LIME-EC: UM MÉTODO AGNÓSTICO PARA QUANTIFICAÇÃO DA INTERPRETABILIDADE DE MODELOS POR MEIO DA CLUSTERIZAÇÃO DE EXPLICAÇÕES

CÁSSIO SOARES CARVALHO[1]; MARILTON SANCHOTENE DE AGUIAR[2]; JÚLIO CARLOS BALZANO DE MATTOS[3]

[1]*Universidade Federal de Pelotas – cassio.carvalho@inf.ufpel.edu.br*
[2]*Universidade Federal de Pelotas – marilton@inf.ufpel.edu.br*
[3]*Universidade Federal de Pelotas – julius@inf.ufpel.edu.br*

1. INTRODUÇÃO

O uso de técnicas de Aprendizado de Máquina (AM) permite a obtenção de modelos de predição cada vez mais eficientes. Ao mesmo tempo, em que a Inteligência Artificial (IA) torna-se onipresente em nossas vidas, os modelos utilizados por aplicações de IA são cada vez mais complexos e de difícil compreensão (CARVALHO; PEREIRA; CARDOSO, 2019; LINARDATOS; PAPASTEFANOPOULOS; KOTSIANTIS, 2021). Dada a ampla utilização de sistemas com tais características, emergem preocupações a respeito da transparência dos modelos e a potencial introdução de vieses (SAXENA; HUANG; DEFILIPPIS; RADANOVIC; PARKES; LIU, 2019).

A transparência de modelos está relacionada à interpretabilidade[1], a qual representa o grau que um ser humano pode entender a causa de uma decisão ou prever consistentemente o resultado de um modelo (MOLNAR, 2022). Já o viés, está relacionado à justiça algorítmica, a qual pode ser entendida como ausência de qualquer viés baseado em características inerentes a um indivíduo e irrelevantes na tomada de decisão (SAXENA; HUANG; DEFILIPPIS; RADANOVIC; PARKES; LIU, 2019).

A Mineração de Dados Educacionais (MDE) realiza descobertas em dados provenientes de ambientes educacionais para compreender o comportamento de alunos no processo de aprendizagem (BAKER et al., 2010), tendo a predição de desempenho como uma das aplicações mais relevantes (BAKHSHINATEGH; ZAIANE; ELATIA; IPPERCIEL, 2018). O estado da arte da interpretabilidade no contexto da MDE é majoritariamente limitado pela aplicação de métodos, necessitando de trabalhos com foco na comparação de modelos sob aspectos da interpretabilidade.

Nesse sentido, realiza um estudo de caso no contexto da evasão estudantil, reconhecidamente um problema internacional que representa desperdícios sociais, acadêmicos e econômicos (SILVA FILHO; MOTEJUNAS; HIPÓLITO; LOBO, 2007). Naturalmente, a evasão tem sido objeto frequente de estudo em Instituições de Ensino brasileiras (COLPO; PRIMO; AGUIAR, 2021).

Este estudo parte da hipótese de que aspectos de interpretabilidade podem subsidiar o processo de escolha por modelos de predição adequados. Nessa perspectiva, foram definidas as seguintes Questões de Pesquisa (QP): “Como explicar modelos de predição, para compará-los quantitativamente sob aspectos de interpretabilidade?”; “De que forma a explicabilidade de um modelo pode estar relacionada ao desempenho preditivo de novas instâncias?”; “De que forma a

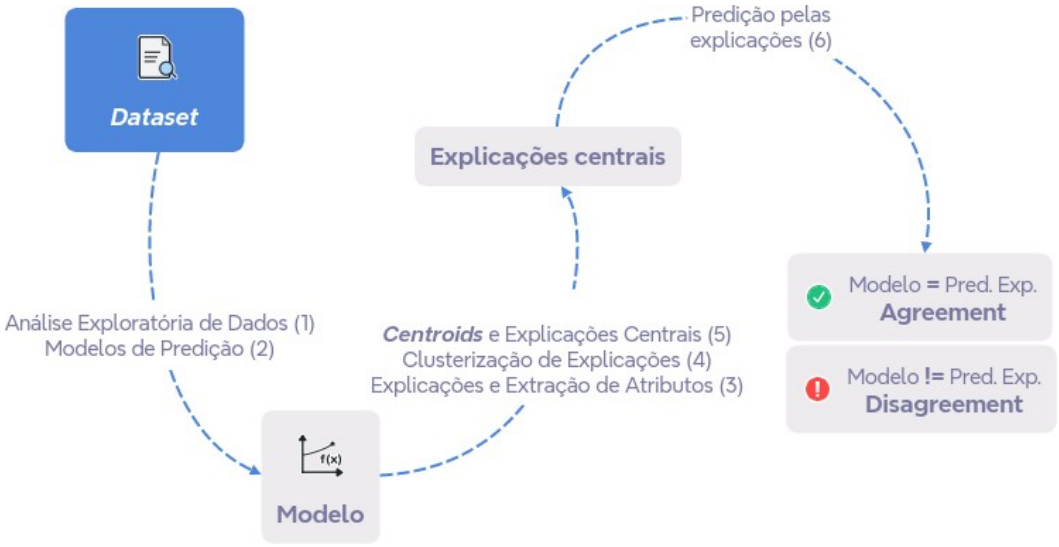
1 Os termos interpretabilidade e explicabilidade são frequentemente utilizados de forma intercambiável.

explicabilidade de um modelo pode estar relacionada a qualidade das explicações de novas predições?"; "De que forma a explicabilidade de modelos de predição pode estar associada a justiça algorítmica?".

2. METODOLOGIA

Nesse contexto, propõe-se o LIME *Explanation Clustering* (LIME-EC), um método de interpretabilidade agnóstico que permite comparar quantitativamente modelos sob aspectos de interpretabilidade, subsidiando o processo de escolha por modelos de predição adequados. A Figura 1 apresenta resumidamente o método proposto, o qual consiste em analisar um conjunto de explicações por meio de aprendizado de máquina não supervisionado, produzindo explicações centrais que descrevem o comportamento do modelo original.

Figura 1: Etapas do método LIME-EC



Ao utilizar essas explicações centrais de forma independente para predizer novas instâncias, dois conjuntos são obtidos: *agreement* e *disagreement*. O conjunto *agreement* contém as instâncias nas quais a predição do modelo é igual à predição utilizando as explicações centrais, enquanto o conjunto *disagreement* contém as instâncias nas quais a predição do modelo difere da predição pelas explicações centrais. O percentual de concordância (percentual de *agreement*) é, portanto, apresentado como uma métrica de interpretabilidade.

Para avaliar a interpretabilidade e sua relação com a justiça algorítmica, são calculados: Desempenho médio de cada técnica nos conjuntos de teste[2], teste2[3], *agreement* e *disagreement;* Percentual médio de *agreement* e *disagreement*; Desempenho médio de cada técnica no contexto das classes de cada uma das variáveis sensíveis, como sexo, raça, cota e origem em escola pública; Desvio padrão do desempenho de cada execução para as diferentes classes de cada grupo sócio demográfico; Desvio padrão médio de desempenho de cada técnica para cada grupo sócio demográfico. Esse valor passa a ser uma medida de justiça em relação a um grupo sócio demográfico (definido por variável sensível).

São realizados também: Avaliação dos desempenhos médios (global e por grupo sociodemográfico) para verificar se há diferença significativa entre os

2 Conjunto de instâncias para avaliar os modelos e para gerar explicações.
3 Conjunto de instâncias para avaliar o método de interpretabilidade,

resultados dos conjuntos *agreement* e *disagreement* em relação ao conjunto teste2; Verificação se há diferença significativa entre os percentuais médios de *agreement* das diferentes técnicas; Verificação se há diferença significativa entre as médias dos desvios padrões em cada grupo sócio demográfico.

3. RESULTADOS E DISCUSSÃO

A primeira avaliação do método LIME-EC foi apresentada em CARVALHO; MATTOS; AGUIAR (2023). Um segundo artigo, investigando a relação entre interpretabilidade e justiça algorítmica, foi aceito e será apresentado no mesmo simpósio em 2024. Resultados das publicações estão disponíveis no GitHub[45].

Tabela 1: 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	↓	↓	=	=	

A Tabela 1 relaciona os resultados da explicabilidade com a justiça algorítmica. Observa-se 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 obteve um *agreement* com desempenho médio superior, entretanto, sem 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.

4. CONCLUSÕES

A proposição de um método de interpretabilidade e uma métrica de interpretabilidade é parte central para atingir os objetivos e responder às QPs propostas. Nesse sentido, apresentou-se o LIME-EC, um método de interpretabilidade que permite encontrar explicações centrais que melhor descrevem o comportamento de um modelo de predição em diferentes regiões do espaço de busca. Os resultados demonstram que a métrica de interpretabilidade permite encontrar regiões onde o modelo original apresenta melhor desempenho preditivo e avaliar a qualidade das explicações em novas predições. A análise também permitiu relacionar diretamente o desempenho em relação à métrica

4 https://github.com/cassiocarvalho/clustering-lime-explanations
5 https://github.com/cassiocarvalho/interpretability-and-fairness

"percentual de *agreement*" e à justiça algorítmica, identificando em que situações essa justiça se mantém similar, diminui ou aumenta. Os 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. O estudo contribui para entender como ajustar modelos de IA para melhorar a sua transparência em contextos educacionais. Acredita-se que o método LIME-EC tenha o potencial de quantificar a interpretabilidade de um modelo de predição, de forma a subsidiar a escolha por modelos adequados. Como trabalho futuro pretende-se ampliar o escopo das avaliações em algumas frentes como avaliação de outras técnicas de AM, ajuste fino de hiperparâmetros, avaliação de métricas de justiça, mitigação de viés e validação do método em um contexto diferente ao da MDE.

5. REFERÊNCIAS BIBLIOGRÁFICAS

BAKER, R. et al. Data mining for education. International encyclopedia of education, [S.l.], v.7, n.3, p.112–118, 2010.

BAKHSHINATEGH, B.; ZAIANE, O. R.; ELATIA, S.; IPPERCIEL, D. Educational data mining applications and tasks: A survey of the last 10 years. Education and Information Technologies, [S.l.], v.23, p.537–553, 2018.

CARVALHO, C.; MATTOS, J.; AGUIAR, M. Avaliação da interpretabilidade de modelos por meio da clusterização de explicações no contexto da predição de evasão no ensino superior. In: XXXIV SIMPóSIO BRASILEIRO DE INFORMáTICA NA EDUCAçãO, 2023, Porto Alegre, RS, Brasil. Anais. . . SBC, 2023. p.1191–1201.

CARVALHO, D. V.; PEREIRA, E. M.; CARDOSO, J. S. Machine Learning Interpretability: A Survey on Methods and Metrics. Electronics, [S.l.], v.8, n.8, 2019.

COLPO, M.; PRIMO, T.; AGUIAR, M. Predição da evasão estudantil: uma análise comparativa de diferentes representações de treino na aprendizagem de modelos genéricos. In: XXXII SIMPóSIO BRASILEIRO DE INFORMáTICA NA EDUCAçãO, 2021, Porto Alegre, RS, Brasil. Anais. . . SBC, 2021. p.873–884.

LINARDATOS, P.; PAPASTEFANOPOULOS, V.; KOTSIANTIS, S. Explainable AI: A Review of Machine Learning Interpretability Methods. Entropy, [S.l.], v.23, n.1, 2021.

MOLNAR, C. Interpretable Machine Learning. 2.ed. [S.l.: s.n.], 2022.

SAXENA, N. A.; HUANG, K.; DEFILIPPIS, E.; RADANOVIC, G.; PARKES, D. C.; LIU, Y. How do fairness definitions fare? Examining public attitudes towards algorithmic definitions of fairness. In: AAAI/ACM CONFERENCE ON AI, ETHICS, AND SOCIETY, 2019., 2019. Proceedings. . . [S.l.: s.n.], 2019. p.99–106.

SILVA FILHO, R. L. L. e.; MOTEJUNAS, P. R.; HIPóLITO, O.; LOBO, M. B. d. C. M. A evasão no ensino superior brasileiro. Cadernos de Pesquisa, [S.l.], v.37, n.132, p.641–659, Sep 2007.