Interpretabilidade e Justiça Algorítmica: Avançando na Transparência de Modelos Preditivos de Evasão Escolar

Cássio S. Carvalho[1], **Júlio C. B. Mattos**[1], **Marilton S. Aguiar**[1]

[1]Programa de Pós-Graduação em Computação – Universidade Federal de Pelotas (UFPel)
Rua Gomes Carneiro, 1 – 96.010-610 – Pelotas – RS – Brasil

{cassio.carvalho,julius,marilton}@inf.ufpel.edu.br

Abstract. *With the ubiquity of Artificial Intelligence (AI), concerns arise about the transparency of models and the introduction of biases. This study examines the relationship between interpretability and algorithmic fairness in predictive models of early school dropout. An evolution of the LIME explanation clustering method is presented, analyzing results fairly on sensitive attributes such as gender, race, quota, and educational origin. The findings show that the "agreement" interpretability metric can relate to variation in algorithmic fairness, identifying regions with varying performance and fairness. Analytics help tune AI models to improve their transparency in educational contexts.*

Resumo. *Com a onipresença da Inteligência Artificial (IA), surgem preocupações sobre a transparência dos modelos e a introdução de vieses. Este estudo examina a relação entre interpretabilidade e justiça algorítmica em modelos preditivos de evasão escolar precoce. É apresentada uma evolução do método de clusterização de explicações LIME, analisando resultados com justiça em atributos sensíveis como gênero, raça, cota e origem escolar. Os achados mostram que a métrica de interpretabilidade "agreement" pode se relacionar com a variação na justiça algorítmica, identificando regiões com desempenho e justiça variados. A análise ajuda a ajustar modelos de IA para melhorar a sua transparência em contextos educacionais.*

1. Introdução

A Mineração de Dados Educacionais (MDE) é a área de pesquisa dedicada ao desenvolvimento de métodos para realizar descobertas em tipos distintos de dados provenientes de ambientes educacionais, e à utilização desses métodos para melhor compreender os alunos e os contextos em que eles aprendem [Baker et al. 2010]. Trata-se de um campo que explora algoritmos estatísticos, de Aprendizado de Máquina (AM) e de Mineração de Dados (MD) aplicados a diferentes tipos de dados educacionais [Romero and Ventura 2010], por meio de um processo cíclico de descoberta de conhecimento [Romero and Ventura 2020].

As aplicações na área da MDE são variadas e podem ser categorizadas com base em diferentes critérios [Baker and Yacef 2009, Romero and Ventura 2010, Peña-Ayala 2014, Hegazi and Abugroon 2016], incluindo o usuário final, como alunos, educadores, administradores e pesquisadores [Romero and Ventura 2013]. A predição de desempenho está entre as aplicações mais relevantes, permitindo a identificação antecipada do risco de reprovação ou desistência no processo de aprendizagem

DOI: 10.5753/sbie.2024.242289

[Xiao et al. 2022, Bakhshinategh et al. 2018]. A evasão no ensino superior é um problema internacional que representa desperdícios sociais, acadêmicos e econômicos [Silva Filho et al. 2007], sendo frequentemente objeto de estudo por instituições de ensino brasileiras [Kantorski et al. 2023, Oliveira and Medeiros 2024].

Os desempenhos cada vez mais avançados dos algoritmos de AM impulsionam a adoção generalizada de ferramentas de Inteligência Artificial (IA). Contudo, os modelos utilizados nessas aplicações apresentam maior complexidade e falta de transparência [Carvalho et al. 2019, Linardatos et al. 2021]. Sistemas com essas características são conhecidos como "caixas pretas" e são difíceis de confiar, especialmente quando usados em áreas que envolvem questões morais e de justiça [Linardatos et al. 2021]. No que diz respeito às decisões de alto risco, o problema é ainda mais grave, pois confiar decisões essenciais a um sistema que não pode se explicar e não pode ser explicado por humanos representa perigos evidentes [Adadi and Berrada 2018]. Especificamente na área da educação, há armadilhas no uso de modelos de caixa preta: falta de transparência, potencial de viés, interpretabilidade limitada, dependência da qualidade dos dados e dificuldade de adaptação às novas circunstâncias [Samek et al. 2019].

Interpretabilidade e explicabilidade são conceitos intimamente relacionados e frequentemente utilizados de forma intercambiável [Molnar 2022] para representar o grau em que um ser humano pode entender a causa de uma decisão [Miller 2019] ou o grau em que um ser humano pode prever consistentemente o resultado de um modelo [Kim et al. 2016]. Diferentes métodos de interpretabilidade são aplicados na área de AM [Vieira and Digiampietri 2022, Araujo 2021], inclusive na educação [Alamri and Alharbi 2021, Rachha and Seyam 2023].

Outro aspecto sensível em relação aos modelos de AM é a justiça algorítmica (*fairness*). Justiça é a ausência de qualquer viés baseado em características inerentes ou adquiridas de um indivíduo irrelevantes no contexto particular da tomada de decisão [Saxena et al. 2019]. A injustiça no aprendizado de máquina pode surgir potencialmente dos dados (quando vieses nos dados distorcem o que é aprendido pelos algoritmos de AM) e dos algoritmos (quando nuances na forma como os algoritmos operam impedem-nos de tomar decisões justas, mesmo quando os dados não são tendenciosos). Além disso, resultados algorítmicos tendenciosos podem impactar a experiência do usuário, criando um ciclo de *feedback* entre dados, algoritmos e usuários que perpetua e amplifica as fontes existentes de viés [Mehrabi et al. 2021].

Definir justiça em problemas de classificação é essencial para combater o preconceito e atingir a justiça algorítmica. Exemplos de definições incluem *Equalized Odds* e *Equal Opportunity* [Hardt et al. 2016], Statistical Parity [Dwork et al. 2012], *Fairness Through Awareness* [Dwork et al. 2012, Kusner et al. 2017], *Fairness Through Unawareness* [Grgic-Hlaca et al. 2016, Kusner et al. 2017] e *Absolute Between-ROC Area* (ABROCA) [Gardner et al. 2019]. Essas definições permitem analisar o desempenho do modelo no contexto de grupos sociodemográficos. Os atributos utilizados para a definição desses grupos são chamados de atributos sensíveis, como, por exemplo, raça, sexo, poder aquisitivo, entre outros.

Dada a relevância das considerações apresentadas, este trabalho é um esforço para investigar aspectos de justiça algorítmica e interpretabilidade no contexto da evasão de es-

tudantes no ensino superior. Baseado em nosso trabalho anterior [Carvalho et al. 2023], propõe-se uma evolução do método de clusterização de explicações visando endereçar a seguinte Questão de Pesquisa (QP): "De que forma a interpretabilidade de um modelo de AM está ligada ao desempenho preditivo entre grupos sociodemográficos?". Para responder à questão de pesquisa proposta, o método original foi aprimorado mediante a inclusão de validação estatística para diferentes amostragens de conjuntos de treino e teste. A consistência das explicações geradas por diferentes técnicas de AM é analisada para relacionar essas explicações com a justiça algorítmica.

O restante do artigo está organizado da seguinte forma. A Seção 2 destaca trabalhos relacionados que abordam justiça algorítmica ou interpretabilidade no contexto da MDE. A Seção 3 descreve o *dataset* utilizado, a análise exploratória de dados, os demais materiais e métodos aplicados. A Seção 4 é dedicada à apresentação dos experimentos e à discussão dos resultados obtidos. Por fim, a Seção 5 apresenta as conclusões finais.

2. Trabalhos Relacionados

Métodos de interpretabilidade são aplicados no contexto da MDE. No caso de métodos agnósticos, podem-se citar estudos que utilizaram SHAP [Qu et al. 2022, Colak Oz et al. 2023], LIME [Ribeiro et al. 2016, Alwarthan et al. 2022, Chou 2023] e também explicações contrafactuais [Tsiakmaki and Ragos 2021]. Existem também estudos específicos para determinadas classes de modelos, como no caso das Redes Neurais Artificiais [Matetic 2019, Jeon et al. 2019]. Esses estudos abordam interpretabilidade sem, contudo, endereçar a justiça algorítmica.

Simultaneamente, muitos trabalhos que abordam a justiça algorítmica não tratam concomitantemente da interpretabilidade dos modelos, conforme descrito nos estudos a seguir. O estudo de [Le Quy et al. 2023] avaliou sete medidas de justiça para problemas de predição de desempenho de alunos, considerando atributos sensíveis como raça e gênero. Os experimentos refletem variações e correlações de medidas de justiça entre conjuntos de dados e modelos preditivos.

A pesquisa de [Hu and Rangwala 2020] desenvolveu um modelo individualmente justo para identificar alunos em risco de baixo desempenho. A métrica de justiça individual pressupõe que a qualificação de um indivíduo não deve mudar se seu atributo sensível for alterado. O modelo consiste em dois classificadores (baseados em Redes Neurais Artificiais), cada um correspondente a um grupo sensível, com a pontuação de justiça sendo a diferença entre os resultados dos classificadores. Cada modelo é treinado considerando seu resultado combinado com o valor comum do escore de justiça (*fairness score*). O modelo proposto demonstrou remover efetivamente o viés das previsões, tornando-as úteis para apoiar todos os alunos.

A investigação de [Xiang et al. 2022] utilizou dois conjuntos de dados educacionais públicos para abordar problemas de regressão e classificação, revelando vieses na previsão do desempenho dos alunos. A abordagem envolveu duas etapas: na primeira, os modelos foram avaliados apenas com base na precisão; na segunda, os melhores modelos foram ajustados considerando tanto a imparcialidade quanto a precisão. A justiça foi medida pelo desvio padrão em grupos sensíveis (sexo, região, nível de ensino, faixa de renda, faixa etária, deficiência). As conclusões sugeriram a possibilidade de ajustar modelos para considerar a equidade, embora com um sensível prejuízo na precisão, ainda

que esses modelos pudessem alcançar um desempenho mais justo em quase todas as subcategorias de alunos.

Por fim, existem trabalhos que abordam os dois temas inter-relacionadamente. O estudo de [Sahlaoui et al. 2023] apresenta diretrizes para mitigar o desbalanceamento de dados e melhorar a justiça algorítmica na educação. Variações do SMOTE são testadas na geração de modelos de predição, ao mesmo tempo em que são comparados tecnicamente os métodos LIME e SHAP. O estudo de [Pei and Xing 2022] fornece um *pipeline* completo e acionável para permitir intervenções personalizadas, mas alerta sobre a equidade dessas intervenções. Destaca que desequilíbrios nas informações demográficas podem introduzir discriminações não intencionais nos modelos, fazendo com que os resultados produzidos não representem os padrões reais existentes em grupos desfavorecidos.

Uma breve relação entre os dois tópicos foi apresentada por [Afrin et al. 2022]. O estudo abordou a predição de sucesso estudantil a partir do uso de mídias sociais, incluindo também informações demográficas e de *background* (WPA – *weighted average mark*). Utilizando explicações fornecidas pelo método SHAP, identificou que variáveis sensíveis como idade, gênero e WPA tiveram alto valor de influência nas predições. Por fim, a realização de um *survey* permitiu identificar que a percepção de justiça no uso de determinados atributos varia entre diferentes grupos de participantes.

O estudo de [Kung and Yu 2020] faz uma comparação de modelos (interpretáveis e de "caixa preta") frequentemente utilizados na predição de desempenho de estudantes, relacionando-os à justiça algorítmica. Foram utilizados atributos como etnia, gênero, *status* de baixa renda, *status* de estudante universitário de primeira geração e quartil GPA (*Grade Point Average*) do ensino médio. Resultados mostraram que modelos interpretáveis não comprometem acurácia ou justiça algorítmica em comparação a modelos mais complexos. Por outro lado, mesmo nos modelos mais justos, o preconceito algorítmico persiste, especialmente contra minorias étnicas e alunos academicamente despreparados.

A pesquisa de [Dsilva et al. 2023] explora o uso de *Explainable Boosting Machines* (EBMs) em tarefas de predição de risco acadêmico, utilizando o *dataset Open University Learning Analytics Dataset* (OULAD) [Kuzilek et al. 2017] para prever reprovações e desistências. Os resultados demonstram que EBMs têm desempenho similar a outras abordagens, enquanto se mostram competitivos em métricas como precocidade (*earliness*), estabilidade (*stability*), justiça (*fairness*) e fidelidade (*faithfulness*) das explicações. A avaliação de justiça considerou atributos como gênero e incapacidades por meio das métricas *Statistical Parity Difference* (SPD) e *Equal Opportunity Difference* (EOD). Para explicabilidade, foram considerados os métodos LIME, SHAP e o modelo inerentemente interpretável de Regressão Logística, sendo que as EBMs geralmente apresentaram melhor fidelidade nas explicações.

O trabalho de [Bhargava et al. 2020] visa melhorar a justiça de classificadores reduzindo suas dependências por atributos sensíveis. Apresenta o *LimeOut*, inspirado na técnica *dropout* de Redes Neurais Artificiais e combinado com uma abordagem *ensemble*. Explicações LIME são utilizadas para determinar se o classificador é justo em relação a determinados atributos sensíveis, os quais são então escolhidos para remoção. Cada remoção de atributo resulta em um novo classificador. O conjunto de classifica-

dores resultantes, quando utilizados combinadamente (*ensemble*), demonstra ser menos dependente dos atributos sensíveis e mantém, ou aumenta, sua acurácia. O *framework* proposto foi revisitado em [Alves et al. 2021] para avaliar seu desempenho em outros *datasets* (incluindo o LSAC – *Law School Admissions Council*), técnicas de AM e em relação a métricas de justiça. *LimeOut* mostrou-se robusto para diferentes pontos de vista de justiça, sem comprometer sua acurácia.

O presente trabalho diferencia-se dos anteriores ao estabelecer uma relação direta entre interpretabilidade de modelos e justiça algorítmica no contexto da MDE. As relações identificadas são apresentadas detalhadamente, destacando que uma métrica de interpretabilidade permite identificar regiões do espaço amostral com melhor desempenho e justiça similar, enquanto outras regiões apresentam desempenho e justiça inferiores.

3. Materiais e Métodos

Experimentos foram realizados utilizando Python no ambiente Google Colaboratory[1]. Os classificadores Random Forest (RF) e Redes Neurais Artificiais (RNA) foram importados do scikit-learn[2], enquanto o XGBoost[3] (XGB) foi importado de pacote específico. O balanceamento de dados foi realizado usando SMOTE[4]. Para a clusterização das explicações LIME, foram utilizados os pacotes LIME[5] [Ribeiro et al. 2016] e k-prototype[6]. Relatórios para análise exploratória de dados foram gerados por meio do pacote YData Profiling[7].

3.1. *Dataset*

O *dataset* utilizado consiste de 17.689 alunos da Universidade Federal de Pelotas (UFPel), elaborado com o intuito de viabilizar a identificação precoce de alunos em risco de evasão. As instâncias foram coletadas considerando o primeiro semestre de estudantes de todos os cursos de graduação, com ingresso entre 2014 e 2017, e saída (tanto evasão quanto conclusão) a partir do segundo semestre. Os atributos disponíveis e utilizados são descritos na Tabela 1. Durante o pré-processamento, valores faltantes foram substituídos pelo valor mais frequente no *dataset*. Por fim, os atributos categóricos foram mapeados utilizando a técnica de *OneHotEncoder* (OHE).

Tabela 1. Atributos do *dataset*.

#	Atributo	Descrição	#	Atributo	Descrição
1	idade_ingresso	Idade no momento do ingresso	11	média_semestre	Média no primeiro semestre
2	anos_entre_grad_médio	Anos entre graduação e conclusão do ensino médio	12	num_disciplinas	Número de disciplinas no primeiro semestre
3	sexo	Masculino / Feminino	13	num_créditos	Número de créditos no primeiro semestre
4	raça	Código IBGE para cor	14	num_créditos_aprovados	Número de créditos aprovados no primeiro semestre
5	estado_civil	Estado civil	15	num_créditos_dispensados	Número de créditos dispensados no primeiro semestre
6	cota	Categoria no sistema de cotas	16	num_créditos_reprovados	Número de créditos reprovados no primeiro semestre
7	de_escola_pública	Aluno oriundo de escola pública	17	num_créditos_infrequentes	Número de créditos infrequentes no primeiro semestre
8	possui_graduação	Aluno com diploma de graduação	18	num_créditos_trancados	Número de créditos trancados no primeiro semestre
9	turno	Turno do curso	19	num_créditos_sem_ref	Número de créditos obrigatórios do primeiro semestre do curso
10	area_fundamental	Área fundamental do curso	20	num_exames	Número de exames no primeiro semestre
			21	situação	Atributo alvo. Graduado (0) ou Evadido (1)

A ferramenta *YData Profiling* permitiu a realização da análise dos dados. É importante destacar que os valores faltantes para as variáveis sensíveis foram 2.521 (14,3%) para raça, 1.783 (10,1%) para cota, 478 (2,7%) para origem em escola pública e 0 (0%) para sexo. Observou-se um desbalanceamento[8] acima de 50% para raça (52,6%), estado

[1] https://colab.research.google.com [2] https://scikit-learn.org [3] https://xgboost.readthedocs.io
[4] https://imbalanced-learn.org [5] https://github.com/marcotcr/lime [6] https://github.com/nicodv/kmodes
[7] https://docs.profiling.ydata.ai/ [8] Na ferramenta *YData Profiling*, o percentual de desbalanceamento é estimado por uma função de entropia que considera a frequência de classes

civil (66,9%) e portador de diploma de graduação (91,5%). A distribuição das classes para o atributo situação foi de 39,7% para Graduado e 60,3% para Evadido. Conforme a configuração padrão da ferramenta, nenhum atributo apresentou distribuição enviesada.

No que diz respeito à correlação entre variáveis, destaca-se o atributo situação, que apresenta alta correlação com o número de créditos aprovados e média do semestre, com valores de 0,525 e 0,516, respectivamente. Considerando os atributos sensíveis, a única correlação significativa é entre origem em escola pública e cota, com valor de 0,517. A distribuição dos dados, correlações e demais informações sobre esse relatório podem ser verificadas na Seção Disponibilidade de Artefatos.

3.2. Método

Conforme apresentado na Seção 1, e baseado em nosso trabalho anterior [Carvalho et al. 2023], esta pesquisa propõe uma evolução do método de clusterização de explicações LIME com intuito de investigar a relação entre justiça algorítmica e interpretabilidade de modelos no contexto da evasão de estudantes no ensino superior. Nesse sentido, a Seção 3.2.1 descreve o método original de clusterização de explicações, enquanto a Seção 3.2.2 apresenta a evolução proposta.

3.2.1. Clusterizando explicações LIME

O método de clusterização de explicações LIME avalia a interpretabilidade de modelos de predição no contexto da evasão de estudantes no ensino superior, fornecendo uma métrica de interpretabilidade. O processo ocorre em duas etapas, conforme Figura 1 e descrito a seguir.

Figura 1. Fluxo do método de clusterização de explicações LIME.

Na etapa de predição, o *dataset* é dividido em conjuntos de treino (60%), teste (20%) e teste 2 (20%). Modelos de predição são obtidos para diferentes técnicas de AM utilizando validação cruzada e balanceamento de dados (para o atributo situação) no conjunto de treino, sendo posteriormente avaliados no conjunto de teste.

Na etapa de explicação, o método LIME é aplicado a todas as instâncias do conjunto de teste, seguido de uma extração de atributos, para obter um conjunto de dados tabulares não rotulados que representam as explicações do modelo para diferentes instâncias. Esse conjunto de explicações é então clusterizado por meio de AM não supervisionado, e os *centroids* dos *clusters* resultantes fornecem explicações centrais que elucidam o modelo em diferentes regiões do espaço amostral.

Ao utilizar essas explicações centrais de forma independente para predizer novas instâncias (no caso, do conjunto de teste 2), dois conjuntos são obtidos: *agreement* e *disagreement*. O conjunto *agreement* contém as instâncias em que a predição do modelo é igual à predição utilizando as explicações centrais, enquanto o conjunto *disagreement*

contém as instâncias em que a predição do modelo difere da predição pelas explicações centrais.

O número total de instâncias do conjunto de teste 2 é chamado de suporte, enquanto o número de instâncias incluídas no conjunto *agreement* é identificado como novo suporte. Portanto, quanto maior o novo suporte, maior a concordância entre o modelo e as explicações centrais. A relação entre o novo suporte e o suporte ($novo\_suporte/suporte$) é utilizada como métrica de interpretabilidade, sendo chamada de percentual de *agreement*.

Ao avaliar o desempenho do modelo em cada um dos conjuntos (*agreement* e *disagreement*), observa-se que o desempenho no *agreement* é normalmente superior ao desempenho no conjunto de teste 2 completo, enquanto o desempenho no *disagreement* é geralmente inferior. Dessa forma, as explicações centrais do modelo permitem dividir o espaço amostral em duas regiões, sendo que uma delas apresenta desempenho superior e está em concordância com as explicações centrais.

3.2.2. Evolução do método

Para investigar as relações entre a interpretabilidade de modelos e a justiça algorítmica, propõe-se uma evolução no método de clusterização de explicações LIME, conforme descrito a seguir:

- Inclusão de uma validação estatística envolvendo a etapa de divisão dos dados. Como resultado, cada técnica de AM será executada 10 vezes. A única diferença entre cada execução de uma mesma técnica é a *seed*[9] utilizada para o processo de divisão dos dados em treino, teste e teste 2. A *seed* 1 será utilizada pela execução 1 de todas as técnicas, a *seed* 2 será utilizada para a execução 2 de todas as técnicas, e assim por diante;
- Cálculo do desempenho médio de cada técnica nos conjuntos de teste, teste 2, *agreement* e *disagreement*, bem como do percentual médio de *agreement* e *disagreement*;
- Cálculo do desempenho médio de cada técnica no contexto das classes de cada uma das variáveis sensíveis: sexo, raça, cota e origem em escola pública;
- Avaliação do desempenho médio (global e por grupo sociodemográfico) pela aplicação do teste T de Student[10], para verificar se há diferença significativa entre os resultados dos conjuntos *agreement* e *disagreement* em relação ao conjunto de teste 2;
- Aplicação do teste T de Student para verificar se há diferença significativa entre os percentuais médios de *agreement* das diferentes técnicas;
- Cálculo do desvio padrão do desempenho de cada execução para as diferentes classes de cada grupo sócio demográfico;
- Cálculo do desvio padrão médio de desempenho de cada técnica para cada grupo sócio demográfico. Esse valor passar a ser uma medida de justiça para uma técnica de AM em relação a um grupo sócio demográfico (variável sensível).

[9] Semente para geração de números aleatórios e reprodutibilidade. Utilizada apenas no processo de divisão dos dados. [10] A escolha pelo teste T de Student busca evitar resultados ao acaso relacionados ao processo de divisão dos dados.

- Aplicação do teste T de Student para verificar se há diferença significativa entre as médias dos desvios padrões em cada grupo sócio demográfico.

4. Experimentos e Resultados

Esta seção descreve os resultados obtidos para todos os experimentos realizados, conforme detalhado na Seção 3, e apresenta uma discussão para responder à questão de pesquisa proposta na Seção 1.

As Tabelas 2, 3 e 4 apresentam os desempenhos de cada técnica de AM e estão organizadas conforme detalhamento a seguir. Os valores presentes nas colunas "teste", "teste 2", "*Agreement*" e "*Disagreement*" são referentes à métrica AUC (Área Sob a Curva ROC). A coluna "*clusters*" informa o número de *clusters* resultantes ao final do agrupamento das explicações. A coluna "Novo suporte" apresenta o número de instâncias atribuídas ao conjunto *agreement*, enquanto "Suporte" é o número de instâncias no conjunto teste 2. "Percentual (%) de *agreement*" é a relação entre novo suporte e suporte. "Média" é o desempenho médio para as 10 execuções. Por último, o *p-value* é o resultado do teste T de Student dos conjuntos *agreement* e *disagreement* em relação ao conjunto de teste 2.

Tabela 2. Resultados das execuções da técnica Random Forest (RF).

Exec.	Técnica	teste	teste 2	*Agrmt.*	*Disagrmt.*	*Clusters*	Novo suporte	Suporte	*% Agrmt.*
1	RF	0,849	0,852	0,888	0,629	3	2609	3538	73,74%
2	RF	0,848	0,851	0,862	0,654	3	2318	3538	65,52%
3	RF	0,852	0,859	0,891	0,689	3	2566	3538	72,53%
4	RF	0,859	0,850	0,892	0,658	3	2636	3538	74,51%
5	RF	0,845	0,858	0,882	0,678	3	2386	3538	67,44%
6	RF	0,851	0,854	0,887	0,678	3	2547	3538	71,99%
7	RF	0,850	0,856	0,890	0,649	3	2678	3538	75,69%
8	RF	0,859	0,851	0,884	0,709	3	2643	3538	74,70%
9	RF	0,849	0,852	0,883	0,686	3	2680	3538	75,75%
10	RF	0,856	0,853	0,888	0,667	3	2692	3538	76,09%
Média		0,8518	0,8536	0,8847	0,6697				72,80%
p-value				9,11E-07	1,03E-09				

Tabela 3. Resultados das execuções da técnica XGBoost (XGB).

Exec.	Técnica	teste	teste 2	*Agrmt.*	*Disagrmt.*	*Clusters*	Novo suporte	Suporte	*% Agrmt.*
1	XGB	0,858	0,862	0,881	0,668	3	2413	3538	68,20%
2	XGB	0,851	0,859	0,852	0,690	4	2205	3538	62,32%
3	XGB	0,862	0,864	0,867	0,683	3	2249	3538	63,57%
4	XGB	0,865	0,863	0,886	0,790	3	2369	3538	66,96%
5	XGB	0,861	0,866	0,870	0,710	3	2345	3538	66,28%
6	XGB	0,857	0,862	0,878	0,660	4	2291	3538	64,75%
7	XGB	0,856	0,863	0,892	0,753	3	2591	3538	73,23%
8	XGB	0,865	0,860	0,893	0,710	3	2525	3538	71,37%
9	XGB	0,855	0,857	0,884	0,682	3	2414	3538	68,23%
10	XGB	0,861	0,863	0,865	0,782	3	2264	3538	63,99%
Média		0,8591	0,8619	0,8768	0,7128				66,89%
p-value				7,09E-03	2,82E-06				

Em relação ao desempenho preditivo apresentado nas Tabelas 2, 3 e 4, a técnica XGB apresentou os melhores valores médios para a métrica AUC, tanto no conjunto de teste com $\overline{AUC_{\text{teste}}} = 0,8591$ quanto no conjunto de teste 2 com $\overline{AUC_{\text{teste2}}} = 0,8619$. Em seguida, aparece RF com $\overline{AUC_{\text{teste}}} = 0,8518$ e $\overline{AUC_{\text{teste2}}} = 0,8536$. Por último, a técnica

Tabela 4. Resultados das execuções da técnica Redes Neurais Artificiais (RNA).

Exec.	Técnica	teste	teste 2	*Agrmt.*	*Disagrmt.*	*Clusters*	Novo suporte	Suporte	*% Agrmt.*
1	RNA	0,840	0,848	0,889	0,744	3	1824	3538	51,55%
2	RNA	0,838	0,842	0,737	0,798	4	1777	3538	50,23%
3	RNA	0,838	0,842	0,678	0,802	4	1781	3538	50,34%
4	RNA	0,849	0,837	0,670	0,796	3	1812	3538	51,22%
5	RNA	0,841	0,847	0,738	0,795	4	2005	3538	56,67%
6	RNA	0,848	0,840	0,661	0,816	4	1705	3538	48,19%
7	RNA	0,835	0,842	0,677	0,794	4	1795	3538	50,73%
8	RNA	0,851	0,840	0,634	0,800	4	1562	3538	44,15%
9	RNA	0,838	0,835	0,808	0,732	4	2261	3538	63,91%
10	RNA	0,836	0,836	0,679	0,781	4	1702	3538	48,11%
Média		0,8414	0,8409	0,7171	0,7858				51,51%
p-value				6,52E-04	1,13E-04				

RNA obteve valores de $\overline{AUC_{\text{teste}}} = 0,8414$ e $\overline{AUC_{\text{teste2}}} = 0,8409$. Todas essas médias são significativamente diferentes entre si, com *p-value* bem abaixo de 0,05, conforme pode ser conferido no material disponibilizado na Seção Disponibilidade de Artefatos.

Nos resultados referentes à clusterização de explicações, observou-se que a técnica RF sempre obteve 3 clusters, enquanto XGB obteve 4 clusters em duas das execuções, e a técnica RNA variou entre 3 e 4 clusters. O desempenho médio do conjunto *agreement* foi superior ao do conjunto teste 2 tanto para RF quanto para XGB, com valores de *p-value* inferiores a 0,05. Ao mesmo tempo, o desempenho médio do conjunto *disagreement* foi inferior ao do teste 2 para RF e XGB, com *p-value* menor que 0,05. Por outro lado, a técnica RNA apresentou tanto o *agreement* quanto o *disagreement* com valores de desempenho médio significativamente inferiores ao do conjunto de teste 2.

Ao verificar a métrica de interpretabilidade "percentual de *agreement*", observa-se que o percentual médio de *agreement* foi maior para RF, seguido de XGB e RNA, sempre significativamente diferentes entre si, com *p-value* bem inferior a 0,05.

Para início da análise em relação à justiça algorítmica, apresentam-se as Tabelas 5, 6, 7 e 8 com os resultados de cada técnica no contexto dos grupos sociodemográficos sexo, raça, cota e origem em escola pública. Cada tabela refere-se a um dos atributos sensíveis e contém as informações organizadas da seguinte forma: para cada técnica, são listados os valores de desempenho médio do modelo nos conjuntos teste 2, *agreement* e *disagreement*. A primeira coluna apresenta o desempenho médio global para determinado conjunto, o que significa que considera todas as instâncias deste conjunto independentemente da classe no atributo sensível. As demais colunas apresentam o desempenho considerando as classes individuais do atributo sensível. A última coluna, $\overline{\sigma}$, apresenta o desvio padrão médio entre as classes do atributo sensível. Importante destacar que todos os desempenhos apresentados são médios em relação às 10 execuções, e os valores de *p-value* são sempre dos conjuntos *agreement* ou *disagreement* em relação ao conjunto de teste 2.

Para fins de exemplo, será detalhada a Tabela 6 para o atributo sensível raça. A tabela apresenta uma região para cada uma das técnicas (RF, XGB e RNA). A técnica RF apresentou desempenho global médio de $\overline{AUC_{\text{teste2}}} = 0,8536$ no conjunto de teste 2, $\overline{AUC_{\text{ag}}} = 0,8847$ no conjunto *agreement* e $\overline{AUC_{\text{dis}}} = 0,6697$ no conjunto *disagree-*

Tabela 5. Justiça algorítmica para o atributo sexo.

Técnica	Análise	Global	Masculino	Feminino	$\overline{\sigma}$
RF	teste 2	0,8536	0,8490	0,8526	6,22E-03
	Agrmt.	0,8847	0,8833	0,8841	6,36E-03
	p-value	9,11E-07	1,84E-05	2,33E-08	9,57E-01
	Disagrmt.	0,6697	0,6470	0,6750	2,52E-02
	p-value	1,03E-09	0,00E+00	1,42E-08	8,30E-03
XGB	teste 2	0,8619	0,8575	0,8610	5,30E-03
	Agrmt.	0,8768	0,8733	0,8788	9,69E-03
	p-value	7,09E-03	2,17E-02	9,95E-04	8,72E-02
	Disagrmt.	0,7128	0,6928	0,7207	2,50E-02
	p-value	2,82E-06	1,39E-06	7,38E-06	5,43E-03
RNA	teste 2	0,8409	0,8424	0,8356	4,81E-03
	Agrmt.	0,7171	0,7218	0,7112	1,27E-02
	p-value	6,52E-04	1,28E-03	4,82E-04	2,49E-02
	Disagrmt.	0,7858	0,7869	0,7845	1,26E-02
	p-value	1,13E-04	6,86E-04	2,61E-04	4,21E-03

ment. O desempenho global médio no *agreement* é significativamente[11] superior ao do teste 2, com *p-value* $= 9,11E-07$. Ao mesmo tempo, o desempenho global médio no *disagreement* é significativamente inferior ao do teste 2, com *p-value* $= 1,03E-09$.

Tabela 6. Justiça algorítmica para o atributo raça.

Técnica	Análise	Global	Branco	Pardo	Preto	PNDC[12]	Indígena	Amarelo	$\overline{\sigma}$
RF	teste 2	0,8536	0,8563	0,8222	0,8503	0,8678	0,8853	0,7665	7,76E-02
	Agrmt.	0,8847	0,8891	0,8506	0,8717	0,8951	0,8332	0,7558	1,06E-01
	p-value	9,11E-07	7,65E-08	1,42E-03	1,19E-02	1,34E-02	N/A	8,35E-01	1,00E-01
	Disagrmt.	0,6697	0,6693	0,6483	0,6680	0,6670	0,6667	0,8755	1,51E-01
	p-value	1,03E-09	1,17E-09	1,45E-07	6,83E-06	9,55E-04	N/A	2,77E-02	3,31E-02
XGB	teste 2	0,8619	0,8655	0,8306	0,8564	0,8660	0,8541	0,7663	7,83E-02
	Agrmt.	0,8768	0,8821	0,8409	0,8633	0,8862	0,8125	0,7556	1,09E-01
	p-value	7,09E-03	3,82E-03	3,37E-01	3,05E-01	2,76E-02	N/A	8,88E-01	2,88E-01
	Disagrmt.	0,7128	0,7124	0,7101	0,7115	0,6759	0,8333	0,7437	1,24E-01
	p-value	2,82E-06	3,71E-06	1,74E-07	8,87E-04	1,82E-04	N/A	7,38E-01	4,73E-02
RNA	teste 2	0,8409	0,8466	0,7993	0,8337	0,8376	0,8471	0,7081	8,20E-02
	Agrmt.	0,7171	0,7210	0,6862	0,6880	0,7429	0,8	0,6852	1,29E-01
	p-value	6,52E-04	5,88E-04	3,08E-03	1,13E-03	1,31E-02	N/A	N/A	1,36E-01
	Disagrmt.	0,7858	0,7928	0,7490	0,7868	0,7913	0,6665	0,5928	1,35E-01
	p-value	1,13E-04	2,52E-04	2,19E-03	1,90E-02	3,75E-02	N/A	N/A	5,47E-03

A mesma análise é possível para cada uma das classes no atributo sensível. Por exemplo, o desempenho entre os pardos também é significativamente superior no conjunto *agreement* quando comparado ao conjunto teste 2, e significativamente inferior no *disagreement* quando comparado ao conjunto teste 2. A coluna $\overline{\sigma}$ apresenta o desvio padrão médio do desempenho entre as diferentes classes do atributo sensível. Neste caso, pode-se observar que o desvio padrão médio no teste 2 foi $\overline{\sigma} = 7,76E-02$, enquanto no *agreement* foi $\overline{\sigma} = 1,06E-01$ e no *disagreement* foi $\overline{\sigma} = 1,51E-01$.

Ao aplicar o teste T nos conjuntos de desvio padrão das 10 execuções que geraram esses valores médios, verificou-se que o desvio padrão médio no *agreement* não mudou significativamente em relação ao teste 2 (*p-value* $= 1,00E-01$). Entretanto, foi significativamente superior no *disagreement* (*p-value* $= 3,31E-02$). Quando o desvio padrão médio não sofre mudanças significativas de um conjunto para o outro, a justiça algorítmica se mantém similar ($=$). Se o desvio padrão médio aumenta, a justiça diminui ($\downarrow$). No caso do desvio padrão médio diminuir, a justiça aumenta ($\uparrow$).

É possível resumir a análise dessa tabela, no que diz respeito à técnica RF, da seguinte forma. A explicabilidade gerou um conjunto *agreement* com desempenho médio superior ao teste 2, e um conjunto *disagreement* com desempenho médio inferior ao teste 2. Essa característica de desempenho também é observada no contexto de cada classe

[11] O limiar de comparação para *p-value* será sempre $0,05$. [12] Prefere não declarar.

Tabela 7. Justiça algorítmica para o atributo cota.

Técnica	Análise	Global	AC	L01	L02	L05	L06	$\overline{\sigma}$
RF	teste 2	0,8536	0,8588	0,8390	0,8351	0,8455	0,8449	1,92E-02
	Agrmt.	0,8847	0,8878	0,8738	0,8587	0,8825	0,8698	2,04E-02
	p-value	9,11E-07	1,03E-04	7,91E-07	1,46E-02	3,12E-05	4,21E-04	6,15E-01
	Disagrmt.	0,6697	0,6724	0,6742	0,6825	0,6478	0,6575	5,00E-02
	p-value	1,03E-09	0,00E+00	2,55E-06	4,30E-07	7,11E-06	9,24E-06	5,41E-04
XGB	teste 2	0,8619	0,8664	0,8485	0,8448	0,8579	0,8481	1,65E-02
	Agrmt.	0,8768	0,8813	0,8635	0,8481	0,8767	0,8509	2,79E-02
	p-value	7,09E-03	9,07E-03	1,86E-02	8,11E-01	4,83E-02	7,80E-01	2,44E-02
	Disagrmt.	0,7128	0,7107	0,7215	0,7131	0,7014	0,7130	3,60E-02
	p-value	2,82E-06	5,37E-06	3,06E-06	8,62E-04	1,53E-06	2,03E-05	1,07E-02
RNA	teste 2	0,8409	0,8477	0,8178	0,8142	0,8452	0,8205	2,25E-02
	Agrmt.	0,7171	0,7181	0,7149	0,6912	0,7341	0,6970	3,03E-02
	p-value	6,52E-04	4,47E-04	2,20E-03	2,14E-03	1,80E-03	9,09E-04	4,84E-02
	Disagrmt.	0,7858	0,7935	0,7594	0,7651	0,7806	0,7555	3,62E-02
	p-value	1,13E-04	2,83E-05	6,46E-03	1,09E-02	2,93E-03	5,43E-02	7,15E-02

Tabela 8. Justiça algorítmica para o atributo origem em escola pública.

Técnica	Análise	Global	Não	Sim	$\overline{\sigma}$
RF	teste 2	0,8536	0,8616	0,8500	1,12E-02
	Agrmt.	0,8847	0,8941	0,8804	1,18E-02
	p-value	9,11E-07	5,85E-06	1,10E-06	7,68E-01
	Disagrmt.	0,6697	0,6944	0,6596	3,18E-02
	p-value	1,03E-09	2,01E-08	5,56E-09	2,15E-03
XGB	teste 2	0,8619	0,8694	0,8586	8,91E-03
	Agrmt.	0,8768	0,8925	0,8679	1,84E-02
	p-value	7,09E-03	2,14E-03	7,01E-02	1,03E-02
	Disagrmt.	0,7128	0,7204	0,7076	1,99E-02
	p-value	2,82E-06	9,08E-08	9,22E-06	2,35E-02
RNA	teste 2	0,8409	0,8499	0,8370	1,15E-02
	Agrmt.	0,7171	0,7243	0,7129	1,90E-02
	p-value	6,52E-04	4,35E-04	7,68E-04	6,83E-02
	Disagrmt.	0,7858	0,7894	0,7848	1,46E-02
	p-value	1,13E-04	1,18E-04	3,47E-04	1,80E-01

do atributo sensível. Por fim, pode-se afirmar que a explicabilidade gerou um conjunto *agreement* com justiça algorítmica similar ao conjunto teste 2, enquanto o conjunto *disagreement* se mostrou menos justo que o teste 2.

Tabela 9. Sumarização Interpretabilidade vs. justiça algorítmica

Técnica	Conjunto	Desempenho	Sexo	Raça	Cota	Escola pública	*% Agrmt.*
RF	*Agrmt.*	Superior	=	=	=	=	72,80%
	Disagrmt.	Inferior	↓	↓	↓	↓	
XGB	*Agrmt.*	Superior	=	=	↓	↓	66,89%
	Disagrmt.	Inferior	↓	↓	↓	↓	
RNA	*Agrmt.*	Inferior	↓	=	↓	=	51,51%
	Disagrmt.	Inferior	↓	↓	=	=	

Ao realizar essa mesma análise para todas as técnicas das Tabelas 5, 6, 7 e 8, apresenta-se um resumo para relacionar os resultados da explicabilidade com a justiça algorítmica, conforme detalhado na Tabela 9. Pode-se observar que a técnica RF gerou um conjunto *agreement* com desempenho médio superior, ao mesmo tempo, mantendo a justiça algorítmica similar entre as classes dos diferentes atributos sensíveis. Concomitantemente, a técnica RF apresentou o melhor valor médio para a métrica percentual de *agreement*. Ainda para a técnica RF, o conjunto *disagreement* permite identificar uma região com menor desempenho e menor justiça para todos os atributos sensíveis.

A técnica XGB conseguiu obter um *agreement* com desempenho médio superior, entretanto, não conseguiu manter a justiça algorítmica nos atributos cota e escola pública. Em relação ao *disagreement*, identificou uma região com menor desempenho e menor justiça para todos os atributos sensíveis. Por último, a técnica RNA não obteve *agreement* com desempenho superior, ao mesmo tempo que não manteve a justiça nos atributos sexo

e cota. Já para o conjunto *disagreement*, identificou uma região com menor desempenho e apresentando menor justiça para os atributos sexo e raça.

A partir dos resultados apresentados e retomando a questão de pesquisa proposta na Seção 1 ("De que forma a interpretabilidade de um modelo de AM está ligada ao desempenho preditivo entre grupos sociodemográficos?"), demonstra-se que, por meio da interpretabilidade de modelos, é possível obter conjuntos (espaços amostrais) onde o modelo apresenta melhores desempenhos no contexto de diferentes classes de grupos sociodemográficos. Ao mesmo tempo, foi possível analisar de que forma a justiça algorítmica varia entre diferentes conjuntos de instâncias. A Tabela 9 resume exatamente essa relação entre interpretabilidade e justiça algorítmica, demonstrando em que situações a justiça do modelo aumenta, diminui ou se mantém similar, ao passo que associa essas variações à métrica de percentual de *agreement*.

5. Conclusões

Este trabalho propôs investigar a relação entre interpretabilidade de modelos e a justiça algorítmica (*fairness*) no contexto da predição precoce de evasão de estudantes no ensino superior. Visando responder à questão de pesquisa proposta, apresentou-se uma evolução do método de clusterização de explicações LIME por meio da inclusão de validação estatística e avaliação de desempenho no contexto de grupos sociodemográficos, a saber, sexo, raça, cota e origem em escola pública.

Propôs-se também medir a justiça algorítmica a partir do desvio padrão do desempenho entre as diferentes classes de uma variável sensível. O conjunto de dados utilizado consistiu em 17.689 instâncias de estudantes de cursos de graduação, incluindo informações sociodemográficas e de desempenho acadêmico referente ao primeiro semestre do aluno no seu curso. A análise e discussão dos resultados permitiu relacionar diretamente o desempenho em relação à métrica "percentual de *agreement*" e a justiça algorítmica, identificando em que situações essa justiça se mantém similar, diminui ou aumenta.

Esses achados sugerem que a interpretabilidade pode identificar regiões do espaço amostral com melhor desempenho e justiça similar ao conjunto de teste, enquanto outras regiões apresentam desempenho e justiça inferiores. Tais relações foram sumarizadas e apresentadas tabularmente, respondendo à questão de pesquisa na Seção 4. A técnica RF destacou-se ao apresentar: maior percentual de *agreement*; conjunto *agreement* com desempenho médio superior (global e por grupo sociodemográfico) e justiça similar em todos atributos sensíveis; conjunto *disagreement* com desempenho médio inferior (global e por grupo sociodemográfico) e justiça inferior para todos atributos sensíveis.

Por fim, a análise contribui para entender como ajustar modelos de IA para melhorar a sua transparência em contextos educacionais. Como trabalho futuro pretende-se avaliar o método em relação a outras métricas de justiça e fidelidade de explicações.

Disponibilidade de Artefatos

Os resultados e análises deste estudo estão disponíveis no repositório *GitHub*, `https://github.com/cassiocarvalho/interpretability-and-fairness`

Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

Referências

Adadi, A. and Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138–52160.

Afrin, F., Hamilton, M., and Thevathyan, C. (2022). On the explanation of ai-based student success prediction. In Groen, D., de Mulatier, C., Paszynski, M., Krzhizhanovskaya, V. V., Dongarra, J. J., and Sloot, P. M. A., editors, *Computational Science – ICCS 2022*, pages 252–258, Cham. Springer International Publishing.

Alamri, R. and Alharbi, B. (2021). Explainable student performance prediction models: A systematic review. *IEEE Access*, 9:33132–33143.

Alves, G., Bhargava, V., Couceiro, M., and Napoli, A. (2021). Making ml models fairer through explanations: The case of limeout. In van der Aalst, W. M. P., Batagelj, V., Ignatov, D. I., Khachay, M., Koltsova, O., Kutuzov, A., Kuznetsov, S. O., Lomazova, I. A., Loukachevitch, N., Napoli, A., Panchenko, A., Pardalos, P. M., Pelillo, M., Savchenko, A. V., and Tutubalina, E., editors, *Analysis of Images, Social Networks and Texts*, pages 3–18, Cham. Springer International Publishing.

Alwarthan, S., Aslam, N., and Khan, I. U. (2022). An explainable model for identifying at-risk student at higher education. *IEEE Access*, 10:107649 – 107668. All Open Access, Gold Open Access.

Araujo, I. (2021). Uma revisao sobre o uso de frameworks de interpretabilidade em aprendizado de máquina. In *Anais do XIV Encontro Unificado de Computaçao do Piaui e XI Simposio de Sistemas de Informaçao*, pages 105–112, Porto Alegre, RS, Brasil. SBC.

Baker, R. et al. (2010). Data mining for education. *International encyclopedia of education*, 7(3):112–118.

Baker, R. S. and Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, 1(1):3–17.

Bakhshinategh, B., Zaiane, O. R., ElAtia, S., and Ipperciel, D. (2018). Educational data mining applications and tasks: A survey of the last 10 years. *Education and Information Technologies*, 23:537–553.

Bhargava, V., Couceiro, M., and Napoli, A. (2020). Limeout: An ensemble approach to improve process fairness. In Koprinska, I., Kamp, M., Appice, A., Loglisci, C., Antonie, L., Zimmermann, A., Guidotti, R., Özgöbek, Ö., Ribeiro, R. P., Gavaldà, R., Gama, J., Adilova, L., Krishnamurthy, Y., Ferreira, P. M., Malerba, D., Medeiros, I., Ceci, M., Manco, G., Masciari, E., Ras, Z. W., Christen, P., Ntoutsi, E., Schubert, E., Zimek, A., Monreale, A., Biecek, P., Rinzivillo, S., Kille, B., Lommatzsch, A., and Gulla, J. A., editors, *ECML PKDD 2020 Workshops*, pages 475–491, Cham. Springer International Publishing.

Carvalho, C., Mattos, J., and Aguiar, M. (2023). Avaliação da interpretabilidade de modelos por meio da clusterização de explicações no contexto da predição de evasão no

ensino superior. In *Anais do XXXIV Simpósio Brasileiro de Informática na Educação*, pages 1191–1201, Porto Alegre, RS, Brasil. SBC.

Carvalho, D. V., Pereira, E. M., and Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8).

Chou, T.-N. (2023). Apply an integrated responsible ai framework to sustain the assessment of learning effectiveness. volume 2, page 142 – 149. All Open Access, Hybrid Gold Open Access.

Colak Oz, H., Güven, Ç., and Nápoles, G. (2023). School dropout prediction and feature importance exploration in malawi using household panel data: machine learning approach. *Journal of Computational Social Science*, 6(1):245 – 287.

Dsilva, V., Schleiss, J., and Stober, S. (2023). Trustworthy academic risk prediction with explainable boosting machines. In Wang, N., Rebolledo-Mendez, G., Matsuda, N., Santos, O. C., and Dimitrova, V., editors, *Artificial Intelligence in Education*, pages 463–475, Cham. Springer Nature Switzerland.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, page 214–226, New York, NY, USA. Association for Computing Machinery.

Gardner, J., Brooks, C., and Baker, R. (2019). Evaluating the fairness of predictive student models through slicing analysis. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, LAK19, page 225–234, New York, NY, USA. Association for Computing Machinery.

Grgic-Hlaca, N., Zafar, M. B., Gummadi, K. P., and Weller, A. (2016). The case for process fairness in learning: Feature selection for fair decision making. In *NIPS symposium on machine learning and the law*, page 11. Barcelona, Spain.

Hardt, M., Price, E., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Hegazi, M. O. and Abugroon, M. A. (2016). The state of the art on educational data mining in higher education. *International Journal of Computer Trends and Technology*, 31(1):46–56.

Hu, Q. and Rangwala, H. (2020). Towards fair educational data mining: A case study on detecting at-risk students. page 431 – 437.

Jeon, B., Shafran, E., Breitfeller, L., Levin, J., and Rosé, C. P. (2019). Time-series insights into the process of passing or failing online university courses using neural-induced interpretable student states. page 330 – 335.

Kantorski, G., Martins, R., Balejo, A., and Frick, M. (2023). Mineração de dados educacionais para predição da evasão em cursos de graduação presenciais no ensino superior. In *Anais do XXXIV Simpósio Brasileiro de Informática na Educação*, pages 1133–1142, Porto Alegre, RS, Brasil. SBC.

Kim, B., Khanna, R., and Koyejo, O. O. (2016). Examples are not enough, learn to criticize! criticism for interpretability. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Kung, C. and Yu, R. (2020). Interpretable models do not compromise accuracy or fairness in predicting college success. In *Proceedings of the Seventh ACM Conference on Learning @ Scale*, L@S '20, page 413–416, New York, NY, USA. Association for Computing Machinery.

Kusner, M. J., Loftus, J., Russell, C., and Silva, R. (2017). Counterfactual fairness. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Kuzilek, J., Hlosta, M., and Zdrahal, Z. (2017). Open university learning analytics dataset. *Scientific Data*, 4(1).

Le Quy, T., Nguyen, T. H., Friege, G., and Ntoutsi, E. (2023). Evaluation of group fairness measures in student performance prediction problems. In Koprinska, I., Mignone, P., Guidotti, R., Jaroszewicz, S., Fröning, H., Gullo, F., Ferreira, P. M., Roqueiro, D., Ceddia, G., Nowaczyk, S., Gama, J., Ribeiro, R., Gavaldà, R., Masciari, E., Ras, Z., Ritacco, E., Naretto, F., Theissler, A., Biecek, P., Verbeke, W., Schiele, G., Pernkopf, F., Blott, M., Bordino, I., Danesi, I. L., Ponti, G., Severini, L., Appice, A., Andresini, G., Medeiros, I., Graça, G., Cooper, L., Ghazaleh, N., Richiardi, J., Saldana, D., Sechidis, K., Canakoglu, A., Pido, S., Pinoli, P., Bifet, A., and Pashami, S., editors, *Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pages 119–136, Cham. Springer Nature Switzerland.

Linardatos, P., Papastefanopoulos, V., and Kotsiantis, S. (2021). Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1).

Matetic, M. (2019). Mining learning management system data using interpretable neural networks. page 1282 – 1287.

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

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.

Molnar, C. (2022). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. 2 edition.

Oliveira, R. d. S. and Medeiros, F. P. A. d. (2024). Modelo de predição de evasão escolar com base em dados de autoavaliação de cursos de graduação. *Revista Brasileira de Informática na Educação*, 32:1–21.

Pei, B. and Xing, W. (2022). An interpretable pipeline for identifying at-risk students. *Journal of Educational Computing Research*, 60(2):380–405.

Peña-Ayala, A. (2014). *Educational Data Mining: Applications and Trends*, volume 524. Springer International Publishing.

Qu, Y., Li, F., Li, L., Dou, X., and Wang, H. (2022). Can we predict student performance based on tabular and textual data? *IEEE Access*, 10:86008 – 86019.

Rachha, A. and Seyam, M. (2023). Explainable ai in education : Current trends, challenges, and opportunities. In *SoutheastCon 2023*, pages 232–239.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Romero, C. and Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(6):601–618.

Romero, C. and Ventura, S. (2013). Data mining in education. *WIREs Data Mining and Knowledge Discovery*, 3(1):12–27.

Romero, C. and Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *WIREs Data Mining and Knowledge Discovery*, 10(3):e1355.

Sahlaoui, H., Alaoui, E. A. A., Agoujil, S., and Nayyar, A. (2023). An empirical assessment of smote variants techniques and interpretation methods in improving the accuracy and the interpretability of student performance models. *Education and Information Technologies*, 29(5):5447–5483.

Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., and Müller, K.-R. (2019). *Explainable AI: interpreting, explaining and visualizing deep learning*, volume 11700. Springer Nature.

Saxena, N. A., Huang, K., DeFilippis, E., Radanovic, G., Parkes, D. C., and Liu, Y. (2019). How do fairness definitions fare? examining public attitudes towards algorithmic definitions of fairness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 99–106.

Silva Filho, R. L. L. e., Motejunas, P. R., Hipólito, O., and Lobo, M. B. d. C. M. (2007). A evasão no ensino superior brasileiro. *Cadernos de Pesquisa*, 37(132):641–659.

Tsiakmaki, M. and Ragos, O. (2021). A case study of interpretable counterfactual explanations for the task of predicting student academic performance. page 120 – 125.

Vieira, C. and Digiampietri, L. (2022). Machine learning post-hoc interpretability: a systematic mapping study. In *Anais do XVIII Simposio Brasileiro de Sistemas de Informaçao*, Porto Alegre, RS, Brasil. SBC.

Xiang, F., Zhang, X., Cui, J., Carlin, M., and Song, Y. (2022). Algorithmic bias in a student success prediction models: Two case studies. In *2022 IEEE International Conference on Teaching, Assessment and Learning for Engineering (TALE)*, pages 310–315.

Xiao, W., Ji, P., and Hu, J. (2022). A survey on educational data mining methods used for predicting students' performance. *Engineering Reports*, 4(5):e12482.