# Análise de Modelos de Aprendizado de Máquina para a Predição do Desempenho de Alunos com Enfoque na Detecção de Viés Algorítmico

**Matias Oliveira**[1], **Luciano de Souza Cabral**[1,2], **Rafael Ferreira Mello**[2,3,4]

[1]Campus Jaboatão dos Guararapes – Instituto Federal de Pernambuco (IFPE)
Jaboatão dos Guararapes – PE – Brasil

[2]Núcleo de Excelência em Tecnologias Sociais (NEES) – Instituto de Computação (IC)
Universidade Federal de Alagoas (UFAL) – Maceió – AL – Brasil

[3]C.E.S.A.R. Inovation Center – Recife – PE – Brasil

[4]Departamento de Sistemas e Computação
Universidade Federal Rural de Pernambuco (UFRPE)
Recife – PE – Brasil

{mco5@discente,luciano.cabral@jaboatao}.ifpe.edu.br, rflm@cesar.org.br

***Abstract.*** *In the current educational landscape, the abundance of data has become crucial. Studies show that variables such as academic history, behavior, and socioeconomic context are correlated with students' future performance. Educational institutions benefit from analyzing this data, preventing dropout rates and efficiently allocating resources. The use of machine learning algorithms (ML) has proven effective, but faces the challenge of algorithmic bias, which can harm underrepresented groups. This study compares algorithms using algorithmic fairness frameworks to measure bias. The results indicate that the K-Nearest Neighbors algorithm is effective in fairly predicting student performance, demonstrating high overall accuracy and low bias.*

***Resumo.*** *No atual panorama educacional, a disponibilidade abundante de dados tornou-se essencial. Pesquisas revelam que fatores como histórico escolar, comportamento e contexto socioeconômico estão diretamente ligados ao sucesso futuro dos alunos. Ao analisar esses dados, as instituições de ensino podem otimizar seus recursos, prevenindo a evasão escolar e promovendo uma alocação eficiente de recursos. Embora o uso de algoritmos de aprendizado de máquina (ML) tenha mostrado eficácia nesse contexto, surge o desafio do viés algorítmico, que pode marginalizar grupos sub-representados. Este estudo se propõe a comparar algoritmos usando frameworks de justiça algorítmica para quantificar e mitigar esse viés. Os resultados indicam que o algoritmo K-Nearest Neighbors se destaca por sua capacidade de prever o desempenho dos alunos de maneira justa, demonstrando alta acurácia global e baixo viés.*

## 1. Introdução

No cenário atual, a disponibilidade abundante de dados em todos os aspectos da vida humana tem aumentado consideravelmente, tornando-se fundamental para diversos setores, inclusive na educação. Estudos como os mencionados por [Aleem and Gore 2020] e [Khanna et al. 2016] demonstram que uma variedade de variáveis, como histórico escolar, comportamento, traços

psicológicos, habilidades, interesses e contexto socioeconômico, estão correlacionadas com o desempenho futuro dos estudantes.

Diversas instituições, como escolas e universidades, colhem benefícios ao estudar os dados de seus estudantes. Uma análise precisa pode identificar individualmente os alunos que precisam de mais apoio acadêmico, permitindo uma alocação mais eficiente de recursos e contribuindo para a prevenção da evasão escolar. Isso beneficia tanto os alunos quanto professores e gestores. O uso desses dados é conhecido como Mineração de Dados Educacionais *(Educational Data Mining - EDM)*.

Uma estratégia cada vez mais utilizada para análise de dados e previsão de resultados futuros é o emprego de algoritmos de aprendizagem de máquina *(machine learning)*. Estes algoritmos têm demonstrado eficiência em cumprir suas funções, como destacado por [Issah et al. 2023]. Por exemplo, o estudo conduzido por [Burgos et al. 2018] apresentou um plano de tutoria voltado para os estudantes com maiores necessidades de auxílio, identificados pelo sistema, baseado em um algoritmo de regressão logística. A implementação desse plano resultou em uma redução de 14% na taxa de evasão dos alunos em um curso online oferecido por uma universidade espanhola.

No entanto, o uso da Aprendizagem de Máquina enfrenta um grande desafio quando se trata do tratamento justo de pessoas em sistemas. Principalmente devido ao desbalanceamento de dados, como destacado por [Yan et al. 2020], frequentemente refletindo problemas sociais, muitos grupos são sub-representados ou mesmo ausentes nas informações utilizadas para alimentar os algoritmos. Esse cenário resulta em previsões refletindo o que há nos dados, o que pode afetar adversamente a vida de muitas pessoas. Esse fenômeno é conhecido como viés *(bias)*.

No campo da previsão do desempenho dos alunos, essa questão não é exceção, já que os sistemas podem influenciar negativamente os estudantes com base em características como sexo, idade, etnia ou país de origem. Para que gestores e professores possam prevenir os efeitos do viés algorítmico, é primordial detectá-lo, embora nem sempre seja uma tarefa simples. Uma das soluções amplamente adotadas é o uso de bibliotecas *frameworks* e métricas de justiça *(fairness)* para avaliar o nível de viés *(bias)* de um determinado modelo [Vasquez Verdugo et al. 2022a].

O objetivo deste estudo é comparar algoritmos utilizando *frameworks* e métricas *fairness* utilizadas na literatura para medir o viés algorítmico. A meta é determinar quais algoritmos são mais eficazes na predição sem comprometer a justiça algorítmica. Busca-se utilizar uma base pública e experimentos abertos para a implementação de práticas mais justas em sistemas de predição de performance de alunos através de algoritmos de *machine learning*, garantindo que decisões baseadas em dados beneficiem todos os estudantes de maneira imparcial, além de garantir a reprodutibilidade do estudo por terceiros *a posteriori*.

## 2. Revisão da Literatura

Vários estudos têm empregado algoritmos de machine learning para prever o desempenho futuro dos alunos, como exemplificado no artigo de [Pallathadka et al. 2023]. Neste estudo, foram utilizados os algoritmos *Support Vector Machine (SVM), Naive Bayes (NB)*, Árvores de Decisão *(Decision Trees)* C4.5 e ID3 para prever o desempenho dos alunos da Universidade do Minho, em Portugal, tendo o SVM alcançado os melhores resultados. Por outro lado, em [Issah et al. 2023], os pesquisadores realizaram uma revisão da literatura, focando principalmente em identificar os atributos mais relevantes na previsão do desempenho dos alunos. Os

atributos mais significativos encontrados foram de natureza acadêmica, sociodemográfica e comportamental. Os algoritmos mais eficazes identificados pelos autores incluíram *Decision-Tree (DT), Bayesian Networks (BN), Artificial Neural Network (ANN), Random Forest (RF), Logistic Regression (LR), K-Nearest Neighbor (KNN)* e *SVM*.

O artigo de [Farissi et al. 2020] concentra-se na resolução do problema da seleção de características em conjuntos de dados de alta dimensão com classes desbalanceadas, comumente encontrados em conjuntos de dados de EDM. A solução proposta pelos autores envolve o uso de um modelo que combina um algoritmo genético para a seleção de características com uma árvore de decisão *(Decision Tree)* para a classificação. O estudo demonstra uma melhoria significativa ao comparar o desempenho dos modelos que utilizam essa técnica com modelos *baseline*, que não incluem a seleção de características por algoritmo genético.

Por outro lado, o estudo realizado por [Baig et al. 2023] sugere a utilização de uma combinação do algoritmo *Fuzzy C-Means* (FCM) para clusterização com diversos algoritmos de *machine learning*, como *Multi-Layer Perceptron* (MLP), *LR* e *RF*. Os resultados mais promissores foram obtidos com as combinações FCM-MLP e FCM-LR.

Apesar da existência de artigos que empregam diversos algoritmos para avaliar sua eficácia na predição do desempenho dos alunos, há uma lacuna em relação à preocupação com o viés algorítmico ao qual esses modelos podem estar sujeitos. A maioria dos conjuntos de dados que representam o desempenho dos alunos contém informações sensíveis ao viés e frequentemente são desbalanceados. Mesmo estudos como a revisão conduzida por [Issah et al. 2023], que destacam a forte correlação entre dados sociodemográficos e o desempenho dos alunos, não abordam especificamente essa questão.

No estudo de [Li et al. 2023], os autores realizaram uma revisão sistemática com foco na identificação dos principais atributos sensíveis, métricas de *fairness* e estratégias para aprimorar a equidade dos modelos. Os atributos sensíveis mais comuns identificados foram sexo / gênero e raça / etnia. A métrica de *fairness* mais recorrente foi a ABROCA (Área sob a Curva ROC Balanceada). As estratégias mais frequentes para melhorar os modelos geralmente envolvem técnicas de pré-processamento. Embora essa revisão tenha sido abrangente e bem conduzida, seu caráter generalista na área de EDM não abordou especificamente as necessidades da predição de desempenho dos alunos.

Um artigo que apresentou uma abordagem interessante foi [Hu and Rangwala 2020], focado em detectar quais estudantes estão em risco de falhar em um curso universitário com base em suas notas anteriores, treinando modelos individualmente para cada curso. O modelo proposto, chamado *Multiple Cooperative Classifier Model* (MCCM), foi comparado a outros algoritmos com foco em *fairness*, além de um modelo de LR sem restrições de *fairness*. Os resultados variaram significativamente entre os cursos, com diferentes grupos sendo considerados sensíveis ao viés em diferentes contextos.

O artigo [Vasquez Verdugo et al. 2022b] propõe um *framework* denominado *FairEd*, que busca medir, gerar explicações e sugerir técnicas de mitigação de viés. Para isso, foram utilizados os algoritmos *SVM*, *LR*, *RF*, *DT* e *KNN*.

# 3. Metodologia

## 3.1. Dataset

O conjunto de dados utilizado neste trabalho é um *dataset* público disponível no Kaggle, denominado *Students' Academic Performance Dataset* (xAPI-Edu-Data)[1]. Este *dataset* é composto por 480 linhas e 17 colunas, contendo variáveis com informações sobre os estudantes, além de uma coluna alvo *(target)* que se deseja prever. A coluna alvo é categorizada em três classes: *Low-level*, *Middle-level* e *High-level*, que foram convertidas em *High-Level* e *Not-High-Level* para a realização da classificação binária. Este *dataset* é amplamente utilizado na literatura, sendo empregado em artigos como [Farissi et al. 2020, Nabil et al. 2021, Baig et al. 2023].

## 3.2. Algoritmos

A partir de nossas pesquisas, escolheu-se quatro algoritmos para testes e experimentos: *LR*, *RF*, *KNN* e *SVM*.

- **Support Vector Machine:** O *SVM* é um conjunto de algoritmos supervisionados. Segundo [Pradhan 2012], "a ideia principal por trás do SVM é a construção de um hiperplano ótimo, que pode ser utilizado para classificação, especialmente em padrões linearmente separáveis."
- **Random Forest:** O *Random Forest* é um algoritmo composto por várias instâncias do algoritmo *Decision Tree*. Cada árvore faz suas predições individualmente e, ao final, suas decisões são combinadas através de um sistema de votação majoritária. Por ser constituído de múltiplos algoritmos, o *Random Forest* é classificado como um algoritmo *ensemble* [Biau and Scornet 2016].
- **Logistic Regression:** O algoritmo *Logistic Regression* é descrito por [Nettleton 2014] da seguinte forma: "é um processo de modelar a probabilidade de uma saída discreta dada uma variável de entrada, sendo as saídas binárias as mais comuns".
- **K-Nearest Neighbors:** O algoritmo *K-Nearest Neighbors*, segundo [Kramer 2013], "é baseado na ideia de que os padrões mais próximos de um padrão alvo $x$, para o qual busca-se o rótulo, fornecem informações úteis sobre esse rótulo."

## 3.3. Algorithmic Fairness Frameworks

- **DALEX:** O DALEX é um pacote em Python e R utilizado principalmente para gerar explicações em modelos *black-box*, visando uma melhor compreensão do funcionamento do modelo. Além disso, o pacote possui uma extensão de *fairness*, que é utilizada para avaliar o viés algorítmico sob diversas métricas [Baniecki et al. 2021].
- **AIF360:** O AI Fairness 360 (AIF360) é um *framewwork* Python desenvolvido pela IBM para a detecção e mitigação de viés algorítmico. Ele oferece diversas métricas para avaliação de *fairness*, tanto a nível individual quanto de grupo [Bellamy et al. 2018].

## 3.4. Métricas

As tabelas 1 e 2 abaixo apresentam as métricas de justiça algorítmica utilizadas para avaliar a equidade e imparcialidade dos modelos de aprendizado de máquina.

- $\hat{Y}$: Representa a variável de saída ou predição do modelo.
- $Y$: Refere-se à variável de saída real ou verdadeira. Esta é a observação real do resultado que está sendo previsto pelo modelo.
- $A$: Representa o atributo sensível ou protegido, como raça, gênero, idade, etc.

[1]https://www.kaggle.com/datasets/aljarah/xAPI-Edu-Data

- $P(\cdot)$: Indica a probabilidade de uma determinada ocorrência.
- $|$: Simboliza ”dado”, ”dado que”ou ”condicional em”.
- $0$: Indica o valor falso.
- $1$: Indica o valor verdadeiro.
- $TP$: Representa *True Positive* (Verdadeiro Positivo).
- $TN$: Significa *True Negative* (Verdadeiro Negativo).
- $FP$: Indica *False Positive* (Falso Positivo).
- $FN$: Refere-se à *False Negative* (Falso Negativo).

| **Métrica de Fairness** | **Fórmula** |
|---|---|
| **Equal Opportunity** | $P(\hat{Y} = 1\|Y = 1, A = 0) = P(\hat{Y} = 1\|Y = 1, A = 1)$ |
| **Predictive Parity** | $P(Y = 1\|\hat{Y} = 1, A = 0) = P(Y = 1\|\hat{Y} = 1, A = 1)$ |
| **Predictive Equality** | $P(\hat{Y} = 1\|Y = 0, A = 0) = P(\hat{Y} = 1\|Y = 0, A = 1)$ |
| **Accuracy Equality** | $P(\hat{Y} = Y\|A = 0) = P(\hat{Y} = Y\|A = 1)$ |
| **Statistical Parity** | $P(\hat{Y}\|A = 0) = P(\hat{Y}\|A = 1)$ |
| **Disparate Impact** | $\frac{P(\hat{Y}=1\|A=0)}{P(\hat{Y}=1\|A=1)}$ |

**Tabela 1. Métricas *fairness* segundo [Makhlouf et al. 2021, Zafar et al. 2017]**

| **Métrica de Classificação** | **Fórmula** |
|---|---|
| **True Positive Rate (TPR)** | $\frac{TP}{TP+FN}$ |
| **Accuracy (Acurácia)** | $\frac{TP+TN}{TP+TN+FP+FN}$ |
| **False Positive Rate (FPR)** | $\frac{FP}{FP+TN}$ |
| **Positive Predictive Value (PPV)** | $\frac{TP}{TP+FP}$ |

**Tabela 2. Métricas de classificação e suas fórmulas [Makhlouf et al. 2021]**

## 4. Resultados

Nesta seção, utilizaremos os *frameworks* e métricas apresentados anteriormente para avançar em nossas análises. O objetivo é aplicar as ferramentas e métodos descritos para avaliar a eficácia e a justiça algorítmica dos modelos de machine learning selecionados.

### 4.1. DALEX

Utilizou-se o *framework* DALEX para conduzir experimentos com os algoritmos descritos na seção 3.2, tanto para a explicação dos modelos quanto para a detecção de possíveis vieses. As métricas empregadas para essa análise foram *Equal Opportunity*, *Predictive Equality*, *Accuracy Equality* e *Statistical Parity*, conforme detalhado na seção 3.4. No contexto dos experimentos, o grupo protegido foi composto por estudantes do gênero feminino (F), enquanto o grupo privilegiado consistiu em estudantes do gênero masculino (M). A base de dados utilizada apresenta uma distribuição de 64% de estudantes do gênero masculino e 36% do gênero feminino. A literatura indica que as mulheres frequentemente enfrentam uma posição desfavorável em muitos

tipos de avaliações preditivas, especialmente quando a base de dados é desbalanceada como a utilizada neste estudo [Barocas and Selbst 2016].

A avaliação dos modelos através do *framework* DALEX revelou que o *Random Forest*, *Support Vector Machine* (SVM) e a *Logistic Regression* apresentaram problemas de justiça algorítmica, especialmente em relação à métrica de *Statistical Parity* (STP). O *Random Forest* foi considerado não justo por exceder o limite de *epsilon* em uma métrica, enquanto o SVM e a *Logistic Regression* foram considerados injustos por ultrapassarem os limites em múltiplas métricas, incluindo *False Positive Rate* (FPR) e STP. Em contraste, o *K-Nearest Neighbors* foi o único algoritmo que passou em todas as métricas de *fairness*, sendo considerado justo pelo *framework* em todas as avaliações realizadas. Os valores podem ser observados em detalhes na tabela abaixo.

| **Métrica \Algoritmo** | **RF** | **SVM** | **KNN** | **LR** |
|---|---|---|---|---|
| **TPR** | 1.0 | 0.928 | 0.974 | 0.943 |
| **ACC** | 1.0 | 0.950 | 0.927 | 0.961 |
| **PPV** | 1.0 | 0.938 | 0.868 | 0.950 |
| **FPR** | NaN | 0.714 | 0.975 | 0.656 |
| **STP** | 0.732 | 0.725 | 0.820 | 0.727 |

**Tabela 3. Métricas *fairness* de avaliação dos algoritmos. Autoria Própria (2024).**

O KNN foi considerado o modelo mais justo no experimento. Seguido pelo RF, que apenas excedeu 1 critério por um limiar bem pequeno (0,067949). Os dois modelos mais injustos foram o LR e o SVC.

### 4.1.1. Comparativo dos modelos

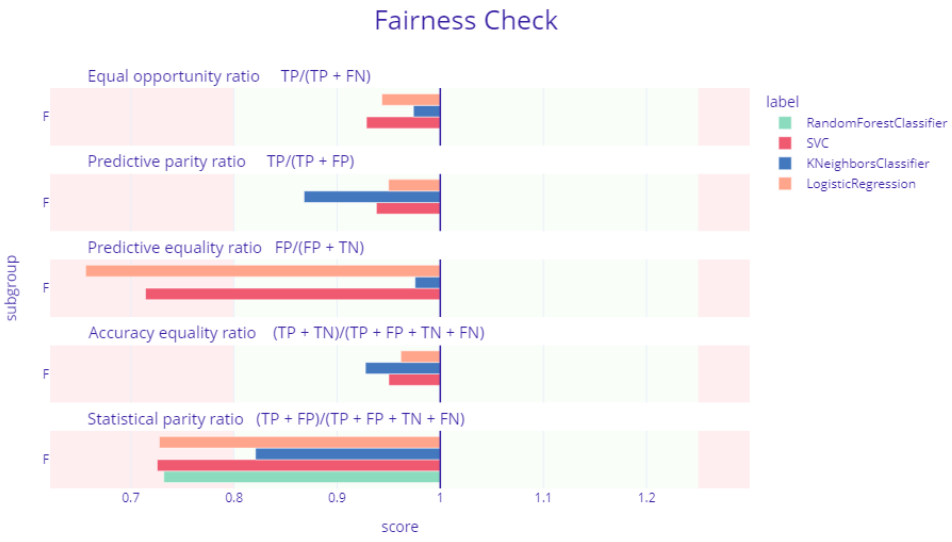


**Figura 1. *Fairness Check* comparando métricas dos modelos. Autoria Própria(2024).**

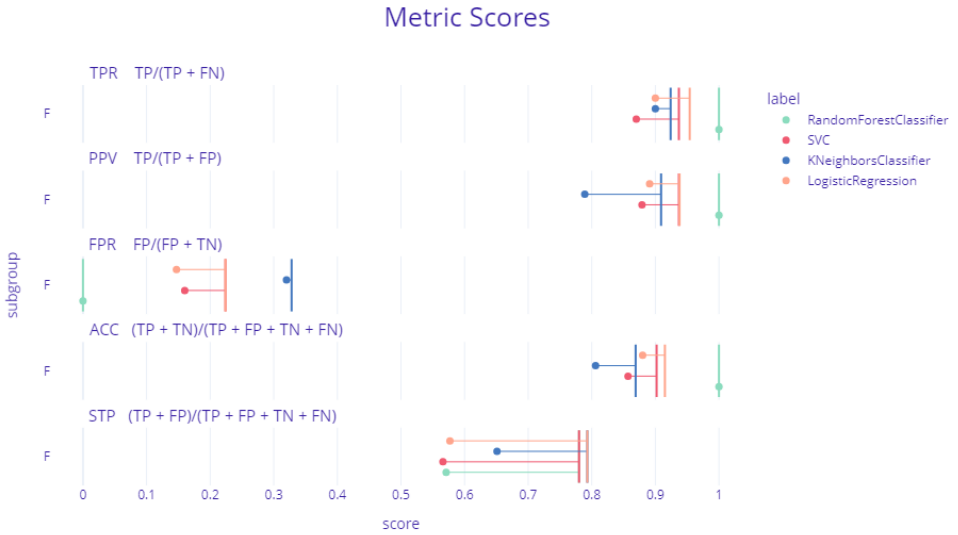


**Figura 2. *Metric Score* comparando as métricas dos modelos. Autoria Própria (2024).**

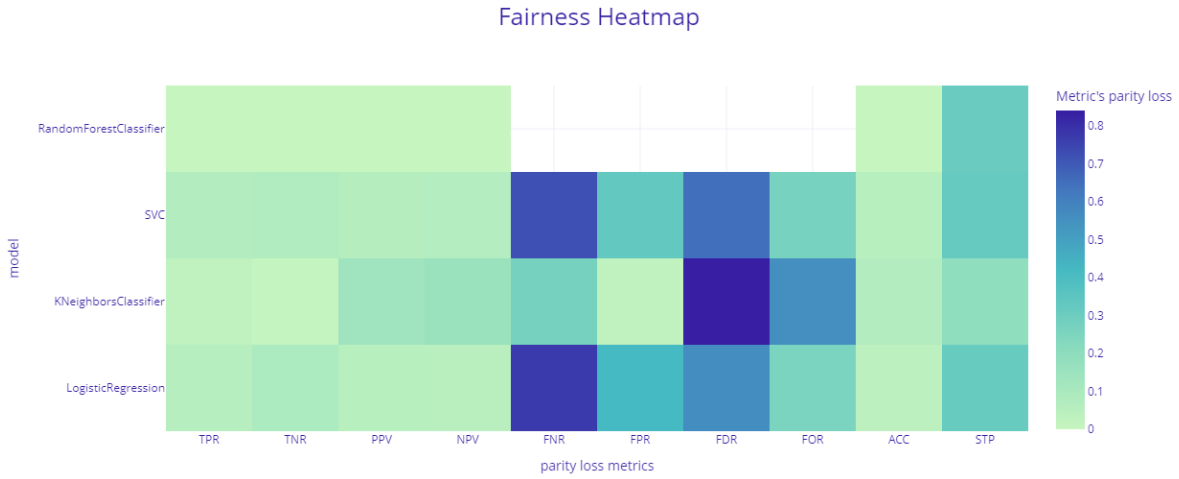


**Figura 3. *Heatmap* comparando os resultados dos modelos. Autoria Própria (2024).**

## 4.2. AIF360

De forma similar, empregou-se o *framework AI Fairness 360* (AIF360) da IBM para conduzir experimentos com os algoritmos mencionados na seção 3.2, visando detectar possíveis vieses. Utilizou-se as métricas *Statistical Parity Difference, Disparate Impact* e *Equal Opportunity Difference*, conforme formuladas na seção 3.4. O grupo protegido foi composto por estudantes do gênero feminino (F), enquanto o grupo privilegiado consistiu de estudantes do gênero masculino (M).

A avaliação dos algoritmos utilizando o *framework* AIF360 indicou que tanto o *Random Forest* quanto o *Support Vector Machine* (SVM) e a *Logistic Regression* apresentaram vieses consideráveis, especialmente nas métricas *Statistical Parity Difference* (SPD) e *Disparate Impact* (DI). Esses algoritmos tiveram valores fora dos intervalos aceitáveis para essas métricas, o que indica uma tendência desfavorável para certos grupos protegidos. Em contrapartida, o algoritmo *K-Nearest Neighbors* foi considerado justo, pois todas as suas métricas permaneceram dentro dos limites aceitáveis, evidenciando uma menor propensão ao viés algorítmico em comparação com os outros modelos. A tabela abaixo descreve os resultados.

| Métrica \Algoritmo | RF | SVM | LR | KNN |
|---|---|---|---|---|
| **SPD** | -0.266 | -0.144 | -0.218 | -0.058 |
| **DI** | 0.660 | 0.798 | 0.730 | 0.928 |
| **EO** | -0.004 | -0.121 | -0.040 | -0.054 |

**Tabela 4. Resultados das métricas para diferentes algoritmos. Autoria Própria (2024).**

## 5. Conclusão

O objetivo deste estudo foi comparar algoritmos frequentemente citados na literatura no campo da predição de desempenho estudantil, analisando-os tanto em termos de acurácia e precisão dos modelos quanto em relação ao viés algorítmico. Em nossos experimentos, utilizamos quatro algoritmos de *machine learning* e diversas métricas de *fairness*.

Embora o algoritmo Random Forest não tenha sido explicitamente considerado injusto, pois apresentou desempenho insatisfatório em apenas uma métrica segundo o framework DALEX, o algoritmo que obteve o melhor desempenho geral foi o K-Nearest Neighbors. Este algoritmo permaneceu dentro dos limites aceitáveis para todas as métricas de *fairness* avaliadas pelos frameworks utilizados e alcançou uma acurácia superior a 0.96, conforme relatado pelo framework DALEX.

Esses resultados sugerem que o K-Nearest Neighbors pode ser uma opção mais eficaz para a predição de desempenho estudantil sem comprometer a justiça algorítmica, especialmente no contexto da comparação entre estudantes do sexo feminino e masculino. Assim, este estudo oferece uma base para a implementação de práticas mais equitativas em sistemas de predição de desempenho acadêmico, ajudando a combater o sexismo nas avaliações e garantindo que as decisões baseadas em dados beneficiem todos os estudantes de maneira imparcial.

Apesar dos resultados promissores, este estudo apresenta algumas limitações. A quantidade de dados utilizada foi relativamente pequena, o que pode impactar a capacidade de generalização dos modelos desenvolvidos. Além disso, a inclusão de características mais detalhadas e abrangentes poderia proporcionar uma visão mais completa do desempenho dos estudantes. Aspectos adicionais, como histórico acadêmico mais detalhado e fatores socioeconômicos, poderiam enriquecer as análises e contribuir para a construção de modelos mais robustos e justos. A restrição na quantidade de dados e na diversidade de características ressalta a necessidade de futuros estudos com conjuntos de dados mais amplos e detalhados, visando aprimorar os resultados deste trabalho.

## 6. Agradecimentos

O presente trabalho foi realizado com apoio do CNPq, Conselho Nacional de Desenvolvimento Científico e Tecnológico - Brasil.

## Referências

Aleem, A. and Gore, M. M. (2020). Educational data mining methods: A survey. In *2020 IEEE 9th International Conference on Communication Systems and Network Technologies (CSNT)*, pages 182–188.

Baig, M., Shaikh, S., Khatri, K., Shaikh, M., Khan, M. Z., and Rauf, M. (2023). Prediction of students performance level using integrated approach of ml algorithms. *International Journal of Emerging Technologies in Learning (iJET)*, 18:216–234.

Baniecki, H., Kretowicz, W., Piatyszek, P., Wisniewski, J., and Biecek, P. (2021). dalex: Responsible machine learning with interactive explainability and fairness in python. *Journal of Machine Learning Research*, 22(214):1–7.

Barocas, S. and Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, 104:671. Available at SSRN: https://ssrn.com/abstract=2477899.

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J. T., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., and Zhang, Y. (2018). AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *CoRR*, abs/1810.01943.

Biau, G. and Scornet, E. (2016). A random forest guided tour. *TEST*, 25(2):197–227.

Burgos, C., Campanario, M. L., de la Peña, D., Lara, J. A., Lizcano, D., and Martínez, M. A. (2018). Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout. *Computers & Electrical Engineering*, 66:541–556.

Farissi, A., Dahlan, H. M., and Samsuryadi (2020). Genetic algorithm based feature selection for predicting student's academic performance. In Saeed, F., Mohammed, F., and Gazem, N., editors, *Emerging Trends in Intelligent Computing and Informatics*, pages 110–117, Cham. Springer International Publishing.

Hu, Q. and Rangwala, H. (2020). Towards fair educational data mining: A case study on detecting at-risk students. In *Educational Data Mining*.

Issah, I., Appiah, O., Appiahene, P., and Inusah, F. (2023). A systematic review of the literature on machine learning application of determining the attributes influencing academic performance. *Decision Analytics Journal*, 7:100204.

Khanna, L., Singh, S. N., and Alam, M. (2016). Educational data mining and its role in determining factors affecting students academic performance: A systematic review. In *2016 1st India International Conference on Information Processing (IICIP)*, pages 1–7.

Kramer, O. (2013). *K-Nearest Neighbors*, pages 13–23. Springer Berlin Heidelberg, Berlin, Heidelberg.

Li, L., Sha, L., Li, Y., Raković, M., Rong, J., Joksimovic, S., Selwyn, N., Gašević, D., and Chen, G. (2023). Moral machines or tyranny of the majority? a systematic review on predictive bias in education. In *LAK23: 13th International Learning Analytics and Knowledge Conference*, LAK2023, page 499–508, New York, NY, USA. Association for Computing Machinery.

Makhlouf, K., Zhioua, S., and Palamidessi, C. (2021). On the applicability of machine learning fairness notions. *SIGKDD Explor. Newsl.*, 23(1):14–23.

Nabil, A., Seyam, M., and Abou-Elfetouh, A. (2021). Prediction of students' academic performance based on courses' grades using deep neural networks. *IEEE Access*, PP:1–1.

Nettleton, D. (2014). Chapter 9 - data modeling. In Nettleton, D., editor, *Commercial Data Mining*, pages 137–157. Morgan Kaufmann, Boston.

Pallathadka, H., Wenda, A., Ramirez-Asís, E., Asís-López, M., Flores-Albornoz, J., and Phasinam, K. (2023). Classification and prediction of student performance data using various machine learning algorithms. *Materials Today: Proceedings*, 80:3782–3785. SI:5 NANO 2021.

Pradhan, A. (2012). Support vector machine-a survey. *IJETAE*, 2.

Vasquez Verdugo, J., Gitiaux, X., Ortega, C., and Rangwala, H. (2022a). Faired: A systematic fairness analysis approach applied in a higher educational context. In *LAK22: 12th International Learning Analytics and Knowledge Conference*, LAK22, page 271–281, New York, NY, USA. Association for Computing Machinery.

Vasquez Verdugo, J., Gitiaux, X., Ortega, C., and Rangwala, H. (2022b). Faired: A systematic fairness analysis approach applied in a higher educational context. In *LAK22: 12th International Learning Analytics and Knowledge Conference*, LAK22, page 271–281, New York, NY, USA. Association for Computing Machinery.

Yan, S., Kao, H.-t., and Ferrara, E. (2020). Fair class balancing: Enhancing model fairness without observing sensitive attributes. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 1715–1724, New York, NY, USA. Association for Computing Machinery.

Zafar, M. B., Valera, I., Gomez Rodriguez, M., and Gummadi, K. P. (2017). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, page 1171–1180, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.