Análise de Justiça em Modelos de Aprendizado de Máquina

Daniel Oliveira dos Santos[*]
Amarildo Maia Rolim Filho[†]
Valdeclebio Farrapo Costa[‡]
Lucas Bezerra de Sena[§]

1

Resumo. *Este artigo apresenta uma análise abrangente da imparcialidade em modelos de aprendizado de máquina usando o UCI Adult Dataset. O estudo se concentra em mitigar vieses relacionados a atributos sensíveis, como raça e gênero, reduzindo a dimensionalidade do conjunto de dados. Avaliamos o desempenho e a imparcialidade de três modelos populares de aprendizado de máquina — Regressão Logística, Floresta Aleatória e Gradient Boosting — Na primeira etapa os modelos foram treinado com objetivo de classificar se a renda é menor, igual a US$ 50.000 ou maior que US$ 50.000. A variável de destino é convertida para binária, porem foi usado outro modelo que tem o ituido de corrigir a diferença em ajustar e otimizar os vieses dos dados sensiveis corrigindo essa falha, que os resultados tanto das métricas e das justiça venha ser iguais. para isso foi usado o modelo Rede Neural Generativa Adversária conhecida como Rede Neural Feedforward Densa.*

1. Introdução

A aplicação crescente de modelos de aprendizado de máquina em atividades empresariais levou a avanços significativos em várias áreas, como análise de sentimentos e classificação de áudio. Apesar desses avanços, há cenários em que a implantação de modelos de aprendizado de máquina requer atenção cuidadosa para prevenir potenciais vieses que podem levar à discriminação e efeitos adversos para a empresa. Garantir justiça e mitigar vieses em modelos de aprendizado de máquina é essencial para manter a integridade e os padrões éticos das operações comerciais.

O foco na justiça no aprendizado de máquina ganhou destaque, com novas técnicas sendo desenvolvidas para detectar e mitigar vieses nesses modelos [0]. Isso garante que os aplicativos de aprendizado de máquina permaneçam éticos e confiáveis, promovendo confiança e resultados equitativos em diferentes demografias.

Estudos recentes destacaram a importância de abordar o viés em modelos de aprendizado de máquina, particularmente em aplicações que envolvem atributos sensíveis. Uma dessas abordagens é o Protected Attribute Suppression System (PASS), que atenua o viés no reconhecimento facial reduzindo a codificação de atributos protegidos, como gênero e tom de pele, sem exigir o retreinamento de ponta a ponta de todo o modelo. Outro trabalho significativo investiga a redução do viés usando o Ensemble Learning no

[*]Universidade Federal do Ceará (UFC-Ceará)
[†]Universidade Federal do Ceará (UFC-Ceará)
[‡]Universidade Federal do Ceará (UFC-Ceará)
[§]Universidade Federal do Ceará (UFC-Ceará)

conjunto de dados UCI Adult, com foco no viés de gênero na previsão salarial e empregando a divergência de Kullback-Leibler (KL) para medir o viés [0].

Neste trabalho, comparamos a imparcialidade de modelos treinados com e sem atributos protegidos usando o UCI Adult Dataset, um benchmark amplamente usado para avaliar a imparcialidade em modelos de aprendizado de máquina. Ao analisar o impacto da inclusão ou exclusão de atributos sensíveis como raça e gênero, buscamos identificar vieses e avaliar sua influência no desempenho do modelo. Nossas descobertas contribuem para o discurso mais amplo sobre a criação de sistemas de aprendizado de máquina mais justos e transparentes, alinhando avanços tecnológicos com padrões éticos em operações comerciais.

2. Fundamentação Teórica

Esta seção fornece o conhecimento de base essencial necessário para entender os conceitos e metodologias empregados neste trabalho. Neste trabalho, adotamos três noções de justiça propostas na literatura: Oportunidade Igualitária, Igualdade Preditiva e Paridade Demográfica [0].

2.1. Equal Opportunity (Oportunidade Igualitária)

Equal Opportunity foca na igualdade das taxas de verdadeiros positivos (True Positive Rate - TPR) entre diferentes grupos sensíveis. Em termos práticos, isso significa que o modelo deve ter a mesma capacidade de identificar corretamente os eventos positivos para todos os grupos. Tem como objetivo de Garantir que a taxa de acertos para eventos positivos seja equivalente entre todos os grupos, independentemente do grupo ao qual o indivíduo pertence. A importância dessa métrica é crucial quando os custos e benefícios de erros de classificação positivos são significativos e devem ser tratados igualmente para todos os grupos.

Taxa de Verdadeiros Positivos (TPR) para um grupo g:

$$\text{TPR}_g = \frac{\text{TP}_g}{\text{TP}_g + \text{FN}_g}$$

Para garantir Equal Opportunity entre grupos g1,g2,g3...,gk:

$$\text{TPR}_{g_1} = \text{TPR}_{g_2} = \ldots = \text{TPR}_{g_k}$$

2.2. Predictive Equality (Igualdade Preditiva)

Predictive Equality busca a igualdade nas taxas de falsos positivos (False Positive Rate - FPR) entre diferentes grupos. Isso implica que a probabilidade de um indivíduo receber uma previsão positiva, quando na verdade ele é negativo, deve ser a mesma para todos os grupos. Tem como objetivo de Reduzir disparidades nas previsões incorretas entre grupos sensíveis. A importância dessa métrica é importante em contextos onde o custo de falsos positivos pode ter impacto negativo significativo, como em sistemas de justiça criminal.

Taxa de Falsos Positivos (FPR) para um grupo g:

$$\text{FPR}_g = \frac{\text{FP}_g}{\text{FP}_g + \text{TN}_g}$$

Para garantir Predictive Equality entre grupos g1,g2,...,gk:

$$\text{FPR}_{g_1} = \text{FPR}_{g_2} = \ldots = \text{FPR}_{g_k}$$

2.3. Demographic Parity (Paridade Demográfica)

Demographic Parity exige que a proporção de previsões positivas feitas pelo modelo seja a mesma para todos os grupos sensíveis, independentemente dos verdadeiros rótulos. Em outras palavras, a distribuição das previsões positivas deve ser equitativa entre diferentes grupos. Tem como objetivo de assegurar que todos os grupos sensíveis sejam tratados igualmente em termos de proporção de previsões positivas. A importância dessa métrica é fundamental quando a equidade nas oportunidades de receber uma previsão positiva é um objetivo primário, como em decisões de contratação ou concessão de empréstimos.

Proporção de Previsões Positivas (PPV) para um grupo g:

$$\text{PPV}_g = \frac{\text{TP}_g + \text{FP}_g}{N_g}$$

Para garantir Demographic Parity entre grupos g1,g2,g3...,gk:

$$\text{PPV}_{g_1} = \text{PPV}_{g_2} = \ldots = \text{PPV}_{g_k}$$

3. Autoencoder

3.1. Teoria

Objetivo: O autoencoder é um tipo de rede neural usada principalmente para aprender uma representação compacta dos dados. É um modelo de aprendizado não supervisionado que aprende a codificar os dados de entrada em uma forma compacta e, em seguida, decodifica essa forma compacta para reconstruir os dados de entrada. O objetivo é minimizar a diferença entre os dados de entrada e a reconstrução.

Arquitetura:

- **Codificador (Encoder):** A parte do autoencoder que transforma a entrada em uma representação de menor dimensionalidade. Pode ser visto como uma rede neural com camadas que reduzem a dimensionalidade dos dados.
- **Decodificador (Decoder):** A parte que tenta reconstruir a entrada original a partir da representação compacta. Pode ser visto como uma rede neural com camadas que expandem a dimensionalidade de volta ao tamanho original dos dados.

Função de Perda: O autoencoder usa a perda de erro quadrático médio (MSE) para medir a diferença entre a entrada original e a reconstrução. A função de perda é dada por:

$$\text{Loss} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x}_i)^2 \tag{1}$$

Onde x_i é o valor original e $\hat{x}_i$ é o valor reconstruído.

3.2. Treinamento

- **Pré-processamento:** Normalize os dados de entrada para ter valores entre 0 e 1, que é comum quando se usa uma função de ativação sigmoid na camada de saída.
- **Construção do Modelo:** Defina a arquitetura do autoencoder, especificando as camadas codificadoras e decodificadoras.
- **Compilação:** Compile o modelo com o otimizador, como Adam, e defina a função de perda (MSE).
- **Treinamento:** Treine o autoencoder usando os dados de entrada. O modelo tentará minimizar a perda ajustando os pesos das camadas para melhorar a qualidade da reconstrução.
- **Avaliação:** Após o treinamento, avalie o desempenho do autoencoder verificando a capacidade de reconstruir os dados de entrada.

4. Classificador

4.1. Teoria

Objetivo: O classificador é um modelo de aprendizado supervisionado projetado para prever a classe ou rótulo de uma amostra com base em suas características. No seu caso, ele está sendo usado para prever se a variável alvo `income` pertence à classe positiva ou negativa.

Arquitetura:

- **Camadas Ocultas:** O classificador contém camadas densas (Dense) que aprendem representações de características complexas dos dados. A função de ativação ReLU é comum nessas camadas para introduzir não linearidade.
- **Regularização:** Regularizadores como L2 e Dropout são usados para evitar overfitting, garantindo que o modelo generalize bem para dados não vistos.
- **Camada de Saída:** A camada final usa a função de ativação sigmoide para gerar uma probabilidade entre 0 e 1, que pode ser interpretada como a probabilidade de uma amostra pertencer à classe positiva.

Função de Perda: A função de perda utilizada é a `binary_crossentropy`, adequada para problemas de classificação binária. É dada por:

$$\text{Loss} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)\right] \tag{2}$$

Onde y_i é o rótulo verdadeiro e $\hat{y}_i$ é a probabilidade prevista.

4.2. Treinamento

- **Pré-processamento:** Normalizar ou padronizar os dados e realizar one-hot encoding, se necessário.
- **Construção do Modelo:** Defina a arquitetura do classificador com camadas densas, funções de ativação e técnicas de regularização.
- **Compilação:** Compile o modelo com o otimizador (por exemplo, Adam), a função de perda (`binary_crossentropy`) e métricas de avaliação (por exemplo, accuracy).

- **Treinamento:** Treine o classificador usando os dados de treino. O modelo ajustará seus pesos para minimizar a perda e melhorar a precisão das previsões.
- **Avaliação:** Avalie o desempenho do classificador nos dados de teste usando métricas como acurácia, precisão, recall e F1-score.

5. rede neural feedforward artificial (FAN)

5.1. Características

- **Camadas Densas:**
  - **Camadas de Entrada:** A primeira camada recebe as entradas, que são os dados de treinamento.
  - **Camadas Ocultas:** Camadas densas que transformam as entradas em representações intermediárias. Essas camadas geralmente utilizam funções de ativação não lineares, como ReLU, para capturar complexidades nos dados.
  - **Camada de Saída:** A última camada gera as previsões. No seu caso, é uma camada com uma função de ativação sigmoide para a classificação binária.
- **Função de Ativação:**
  - **ReLU (Rectified Linear Unit):** Usada nas camadas ocultas para introduzir não linearidade e permitir que a rede aprenda representações mais complexas dos dados.
  - **Sigmoide:** Usada na camada de saída para prever probabilidades entre 0 e 1, apropriada para problemas de classificação binária.
- **Regularização:**
  - **L2 Regularization:** Adiciona uma penalização aos pesos do modelo para evitar overfitting.
  - **Dropout:** Desativa aleatoriamente uma porcentagem das unidades durante o treinamento para ajudar a prevenir overfitting.
- **Função de Perda:**
  - **Binary Cross-Entropy:** Utilizada para medir a diferença entre as previsões do modelo e as classes verdadeiras. É apropriada para problemas de classificação binária.
- **Otimização:**
  - **Adam Optimizer:** Um otimizador eficiente que ajusta os pesos da rede para minimizar a função de perda durante o treinamento.

5.2. Aplicação do Modelo

- **Autoencoder:**
  - **Objetivo:** Reduzir a dimensionalidade dos dados e aprender uma representação compacta dos dados de entrada. A reconstrução dos dados é usada como uma métrica para otimização.
- **Classificador:**
  - **Objetivo:** Prever a classe de uma amostra, que no seu caso é a variável alvo `income`. O modelo é treinado para minimizar a perda de classificação e melhorar a acurácia das previsões.

6. Trabalho Relacionado

A questão da justiça em modelos de aprendizado de máquina tem atraído cada vez mais atenção na comunidade de pesquisa, com várias abordagens e metodologias sendo propostas para mitigar vieses e promover decisões mais equitativas. Barocas et al. [0] propõem o Protected Attribute Suppression System (PASS) para reduzir o viés no reconhecimento facial suprimindo a codificação de atributos protegidos, como gênero e tom de pele. Seu método opera em descritores faciais de redes pré-treinadas, obtendo alta precisão de verificação sem a necessidade de retreinamento de ponta a ponta. Ao contrário do estudo de Dhar et al. [0], nosso estudo se concentra na avaliação da discriminação no UCI Adult Dataset. Analisamos o viés usando métricas de justiça, como Diferença de Paridade Demográfica e Diferença de Probabilidades Equalizadas, fornecendo uma investigação detalhada sobre estratégias de mitigação de viés dentro deste contexto específico.

Vários estudos abordaram o viés em modelos de aprendizado de máquina sobre atributos sensíveis. O trabalho de Girhepuje [0] examina o viés de gênero na predição salarial usando o Ensemble Learning no conjunto de dados UCI Adult, revelando disparidades significativas e maior viés em modelos baseados em árvore. Em contraste, nosso estudo avalia a discriminação no conjunto de dados UCI Adult analisando duas métricas de justiça: Diferença de Paridade Demográfica e Diferença de Probabilidades Equalizadas, investigando especificamente vieses relacionados à raça e gênero para uma análise abrangente da mitigação de viés.

7. Metodologia

Nesta seção, você encontrará uma visão geral da nossa abordagem para avaliar a imparcialidade em modelos de machine learning. Detalhamos o conjunto de dados usado, o design experimental e as métricas específicas para desempenho e imparcialidade.

7.1. Conjunto de Dados

O conjunto de dados UCI Adult, também conhecido como conjunto de dados "Census Income", é um benchmark popular para machine learning. Ele vem do Censo dos EUA de 1994 e é usado para prever se a renda de uma pessoa excede US$ 50.000 por ano com base em informações demográficas, como idade, educação, ocupação, raça e gênero. O conjunto de dados contém 48.842 entradas com 14 atributos. A variável de destino é binária, indicando se a renda é menor ou igual a US$ 50.000 ou maior que US$ 50.000. O conjunto de dados UCI Adult é amplamente usado para tarefas de classificação, análise de imparcialidade para avaliar e mitigar vieses e benchmarking de diferentes algoritmos de aprendizado de máquina.

7.2. Design Experimental

O principal objetivo do design experimental visam garantir que os modelos de aprendizado de máquina não apenas performem bem em termos de acurácia, mas também sejam justos e imparciais ao lidar com dados sensíveis. A análise ajuda a identificar e mitigar vieses que podem impactar negativamente a equidade das previsões, contribuindo para o desenvolvimento de sistemas mais justos e éticos.

7.3. Modelos e Métricas

Para avaliar o desempenho e a imparcialidade do sistema, utilizamos três modelos amplamente conhecidos de aprendizado de máquina: Regressão Logística, Floresta Aleatória e Gradiente Boosting. Esses modelos foram selecionados por sua versatilidade, aplicabilidade em uma ampla gama de cenários e diferenças na forma como abordam a modelagem de dados. Enquanto a Regressão Logística é um modelo linear, a Floresta Aleatória e o Gradiente Boosting são modelos baseados em árvores, conhecidos por capturarem interações complexas entre variáveis.

A avaliação de desempenho foi realizada com base em métricas clássicas de classificação, incluindo:

- **Acurácia**: a proporção de previsões corretas sobre o total de previsões.
- **Precisão**: a proporção de verdadeiros positivos entre todos os exemplos classificados como positivos.
- **Recall**: a proporção de verdadeiros positivos sobre todos os exemplos que realmente são positivos.
- **F1-score**: a média harmônica entre precisão e recall, proporcionando uma medida equilibrada entre as duas.

Além do desempenho preditivo, avaliamos a imparcialidade dos modelos através de métricas específicas de *fairness*, como:

- **Paridade Demográfica (Demographic Parity)**: verifica se a taxa de previsão positiva é a mesma entre diferentes grupos, sem levar em conta o rótulo verdadeiro.
- **Oportunidade Igualitária (Equal Opportunity)**: garante que a taxa de verdadeiros positivos seja igual entre diferentes grupos. Ou seja, a probabilidade de um indivíduo positivo ser corretamente classificado como tal deve ser a mesma independentemente do grupo.
- **Igualdade Preditiva (Predictive Equality)**: assegura que a taxa de falsos positivos seja semelhante entre os grupos. Essa métrica é crucial para minimizar o impacto desproporcional de erros em certos grupos demográficos.

Essas métricas fornecem uma visão abrangente do desempenho preditivo e da equidade dos modelos, permitindo identificar e mitigar possíveis vieses presentes nas predições.

8. Resultados

Nesta seção, apresentamos e discutimos os resultados obtidos a partir dos experimentos. As tabelas abaixo mostram a performance dos modelos e as métricas de justiça associadas.

8.1. Resultados do Desempenho dos Modelos

Accuracy: 0.87

	precision	recall	f1-score	support
0	0.89	0.95	0.92	7640
1	0.79	0.61	0.69	2360
accuracy			0.87	10000
macro avg	0.84	0.78	0.80	10000
weighted avg	0.86	0.87	0.86	10000

Figura 1. Métricas do Modelo Gradient Boosting Classifier.

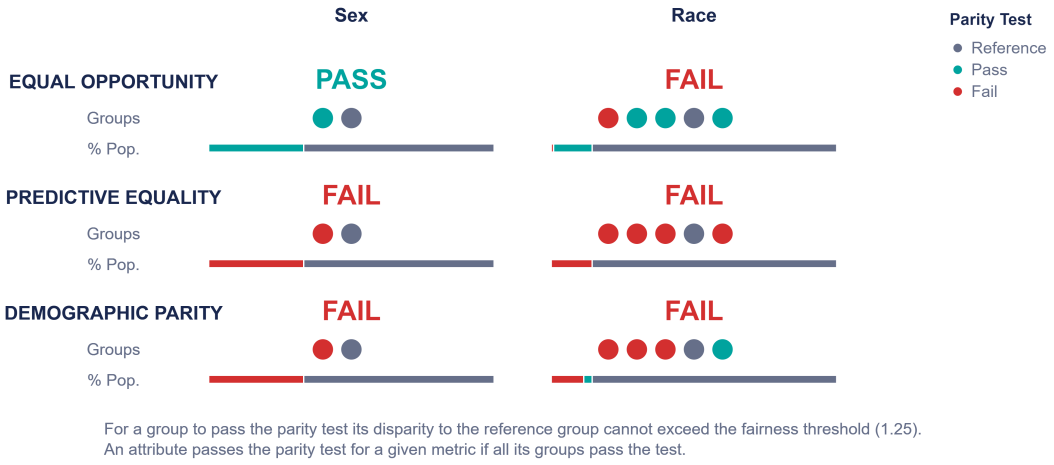

Figura 2. Gráfico de Desempenho do Modelo Gradient Boosting Classifier.

Random Forest Accuracy: 0.8563

	precision	recall	f1-score	support
0	0.87	0.96	0.91	7640
1	0.80	0.53	0.63	2360
accuracy			0.86	10000
macro avg	0.83	0.74	0.77	10000
weighted avg	0.85	0.86	0.85	10000

Figura 3. Métricas do Modelo Random Forest.

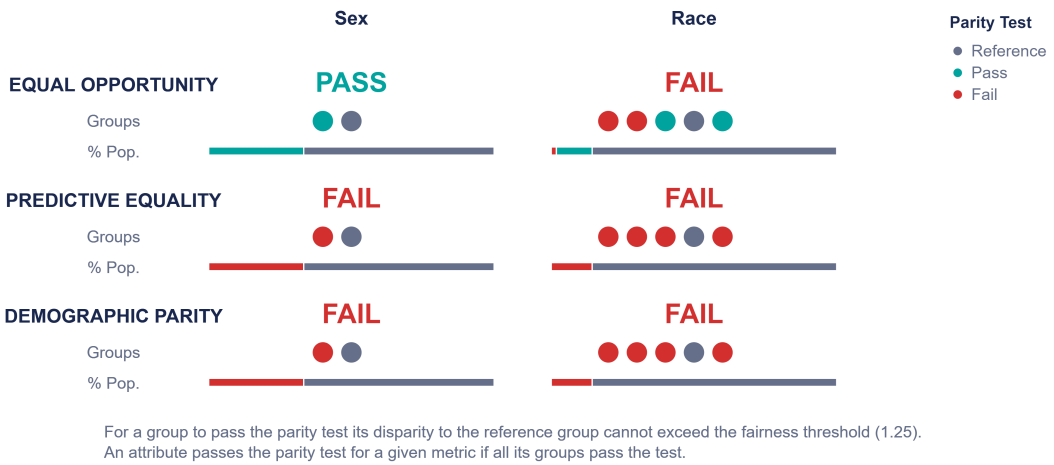

Figura 4. Gráfico de Desempenho do Modelo Random Forest.

Accuracy: 0.85

	precision	recall	f1-score	support
0	0.88	0.93	0.90	7640
1	0.72	0.59	0.65	2360
accuracy			0.85	10000
macro avg	0.80	0.76	0.78	10000
weighted avg	0.84	0.85	0.84	10000

Figura 5. Métricas do Modelo Regressão Logística.

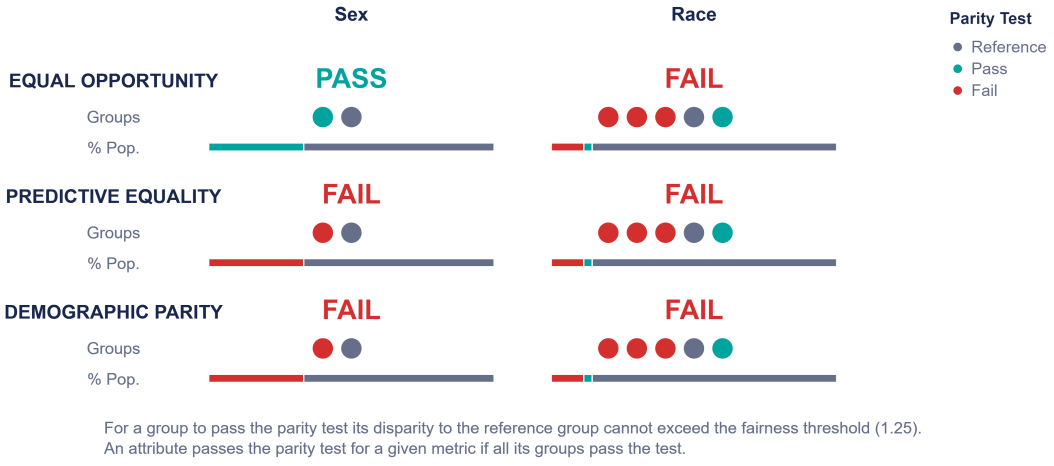

Figura 6. Gráfico de Desempenho do Modelo Regressão Logística.

8.2. Rede Neural Feedforward Artificial (FAN)

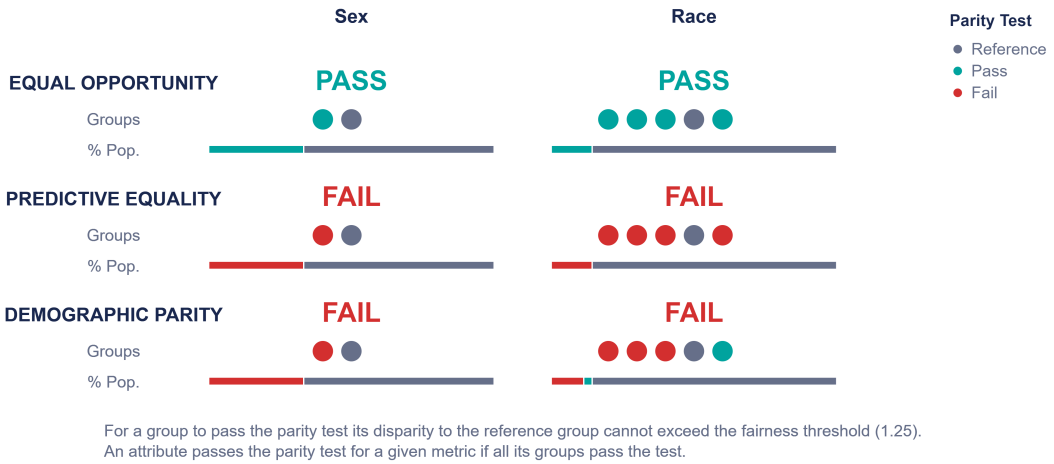

Figura 7. Gráfico de Desempenho do Modelo Rede Neural Feedforward Artificial.

9. Discussão

O modelo apresentou sucesso em Equal Opportunity tanto para raça quanto para gênero. Isso significa que, independentemente da raça ou do gênero, o modelo conseguiu identificar corretamente as pessoas que realmente ganham mais de 50 mil dólares por ano, mostrando que não houve discriminação na taxa de verdadeiros positivos. Esse é um resultado muito positivo, pois indica que o modelo não está subestimando a renda de certos grupos, o que é uma preocupação comum em análises socioeconômicas. O fato de o modelo ter sucesso em Equal Opportunity tanto para raça quanto para gênero é um indicador muito forte de que ele trata diferentes grupos de maneira justa em termos de verdadeiros positivos. Isso é crucial em modelos de renda, já que garante que o modelo não está sendo discriminatório ao identificar corretamente as pessoas de diferentes raças e gêneros que ganham mais de 50 mil dólares. Contudo, para garantir uma justiça completa, o modelo ainda deve ser avaliado e, possivelmente, aprimorado em outras métricas, como a Paridade Demográfica e a Igualdade Preditiva, especialmente se houver diferenças na distribuição de falsos positivos entre os grupos. Essa análise completa demonstra a importância de utilizar múltiplas métricas de justiça para garantir que o modelo de machine learning não está introduzindo viéses prejudiciais em suas previsões, especialmente em problemas socioeconômicos sensíveis como o predito pelo UCI Adult dataset.

10. Conclusão

O estudo revelou que a inclusão de características sensíveis no treinamento de modelos de aprendizado de máquina pode melhorar a justiça em termos de Paridade Demográfica e Probabilidades Equalizadas. No entanto, é importante considerar o equilíbrio entre justiça e desempenho, especialmente quando se trata de atributos sensíveis que podem impactar a equidade das decisões do modelo. Nossos resultados destacam a necessidade de técnicas de mitigação de viés eficazes que considerem tanto a justiça quanto o desempenho dos modelos.

Referências

Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and machine learning: Limitations and opportunities*. MIT Press.

Caton, S., & Haas, C. (2024). Fairness in machine learning: A survey. *ACM Computing Surveys*, 56(7), 1–38.

Chaves, I. C., Martins, A. D. F., Praciano, F. D., Brito, F. T., Monteiro, J. M., & Machado, J. C. (2022). BPA: A multilingual sentiment analysis approach based on BiLSTM. In *Proceedings of ICEIS (1)* (pp. 553–560).

Dhar, P., Gleason, J., Roy, A., Castillo, C. D., & Chellappa, R. (2021). PASS: Protected attribute suppression system for mitigating bias in face recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 15087–15096).

Girhepuje, S. (2023). Identifying and examining machine learning biases on adult dataset. arXiv preprint arXiv:2310.09373.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems* (Vol. 29).

Koumeri, L. K., Legast, M., Yousefi, Y., Vanhoof, K., Legay, A., & Schommer, C. (2023). Compatibility of fairness metrics with EU non-discrimination laws: Demographic parity & conditional demographic disparity. arXiv preprint arXiv:2306.08394.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6), 1–35.

Sena, L. B., Praciano, F. D., Chaves, I. C., Brito, F. T., Neto, E. R. D., Monteiro, J. M., & Machado, J. C. (2022). Audio-MC: A general framework for multi-context audio classification. In *Proceedings of ICEIS (1)* (pp. 374–383).

Stoyanovich, J., Howe, B., & Jagadish, H. V. (2020). Responsible data management. *Proceedings of the VLDB Endowment*, 13(12).

Žliobaitė, I. (2017). Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery*, 31(4), 1060–1089.