Explicabilidade da Inteligência Artificial e Tomada de Decisão Educacional: Predição de Abandono Universitário

Lucas Della Giustina Schiochet, UCS, ldgschiochet@ucs.br, 0009-0001-1857-9343
Tiago Luiz Schmitz, UDESC, tiago.schmitz@udesc.br, 0009-0001-8389-4745
Carine Geltrudes Webber, UCS, cgwebber@ucs.br, 0000-0001-7778-6740

Resumo. *Este artigo apresenta o desenvolvimento e a avaliação de um produto educacional voltado à integração da Inteligência Artificial Explicável (XAI) no contexto educacional, aplicado à predição de abandono universitário. A pesquisa, de natureza aplicada e abordagem qualitativa, fundamenta-se na necessidade de promover transparência e compreensão nos modelos de IA utilizados na educação, aproximando práticas de análise preditiva de finalidades pedagógicas e institucionais. Foram empregados os métodos SHAP e LIME para interpretar predições de uma rede neural voltada à identificação de risco de evasão. A análise baseou-se nas dimensões de explicabilidade e interpretabilidade, considerando sua relevância cognitiva e prática para professores e gestores. Os resultados indicam que a XAI pode favorecer a compreensão dos fatores de risco e apoiar decisões mais informadas e éticas na gestão educacional. Apenas dados anônimos foram utilizados, respeitando os princípios éticos de pesquisa. Mais do que relatar um experimento técnico, o artigo propõe um produto educacional que instrumentaliza o uso pedagógico da explicabilidade da IA na formação e na prática docente.*

***Palavras-chave**: IA aplicada à Educação, IA Explicável, Redes Neurais Artificiais.*

Explainable Artificial Intelligence and Educational Decision-Making: University Dropout Prediction

Abstract. *This paper presents the development and evaluation of an educational product designed to integrate Explainable Artificial Intelligence (XAI) into the educational context, applied to university dropout prediction. The applied research, with a qualitative approach, is grounded in the need to promote transparency and understanding in AI models used in education, bridging predictive analysis with pedagogical and institutional purposes. SHAP and LIME methods were employed to interpret neural network predictions related to dropout risk. The analysis considered the dimensions of explainability and interpretability, emphasizing their cognitive and practical relevance for teachers and educational managers. The results indicate that XAI can enhance the understanding of risk factors and support more informed and ethical decision-making in educational management. Only anonymized data were used, ensuring ethical research standards. Beyond a technical experiment, this article proposes an educational product that fosters the pedagogical use of AI explainability in teaching and professional development.*

***Keywords:** AI in Education, Explainable AI, Artificial Neural Networks.*

1. Introdução

A Inteligência Artificial (IA) vem assumindo papel crescente na gestão educacional e na personalização das práticas pedagógicas, ampliando a capacidade das instituições de compreender e responder a fenômenos complexos como o abandono

escolar. Contudo, o uso de modelos preditivos baseados em algoritmos de aprendizagem de máquina ainda apresenta um desafio central: a falta de transparência nas decisões automatizadas, fenômeno conhecido como "caixa-preta". Essa opacidade compromete a confiança, a ética e a utilização pedagógica dos resultados, especialmente quando decisões institucionais podem afetar trajetórias estudantis.

Nesse contexto, a Inteligência Artificial Explicável (XAI) surge como um campo promissor para a educação, ao buscar tornar compreensíveis os processos de inferência de modelos complexos. Mais do que uma questão técnica, a explicabilidade tem implicações pedagógicas diretas, pois permite que educadores e gestores compreendam os fatores que influenciam a aprendizagem e o desempenho dos estudantes, convertendo previsões em ações educativas contextualizadas.

Diversos estudos recentes (Holmes et al., 2022; Luckin et al., 2024; OECD, 2025) reforçam que a integração ética e transparente da IA em contextos educacionais deve priorizar o protagonismo docente e o direito à compreensão das decisões automatizadas. Ao compreender "por que" e "como" um modelo chega a determinada previsão, é possível criar intervenções pedagógicas mais justas, precisas e sensíveis às diversidades sociocognitivas.

Neste artigo, investigam-se as potencialidades da explicabilidade da IA aplicada à predição de abandono universitário, por meio do uso de métodos SHAP e LIME integrados a um modelo de rede neural artificial (HAYKIN, 2009). Diferentemente de estudos puramente técnicos, este trabalho adota uma perspectiva educacional, compreendendo a explicabilidade como uma prática de apoio à tomada de decisão pedagógica e institucional. O estudo, portanto, propõe-se a analisar como a XAI pode contribuir para a interpretação dos fatores associados ao abandono e para o desenho de estratégias de permanência que dialoguem com princípios éticos e formativos da educação superior.

2. Fundamentação Teórica

A Inteligência Artificial Explicável (XAI) é uma abordagem voltada a tornar os sistemas de IA mais transparentes, auditáveis e compreensíveis para seus usuários. Em um cenário em que os modelos de IA são frequentemente percebidos como "caixas-pretas", a XAI oferece mecanismos interpretativos capazes de revelar a lógica subjacente às decisões algorítmicas, favorecendo a confiança, a ética e a conformidade regulatória (RIBEIRO; SAMEER, 2016; SETZU, 2021; LONGO et al., 2024).

Autores como Adadi e Berrada (2018) e Doran, Schulz e Besold (2017) destacam que a XAI tem papel estratégico na construção de sistemas de IA confiáveis, equilibrando desempenho técnico e compreensibilidade. No campo educacional, essa transparência é especialmente relevante: compreender as razões pelas quais um modelo prediz o risco de evasão, por exemplo, permite que docentes e gestores interpretem os dados de modo formativo e orientem intervenções pedagógicas mais contextualizadas (HOLMES et al., 2022; LUCKIN et al., 2024).

Um sistema é considerado explicável quando demonstra capacidade interpretativa — ou seja, quando justifica suas predições com base em estratégias textuais, visuais ou simbólicas que traduzem processos complexos em representações cognitivamente acessíveis (ALVES, 2021). Tal capacidade deve não apenas explicitar o que o sistema faz, mas também por que o faz, tornando o raciocínio do modelo compreensível a não especialistas. Essa perspectiva é fundamental para que a IA, no contexto educacional, se torne uma ferramenta de apoio à tomada de decisão, e não uma autoridade opaca (SIEMENS, 2022).

Segundo Philips et al. (2021), a explicabilidade em IA estrutura-se em quatro dimensões: explicação, acurácia, significado e limites. A explicação refere-se à clareza e coerência das justificativas produzidas; a acurácia assegura que as explicações reflitam o funcionamento real do modelo; o significado garante relevância cognitiva e contextual; e os limites definem as fronteiras da confiança nas explicações. Esses princípios são essenciais para a educação, pois asseguram que as explicações produzidas pelos sistemas de IA possam ser interpretadas e utilizadas por diferentes públicos — docentes, gestores, alunos e formuladores de políticas (BAKER; INVENTADO, 2019).

A explicabilidade pode assumir múltiplas formas — numérica, textual, visual ou baseada em regras — dependendo do público e do contexto de aplicação. O exemplo clássico em que um Husky foi incorretamente classificado como lobo devido ao fundo nevado (RIBEIRO et al., 2016; ROSSETTI; ANGELUCI, 2021) evidencia a importância de compreender não apenas o resultado de um modelo, mas também as razões que o produzem. Essa compreensão é vital em ambientes educacionais, em que decisões algorítmicas devem ser justificáveis perante princípios de equidade e inclusão (HOLMES; TUOMI, 2023).

Entre as abordagens explicativas mais difundidas estão as numéricas, que mensuram a contribuição de cada variável (como SHAP e LIME), as baseadas em regras, que extraem lógicas compreensíveis de decisão, as textuais, que traduzem as inferências em linguagem natural, e as visuais, que utilizam mapas e gráficos para representar o raciocínio do modelo (GUNNING et al., 2019; ZANON et al., 2024). No contexto desta pesquisa, optou-se pela análise das abordagens numéricas SHAP e LIME, devido à sua aplicabilidade em modelos preditivos educacionais, especialmente na detecção e interpretação de fatores associados ao abandono universitário (DI NOIA et al., 2022).

2.1 SHAP: Shapley Additive Explanations

O método SHAP (SHapley Additive exPlanations) baseia-se nos valores de Shapley, oriundos da teoria dos jogos cooperativos, para quantificar a contribuição de cada variável de entrada em uma predição (LUNDBERG; LEE, 2017; HART, 2024). O valor SHAP estima, em média, quanto cada atributo influencia o resultado final, considerando todas as combinações possíveis de variáveis. Essa abordagem garante uma distribuição equitativa da importância de cada atributo e proporciona uma interpretação matemática transparente do "peso" de cada variável (MOLNAR, 2022).

Apesar de seu rigor teórico, o SHAP apresenta limitações práticas, como alta demanda computacional e dependência de acesso ao conjunto original de dados. Mais importante, porém, é reconhecer que a explicabilidade matemática não equivale à interpretabilidade cognitiva: um gráfico SHAP pode ser tecnicamente preciso, mas pouco útil para professores ou gestores sem formação em IA. Em contextos educacionais, a utilidade pedagógica da explicabilidade depende da capacidade de traduzir as análises em narrativas compreensíveis, que revelem caminhos de intervenção e reflexão didática.

2.2 LIME: Local Interpretable Model-agnostic Explanations

O método LIME, proposto por Ribeiro et al. (2016), busca explicar predições de modelos complexos por meio de aproximações locais mais simples — como regressões lineares ou árvores de decisão. Ao gerar perturbações artificiais em torno de uma instância e observar as respostas do modelo original, o LIME constrói uma explicação interpretável baseada na importância relativa de cada atributo.

Entre suas vantagens, destacam-se a independência de modelo (model-agnostic), o baixo custo computacional e a não exigência de acesso direto ao dataset original. Essas características tornam o LIME adequado para aplicações educacionais com restrições de dados sensíveis, comuns em contextos institucionais. Entretanto, suas limitações — especialmente a inconsistência entre explicações locais e a sensibilidade a parâmetros internos — exigem mediação interpretativa. Em educação, a utilidade da explicação depende de como ela é traduzida para decisões pedagógicas: a interpretação de um risco de evasão, por exemplo, precisa ser compreendida à luz das dimensões socioeconômicas e motivacionais do estudante.

Assim, mais do que descrever as técnicas SHAP e LIME, este estudo as compreende como instrumentos de mediação pedagógica, capazes de transformar dados preditivos em conhecimento educacional explicável, ético e acionável.

3. Materiais e Método

A pesquisa foi desenvolvida em três etapas principais:
(i) levantamento e preparação dos dados de abandono em cursos universitários;
(ii) concepção, treinamento e avaliação de uma rede neural artificial voltada à predição de abandono universitário; e
(iii) aplicação e análise de métodos de Inteligência Artificial Explicável (XAI), com o objetivo de identificar os atributos que mais contribuem para a previsão de casos de abandono.

Todas as implementações foram realizadas pelos autores utilizando a linguagem de programação **Python**, com o ambiente **Anaconda** e as bibliotecas **Pandas** e **Scikit-learn**.

3.1 Etapa I – Levantamento e preparação dos dados

Os dados utilizados referem-se ao histórico acadêmico de estudantes de um curso de Licenciatura em Ciências Biológicas, ofertado na modalidade a distância por uma instituição pública de ensino superior no período de 2017 a 2025. O conjunto inclui informações como distância do polo de apoio, nota final, forma de ingresso e tipo de cota. Todos os dados foram anonimizados, garantindo a preservação da identidade dos indivíduos.

O dataset é composto por 4.547 instâncias relativas a 438 discentes e inclui os seguintes atributos:

- Raça/etnia (Amarela, Branca, Indígena, Não declarada, Negra, Parda);
- Distância do polo (em quilômetros e horas);
- Código da disciplina;
- Situação da matrícula na disciplina (Ativo, Cancelado, Remanescente);
- Situação do resultado na disciplina (Aprovado, Cursando, Exame, Matrícula Cancelada, Não Concluiu, Reprovado por Frequência, Reprovado por Nota, Transferido);
- Nota final da disciplina;
- Situação atual no curso (Abandono, Cancelado, Formado, Vinculado);
- Idade;

- Forma de ingresso (Processo Seletivo, Transferência Compulsória, Vestibular);
- Tipo de cota de ingresso (Cotas por ensino público, Cotas por raça, Não cotista).

Para o pré-processamento, os atributos categóricos foram codificados via *One-Hot Encoding* e os atributos numéricos normalizados para otimizar o desempenho dos algoritmos. O atributo "Situação Atual no Curso" foi utilizado como variável de saída (*target*), com quatro classes possíveis: *Abandono*, *Cancelado*, *Formado* e *Vinculado*.

3.2 Etapa II – Modelagem e treinamento da rede neural

Foram testadas diversas configurações de redes neurais com o objetivo de identificar a arquitetura mais adequada ao problema. O modelo final foi composto por:

- Uma camada de entrada correspondente ao número de atributos do dataset;
- Duas camadas ocultas com 64 e 32 neurônios, respectivamente, utilizando a função de ativação ReLU;
- Uma camada de saída com ativação Softmax, adequada à classificação multiclasse.

O modelo foi treinado com o otimizador Adam e a função de perda categórica *cross-entropy*, utilizando validação cruzada para garantir a robustez dos resultados.

3.3 Etapa III – Avaliação e explicabilidade do modelo

Nesta fase, foram aplicadas as bibliotecas SHAP (SHapley Additive exPlanations) e LIME (Local Interpretable Model-agnostic Explanations) para interpretar as predições do modelo e identificar os principais atributos determinantes do abandono universitário. As análises buscaram verificar a coerência das explicações entre os dois métodos e examinar o potencial das abordagens de XAI como instrumentos de apoio à decisão educacional.

A partir dessas análises, foram geradas visualizações que evidenciam a influência relativa de cada variável, permitindo compreender de maneira mais intuitiva os padrões de risco preditos pela rede neural.

3.4 Aspectos Éticos

O presente estudo foi conduzido em conformidade com os princípios éticos da pesquisa científica e educacional. Todos os dados utilizados foram previamente anonimizados, não sendo possível identificar individualmente os participantes ou suas instituições de origem. As informações analisadas corresponderam exclusivamente a registros acadêmicos disponibilizados pela instituição de ensino, com autorização institucional e finalidade estritamente científica.

Não houve coleta direta de dados pessoais nem qualquer forma de intervenção com seres humanos. Dessa forma, o estudo não se enquadra nas categorias que exigem apreciação por Comitê de Ética em Pesquisa (Resolução nº 510/2016, Conselho

Nacional de Saúde). Todos os procedimentos observaram as normas de proteção de dados e confidencialidade, em alinhamento com a Lei Geral de Proteção de Dados (Lei nº 13.709/2018).

Para fins exclusivos de treinamento e avaliação do modelo, os dados anonimizados foram processados em ambiente controlado; não foram armazenados de forma permanente e, ao término do processamento, foram descartados, mantendo-se apenas os artefatos técnicos derivados (pesos do modelo, métricas agregadas e explicações SHAP/LIME), que não permitem a reidentificação de indivíduos.

4. Resultados e Discussão

Os resultados obtidos indicam um bom desempenho global do modelo, com acurácia geral de 83,94%, o que demonstra sua capacidade promissora para apoiar a identificação de padrões acadêmicos em larga escala. A classe “Formado” (classe 2 - estudantes que graduaram e concluíram o curso) obteve excelente *recall* (97,6%), mostrando que o modelo é altamente eficaz em reconhecer trajetórias de sucesso acadêmico, um indicativo relevante de sua capacidade de captar sinais positivos no percurso estudantil.

A classe “Cancelado” (classe 1 - estudantes que cancelaram o curso) também apresentou desempenho sólido, com equilíbrio entre precisão e recall (ambos em torno de 86,7%), refletindo estabilidade na detecção desse grupo específico. A classe “Vinculado” (classe 3 - estudantes matriculados regularmente) alcançou bons níveis de acerto (recall de 86,0%), mesmo diante da natural sobreposição com perfis de evasão e conclusão, o que aponta para a complexidade do cenário, mas também para a competência do modelo em lidar com categorias fronteiriças.

No caso da classe “Abandono” (classe 0 - estudantes que não estão vinculados ao curso), embora o *recall* esteja mais baixo (70,4%), os resultados revelaram um ponto de atenção importante e uma oportunidade de aprimoramento. Parte dos erros se deve à semelhança deste perfil com estudantes vinculados ou em vias de formação, o que sugere a necessidade de refinamento das variáveis utilizadas para capturar melhor o comportamento associado à evasão.

Esses resultados justificam a aplicação das técnicas de **XAI (Explainable Artificial Intelligence)**, como **SHAP** e **LIME**, empregadas para identificar e compreender os fatores que mais contribuíram para as decisões do modelo. Essa etapa de análise explicável não apenas aprimora a **transparência e auditabilidade**, mas também amplia o potencial de aplicação pedagógica e institucional dos achados, fornecendo subsídios concretos para ações preventivas.

	precision	recall	f1-score	support
Abandono	88.44	70.4	78.39	402.0
Cancelado	86.76	86.76	86.76	136.0
Formado	76.57	97.57	85.8	288.0
Vinculado	85.58	86.06	85.82	538.0

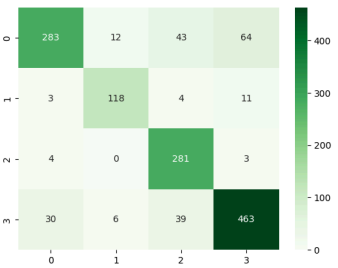

Figura 1. Métricas de avaliação e a matriz de confusão.

4.1 - Resultados do método SHAP

As visualizações baseadas em valores SHAP fornecem uma perspectiva detalhada sobre a influência de cada atributo nas predições do modelo. Cada ponto representado nos gráficos violinos (figura 2) corresponde a uma instância do conjunto de dados, e sua posição no eixo horizontal reflete a magnitude e o sinal do valor SHAP associado a determinada variável — indicando se ela aumentou ou reduziu a probabilidade da predição para uma classe específica. As cores auxiliam na leitura semântica: tons vermelhos indicam valores altos do atributo, enquanto tons azuis indicam valores baixos.

A análise revelou que os atributos “Nota Final”, “Distância do Polo” e “Forma de Ingresso” estão entre os mais relevantes na predição de abandono universitário. De maneira consistente com evidências empíricas, o modelo indicou que notas finais mais baixas e maior distância até o polo presencial estão fortemente associadas a maiores probabilidades de evasão. A forma de ingresso também emergiu como um fator discriminante, possivelmente refletindo o impacto de políticas de acesso e inclusão no desempenho e permanência acadêmica.

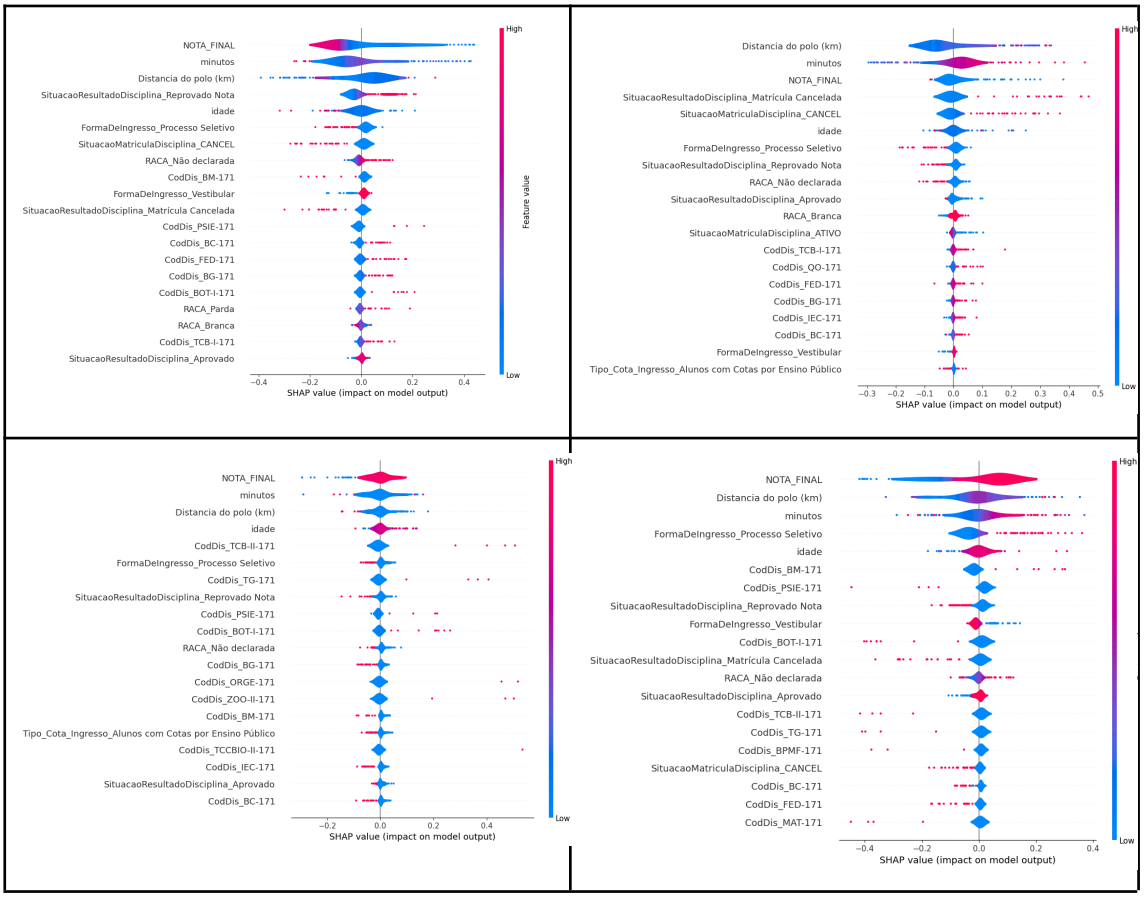

Figura 2. Gráfico violino para as classes abandono, cancelado, bem-sucedido e vinculado, respectivamente.

No caso específico da **classe abandono**, **notas baixas** e **distância elevada** são os fatores mais fortemente associados à evasão. Para **cancelamentos**, a **distância do polo** manteve-se como principal variável, enquanto o peso da **nota final** foi menor. Já entre os **estudantes formados**, observou-se o inverso: **notas altas e proximidade geográfica** reduziram a probabilidade de evasão. Por fim, entre os **alunos ainda vinculados**, surgiram padrões intermediários — notas médias, distância moderada e crescente relevância da forma de ingresso. Esses resultados reforçam a importância da **explicabilidade** como instrumento para revelar padrões sutis e orientar estratégias institucionais de acompanhamento.

4.2 Resultados do Método LIME

A aplicação do método **LIME (Local Interpretable Model-agnostic Explanations)** complementou as análises SHAP ao fornecer explicações **locais e individualizadas** para cada decisão do modelo. Em instâncias específicas — como a predição de um aluno identificado como provável caso de abandono — o LIME destacou exatamente quais atributos tiveram maior influência naquela predição (Figura 3).

Os resultados confirmaram a predominância dos atributos **nota final**, **forma de ingresso** e **distância do polo**, embora em proporções distintas entre indivíduos. Essa granularidade permite interpretações personalizadas e fornece **apoio direto à tomada de decisão pedagógica**. Por utilizar modelos lineares simples ou árvores de decisão para explicar o comportamento local da rede neural, o LIME torna as explicações **mais acessíveis para gestores e docentes**, ampliando o potencial de aplicação prática dos resultados da XAI na gestão educacional.

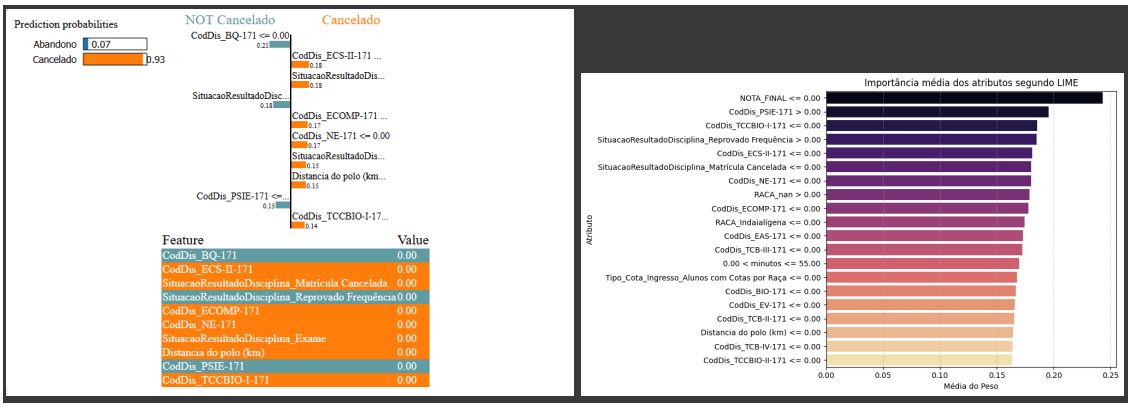

Figura 3. Resultados da aplicação do método LIME

Os resultados obtidos com o uso combinado das técnicas SHAP e LIME evidenciam não apenas a eficácia do modelo na predição do abandono universitário, mas também o potencial transformador da explicabilidade como recurso pedagógico e institucional. A análise dos fatores explicativos oferece subsídios concretos para que gestores, professores e desenvolvedores de políticas educacionais compreendam por que e como determinados padrões de evasão emergem, favorecendo uma cultura de decisão baseada em evidências. Nesse sentido, o produto educacional proposto — um guia prático para uso de técnicas de XAI na educação — traduz a complexidade algorítmica

em conhecimento acessível, promovendo a apropriação docente dessas ferramentas e estimulando o desenvolvimento de competências analíticas nos ambientes educacionais. Assim, mais do que uma validação técnica, este trabalho representa um avanço na integração entre ciência de dados e prática educativa, reforçando a importância da explicabilidade como ponte entre a inteligência artificial e a inteligência humana no contexto da aprendizagem e da permanência estudantil.

5. Considerações Finais

Os resultados deste estudo reforçam a relevância de integrar abordagens de XAI no desenvolvimento de sistemas preditivos voltados à permanência estudantil. Ao demonstrar como SHAP e LIME podem tornar modelos de *machine learning* mais transparentes e interpretáveis, a pesquisa avança não apenas no plano técnico, mas também ético e pedagógico, ao evidenciar a possibilidade concreta de uma IA educacional responsável, auditável e socialmente orientada.

A combinação das técnicas SHAP e LIME mostrou-se altamente promissora para aumentar a transparência e a confiança na utilização de modelos preditivos em contextos educacionais. O SHAP forneceu uma visão global e matematicamente consistente das variáveis mais determinantes para o abandono universitário, permitindo análises abrangentes sobre o comportamento do modelo. Complementarmente, o LIME proporcionou explicações locais e individualizadas, úteis para a compreensão de casos específicos e para o planejamento de ações pedagógicas direcionadas.

Em conjunto, essas abordagens potencializam o valor dos sistemas de IA aplicados à educação, tornando-os menos opacos (*"caixa-preta"*) e mais auditáveis, éticos e compreensíveis para docentes, gestores e formuladores de políticas públicas. Além de fortalecer o rigor técnico dos modelos de *machine learning*, a adoção de métodos de XAI contribui para orientar políticas educacionais baseadas em evidências, oferecendo subsídios concretos para estratégias preventivas contra a evasão.

Por fim, ao traduzir a complexidade algorítmica em conhecimento acessível e útil para professores, essa pesquisa reafirma o papel da XAI como ponte entre a inteligência artificial e a inteligência humana, promovendo uma educação mais justa, informada e capaz de compreender as causas que sustentam o sucesso ou o abandono escolar.

Referências

ADADI, A.; BERRADA, M. *Peeking Inside the Black Box: A Survey on Explainable Artificial Intelligence (XAI).* IEEE Access, v. 6, p. 52138-52160, 2018.

ALVES, J. *Explicabilidade e interpretabilidade em modelos de aprendizado de máquina: fundamentos e perspectivas éticas.* Revista Brasileira de Computação Aplicada, v. 13, n. 2, p. 75-92, 2021.

DI NOIA, T. et al. *Explaining Predictions in Educational Data Mining: A Review of XAI Approaches for Student Dropout.* Computers & Education: Artificial Intelligence, v. 3, p. 100067, 2022.

DORAN, D.; SCHULZ, S.; BESOLD, T. *What Does Explainable AI Really Mean?* In: *Proceedings of the 2017 International Joint Conference on Artificial Intelligence.* p. 169-175, 2017.

GUNNING, D.; AHA, D. *DARPA's Explainable Artificial Intelligence (XAI) Program.* AI Magazine, v. 40, n. 2, p. 44-58, 2019.

HART, S. Shapley Value. In: Game Theory. [s.l.]: Stanford University, 2016. Disponível em: https://cs.stanford.edu/people/eroberts/courses/soco/projects/1998-99/game-theory/shapley.html. Acesso em: 25 out. 2024.

HAYKIN, S. Neural Networks and Learning Machines. 3. ed. [s.l.]: Pearson, 2009.

LONGO, L.; LAPUSCHKIN, S.; SEIFERT, C. Explainable Artificial Intelligence. Second World Conference, XAI 2024. Valletta, Malta, July 17-19, 2024. [s.l.]: Springer, 2024.

LUNDBERG, S. M.; LEE, S.-I. *A Unified Approach to Interpreting Model Predictions.* In: *Advances in Neural Information Processing Systems (NeurIPS).* p. 4765-4774, 2017.

MOLNAR, C. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable.* 2. ed. Lulu Press, 2022.

PHILLIPS, P. J. et al. *Four Principles of Explainable Artificial Intelligence.* National Institute of Standards and Technology (NIST), U.S. Department of Commerce, 2021. DOI: 10.6028/NIST.IR.8312.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. *"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.* p. 1135-1144, 2016.

ROSSETTI, L.; ANGELUCI, A. *Challenges in Applying Explainable AI in Educational Contexts.* Journal of Educational Technology & Society, v. 24, n. 4, p. 39-53, 2021.

SETZU, M. *Explainable Artificial Intelligence: State of the Art, Challenges and Opportunities.* Expert Systems, v. 38, n. 5, p. e12743, 2021.

ZANON, D. et al. *Visual Explainability in Deep Learning: Emerging Educational Applications.* Computers in Human Behavior Reports, v. 10, p. 101282, 2024.