Fairness em Machine Learning: Uma análise baseada na Teoria de Resposta ao Item

Thiago Paracampo de Castro[1], **Lucas Ferraro Cardoso**[1,2], **Vitor Cirilo Santos**[2], **Ronnie Oliveira Alves**[2], **Regiane Kawasaki Francês**[1]

[1]Universidade Federal do Pará (UFPA)
Belém – PA – Brasil

[2]Instituto Tecnológico Vale (ITV)
Belém – PA – Brasil.

{thiago.castro, lucas.cardoso}@icen.ufpa.br

vitor.cirilo3@gmail.com, ronnie.alves@itv.org, kawasaki@ufpa.br

Abstract. *The popularization of Machine Learning tools has made a new problem evident: unfair models applied in sensitive contexts. In view of this, this article explores the use of Item Response Theory in the analysis of models trained on sensitive context data. As a case study, the Credit-G dataset was used to analyze the impact of an existing injustice in the dataset. The data presents a gender representation bias (male and female), therefore, the generation of synthetic data was used to mitigate this injustice. The evaluation of the item parameters for the two groups (biased and mitigated) by the IRT concepts indicate that biased data can be perceived by the occurrence of instances with negative discrimination. The results obtained show that the set of female instances presented average values of* $-1,675$ *of discrimination, while the male instances presented* 3.2*, for the mitigated set.*

Resumo. *A popularização das ferramentas de Machine Learning tornou evidente um novo problema: modelos injustos aplicados em contextos sensíveis. Diante disso, este artigo explora a utilização da Teoria de Resposta ao Item na análise de modelos treinados em dados de contexto sensível. Como estudo de caso, o dataset Credit-G foi utilizado para analisar o impacto de uma injustiça existente em si. Os dados apresentam um viés de representação de gênero, portanto, a geração de dados sintéticos foi utilizada para mitigar essa injustiça. A avaliação pelos conceitos da TRI dos parâmetros de item para os dois grupos (enviesados e mitigados) indicam que dados enviesados podem ser percebidos pela ocorrência de instâncias com discriminação negativa. Os resultados obtidos mostram que, para o conjunto mitigado, as instâncias femininas apresentaram valores médios de* -1.675 *de discriminação, enquanto as masculinas apresentaram* 3.2*.*

1. Introdução

Na última década, sistemas que usam Inteligência Artificial (IA) se demonstraram extremamente úteis no auxílio e exploração dos diversos campos do conhecimento. Um de seus subdomínios, o Aprendizado de Máquina, ou *Machine Learning* (ML), se tornou

ubíquo na procura de soluções científicas, médicas e empresariais, em função de sua capacidade de observar padrões e prever resultados. Embora seu uso torne o trabalho mais eficiente, conforme a distribuição desses modelos de ML se intensifica, suas predições impactam cada vez mais a vida das pessoas. Por exemplo, as esferas médicas e jurídicas são alvos de diversas promessas em relação ao uso de ML no avanço de seus diversos processos [Sidey-Gibbons and Sidey-Gibbons 2019, Sengupta and Dave 2022], entretanto, é necessário cautela ao tratar de situações críticas como essas.

Esses ambientes críticos são denominados como "contextos sensíveis", tendo em vista que a decisão de modelos preditivos impacta significativamente a vida de algum indivíduo. Por essa razão, diversas discussões éticas sobre a aplicação dessas ferramentas se difundiram. [Vayena et al. 2018] aborda o meio da medicina, enquanto [Diakopoulos 2021] tem um âmbito mais geral, mas é consenso que um ponto deve ser levado em consideração: que o modelo seja justo. Na área de ML, embora a definição precisa de *fairness* seja alvo de discussão [Mitchell et al. 2021], o termo *fair* pode ser utilizado para definir o sistema que realiza previsões de modo eticamente justo. Por exemplo, ao tomar uma decisão judicial, espera-se que um modelo não tenha sua acurácia alterada por vieses étnico-culturais, como o caso do software judicial americano COMPAS [Dressel and Farid 2018].

Sabe-se que o desempenho de um modelo é bastante atrelado aos seus dados de treinamento. Desse modo, um dos meios de contornar esse problema é analisá-los e impedir a propagação de injustiças. No entanto, as métricas clássicas de avaliação são insuficientes, pois não levam em consideração a natureza desses dados, logo, a aplicação de metodologias como a Teoria de Resposta ao Item (TRI) podem ser úteis nesse contexto. O objetivo desse trabalho, portanto, é investigar o uso da TRI em um contexto sensível injusto, para apontar se é possível utilizá-la na identificação ou mitigação de uma injustiça. Como estudo de caso, o *dataset* Credit-G [Hofmann 1994] foi utilizado por apresentar um forte viés [Corrales-Barquero et al. 2021], no qual ocorre a subrepresentação da população feminina.

O restante deste trabalho está organizado da seguinte forma: Seção 2, apresenta a base teórica sobre o impacto da injustiça em *datasets* de ML em contextos sensíveis e explana o funcionamento da TRI; Seção 3, explica a metodologia utilizada para a elaboração deste estudo; Seção 4, traz os resultados obtidos e discussões sobre eles; Seção 5, apresenta as considerações finais do artigo.

2. Referencial Teórico

2.1. Contextos Sensíveis e Injustiça no Machine Learning

A popularização e aplicabilidade variada das ferramentas de ML trouxe à frente um problema na sua utilização em ambientes de contexto sensível, nos quais modelos impactam significativamente a vida de pessoas. Embora a segurança dos dados seja essencial, a integridade moral e a justiça de um modelo é um fator preocupante.

Um caso de injustiça aborda o *Correctional Offender Management Profiling for Alternative Sanctions* (COMPAS), um software utilizado pela corte americana desde 2000 para prever reincidência criminal de detentos em seu sistema penal. Em [Dressel and Farid 2018], demonstra-se que o sistema apresenta uma injustiça atrelada à

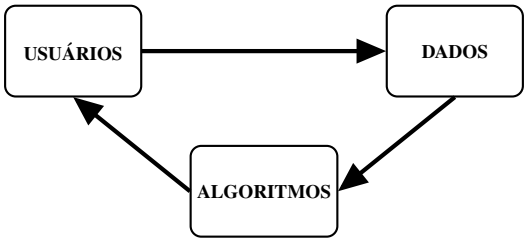


Figura 1. Ciclo de perpetuação de viés moralmente injusto

etnia do réu, dado que há grande disparidade na taxa de falsos positivos e falsos negativos entre réus afro-americanos e brancos, que são mais favorecidos pelo sistema.

Levando em conta essa disparidade e dado o período em que o modelo ficou em funcionamento, pessoas foram diretamente afetadas de forma negativa por um modelo injusto. Portanto, o campo de *fairness*, ou justiça, em ML é indispensável para criar modelos acurados e moralmente corretos, tornando seu uso fundamental na aplicação da tecnologia em contextos sensíveis.

2.2. Teoria de Resposta ao Item (TRI)

A TRI é um modelo de avaliação originado no campo da psicometria com o objetivo de avaliar respondentes de testes de forma mais precisa. Ela surgiu como um contraste à metodologia clássica de avaliação, que considera como relevante somente a quantidade de erros e acertos de um indivíduo. Seu principal objetivo é mensurar a real habilidade latente dos respondentes, levando em conta as diferentes características de cada item presente no teste [Baker 2001].

Em ML, as métricas clássicas são utilizadas de forma extensiva para avaliação de modelos. No entanto, a TRI pode trazer maior precisão nessa mensuração, além de revelar *insights* valiosos sobre as características dos dados [Martínez-Plumed et al. 2016]. Para aplicar esse método, uma simples analogia é suficiente: considera-se o *respondente* como o modelo de ML e os *itens* como instâncias de teste. Além disso, o uso desse método já se provou promissor nesse contexto de *Machine Learning* em trabalhos anteriores, vide [Cardoso et al. 2020].

Sendo o ponto central da teoria, o *item* é associado a três parâmetros que identificam a sua natureza. Considerando um item i, esses são:

- (a_i) **discriminação**: indica o quão bem aquele item diferencia respondentes com uma habilidade alta de uma baixa, da mesma forma também mede a qualidade do item para realizar avaliações;
- (b_i) **dificuldade**: constitui quanta habilidade é necessária para que o respondente acerte o item, portanto, o quão difícil ele é;
- (c_i) **adivinhação**: a chance mínima de um item ser respondido corretamente, o que indica a chance de um respondente acertá-lo ao acaso.

A TRI tem vários modelos logísticos para cálculo de habilidade. Elas variam principalmente na quantidade de parâmetros associados aos itens, sendo o Modelo Logístico de Três Parâmetros (3PL) o que utiliza todos os parâmetros supracitados. O 3PL para cálculo da resposta correta de itens dicotômicos U_{ij} – que podem estar errados (0) ou

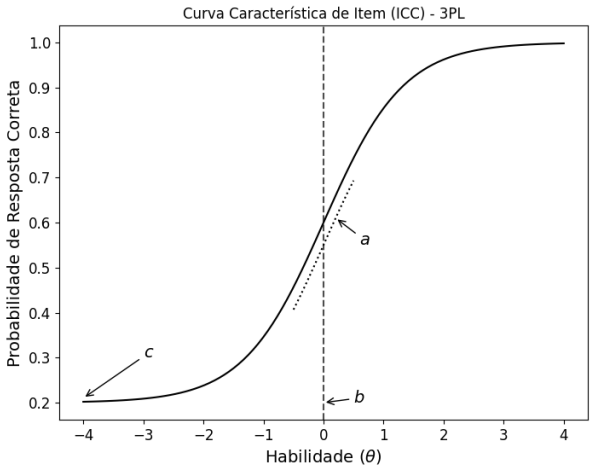


Figura 2. Uma CCI do modelo 3PL de exemplo

certos (1) – é descrito na Equação 1, onde i é o *item* e θ_j a *habilidade do respondente* j.

$$P(U_{ij} = 1|\theta_j) = c_i + (1 - c_i)\frac{1}{1 + e^{-a_i(\theta_j - b_i)}} \tag{1}$$

Por fim, para gerar uma avaliação única para cada respondente, é comum o uso do *True Score* [Lord and Wingersky 1984], uma equação composta pela soma das probabilidades de uma resposta correta e que é calculado para cada respondente, gerando um *score* único. A sua definição é encontrada na Equação 2.

$$TrueScore_j = \sum_{i=1}^{N} P(U_{ij} = 1|\theta_j) \tag{2}$$

Além disso, é possível analisar a relação entre os parâmetros de item e a probabilidade de uma resposta correta em função da habilidade, a partir da Curva Característica de Item (CCI), ou *Item Characteristic Curve* (ICC), presente na Figura 2. Veja que, no gráfico, os parâmetros estão presentes e influenciam toda sua estrutura: a discriminação (a) reflete a inclinação da curva no ponto de dificuldade; a dificuldade (b) divide o gráfico e define a habilidade necessária para acertar o item; e a adivinhação (c) indica a menor probabilidade de acerto do item.

3. Metodologia

3.1. *Dataset* Credit-G

O primeiro passo deste estudo foi a seleção do conjunto de dados. Optou-se pelo Credit-G [Hofmann 1994], um *dataset* amplamente reconhecido e estudado na literatura, o qual apresenta um viés de gênero em um contexto sensível [Corrales-Barquero et al. 2021]. O Credit-G é um *dataset* binário de análise de viabilidade de crédito para indivíduos e é composto por 700 instâncias positivas (análise de crédito *"good"*) e 300 negativas (análise de crédito *"bad"*). Cada instância é composta por 20 *features* de informações pessoais sobre os indivíduos que solicitaram crédito, tais como: idade, tempo de trabalho, etc.

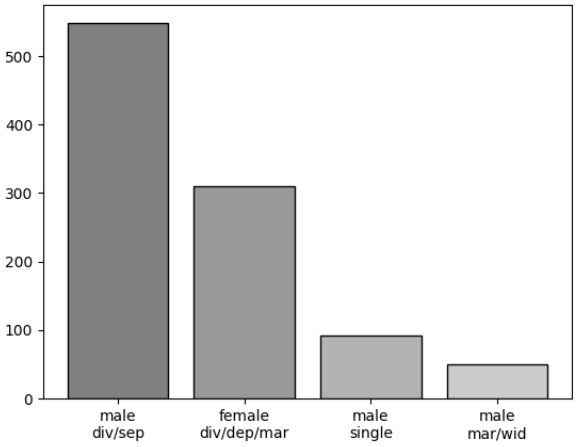


Figura 3. Histograma das categorias de instâncias de Estado Civil no Credit-G

O viés de representação pode ser encontrado na análise da *feature* categórica *"personal_status"*, que informa o sexo do indivíduo e seu estado civil, onde:

- **single** representa solteiro;
- **div** representa divorciado;
- **dep** representa dependente;
- **mar** representa casado e
- **wid** representa viúvo.

A partir da análise, nota-se que o grupo de indivíduos feminino possui 201 instâncias positivas e 109 negativas, enquanto para o masculino esses valores são 499 e 191, respectivamente. Além da quantidade de instâncias femininas ser notavelmente inferior (compõem somente 31% de todo o *dataset*), também se nota a presença de 12% a mais de instâncias negativas no conjunto feminino.

A partir da observação dos dados, nota-se um claro Viés de Representação – como definido em [Mehrabi et al. 2019] – já que a presença de instâncias do sexo feminino é significativamente menor, além de apresentar informações incompletas de seu estado civil, pois toda a população feminina é representada como uma única categoria de estado civil (divorciado, dependente e casado), enquanto as instâncias masculinas são representadas em diferentes categorias (ver Figura 3).

3.2. Estimadores da TRI para ML

Para analisar o comportamento dos dados enviesados e dos modelos de ML pelo ponto de vista da TRI, utilizou-se um método *ad-hoc*. Esse método consiste em fazer uma análise dividida, gerando dois conjuntos de parâmetros para cada item: um com influência da injustiça e outro com ela mitigada. Como o tipo de viés é conhecido, é possível tentar contorná-lo a partir da criação de instâncias artificiais para reforçar o treinamento dos modelos para o grupo sub-representado.

A criação de dados artificiais, exclusivamente do sexo feminino, foi realizada por meio da biblioteca SDV (Synthetic Data Vault) [Patki et al. 2016], onde o *GaussianCopulaSynthesizer* foi escolhido, já que ele gera dados estatisticamente alinhados com o *dataset*

provido. Além disso, para gerar os parâmetros de item da TRI, este trabalho se baseou no estudo de [Cardoso et al. 2024], o qual apresenta uma metodologia e implementação em Python para a aplicação da TRI no contexto do ML.

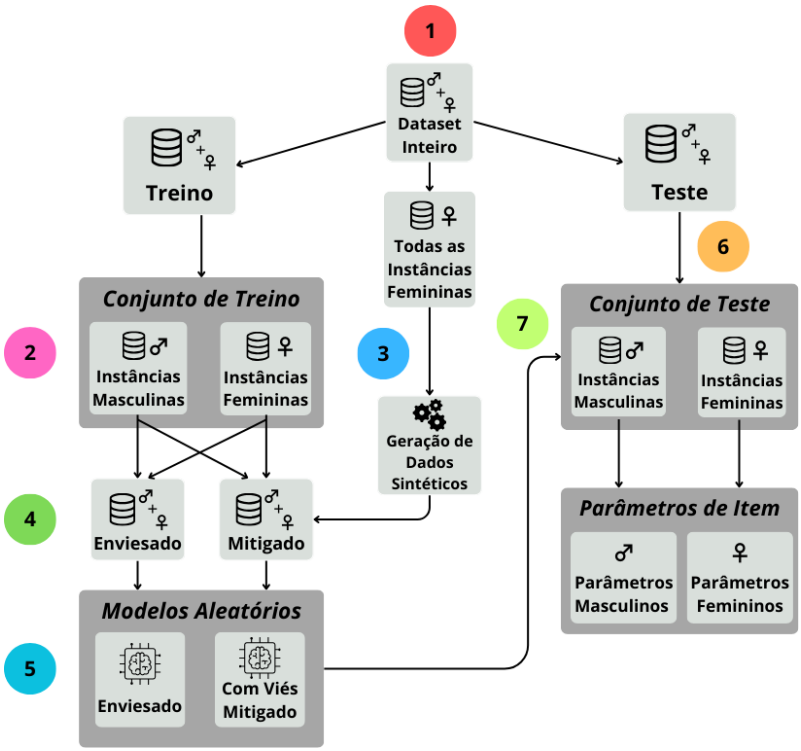


Figura 4. Diagrama para geração de parâmetros de item

A Figura 4 apresenta um diagrama da metodologia de geração dos parâmetros a serem comparados. Essa metodologia é dividida em 7 passos:

1. O *dataset* inteiro é dividido de modo estratificado em treino (70%) e teste (30%);
2. No conjunto de treino, as instâncias são divididas em dois subconjuntos pelo seu gênero, gerando o subconjunto com instâncias femininas e o subconjunto com as masculinas;
3. Utilizando todas as instâncias femininas do *dataset*, dados sintéticos, também femininos, são gerados por um método determinado. Neste trabalho, o *GaussianCopulaSynthetizer* foi utilizado;
4. A partir desses três subconjuntos (femininos, masculinos e sintéticos), dois conjuntos serão criados: o Enviesado e o Mitigado. O primeiro consiste nos dados reais do *dataset* original, assegurando o viés de representação, enquanto o segundo contém os dados originais com a adição de 300 dados sintéticos femininos, o que garante a representação balanceada da população de cada sexo (cerca de 50% de cada);
5. Em seguida, utilizando os dois conjuntos gerados no passo anterior, ocorre o treinamento de duas *pools* de modelos diferentes, cada uma contendo N modelos dos tipos *Decision Tree* (DT), *Random Forest* (RF), *Ada Boost* (ADA), *Gradient Boosting* (GB), *Bagging* (BAG), *Multi-layer Perceptron* (MLP), *k-Nearest Neighbors* (KNN), *Support Vector Machine* (SVM), *Linear Support Vector Machine* (LSVM)

e *Linear Discriminant Analysis* (LDA). A primeira *pool* corresponde aos modelos treinados com dados enviesados e a segunda com os dados mitigados.

6. Com os modelos em mãos, o conjunto de teste tem suas instâncias separadas em dois subconjuntos, da mesma forma descrita no passo 2;
7. Por fim, as duas *pools* de modelos utilizam os dados do passo anterior para gerar, separadamente, os parâmetros de item das instâncias de teste das duas categorias.

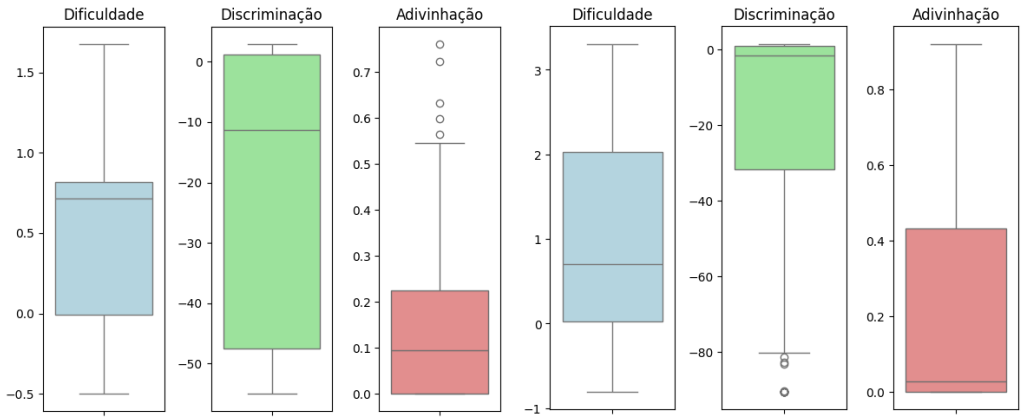
Vale lembrar que somente os dados de teste terão seus parâmetros definidos. Além disso, modelos selecionados com *Grid Search* e *Cross Validation* (dos mesmos tipos apresentados no passo 5) serão treinados e avaliados conforme a TRI. Por fim, tais parâmetros e avaliações serão analisados, comparados e discutidos, a fim de apresentar a influência de dados injustos na TRI e como seus conceitos podem ser aplicados nesse tipo de cenário.

4. Resultados e Discussão

Conforme explanado na seção 3.1, este trabalho visa utilizar a TRI como uma nova ferramenta para análise de dados e modelos enviesados de contexto sensível. Como estudo de caso, foi utilizado o *dataset* Credit-G, por apresentar o viés de representação. Esta seção apresenta os resultados obtidos após a execução da metodologia apresentada na seção 3.2[1] e está dividida em duas partes, das quais a primeira aborda diretamente, por meio da visão dos parâmetros de item, os dados enviesados e mitigados, enquanto a segunda analisa o desempenho dos modelos ajustados com *Grid-Search*.

4.1. Credit-G pela TRI

Para facilitar a análise dos dados pelas lentes da TRI, os resultados foram separados em dois conjuntos de dados: femininos e masculinos. Isso é feito tanto para o conjunto com viés quanto para o conjunto mitigado. Dessa forma, é possível comparar o comportamento dos parâmetros de item para as instâncias de cada grupo representativo.



(a) Dados Enviesados. (b) Dados Mitigados.

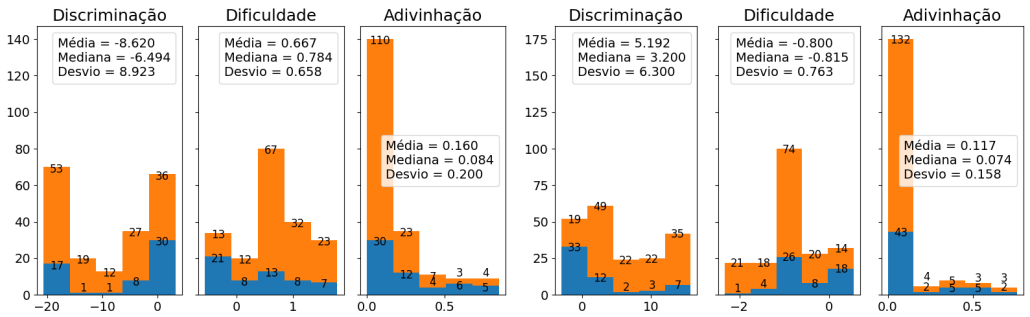
Figura 5. Parâmetros de Item – Conjunto Feminino.

A Figura 5 demonstra a distribuição dos parâmetros de item para os conjuntos de dados do grupo feminino. A diferença mais notória está na discriminação: enquanto

[1]Os códigos executados e arquivos gerados podem ser encontrados no repositório disponível em `https://github.com/Stopfield/IRT_BiasStudy`

o conjunto de dados enviesado possui uma mediana de -11.279, o conjunto de dados mitigado apresenta uma mediana de -1.675. Na TRI, itens com discriminação negativa são comumente vistos como um sinal de alerta. Pelos conceitos da discriminação (ver seção 2.2), um item com discriminação negativa significa que os respondentes de menor habilidade são os que têm maior probabilidade de acerto, e tal condição é contraintuitiva.

Na psicometria, esses valores negativos costumam acontecer devido a inconsistências como ambiguidade ou má formulação do item. Mas também é possível devido a um ensino ruim, i.e., se o ensino que o aluno teve foi ruim ou confuso, o aluno pode responder um item de forma esperada segundo o ensino, porém errada para o teste. No âmbito do ML, isso pode ser o indicativo de viés nos dados, de forma que os dados de treino (ensino) estão confusos para os modelos (alunos), como resultado disso os modelos são induzidos ao erro.



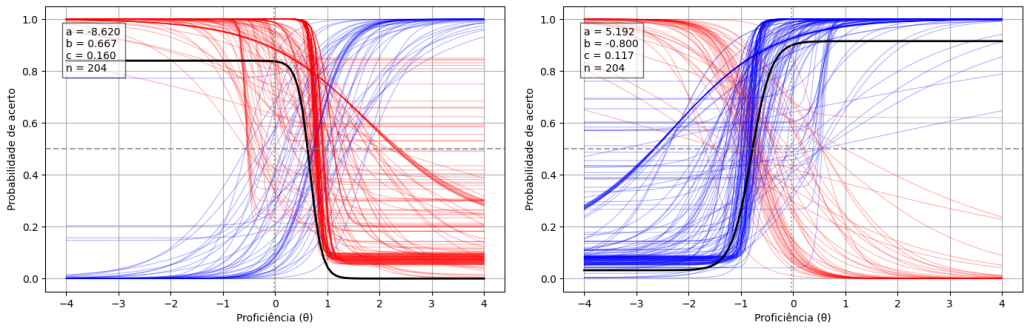
(a) Dados Enviesados. (b) Dados Mitigados.

Figura 6. Parâmetros de Item – Conjunto Masculino.

Conforme demonstrado, a discriminação aumentou consideravelmente para os dados com viés mitigado por dados sintéticos. Para comparar, a Figura 6 apresenta o comportamento dos parâmetros de item para o conjunto masculino, na qual a cor amarela representa a classe majoritária e a azul, a classe minoritária. Nota-se que a discriminação se tornou positiva após a mitigação do viés, de forma que a quantidade de instâncias com discriminação positiva aumentou de 21.56% para 76.96%. Isso indica que o "ensino" das instâncias masculinas também melhorou.

Isso também se reflete na dificuldade, cuja mediana diminuiu de 0.784 para -0.815, o que significa que esse conjunto se tornou mais fácil de classificar. Para o conjunto feminino, no entanto, não houve grande variação de dificuldade (0.718 com viés, 0.704 mitigado). A Figura 7 ilustra como esse impacto é visto na TRI a partir das Curvas Características de cada instância masculina, na qual as linhas vermelhas representam instâncias com discriminação negativa, as azuis, com discriminação positiva, e a preta, os valores médios.

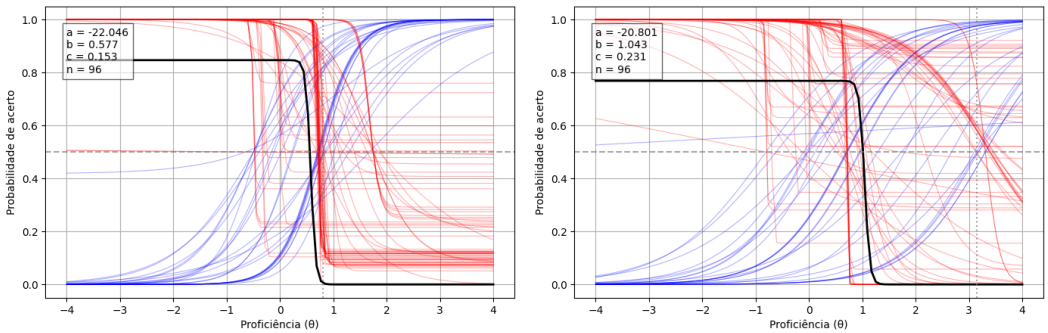
Para o conjunto feminino, a diferença não é tão expressiva (ver Figura 8). Apesar da diminuição dos valores médios, não houve mudança na quantidade de instâncias com discriminação negativa, pois para ambos os conjuntos (enviesado e mitigado) existem 69 instâncias com valores negativos. Entretanto, é possível notar uma menor inclinação das curvas de discriminação negativa, como já era esperado devido à diminuição de seus valores médios. Isso pode ser entendido como um indicativo de que os dados sintéticos



(a) Dados Enviesados. (b) Dados Mitigados.

Figura 7. CCIs – Conjunto Masculino.

ajudaram de fato a mitigar o viés, apesar de ainda não o solucionar. Um trabalho futuro pode explorar as instâncias geradas sinteticamente e medir sua qualidade com a TRI.



(a) Dados Enviesados. (b) Dados Mitigados.

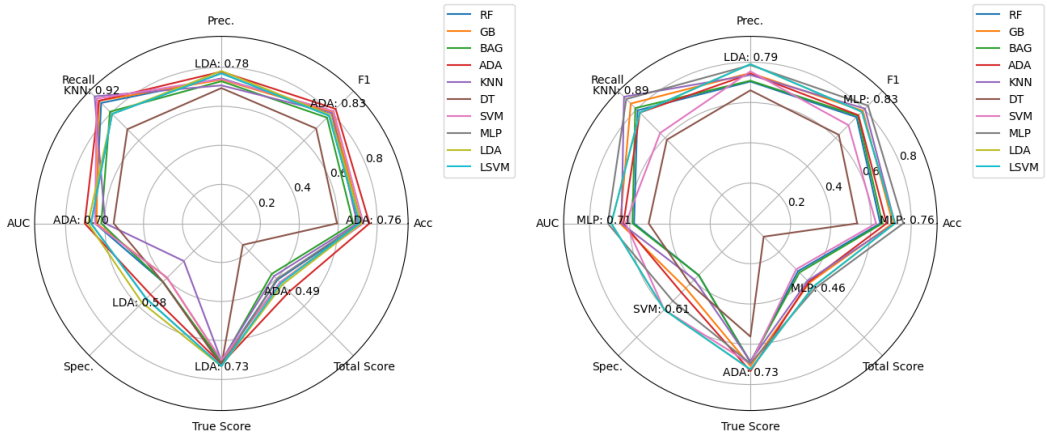
Figura 8. CCIs – Conjunto Feminino.

4.2. Avaliação dos Modelos

Aqui será explorado de forma mais aprofundada o resultado dos modelos selecionados com o *GridSearch* e *CrossValidation* com os dados enviesados e mitigados pela perspectiva da TRI. Para isso, foram calculadas métricas clássicas de ML e os estimadores da TRI: o *True Score* e o *Total Score*.

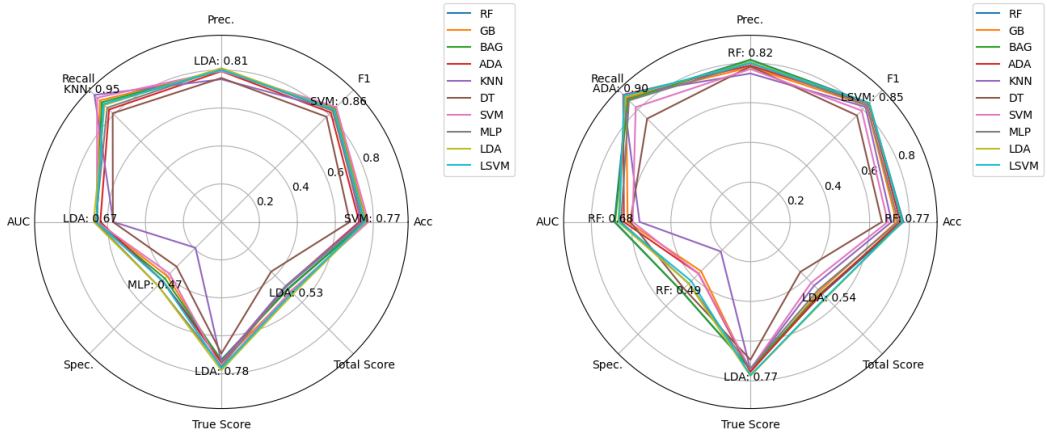
A Figura 9 apresenta o resultado dos modelos para o conjunto de instâncias femininas nos cenários enviesado e mitigado, na qual se destaca os modelos que obtiveram maior *score* para cada métrica. Em suma, não houve grande mudança no desempenho médio dos modelos, apenas se observa que ocorreu alteração de posição deles. E.g., para o conjunto enviesado o modelo com maior score de *F1* foi o ADA com 0.83, paralelamente, o modelo com maior *F1* para o conjunto mitigado foi o MLP com o mesmo resultado. Apesar de a mitigação ter resultado em uma melhoria na discriminação, conforme demonstrado na seção 4.1, não foi suficiente para que surtisse impacto positivo suficiente no desempenho dos modelos.

A mesma situação ocorre com os conjuntos de instâncias masculinas. Não é evidente nenhuma variação significativa no desempenho geral dos modelos ao treiná-los com



(a) Dados Enviesados. (b) Dados Mitigados.

Figura 9. Modelos – Conjunto Feminino.



(a) Dados Enviesados. (b) Dados Mitigados.

Figura 10. Modelos – Conjunto Masculino.

dados enviesados ou mitigados. A diferença mais significativa só aparece ao comparar os conjuntos de dados masculinos aos femininos, nos quais o *Total Score* da TRI, precisamente no conjunto de dados mitigado, aumenta sua pontuação de 0.46 para 0.54.

A grande vantagem da aplicação da TRI na análise de resultados de teste está na capacidade de avaliação individual de cada item. Enquanto as métricas gerais não apontam melhoria na performance dos modelos, o comportamento das CCIs mostra que há alguma diferença na forma como o dado é visto pelos modelos. Por exemplo, isso pode indicar que os modelos treinados com dados mitigados são mais confiáveis, pois foram testados com instâncias de maior qualidade de avaliação (discriminação positiva, ver Figura 7).

5. Considerações Finais

Este trabalho analisou o comportamento das métricas da TRI em um cenário de contexto sensível sob influência de um viés de representação. Para isso, comparou-se uma condição

de enviesamento a outra em que esse viés foi contornado por meio da geração de dados sintéticos. A hipótese central é que, embora as métricas clássicas ainda sejam fundamentais na avaliação de modelos de ML, uma abordagem centrada nas instâncias permite uma maior compreensão da natureza dos dados e como eles influenciam seu desempenho.

O *dataset* escolhido para análise foi o Credit-G, que possui uma sub-representação da população feminina. A partir da comparação das métricas enviesadas e mitigadas para os grupos masculino e feminino, observa-se a utilidade da teoria na análise desse viés. O impacto dessa distorção se manifesta principalmente nos valores médios de discriminação em ambos os grupos, que indicam uma má formação dos dados, sobretudo para o feminino.

Outro resultado interessante é o aumento dos valores de discriminação para o grupo masculino, que teve sua média de discriminação elevada de -8.62 para 5.192 (Figura 6) de forma a tornar a maioria de suas instâncias em itens com discriminação superior a zero. Isso pode estar relacionado à melhor representação do grupo, o que resultou em um entendimento mais razoável da estrutura desses dados. Ainda que não tenha impactado expressivamente o desempenho dos modelos, pode-se inferir que eles se tornaram mais confiáveis, em virtude do seu treinamento com instâncias de maior qualidade.

Também é possível observar que a tentativa de mitigação do grupo feminino não gerou resultados significativos, pois não houve aumento relevante na discriminação das instâncias desse grupo. Mesmo após sua mitigação, a maioria dos itens ainda apresenta uma discriminação negativa, sugerindo a possibilidade de o viés de representação não ser o único enviesamento presente. Além disso, isso também pode indicar que os dados femininos do conjunto original são profundamente mal estruturados e que, portanto, mais dados reais são necessários para avaliação.

Desse modo, pode-se dizer que a TRI é uma potencial aliada na identificação de cenários enviesados. Por considerar a natureza das instâncias, seu uso permite gerar mais informação sobre os dados e o desempenho dos modelos, o que facilita a detecção de tendências injustas no contexto. Do mesmo modo, os artefatos da teoria apontam para a qualidade inferior dos dados no cenário enviesado e como uma estratégia de mitigação pode aprimorá-los, contribuindo para a criação de modelos mais confiáveis.

Por outro lado, o trabalho também apresentou limitações, pois o método de mitigação utilizado não é robusto o suficiente para contornar o viés apresentado. Entretanto, mesmo com uma abordagem simples de mitigação, é possível observar uma alteração significativa na natureza dos dados, o que sugere a necessidade de investigações futuras do seu comportamento em estratégias mais robustas.

Apesar dos resultados favoráveis ao seu uso, mais testes são necessários para definir uma metodologia robusta que utilize a TRI na identificação de vieses moralmente injustos. Desse modo, um trabalho futuro pode utilizar conjuntos de dados e estratégias de mitigação diferentes para a observação desses parâmetros.

Referências

[Baker 2001] Baker, F. B. (2001). *The basics of item response theory*. ERIC.

[Cardoso et al. 2024] Cardoso, L. F., Ribeiro Filho, J. d. S., Santos, V. C., Kawasaki Francês, R. S., and Alves, R. C. (2024). Standing on the shoulders of giants. In *Brazilian*

Conference on Intelligent Systems, pages 416–430. Springer.

[Cardoso et al. 2020] Cardoso, L. F., Santos, V. C., Francês, R. S. K., Prudêncio, R. B., and Alves, R. C. (2020). Decoding machine learning benchmarks. In *Brazilian conference on intelligent systems*, pages 412–425. Springer.

[Corrales-Barquero et al. 2021] Corrales-Barquero, R., Marín-Raventós, G., and Barrantes, E. G. (2021). A review of gender bias mitigation in credit scoring models. In *2021 Ethics and Explainability for Responsible Data Science (EE-RDS)*, pages 1–10.

[Diakopoulos 2021] Diakopoulos, N. (2021). Algorithm auditing: Managing the legal, ethical, and technological risks of artificial intelligence, machine learning, and associated algorithms. *IEEE Transactions on Technology and Society*, 2(3):143–148.

[Dressel and Farid 2018] Dressel, J. and Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1).

[Hofmann 1994] Hofmann, H. (1994). Uci machine learning repository: Statlog (german credit data) data set. *Institut fur Statistik und "Okonometrie Universit" at Hamburg*.

[Lord and Wingersky 1984] Lord, F. M. and Wingersky, M. S. (1984). Comparison of irt true-score and equipercentile observed-score"equatings". *Applied psychological measurement*, 8(4):453–461.

[Martínez-Plumed et al. 2016] Martínez-Plumed, F., Prudêncio, R. B. C., Martínez-Usó, A., and Hernández-Orallo, J. (2016). Making sense of item response theory in machine learning. In *Proceedings of the 22nd European Conference on Artificial Intelligence (ECAI 2016)*, pages 1140–1148.

[Mehrabi et al. 2019] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2019). A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*.

[Mitchell et al. 2021] Mitchell, S., Potash, E., Barocas, S., D'Amour, A., and Lum, K. (2021). Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, 8:141–163.

[Patki et al. 2016] Patki, N., Wedge, R., and Veeramachaneni, K. (2016). The synthetic data vault. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 399–410.

[Sengupta and Dave 2022] Sengupta, S. and Dave, V. (2022). Predicting applicable law sections from judicial case reports using legislative text analysis with machine learning. *Journal of Computational Social Science*, 5:503–516.

[Sidey-Gibbons and Sidey-Gibbons 2019] Sidey-Gibbons, J. A. and Sidey-Gibbons, C. J. (2019). Machine learning in medicine: a practical introduction. *BMC medical research methodology*, 19:1–18.

[Vayena et al. 2018] Vayena, E., Blasimme, A., and Cohen, I. G. (2018). Machine learning in medicine: Addressing ethical challenges. *PLOS Medicine*, 15(11):e1002689.