Aplicação de Otimização Multi-Objetivo em Problemas de Aprendizagem de Máquinas para Diminuição de Discriminação

Vitória Aquino Guardieiro
Orientador: Jorge Poco
Coorientador: Marcos Raimundo

FGV/EMAp - Escola de Matemática Aplicada

07/12/2020

Introdução

- A utilização de modelos de aprendizagem de máquinas para tomar decisões está se popularizando e expandindo nas políticas públicas, contratação de funcionários, análises de crédito, entre outras aplicações que tem um impacto significativo na vida das pessoas e da sociedade como um todo.
- Entretanto, tais modelos podem propagar e perpetuar discriminações, caso sejam treinados com dados implicitamente discriminatórios ou caso ser discriminatório resulte em erros menores para os modelos.
- A partir da importância ética de não se discriminar, assim como das iminentes regulamentações em relações ao tópico, surgiu a subárea de discriminação ou *fairness* nas pesquisas que envolvem inteligências artificiais.
- Inicialmente, acreditava-se que treinar um modelo sem fornecer a ele informações potencialmente discriminatórias (chamadas de **sensíveis**), como raça e gênero, era suficiente para garantir que seu resultado não seria discriminatório. Mas isso já se provou falso, por conta existirem características bastante correlacionadas com as sensíveis e que são necessárias para a aplicação, como endereço e raça.
- Assim se tornou necessário o desenvolvimento de novas estratégias de treinamento para os modelos, que considerasse a discriminação e não apenas o desempenho resultante.

Motivação

- Quando restringimos ou otimizamos a discriminação de um modelo, isso geralmente resulta em uma perda em seu desempenho, o que já foi apresentado em diversos estudos.
- Com tal conflito entre as métricas de discriminação e desempenho, assim como o desejo de otimizar ambas simultaneamente, temos um cenário bastante similar ao que os métodos de otimização multi-objetivo buscam resolver.
- Alguns trabalhos em *fairness* aplicaram tais métodos de otimização no treinamento de modelos de aprendizagem de máquinas, mas eles se limitam a encontrar um único modelo que é determinado como o melhor, não deixando o utilizador do modelo escolher explicitamente o melhor modelo dado o conflito entre desempenho e discriminação.
- Este trabalho busca utilizar da otimização multi-objetivo para explicitar o conflito entre discriminação e desempenho, permitindo que o utilizador escolha qual o modelo dentre eles consideração sua aplicação.

Objetivos

- 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.
- Além disso, outro objetivo é comparar os resultados obtidos pelas propostas deste trabalho com estratégias similares da literatura de *fairness*.
- E, por fim, analisar a utilização 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.

Metodologia

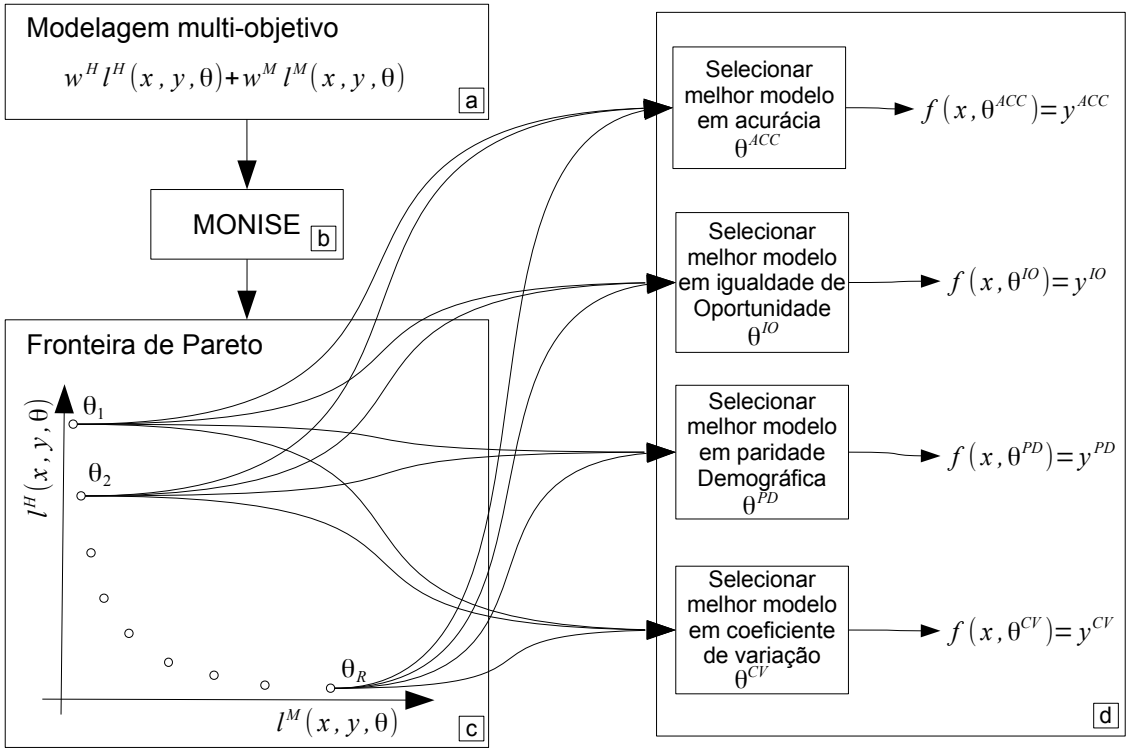


Figura 1: Descrição visual da metologia proposta neste trabalho. Em (a) temos a modelagem por soma ponderada dos objetivos que minimizaremos, em (b) utilizamos a abordagem MONISE para resolver o problema multiobjetivo, resultando na Fronteira de Pareto apresentada em (c). Em (d), selecionamos dentre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

Otimização Multi-objetivo

- A otimização multi-objetivo consiste em minimizar ou maximizar m funções $f_1, f_2, \ldots, f_m$ simultaneamente, com $f_i(\vec{x}) : \Omega \to \mathbb{R}$ e $\vec{x} \in \Omega, \Omega \subset \mathbb{R}^d$, sendo Ω o subespaço de valores possíveis para x.
- Devido à existência de diversos objetivos, a otimização multi-objetivo não busca encontrar um único $\vec{x}$ que otimiza todos os objetivos simultaneamente, mas sim um conjunto ótimo, chamado de **Fronteira de Pareto**, tal que os elementos desse conjunto otimizem o conflito/*tradeoff* entre os objetivos.
- No trabalho, apliquei a abordagem *MONISE - Many-Objective Non-Inferior Set Estimation* para resolver os problemas de otimização propostos. Ela utiliza do método de soma ponderada, que consiste em otimizar diversas vezes $\vec{w}^T f(\vec{x})$, determinando vetores w de forma a obter um conjunto representativo da Fronteira de Pareto.

Regressão Logística

- Um problema de classificação binária é dado por um conjunto de N amostras, onde $\vec{x_i} \in \mathbb{R}^d : i \in \{1, \ldots, N\}$, e $\vec{X} = \{\vec{x_1}, \ldots, \vec{x_N}\}$ consiste nos atributos de entrada e $y_i \in \{0, 1\} : i \in \{1, \ldots, N\}$ é o valor de saída que se deseja predizer. 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 $f(x, \theta) = \frac{e^{\theta^\top \phi(\vec{x})}}{1+e^{\theta^\top \phi(\vec{x})}} \in [0, 1]$, que descreve a probabilidade de uma nova amostra $\vec{x}$ 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 sigmóide melhor se adequar aos dados:

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

Primeira proposta: Erro por grupo

Na primeira proposta do trabalho, é utilizada a função de perda da regressão logística, mas o valor de perda é calculado para cada grupo com base na característica sensível $g \in \{1, \ldots, G\}$:

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

Com isso, o treinamento do modelo consiste em otimizar a seguinte soma ponderada:

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

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

Segunda proposta: Probabilidade por grupo

Na segunda proposta, ao invés de utilizarmos a perda por grupo, utilizamos da probabilidade de aceitação por grupo, dada por:

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

Assim, é otimizada a seguinte soma ponderada:

$$\min_{\theta} \quad \sum_{g=1}^{G} \vec{w}_g a^g(\vec{X}, \theta) + \vec{w}_{G+1} l(\vec{X}, \vec{y}, \theta) + \vec{w}_{G+2} \|\theta\|_2 \tag{5}$$

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

Métricas de *Fairness*

Para compararmos e avaliarmos os métodos propostos, utilizaremos as seguintes métricas de discriminação/*fairness*, que representam noções distintas sobre o que é discriminação:

	Noção de Justiça	Condição de Justiça
Justiça por 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	Coeficiente de Variação	Indivíduos que merecem resultados similares recebem resultados similares

Tabela 1: Visão geral das métricas de desigualdade.

Ensemble Learning

- Os métodos de *ensemble learning*, ou aprendizado por agrupamento, consistem 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.
- No trabalho, utilizei do método de *ensemble learning* de votação simples suave, onde 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 $f(x, \theta_i)$. Com o método suave (*soft*), 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.

Experimentos

Foram realizados três experimentos com as estratégias propostas no trabalho. Para eles, foram utilizados os seguintes conjuntos de dados:

German O conjunto de dados sobre solicitações de crédito, cujo objetivo é classificar se uma solicitação deve ser aceita ou não. A característica sensível é o gênero.

Adult Contém informação sobre indivíduos do Censo dos Estados Unidos de 1994. A tarefa proposta é prever se um certo indivíduo recebe mais ou menos que $50.000 por ano, sendo raça a característica sensível.

LSAC Contém informações de exames e raça (utilizado como característica sensível). O objetivo é identificar se o aluno passou no exame de ordem ou não.

ProPublica Inclui informações sobre indivíduos que foram presos, incluindo o grau do incidente e raça (utilizado como característica sensível). O objetivo é prever se o indivíduo voltará a ser preso em dois anos.

Experimentos

Nos experimentos, os resultados obtidos pelas estratégias propostas foram comparadas com outras estratégias:

Regressão Logística Como definida na métodologia, será utilizada como base de comparação.

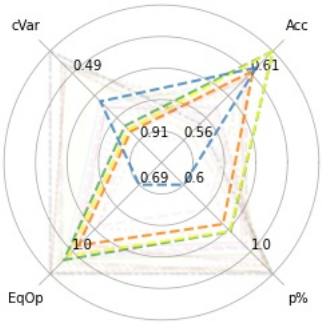
Reweighting Consiste em treinar um modelo a partir da regressão logística, mas considera pesos para cada amostra de acordo com a característica sensível.

Classificador de Paridade Demográfica Treina um modelo de regressão logística considerando uma restrição para o valor da métrica de Paridade Demográfica (ou P porcento).

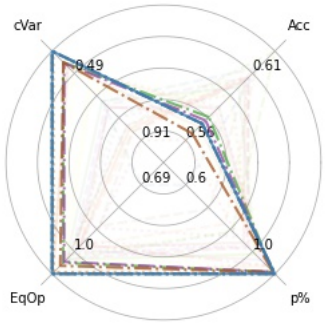
Classificador de Igualdade de Oportunidade Treina um modelo de regressão logística considerando uma restrição para o valor da métrica de Igualdade de Oportunidade.

Minimax Estratégia de otimização que encontra o modelo que minimiza o erro máximo para os grupos gerados a partir da característica sensível.

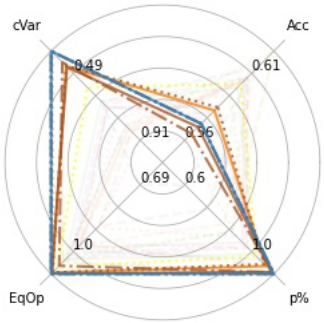
Experimento 1 - Otimização Individual



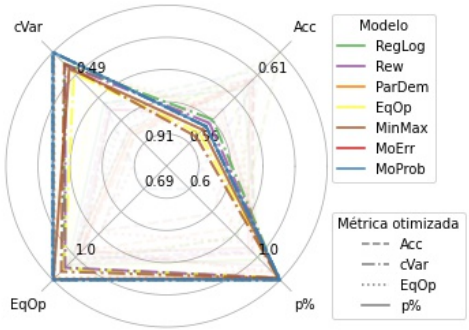
(a) Acurácia



(b) Coeficiente de variação



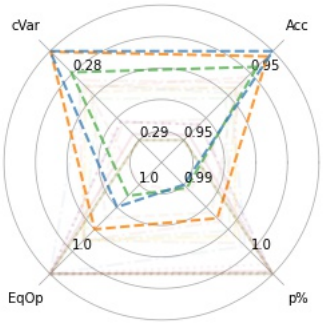
(c) Igualdade de oportunidade



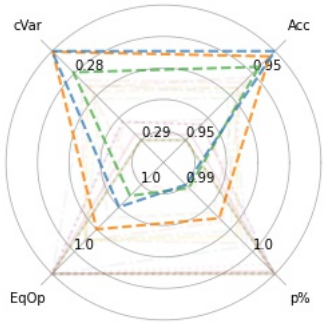
(d) Paridade demográfica

Figura 2: Melhores resultados obtidos em cada métrica para o conjunto de dados *German* para as estratégias propostas e comparadas.

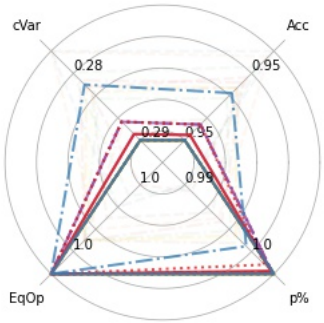
Experimento 1 - Otimização Individual



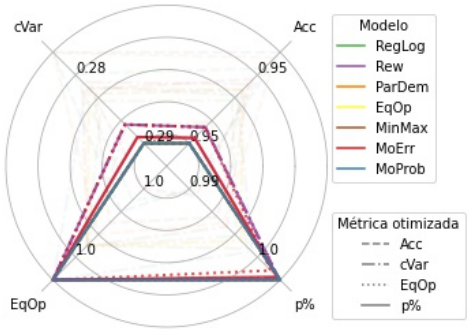
(a) Acurácia



(b) Coeficiente de variação



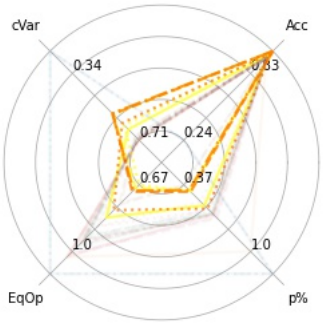
(c) Igualdade de oportunidade



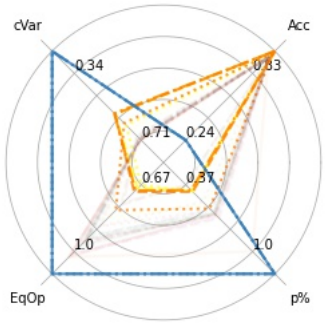
(d) Paridade demográfica

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

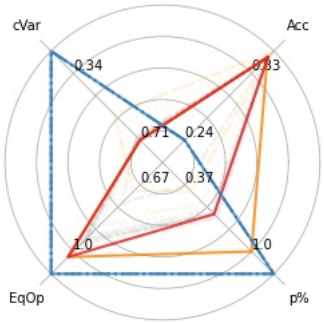
Experimento 1 - Otimização Individual



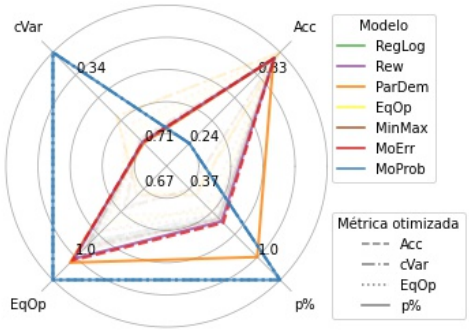
(a) Acurácia



(b) Coeficiente de variação



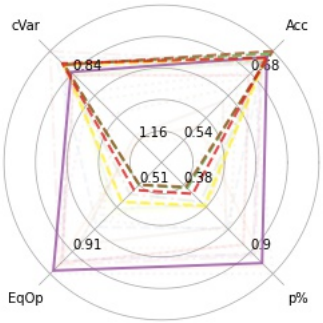
(c) Igualdade de oportunidade



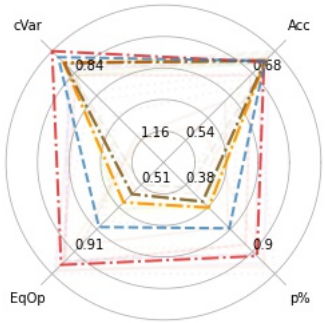
(d) Paridade demográfica

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

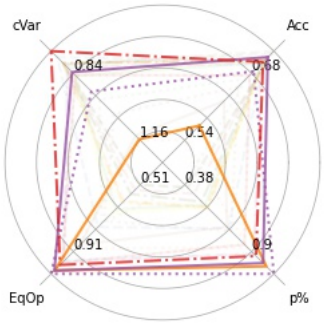
Experimento 1 - Otimização Individual



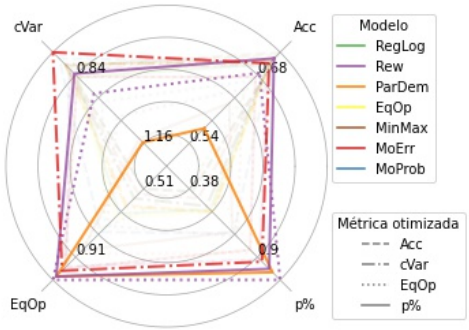
(a) Acurácia



(b) Coeficiente de variação



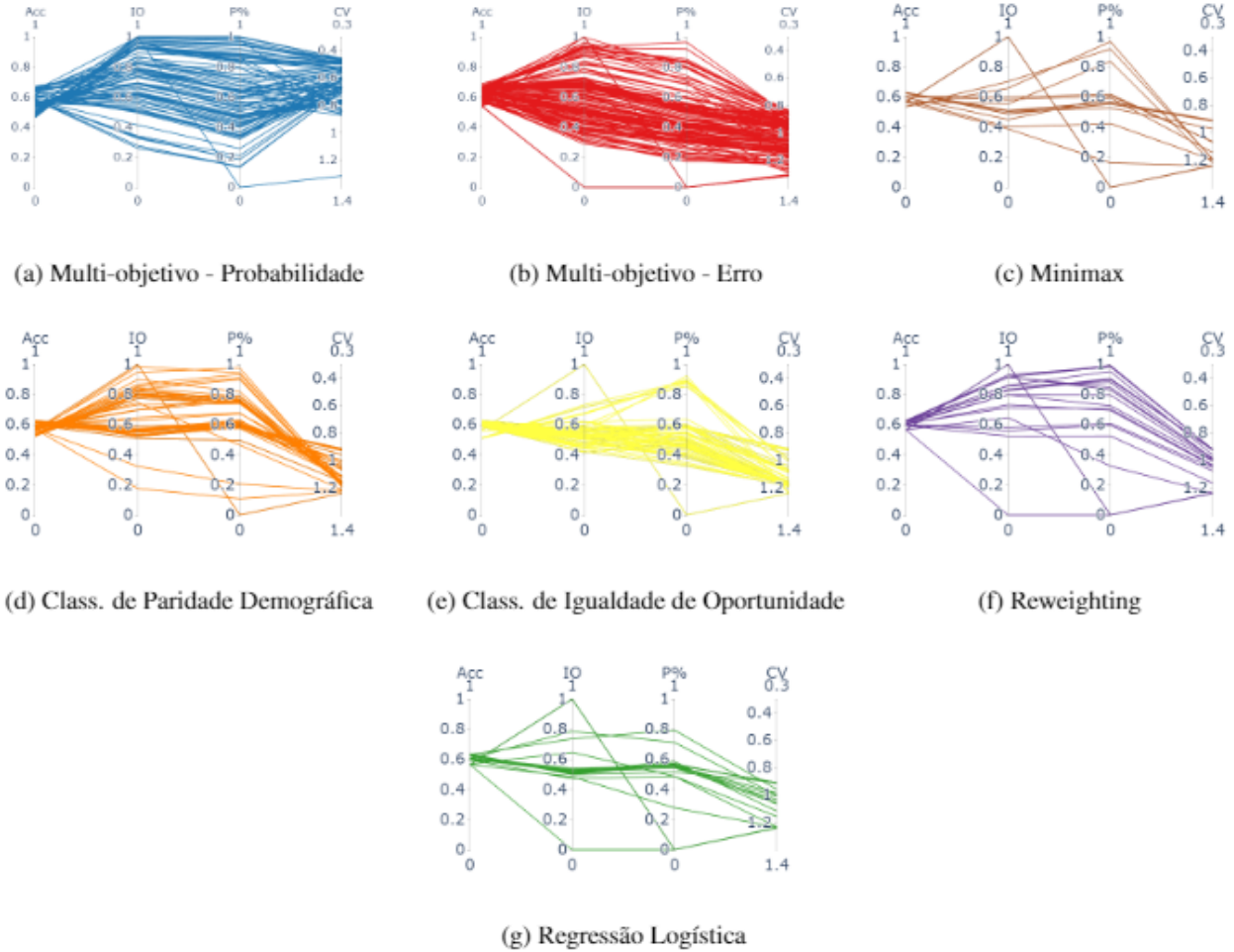
(c) Igualdade de oportunidade



(d) Paridade demográfica

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

Experimento 2 - Diversidade



(a) Multi-objetivo - Probabilidade

(b) Multi-objetivo - Erro

(c) Minimax

(d) Class. de Paridade Demográfica

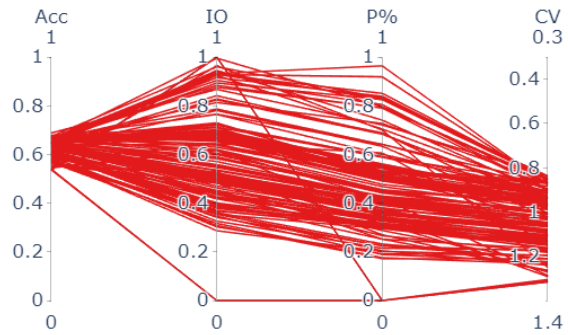
(e) Class. de Igualdade de Oportunidade

(f) Reweighting

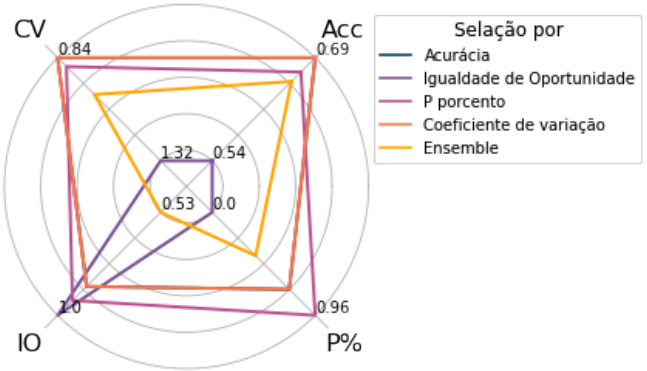
(g) Regressão Logística

Figura 6: Valores encontrados para métricas de desempenho e discriminação para os modelos resultantes de cada estratégia utilizando o conjunto de dados COMPAS.

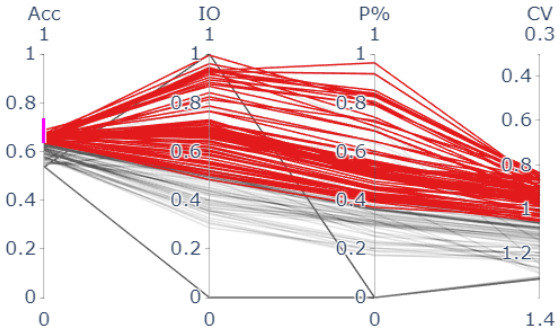
Experimento 3 - *Ensemble Learning*



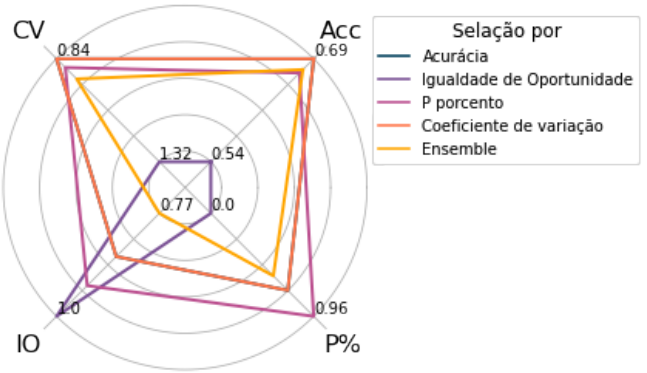
(a) Seleção dos modelos



(b) Resultados obtidos

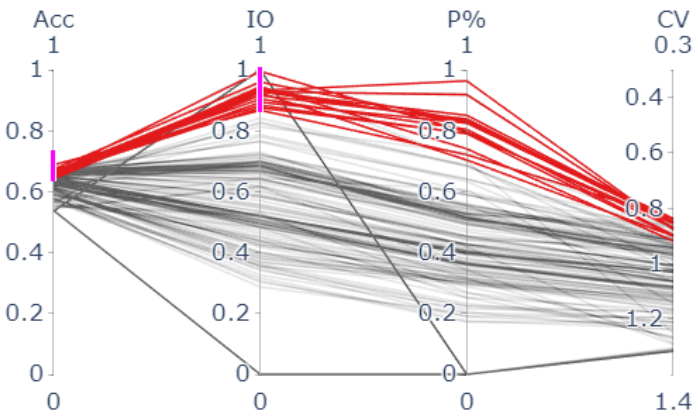


(c) Seleção dos modelos

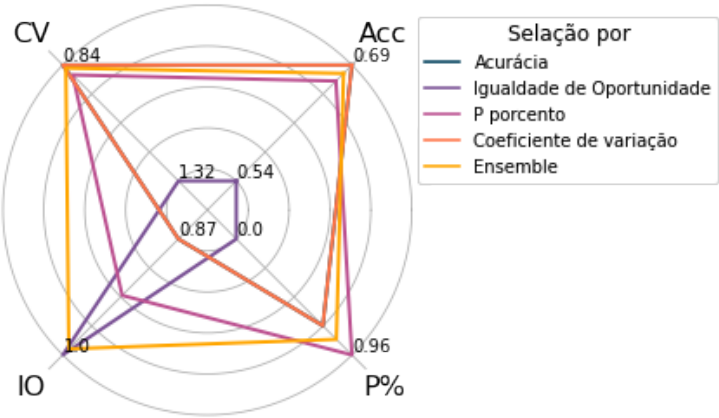


(d) Resultados obtidos

Experimento 3 - *Ensemble Learning*



(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 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.

Trabalhos Futuros

Possíveis trabalhos futuros:

- Estender a implementação e estudo para características sensíveis que não sejam binárias, por exemplo usando mais variações de raça além de branco/não-branco ou utilizar duas características sensíveis simultaneamente, como gênero e raça.
- Análise da Fronteira de Pareto resultante das estratégias propostas, comparando com as fronteiras encontradas por outras estratégias.
- Aplicação dessas estratégias para outros problemas além da classificação binária, como regressões lineares.