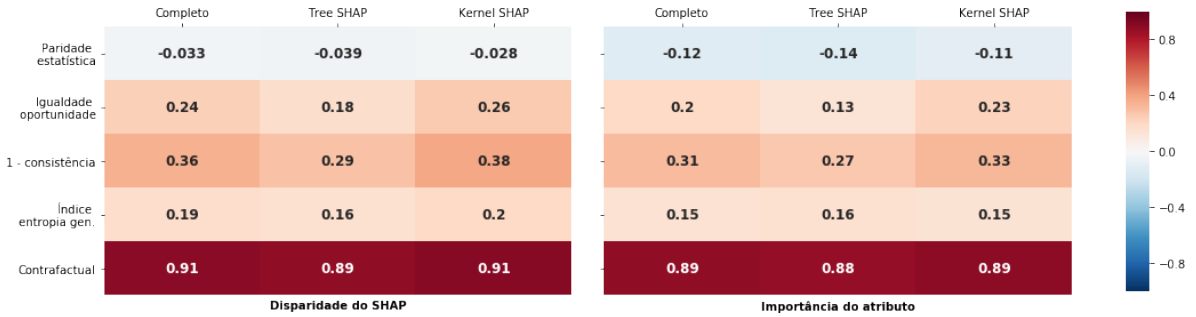


Figura 19: Correlação da variação de resultados dos modelos com e sem mitigação de viés. Avaliação separado pelas técnicas do SHAP, completo mostra a correlação com todos os resultados. No gráfico da esquerda foi calculada a correlação entre a variação das medidas de justiça e da disparidade do SHAP. No gráfico da direita foi calculada a correlação entre a variação das medidas de justiça e da importância do atributo.

Os resultados dos dois heatmaps da Figura 19 mostram que para todas as metodologias do SHAP a correlação da medida contrafactual com a disparidade do SHAP e com a importância do atributo é consideravelmente maior do que a correlação com as outras medidas de justiça. Enquanto que a medida de paridade estatística tem a menor correlação com as medidas do SHAP, sendo desprezível para todas as metologias do SHAP.

Quando a correlação obtida com as diferentes técnicas do SHAP é comparada, apesar de sempre ser obtida a mesma conclusão de qual medida de justiça tem a maior correlação e qual medida tem a menor correlação, verifica-se que o resultado varia um pouco entre as técnicas. Por exemplo, a correlação da igualdade de oportunidade em relação as duas medidas do SHAP é consideravelmente maior com o Kernel SHAP do que com o Tree SHAP. Para entender melhor essa diferença, na próxima seção será avaliado com mais detalhe os resultados obtidos com cada tipo de modelo.

5.4 Comparação das metodologias do SHAP

O Kernel SHAP é a única metodologia do SHAP agnóstica ao tipo de modelo, porém para tornar a metodologia genérica para qualquer modelo o seu cálculo é mais custoso. Por isso, existem metodologias do SHAP que são específicas para certos modelos e com isso é possível tornar o cálculo mais eficiente. Umas destas metodologias é o Tree SHAP, que é específico para modelos baseados em árvore. Todas as metodologias do SHAP

tem como base o conceito do valor de Shapley, por isso seria esperado obter os mesmos resultados com elas.

Porém, os resultados das seções anteriores mostram que em alguns casos os resultados do Kernel e Tree SHAP foram diferentes. Um dos fatores que pode justificar essa diferença é que o Kernel SHAP e o Tree SHAP foram usados com diferentes tipos de modelos. Para avaliar se de fato existe diferença entre o Kernel e Tree SHAP nesta seção a comparação será restrita a resultados gerados pelos mesmos tipos de modelos. Como a metodologia do Tree SHAP é específica para modelos de árvore, nessa pesquisa ela foi utilizada apenas com os modelos Random Forest e Gradient Boosting.

Além disso, nesta seção será comparado os resultados do Linear e Kernel SHAP. Todos os resultados do Linear SHAP foram calculados com Regressão Logística, por isso serão selecionados apenas os resultados do Kernel SHAP obtidos com esse tipo de modelo. Para permitir melhor entendimento de possíveis diferenças entre as metodologias do SHAP, nesta seção será avaliado com mais detalhe resultados específicos obtidos com alguns modelos, e será analisado os gráficos que são gerados pela biblioteca do SHAP para os casos selecionados.

5.4.1 Kernel e Tree SHAP

Para comparar o Tree e o Kernel SHAP inicialmente foi feita a avaliação dos cenários de justiça para verificar se os modelos com mitigação de viés ficaram mais justos, e em todos os casos foram obtidos os mesmos cenários com o Tree SHAP e o Kernel SHAP. Depois foi feita a análise da correlação entre a variação das medidas de justiça e do SHAP.

A Figura 20 mostra os heatmaps calculados, sendo o da direita a correlação entre as medidas de justiça com a disparidade do SHAP, e o da esquerda a correlação entre as medidas de justiça e a importância do atributo. Diferentemente dos resultados obtidos na Figura 19, agora que está sendo avaliado apenas os resultados gerados pelos modelos de árvore, verifica-se que as correlações obtidas com o Tree e o Kernel SHAP são muito semelhantes.

As Figuras 32 e 33 do Apêndice A mostram o resultado completo obtido com todas as medidas de justiça e do SHAP. Avaliando os gráficos, verifica-se que de forma geral as medidas do SHAP calculadas com o Kernel e Tree SHAP para o mesmo modelo foram semelhantes. Como esse padrão de semelhança se repete entre os conjuntos de dados e pela grande quantidade de resultados gerados, a comparação será restrita ao conjunto de

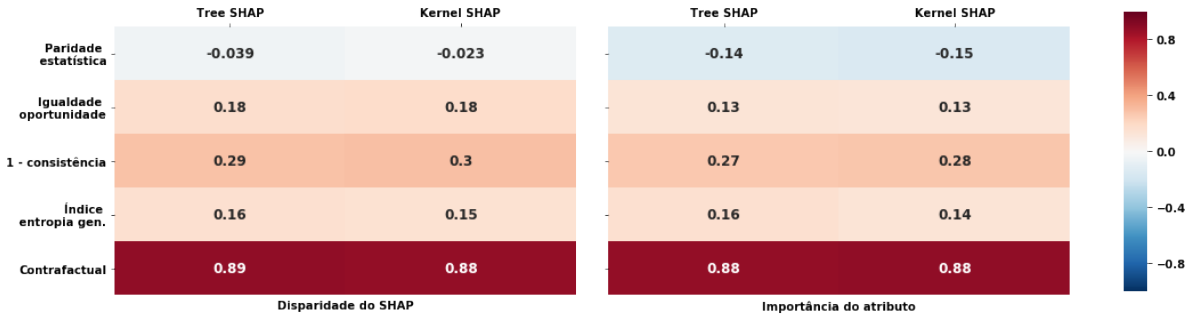


Figura 20: Correlação da variação de resultados dos modelos de árvore com e sem mitigação de viés. Avaliação separado pelas técnicas do Kernel SHAP e Tree SHAP. No heatmap da esquerda foi calculada a correlação entre a variação das medidas de justiça e da disparidade do SHAP. No heatmap da direita foi calculada a correlação entre a variação das medidas de justiça e da importância do atributo.

dados do PNAD com variável sensível raça.

A Figura 21 mostra os resultados das medidas do SHAP para esse conjunto de dados. O eixo x indica a metodologia do SHAP e a abreviação do modelo, sendo que *gb* se refere a Gradient Boosting e rf a Random Forest. A legenda indica o tipo de modelo utilizado, a linha vermelha (*orig*) mostra o resultado do modelo sem mitigação de viés, a linha azul com quadrado (*rw*) mostra o resultado do modelo com aplicação da reponderação, a linha cinza com triângulo (*usd-1*) mostra o resultado do modelo com aplicação da subamostragem com $d = -1$, e a linha azul com círculo (*usd-0*) mostra o resultado do modelo com aplicação da subamostragem com $d = 0$.

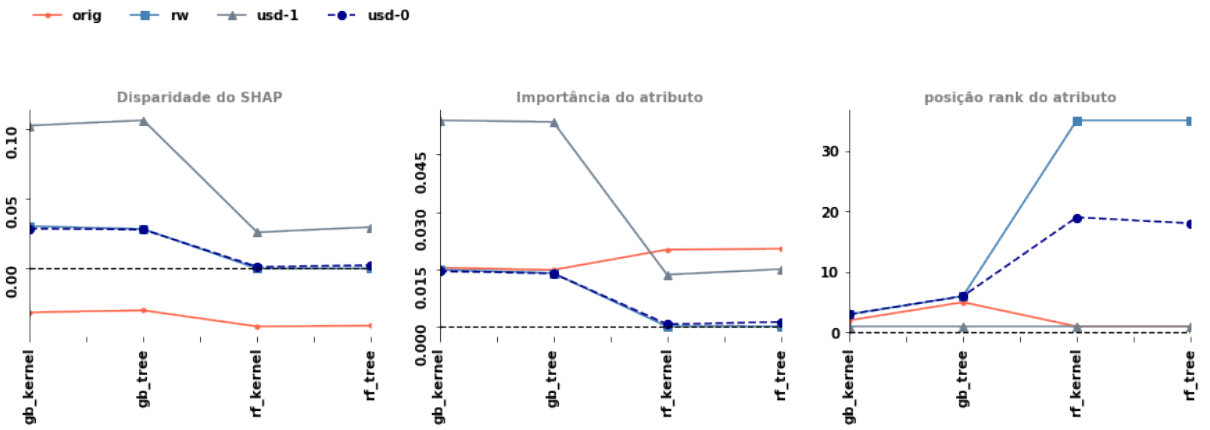


Figura 21: Avaliação de medidas de justiça e medidas obtidas com SHAP com conjunto de dados do PNAD para os modelos de Gradient Boosting e Random Forest. As medidas do SHAP foram calculadas com duas metologias: Kernel SHAP e Tree SHAP.

Analisando os resultados do Kernel e Tree SHAP, verifica-se que a disparidade do SHAP e a importância do atributo foram muito semelhantes. Enquanto que posição do

atributo variou um pouco para alguns modelos, esses modelos serão selecionados para análise com mais detalhe. A maior variação na posição do atributo foi verificada com o modelo de Gradient Boosting sem mitigação de viés (*orig*), e com o Gradient Boosting com subamostragem com $d = 0$ (*usd0*). Para avaliar melhor esses casos foram gerados os gráficos de *summary plot* nas Figuras 22 e 24, e os gráficos de *partial dependence plot* (PDP) nas Figuras 23 e 25.

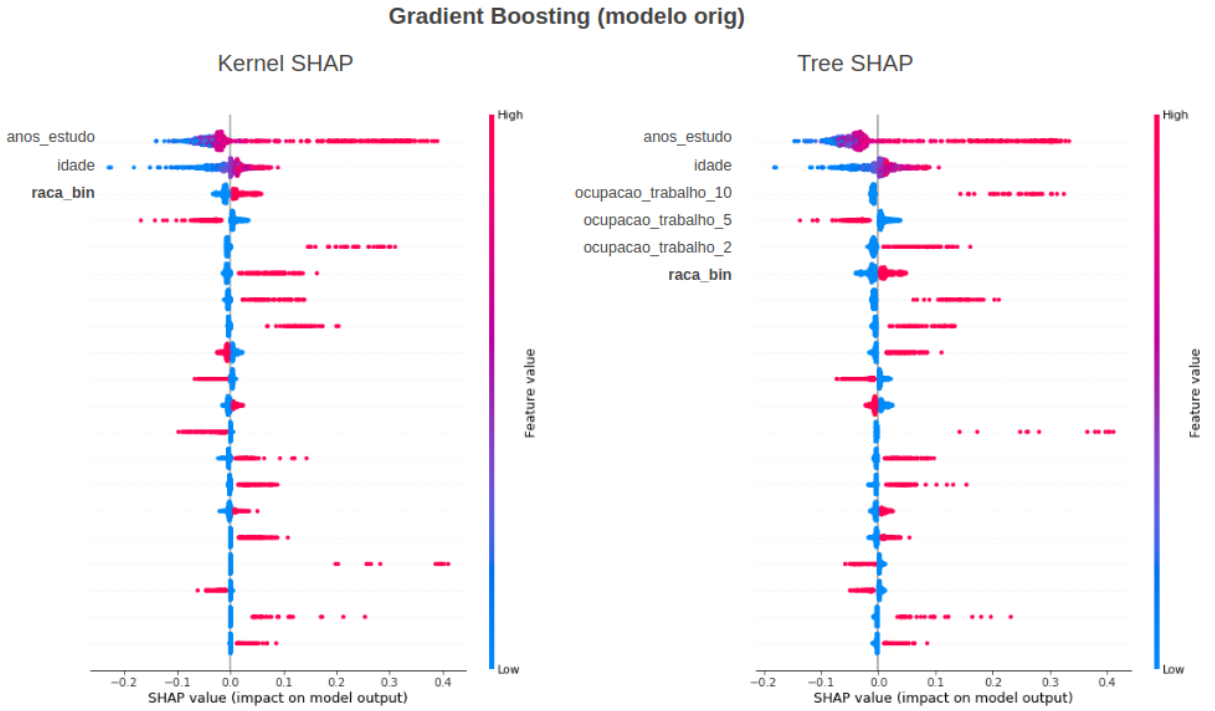


Figura 22: Gráficos de summary plot gerados para um modelo de Gradient Boosting sem mitigação de viés. O gráfico da esquerda foi obtido com a metodologia do Kernel SHAP, e o da direita com o Tree SHAP.

Comparando os gráficos de summary plot do Kernel e Tree SHAP na Figura 22 para o modelo de Gradient Boosting sem mitigação de viés, percebe-se uma diferença significativa na posição do atributo sensível raça com as duas metodologias. Pelo resultado do Kernel SHAP, raça foi o terceiro atributo mais importante no modelo, enquanto que pelo resultado do Tree SHAP foi o sexto atributo mais importante no modelo. Apesar da diferença na posição do atributo, comparando os gráficos de PDP na Figura 23 para o atributo raça, nos dois gráficos de PDP vemos um comportamento muito semelhante com as metodologias do Kernel e Tree SHAP. Pois com ambas as metodologias a faixa de variação de valores é semelhante, e o atributo anos de estudo foi considerado o mais correlacionado com raça.

Para o modelo de Gradient Boosting que foi aplicado subamostragem para mitigação de viés também existe diferença significativa na posição do atributo raça. Os gráficos de summary plot da Figura 24 evidenciam essa diferença, onde com o Kernel SHAP raça foi

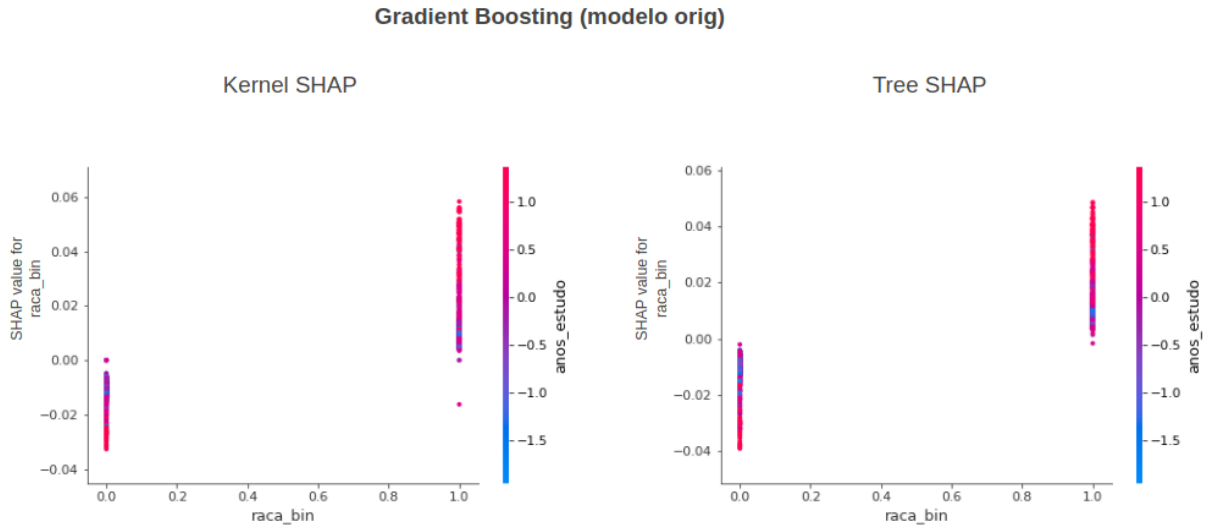


Figura 23: Gráficos de PDP do atributo raça do conjunto de dados do PNAD que foi treinado com um modelo de Gradient Boosting. O gráfico da esquerda foi gerado com a metodologia do Kernel SHAP, e o da direita com o Tree SHAP.

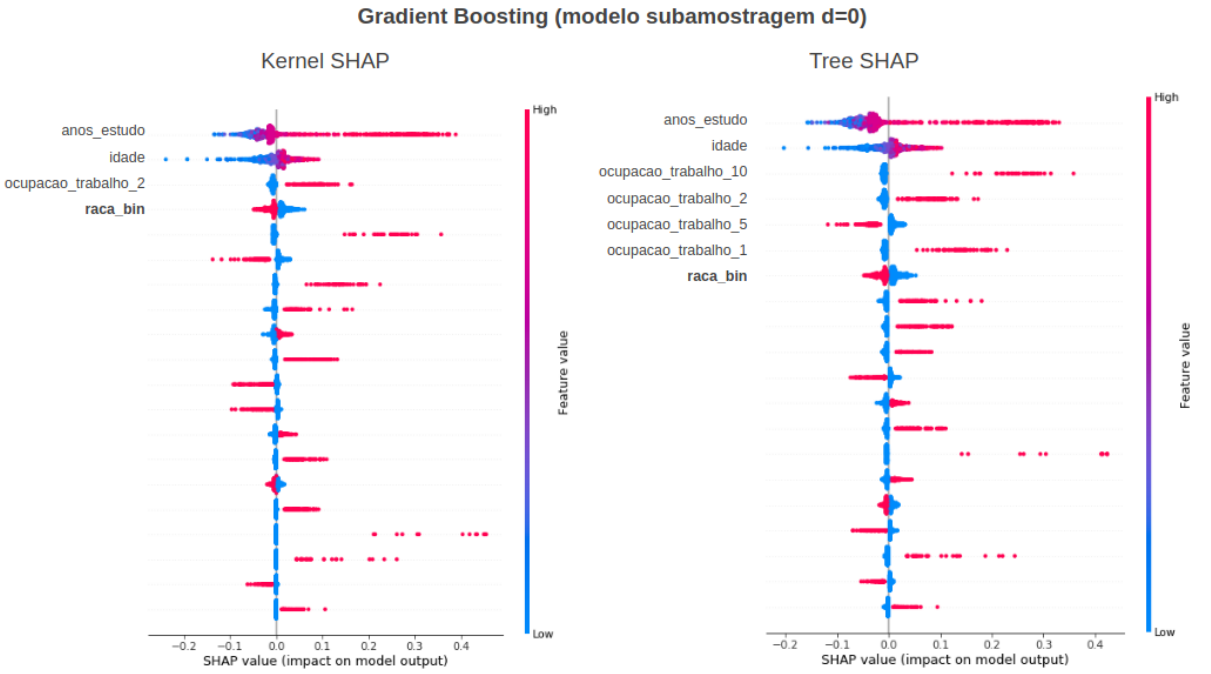


Figura 24: Gráficos de summary plot gerados com o conjunto de dados do PNAD para um modelo de Gradient Boosting que foi aplicado a técnica de subamostragem para mitigação de viés. A esquerda o resultado foi gerado com a metodologia do Kernel SHAP, e a direita com o Tree SHAP.

o quarto atributo mais importante, enquanto que com o Tree SHAP foi o sétimo atributo mais importante. Apesar dessa diferença de posição com as metodologias, os gráficos de PDP da Figura 25 mostram um comportamento semelhante nos valores de SHAP. Com o Kernel e Tree SHAP o range de variação de valores é semelhante, e o atributo mais

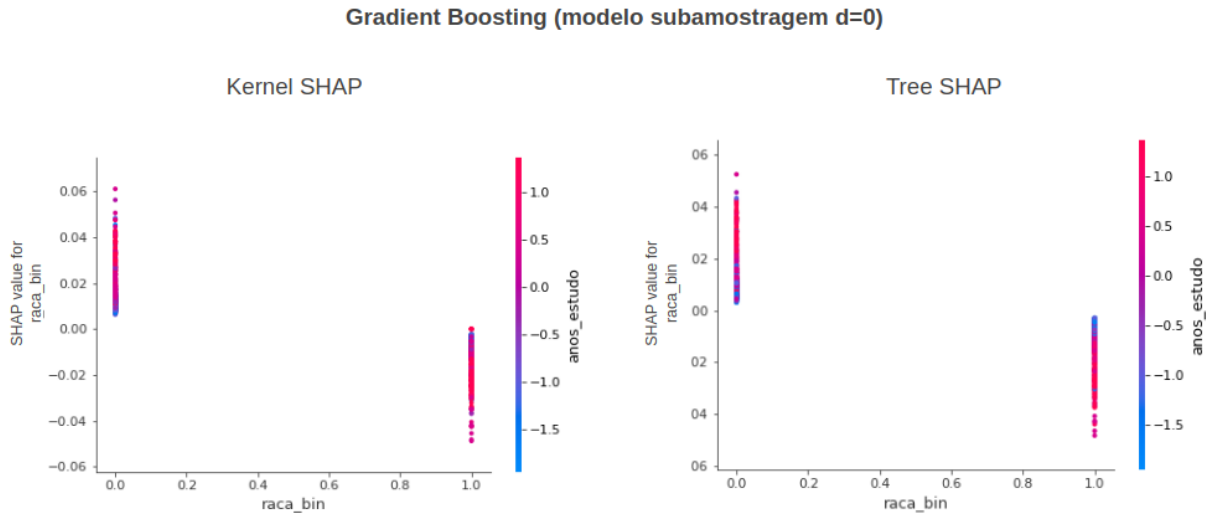


Figura 25: Gráficos de PDP do atributo raça gerados com um modelo de Gradient Boosting com subamostragem treinado com o conjunto de dados do PNAD. O gráfico da esquerda foi gerado com o Kernel SHAP e o da direita com o Tree SHAP.

correlacionada é anos de estudos.

5.4.2 Kernel e Linear SHAP

Agora o Kernel e Linear SHAP serão comparados, em todas as análises feitas anteriormente os resultados do Linear SHAP não foram avaliados pois esta metodologia calcula os valores de SHAP aproximando sua soma pelo log odds previsto do modelo. Enquanto que as outra metodologias do SHAP utilizadas nessa pesquisa é possível configurar para aproximar a probabilidade prevista pelo modelo. Por isso, a escala dos valores de SHAP obtidos com o Linear SHAP é consideravelmente maior que a obtido com as outras metodologias. Dessa forma, nessa seção o foco será em comparar padrões existentes do Kernel e Linear SHAP, diferentemente do que foi feito em seções anteriores onde foram comparados os valores de SHAP obtidos.

O Linear SHAP é específico para modelos lineares, como o único modelo linear que utilizamos foi a Regressão Logística, selecionaremos apenas os resultados do Kernel SHAP gerados para este tipo de modelo. Existem duas variações do Linear SHAP, uma que considera a dependência entre os atributos, e outra que considera independência aplicando intervenção nos dados, vamos usar estas duas variações do Linear SHAP.

A Figura 26 mostra os heatmaps obtidos com a análise da correlação entre a variação das medidas de justiça e do SHAP. Nestes heatmaps os resultados foram separados por metodologias do SHAP. O heatmap da esquerda mostra a correção entre as medidas

de justiça e a disparidade do SHAP, e o heatmap da direita mostra a correlação entre as medidas de justiça e a importância do atributo. Avaliando as correlações obtidas percebe-se que as duas variações do Linear SHAP tiverem resultados consideravelmente diferentes do Kernel SHAP.

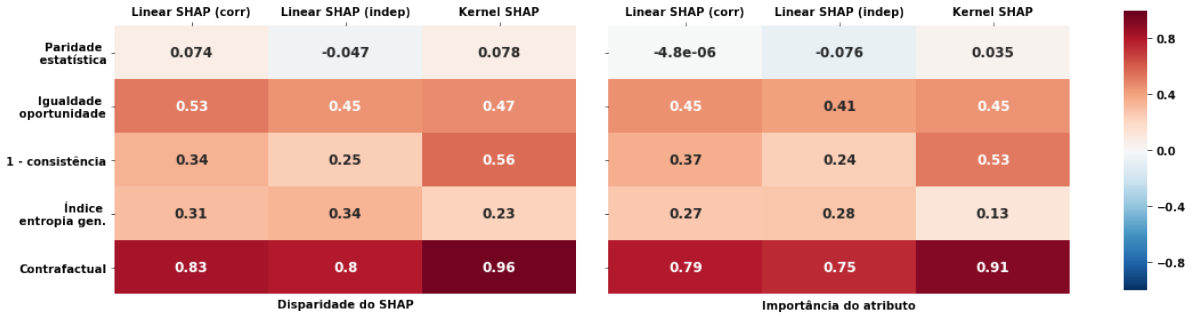


Figura 26: Correlação da variação de resultados dos modelos de Regressão Logística com e sem mitigação de viés. Avaliação separado pelas técnicas do Kernel SHAP e duas variações do Linear SHAP, uma que considera correlação entre os atributos (*corr*) e outra que considera independência entre os atributos (*indep*). No gráfico da esquerda foi calculada a correlação entre a variação das medidas de justiça e da disparidade do SHAP. No gráfico da esquerda foi calculada a correlação entre a variação das medidas de justiça e da importância do atributo.

Outra análise realizada foi a comparação entre o cenário de justiça obtido com cada medida do Kernel e do Linear SHAP do. Nesta comparação foi visto que em sete modelos o cenário de justiça foi diferente entre as metodologias do SHAP, serão selecionados três destes modelos para análise mais detalhada: o modelo sem mitigação de viés treinado com o conjunto de dados do COMPAS, o modelo com reamostragem treinado com o conjunto de dados do Adult e o modelo com subamostragem treinado com o conjunto de dados do PNAD. Os gráficos de summary plot destes casos são mostrados na Figuras 27 e 29, e os de PDP na Figuras 28, 30 e 31.

Os gráficos de summary plot e PDP nas Figuras 27 e 28 respectivamente, são de uma Regressão Logística treinada com o conjunto de dados do COMPAS. O gráfico de PDP foi gerado para o atributo sensível raça. Nestas figuras, a esquerda é mostrado o resultado do Kernel SHAP e a direita do Linear SHAP (considerando correlação entre os atributos).

Os gráficos de summary plot e PDP mostram diferenças significativas de resultado pelas metodologias do SHAP. Nos gráficos de summary plot da Figura 27, o Kernel SHAP considerou raça como o 8º atributo mais importante no modelo, enquanto que o Linear SHAP considerou raça como o 12º (atributo menos importante do modelo). Nos gráficos de PDP da Figura 28 pelo resultado do Kernel SHAP o grupo privilegiado de raça (valor 1) tem valores de SHAP negativo, por isso são responsáveis por diminuir a previsão, e o

grupo desprivilegiado (valor 0) tem valores de SHAP positivo, e por isso são responsáveis por aumentar a previsão. Já no gráfico de PDP do Linear SHAP foi obtido o contrário dessa relação, o grupo privilegiado tem valores de SHAP positivo, e o desprivilegiado tem valores de SHAP negativo.

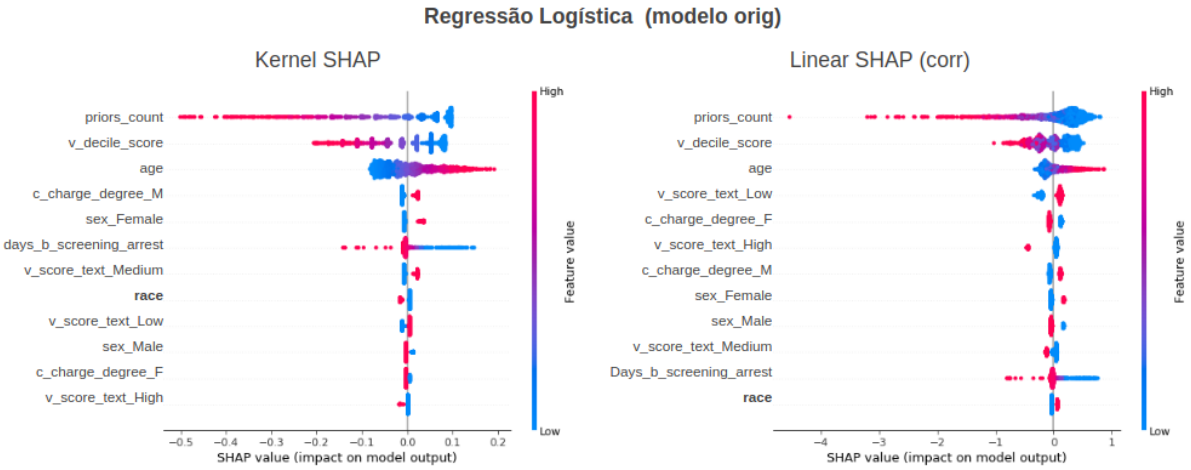


Figura 27: Gráficos de summary plot obtidos com uma Regressão Logística treinada com o conjunto de dados do COMPAS. A esquerda foi gerado o resultado com o Kernel SHAP, e a direita com o Linear SHAP (considerando correlação entre os atributos).

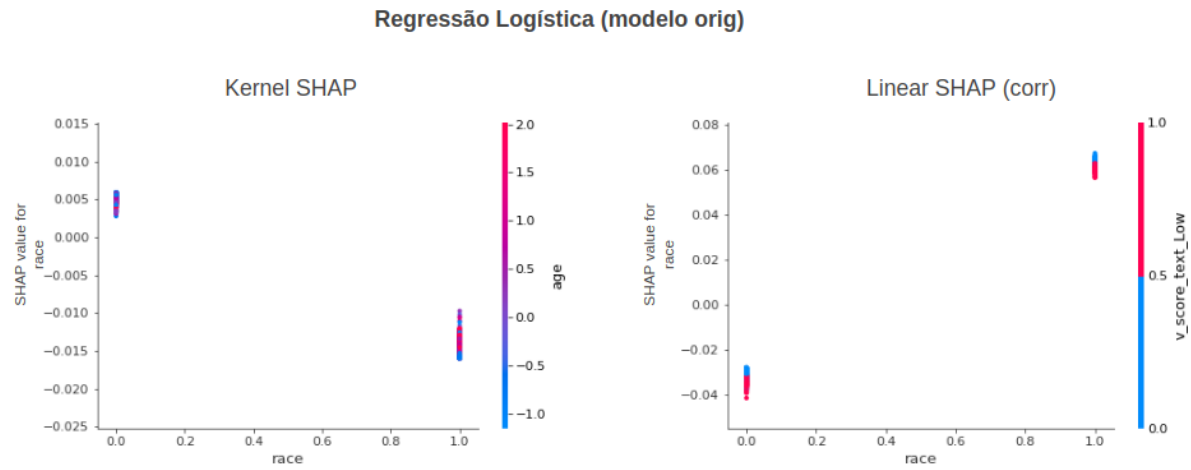


Figura 28: Gráficos de PDP obtidos com uma Regressão Logística que foi treinada com o conjunto de dados do COMPAS. O gráfico de PDP da esquerda foi gerado com o Kernel SHAP, e o gráfico da direita com o Linear SHAP (considerando correlação entre os atributos).

As Figuras 29 e 30 mostram os resultados do Kernel SHAP e do Linear SHAP (considerando correlação entre os atributos) para o modelo de Regressão Logística com reamostragem treinado com conjunto de dados do Adult. Nestes gráficos existem diferenças significativas entre as metodologias do SHAP.

Nos gráficos de summary plot do Kernel SHAP na Figura 29 o atributo sensível gênero foi o 10º mais importante, enquanto que pela metodologia do Linear SHAP foi o 28º mais importante (não é mostrado no gráfico de summary plot pois ele mostra apenas os 20 atributos mais importantes do modelo). O gráfico de PDP mostra que o Kernel SHAP considerou que o grupo desprivilegiado do atributo gênero têm impacto de fazer a previsão aumentar e o desprivilegiado de fazer a previsão diminuir. Entretanto o gráfico de PDP do Linear SHAP mostra que os dois grupos do atributo gênero têm impactos semelhantes no modelo.

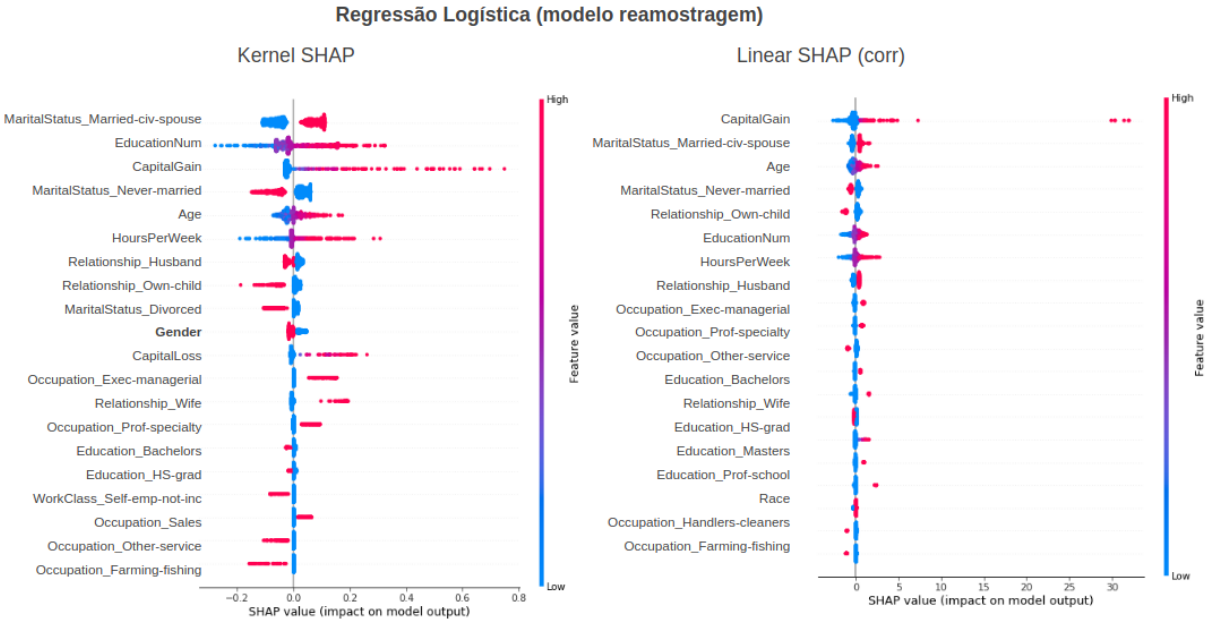


Figura 29: Gráficos de summary plot gerados para um modelo de Regressão Logística com aplicação da técnica de reamostragem para mitigação de viés. Este modelo foi treinado com o conjunto de dados do Adult. O gráfico da esquerda foi obtido com o Kernel SHAP, e o gráfico da direita com o Linear SHAP (considerando correlação entre os atributos).

Outra diferença interessante presente nos gráficos de summary plot da Figura 29 são os valores de SHAP extremos do atributo *CapitalGain* obtidos com o Linear SHAP, que não foram obtidos com o Kernel SHAP. Essa diferença de valores pode ser explicada porque o Kernel SHAP calcula os valores de SHAP aproximando sua soma pela probabilidade prevista pelo modelo, enquanto que o Linear SHAP aproximando pelo log odds previsto.

A probabilidade prevista é calculada pela Equação (5.1) e o log odds pela Equação (5.2). Por estas fórmulas verifica-se que a probabilidade é limitada entre 0 e 1, enquanto que o log odds não é limitado; por isso apenas com o log odds é possível obter valores extremos de previsão, que consequentemente gerariam valores extremos do SHAP. A Tabela 10 mostra os dez maiores valores em log odds previsto pelo modelo. Nela percebe-se que os seis maiores valores previstos foram muito maiores do que os outro valores da tabela,

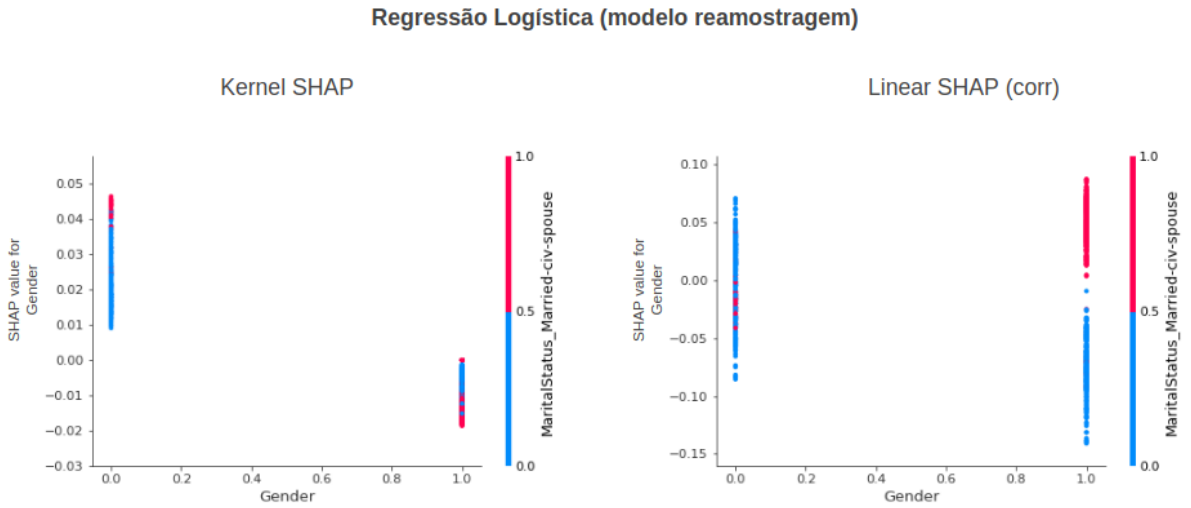


Figura 30: Gráficos de PDP gerados a esquerda com o Kernel SHAP e a direita com o Linear SHAP (considerando correlação entre os atributos). Os resultados destes gráficos foram gerados para o atributo gênero do conjunto de dados do Adult de um modelo de Regressão Logística com aplicação do pré processamento de reamostragem.

e foram estes casos que causaram os valores extremos com o Linear SHAP. Enquanto que para estes mesmos casos a probabilidade prevista ficou limitada em 1.

$$P = \frac{1}{1 + e^{-(b_0 + b_1 x)}}. \tag{5.1}$$

$$\log\left(\frac{P}{1-P}\right) = \log\left(e^{b_0 + b_1 x}\right) = b_0 + b_1 x. \tag{5.2}$$

A Figura 31 mostra gráficos de PDP calculados para o atributo gênero do conjunto de dados do PNAD de uma Regressão Logística com subamostragem. Os gráficos do Kernel SHAP (esquerda) e do Linear SHAP (direita) mostram resultados distintos. Pelo gráfico do Kernel SHAP verifica-se que os grupos privilegiados e desprivilegiados do atributo raça têm importância semelhante no modelo, enquanto que pelo resultado do Linear SHAP verifica-se que o grupo desprivilegiado têm impacto de aumentar a previsão e o desprivilegiado de diminuir a previsão.

log_odds	prob	shap_value_CapitalGain	CapitalGain
33.263293	1.000000	30.434430	13.534688
33.102805	1.000000	31.443525	13.534688
32.864660	1.000000	29.979339	13.534688
32.259133	1.000000	30.420780	13.534688
31.419960	1.000000	32.031086	13.534688
31.234420	1.000000	31.978840	13.534688
7.541124	0.999469	7.325155	3.661227
6.337542	0.998234	2.874443	1.909556
5.914607	0.997308	2.900332	1.909556
5.841082	0.997103	2.853630	1.909556

Tabela 10: Tabela com os dez maiores valores de log odds previstos por um modelo de Regressão Logística treinada com o conjunto de dados do Adult que foi aplicada técnica de reamostragem. Os seguintes valores são mostrados: log odds previsto pelo modelo (*log_odds*), probabilidade prevista pelo modelo (*prob*), valores de SHAP do atributo CapitalGain (*shap_value_CapitalGain*) e valores do atributo CapitalGain (*CapitalGain*).

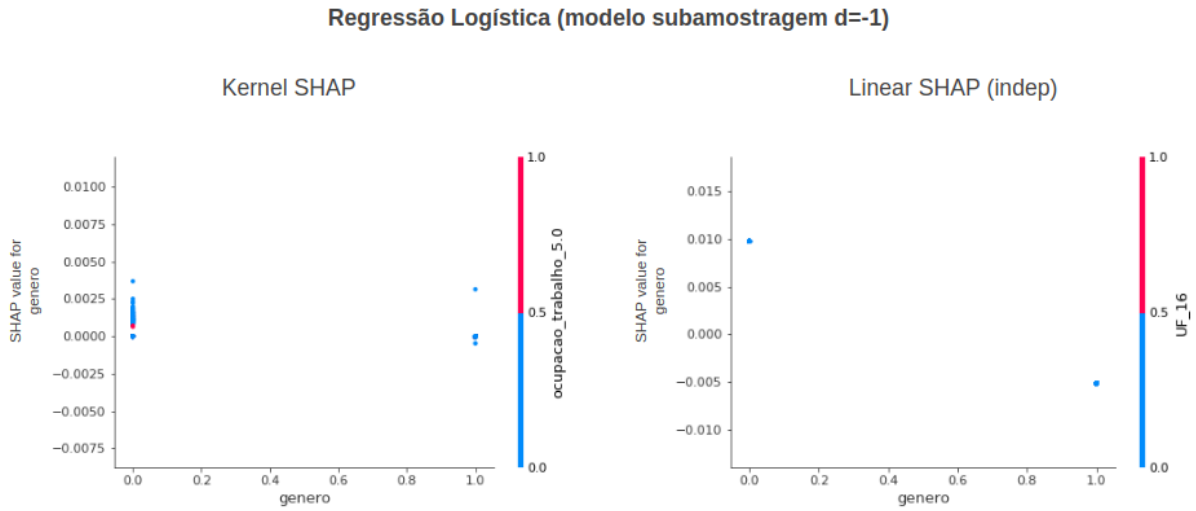


Figura 31: Gráficos de PDP do atributo gênero do conjunto de dados do PNAD de um modelo de Regressão Logística com aplicação da técnica de subamostragem com $d = -1$ para mitigação de viés. O gráfico da esquerda foi gerado com a metodologia do Kernel SHAP, e o gráfico da direita foi gerado com o Linear SHAP.

5.5 Discussão dos resultados

Na avaliação de resultados foi comparado o cenário de justiça obtido com as medidas de justiça e com as medidas do SHAP. Além disso, foi calculado a variação de valores das medidas nos modelos com e sem mitigação de viés para verificar se alguma medida de justiça variou de forma semelhante a medida do SHAP. Nestas duas avaliações foi visto que a medida contrafactual é muito consistente com as medidas de disparidade do SHAP e importância do atributo. Enquanto que a medida de paridade estatística apresenta menor consistência, isso ficou mais evidente na comparação da variação de valores das medidas.

Outro tipo de avaliação realizada foi para comparar a consistência nos resultados quando utilizado diferentes metodologias do SHAP com o mesmo modelo. Nessa avaliação foi comparado o Kernel e o Tree SHAP, e o Kernel e o Linear SHAP. Na comparação entre o Kernel e Tree SHAP foi visto que de forma geral ambas as metodologias apresentam resultados muito semelhante. Enquanto que ao comparar o Kernel SHAP com o Linear SHAP foi verificado maior divergência de resultados.

6 CONCLUSÃO

Em aplicações de aprendizagem de máquina nas quais existem preocupação com discriminação social, os modelos costumam ser avaliados com medida de justiça e com técnicas de interpretabilidade. Essas duas metodologias são complementares, pois a medida de justiça fornece um número que permite comparar o quanto um modelo é mais justo que o outro, enquanto que a técnica de interpretabilidade serve para dar mais confiança no modelo, permitindo entender mais profundamente o seu comportamento e a influência das seus atributos.

Além disso, a interpretabilidade permite explicar resultados específicos gerados pelo modelo, o que é importante para o caso de alguém que se sentiu afetado pela decisão do modelo. Como essas metodologias são usadas em conjunto, o esperado seria que elas mostrassem a mesma relação, só que avaliadas de ponto de vista diferentes. Porém, parece não existir nenhum estudo que compara se de fato os resultados dessas duas metodologias refletem a mesma relação. Por isso, nessa dissertação foi apresentado um framework que permite comparar esses dois resultados.

Foi usado o resultado de diversas definições de justiça para comparar com o resultado do SHAP como metodologia de interpretabilidade, que quantifica a importância do atributo por amostra. A ideia básica é examinar como as definições de justiça e o resultado do SHAP variam ao alterar o efeito da variável sensível no modelo. Com isso pretendia-se identificar se existe alguma medida de justiça que seja mais consistente com o resultado do SHAP. Apesar de o foco ter sido no uso do SHAP, esse mesmo framework poderia ser usado com outra técnica de interpretabilidade, para identificar a medida de justiça mais relacionada com a técnica.

Os resultados obtidos mostraram que existe um consenso grande entre o resultado do SHAP e da medida contrafactual, sendo consideravelmente superior ao obtido com outras medidas. Por outro lado, a medida de justiça com menor consenso com o SHAP foi a paridade estatística. Por isso, em uma aplicação que fosse usada a medida de justiça contrafactual, seria preferível o uso do SHAP como técnica de interpretabilidade. Enquanto

que em uma aplicação que fosse usada a medida de justiça de paridade estatística não seria recomendável o uso do SHAP.

Existem diversas variações do SHAP. Apenas o Kernel SHAP é agnóstica ao modelo, e as outras são específicas para certos modelos, criadas para otimizar o cálculo. Usamos o Kernel SHAP em todos os modelos, e para os modelos que existiam alguma variação nós também a usamos. Com isso alguns modelos foram avaliados com mais de um tipo de técnica do SHAP, nesses casos nós comparamos se os resultados foram equivalentes.

Vimos que o resultado do Kernel e Tree SHAP foram muito semelhantes, e o resultado do Kernel e Linear SHAP apresentaram algumas diferenças. Apesar de algumas diferenças, com todas as técnica do SHAP obtivemos a mesma conclusão da medida de justiça com maior e menor consenso.

Nessa pesquisa usamos a variável sensível diretamente no modelo, pois com o SHAP conseguimos interpretar apenas os resultados das variáveis usadas no modelo. Porém não é sempre recomendável, e em alguns casos não é permitido, usar a variável sensível para treinar o modelo. Fizemos isso em um primeiro momento para validar o framework. Por isso, uma extensão dessa pesquisa é deixar de usar a variável sensível no modelo e calcular as medidas do SHAP em variáveis proxy da variável sensível que foram usadas no modelo.

Outra extensão importante seria remover algumas restrições. Nessa pesquisa foi usado apenas conjunto de dados com uma única variável sensível binária para problemas de classificação com uma classe. Estas restrições foram necessária devido as medidas de justiça e técnicas de mitigação de viés que foram usadas. Por isso um trabalho futuro interesse seria aplicar esse mesmo framework em casos que não foram testados.

REFERÊNCIAS

[1] Aas, K., Jullum, M., Løland, A.: Explaining Individual Predictions when Features are Dependent: More Accurate Approximations to Shapley Values. arXiv Preprint arXiv:1903.10464 (2019)

[2] Adebayo, J., Kagal, L.: Iterative Orthogonal Feature Projection for Diagnosing Bias in Black-Box models. Workshop on Fairness, Accountability, and Transparency in Machine Learning (2016)

[3] Alvarez-Melis, D., Jaakkola, T.S.: On the Robustness of Interpretability Methods. In: ICML Workshop on Human Interpretability in Machine Learning. pp. 66–71 (2018)

[4] Angwin, J., Larson, J., Mattu, S., Kirchner, L.: Machine Bias (2016 - acessado 13/5/2019), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[5] Angwin, J., Larson, J., Mattu, S., Kirchner, L.: How We Analyzed the COMPAS Recidivism Algorithm (2016 - acessado 18/5/2019), https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

[6] Arya, V., Bellamy, R.K.E., Chen, P.Y., Dhurandhar, A., Hind, M., Hoffman, S.C., Houde, S., Liao, Q.V., Luss, R., Mojsilović, A., Mourad, S., Pedemonte, P., Raghavendra, R., Richards, J., Sattigeri, P., Shanmugam, K., Singh, M., Varshney, K.R., Wei, D., Zhang, Y.: One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques. arXiv Preprint arXiv:1909.03012 (2019)

[7] Augasta, M.G., Kathirvalavakumar, T.: Reverse Engineering the Neural Networks for Rule Extraction in Classification Problems. Neural Processing Letters pp. 131–150 (2011)

[8] Auret, L., Aldrich, C.: Empirical Comparison of Tree Ensemble Variable Importance Measures. Chemometrics and Intelligent Laboratory Systems pp. 157 – 170 (2011)

[9] Bastani, O., Kim, C., Bastani, H.: Interpreting Blackbox Models via Model Extraction. arXiv Preprint arXiv:1705.08504 (2017)

[10] Bellamy, R.K.E., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K.N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K.R., Zhang, Y.: AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. IBM Journal of Research and Development pp. 1–15 (2019)

[11] Berk, R., Heidari, H., Jabbari, S., Joseph, M., Kearns, M.J., Morgenstern, J., Neel, S., Roth, A.: A Convex Framework for Fair Regression. Workshop on Fairness, Accountability, and Transparency in Machine Learning (2017)

[12] Binns, R.: Fairness in Machine Learning: Lessons from Political Philosophy. In: Proceedings of Machine Learning Research. pp. 1–11 (2018)

[13] Botari, T., Izbicki, R., de Carvalho, A.: Local Interpretation Methods to Machine Learning Using the Domain of the Feature Space. In: Machine Learning and Knowledge Discovery in Databases. pp. 241–252 (2020)

[14] Brasil: Lei Geral de Proteção de Dados Pessoais (2018 - acessado 19/04/2020), http://www.planalto.gov.br/ccivil_03/_Ato2015-2018/2018/Lei/L13709.htm

[15] Bravo-Marmolejo, S.P., Moreno, J., Gomez, J.C., Pérez-Martínez, C., Ibarra-Manzano, M.A., Almanza-Ojeda, D.L.: Identification of Age and Gender in Pinterest by Combining Textual and Deep Visual Features. In: Information and Software Technologies. pp. 321–332 (2019)

[16] Breiman, L., Friedman, J., Stone, C.J., Olshen, R.: Classification and Regression Trees. Taylor Francis, (1984)

[17] Calders, T., Kamiran, F., Pechenizkiy, M.: Building Classifiers with Independency Constraints. In: International Conference on Data Mining Workshops. pp. 13–18 (2009)

[18] Cesaro, J., Cozman, F.G.: Measuring Unfairness Through Game-Theoretic Interpretability. In: Machine Learning and Knowledge Discovery in Databases. pp. 253–264 (2020)

[19] Chen, T., Guestrin, C.: XGBoost: A Scalable Tree Boosting System. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 785–794 (2016)

[20] Coeckelbergh, M.: AI Ethics. The MIT Press (2020)

[21] Craven, M.W., Shavlik, J.W.: Extracting Tree-Structured Representations of Trained Networks. In: Advances in Neural Information Processing Systems. pp. 24–30 (1995)

[22] Dastin, J.: Amazon Scraps secret AI Recruiting Tool that Showed Bias Against Women (2018 - acessado 5/5/2019), https://reut.rs/2Od9fPr

[23] Datta, A., Sen, S., Zick, Y.: Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems. In: Symposium on Security and Privacy. pp. 598–617 (2016)

[24] Dua, D., Taniskidou, E.K.: UCI Machine Learning Repository (2018 - acessado 14/05/2019), https://archive.ics.uci.edu/ml

[25] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.S.: Fairness Through Awareness. In: Innovations in Theoretical Computer Science Conference. pp. 214–226 (2011)

[26] Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C., Venkatasubramanian, S.: Certifying and Removing Disparate Impact. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 259–268 (2015)

[27] Fisher, A.J., Rudin, C., Dominici, F.: All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. J. Mach. Learn. Res. pp. 1–81 (2019)

[28] Galhotra, S., Brun, Y., Meliou, A.: Fairness Testing: Testing Software for Discrimination. In: Joint Meeting on Foundations of Software Engineering. pp. 498–510 (2017)

[29] Guidotti, R., Monreale, A., Turini, F., Pedreschi, D., Giannotti, F.: A Survey Of Methods For Explaining Black Box Models. ACM Comput. Surv. pp. 1–42 (2018)

[30] Hardt, M., Price, E., Srebro, N.: Equality of Opportunity in Supervised Learning. In: International Conference on Neural Information Processing Systems. pp. 3323–3331 (2016)

[31] Hutchinson, B., Mitchell, M.: 50 Years of Test (Un)fairness: Lessons for Machine Learning. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. pp. 49–58 (2018)

[32] Janzing, D., Minorics, L., Bloebaum, P.: Feature Relevance Quantification in Explainable AI: A Causal Problem. In: International Conference on Artificial Intelligence and Statistics. pp. 2907–2916 (2020)

[33] Kamiran, F., Calders, T.: Classifying without discriminating. In: International Conference on Computer, Control and Communication. pp. 1–6 (2009)

[34] Kamiran, F., Calders, T.: Data Preprocessing Techniques for Classification Without Discrimination. Knowledge and Information Systems pp. 1–33 (2012)

[35] Kamishima, T., Akaho, S., Asoh, H., Sakuma, J.: Fairness-Aware Classifier with Prejudice Remover Regularizer. In: Learning and Knowledge Discovery in Databases. pp. 35–50 (2012)

[36] Kearns, M., Roth, A.: The Ethical Algorithm: The Science of Socially Aware Algorithm Design. OUP USA (2019)

[37] Kilbertus, N., Rojas Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., Schölkopf, B.: Avoiding Discrimination through Causal Reasoning. In: Advances in Neural Information Processing Systems, pp. 656–666 (2017)

[38] Kusner, M.J., Loftus, J., Russell, C., Silva, R.: Counterfactual Fairness. In: Advances in Neural Information Processing Systems, pp. 4066–4076 (2017)

[39] Laugel, T., Renard, X., Lesot, M., Marsala, C., Detyniecki, M.: Defining Locality for Surrogates in Post-hoc Interpretablity. In: ICML Workshop on Human Interpretability in Machine Learning. pp. 47–53 (2018)

[40] Lavalle, A.G., Medeiros, B.F., Kirschbaum, C., Marques, E.C.L., Lotta, G.S., Phillips, J., da Silva Arretche, M.T., Bichir, R.M., Peres, U.D., Coelho, V.S.R.P.: Centro de Estudos da Metrópole da USP (2018 - acessado 07/06/2020), http://centrodametropole.fflch.usp.br/pt-br/download-de-dados

[41] Leslie, D.: Understanding Artificial Intelligence Ethics and Safety. The Alan Turing Institute (2019)

[42] Lipovetsky, S., Conklin, M.: Analysis of Regression in Game Theory Approach. Applied Stochastic Models in Business and Industry pp. 319–330 (2001)

[43] Liu, H.W., Lin, C.F., Chen, Y.J.: Beyond State v. Loomis: Artificial Intelligence, Government Algorithmization, and Accountability. International Journal of Law and Information Technology pp. 122–141 (2019)

[44] Louppe, G., Wehenkel, L., Sutera, A., Geurts, P.: Understanding Variable Importances in Forests of Randomized Trees. In: Advances in Neural Information Processing Systems, pp. 431–439 (2013)

[45] Lundberg, S.M.: Question About Missingness (2018 - acessado 23/12/2019), https://github.com/slundberg/shap/issues/175

[46] Lundberg, S.M.: Math Behind Linear Explainer with Correlation Feature Perturbation (2019 - acessado 26/12/2019), https://github.com/slundberg/shap/blob/master/notebooks/linear_explainer/Math%20behind%20LinearExplainer%20with%20correlation%20feature%20perturbation.ipynb

[47] Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.I.: From Local Explanations to Global Understanding with Explainable AI for Trees. Nature Machine Intelligence pp. 2522–5839 (2020)

[48] Lundberg, S.M., Lee, S.I.: A Unified Approach to Interpreting Model Predictions. In: Advances in Neural Information Processing Systems, pp. 4765–4774 (2017)

[49] Madiega, T.: EU Guidelines on Ethics in Artificial Intelligence: Context and Implementation (2019 - acessado 01/05/2020), https://www.europarl.europa.eu/RegData/etudes/BRIE/2019/640163/EPRS_BRI(2019)640163_EN.pdf

[50] Miller, D.: Justice. In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University (2017)

[51] Molnar, C.: Interpretable Machine Learning - A Guide for Making Black Box Models Explainable (2019)

[52] Nicolae, M.I., Sinn, M., Tran, M.N., Buesser, B., Rawat, A., Wistuba, M., Zantedeschi, V., Baracaldo, N., Chen, B., Ludwig, H., Molloy, I., Edwards, B.: Adversarial Robustness Toolbox v1.0.0. arXiv Preprint arXiv:1807.01069v4 (2019)

[53] Partnership on IA: About us (2016 - acessado 01/05/2020), https://www.partnershiponai.org/about/

[54] Pedreshi, D., Ruggieri, S., Turini, F.: Discrimination-aware Data Mining. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 560–568 (2008)

[55] Rawls, J.: A Theory of Justice. The Belknap Press (1971)

[56] Ribeiro, M.T., Singh, S., Guestrin, C.: "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1135–1144 (2016)

[57] Riols, F.: NeurIPS 2019 - A Recap (2020 - acessado 01/05/2020), https://www.elementai.com/news/2020/neurips-2019-a-recap

[58] Saabas, A.: Interpreting Random Forests (2014 - acessado 5/5/2019), http://blog.datadive.net/interpreting-random-forests

[59] Shapley, L.S.: A Value for N-Person Games. In: Contributions to the Theory of Games, pp. 307–317 (1953)

[60] Speicher, T., Heidari, H., Grgic-Hlaca, N., Gummadi, K.P., Singla, A., Weller, A., Zafar, M.B.: A Unified Approach to Quantifying Algorithmic Unfairness: Measuring Individual Group Unfairness via Inequality Indices. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 2239–2248 (2018)

[61] Tan, S., Caruana, R., Hooker, G., Lou, Y.: Distill-and-Compare: Auditing Black-Box Models Using Transparent Model Distillation. In: AAAI/ACM Conference on AI, Ethics, and Society. pp. 303–310 (2018)

[62] Verma, S., Rubin, J.: Fairness Definitions Explained. In: Proceedings of the International Workshop on Software Fairness. pp. 1–7 (2018)

[63] Wadsworth, C., Vera, F., Piech, C.: Achieving Fairness through Adversarial Learning: an Application to Recidivism Prediction. Workshop on Fairness, Accountability, and Transparency in Machine Learning (2018)

[64] Wakefield, J.: Microsoft Chatbot is Taught to Swear on Twitter (2016 - acessado 16/11/2019), https://www.bbc.com/news/technology-35890188

[65] Wexler, J., Pushkarna, M., Bolukbasi, T., Wattenberg, M., Viégas, F.B., Wilson, J.: The What-If Tool: Interactive Probing of Machine Learning Models. IEEE Transactions on Visualization and Computer Graphics pp. 56–65 (2019)

[66] Young, H.P.: Monotonic Solutions of Cooperative Games. International Journal of Game Theory pp. 65–72 (1985)

[67] Zelaya, V.G., Missier, P., Prangle, D.: Parametrised Data Sampling for Fairness Optimisation. KDD XAI Workshop (2019)

[68] Zemel, R., Wu, Y., Swersky, K., Pitassi, T., Dwork, C.: Learning Fair Representations. In: International Conference on Machine Learning. pp. 325–333 (2013)

[69] Zhang, B.H., Lemoine, B., Mitchell, M.: Mitigating Unwanted Biases with Adversarial Learning. In: AAAI/ACM Conference on AI, Ethics, and Society. pp. 335–340 (2018)

[70] Zilke, J.R., Mencía, E.L., Janssen, F.: DeepRED – Rule Extraction from Deep Neural Networks. In: Discovery Science. pp. 457–473 (2016)

[71] Zliobaite, I.: On the Relation between Accuracy and Fairness in Binary Classification. Workshop on Fairness, Accountability, and Transparency in Machine Learning (2015)

[72] Štrumbelj, E., Kononenko, I.: Explaining Prediction Models and Individual Predictions with Feature Contributions. Knowledge and Information Systems pp. 647–665 (2013)

APÊNDICE A – GRÁFICOS

As Figuras 32 e 33 mostram gráficos com os resultados obtidos para as medidas de justiça e do SHAP. Nestes gráficos os resultados foram separados por conjunto de dados. Na Figura 32 estão presentes os resultados para os conjuntos de dados: Adult, Compas e German. Na Figura 33 para os conjuntos de dados: Default e PNAD. No caso do PNAD alguns modelos foram treinados usando apenas a variável sensível gênero, e outros com apenas a variável sensível raça.

Nos gráficos o eixo x indica a metodologia do SHAP e a abreviação do modelo, sendo que *gb* se refere a Gradient Boosting, *lr* a Regressão Logística, rf a Random Forest e *svm* a Support vector machine (SVM). A legenda indica o tipo de modelo utilizado, sendo que a linha vermelha (*orig*) mostra o resultado do modelo sem mitigação de viés, a linha azul com quadrado (*rw*) mostra o resultado do modelo com aplicação da reponderação, a linha cinza com triângulo (*usd-1*) mostra o resultado do modelo com aplicação da subamostragem com $d = -1$, e a linha azul com círculo (*usd-0*) mostra o resultado do modelo com aplicação da subamostragem com $d = 0$.

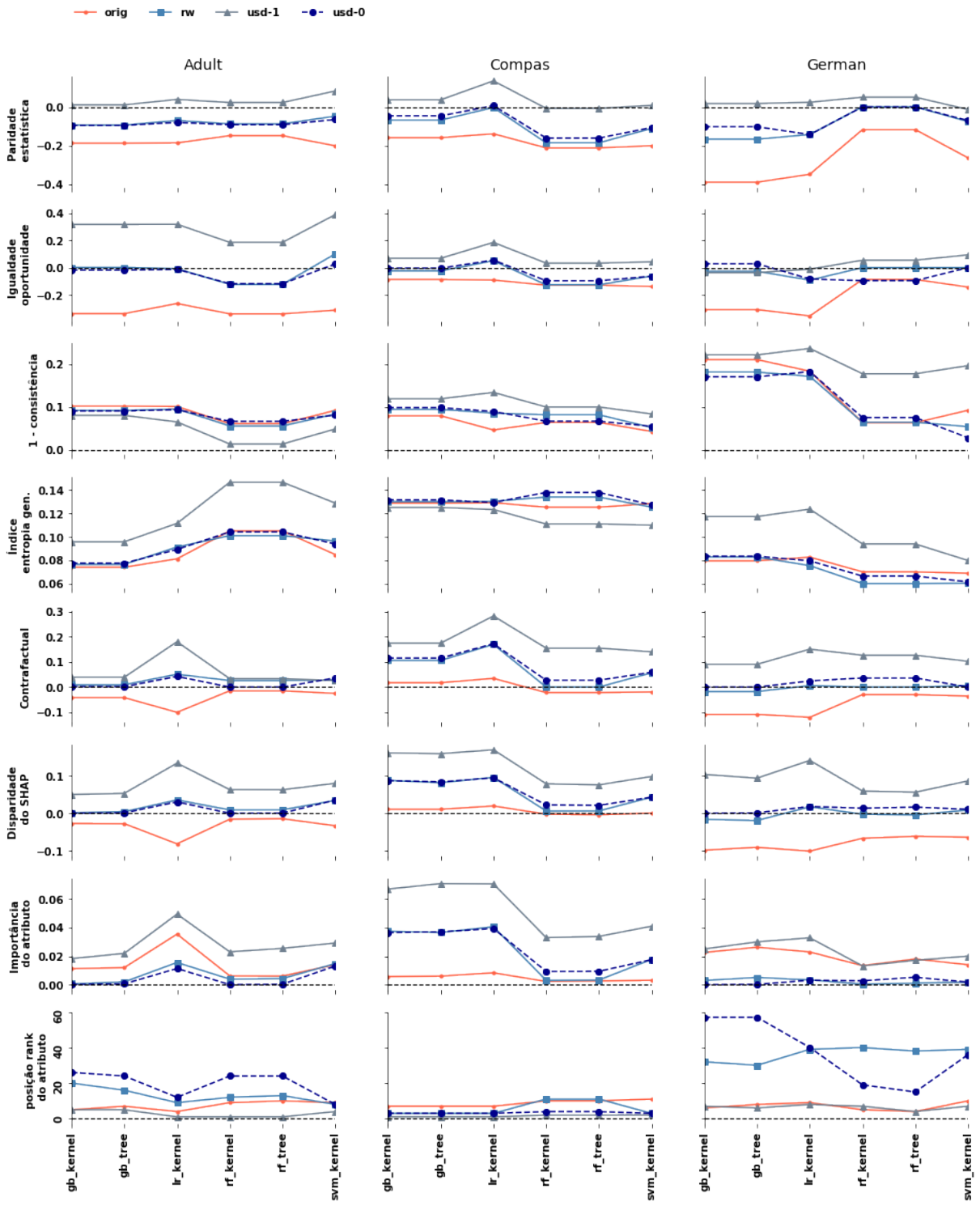


Figura 32: Avaliação de medidas de justiça e medidas obtidas com SHAP para os conjuntos de dados: Adult, COMPAS e German, com quatro tipos de modelos: Regressão Logística, Random Forest, Gradient Boosting e SVM.

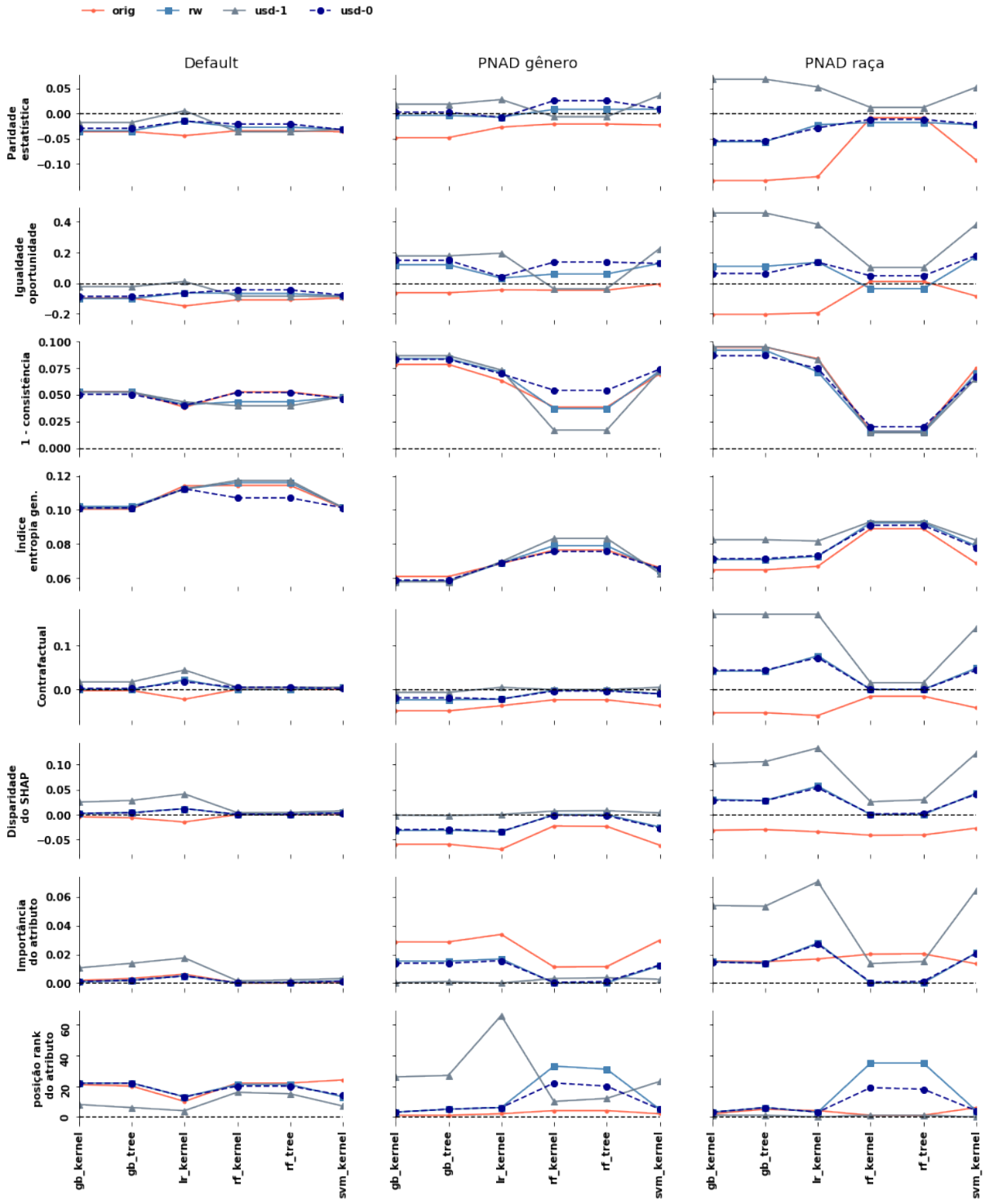


Figura 33: Avaliação de medidas de justiça e medidas obtidas com SHAP para os conjuntos de dados: Default e PNAD (dois tipos de modelos, um com variável sensível gênero e outro com raça), com quatro tipos de modelos: Regressão Logística, Random Forest, Gradient Boosting e SVM.