

Relações entre equidade no processo de redução de dimensionalidade e equidade em tarefas de classificação

Felipe Tetsuo Yamada, Guilherme Dean Pelegrina, Leonardo Tomazeli Duarte

Faculdade de Ciências Aplicadas - Universidade Estadual de Campinas

Rua Pedro Zaccaria, 1300, Limeira - SP, Brasil, CEP 13484-350

felipeyamada@gmail.com, guidean@unicamp.br,

leonardo.duarte@fca.unicamp.br

RESUMO

Métodos de aprendizado de máquina estão cada vez mais presentes no cotidiano das pessoas. Decisões automatizadas são frequentes e, em muitos casos, tendenciosas. Neste contexto, trabalhos recentes vêm explorando técnicas que procuram mitigar resultados enviesados, de maneira a garantir a equidade entre diferentes grupos de indivíduos. Via de regra, a literatura aborda essa questão em uma das etapas de construção do modelo. Neste estudo, investigamos a influência que a busca por equidade na etapa de pré-processamento, baseada na análise de componentes principais, exerce sobre a etapa de classificação. Mais precisamente, aplicamos uma técnica baseada na otimização multiobjetivo, a qual busca reduzir a disparidade entre diferentes grupos na redução da dimensão dos dados, e verificamos se também houve aumento da equidade na saída do classificador. Avaliamos essa relação a partir de experimentos numéricos realizados em um conjunto de dados de análise de crédito.

PALAVRAS CHAVE. Aprendizado de máquina, Equidade, Análise de componentes principais.

IC – Inteligência Computacional

ABSTRACT

Machine learning methods are increasingly present in people's daily lives. Automated decisions are frequent and, in many cases, biased. In this context, recent works have been exploring techniques whose goal are to mitigate biased results, in order to guarantee fairness between different groups of individuals. Usually, the literature addresses this issue in one of the steps in the model construction. In this study, we investigate the influence that the search for fairness in the pre-processing step, based on principal component analysis, has in the classification step. More precisely, we apply a multi-objective-based technique, which aim is to reduce the disparity between different groups in dimension reduction, and verify if there was also an increase in fairness in the classifier output. We evaluate this relation through numerical experiments performed on a credit analysis dataset.

KEYWORDS. Machine learning, Fairness, Principal component analysis.

IC - Computational Intelligence

1. Introdução

Os avanços tecnológicos conquistados nas últimas décadas e a enorme produção de dados gerada pela atual sociedade possibilitaram o surgimento de um ambiente propício para o desenvolvimento e aplicação de modelos de aprendizado de máquina (ML - *Machine Learning*). A adoção de ferramentas inteligentes, capazes de automatizar o apoio à tomada de decisões e gerar escalabilidade a processos rotineiros, tornou-se uma prática possível e atraente para muitas instituições. No entanto, apesar dos benefícios gerados, há diversas questões éticas que devemos nos atentar quando implementamos tal sistema em situações envolvendo pessoas [Kearns e Roth, 2019].

Recentemente, um novo tema de estudo em inteligência artificial, chamado *fairness* (traduzido, aqui, por equidade), deu início às discussões sobre os possíveis efeitos negativos e riscos decorrentes da aplicação de modelos de aprendizado de máquina em tarefas do cotidiano [Barocas et al., 2019]. Por exemplo, um estudo realizado pela ProPublica [Angwin et al., 2016] trouxe luz ao tema ao revelar que *softwares* de predição para reincidência criminal, utilizados por juízes como ferramenta de apoio, possuíam uma tendência a prejudicar injustamente indivíduos negros. No caso, a porcentagem de negros classificados como possíveis reincidentes criminais, mas que não vinham a cometer novos crimes, era consideravelmente maior em relação aos brancos. O uso de modelos enviesados pode também afetar o acesso às diversas oportunidades, como o pedido de crédito em uma instituição financeira, a procura por um emprego, ou mesmo o ingresso em uma boa universidade [O'Neil, 2016]. Torna-se evidente, portanto, o poder de influência que essas ferramentas possuem sobre a população e a necessidade de corrigir as devidas falhas envolvidas em suas aplicações. Nesse contexto, o estudo do tema equidade visa contribuir com melhores práticas no universo da inteligência artificial, comportamento decisivo para o desenvolvimento de soluções e procedimentos mais transparentes e imparciais.

Atualmente, soluções para problemas de equidade em ML vem sendo adotadas em três etapas da construção do modelo: pré-processamento, treinamento e pós-processamento [Mehrabi et al., 2019; Caton e Haas, 2020]. Em particular, o presente trabalho tem como foco métodos de pré-processamento voltados à redução de dimensionalidade, como a Análise de Componentes Principais (PCA - *Principal Component Analysis*) [Jolliffe, 2002]. Apesar de amplamente utilizada, a PCA pode gerar um problema de equidade ao criar disparidades na representação para grupos distintos de amostras. Em outras palavras, durante o processo de compactação dos dados, diferentes grupos podem sofrer perdas desiguais de suas informações referentes às suas características originais.

Frente a esse risco de soluções PCA que não são equitáveis, trabalhos recentes foram desenvolvidos com o intuito de combater a perda desigual de informações durante o processo de redução de dimensionalidade [Samadi et al., 2018; Kamani et al., 2019; Pelegrina et al., 2021]. No entanto, a literatura pouco aborda a relação entre o impacto que a adoção de técnicas equitáveis no pré-processamento tem na etapa de classificação. Nesse sentido, buscamos no presente trabalho investigar se soluções que tendem a ser mais imparciais em relação à redução da dimensão dos dados podem, também, trazer melhorias nas métricas de equidade na classificação. Para isso, exploramos a técnica MOFPCA (Multi-Objective Fair Principal Component Analysis) [Pelegrina et al., 2021], a qual lida com o problema de equidade em PCA através da otimização multiobjetivo, e verificamos a relação entre equidade nas etapas de pré-processamento e classificação. Usamos como exemplo um problema de classificação de possíveis inadimplentes na solicitação de crédito bancário.

Com relação à organização deste artigo, descrevemos os principais aspectos teóricos usados neste estudo na Seção 2. Em seguida, na Seção 3, discutimos a metodologia usada para analisar a relação entre as medidas de equidade consideradas nas duas etapas do modelo de aprendizado de máquina. Os experimentos e os resultados obtidos são apresentados na Seção 4. Por fim, as

conclusões deste estudo e as futuras perspectivas são discutidas na Seção 5.

2. Fundamentação teórica

Como mencionado na seção anterior, este estudo procura investigar a relação entre explorar soluções *fair* na etapa de pré-processamento e a as métricas de equidade calculadas na classificação. Nesta seção, discutimos os principais aspectos teóricos associados a esses dois tópicos.

2.1. Métricas de desempenho e equidade em problemas de classificação

A fim de facilitar o entendimento das métricas de desempenho e equidade usadas em problemas de classificação, considere, à título de exemplo, que uma instituição financeira precise avaliar pedidos de acesso a crédito. Para isso, o banco precisará das informações (dados) pessoais dos clientes que já receberam créditos (e pagaram ou não) a fim de construir um modelo de aprendizado de máquina que classifique cada novo cliente como possível adimplente ou inadimplente. Vale ressaltar que decisões de empréstimos são importantes ao banco e erros cometidos nesse processo podem limitar potenciais ganhos da instituição através de duas formas: negando-se empréstimos a clientes com perfis adimplentes ou cedendo crédito a clientes com perfis inadimplentes. Logo, deseja-se que o modelo de classificação atue com as menores taxas de erros possíveis e, para isso, torna-se relevante a execução de uma etapa de validação do modelo.

Matematicamente, podemos definir a matriz de dados como $\mathbf{A} \in \mathbb{R}^{m \times n}$, onde cada linha $i = 1, \dots, m$ representa um cliente (ou amostra do modelo) e cada coluna $j = 1, \dots, n$ representa uma informação (ou atributo) sobre o cliente. Além disso, no treinamento do modelo, temos o vetor de m rótulos $\mathbf{b}$, o qual indica a classe a qual cada amostra pertence. Em nosso exemplo, podemos definir $b_i = 0$ ou $b_i = 1$ quando o cliente i é tido como inadimplente ou adimplente, respectivamente. As principais métricas de validação para problemas de classificação são baseadas na matriz de confusão, estratégia utilizada na Tabela 1. Observa-se que as respostas 1 ou 0 do classificador estão distribuídas ao longo da primeira ou da segunda linha, respectivamente. No cruzamento de cada rótulo com a saída do classificador temos as seguintes métricas:

- verdadeiro positivo (TP): número de clientes classificados corretamente como inadimplentes;
- falso positivo (FP): número de clientes classificados como adimplentes quando deveriam ser inadimplentes;
- verdadeiro negativo (TN): número de clientes classificados corretamente como adimplentes;
- falso negativo (FN): número de clientes classificados como inadimplentes quando deveriam ser adimplentes.

Tabela 1: Matriz de confusão.

		Rótulo	
		1 (inadimplente)	0 (adimplente)
Classif.	1 (inadimplente)	TP	FP
	0 (adimplente)	FN	TN

Neste estudo, uma métrica de desempenho que terá grande importância na etapa de validação é a Taxa de Verdadeiro Positivo (TPR - *True Positive Rate*) [Alpaydin, 2014], dada por

$$TPR = \frac{TP}{TP + FN}. \quad (1)$$

Embora a TPR indique a acurácia do modelo em classificar os rótulos “positivos” (ou inadimplentes em nosso exemplo), essa métrica não considera diferentes grupos de indivíduos em seu cálculo. Dessa forma, ao olhar apenas para essa métrica na construção do modelo, podemos ter resultados injustos em relação a diferentes grupos de indivíduos. Neste contexto, uma forma de avaliar possíveis disparidades entre dois grupos G_1 e G_2 consiste em calcular as TPRs de cada um e comparar os valores encontrados. Em um cenário equitável para ambos os grupos espera-se que

$$TPR_{G_1} = TPR_{G_2}. \quad (2)$$

A igualdade apresentada na Equação (2) representa o conceito de igual oportunidade (EO ou *Equal Opportunity*), o qual se associa à saída “positiva”(ou 1) do classificador. Note que a EO pode não refletir uma classificação que resulte nos mesmos “benefícios” para os indivíduos de grupos diferentes. Por exemplo, ao considerar o rótulo 1 como inadimplente no problema da avaliação de crédito, $TPR_{G_1} = TPR_{G_2}$ indica uma igualdade em relação à negação de crédito por parte dos indivíduos corretamente classificados como inadimplentes.

2.2. Equidade na redução da dimensão dos dados e o método MOFPCA

Na seção anterior, apresentamos uma métrica de equidade usada na etapa de classificação do modelo de aprendizado de máquina. Nesta seção, será discutida a equidade na etapa de pré-processamento. Sendo o foco deste estudo avaliar se a busca por soluções *fair* na etapa de pré-processamento implica em melhores métricas de equidade na classificação, abordaremos com maior detalhe os procedimentos usados na primeira etapa. Mais especificamente, apresentaremos o método clássico da análise de componentes principais e a proposta MOFPCA [Pelegrina et al., 2021], a qual busca por projeções equitáveis na redução da dimensão dos dados.

2.2.1. Análise de componentes principais

A PCA [Jolliffe, 2002] é uma ferramenta amplamente utilizada em problemas de redução da dimensão dos dados. Seja $\mathbf{A} \in \mathbb{R}^{m \times n}$ uma matriz formada por m amostras e n atributos. A PCA busca minimizar o erro de representação entre a matriz de dados original $\mathbf{A}$ e sua projeção $\hat{\mathbf{A}}$ de posto $k \leq n$. Logo, a PCA pode ser formulada como o seguinte problema de otimização:

$$\begin{aligned} \min_{\mathbf{W}} \quad & \|\mathbf{A} - \mathbf{A}\mathbf{W}\mathbf{W}^T\|_F^2 \\ \text{sujeito a} \quad & \mathbf{W}^T\mathbf{W} = \mathbf{I}, \end{aligned} \quad (3)$$

onde a matriz $\mathbf{W} \in \mathbb{R}^{n \times k}$ mapeia as amostras da base original $\mathbf{A}$ até sua representação k -dimensional $\hat{\mathbf{A}} = \mathbf{A}\mathbf{W}$, $\|\mathbf{A} - \mathbf{A}\mathbf{W}\mathbf{W}^T\|_F$ indica a norma de Frobenius [Golub e Loan, 1996] e $\mathbf{I}$ é a matriz identidade. A matriz responsável por reduzir o posto da matriz $\mathbf{A}$ será denominada matriz de projeção, sendo representada por $\mathbf{P} = \mathbf{W}\mathbf{W}^T$, onde $\mathbf{P} \in \mathbb{R}^{n \times n}$.

2.2.2. Uma proposta para soluções equitáveis em PCA

Considere que a matriz $\mathbf{A}$ possa ser dividida em

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{G_1} \\ \mathbf{A}_{G_2} \end{bmatrix}. \quad (4)$$

onde $\mathbf{A}_{G_1} \in \mathbb{R}^{m_1 \times n}$ e $\mathbf{A}_{G_2} \in \mathbb{R}^{m_2 \times n}$ representam as amostras pertencentes aos grupos G_1 (com m_1 amostras) e G_2 (com m_2 amostras), respectivamente. É notável que a PCA clássica não leva em conta diferentes grupos quando realizamos a redução da dimensão. Sendo assim, o método pode promover uma disparidade entre indivíduos, uma vez que, no espaço de dimensão reduzida, um

grupo pode ter uma melhor representação do que o outro. É neste contexto que o método MOFPCA [Pelegrina et al., 2021] visa incluir uma medida de equidade quando da busca pela redução da dimensão através da análise de componentes principais. Mais precisamente, o método MOFPCA aborda o problema a partir da otimização multiobjetivo [Miettinen, 1999], ou seja, otimizando tanto o erro de representação $\mathcal{R}(\mathbf{W}) = \|\mathbf{A} - \mathbf{A}\mathbf{W}\mathbf{W}^T\|_F^2$ quanto a medida de equidade

$$\mathcal{F}(\mathbf{W}) = \left(\frac{1}{m_1} \|\mathbf{A}_{G1} - \mathbf{A}_{G1}\mathbf{W}\mathbf{W}^T\|_F^2 - \frac{1}{m_2} \|\mathbf{A}_{G2} - \mathbf{A}_{G2}\mathbf{W}\mathbf{W}^T\|_F^2 \right)^2 \quad (5)$$

simultaneamente. Observe que a medida de equidade adotada pelo método MOFPCA tende a zero quando estamos em um cenário em que não há disparidades entre os grupos $G1$ e $G2$ na redução da dimensão. Sendo assim, o método MOFPCA pode ser resumido através do seguinte problema de otimização multiobjetivo¹:

$$\begin{aligned} \min_{\mathbf{W}} \quad & [\mathcal{R}(\mathbf{W}), \mathcal{F}(\mathbf{W})] \\ \text{sujeito a} \quad & \mathbf{W}^T \mathbf{W} = \mathbf{I}. \end{aligned} \quad (6)$$

É importante salientar que os objetivos do problema (6) tendem a possuir uma relação de compromisso, não sendo possível encontrar uma solução que otimize ambos simultaneamente. Alterações positivas em um deles provavelmente gerará resultados negativos para o outro. Neste cenário, a abordagem multiobjetivo fornece um conjunto de soluções, chamadas de soluções não-dominadas ou Pareto-ótimas, as quais são obtidas a partir da relação de dominância, definida como [Miettinen, 1999]:

Definição 1 (Relação de dominância): Considere duas soluções factíveis a e a' . Dizemos que a domina a' se a é tão boa quanto a' em todos os objetivos e a é estritamente melhor que a' em pelo menos um objetivo.

Então, no caso do problema da redução da dimensão equitável, uma solução é dita não-dominada se nenhum dos objetivos (erro de reconstrução ou medida de equidade) pode ser melhorado sem que haja uma piora no outro objetivo. Em outras palavras, se há uma solução com um menor erro de reconstrução, necessariamente, esta solução terá um maior valor de $\mathcal{F}(\mathbf{W})$. Nota-se, portanto, que dentro do conjunto de soluções não-dominadas há as soluções que minimizam cada um dos objetivos individualmente e as soluções de compromissos entre eles. Retomaremos a discussão sobre essas soluções na próxima seção.

3. Metodologia proposta

3.1. Bases de dados utilizadas

A base de dados escolhida para este estudo contém informações bancárias de trinta mil clientes de um banco taiwanês [Yeh e Lien, 2009]. Cada cliente é descrito por 23 atributos, desde informações pessoais (idade, gênero, nível educacional, ano de nascimento, estado civil e escolaridade) até informações sobre uso de crédito (registradas em um intervalo de seis meses, entre abril e setembro de 2005). Nesse contexto, EDUCATION é definido como atributo sensível e dividido em dois grupos: clientes com nível superior (*high_ed*) e clientes sem nível superior (*low_ed*). A base de dados conta com uma coluna de rótulos, informando a condição de inadimplência de cada cliente

¹No problema de otimização (6), a restrição não se faz necessário uma vez que o método proposto por Pelegrina et al. [2021] usa as componentes principais obtidas pelo PCA clássico, ou seja, vetores que já são ortogonais. No entanto, a fim de facilitar o entendimento deste estudo, preferimos ser redundantes na definição do problema.

(1: inadimplente, 0: adimplente) em relação ao último pagamento registrado, referente ao mês de outubro de 2005.

O fluxograma da Figura 1a revela a composição da base de dados original. Note que há uma divisão a partir dos grupos de rótulos (*Rótulo 1* ou *Rótulo 0*) e dos grupos do atributo EDUCATION (*high_ed* ou *low_ed*). Além disso, há um desequilíbrio na proporção entre os grupos de rótulos e os grupos de atributos EDUCATION. Levando em conta que bases desequilibradas podem interferir nos estudo de equidade [Hinnefeld et al., 2018], este estudo fará uso de uma segunda base de dados, denominada base proporcional (Figura 1b). Sua construção inicia-se com a escolha do menor grupo de amostras dentre os subgrupos g_1 , g_2 , g_3 e g_4 apresentados na Figura 1a. Observe que o grupo g_2 atende à restrição, pois apresenta o menor grupo de amostras (1.270 unidades). A base proporcional é, então, gerada a partir de modificações na base original. Reduz-se, aleatoriamente, a quantidade de amostras dos outros grupos (g_1 , g_3 e g_4) até esse mínimo de 1.270 amostras. Note que, na Figura 1b, a base proporcional apresenta grupos equilibrados, tanto entre os grupos de rótulos, quanto entre os grupos do atributo sensível.

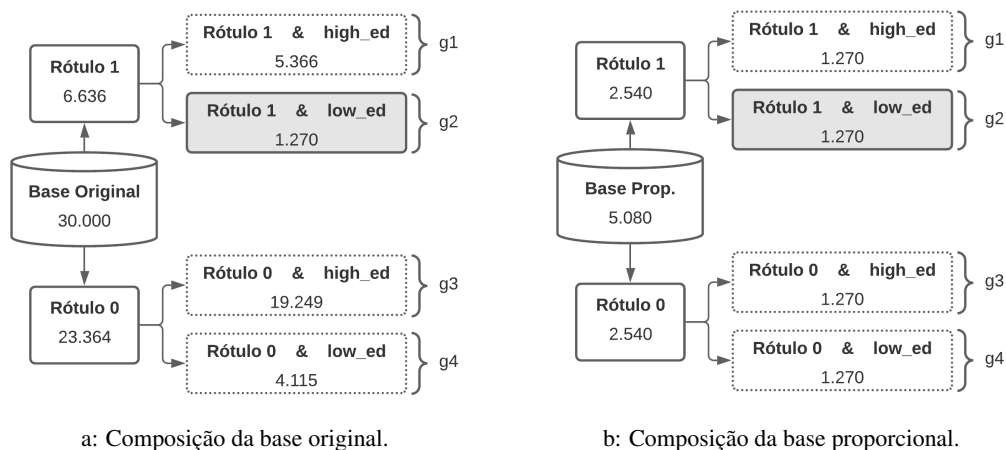


Figura 1: Representação das bases de dados utilizadas no estudo.

As bases de dados são divididas em cinco conjuntos (*folds*), cada qual contendo amostras de treino e teste. As amostras da base original são divididas em *folds* através do método *Stratified K-fold* [Ojala e Garriga, 2010]. Por outro lado, a divisão da base proporcional é feita de acordo com a estratégia demonstrado na Figura 2, pois há interesse em se preservar a mesma proporção entre os grupos sensíveis e rótulos.

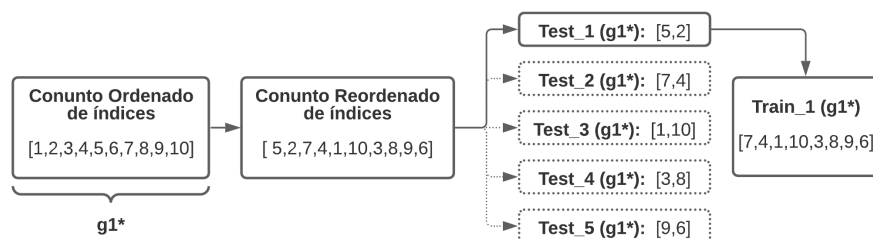


Figura 2: Folds.

Observa-se, na Figura 2, um pequeno grupo de dez amostras, denominado $g1^*$. Cada amostra é associada a um índice específico, formando um conjunto ordenado de índices. Estes são posteriormente reordenados através de um sorteio aleatório. Em seguida, o conjunto reordenado de amostras é então dividido em cinco intervalos consecutivos, formando cinco subconjuntos de amostras de teste. Obtém-se as amostras de treinamento para cada *fold* ao eliminar do conjunto total de amostras o respectivo subconjunto de teste. As amostras de treino e teste para a base proporcional são então obtidas aplicando-se esse método de divisão aos subgrupos $g1'$, $g2'$, $g3'$ e $g4'$ da Figura 1b.

3.2. Comparação entre métricas

A Figura 3a mostra um exemplo de resultado produzido pelo método MOFPCA: um conjunto de três soluções não-dominadas, distribuídas ao longo da fronteira de Pareto. Cada solução representa uma respectiva matriz de projeção, avaliada a partir do erro de reconstrução ($\mathcal{R}$) e da métrica de equidade ($\mathcal{F}$). Observa-se que esse conjunto de soluções agrupa casos extremos, contendo tanto o menor, quanto o maior valor de $\mathcal{F}$ (F1 e F3, respectivamente). Cada solução produz uma nova matriz de posto reduzido, utilizada como base de treinamento para modelos de classificação. De acordo com a Equação (5), baixos valores de $\mathcal{F}$ representam maior equidade na etapa de redução de dimensionalidade. Logo, caso haja relação entre as métricas de equidade no pré-processamento e na classificação, espera-se que os valores de TPR entre grupos distintos se aproximem à medida que os valores de $\mathcal{F}$ diminuam, como exemplificado na Figura 3b. Dessa forma, a equidade na classificação será avaliada através da diferença (em módulo) entre as TPRs dos grupos distintos, *i.e.*, $|TPR_{G_1} - TPR_{G_2}|$, sendo que essa métrica tende a ser menor em um cenário de maior equidade.

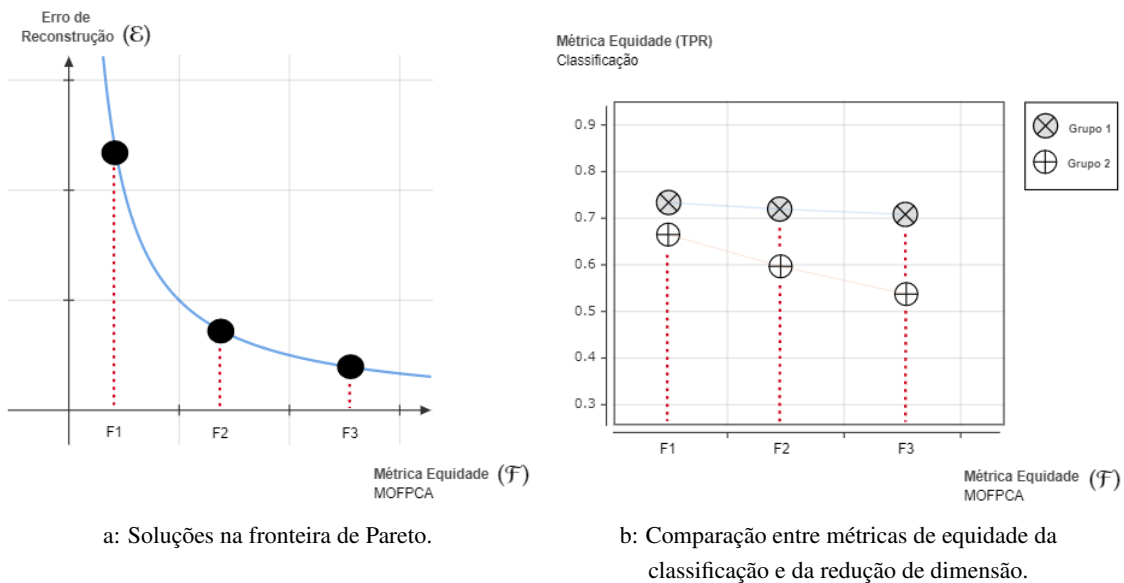


Figura 3: A figura à esquerda (a) representa a distribuição de soluções ao longo da fronteira de Pareto. A figura à direita (b) compara possíveis relações entre diferentes métricas de equidade.

A Figura 4 traduz o *pipeline* dos experimentos. Nota-se que o fluxograma inicia-se com o exemplo de uma base proporcional, sendo esta dividida em 5 *folds*. Cada *fold* apresentará um conjunto de amostras de treino e teste. Tendo em mãos as amostras de treino, inicia-se o método

MOFPCA, com o intuito de gerar um conjunto de matrizes de projeções (fronteira de Pareto). Em seguida, cada uma dessas matrizes P^* poderão gerar conjuntos de X'_{train} e X'_{test} , bases de posto reduzido. O modelo de predição será então treinado utilizando-se da base X'_{train} . Por fim, a predição é realizada com ambas as bases de posto reduzido (X'_{train} e X'_{test}) e resultados de TPR_{G1} , TPR_{G2} e $\mathcal{F}$ são gerados para as amostras de treino e teste.

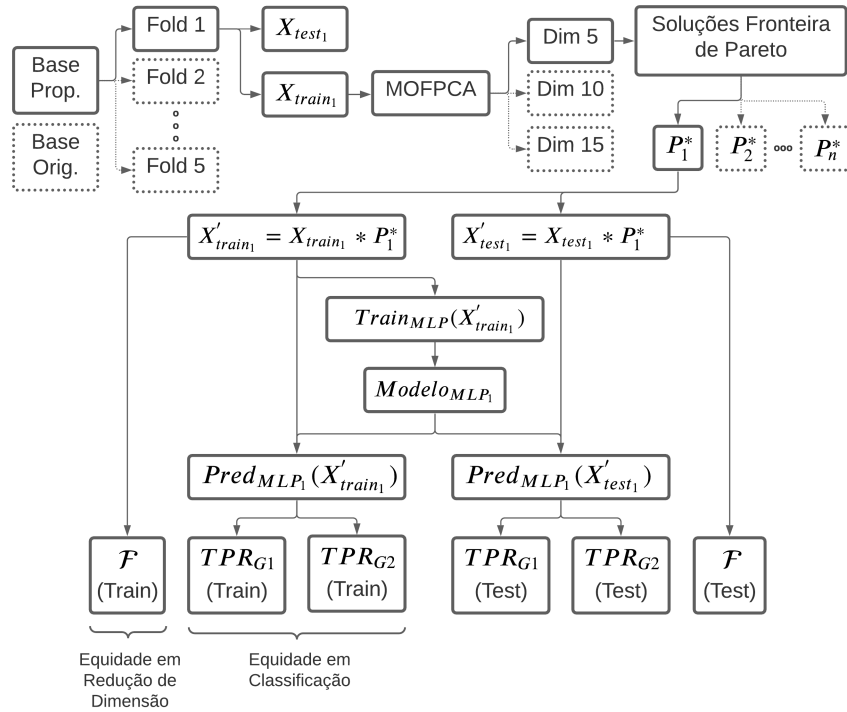


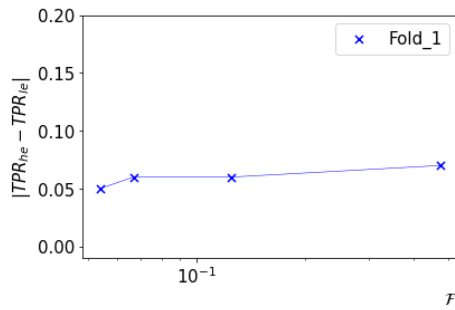
Figura 4: Fluxograma do processo de treinamento e teste de modelos preditivos.

4. Experimentos numéricos

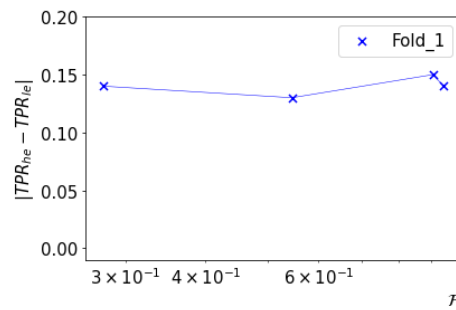
Tendo em vista a metodologia demonstrada na Figura 4, esta seção trará 4 resultados distintos. Vale lembrar que o estudo levou em conta a análise de bases proporcionais e originais, cada qual analisada para postos 5, 10 e 15. No entanto, dada a limitação do número de páginas, esta seção mostrará apenas os resultados para as bases de posto 5. A organização das imagens, contendo o resultado para cada experimento, é esclarecida pela Tabela 2. Por exemplo, em relação ao primeiro resultado (primeira coluna da Figura 5), adotamos a base proporcional e usamos as amostras X'_{train} e X'_{test} para a análise dos resultados acerca da redução da dimensão dos dados para posto 5.

Tabela 2: Organização dos 4 resultados distribuídos nas Figuras 5 e 6.

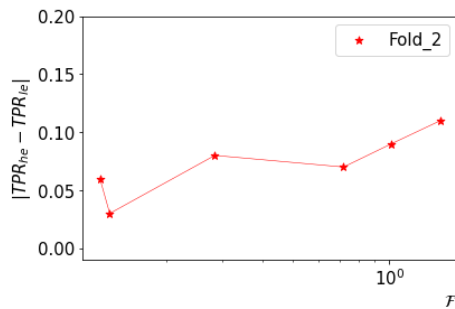
Resultado	Base	Posto	Amostras	Figura	Imagens
1	Proporcional	5	X'_{train}	5	a, c, e, g e i (1ª coluna da Figura 5)
2		5	X'_{test}	5	b, d, f, h e j (2ª coluna da Figura 5)
3	Original	5	X'_{train}	6	a, c, e, g e i (1ª coluna da Figura 6)
4		5	X'_{test}	6	b, d, f, h e j (2ª coluna da Figura 6)



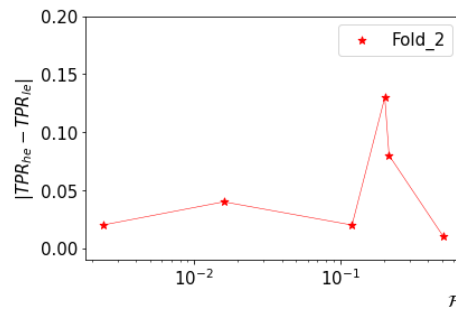
a: Posto 5 — Fold 1 — Xtrain.



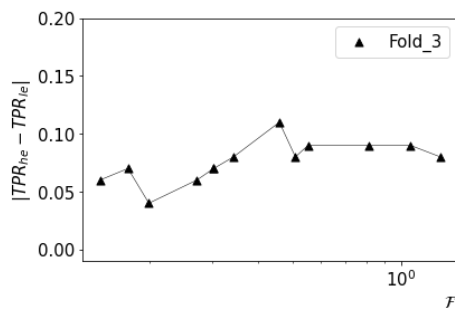
b: Posto 5 — Fold 1 — Xtest.



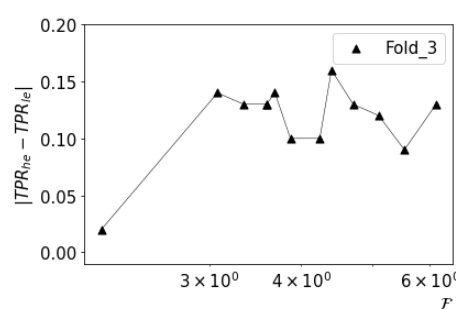
c: Posto 5 — Fold 2 — Xtrain.



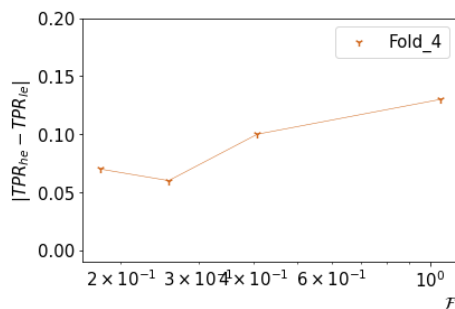
d: Posto 5 — Fold 2 — Xtest.



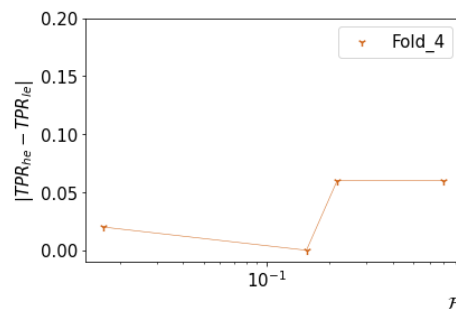
e: Posto 5 — Fold 3 — Xtrain.



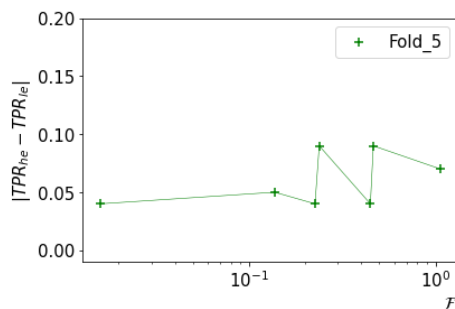
f: Posto 5 — Fold 3 — Xtest.



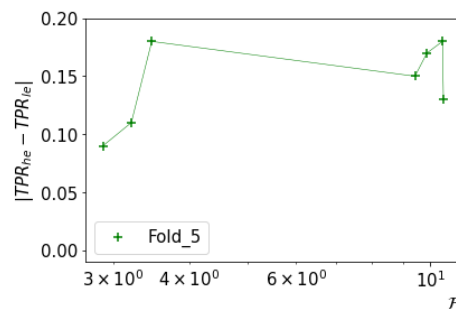
g: Posto 5 — Fold 4 — Xtrain.



h: Posto 5 — Fold 4 — Xtest.

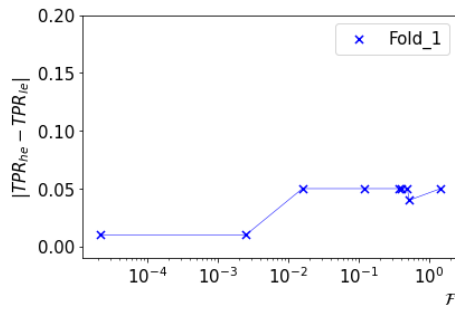


i: Posto 5 — Fold 5 — Xtrain.

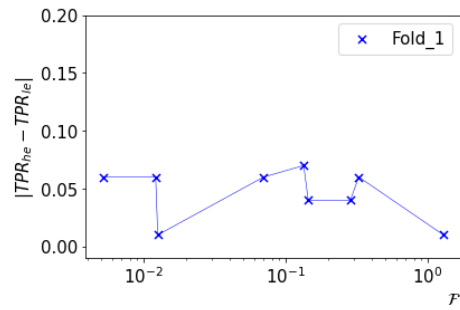


j: Posto 5 — Fold 5 — Xtest.

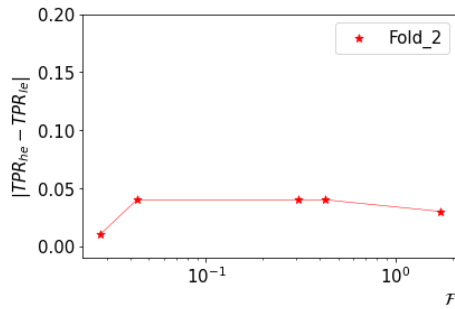
Figura 5: Resultados da base proporcional. Coluna à esquerda representam predições feitas sobre os dados de treinamento e coluna à direita representam predições sobre os dados de teste.



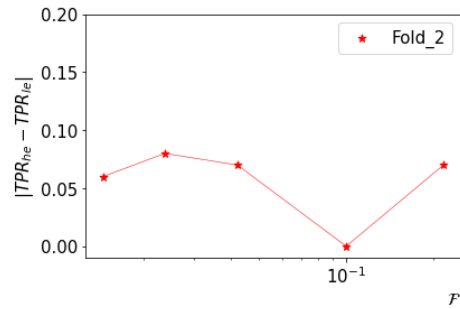
a: Posto 5 — Fold 1 — Xtrain.



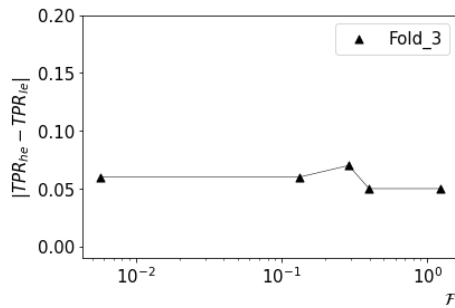
b: Posto 5 — Fold 1 — Xtest.



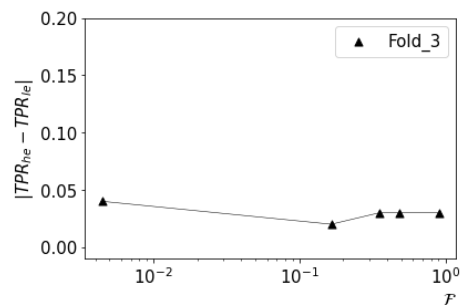
c: Posto 5 — Fold 2 — Xtrain.



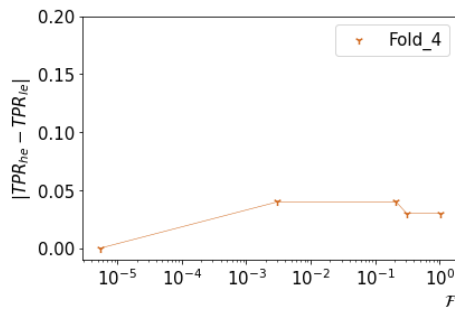
d: Posto 5 — Fold 2 — Xtest.



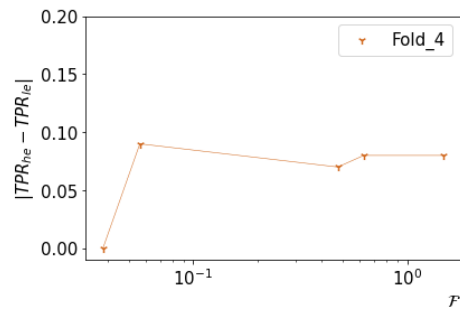
e: Posto 5 — Fold 3 — Xtrain.



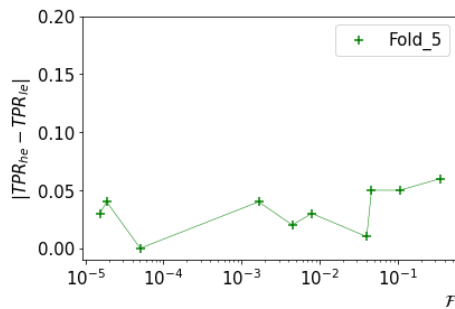
f: Posto 5 — Fold 3 — Xtest.



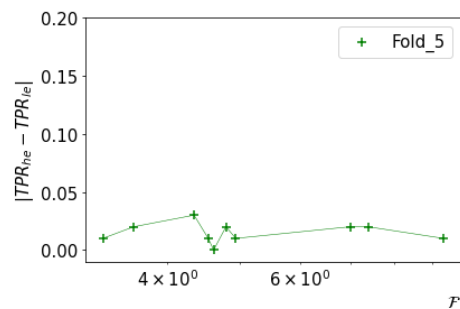
g: Posto 5 — Fold 4 — Xtrain.



h: Posto 5 — Fold 4 — Xtest.



i: Posto 5 — Fold 5 — Xtrain.



j: Posto 5 — Fold 5 — Xtest.

Figura 6: Resultados da base original. Coluna à esquerda representam predições feitas sobre os dados de treinamento e coluna à direita representam predições sobre os dados de teste.

Algumas conclusões preliminares interessantes podem ser retiradas dos resultados apresentados nas Figuras 5 e 6. Levando em conta a base proporcional (Figura 5), nota-se que, no treinamento, alguns *folds* aparentam corroborar a hipótese exemplificada pela Figura 3b (Figuras 5c, 5e e 5g, no caso). No entanto, não temos a mesma consistência de resultados na aplicação do modelo nos dados de teste. Por exemplo, embora a curva apresentada na Figura 5c pareça decrescer na medida em que se aumenta a equidade no pré-processamento (ou seja, valores baixos para $\mathcal{F}$ implicaria em menores valores para $|TPR_{G_1} - TPR_{G_2}|$), essa relação não é confirmada nos respectivos dados de teste (Figura 5d). Além da possível ausência de relação entre equidade nas duas etapas, uma hipótese para esse resultado seria a questão que, embora o método MOFPCA forneça as soluções não-dominadas na etapa de treinamento, as projeções para os dados de teste podem não ser não-dominadas. Essa disparidade, então, pode mitigar o efeito que as soluções mais equitáveis na redução da dimensão exerce sobre a equidade na classificação.

Outra conclusão parcial que podemos extrair dos resultados consiste na comparação entre as bases de dados. No caso, a possível ausência da relação entre equidade nas duas etapas consideradas parece ser evidenciada quando adotamos a base de dados original. A grande maioria dos resultados apresentados na Figura 5 não indicam um padrão que confirme a relação investigada.

5. Considerações finais

Este trabalho propôs avaliar se há uma relação direta entre buscar soluções mais equitáveis na etapa de redução de dimensão dos dados e a obtenção de soluções mais equitáveis em relação à métrica TPR. Para isso, consideramos as projeções *fair* obtidas a partir do método MOFPCA, o qual aborda a análise de componentes principais a partir da otimização multiobjetivo, a verificamos a equidade na etapa da classificação para as diferentes soluções não-dominadas. Vale lembrar que, dentre as soluções não-dominadas, há a que otimiza a equidade na redução de dimensão, a que minimiza o erro de reconstrução e as soluções de compromisso entre esses dois objetivos.

A partir dos experimentos realizados neste estudo e dos resultados alcançados, não podemos afirmar a hipótese de que um possível ganho em equidade na etapa de redução de dimensionalidade implicará em um ganho de equidade em uma etapa posterior de classificação. Embora alguns casos na etapa de treinamento aparenta indicar tal relação, eles não foram consistentes em relação aos dados de teste. Além disso, houve diferenças significativas entre os resultados levando em conta a base de dados proporcional e a original.

Sendo assim, trabalhos futuros podem ser conduzidos a fim de investigar com maior profundidade essa relação, levando em conta, por exemplo, outras métricas de equidade. Também podem ser consideradas diferentes bases de dados, principalmente com mais atributos, a fim de investigar se há uma diferença expressiva quando nos deparamos com dimensões de dados muito elevadas. Outra perspectiva futura seria, como já apontado na seção anterior, investigar se as projeções obtidas pelo MOFPCA também levaria a soluções não-dominadas nos dados de teste.

Agradecimentos

Os autores gostariam de agradecer aos processos n. 2020/10572-5 e 2020/01089-9, Fundação de Amparo a Pesquisa do Estado de São Paulo (FAPESP), e ao processo n. 312228/2020-1, Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

Referências

Alpaydin, E. (2014). *Introduction to machine learning*. MIT Press, Cambridge, USA, 3 edition.

- Angwin, J., Larson, J., Mattu, S., e Kirchner, L. (2016). Machine Bias - ProPublica. URL <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.
- Barocas, S., Hardt, M., e Narayanan, A. (2019). *Fairness in machine learning*. fairmlbook.org. URL <http://www.fairmlbook.org>.
- Caton, S. e Haas, C. (2020). Fairness in machine learning: A survey. *ArXiv ID: 2010.04053*. URL <http://arxiv.org/abs/2010.04053>.
- Golub, G. H. e Loan, C. F. V. (1996). *Matrix computations*. The Johns Hopkins University Press.
- Hinnefeld, J. H., Cooman, P., Mammo, N., e Deese, R. (2018). Evaluating Fairness Metrics in the Presence of Dataset Bias. *arXiv preprint arXiv:1809.09245*. URL <http://arxiv.org/abs/1809.09245>.
- Jolliffe, I. T. (2002). *Principal component analysis*. Springer-Verlag, New York, USA.
- Kamani, M. M., Haddadpour, F., Forsati, R., e Mahdavi, M. (2019). Efficient fair principal component analysis. *ArXiv ID: 1911.04931v2*. URL <http://arxiv.org/abs/1911.04931v2>.
- Kearns, M. e Roth, A. (2019). *The ethical algorithm: The science of socially aware algorithm design*. Oxford University Press, New York, USA.
- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., e Galstyan, A. (2019). A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*. URL <http://arxiv.org/abs/1908.09635>.
- Miettinen, K. M. (1999). *Nonlinear multiobjective optimization*. International Series in Operations Research and Management Science 12, Kluwer Academic Publishers, Norwell, MA, USA.
- Ojala, M. e Garriga, G. C. (2010). Permutation tests for studying classifier performance. *Journal of Machine Learning Research*, 11(6).
- O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown, New York, USA.
- Pelegriana, G. D., Brotto, R. D. B., Duarte, L. T., Attux, R., e Romano, J. M. T. (2021). A novel multi-objective-based approach to analyze trade-offs in Fair Principal Component Analysis. *arXiv preprint arXiv:2006.06137*. URL <http://arxiv.org/abs/2006.06137v2>.
- Samadi, S., Tantipongpipat, U., Morgenstern, J., Singh, M., e Vempala, S. (2018). The price of fair PCA: One extra dimension. In *Advances in Neural Information Processing Systems*, p. 10976–10987.
- Yeh, I. C. e Lien, C.-h. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2 PART 1):2473–2480.