

FUNDAÇÃO GETULIO VARGAS
ESCOLA DE MATEMÁTICA APLICADA
CURSO DE GRADUAÇÃO EM MATEMÁTICA APLICADA

VITÓRIA AQUINO GUARDIEIRO

**APLICAÇÃO DE OTIMIZAÇÃO MULTI-OBJETIVO EM PROBLEMAS DE
APRENDIZAGEM DE MÁQUINAS PARA DIMINUIÇÃO DE DISCRIMINAÇÃO**

Rio de Janeiro - Brasil
2020

VITÓRIA AQUINO GUARDIEIRO

**APLICAÇÃO DE OTIMIZAÇÃO MULTI-OBJETIVO EM PROBLEMAS DE
APRENDIZAGEM DE MÁQUINAS PARA DIMINUIÇÃO DE DISCRIMINAÇÃO**

Trabalho de conclusão de curso apresentado à Escola de Matemática Aplicada, como parte dos requisitos para a obtenção do título de Bacharel em Matemática Aplicada.

Orientador: Jorge Poco

Coorientador: Marcos Raimundo

Rio de Janeiro - Brasil

2020

VITORIA AQUINO GUARDIEIRO

**APLICAÇÃO DE OTIMIZAÇÃO MULTI-OBJETIVO EM PROBLEMAS DE
APRENDIZAGEM DE MÁQUINAS PARA DIMINUIÇÃO DE DISCRIMINAÇÃO**

Trabalho de Conclusão apresentado à Escola de Matemática Aplicada como requisito para a
obtenção do grau de bacharel em Matemática Aplicada

Aprovado em 07 de Dezembro de 2020



Jorge Poco

Escola de Matemática Aplicada – FGV



Marcos Raimundo

Escola de Matemática Aplicada – FGV



Romis R. de Faissol Attux

Faculdade de Engenharia Elétrica e de Computação – UNICAMP

Resumo

Um dos principais desafios ao criar um simulador do comportamento humano consiste em suprimir a perpetuação de vieses discriminatórios presentes na tomada de decisão humana. Diversas abordagens para lidar com tal problema já foram propostas na literatura de Inteligência Artificial. Tais propostas geralmente consistem na definição de métricas desenvolvidas para quantificar a discriminação decorrente de um modelo de inteligência artificial e na aplicação de tais métricas na geração do modelo de forma a limitar os efeitos discriminatórios gerados. Entretanto, a limitação da discriminação de um modelo faz com que o erro médio dele aumente, gerando um cenário de conflito entre essas métricas. Neste trabalho, propomos uma metodologia que utiliza de otimização multi-objetivo que simultaneamente otimiza o desempenho dos modelos e a qualidade da predição nos grupos sensíveis (que podem sofrer discriminação). Focando em problemas de classificação binária, buscamos encontrar o conjunto dos melhores modelos para tais métricas. Comparamos nossa abordagem com outras metodologias bem conhecidas da área, usando a Regressão Logística como referência de comparação. A metodologia proposta obteve resultados promissores, sendo capaz de diminuir a discriminação dos modelos de inteligência artificial conjuntamente com a manutenção do desempenho em acurácia.

Palavras-chave: Inteligência Artificial. Ética. Aprendizado de Máquina. Otimização Multi-objetivo. Classificação.

Abstract

One of the main challenges when creating a simulator of human behavior is to suppress the perpetuation of discriminatory biases present in human decision making. Several approaches to deal with this problem have already been proposed in Artificial Intelligence literature. Such proposals generally consist of the definition of metrics developed to quantify the discrimination resulting from an artificial intelligence model and the application of such metrics in the generation of the model to limit the discriminatory effects generated. However, limiting the discrimination of a model causes the average error of it to increase, generating a scenario of conflict between these metrics. In this project, we propose a methodology that uses multi-objective optimization that simultaneously optimizes the performance of the models and the quality of the prediction in sensitive groups (which may suffer discrimination). Focusing on binary classification problems, we seek to find the set of the best models for such metrics. We compared our approach with other well-known methodologies in the area, using Logistic Regression as a comparison reference. The proposed methodology obtained promising results, being able to reduce the discrimination of artificial intelligence models together with the maintenance of accurate performance.

Keywords: Artificial intelligence. Ethic. Machine Learning. Multi-objective optimization. Classification.

Sumário

1	Introdução	6
1.1	Motivação	7
1.2	Objetivo	8
1.3	Esboço do trabalho	8
2	Trabalhos relacionados	10
2.1	Contextualização em <i>Fairness</i>	10
2.2	<i>Fairness</i> com otimização	11
2.3	<i>Fairness</i> com otimização multi-objetivo	12
2.4	Considerações finais	13
3	Metodologia	14
3.1	Otimização multi-objetivo	15
3.1.1	MONISE - Many-Objective Non-Inferior Set Estimation	16
3.2	Modelos de aprendizado e o conflito entre objetivos	17
3.2.1	Primeira proposta: Erro por grupo	18
3.2.2	Segunda proposta: Aceitação por grupo	19
3.3	Métricas de discriminação	20
3.4	<i>Ensemble Learning</i>	22
4	Experimentos	24
4.1	Configurações dos experimentos	24
4.1.1	Dados	24
4.1.2	Abordagens comparadas	25
4.2	Experimento de otimização individual	27
4.3	Experimento de diversidade	33
4.4	Experimento de <i>ensemble learning</i>	35
4.5	Conclusão dos experimentos	39
5	Conclusões e trabalhos futuros	40
	Referências	42
	Apêndice	46

1 Introdução

A aplicação de inteligência artificial tem se expandido na automatização na tomada de decisão. Esses modelos têm guiado políticas públicas, análise de crédito, análise de contratação de empregados, entre outras tomadas de decisão que tem um impacto significativo na vida das pessoas e da sociedade como um todo. Por conta desses modelos buscarem simular o comportamento humano apresentado a eles através dos dados, se os próprios dados fornecidos contiverem decisões injustas ou discriminatórias, os sistemas de inteligência artificial também podem propagar tais injustiças. Ainda que os dados não forneçam informações explícitas dos grupos discriminados, ela ainda assim pode se tornar injusta ao capturar padrões implícitos que auxiliem na minimizar o erro médio de dados que contém discriminação na sua rotulação.

Diversos escândalos em relação a inteligências artificiais injustas e discriminatórias já ocorreram. Um caso comum acontece com modelos de reconhecimento facial que possuem erro bastante superior para pessoas de pele escura em detrimento de pessoas de pele clara e também acontece com modelos utilizados no sistema judiciário para prever a chance de um criminoso voltar a cometer algum crime. Neste último caso, modelos de IA predizem probabilidades maiores para pessoas negras em relação a pessoas brancas com demais características similares. Existem diversas aplicações e formas com que esses modelos de inteligências artificiais causam e propagam injustiças e discriminações nas mais diversas áreas [1, 2]. Nesse cenário, se torna necessário olhar esses modelos sob uma nova ótica, buscando não somente modelos que tenham alta assertividade, mas que sejam capazes de reduzir os vieses contidos na sociedade que são retratados nos dados.

Considerando o exemplo do problema de classificação para aprovar ou rejeitar um pedido de empréstimo, como o apresentado no conjunto de dados *Statlog (German Credit Data)* [3]. Há diversas informações sobre um pedido de empréstimo, como seu valor, e informações também sobre a pessoa que está solicitando o empréstimo, dentre elas o gênero da pessoa. Ao ajustarmos um classificador com esses dados, é possível que o classificador leve em conta o gênero da pessoa para atribuir chances diferentes de receber o crédito, mesmo com as demais características similares [23], o que pode ser considerado injusto. Uma opção para resolver esse problema seria remover esse atributo de gênero, também chamado de atributo sensível, mas em muitos casos isso não é suficiente para assegurar que esse modelo seja justo [4]. Podem existir outros atributos que são altamente correlacionados com o atributo sensível (por exemplo, as mulheres terem menos dinheiro guardado em suas contas bancárias do que os homens, por ser comum que as finanças fiquem sob responsabilidade do marido), e o modelo pode usar essa informação e continuar gerando discrepâncias entre os grupos [4]. A aborda-

gem de treinar modelos de inteligência artificial considerando a potencial discriminação que eles podem gerar é justamente o que constitui a área de justiça/*fairness*.

Muitas das discriminações geradas pelos modelos de aprendizado de máquina se dão pelo treinamento deles focarem apenas na minimização do erro gerado, com a discriminação possível não sendo considerada. As abordagens de *fairness* em aprendizado de máquina fazem com que a discriminação passe a ser considerada ao treinar e escolher o melhor modelo. Falaremos um pouco mais sobre elas na Seção 2 - Trabalhos relacionados, mas em geral elas se limitam a tentar remover a discriminação dos conjuntos de dados que serão utilizados para treinamento, tratar a discriminação como restrição no treinamento do modelo ou mudar o modelo pós-treinamento de forma a diminuir a discriminação. Neste trabalho, abordaremos o treinamento dos modelos de aprendizado de máquina enquanto um problema de otimização multi-objetivo, de forma a minimizar não apenas o erro, mas também a discriminação gerada por ele.

1.1 Motivação

Ao treinar um classificador considerando métricas de discriminação, muitas vezes obteremos um preditor com um desempenho global diferente de um modelo que foca somente em acurácia. Isso é devido ao conflito entre o aprendizado de algo acurado e algo justo, como observado nos trabalhos [9], [13] e [14]. Não obstante, também existe um conflito dentre as várias métricas de discriminação e justiça [13]. Com isso, o que desejamos é encontrar um compromisso ótimo entre essas métricas, encontrando todos os modelos (ou um conjunto representativo deles) que possuam as melhores trocas entre as métricas.

Se temos dois modelos, θ^A e θ^B , com os erros e_A e e_B e as métricas de discriminação d_A e d_B , e $e_A < e_B$ e $d_A < d_B$, então é claro que o modelo θ^A é melhor que o modelo θ^B , por ser melhor em todas as métricas. Mas se temos que $e_A < e_B$ e $d_A > d_B$, não é possível arbitrar qual dos dois modelos é o melhor dentre eles. Sendo assim, quando não existem modelos melhores que θ^A e θ^B , nós desejamos encontrar ambos os modelos, e somente decidir qual dos dois é melhor para nossa aplicação após o treinamento. Tal processo de treinamento que busca otimizar modelos para diversas métricas pode ser abordado através de otimização multi-objetivo.

Essa pesquisa tem como objetivo estudar o impacto da utilização de otimização multi-objetivo em modelos de aprendizado de máquina, isso nos habilita investigar como o aprendizado nos grupos discriminados afeta o aprendizado como um todo. Ao usar metodologias de otimização multi-objetivo *a posteriori* é possível obter a fronteira de Pareto, que consiste num conjunto de soluções com diferentes compromissos de aprendizado para os vários gru-

pos estudados. Dessa forma, é possível escolher, dentre essas soluções, quais delas são mais apropriadas para o nosso objetivo de reduzir a desigualdade dentre os grupos sociais.

1.2 Objetivo

O objetivo principal deste projeto é obter uma metodologia capaz ajustar modelos de inteligência artificial para classificação binária minimizando não somente o erro de aprendizado, mas também a discriminação. Isso será abordado utilizando da otimização multi-objetivo, que permite a otimização simultânea de diversas funções objetivo. Entretanto, como tais métricas de discriminação são não convexas, o que é demandado pelo otimizador utilizado, exploramos duas propostas diferentes de conjuntos de funções objetivo para otimizarmos, indiretamente, as métricas de discriminação.

Focamos em classificação binária devido a esta aplicação ser a mais explorada dentro da literatura da área, possuindo diversas métricas de justiça dedicadas exclusivamente à esta classificação. Nossa metodologia pode ser aplicada a outros tipos de modelo com algumas adaptações. Tais aplicações podem ser exploradas em trabalhos futuros.

Os objetivos específicos consistem em:

- Modelar o classificador de regressão logística utilizando da otimização multi-objetivo tomando em consideração o conflito de objetivos de desempenho de cada grupo e a complexidade do modelo.
- Criação de um arcabouço que seja capaz de ajustar diversos classificadores de forma a reduzir a discriminação sem perdas muito grandes de desempenho.
- Comparar os resultados obtidos pelas propostas deste trabalho com estratégias similares da literatura de *fairness*.
- Utilizar de métodos de *ensemble learning* para, a partir de um conjunto de modelos gerado pela abordagem multi-objetiva, permitir ao tomador de decisão que combine modelos em um ensemble para atender múltiplas métricas.

1.3 Esboço do trabalho

Na **Seção 2 - Trabalhos Relacionados** apresentamos uma contextualização das pesquisas realizadas em discriminação de modelos de inteligência artificial. Além disso, como esse trabalho é focado na otimização multi-objetivo dos modelos, vamos expor trabalhos em otimização de modelos, e mais especificamente em otimização multi-objetivo de modelos. Esses trabalhos adaptam o treinamento de tais modelos com o objetivo de reduzir sua

discriminação e injustiça, e nosso objetivo é abordar as principais diferenças entre tais trabalhos e a metodologia proposta.

Na **Seção 3 - Metodologia** definimos do que se trata a otimização multi-objetivo e apresentamos o método que utilizaremos para resolver os problemas de otimização que iremos propor. Em seguida, formalizamos as duas propostas deste projeto, juntamente com a motivação de cada uma. Ao fim, apresentamos também diversas métricas de discriminação que utilizaremos para avaliar os modelos na seção seguinte.

Na **Seção 4 - Resultados** apresentamos diversos experimentos que realizamos com as metodologias propostas, comparando com outras metodologias utilizadas na literatura de *fairness*. Desta forma, avaliamos como as propostas se comportam em conjuntos de dados bem conhecidos e testamos a validade da nossa proposta.

Na **Seção 5 - Conclusão** apresentamos as considerações finais do trabalho, assim como as perspectivas de pesquisa futura para o tema abordado.

2 Trabalhos relacionados

As pesquisas que buscam estudar a discriminação e injustiça gerada por modelos de inteligência artificial abordam perspectivas diferentes, desde definir o que seria uma discriminação e apresentar métricas que permitem quantificar o quão discriminatório é um modelo, até ao desenvolvimento de estratégias para diminuí-la. Este trabalho aborda duas propostas de estratégias para diminuir a discriminação de modelos de aprendizagem de máquinas utilizando de otimização multi-objetivo.

Na Seção 2.1 apresentamos uma contextualização das estratégias mais comuns na literatura de discriminação de modelos de aprendizagem de máquina. Em seguida, na Seção 2.2, exploramos mais a fundo projetos que realizam modificações no problema de otimização correspondente ao treinamento do modelo. Por fim, em 2.3, apresentamos trabalhos que, assim como este, utilizam de otimização multi-objetivo, ressaltando suas principais diferenças com as estratégias propostas aqui.

2.1 Contextualização em *Fairness*

Inicialmente acreditava-se que desconsiderar os atributos sensíveis, como raça e gênero, ao treinar modelos de aprendizagem de máquina era suficiente para evitar discriminações, o que ficou conhecido como “justiça por desconhecimento” (“*fairness through unawareness*”) [31]. Entretanto, tal abordagem se mostrou inadequada, devido aos atributos sensíveis terem correlação considerável com outros atributos que gostaríamos de utilizar, permitindo, assim, que ocorra discriminação através de tais atributos correlacionados. Por exemplo, código postal é bastante correlacionado com raça [35].

Com isso, diversas estratégias estão sendo desenvolvidas para lidar com a discriminação gerada por inteligências artificiais, levando em consideração as características sensíveis. Em geral, elas realizam alterações nas máquinas de aprendizado já bastante utilizadas na área, como regressões logísticas. Podemos dividi-las em três famílias [32], sendo elas as que realizam mudanças durante o pré-processamento, no próprio processamento e no pós-processamento das modelos de inteligência artificial gerados.

Pré-processamento As estratégias de pré-processamento buscam compensar os vieses das informações contidas na base de dados antes da criação do modelo de inteligência artificial. Isso pode ser atingido através de uma nova representação de cada amostra por uma combinação de protótipos que reduzem o viés [6], ou criando ponderações para cada amostra de forma que o modelo seja forçado a aprender melhor os grupos discriminados [8].

Durante o processamento As estratégias que são aplicadas durante o processamento, ou treinamento, dos modelos reescrevem o problema de otimização que corresponde ao treinamento de forma a considerar explicitamente as métricas de discriminação, podendo, por exemplo, adicionar restrições que forçam o modelo a atender uma métrica de discriminação [9], ou realizar uma sequência de classificações que adicionam uma ponderação para reduzir a discriminação [10].

Pós-processamento As estratégias de pós-processamento mudam componentes do modelo de inteligência artificial após o aprendizado do mesmo. Isso pode ser feito mudando o threshold de classificação [11]; ou usando aprendizado adversarial [12].

Cada uma dessas estratégias partem de diferentes concepções e necessidades durante o desenho de uma máquina de aprendizado. Entretanto, esse trabalho apresenta propostas de adaptações a serem feitas nos modelos durante o processamento deles, através de modificações feitas ao problema de otimização correspondente ao treinamento.

2.2 *Fairness* com otimização

Diversos trabalhos na literatura de *fairness* abordam a discriminação dos modelos de aprendizado de máquinas através da modificação do treinamento deles. De forma geral, um modelo de aprendizado de máquinas é gerado (ou treinado) de forma a minimizar o erro (ou maximizar o desempenho) do modelo para um conjunto de dados fornecido. Essa minimização/maximização é o objeto de estudo da Otimização. Assim, alguns dos trabalhos de *fairness* que abordam mudanças no processo de treinamento de forma a reduzir o viés discriminatório.

Tais trabalhos levam em consideração que o erro de aprendizado (ou desempenho) e discriminação dos modelos são implicitamente conflitantes, isto é, ao maximizarmos o desempenho, teremos uma discriminação maior do que se tivéssemos minimizado a discriminação. Assim, otimizar uma métrica piora a outra. Tal conflito foi abordado nos trabalhos [9], [13] e [14]. Além disso, também existe um conflito dentre as várias métricas de discriminação [13].

Vários dos modelos que utilizam da otimização o fazem otimizando o desempenho do modelo, mas com restrições em relações à discriminação resultante. O trabalho [9], por exemplo, otimiza o desempenho limitando a diferença entre os erros de aprendizado entre os grupos sensíveis e não sensíveis. Já o trabalho [19] limita a diferença entre as chances de obter o recurso desejado entre os grupos.

Entretanto, essas otimizações com restrições não garantem que os modelos treinados terão os melhores valores possíveis para as métricas de desempenho e de discriminação, prin-

principalmente quando são utilizadas mais que duas métricas ao avaliar o modelo. Assim, surgiram diversas abordagens que utilizam métodos de otimização que são focados justamente em otimizar diversas métricas ao mesmo tempo, pertencentes à subárea chamada de otimização multi-objetivo.

2.3 *Fairness* com otimização multi-objetivo

Levando em consideração o conflito existente entre as métricas de desempenho e métricas de discriminação, e também considerando que podemos querer utilizar mais de uma métrica de discriminação ou de desempenho, tentar otimizar tantos objetivos ao mesmo tempo pode gerar alguns problemas. Um deles é o tempo de execução do treinamento dos modelos. Outro, ainda mais impactante que o primeiro, é que podemos acabar com modelos com valores sub-ótimos para as métricas, dependendo da metodologia utilizada para abordar a otimização delas. Tendo isso em vista, alguns trabalhos utilizam da metodologia de otimização multi-objetivo para realizar o treinamento dos modelos, o remodelando de forma a otimizar simultaneamente diversos objetivos, sendo eles o desempenho e a discriminação.

Devido à utilização de mais de um objetivo a ser otimizado, as metodologias de otimização multi-objetivo encontram mais do que um único resultado ótimo, encontrando, na verdade, um conjunto de resultados ótimos chamado de Fronteira de Pareto. Já as estratégias de *fairness* podem retornar tanto essa Fronteira e permitir que o utilizador do modelo escolha qual dos modelos ótimos ele prefere ou então pré-estabelecer uma forma automática de escolher qual o melhor modelo dentre esse conjunto. A Fronteira de Pareto é capaz de gerar diversidade de modelos [24] que podem ser muito importante tanto para gerar um único modelo capaz de ser mais justo quanto para agregar vários modelos se utilizando de metodologias de *ensemble learning* [24].

Um dos trabalhos que utiliza da otimização multi-objetivo considera que o desempenho (ou erro) para cada um dos grupos que podem sofrer discriminação (ou serem beneficiados) são explicitamente conflitantes entre si [20]. Entretanto, essa proposta se limita a encontrar apenas um modelo de inteligência artificial, sendo ele o modelo que minimiza o maior erro dentre os grupos.

O trabalho [34] também utiliza da otimização multi-objetivo para otimizar simultaneamente a performance e discriminação do modelo. Ele aborda isso através da otimização dos hiper parâmetros do modelo, criando uma abordagem que é agnóstica em relação ao modelo e à métrica de *fairness* escolhidos. Assim como este trabalho, ele tem como objetivo encontrar um conjunto de modelos representativo da fronteira de Pareto do compromisso entre performance e discriminação, deixando na mão do utilizador escolher qual o melhor modelo

para a aplicação deseja dado o conjunto ótimo. Todavia, ao otimizar através da seleção dos hiper parâmetros, não modificando a forma do modelo em si, a metodologia está limitada ao modelo de aprendizado que inserido para fazer o ajuste do hiper-parâmetro, dependendo das características do mesmo para reduzir o viés discriminatório.

Com isso, a metodologia proposta não se limita a criar modelos que induzem redução viés [9, 19], nem a tratar o conflito das perdas de aprendizado para cada grupo [20], nem a gerar múltiplos modelos com otimização multi-objetivo [34]. Este trabalho tem todas essas características e é capaz de gerar diversidade de modelos o suficiente para que um usuário (ou uma métrica) selecione um modelo, ou a seleção de múltiplos modelos, permitindo que métodos de *ensemble learning* agreguem essas múltiplas predições.

2.4 Considerações finais

Nessa seção apresentamos as principais categorias de estratégias que abordam o problema da discriminação advinda de modelos de aprendizagem de máquina, focando nas estratégias que realizam modificações no processo de treinamento dos modelos, por ser a categoria que este trabalho pertence. Em seguida motivamos e nos aprofundamos nas estratégias que propõem a utilização de mecanismos de otimização para efetuar as alterações ao treinamento, que podem se dar através do acréscimo de restrições à otimização, como a restrição da discriminação do modelo, ou então da modificação do que será otimizado, como, por exemplo, remodelando o treinamento enquanto um problema de otimização multi-objetivo. Por fim, expomos trabalhos que, assim como este, utilizam da otimização multi-objetivo, destacando suas principais diferenças teóricas e práticas com este.

Na próxima seção definimos a otimização multi-objetivo, juntamente com o método que utilizaremos para resolver os problemas de otimização que definiremos enquanto treinamento do modelo. Na seção seguinte, apresentamos nossas duas propostas de modelagem multi-objetivo para o treinamento de modelos de classificação binária. Ao fim da seção apresentamos, também, as definições e métricas de discriminação que utilizaremos para avaliar os modelos na seção seguinte de resultados.

3 Metodologia

Dado um conjunto de N amostras $x_i \in \mathbb{R}^d : i \in 1, \dots, N$, onde cada x_i corresponde a um indivíduo, e que dentre os d atributos há uma informação sensível $a \in \{1, \dots, G\}$. Essa informação estratifica essas amostras entre G grupos e o aprendizado pode gerar viés favorecendo um certo grupo. Dado a classificação binária $y_i \in \{0, 1\} : i \in 1, \dots, N$, nosso objetivo é encontrar um conjunto de classificadores que otimizem a qualidade de predição ao mesmo tempo que reduzem a discriminação gerada pelo modelo.

Sabendo que a predição da máquina de aprendizado pode oferecer algum benefício para cada um desses grupos, nós consideramos que existe um conflito dentre esses grupos. Isso ocorre pois aumentar o benefício para um certo grupo, pode reduzir o benefício para outro grupo. Com isso, utilizamos da formulação de otimização multi-objetivo (definida na Seção 3.1) para reescrever o problema de treinar os classificadores em um problema com diversos objetivos, agora considerando métricas para o grupo com atributo sensível e o grupo sem atributo sensível, além do termo de regularização. Na formulação apresentada na Figura 1, o grupo sensível é o de mulheres, representado por M , e o grupo não sensível é o de homens, representado por H .

Para resolvermos o problema multi-objetivo, aplicaremos o método MONISE (explicado em 3.1.1), que utiliza da metodologia de soma ponderada para encontrar um subconjunto representativo do conjunto de soluções ótimas do problema, que é chamado de Fronteira de Pareto. Por ser um método de soma ponderada, reescrevemos o objetivo de forma a ter uma única função objetivo (de forma detalhada em 3.2), que é justamente a soma ponderada dos diversos objetivos. Um dos objetivos de soma ponderada que propomos é o apresentado em (a) na Figura 1.

Após utilizar o MONISE (b), teremos a Fronteira de Pareto do problema, que é o conjunto de hiperparâmetros que geram os modelos de regressão logística que possuem os melhores valores para as métricas escolhidas nos grupos (sensível e não sensível) e o termo de regularização. Essa fronteira gerada é representada em (c) na imagem. Na Figura 1 não temos o termo de regularização, para que a figura fique em duas dimensões e, com isso, seja mais facilmente interpretada.

Com esse conjunto de modelos encontrados pelo MONISE, selecionamos os modelos que otimizam unicamente a acurácia e três métricas de justiça, Igualdade de Oportunidade, Paridade Demográfica e Coeficiente de Variação. Essas métricas de justiça são apresentadas em 3.3. Ao final, teremos um modelo para cada métrica, o que é representado em (d) na Figura 1.

Além disso, também exploraremos metodologias de *ensemble* para encontrar, dado o

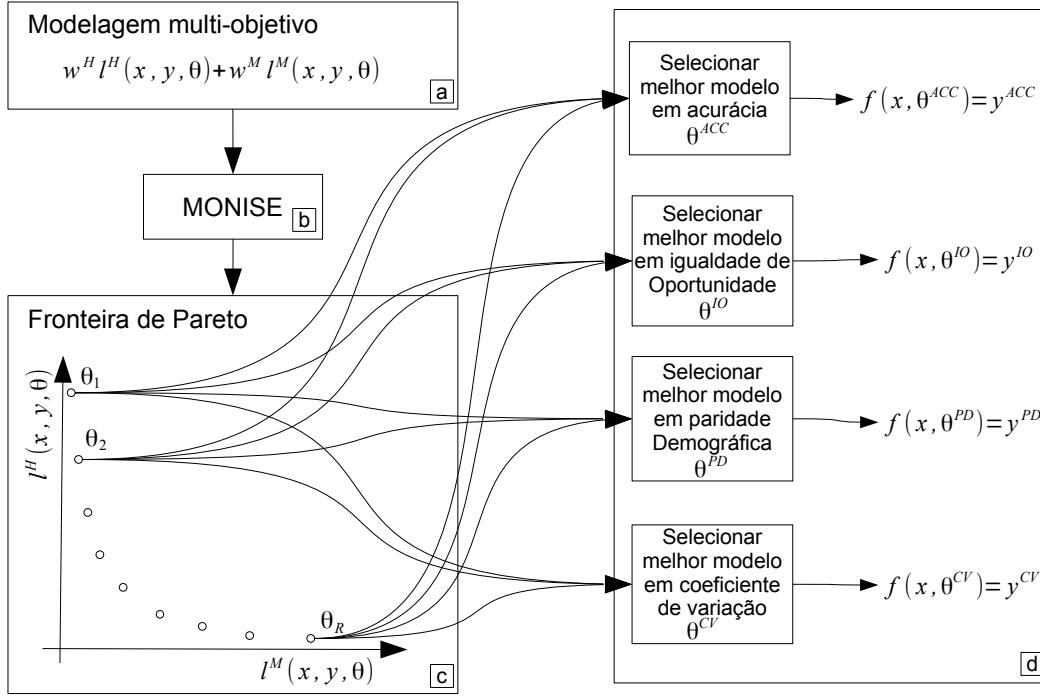


Figura 1: Descrição visual da metodologia proposta neste trabalho. Em (a) temos a modelagem por soma ponderada dos objetivos que minimizaremos, em (b) utilizamos a abordagem MONISE para resolver o problema multiobjetivo, resultando na Fronteira de Pareto apresentada em (c). Em (d), selecionamos dentre os modelos de (c) aqueles que minimizam unicamente cada uma das métricas de desempenho e discriminação escolhidas.

conjunto de modelos, um novo modelo que se baseia nos anteriores, agregando seus resultados, de forma obter resultados que obtenham bom desempenho em todas as métricas utilizadas, não em uma única específica. Na Seção 3.4 detalhamos o método de *ensemble* que utilizaremos neste trabalho.

3.1 Otimização multi-objetivo

Na otimização multi-objetivo, mais de uma função objetivo é otimizada simultaneamente. Sejam f_1, f_2, \dots, f_m m funções a serem otimizadas, com $f_i(\mathbf{x}) : \Omega \rightarrow \mathbb{R}$ e $\mathbf{x} \in \Omega, \Omega \subset \mathbb{R}^d$, então o problema de otimização multi-objetivo é definido como:

Definition 3.1 (*Problema multi-objetivo*) [15]

$$\begin{aligned} & \underset{\mathbf{x}}{\text{minimize}} && f(\mathbf{x}) = \{f_1(\mathbf{x}), f_2(\mathbf{x}), \dots, f_m(\mathbf{x})\} \\ & \text{subject to} && \mathbf{x} \in \Omega, \Omega \subset \mathbb{R}^d \\ & && f(\mathbf{x}) : \Omega \rightarrow \Psi, \Psi \subset \mathbb{R}^m \end{aligned}$$

Dadas duas soluções possíveis $\mathbf{x}^i, \mathbf{x}^j \in \mathbb{R}^d$, diz-se que \mathbf{x}^i **domina** \mathbf{x}^j se $f_k(\mathbf{x}^i) \leq f_k(\mathbf{x}^j), \forall k \in 1, \dots, m$ e $\exists k : f_k(\mathbf{x}^i) < f_k(\mathbf{x}^j)$, isto é, se \mathbf{x}^i é melhor que \mathbf{x}^j em todas as métricas. Neste caso fica claro que \mathbf{x}^i é melhor do que \mathbf{x}^j .

Já se existe um objetivo k onde $f_k(\mathbf{x}^i) < f_k(\mathbf{x}^j)$ e outro objetivo l tal que $f_l(\mathbf{x}^i) > f_l(\mathbf{x}^j)$, isto é, \mathbf{x}^i é melhor que \mathbf{x}^j no objetivo k e \mathbf{x}^j é melhor do que \mathbf{x}^i no objetivo l , torna-se impossível dizer qual é melhor entre \mathbf{x}^i e \mathbf{x}^j . Tal situação é chamada de **não dominância**.

Os métodos de otimização multi-objetivo encontram um conjunto de soluções ótimas de Pareto, que são soluções que não são dominadas por nenhuma outra e que não dominam uma a outra. Esse é o equivalente a soluções ótimas de otimização com apenas um objetivo. Cada uma dessas soluções representa um *tradeoff* entre os objetivos.

Um exemplo de problema multi-objetivo é decidir qual CPU comprar. Tendo as informações de preço e desempenho (*benchmark score*) para várias CPUs diferentes, deseja-se comprar aquela que possui o menor preço e, ao mesmo tempo, o melhor desempenho. Na próxima imagem, cada ponto representa uma CPU, com o eixo horizontal sendo o preço e o vertical sendo o inverso do desempenho. Como esperado, não existe um único computador que possui tanto o melhor desempenho quanto o menor preço. Os pontos mais escuros são parte do conjunto de soluções ótimas de Pareto e os pontos em cinza são dominados pelos pontos escuros.

Para resolver os problemas multi-objetivos, utilizaremos a abordagem chamada MONISE [16].

3.1.1 MONISE - Many-Objective Non-Inferior Set Estimation

A abordagem para problemas multi-objetivos MONISE [16] encontra globalmente o conjunto de soluções ótimas de Pareto, formulando e resolvendo uma sequência de problemas como o seguinte:

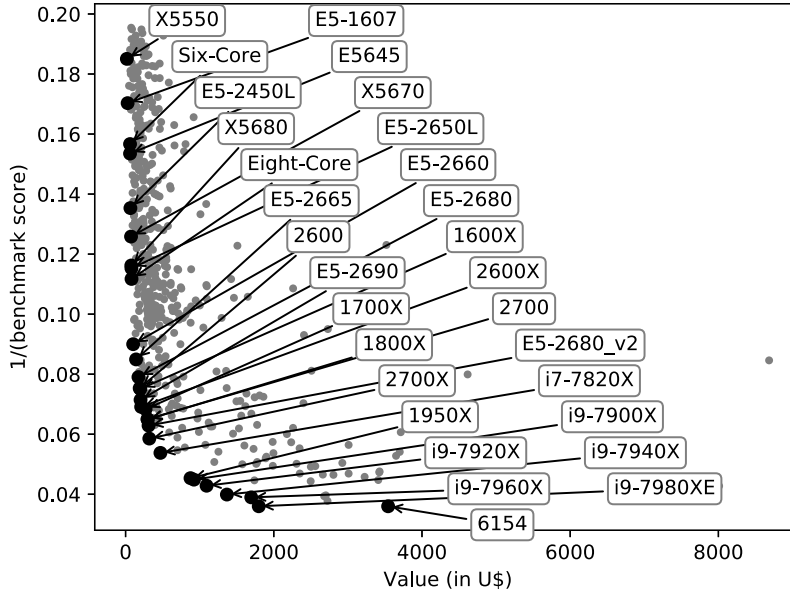


Figura 2: Multi-objective representation of high-end CPUs `cpubenchmark.net` with 485 CPUs and only 29 Pareto-optimal CPUs.

Definition 3.2 *Método de soma ponderada [17]*

$$\begin{aligned}
 & \underset{\mathbf{x}}{\text{minimize}} && \mathbf{w}^T f(\mathbf{x}) \\
 & \text{subject to} && \mathbf{x} \in \Omega, \Omega \subset \mathbb{R}^d \\
 & && f(\mathbf{x}) : \Omega \rightarrow \Psi, \Psi \subset \mathbb{R}^m
 \end{aligned}$$

onde $w_i \geq 0, \forall i \in \{1, 2, \dots, m\}$ e $\mathbf{w}^T \mathbf{1} = 1$.

Para cada \mathbf{w} encontramos uma única solução eficiente, mas deseja-se encontrar um conjunto de soluções de Pareto bem distribuídas. Assim, o MONISE é capaz de encontrar sequencialmente as R soluções mais representativas, que depois serão selecionadas de acordo com a preferência a posteriori do tomador de decisão.

Utilizamos MONISE por nos permitir ter três ou mais funções a serem otimizadas, em detrimento do método NISE (*Non-Inferior Set Estimation*), que também utiliza do método de soma ponderada para encontrar a fronteira de Pareto, mas que não possui bons resultados para problemas com mais de duas funções objetivo.

3.2 Modelos de aprendizado e o conflito entre objetivos

Um problema de classificação binária é dado por um conjunto de N amostras, onde $\mathbf{x}_i \in \mathbb{R}^d : i \in \{1, \dots, N\}$, e $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ consiste nos atributos de entrada e $y_i \in$

$\{0, 1\} : i \in \{1, \dots, N\}$ é o valor de saída que se deseja prever. Em classificação, esse valor indica a pertinência à classe 0 ou à classe 1. A regressão logística consiste em escolher como modelo a função sigmoide $f(x, \theta) = \frac{e^{\theta^\top \phi(\mathbf{x})}}{1 + e^{\theta^\top \phi(\mathbf{x})}} \in [0, 1]^1$, que descreve a probabilidade de uma nova amostra \mathbf{x} ter a sua pertinência vinculada ao grupo 1, e usar o seguinte problema de otimização para encontrar o vetor de parâmetros θ que faz a sigmoide melhor se adequar aos dados:

$$\min_{\theta} - \sum_{i=1}^N \left[y_i \ln \left(\frac{e^{\theta^\top \phi(\mathbf{x}_i)}}{1 + e^{\theta^\top \phi(\mathbf{x}_i)}} \right) + (1 - y_i) \ln \left(1 - \frac{e^{\theta^\top \phi(\mathbf{x}_i)}}{1 + e^{\theta^\top \phi(\mathbf{x}_i)}} \right) \right] + \lambda \|\theta\|_2. \quad (1)$$

sendo $l(\mathbf{X}, \mathbf{y}, \theta) = \sum_{i=1}^N - \left[y_i \ln \left(\frac{e^{\theta^\top \phi(\mathbf{x}_i)}}{1 + e^{\theta^\top \phi(\mathbf{x}_i)}} \right) + (1 - y_i) \ln \left(1 - \frac{e^{\theta^\top \phi(\mathbf{x}_i)}}{1 + e^{\theta^\top \phi(\mathbf{x}_i)}} \right) \right]$ a função de perda. Essa função detecta o quão distante as N amostras estão de realizar predições corretas. O segundo termo de 1, $\|\theta\|_2$, é a função de regularização, que tenta fazer o modelo ser o mais simples possível.

Entretanto, é possível perceber que otimizar esse problema não leva em conta a natureza de cada amostra, considerando todas elas como equivalentes. Neste contexto, pode ser que uma máquina de aprendizado prejudique um grupo desprivilegiado pois o processo de otimização é capaz de acertar mais amostras, em média, ao privilegiar um outro grupo. Assim, propomos duas estratégias diferentes para reescrever o problema de forma a levar em consideração a característica sensível.

3.2.1 Primeira proposta: Erro por grupo

Primeiramente vamos estratificar as amostras em G grupos sensíveis. Assim, podemos criar conjuntos \mathcal{G}^g para cada grupo sensível $g \in \{1, \dots, G\}$. Com isso, é possível modelar a perda no aprendizado para cada um dos grupos g :

$$l^g(\mathbf{X}, \mathbf{y}, \theta) = \sum_{i \in \mathcal{G}^g} - \left[y_i \ln \left(\frac{e^{\theta^\top \phi(\mathbf{x}_i)}}{1 + e^{\theta^\top \phi(\mathbf{x}_i)}} \right) + (1 - y_i) \ln \left(1 - \frac{e^{\theta^\top \phi(\mathbf{x}_i)}}{1 + e^{\theta^\top \phi(\mathbf{x}_i)}} \right) \right] \quad (2)$$

Por fim, é possível afirmar que cada uma das funções de perda $l^g(\mathbf{X}, \mathbf{y}, \theta)$, $g \in \{1, \dots, G\}$ são conflitantes entre si, assim como conflitantes com a regularização $\|\theta\|_2$. Com isso é possível modelar esse problema como otimização multi-objetivo sob a forma da soma ponderada:

¹Por questões de conveniência, nós usamos a função $\phi(\mathbf{x}) = [\phi_0(\mathbf{x}), \dots, \phi_d(\mathbf{x})]^\top$, na qual $\phi_0(\mathbf{x}) = 1$, $\phi_i(\mathbf{x}) = \mathbf{x}_i \forall i \neq 0$, $i \in \{1, \dots, d\}$. Essa notação é usada para adicionar um termo de viés na modelagem.

$$\min_{\theta} \sum_{g=1}^G \mathbf{w}_g l^g(\mathbf{X}, \mathbf{y}, \theta) + \mathbf{w}_{G+1} \|\theta\|_2. \quad (3)$$

sendo \mathbf{w} o vetor de pesos que será encontrado pelo método MONISE.

Com isso, encontramos R soluções para o modelo na Equação 3 utilizando o método MONISE. Existem infinitas soluções para 3, devido aos diferentes possíveis valores de \mathbf{w} , então nos limitamos a um subconjunto com R de tais soluções. O MONISE se encarregará de encontrar os \mathbf{w} de forma que as soluções encontradas sejam as mais representativas do conjunto. Em posse desses modelos, é possível escolher o melhor modelo de acordo com um conjunto de métricas de desempenho e discriminação. Na próxima seção apresentaremos algumas métricas de discriminação bastante utilizadas na literatura de *fairness*.

Ao modificar o problema de forma a considerar o erro para cada grupo de acordo com a característica sensível, encontramos um conjunto de modelos que possui valores diversos de erro para cada grupo, com modelos cujos valores são distantes e também modelos em que os erros para os grupos são próximos. Assim, se estamos considerando enquanto métrica de justiça que queremos que todos os grupos tenham taxa de acerto parecidas, esta proposta tem o potencial de satisfazer bem essa métrica. Entretanto, outras métricas de justiça propõem que o modelo ideal não é focado em diminuir o erro para cada grupo, mas sim que os grupos diferentes possuem taxas de aceitação não muito distoantes. Isto é, considerando o caso de que queremos um modelo para decidir se uma pessoa deve ou não ser contratada para um certo cargo, uma métrica de justiça possível aqui é de que o modelo não deveria aceitar mais homens do que mulheres, mesmo que nos dados fornecidos isto ocorra. Assim, propomos uma outra estratégia para esses casos, que possui como um dos objetivos a serem otimizados a taxa de pessoas aceitas para cada um dos grupos.

3.2.2 Segunda proposta: Aceitação por grupo

Se desejamos que nosso classificador tenha taxas de aceitação parecida entre os grupos, uma abordagem possível para atingir este objetivo é adicionar as taxas de aceitação no conjunto de funções que queremos otimizar. Desta forma, nesta segunda proposta também utilizamos a separação dos grupos.

Para cada grupo, temos a taxa de aceitação dada por:

$$a^g(\mathbf{X}, \theta) = - \sum_{i \in \mathcal{G}^g} \ln \left(\frac{e^{\theta^\top \phi(\mathbf{x}_i)}}{1 + e^{\theta^\top \phi(\mathbf{x}_i)}} \right) \quad (4)$$

Além disso, também utilizamos como objetivo do problema minimizar a perda do

aprendizado de forma geral, sem separar por grupos, definida da seguinte forma:

$$l(\mathbf{X}, \mathbf{y}, \theta) = - \sum_{i=1}^N \left[y_i \ln \left(\frac{e^{\theta^\top \phi(\mathbf{x}_i)}}{1 + e^{\theta^\top \phi(\mathbf{x}_i)}} \right) + (1 - y_i) \ln \left(1 - \frac{e^{\theta^\top \phi(\mathbf{x}_i)}}{1 + e^{\theta^\top \phi(\mathbf{x}_i)}} \right) \right] \quad (5)$$

Consideramos a perda enquanto objetivo pois, caso ela não seja considerada, então os modelos resultantes não possuiriam valor prático, já que focariam apenas em aprovar todos os indivíduos.

Como na estratégia anterior, aqui também adicionamos o termo de regularização $\|\theta\|_2$, o que faz com que nosso problema possa ser escrito da seguinte forma de soma ponderada:

$$\min_{\theta} \sum_{g=1}^G \mathbf{w}_g a^g(\mathbf{X}, \theta) + \mathbf{w}_{G+1} l(\mathbf{X}, \mathbf{y}, \theta) + \mathbf{w}_{G+2} \|\theta\|_2. \quad (6)$$

sendo \mathbf{w} o vetor de pesos que será encontrado pelo método MONISE.

Por fim, utilizamos, da mesma forma que na primeira abordagem, o MONISE para encontrar R soluções para o modelo na Equação 6 e selecionamos, considerando as métricas escolhidas, o melhor modelo dentre estes.

A seguir apresentamos algumas das métricas de discriminação que são bastante utilizadas na literatura. Elas serão de extrema importância para a comparação que faremos entre as duas estratégias apresentadas, além da comparação delas com outras abordagens já bem utilizadas.

3.3 Métricas de discriminação

Para estudar e, eventualmente, minimizar a discriminação gerada por modelos de aprendizado de máquina, é preciso definir o que é uma discriminação, assim como quantificar o quão discriminatório é um modelo.

Dado o atributo sensível A , com $A_i = 1$ indicando que o indivíduo i faz parte do grupo sensível estudado e $A_i = 0$ indicando o contrário, um preditor \hat{y} gerado a partir desses dados pode ser considerado justo, ou injusto, através de diversas definições de justiça. Três das mais utilizadas atualmente são:

Definition 3.3 *Paridade Demográfica (DP)[4]. Um preditor \hat{y} satisfaz a paridade demográfica se $P(\hat{y} = 1 | A = 0) = P(\hat{y} = 1 | A = 1)$.*

Esta definição de justiça, definida para problemas de classificação binária, parte da premissa de que a característica sensível que estamos avaliando não deve interferir na chance

de um indivíduo ser classificado positivamente. Então ainda que na realidade a proporção de indivíduos que são classificados positivamente seja diferente com base na característica sensível, esta definição de justiça propõe que o modelo deve forçar com que esta proporção seja a mesma para o grupo de indivíduos com a característica sensível e o grupo sem ela, resolvendo injustiças já existentes.

Definition 3.4 *Igualdade de Oportunidade (OE)[11]. Um preditor \hat{y} satisfaz a igualdade de oportunidades se $P(\hat{y} = 1|A = 0, y = 1) = P(\hat{y} = 1|A = 1, y = 1)$.*

Assim como a Paridade Demográfica, a Igualdade de Oportunidade parte da análise de grupos com base na característica sensível, mas, diferente da definição anterior, esta definição não força com que o modelo resolva injustiças já existentes. A Igualdade de Oportunidade procura, na verdade, garantir que o modelo continue classificando os indivíduos positivamente da mesma forma que foram nos dados fornecidos, sem valorizar um grupo em detrimento de outro.

Definition 3.5 *Justiça individual (FI)[18]. Um algoritmo é justo se der previsões semelhantes a indivíduos semelhantes. Formalmente, dada uma métrica $d(\cdot, \cdot)$, se os indivíduos i e j são semelhantes sob essa métrica (isto é, $d(i, j)$ é pequeno), então suas previsões devem ser semelhantes: $\hat{y}(x_i, A_i) \approx \hat{y}(x_j, A_j)$.*

Esta definição de justiça parte da premissa de a justiça deve ser avaliada da perspectiva individual, ao invés da perspectiva de grupos.

Através de definições de desigualdade e justiça são definidas métricas de discriminação, que são utilizadas para quantificar o quão injusto um modelo é e, com isso, permitir que procuremos os modelos mais justos para a tarefa proposta. Neste trabalho serão utilizadas as seguintes:

Definition 3.6 *(Métrica de P Porcento [19])*

$$P\% = \min \left(\frac{P(\hat{y} = 1|A = 1)}{P(\hat{y} = 1|A = 0)}, \frac{P(\hat{y} = 1|A = 0)}{P(\hat{y} = 1|A = 1)} \right)$$

A métrica de P porcento vem da definição de Paridade Demográfica, sendo muitas vezes chamada de métrica de Paridade Demográfica, embora o termo P porcento seja mais utilizado.

Definition 3.7 *Métrica de Igualdade de Oportunidade[11]*

$$IO = \min \left(\frac{P(\hat{y} = 1|A = 1, y = 1)}{P(\hat{y} = 1|A = 0, y = 1)}, \frac{P(\hat{y} = 1|A = 0, y = 1)}{P(\hat{y} = 1|A = 1, y = 1)} \right)$$

Definition 3.8 *Índice de Entropia Generalizado[21]*

Sejam N a quantidade de amostras, $b_i = \hat{y}_i - y_i + 1$ e $\mu = \frac{\sum_{i=1}^N b_i}{N}$:

$$\varepsilon_\alpha(b_1, b_2, \dots, b_n) = \begin{cases} \frac{1}{N\alpha(\alpha-1)} \sum_{i=1}^N \left[\left(\frac{b_i}{\mu} \right)^\alpha - 1 \right], & \text{if } \alpha \neq 0, 1 \\ \frac{1}{N} \sum_{i=1}^N \frac{b_i}{\mu} \ln \frac{b_i}{\mu}, & \text{if } \alpha = 1 \\ -\frac{1}{N} \sum_{i=1}^N \ln \frac{b_i}{\mu}, & \text{if } \alpha = 0 \end{cases}$$

O Índice de Entropia Generalizado é, na verdade, um conjunto de diferentes índices para cada valor de α . Neste trabalho, utilizaremos a métrica Coeficiente de Variação, definida como:

Definition 3.9 *Coeficiente de Variação[22]*

$$CV = 2\sqrt{\varepsilon_{\alpha=2}(b_1, b_2, \dots, b_n)}$$

O Índice de Entropia Generalizado, assim como o Coeficiente de Variação, são considerados métricas baseadas em justiça individual, por não separarem e quantificarem estatísticas com base nos grupos definidos a partir da característica sensível, mas sim nos indivíduos em geral.

A seguinte tabela apresenta uma visão geral das métricas que utilizamos, classificando-as em Justiça por Grupos ou Individual e apresentando qual condição cada uma estabelece.

	Noção de Justiça	Condição de Justiça
Justiça por Grupos	Acurácia	Acurácia igual para todos os grupos
	Igualdade de Oportunidade	Taxa de verdadeiros positivos igual para todos os grupos
	Paridade Demográfica	Taxa de aceitação igual para todos os grupos
Justiça Individual	Índice de Entropia Generalizado	Indivíduos que merecem resultados similares recebem resultados similares

Tabela 1: Visão geral das métricas de desigualdade.

3.4 Ensemble Learning

Para melhorar os modelos de aprendizagem de máquinas, muitas vezes são utilizados métodos de *ensemble learning* [38, 39], ou aprendizado por agrupamento, que consiste em, a partir de um conjunto de modelos treinados para uma mesma tarefa, produzir um novo modelo, mais complexo, que tenha performance mais robusta do que os modelos do conjunto. O objetivo dessa estratégia é a minimizar as desvantagens individuais dos modelos mais simples no modelo final.

Como nas propostas deste trabalho encontramos um conjunto de modelos que otimizam métricas de desempenho e de discriminação, é possível utilizar métodos de *ensemble* para, a partir desse conjunto, gerar um único modelo que tenha boa performance em todas as métricas. Conforme mostrado em [24] e explorado nos experimentos que faremos a seguir, a Fronteira de Pareto é capaz de gerar diversidade de modelos, o que favorece a utilização das metodologias de *ensemble*.

Nos experimentos que faremos sobre as propostas, utilizaremos o método de **votação simples** como *ensemble* para gerar um único modelo dada a Fronteira de Pareto encontrada. Nesse método, cada modelo "vota" no que deveria ser o resultado. Como abordamos problemas de classificação binária, então para cada indivíduo x cada modelo i predirá sua classificação $f(x, \theta_i)$. Para gerar o único modelo, podemos utilizar o método rígido (*hard*), onde cada modelo do conjunto utilizado pode votar se o indivíduo faz parte de um grupo ou do outro, com os votos sendo binários, e a classificação do indivíduo é o grupo que recebeu mais votos; ou então podemos utilizar o método suave (*soft*), onde cada modelo dá a probabilidade do indivíduo pertencer a cada um dos dois grupos, com a classificação do indivíduo sendo o grupo que obteve maior soma das probabilidades preditas.

4 Experimentos

Nesta seção avaliaremos comparativamente o desempenho das estratégias de treinamento de modelos de aprendizagem de máquina através de otimização multi-objetivo propostas neste trabalho. Antes de apresentarmos os resultados obtidos para os experimentos, na Seção 4.1 apresentamos os conjuntos de dados que utilizaremos nos experimentos seguintes e os modelos com que compararemos os resultados obtidos nas propostas. Em seguida, na Seção 4.2 comparamos o desempenho da nossa estratégia com baselines da literatura; neste experimento cada seleção dos modelos leva uma única métrica de desempenho por vez, ao invés de otimizá-las simultaneamente. Tal experimento é importante para validar a competitividade da metodologia proposta em comparação às demais metodologias da literatura.

Na Seção 4.3 comparamos a diversidade de modelos que podemos gerar com cada estratégia proposta em levando em conta as métricas analisadas. Tal diversidade é importante tanto para permitir uma maior liberdade de escolha para utilizador ao que tange diferentes compromissos entre discriminação e desempenho do modelo, assim como propiciar a utilização de métodos de *ensemble learning* para agregar com conjunto de modelos buscando uma maior robustez. Na Seção 4.4 aplicamos método de *ensemble learning* para encontrar um único modelo a partir do conjunto de modelos resultante das estratégias multi-objetivas propostas. Buscamos obter um modelo final que seja mais robusto do que os modelos do conjunto, balanceando o conflito entre desempenho e discriminação. Por fim, na Seção 4.5 apresentamos as conclusões obtidas a partir dos experimentos realizados.

4.1 Configurações dos experimentos

Nas seções seguintes, realizaremos diversos experimentos com as metodologias que propomos neste trabalho. Com isso, apresentamos os conjuntos de dados que utilizaremos nesses experimentos, e apresentamos os modelos com os quais iremos comparar os resultados obtidos nos experimentos.

4.1.1 Dados

Para os experimentos utilizamos quatro conjuntos de dados reais que são bastante comuns na literatura de *fairness* no contexto de aprendizado de máquina. Os dados são de aplicações bastante diferentes, abordando crédito, renda, admissão acadêmica e crime. Esses contextos apresentam viés discriminatório mesmo em situações que não utilizem modelos de aprendizagem de máquina.

German O conjunto de dados *German Credit Data* [3] contém 1000 solicitações de crédito, com informações de quantidade de crédito, duração do pagamento, razão para a solicitação e outros dados em relação à pessoa que fez a solicitação, como idade, se trabalha, se tem muito ou pouco dinheiro guardado e o gênero da pessoa (utilizado como característica sensível), além de indicar se o pedido foi aceito ou negado. O objetivo, com esses dados, foi treinar um modelo para classificar, dado um pedido de crédito, se ele deve ser aceito ou não.

Adult O conjunto de dados *Adult Data Set* [3], também conhecido como *Census Income*, contém informação sobre indivíduos do Censo dos Estados Unidos de 1994. São informações de 48842 pessoas com 14 atributos diferentes, dentre eles gênero (não utilizado como atributo sensível) e raça (utilizado como característica sensível). A tarefa proposta é prever se um certo indivíduo recebe mais ou menos que \$50.000 por ano.

LSAC Os dados do LSAC [27] (*Law School Admission Council*) contém dados de 1.823 alunos de direito, com informações de exames e raça (utilizado como característica sensível). O objetivo é identificar se o aluno passou no exame de ordem ou não.

ProPublica Os dados do ProPublica [26] foram coletados da utilização da ferramenta de risco COMPAS (*Correctional Offender Management Profiling for Alternative Sanctions*). Eles incluem informações sobre 6.167 indivíduos que foram presos, incluindo o grau do incidente, sexo (não utilizado como característica sensível) e raça (utilizado como característica sensível). O objetivo é prever se o indivíduo voltará a ser preso em dois anos.

Em seguida, apresentamos as estratégias de *fairness* com que compararemos os resultados obtidos nos experimentos que faremos comparações com as estratégias propostas.

4.1.2 Abordagens comparadas

A fim de avaliar as hipóteses sobre as propostas apresentadas, é necessário comparar os resultados obtidos nos experimentos que fizemos com outras estratégias de *fairness*. Para isso, escolhemos cinco estratégias que possuem similaridades com as apresentadas nesse trabalho, mas que representam categorias diferentes utilizadas na literatura. Todas as estratégias escolhidas abordam classificações binárias e se baseiam na Regressão Logística.

Temos a Regressão Logística que não apresenta foco algum na discriminação do modelo; *Reweighting* realiza mudanças no problema de otimização do treinamento do modelo,

mas sem acrescentar restrições; Classificador de Paridade Demográfica e Classificador de Igualdade de Oportunidade adicionam restrições ao treinamento; *Minimax* modela o treinamento enquanto um problema de otimização multi-objetivo.

Regressão Logística Abordagem que utilizaremos como base nas comparações, ele utiliza como modelagem a função sigmoide $f(x, \theta) = \frac{e^{\theta^\top \phi(x)}}{1 + e^{\theta^\top \phi(x)}} \in [0, 1]$, sendo f a probabilidade do indivíduo x fazer parte do grupo classificado. O processo de treinamento consiste em encontrar o melhor valor para θ de forma a minimizar o erro da classificação para o conjunto de dados utilizado. Essa abordagem não possui nenhum foco em minimizar a discriminação do modelo, focando apenas em maximizar o desempenho.

Reweighting A abordagem de *reweighting* [7] ou reponderação modifica a regressão logística de forma a considerar pesos diferentes para os indivíduos. Ela separa os indivíduos em grupos de acordo com a característica sensível e a classificação que busca replicar (então, se a classificação for binária e a característica sensível também, teremos quatro grupos). Para cada grupo ela calcula o peso que corresponde a razão entre a probabilidade esperada de que aquele grupo tivesse aquela classificação em um mundo sem preconceitos e a probabilidade real apresentada no conjunto de dados. Com esses pesos, é encontrado o parâmetro θ da função sigmoide de forma a minimizar o erro no conjunto de dados do treinamento, com esse erro tendo o peso calculado para cada indivíduo.

Classificador de Paridade Demográfica O classificador de Paridade Demográfica [19] reescreve o problema de otimização da regressão logística de forma a adicionar uma restrição à otimização para que a discriminação do modelo encontrado não ultrapasse um certo valor. Como o nome indica, ele utiliza como métrica de discriminação a Paridade Demográfica. Entretanto, devido à métrica não ser convexa e outras características dela, o classificador não a utiliza enquanto restrição, mas adapta ela em uma restrição que indiretamente limita a discriminação.

Classificador de Igualdade de Oportunidade Parecido com Classificador de Paridade Demográfica, o Classificador de Igualdade de Oportunidade [9] modifica a regressão logística adicionando uma restrição sobre a discriminação gerada pelo modelo, restrição essa baseada na métrica de Igualdade de Oportunidade.

Minimax A abordagem *Minimax Pareto Fairness* [20] modela o treinamento como um problema de otimização multi-objetivo em que as funções a serem otimizadas são o erro do modelo para os grupos de acordo com a característica sensível. Ela busca minimizar o maior dentre os erros dos grupos, encontrando um único modelo.

Além disso, vale lembrar as duas abordagens propostas nesta pesquisa:

Multi-Objetivo de Erros Abordagem proposta que considera que o erro de aprendizado para cada um dos grupos é conflitante entre si, além de ser conflitante com a regularização.

Multi-objetivo de Probabilidades Abordagem proposta que considera que a probabilidade de obter o recurso desejado (saída 1 na classificação binária) em cada um dos grupos é conflitante entre si, além de ser conflitante com o erro de aprendizado e a regularização.

Para mais detalhes, consulte a Seção 3.2.

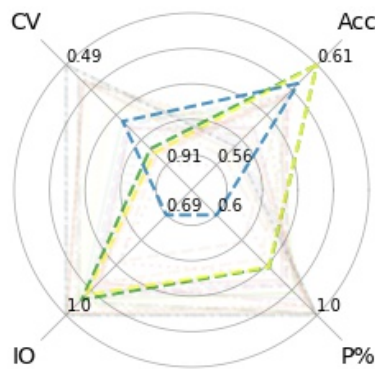
4.2 Experimento de otimização individual

No primeiro experimento testamos a validade da metodologia proposta em comparação à literatura quando se utiliza apenas uma métrica para selecionar os modelos. Esse não é o principal foco da metodologia proposta, cuja principal vantagem é permitir a otimização simultânea das métricas, mas é algo importante a se avaliar. Assim como nos próximos experimentos, serão utilizadas as seguintes métricas para avaliar os modelos: erro de treinamento – acurácia (ACC); igualdade de oportunidade (IO); paridade demográfica (PD); coeficiente de variação (CV). Entretanto, nesse experimento utilizaremos cada uma dessas métricas de forma isolada para selecionar o melhor modelo através da otimização dessa métrica na validação.

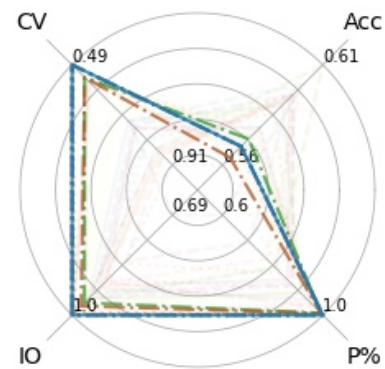
Para realizar o experimento aplicamos nossas propostas em cada um dos conjuntos de dados escolhidos, o que resulta em um conjunto de modelos de aprendizagem de máquina para cada um. Com o conjunto de modelos, escolhemos aquele que otimiza cada uma das métricas. Assim, para cada conjunto de dados ficamos com oito modelos, sendo dois para cada métrica otimizada, um para cada proposta apresentada.

Para as abordagens com que comparamos, utilizamos uma metodologia de ajuste de hiper-parâmetros chamada *optuna* (optuna.org), a qual dado uma métrica de performance na validação, ela um único modelo cujos hiper-parametros são mais adequados para otimizar tal métrica. Por fim teremos 28 modelos escolhidos para cada conjunto de dados. Para este experimento, separamos cada base de dados em três conjuntos, o conjunto de treinamento, que será utilizado para treinar os diversos modelos para cada estratégia, o conjunto de validação, que será utilizado para selecionar os modelos treinados de acordo com suas métricas, e o conjunto de teste, que será utilizado para avaliar os modelos selecionados. No apêndice apresentamos as tabelas de resultado das métricas para o conjunto de teste. Foi utilizada a estratégia de *cross validation* por *K-Folds* em cinco *folds* para criar cinco conjuntos de testes, de forma a fazer uma médias das performance e garantir a consistência dos resultados.

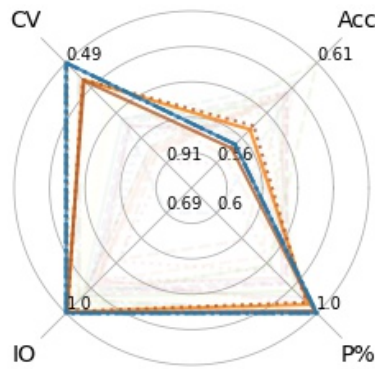
A seguir apresentamos algumas visualizações que representam, para cada uma das métricas, os modelos que tiveram melhor performance na métrica. É importante ressaltar que cada métrica tem uma escala diferente, a fim de representar melhor seus valores. Em especial, o coeficiente de variação (CV) está com eixo invertido, pois menores valores indicam melhores modelos enquanto as demais métricas indicam modelos melhores quanto maior forem os valores. Cada modelo é representado por quatro retas nessa visualização em radar, com a cor indicando qual a estratégia utilizada no modelo, com as nossas propostas estando nas cores vermelho e azul, e o estilo da linha indicando qual foi a métrica otimizada pelo modelo.



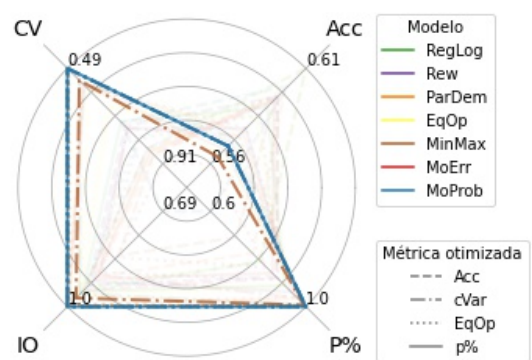
(a) Acurácia



(b) Coeficiente de variação



(c) Igualdade de oportunidade



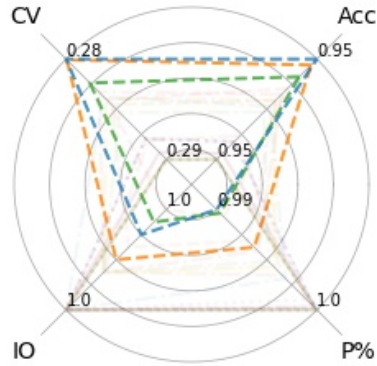
(d) Paridade demográfica

Figura 3: Melhores resultados obtidos em cada métrica para o conjunto de dados *German* para as estratégias propostas e comparadas.

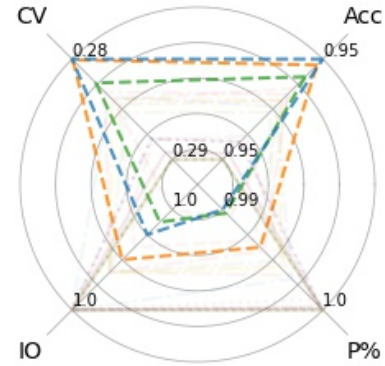
Os resultados apresentados na Figura 3 foram gerados com o conjunto de dados *German* são bastante interessantes. Primeiro, fica claro visualmente o conflito entre o desem-

penho e a discriminação dos modelos, com as visualizações que focam em cada uma das métricas de discriminação possuindo valores de acurácia significativamente inferiores que os obtidos ao focar na acurácia em si e a visualização que foca na acurácia possuindo valores também inferiores para as métricas de discriminação, principalmente em coeficiente de variação e paridade demográfica. Além disso, temos que a regressão logística e a classificação de igualdade de oportunidade possuíram os maiores valores para a acurácia, o que é esperado da regressão logística, já que ele seria o modelo padrão escolhido caso o problema de desigualdade não fosse levado em consideração.

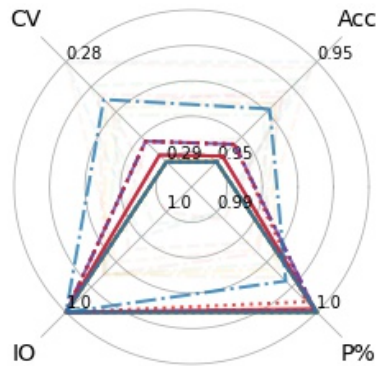
Observando os modelos propostos neste trabalho, temos que as variações do modelo multi-objetivo de probabilidades estão presente entre os melhores modelos para cada uma das métricas, o que indica que ele foi bem sucedido em otimizar individualmente para cada uma das métricas isoladamente. Para as métricas de discriminação, o modelo teve bons resultados para todas simultaneamente, mas com acurácia comparativamente baixa. Já quando ele é focado em acurácia os valores para as métricas de discriminação decrescem bastante, principalmente para Igualdade de Oportunidade e P Porcento.



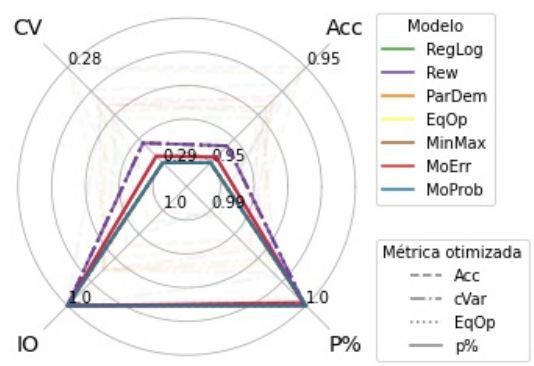
(a) Acurácia



(b) Coeficiente de variação



(c) Igualdade de oportunidade



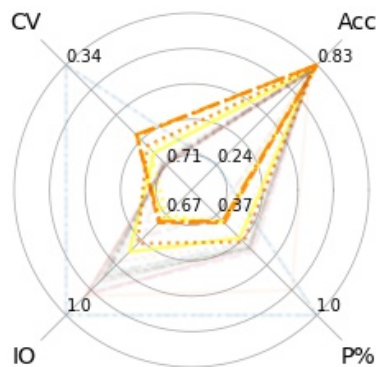
(d) Paridade demográfica

Figura 4: Melhores resultados obtidos em cada métrica para o conjunto de dados LSAC para as estratégias propostas e comparadas.

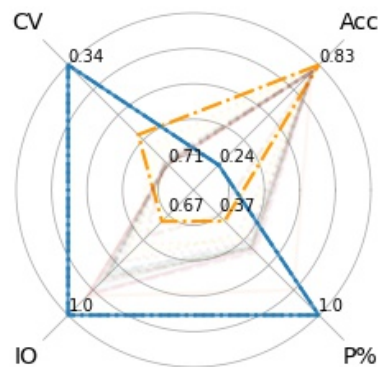
Na Figura 4, com os resultados para o conjunto de dados LSAC, percebemos um padrão diferente em relação à primeira, com as visualizações que focam em acurácia e coeficiente de variação possuindo os mesmos modelos, não indicando conflito entre eles, mas o conflito entre acurácia e paridade demográfica se mantém. Focando na acurácia, temos que a regressão logística também está presente entre os melhores modelos, mas não mais como o melhor, sendo substituída pela nossa proposta de probabilidade. Observando as visualizações que focam nos melhores modelos em igualdade de oportunidade e paridade demográfica, percebe-se que os modelos selecionados por cada um deles são praticamente os mesmos, com a igualdade de oportunidade tendo nossa proposta que utiliza das probabilidades, sendo otimizada para coeficiente de variação, enquanto a paridade demográfica não o contém.

Os modelos propostos neste trabalho também obtiveram bons resultados para este con-

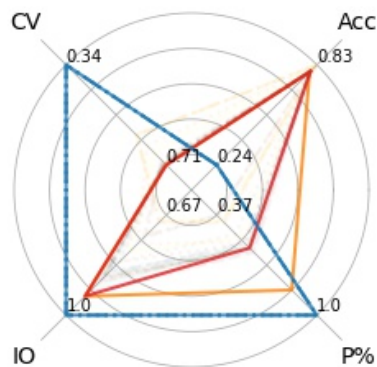
junto de dados. O modelo multi-objetivo de probabilidades focado em acurácia teve valores muito bons para acurácia e coeficiente de variação, enquanto que o focado em coeficiente de variação foi um dos melhores em Igualdade de Oportunidade e o focado em P por cento foi um dos melhores tanto para Igualdade de Oportunidade quanto para P por cento. Já o modelo multi-objetivo obteve bons resultados para Igualdade de Oportunidade e Paridade Demográfica ao focar nelas.



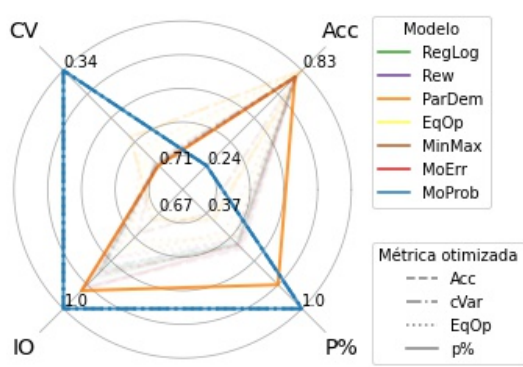
(a) Acurácia



(b) Coeficiente de variação



(c) Igualdade de oportunidade

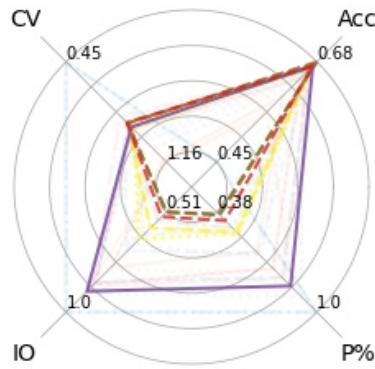


(d) Paridade demográfica

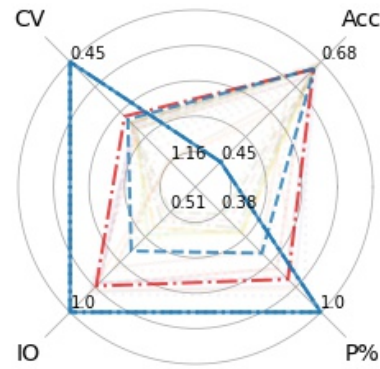
Figura 5: Melhores resultados obtidos em cada métrica para o conjunto de dados *Adult* para as estratégias propostas e comparadas.

Para o conjunto de dados *Adult*, percebemos a partir da Figura 5 que os valores para acurácia são quase todos muito próximos, com os maiores sendo observados para o Classificador de Paridade Demográfica e o Classificador de Igualdade de Oportunidade. Entretanto, para a métrica de discriminação Coeficiente de Variação o modelo Multi-objetivo de Probabilidades possui desempenho significativamente superior aos demais. Tal modelo também

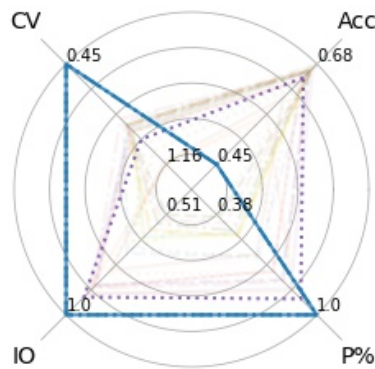
apresenta os melhores valores para Igualdade de Oportunidade e Paridade Demográfica, maximizando simultaneamente todas as métricas de discriminação.



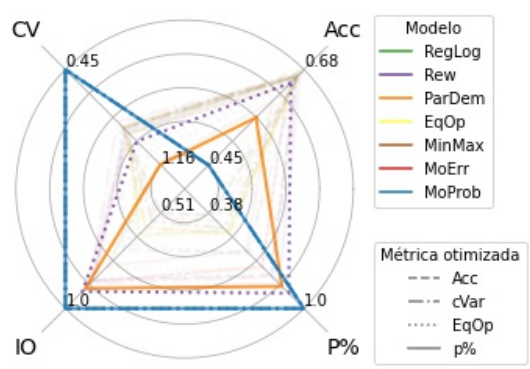
(a) Acurácia



(b) Coeficiente de variação



(c) Igualdade de oportunidade



(d) Paridade demográfica

Figura 6: Melhores resultados obtidos em cada métrica para o conjunto de dados COMPAS para as estratégias propostas e comparadas.

Por fim, para o conjunto de dados COMPAS, como apresentado na Figura 6, o modelo Multi-Objetivo de Probabilidades consegue otimizar simultaneamente as métricas de discriminação, mas com uma perda grande de acurácia. Já o modelo Multi-Objetivo de Erros consegue otimizar a acurácia, sendo um dos modelos com maiores valores para tal métrica, e também consegue manter a acurácia alta tendo valores razoáveis para as métricas de discriminação.

Com esse experimento conseguimos perceber que os modelos propostos neste trabalho conseguem otimizar individualmente cada uma das métricas escolhidas, estando frequentemente entre os com melhores valores para elas. Além disso, temos que o modelo

Multi-objetivo de probabilidades consegue, muitas vezes, otimizar simultaneamente todas as métricas de discriminação escolhidas, mas com uma perda significativa em acurácia. Tal capacidade de otimizar individualmente cada uma das métricas pode indicar uma ampla diversidade. Se tal propriedade for bem explorada por metodologias de ensemble, pode gerar modelos que tenham bom desempenho de aprendizado e sejam capazes de reduzir a discriminação.

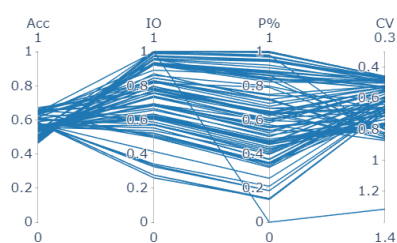
A seguir, realizamos outro experimento, com o objetivo de avaliar a diversidade de modelos que cada estratégia pode gerar dado um mesmo conjunto de dados, o que dá ao utilizador mais liberdade de escolha em relação ao modelo final para sua aplicação, além de favorecer a utilização de estratégias de *ensemble learning*.

4.3 Experimento de diversidade

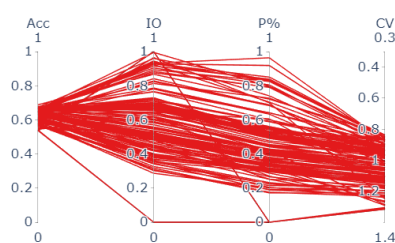
No segundo experimento exploramos a habilidade dos modelos propostos e comparados em encontrar modelos com características diversas para o mesmo conjunto de dados. Tal característica é importante por dois principais motivos: (1) Permitir que o utilizador possa escolher o compromisso entre as métricas de desempenho e discriminação que faça mais sentido para sua aplicação, a diversidade dos modelos fornece ao usuário mais possibilidades a serem escolhidas; (2) Possibilitar a utilização de *ensemble learning* para, a partir do conjunto diverso de modelos, gerar um único modelo mais robusto que os anteriores.

Neste experimento, analisaremos os valores obtidos para as métricas de desempenho e discriminação para os modelos dos conjuntos resultantes das estratégias multi-objetivo propostas. Já para as estratégias com que comparamos, geramos conjuntos de modelos através da variação dos hiper parâmetros delas, da mesma forma que foi feito no primeiro experimento. Diferentemente do primeiro experimento, não selecionaremos o melhor modelo, mas fazemos a geração de diversos modelos para cada metodologia variando os valores dos hiper-parâmetros.

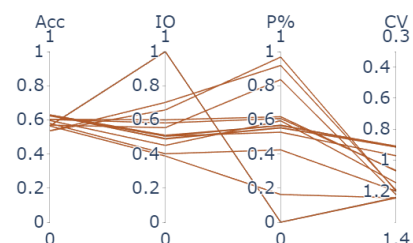
A seguir apresentamos uma visualização dos resultados obtidos para o experimento para o conjunto de dados COMPAS. Os resultados para os demais conjuntos de dados estão disponíveis no apêndice. Tal conjunto de dados foi escolhido por exibir uma maior diversidade de métricas para todas as estratégias analisadas. No apêndice estão presentes as visualizações desta forma para os demais conjuntos de dados analisados.



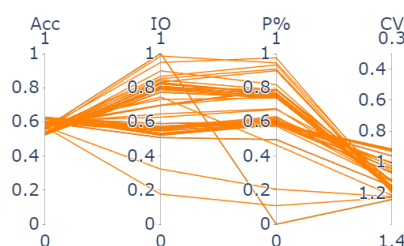
(a) Multi-objetivo - Probabilidade



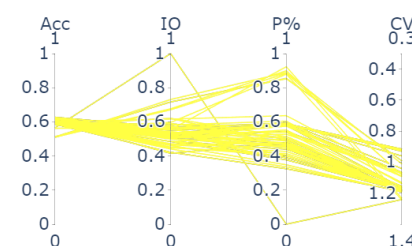
(b) Multi-objetivo - Erro



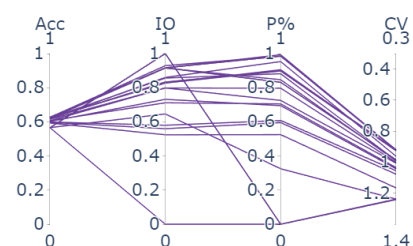
(c) Minimax



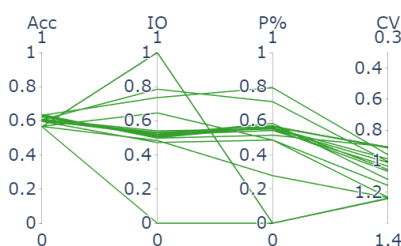
(d) Class. de Paridade Demográfica



(e) Class. de Igualdade de Oportunidade



(f) Reweighting



(g) Regressão Logística

Figura 7: Valores encontrados para métricas de desempenho e discriminação para os modelos resultantes de cada estratégia utilizando o conjunto de dados COMPAS.

Visualmente percebemos na Figura 7 que as estratégias multi-objetivo propostas possuem maior diversidade de valores para as métricas, principalmente as de discriminação. É interessante que os modelos multi-objetivos não são diversos apenas nos valores individuais das métricas, mas também na combinação entre elas. Para este conjunto de dados em específico, temos que a estratégia multi-objetiva que utiliza os erros dos grupos gera modelos que coeficientes de variação melhores do que a estratégia multi-objetiva de probabilidades.

Com a diversidade de modelos gerada pelas estratégias multi-objetivo propostas, exploraremos, no próximo experimento, a capacidade desses modelos de gerarem, com a metodologia de *ensemble learning*, um único modelo que seja mais robusto. Isso será feito através de seleções de subconjuntos de modelos de acordo com os valores obtidos para as métricas.

4.4 Experimento de *ensemble learning*

Nos experimentos anteriores exploramos como as estratégias multi-objetivo propostas para o treinamento de modelos de inteligência artificial que otimizam simultaneamente desempenho e discriminação conseguem bons resultados para as métricas escolhidas em comparação com outras estratégias similares da literatura de *fairness*, além de encontrarem um conjunto de modelos bastante diverso em relação a tais métricas. Entretanto, da forma que as estratégias foram propostas, elas retornam um conjunto de modelos, deixando a cargo do utilizador escolher qual dos modelos será utilizado.

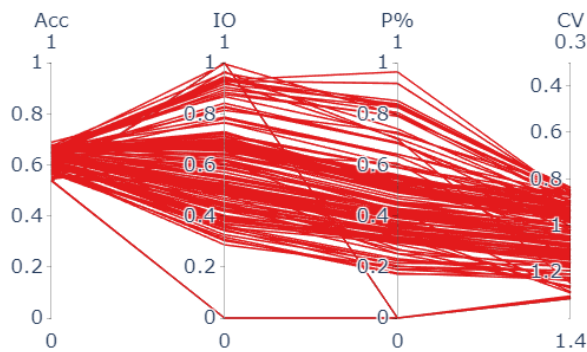
Para encontrarmos um único modelo a partir das estratégias propostas, sem depender da escolha do utilizador, podemos utilizar métodos de *ensemble learning* para gerar, a partir do conjunto de modelos encontrado pela estratégia, um único modelo que seja mais robusto que os anteriores. No terceiro e último experimento deste trabalho, exploraremos a utilização do método de *ensemble learning* de votação simples suave (*soft simple voting*) com os conjuntos de modelos gerados pelas estratégias propostas. Será utilizado o conjunto de dados COMPAS para isso, por ele apresentar maior diversidade de modelos gerados.

Analisaremos o modelo *ensemble* gerado a partir de diversas seleções diferentes de modelos resultantes de cada uma das estratégias propostas, comparando-o com os modelos escolhidos a partir da otimização individual de cada uma das métricas. A seleção dos modelos será feita utilizando os valores obtidos para as métricas no subconjunto de dados de validação e o resultado apresentado será o obtido para o subconjunto de dados de teste. Começamos a análise pela estratégia multi-objetivo de erros:

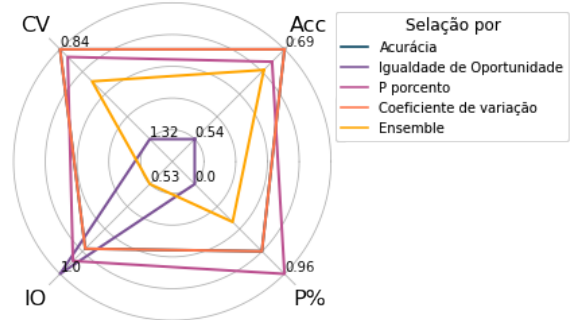
Na Figura 8 temos o resultado para o *ensemble* e os demais modelos de seleção quando não adicionamos restrições ao conjunto de modelos utilizado para o treinamento. Temos que o modelo de *ensemble* possui valores para as métricas de discriminação consideravelmente inferiores aos demais modelos e sua acurácia também é ligeiramente inferior. Seleccionamos, então, os modelos que obtiveram os maiores valores para acurácia no subconjunto de dados de validação.

Seleccionando os melhores modelos em acurácia, conforme apresentado na Figura 9, temos que não só a acurácia do modelo de *ensemble* aumenta, mas também os valores para as métricas de discriminação melhoram. Entretanto, o valor para Igualdade de Oportunidade continua significativamente inferior aos demais modelos seleccionados. Assim, seleccionamos, também, os modelos com os melhores valores para tal métrica.

Na Figura 10 temos os resultados obtidos para a seleção de modelos tanto por acurácia quanto por Igualdade de Oportunidade. Com essa seleção, temos que o modelo de *ensemble* melhora sua Igualdade de Oportunidade em relação às duas seleções anteriores, mas man-

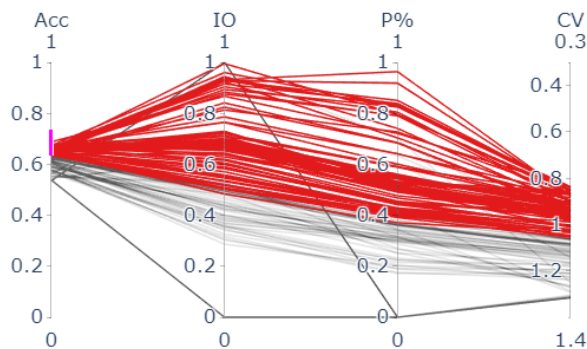


(a) Seleção dos modelos

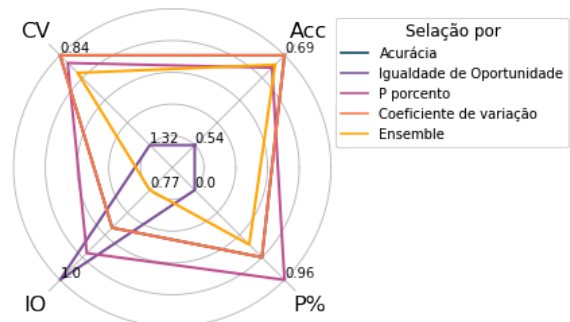


(b) Resultados obtidos

Figura 8: Resultados obtidos em teste para o modelo gerado a partir do método de *ensemble learning* utilizando todos o conjunto de modelos encontrado pela estratégia multi-objetiva de erros para o conjunto de dados COMPAS.



(a) Seleção dos modelos



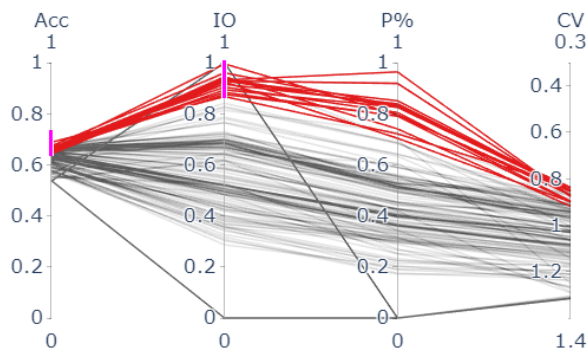
(b) Resultados obtidos

Figura 9: Resultados obtidos em teste para o modelo gerado a partir do método de *ensemble learning* utilizando o subconjunto de modelos encontrado pela estratégia multi-objetiva de erros que possuem maiores valores na métrica de acurácia para o conjunto de dados COMPAS.

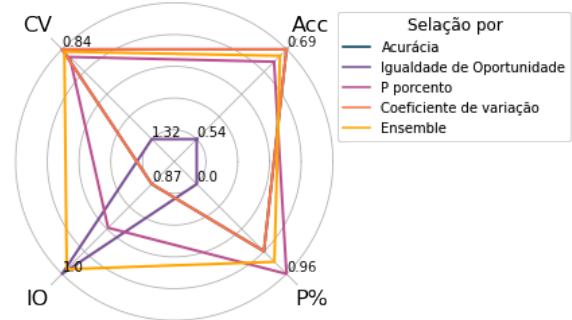
tendo seu bom desempenho em acurácia e nas demais métricas de discriminação.

Em seguida, realizamos o mesmo experimento para o conjunto de modelos encontrados pela estratégia multi-objetiva por probabilidades:

Quando não adicionamos nenhuma restrições aos modelos utilizados para geração do *ensemble*, como apresentado na Figura 11, temos que tal modelo possui os melhores valores

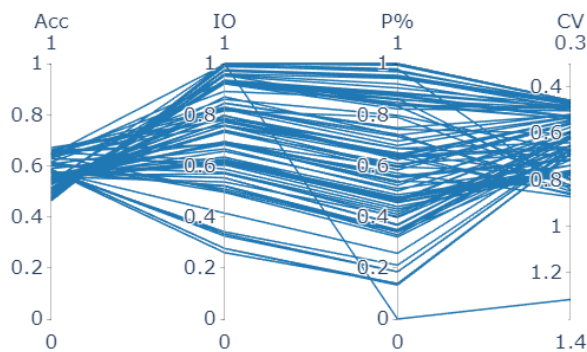


(a) Seleção dos modelos

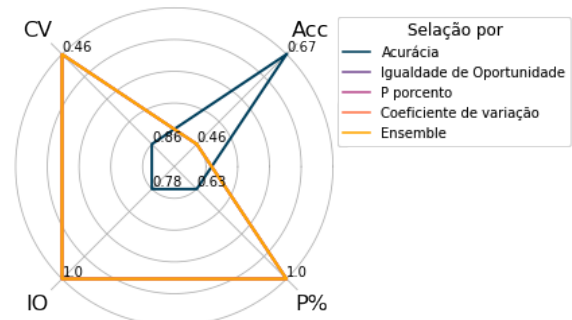


(b) Resultados obtidos

Figura 10: Resultados obtidos em teste para o modelo gerado a partir do método de *ensemble learning* utilizando o subconjunto de modelos encontrado pela estratégia multi-objetiva de erros que possuem maiores valores na métrica de acurácia e igualdade de oportunidade para o conjunto de dados COMPAS.



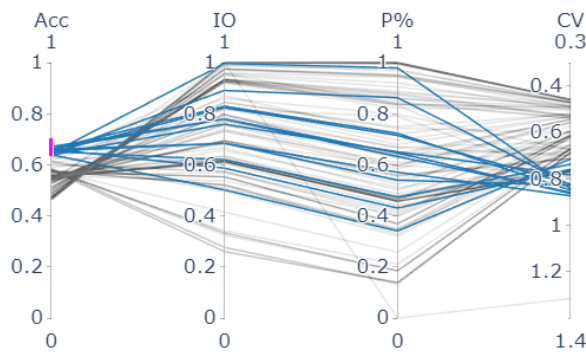
(a) Seleção dos modelos



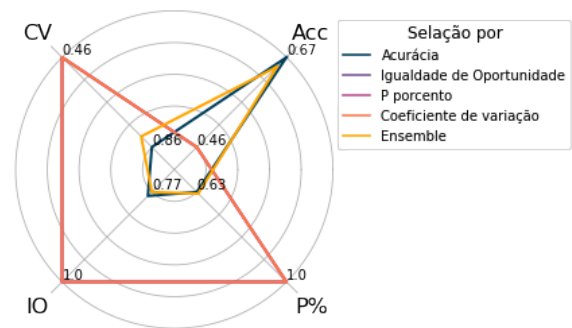
(b) Resultados obtidos

Figura 11: Resultados obtidos em teste para o modelo gerado a partir do método de *ensemble learning* utilizando o subconjunto de modelos encontrado pela estratégia multi-objetiva de probabilidades que possuem maiores valores na métrica de acurácia e igualdade de oportunidade para o conjunto de dados COMPAS.

para as métricas de discriminação, tendo os mesmos valores obtidos pelos modelos que otimizam individualmente as métricas de Igualdade de Oportunidade, P por cento e Coeficiente de variação, mas sua acurácia é significativamente inferior à obtida pelo modelo que otimiza individualmente tal métrica. Por conta disso, selecionamos os modelos que possuem maiores valores para acurácia e avaliamos o modelo de *ensemble* obtido.



(a) Seleção dos modelos



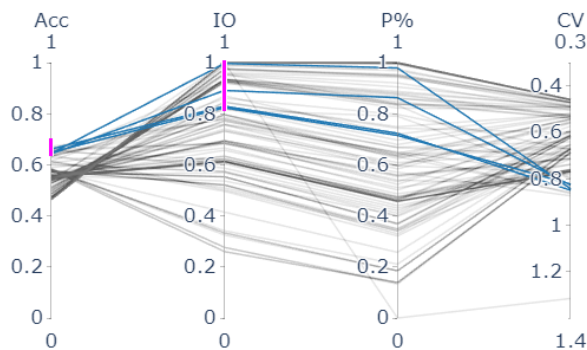
(b) Resultados obtidos

Figura 12: Resultados obtidos em teste para o modelo gerado a partir do método de *ensemble learning* utilizando o subconjunto de modelos encontrado pela estratégia multi-objetiva de probabilidades que possuem maiores valores na métrica de acurácia para o conjunto de dados COMPAS.

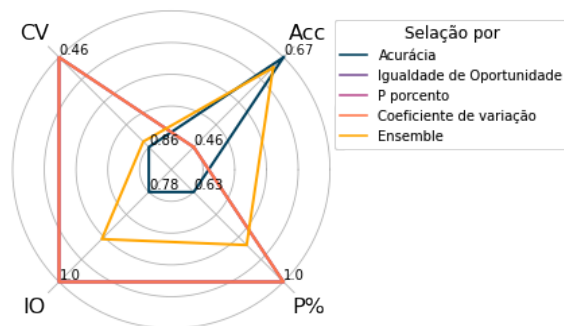
Na Figura 12 selecionamos os modelos pela acurácia. Com isso, temos que o modelo de *ensemble* aumenta seu valor de acurácia, se aproximando bastante da acurácia obtida pelo modelo que otimiza individualmente tal métrica. Entretanto, os valores obtidos para as métricas de discriminação obtidos diminuem bastante. Observando os modelos selecionados podemos perceber que não há muito o que fazer para melhorar o coeficiente de variação, pois todos os modelos da seleção possuem valores altos para tal métrica. Entretanto, é possível melhorarmos as métricas de Igualdade de Oportunidade e P por cento realizando uma segunda seleção.

Por fim, selecionamos os melhores modelos em Igualdade de Oportunidade na Figura 13. Com isso o modelo de *ensemble* resulta em valores melhores para as métricas de Igualdade de Oportunidade e P por cento, mantendo o valor de acurácia obtido na seleção anterior. Tal modelo continua com valores para as métricas de discriminação inferiores aos obtidos pelos modelos que focam em tais métricas, mas ele possui uma melhora significativa neles em comparação com o modelo que foca em acurácia, mantendo a acurácia bem parecida com este.

Com este experimento percebemos que a utilização de métodos de *ensemble learning* para gerar um único modelo a partir do conjunto de modelos resultante das estratégias multi-objetivo propostas neste trabalho consegue resultar um modelo que é mais balanceado em relação ao conflito entre discriminação e desempenho quando comparado com modelos obtidos da otimização individual de tais métricas.



(a) Seleção dos modelos



(b) Resultados obtidos com a seleção (a)

Figura 13: Resultados obtidos em teste para o modelo gerado a partir do método de *ensemble learning* utilizando o subconjunto de modelos encontrado pela estratégia multi-objetiva de probabilidades que possuem maiores valores na métrica de acurácia e igualdade de oportunidade para o conjunto de dados COMPAS.

4.5 Conclusão dos experimentos

Com os experimentos realizados conseguimos evidências de que as estratégias propostas neste trabalho para o treinamento de modelos de aprendizagem de máquinas utilizando otimização multi-objetivo para otimizar simultaneamente desempenho e discriminação possuem potencial para serem utilizadas em casos de uso reais, tendo desempenho similar e, em alguns casos, melhor do que as estratégias similares bastante utilizadas da literatura de *fairness*. Evidenciamos, em 4.2, que as propostas conseguem otimizar individualmente as métricas de desempenho e discriminação de forma similar a estratégias que possuem tal foco. Vale observar que por vezes as abordagens propostas reduziam o viés discriminatório custando a performance do modelo. Entretanto, metodologias com a mesma família eram capazes de terem bom desempenho, levantando a hipótese que outras formas de exploração desses modelos poderiam gerar um classificador mais robusto. Com isso, mostramos em 4.3 que as nossas estratégias geram um conjunto bastante diverso de modelos, apresentando diversidade para toda métrica analisada, o que nos permitiu avaliar, em 4.4 que agregação de modelos através *ensemble learning* geravam modelo mais robustos e capazes de gerar bons resultados para todas as métricas.

A seguir apresentamos as conclusões finais deste trabalho, analisando as limitações das estratégias propostas e possibilidades de trabalhos futuros capazes de lidar com tais limitações.

5 Conclusões e trabalhos futuros

Neste projeto remodelamos a regressão logística enquanto um problema multi-objetivo a ser resolvido através do método de soma ponderada MONISE, apresentando duas propostas de objetivos a serem utilizados para realizar o treinamento dos modelos de aprendizagem de máquina. Na primeira formulação, utilizamos o erro (ou perda) para cada grupo com base na característica sensível (grupo sensível e grupo não sensível), juntamente com o termo de regularização do parâmetro da regressão logística, como conjunto de objetivos do treinamento. Apresentamos também uma segunda proposta de objetivos, utilizando a probabilidade de aceitação de cada grupo, além do erro geral e o termo de regularização.

Com as estratégias propostas, encontramos um conjunto de modelos que representa as melhores trocas entre as métricas de discriminação e desempenho que são conflitantes, sendo este conjunto a Fronteira de Pareto do problema de otimização. A partir desse conjunto utilizamos cinco estratégias para extrair um único modelo, sendo quatro delas otimizando individualmente cada uma das quatro métricas utilizadas e a quinta estratégia é a utilização do método de *ensemble learning* para gerar um único modelo a partir do conjunto de modelos resultantes, de forma que esse novo modelo seja mais robusto e balanceado que os anteriores.

Comparamos as estratégias propostas neste trabalho com diversas outras estratégias bastante utilizadas na literatura de *fairness*, em quatro conjunto de dados diferentes, e com isso obtivemos evidências de que nossas propostas possuem resultados similares e, em alguns casos, superior às estratégias comparadas, mesmo ao otimizar cada métrica individualmente ao invés de simultaneamente, o que não é o propósito de nossas estratégias. Além disso, percebemos que as propostas conseguem obter modelos com acurácia superior às outras estratégias enquanto mantendo valores de discriminação similares aos obtidos pelos modelos focados apenas neles. Entretanto, este trabalho possui algumas limitações.

A primeira limitação que percebemos no projeto é que ele foi modelado apenas a partir da regressão logística, conseguindo ser utilizado apenas em aplicações que realizam classificações binárias, o que é bastante restrito em relação às diversas aplicações de modelos de aprendizagem de máquina. Tal limitação pode ser abordada em trabalhos futuros utilizando da mesma metodologia proposta neste, mas com foco em outros tipos de aplicações através da remodelagem de outros algoritmos bastante utilizados na área de inteligência artificial, como regressões lineares e algoritmos de classificação em mais de dois grupos. Para isso, entretanto, se torna necessária a utilização de métricas de discriminação diferentes das utilizadas neste trabalho, por elas serem definidas apenas para problemas de classificação binária.

Outra limitação percebida é que os resultados foram testados apenas em conjuntos de

dados com uma única característica sensível que é binária, como se o indivíduo é homem ou não, ou se é branco ou não branco. Entretanto, diversas características sensíveis não são binárias, como a própria raça, além de que podemos lidar com mais de uma característica sensível, lidando com diversos grupos com base em tais características. As estratégias propostas neste trabalho não são limitadas a características sensíveis binárias, mas as métricas de discriminação utilizadas para avaliar os modelos, assim como boa parte dos modelos comparados, o são. Assim, tornam-se possíveis trabalhos futuros que lidam com diversos grupos sensíveis simultaneamente utilizando as mesmas estratégias propostas aqui.

Por fim, em trabalhos futuros podemos abordar a acessibilidade de utilização das estratégias propostas, implementando uma biblioteca para facilitar o uso das ferramentas desenvolvidas durante este trabalho, o que inclui a implementação das estratégias propostas no trabalho e a otimização das estratégias comparadas através de seus hiper parâmetros. Tal implementação é bastante importante para permitir a aplicação do trabalho desenvolvido em problemas reais de aprendizagem de máquinas em cenários propícios a discriminação.

Referências

- [1] HOWARD, A. e BORENSTEIN, J. *The ugly truth about ourselves and our robot creations: the problem of bias and social inequity*. Science and engineering ethics, 2018.
- [2] OSOBA, O. A. e WELSER IV, W. *An intelligence in our image: The risks of bias and errors in artificial intelligence*. Rand Corporation, 2017.
- [3] DUA, D. e GRAFF, C. *UCI Machine Learning Repository*. Disponível em: <http://archive.ics.uci.edu/ml>. Acesso em: 07/09/2020.
- [4] VERMA, S. e RUBIN, J. *Fairness Definitions Explained*. ACM/IEEE International Workshop on Software Fairness, 1:7, 2018.
- [5] ZHONG, Z. *A Tutorial on Fairness in Machine Learning*. towardsdatascience, 2018. Disponível em: <https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb>. Acesso em: 07/09/2020.
- [6] ZEMEL, R.; WU, Y.; SWERSKY, K.; PITASSI, T.; DWORK, C. *Learning Fair Representations*. CoRR, abs:1507.05259, 2015.
- [7] KAMIRAN, F; CALDERS, T; PECHENIZKIY, M. *Building Classifiers with Independence Constraints*. IEEE International Conference on Data Mining Workshops, pages 13–18, 2009.
- [8] KAMIRAN, F. e CALDERS, T. *Data preprocessing techniques for classification without discrimination*. Knowledge and Information Systems, 33(1):1–33, 2012.
- [9] ZAFAR, M.; VALERA, I.; RODRIGUEZ, M.; GUMMADI, K. *Fairness Constraints: Mechanisms for Fair Classification*. AISTATS, 2017.
- [10] AGARWAL, A.; BEYGELZIMER, A.; DUDÍK, M.; LANGFORD, J; WALLACH, H. *A reductions approach to fair classification*. Fairness, Accountability, and Transparency in Machine Learning (FATML), 2017.
- [11] HARDT M.; PRICE, E.; SREBRO, N. *Equality of opportunity in supervised learning*. Advances in Neural Information Processing Systems, 2016.
- [12] WADSWORTH, C.; VERA, F.; PIECH, C.. *Achieving Fairness through Adversarial Learning: an Application to Recidivism Prediction*. Fairness, Accountability, and Transparency in Machine Learning (FATML), 2018.

- [13] ZHAO, H and GORDON, G. *Inherent Tradeoffs in Learning Fair Representations*. NeurIPS, 2019.
- [14] ZLIOBAITE, I. *On the relation between accuracy and fairness in binary classification*. CoRR, abs/1505.05723, 2015.
- [15] MARLER, R. e ARORA, J. *Survey of multi-objective optimization methods for engineering*. Structural and multidisciplinary optimization 26 (6), 369–395, 2004.
- [16] RAIMUNDO, M. M.; FERREIRA, P. A. V. e VON ZUBEN, F. J. *An extension of the non-inferior set estimation algorithm for many objectives*. European Journal of Operational Research, Volume 284, Issue 1, 1 July 2020, Pages 53-66.
- [17] COHON, J. *Multiobjective programming and planning*. Mathematics in Science and Engineering, vol. 140. Elsevier, 1978.
- [18] GAJANE, P. e PECHENIZKIY, M. *On Formalizing Fairness in Prediction with Machine Learning*. arXiv, vol. 1710.03184, pp. 1-6, 218.
- [19] ZAFAR, M.; VALERA, I.; RODRIGUEZ, M.; GUMMADI, K.; WELLER, A. *From Parity to Preference-based Notions of Fairness in Classification*. NIPS, 2017.
- [20] MARTINEZ, N.; BERTRAN, M.; SSAPIRO, G. *Minimax Pareto Fairness: A Multi Objective Perspective*. International Conference on Machine Learning, 2020.
- [21] SPEICHER, T.; HEIDARI, H.; GRGIC-HLACA, N.; GUMMADI, K.; SINGLA, A.; WELLER, A.; ZAFAR, M. *A Unified Approach to Quantifying Algorithmic Unfairness: Measuring Individual & Group Unfairness via Inequality Indices*. Association for Computing Machinery, 2018.
- [22] ABDI, H. *Coefficient of variation*. Encyclopedia of research design, 2010.
- [23] USTUN, B.; SPANGHER, A.; LIU, Y. *Actionable Recourse in Linear Classification*. Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19). Association for Computing Machinery, New York, NY, USA, 10–19.
- [24] RAIMUNDO, M. M. e VON ZUBEN, F. J. *Investigating multiobjective methods in multitask classification*. International Joint Conference on Neural Networks (IJCNN), Rio de Janeiro, 2018, pp. 1-9.

- [25] CRUZ, R.; HAFEMANN, L.; SABOURIN, R.; CAVALCANTI, G. *DESlib: A Dynamic ensemble selection library in Python*. Journal of Machine Learning Research (8):1-5, 2020.
- [26] ANGWIN, J.; LARSON, J.; MATTU, S.; KIRCHNER, L. *Machine Bias*. Disponível em: <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>. Acesso em: 02/11/2020.
- [27] WIGHTMAN, L. *LSAC national longitudinal bar passage study*. Technical report, LSAC Research Report Series, 1998.
- [28] MAUGHAN, K. e NEAR, J. P. *Towards a Measure of Individual Fairness for Deep Learning*. arXiv:2009.13650, 2020.
- [29] BECHAVOD, Y. e LIGETT, K. *Penalizing Unfairness in Binary Classification*. arXiv:1707.00044, 2017.
- [30] BICKE, P.; HAMMEL, E; O'CONNELL, J. *Sex Bias in Graduate Admissions: Data from Berkeley*. Science, vol. 187 no. 4175 398-404, 1975.
- [31] CHEN, J.; KALLUS, N.; MAO, X.; SVACHA, G.; UDELL, M. *Fairness Under Unawareness: Assessing Disparity When Protected Class Is Unobserved*. Proceedings of the Conference on Fairness, Accountability, and Transparency. ACM, 339–348, 2019.
- [32] MEHRABI, N.; MORSTATTER, F.; SAXENA, N.; LERMAN, K.; GALSTYAN, A. A *Survey on Bias and Fairness in Machine Learning*. arXiv:1908.09635, 2019.
- [33] MEHRABI, N.; GOWDA, T.; MORSTATTER, F.; PENG, N.; GALSTYAN, A. *Man is to Person as Woman is to Location: Measuring Gender Bias in Named Entity Recognition*. arXiv:1910.10872, 2019.
- [34] CRUZ, A. *Fairness-Aware Hyperparameter Optimization*. .
- [35] CAVES, R. *Encyclopedia of the city*. Routledge, 2005.
- [36] WANNG, Q.; XU, Z.; CHEN, Z.; WANG, Y.; LIU, S.; QU, H. *Visual Analysis of Discrimination in Machine Learning*. Routledge, 2005.
- [37] Gareth James; Daniela Witten; Trevor Hastie; Robert Tibshirani. *An Introduction to Statistical Learning*. Springer, .

- [38] ZHANG, C. e MA, Y. *Ensemble Machine Learning: Methods and Applications*. Springer, 2012.
- [39] NEVES, E. **Modelos de Predição — Ensemble Learning**. Disponível em: <https://medium.com/turing-talks/turing-talks-24-modelos-de-predi%C3%A7%C3%A3o-ensemble-learning-aa02ce01afda>. Acesso em: 26/11/2020.

Apêndice

Apêndice A - Resultados do primeiro experimento

German

Método	Métrica Otimizada	Acurácia	Coefficiente de Variação	Igualdade de Oportunidade	Paridade Demográfica
Multi-objetivo - Probabilidade	ACC	0.60	0.72	0.68	0.59
	CV	0.57	0.48	1.00	1.00
	IO	0.57	0.48	1.00	1.00
	P%	0.57	0.48	1.00	1.00
Multi-objetivo - Erro	ACC	0.59	0.84	0.94	0.88
	CV	0.57	0.75	0.89	0.96
	IO	0.56	0.90	0.83	0.83
	P%	0.59	0.85	0.90	0.85
Minimax	ACC	0.59	0.86	0.83	0.70
	CV	0.56	0.54	0.97	0.99
	IO	0.58	0.55	0.99	0.95
	P%	0.56	0.56	0.99	0.99
Classificador de Igualdade de Oportunidade	ACC	0.60	0.86	0.89	0.77
	CV	0.57	0.48	1.00	1.00
	IO	0.57	0.49	0.99	0.99
	P%	0.57	0.56	0.99	0.96
Classificador de Paridade Demográfica	ACC	0.61	0.85	0.93	0.80
	CV	0.56	0.58	0.96	0.99
	IO	0.59	0.65	0.97	0.92
	P%	0.58	0.67	0.95	0.91
<i>Reweighting</i>	ACC	0.59	0.74	0.93	0.87
	CV	0.57	0.54	0.95	0.99
	IO	0.58	0.80	0.88	0.88
	P%	0.58	0.75	0.91	0.93
Regressão Logística	ACC	0.61	0.84	0.95	0.80
	CV	0.57	0.54	0.96	0.99
	IO	0.59	0.79	0.93	0.81
	P%	0.59	0.72	0.94	0.91

Tabela 2: Comparação da performance dos métodos para o conjunto de dados *German* quando o modelo é escolhido com base na métrica otimizada.



Figura 14: Visualização da comparação da performance dos métodos para o conjunto de dados *German* quando o modelo é escolhido com base na métrica otimizada.

LSAC

Método	Métrica Otimizada	Acurácia	Coefficiente de Variação	Igualdade de Oportunidade	Paridade Demográfica
Multi-objetivo - Probabilidade	ACC	0.95	0.28	0.99	0.99
	CV	0.95	0.28	1.00	1.00
	IO	0.95	0.28	1.00	1.00
	P%	0.95	0.28	1.00	0.99
Multi-objetivo - Erro	ACC	0.95	0.28	0.99	0.99
	CV	0.95	0.28	1.00	0.99
	IO	0.95	0.28	1.00	0.99
	P%	0.95	0.28	0.99	0.99
Minimax	ACC	0.95	0.28	0.99	0.99
	CV	0.95	0.28	0.99	0.99
	IO	0.95	0.28	1.00	1.00
	P%	0.95	0.28	1.00	1.00
Classificador de Igualdade de Oportunidade	ACC	0.95	0.28	0.99	0.99
	CV	0.95	0.28	0.99	0.99
	IO	0.95	0.28	1.00	1.00
	P%	0.95	0.28	1.00	1.00
Classificador de Paridade Demográfica	ACC	0.95	0.28	0.99	0.99
	CV	0.95	0.28	0.99	0.99
	IO	0.95	0.28	1.00	1.00
	P%	0.95	0.28	1.00	1.00
<i>Reweighting</i>	ACC	0.95	0.28	1.00	0.99
	CV	0.95	0.28	1.00	0.99
	IO	0.95	0.28	1.00	0.99
	P%	0.95	0.28	1.00	0.99
Regressão Logística	ACC	0.95	0.28	0.99	0.99
	CV	0.95	0.28	0.99	0.99
	IO	0.95	0.28	1.00	1.00
	P%	0.95	0.28	1.00	1.00

Tabela 3: Comparação da performance dos métodos para o conjunto de dados *LSAC* quando o modelo é escolhido com base na métrica otimizada.

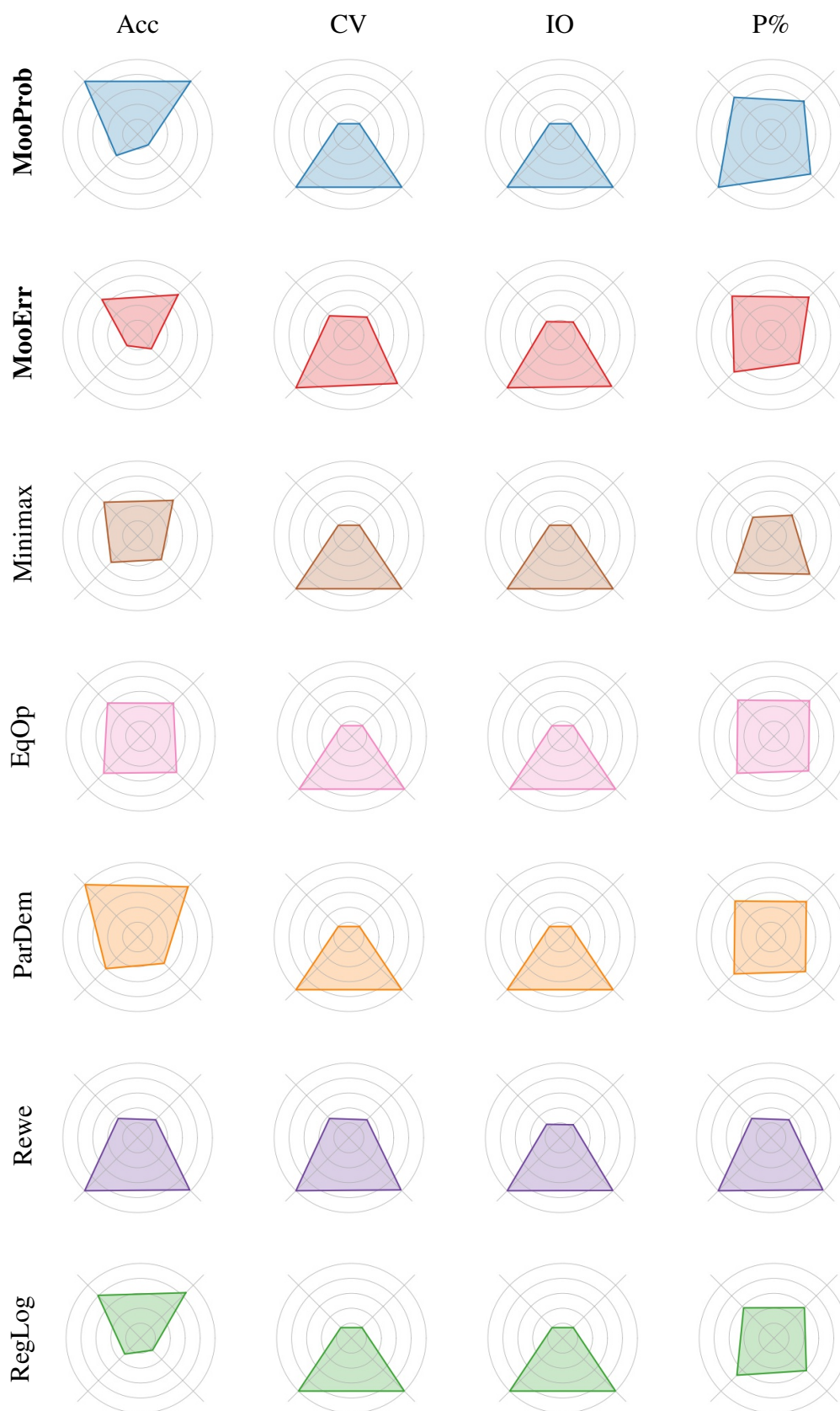


Figura 15: Visualização da comparação da performance dos métodos para o conjunto de dados LSAC quando o modelo é escolhido com base na métrica otimizada.

Adult

Método	Métrica Otimizada	Acurácia	Coefficiente de Variação	Igualdade de Oportunidade	Paridade Demográfica
Multi-objetivo - Probabilidade	ACC	0.79	0.70	0.85	0.56
	CV	0.24	0.34	1.00	1.00
	IO	0.24	0.34	1.00	1.00
	P%	0.24	0.34	1.00	1.00
Multi-objetivo - Erro	ACC	0.79	0.71	0.92	0.60
	CV	0.79	0.71	0.88	0.57
	IO	0.79	0.71	0.82	0.54
	P%	0.79	0.70	0.93	0.57
Minimax	ACC	0.79	0.69	0.74	0.36
	CV	0.79	0.69	0.73	0.37
	IO	0.79	0.70	0.82	0.48
	P%	0.79	0.70	0.87	0.58
Classificador de Igualdade de Oportunidade	ACC	0.83	0.60	0.68	0.41
	CV	0.83	0.60	0.68	0.40
	IO	0.82	0.65	0.66	0.41
	P%	0.81	0.66	0.79	0.51
Classificador de Paridade Demográfica	ACC	0.83	0.60	0.69	0.41
	CV	0.83	0.60	0.68	0.40
	IO	0.82	0.64	0.76	0.53
	P%	0.79	0.71	0.93	0.84
<i>Reweighting</i>	ACC	0.79	0.70	0.88	0.57
	CV	0.79	0.70	0.86	0.54
	IO	0.79	0.71	0.88	0.58
	P%	0.79	0.70	0.92	0.59
Regressão Logística	ACC	0.79	0.70	0.87	0.58
	CV	0.79	0.70	0.87	0.57
	IO	0.79	0.71	0.85	0.56
	P%	0.79	0.70	0.89	0.56

Tabela 4: Comparação da performance dos métodos para o conjunto de dados *Adult* quando o modelo é escolhido com base na métrica otimizada.

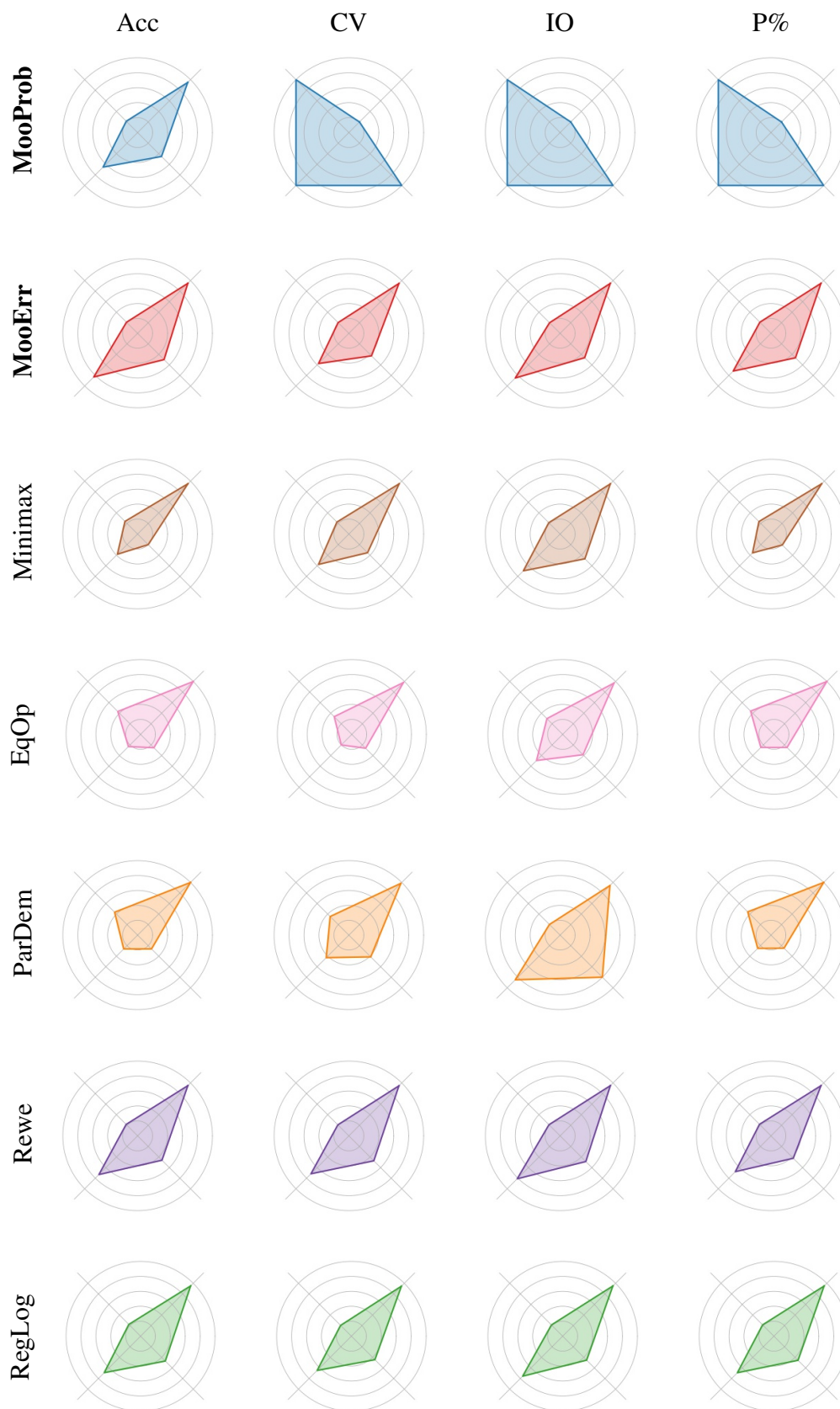


Figura 16: Visualização da comparação da performance dos métodos para o conjunto de dados *Adult* quando o modelo é escolhido com base na métrica otimizada.

COMPAS

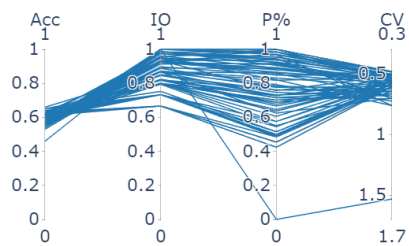
Método	Métrica Otimizada	Acurácia	Coefficiente de Variação	Igualdade de Oportunidade	Paridade Demográfica
Multi-objetivo - Probabilidade	ACC	0.66	0.85	0.70	0.63
	CV	0.44	0.45	1.00	1.00
	IO	0.44	0.45	1.00	1.00
	P%	0.44	0.45	1.00	1.00
Multi-objetivo - Erro	ACC	0.66	0.88	0.53	0.43
	CV	0.66	0.83	0.87	0.79
	IO	0.64	0.92	0.86	0.72
	P%	0.64	0.92	0.86	0.72
Minimax	ACC	0.67	0.88	0.50	0.39
	CV	0.66	0.87	0.53	0.47
	IO	0.62	1.00	0.67	0.65
	P%	0.59	1.14	0.78	0.62
Classificador de Igualdade de Oportunidade	ACC	0.66	0.88	0.58	0.50
	CV	0.66	0.87	0.59	0.51
	IO	0.66	0.92	0.64	0.51
	P%	0.66	0.91	0.63	0.51
Classificador de Paridade Demográfica	ACC	0.66	0.88	0.58	0.50
	CV	0.53	0.87	0.58	0.51
	IO	0.56	1.15	0.79	0.74
	P%	0.66	1.15	0.89	0.85
<i>Reweighting</i>	ACC	0.66	0.89	0.85	0.78
	CV	0.65	0.89	0.85	0.78
	IO	0.64	0.98	0.91	0.90
	P%	0.66	0.91	0.89	0.83
Regressão Logística	ACC	0.67	0.88	0.50	0.39
	CV	0.63	0.87	0.55	0.47
	IO	0.65	1.01	0.68	0.37
	P%	0.66	0.92	0.61	0.49

Tabela 5: Comparação da performance dos métodos para o conjunto de dados *COMPAS* quando o modelo é escolhido com base na métrica otimizada.

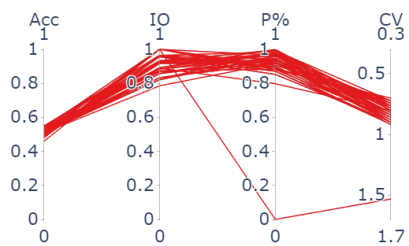


Figura 17: Visualização da comparação da performance dos métodos para o conjunto de dados COMPAS quando o modelo é escolhido com base na métrica otimizada.

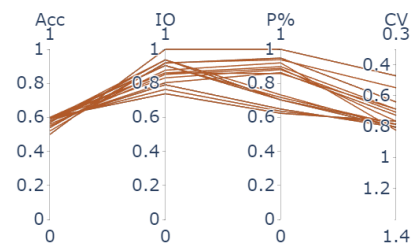
Apêndice B - Diversidade dos modelos



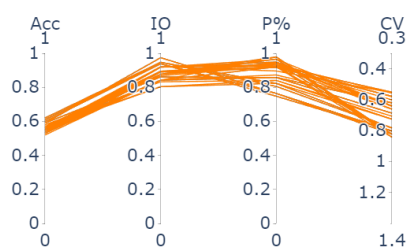
(a) Multi-objetivo - Probabilidade



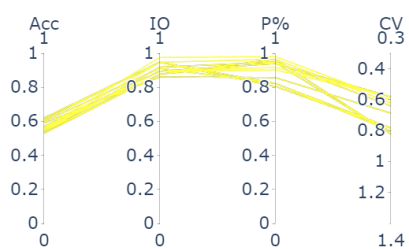
(b) Multi-objetivo - Erro



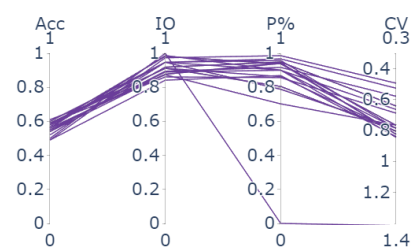
(c) Minimax



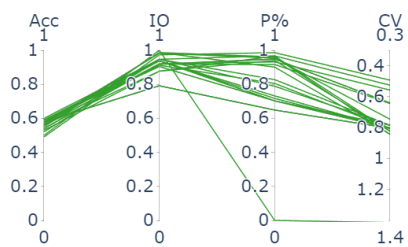
(d) Class. de Paridade Demográfica



(e) Class. de Igualdade de Oportunidade

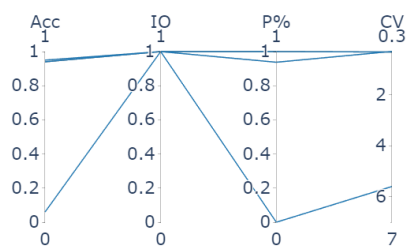


(f) Reweighting

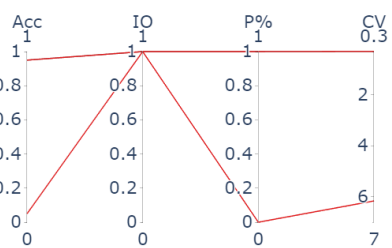


(g) Regressão Logística

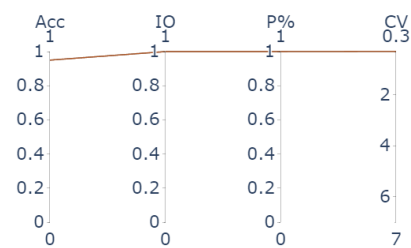
Figura 18: Valores encontrados para métricas de desempenho e discriminação para os modelos resultantes de cada estratégia utilizando o conjunto de dados *German*.



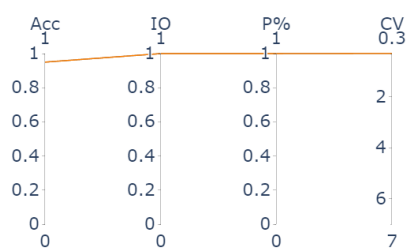
(a) Multi-objetivo - Probabilidade



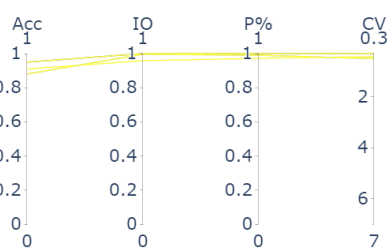
(b) Multi-objetivo - Erro



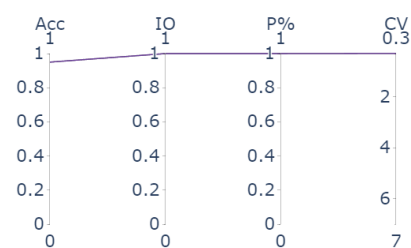
(c) Minimax



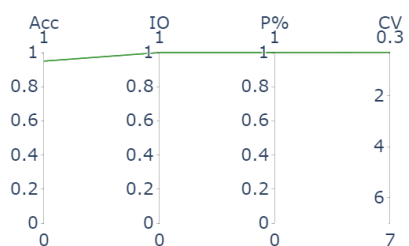
(d) Class. de Paridade Demográfica



(e) Class. de Igualdade de Oportunidade

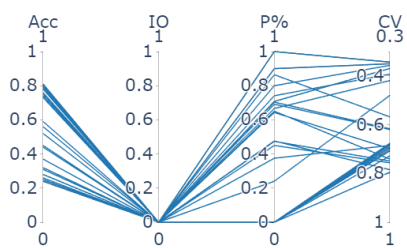


(f) Reweighting

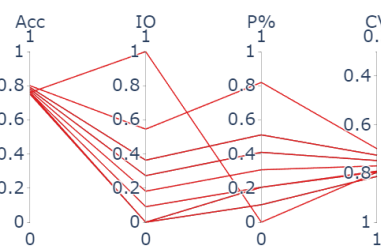


(g) Regressão Logística

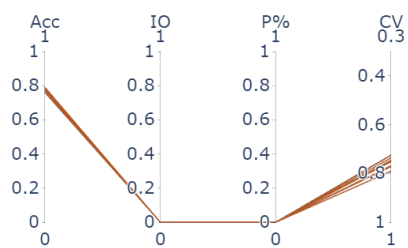
Figura 19: Valores encontrados para métricas de desempenho e discriminação para os modelos resultantes de cada estratégia utilizando o conjunto de dados LSAC.



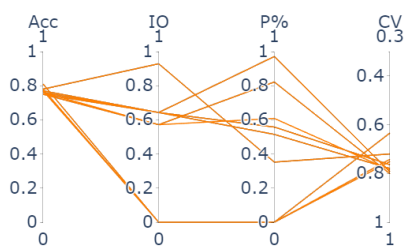
(a) Multi-objetivo - Probabilidade



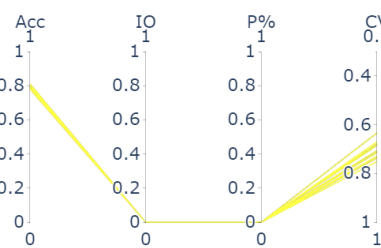
(b) Multi-objetivo - Erro



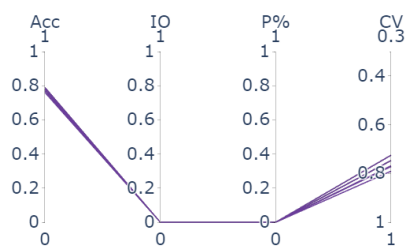
(c) Minimax



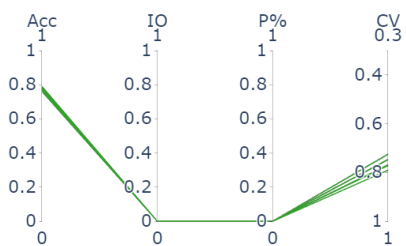
(d) Class. de Paridade Demográfica



(e) Class. de Igualdade de Oportunidade



(f) Reweighting



(g) Regressão Logística

Figura 20: Valores encontrados para métricas de desempenho e discriminação para os modelos resultantes de cada estratégia utilizando o conjunto de dados *Adult*.