MODELAGEM PREDITIVA DE EVASÃO ESCOLAR NO ENSINO SUPERIOR

DOI: 10.37702/2175-957X.COBENGE.2025.6399

Autores: *JULIA ALVES FARIAS,VICTOR CARNEIRO LIMA,MATHEUS SOUZA,RENATO DA ROCHA LOPES*

Resumo: *Este trabalho tem como principal objetivo a construção de um modelo preditivo de evasão no
ensino superior, como foco na interpretabilidade e aplicabilidade dos resultados. A pesquisa realiza uma análise exploratória de dados acadêmicos e socioeconômicos de estudantes da Universidade Estadual de Campinas (UNICAMP), integrando diferentes fontes institucionais. São aplicadas técnicas de pré-processamento, normalização e balanceamento de classes, além da criação de indicadores como coeficiente de progressão e rendimento. Utilizou-se a biblioteca PyCaret para o treinamento de diversos algoritmos de aprendizado de máquina, comparando seus desempenhos com base na acurácia balanceada, recall e AUC. O modelo LightGBM destacou-se por sua robustez em diferentes cenários, apontando variáveis relevantes na predição da evasão. Os resultados obtidos contribuem para a identificação de estudantes em risco, oferecendo subsídios para ações preventivas
e estratégicas na gestão educacional.*

Palavras-chave: *aprendizado de máquina, predição, ciência de dados, acurácia balanceada,evasão,inteligência artificial,predição*

MODELAGEM PREDITIVA DE EVASÃO ESCOLAR NO ENSINO SUPERIOR

1 INTRODUÇÃO

A evasão escolar no ensino superior representa um desafio significativo para gestores educacionais, impactando diretamente a eficiência das políticas institucionais, a alocação de recursos e a formação acadêmica dos estudantes. No contexto brasileiro, o problema é agravado por fatores socioeconômicos, dificuldades acadêmicas e falta de políticas eficazes de retenção estudantil. Diversos estudos têm apontado que a predição da evasão pode auxiliar na implementação de medidas preventivas e no direcionamento de recursos para estudantes em situação de vulnerabilidade acadêmica (AHMAD, 2023).

Nesse cenário, a aplicação de técnicas de ciência de dados e inteligência artificial tem se destacado como uma abordagem promissora para a modelagem da evasão escolar (OECD, 2021). Modelos preditivos baseados em aprendizado de máquina possibilitam a análise de grandes volumes de dados acadêmicos, permitindo identificar padrões que indiquem estudantes com maior risco de abandono.

Este artigo tem então como objetivo construir um modelo preditivo para evasão no ensino superior, com ênfase na interpretabilidade dos resultados. Esse aspecto é essencial, uma vez que decisões automatizadas na educação devem ser auditáveis e transparentes, garantindo que não perpetuem vieses discriminatórios nem resultem em decisões arbitrárias. Além disso, a pesquisa busca identificar os fatores mais influentes na decisão de um estudante de abandonar o curso e explorar como a ciência de dados pode fornecer insights valiosos para que gestores educacionais intervenham de forma proativa.

Este trabalho está dividido da seguinte maneira. Na seção 2 apresentamos a metodologia adotada, com a descrição do conjunto de dados, análise do desbalanceamento entre as classes e os classificadores escolhidos. Na seção 3, detalhamos o pré-processamento, incluindo a integração de dados externos, criação de indicadores, definição da variável alvo e geração do conjunto final utilizado para análise. Na seção 4 descreve a etapa de classificação, abordando o uso da biblioteca PyCaret, o treinamento, as métricas e a comparação entre os desempenhos dos modelos. Na seção 5, são apresentados os resultados obtidos, com base nas métricas e visualizações de desempenho.

2 METODOLOGIA

2.1 Descrição do Dataset

Os dados utilizados neste trabalho foram fornecidos por duas fontes da Universidade Estadual de Campinas (UNICAMP), a Diretoria Acadêmica (DAC) e a Comissão Permanente para os Vestibulares da Unicamp (COMVEST). Cada uma dessas entidades contribuiu com diferentes tipos de informações relevantes para a análise da evasão estudantil.

A DAC forneceu informações de caráter acadêmico, como o histórico escolar dos estudantes, dados sobre os cursos e disciplinas, desempenho ao longo dos semestres (notas, aprovação ou reprovação), além de detalhes sobre o ingresso e a saída do curso, incluindo o motivo da evasão, quando aplicável. Já a COMVEST disponibilizou dados de natureza socioeconômica e demográfica, coletados no momento da inscrição no vestibular. Entre os campos fornecidos estão: renda familiar mensal, renda per capita, tipo e modalidade de ensino médio cursado, participação na renda familiar, atividade remunerada, cor/raça e sexo dos estudantes. Todas as informações fornecidas pelos órgãos da Universidade foram previamente anonimizadas.

A integração de atributos acadêmicos e socioeconômicos permite uma compreensão mais abrangente da trajetória estudantil, ao contemplar tanto o desempenho acadêmico quanto o contexto pessoal de cada aluno. Essa abordagem multidimensional é essencial para revelar padrões que podem estar diretamente relacionados à decisão de evadir.

2.2 Desbalanceamento de dados

No conjunto de dados analisado, a classe correspondente aos alunos que evadiram representa uma fração significativamente menor em comparação aos ativos ou concluintes, o que caracteriza um desbalanceamento na distribuição das classes. Esse desequilíbrio pode comprometer o desempenho dos modelos preditivos, levando-os a favorecer a predição da classe majoritária, o que resulta em métricas ilusoriamente elevadas, porém sem capacidade real de identificar corretamente os casos de evasão (KRAWCZYK, 2016).

Para mitigar os efeitos do desbalanceamento, foram utilizadas estratégias de balanceamento artificial das classes. Neste trabalho, adotou-se a técnica de oversampling, aplicada automaticamente pelo PyCaret, que utiliza a técnica SMOTE (*Synthetic Minority Over-sampling Technique*) (CHAWLA, 2002) no conjunto de treino durante a validação,

garantindo que o balanceamento não afete o conjunto de teste. Essa abordagem permite que os modelos tenham mais exemplos da classe minoritária para aprender, aumentando sua capacidade de identificar casos de evasão e promovendo uma avaliação mais fiel ao desempenho real dos algoritmos. Além disso, incorporamos como índice de mérito a acurácia balanceada e função objetivo ponderada, para penalizar de maneira mais agressiva classificações do tipo falso negativo.

2.3 Classificadores utilizados

Para a construção do modelo preditivo de evasão no ensino superior, empregamos diversos algoritmos de aprendizado supervisionado. A seguir, são descritos os classificadores utilizados:

- **Árvore de Decisão (Decision Tree)**: É uma estrutura hierárquica que realiza classificações com base em perguntas binárias sucessivas sobre os atributos dos dados (BREIMAN, 1984). Cada nó representa uma decisão com base em uma variável, e os ramos levam a outras decisões ou a um resultado final. As árvores de decisão são intuitivas, facilmente interpretáveis e fornecem uma visão clara de como as decisões são tomadas. No entanto, tendem a sofrer com o *overfitting*[1], especialmente em dados complexos ou com ruído.
- **Random Forest**: Conjunto de múltiplas árvores de decisão construídas a partir de diferentes subconjuntos dos dados e de atributos aleatórios (HO, 1995). A predição final é obtida por meio de votação entre as árvores. Essa técnica de *ensemble*[2] melhora a generalização do modelo, reduzindo a variância e mitigando o risco de overfitting que afeta árvores individuais.
- **Extra Trees (Extremely Randomized Trees)**: Variante do Random Forest que introduz ainda mais aleatoriedade no processo de construção das árvores. Ao invés de buscar o melhor ponto de divisão, os pontos são selecionados aleatoriamente, o que tende a reduzir a variância do modelo (GEURTS, 2006). Essa abordagem é particularmente útil quando se deseja alta velocidade de treinamento com boa capacidade de generalização.

[1] *Overfitting* ocorre quando um modelo apresenta desempenho elevado nos dados de treinamento, mas falha em generalizar para dados não vistos, indicando alta variância e baixa capacidade preditiva (GÉRON, 2019).
[2] *Ensemble* é uma abordagem que combina múltiplos modelos preditivos com o objetivo de reduzir erros e melhorar a robustez e acurácia da predição (GÉRON, 2019).

- **LightGBM (Light Gradient Boosting Machine)**: Baseado em *gradient boosting*, técnica de aprendizado de máquina que constrói modelos preditivos combinando várias árvores de decisão simples, treinadas de forma sequencial. Cada nova árvore tenta corrigir os erros das anteriores, resultando em um modelo mais preciso (KE, 2017). Utiliza estratégias como crescimento de árvore orientado por folhas, discretização dos dados por histogramas e suporte nativo a variáveis categóricas.
- **Naive Bayes**: Modelo probabilístico baseado no Teorema de Bayes, que assume independência condicional entre as variáveis preditoras. Embora essa suposição raramente se confirme na prática, o Naive Bayes costuma apresentar bons resultados em diversas aplicações, especialmente quando o número de atributos é grande o suficiente (SCIKIT-LEARN, 2025).
- **Análise Discriminante Quadrática (QDA)**: Técnica estatística que modela a distribuição de probabilidade de cada parâmetro assumindo que elas seguem distribuições normais, mas com diferentes matrizes de covariância (WU, 2022). É apropriado para problemas onde as classes são bem separadas e têm variâncias distintas.
- **Máquinas de Vetores de Suporte (SVM)**: Algoritmo que busca encontrar o hiperplano ótimo que melhor separa as classes no espaço de características. Utiliza margens máximas para garantir uma separação robusta entre as classes e pode ser estendido a classificações não lineares também (CORTES, 1995). É eficaz em espaços de alta dimensionalidade.
- **Regressão Logística**: Modelo estatístico que estima a probabilidade de uma instância pertencer a uma classe com base em uma combinação linear dos atributos. Utiliza a função logística para transformar essa combinação em uma probabilidade entre 0 e 1 (CRAMER, 2002). É bastante eficaz, interpretável e serve como um excelente baseline para tarefas de classificação binária.

3 PRÉ-PROCESSAMENTO DE DADOS

3.1 Operações realizadas

O pré-processamento dos dados envolveu uma série de etapas voltadas à consolidação, enriquecimento e transformação das informações brutas extraídas dos

históricos acadêmicos e bases administrativas. O objetivo dessas operações foi gerar uma base analítica capaz de sustentar as investigações desejadas.

I. Incorporação de dados socioeconômicos

Inicialmente, os dados acadêmicos dos estudantes foram enriquecidos com informações socioeconômicas provenientes da base de ingressantes da universidade. As variáveis incorporadas incluíram renda familiar, escolaridade dos pais, tipo de escola cursada no ensino médio, cor/raça autodeclarada, entre outras. Esse processo exigiu o tratamento de dados ausentes, a padronização de categorias e a verificação de consistência entre os registros.

II. Cálculo de Evolução esperada dos currículos

Com base na estrutura curricular dos cursos, foi realizada uma simulação da trajetória ideal de integralização das disciplinas obrigatórias. Essa simulação consistiu em distribuir, semestre a semestre, os componentes curriculares obrigatórios de acordo com o período recomendado no projeto pedagógico.

Para cada semestre ideal, foram acumulados os créditos obrigatórios esperados até aquele ponto, permitindo o cálculo da progressão curricular esperada como a razão entre os créditos acumulados e o total de créditos obrigatórios do curso. Esse valor ideal de progressão foi utilizado posteriormente como parâmetro comparativo para a evolução real de cada estudante ao longo de sua trajetória acadêmica.

III. Cálculo de Coeficientes de Desempenho dos estudantes

Com os históricos escolares consolidados, foi possível então calcular duas métricas principais associadas ao desempenho acadêmico:

- **Coeficiente de Progressão (CP)**: Indicador da proporção de créditos integralizados em relação à carga horária total exigida para a conclusão do curso. O cálculo levou em conta os créditos aprovados em disciplinas obrigatórias, eletivas e livres, de acordo com os critérios definidos pelo currículo vigente no ano de ingresso do estudante. Em casos de créditos excedentes em uma determinada categoria (por exemplo, eletivas), o excedente foi redistribuído respeitando a hierarquia de

aproveitamento estabelecida institucionalmente, considerando limites mínimos e máximos.

- **Coeficiente de Rendimento (CR)**: Média ponderada das notas obtidas, considerando apenas disciplinas válidas para o cálculo conforme definido pelo regulamento do curso. Foram desconsideradas disciplinas em que o aluno obteve trancamento, cancelamento ou situações equivalentes. A média foi ponderada pela carga horária de cada disciplina.
- **Diferencial de Coeficiente de Progressão**: Medida da diferença entre o coeficiente de progressão real do estudante, e o coeficiente esperado para o perfil, definido por curso e semestre. Calculado através da subtração dos coeficientes, indica se o aluno está adiantado ou atrasado em sua trajetória acadêmica.

IV. Definição da variável alvo

Foram adotadas janelas de predição de dois e quatro semestres para rotular a evasão, com o objetivo de evitar que o aluno seja classificado como evadido em todos os períodos anteriores à sua saída, já que poderia distorcer o aprendizado do modelo, ao mesmo tempo em que permite a detecção com tempo hábil para uma intervenção pedagógica. Essa abordagem permite concentrar a detecção em um momento mais próximo da evasão, favorecendo a identificação de padrões comportamentais recentes que antecedem a decisão do abandono. Assim, a modelagem foca na diferenciação entre os estudantes que estão prestes a evadir e aqueles que seguem ativos, mesmo que todos apresentem histórico semelhante nos primeiros semestres.

V. Geração da base final em CSV para análise

Após o enriquecimento e cálculo dos indicadores, os dados foram reorganizados em um formato tabular (CSV), no qual cada linha representa o desempenho de um estudante em um determinado semestre. As variáveis da base final incluíram:

- Identificação do estudante, por índice (anonimizada)
- Curso, currículo e ano de ingresso
- Informações socioeconômicas
- Semestre em análise (com codificação sequencial para facilitar modelagens temporais)
- Créditos cursados e aprovados por categoria (obrigatória, eletiva, livre)

- CP, CR e Diferencial de CP naquele ponto da trajetória
- Tempo de curso até o semestre em questão
- Número de disciplinas trancadas ou reprovadas
- Indicação de evasão, variável alvo **binária**

Essa base final serviu de insumo para as análises estatísticas, modelos preditivos e visualizações que compõem as seções seguintes deste trabalho.

4 CLASSIFICAÇÃO

Na etapa de classificação, foram utilizados modelos supervisionados para prever a evasão de estudantes com base em seu histórico e nas variáveis derivadas do processo de pré-processamento. Os experimentos de modelagem foram conduzidos utilizando a biblioteca PyCaret, uma ferramenta de machine learning de código aberto que facilita a experimentação com diversos algoritmos e técnicas de pré-processamento (PYCARET, 2025). Para avaliar o impacto das diferentes configurações de inicialização do modelo, fizemos um conjunto de experimentos variando alguns parâmetros como normalização e estratégias de redução de desbalanceamento. Especialmente procurando alternativas para tratar do desbalanceamento de dados. A seguir, são detalhadas as etapas realizadas:

I. Critérios de desempenho

Como critério principal para definição do melhor modelo, adotou-se a acurácia balanceada. Essa métrica corresponde à média entre a taxa de verdadeiros positivos (sensibilidade) e verdadeiros negativos (especificidade), oferecendo uma avaliação mais justa em contextos de desbalanceamento de classes (CARRILLO, 2014). No caso da evasão, onde a maioria dos registros pertence à classe “não evadiu”, o uso da acurácia balanceada evita que o modelo tenha seu desempenho inflado por simplesmente priorizar a classe majoritária, possibilitando uma análise mais fiel da capacidade de identificar corretamente os casos de evasão.

Além da acurácia balanceada, também foi considerada a métrica de recall, que corresponde à proporção de instâncias positivas corretamente classificadas (GÉRON, 2019). Em problemas como a previsão de evasão, o recall é relevante, pois busca minimizar falsos negativos - ou seja, casos de estudantes evadidos que seriam incorretamente classificados como não evadidos.

II. Treinamento e comparação de modelos

Após a configuração, avaliamos diversos algoritmos de classificação. A ferramenta utilizada retorna os modelos com melhor desempenho com base em métricas como AUC, acurácia balanceada e recall. Como o objetivo principal era prever a evasão (especialmente reduzir falsos negativos, ou seja, evadidos classificados como não evadidos), métricas como recall e acurácia balanceada foram priorizadas na escolha dos modelos que melhor performaram.

5 RESULTADOS OBTIDOS

Para avaliar a eficácia dos modelos preditivos, foram realizados experimentos em cinco diferentes cenários, variando estratégias de normalização e balanceamento dos dados. Esta seção apresenta os resultados comparativos entre os modelos treinados em cada configuração, bem como uma análise mais detalhada do modelo com melhor desempenho.

Tabela 1 - Com normalização, sem balanceamento, janela de evasão de 2 semestres

Modelo	AUC	Recall	Acurácia balanceada
Decision Tree	0.7932	0.7199	0.7839
Extra Trees	0.7775	0.6129	0.6938
LightGBM	**0.9161**	0.7947	**0.8366**
Logistic Regression	0.8593	0.7478	0.7772
Naive Bayes	0.7334	0.4296	0.6381
QDA	0.7893	0.3666	0.6398
Random Forest	0.8144	0.7419	0.7451
SVM	0.7350	**0.8006**	0.7350

Tabela 2 - Com normalização, com balanceamento, janela de evasão de 2 semestres

Modelo	AUC	Recall	Acurácia balanceada
Decision Tree	0.7941	0.4604	0.6965
Extra Trees	0.7809	0.6745	0.7046
LightGBM	0.8133	0.7478	0.7432
Logistic Regression	0.8553	0.7346	0.7737
Naive Bayes	0.6835	**0.8109**	0.5757
QDA	0.7552	0.6452	0.7076
Random Forest	**0.8632**	0.6012	0.7554
SVM	0.7828	0.7522	**0.7828**

Tabela 3 - Sem normalização, sem balanceamento, janela de evasão de 2 semestres

Modelo	AUC	Recall	Acurácia balanceada
Decision Tree	0.7941	0.7214	0.7846
Extra Trees	0.7775	0.6129	0.6938
LightGBM	**0.9166**	**0.7991**	**0.8394**
Logistic Regression	0.8605	0.7449	0.7753
Naive Bayes	0.7333	0.5894	0.7403
QDA	0.7891	0.6833	0.7403
Random Forest	0.8144	0.7419	0.7451
SVM	0.6053	0.2302	0.6053

Tabela 4 - Sem normalização, com balanceamento, janela de evasão de 2 semestres

Modelo	AUC	Recall	Acurácia balanceada
Decision Tree	0.7941	0.4604	0.6965
Extra Trees	0.7809	0.6745	0.7046
LightGBM	0.8131	0.7522	0.7452
Logistic Regression	0.8569	0.7375	0.7747
Naive Bayes	0.7757	0.7038	0.7312
QDA	0.7974	**0.7830**	0.7346
Random Forest	**0.8628**	0.5924	0.7509
SVM	0.7775	0.7419	**0.7775**

Tabela 5 - Com normalização, sem balanceamento, janela de evasão de 4 semestres

Modelo	AUC	Recall	Acurácia balanceada
Decision Tree	0.7887	0.7263	0.7787
Extra Trees	0.7943	0.6540	0.7161
LightGBM	**0.9266**	0.8201	**0.8443**
Logistic Regression	0.8670	0.7654	0.7847
Naive Bayes	0.7424	0.5122	0.6624
QDA	0.7912	0.5122	0.6407
Random Forest	0.8172	0.7986	0.7623
SVM	0.7574	**0.8211**	0.7574

A análise dos resultados evidencia que o desempenho dos modelos varia de acordo com estratégias de normalização, balanceamento e janela de observação adotada. No cenário 1, ilustrado pela Tabela 1, o LightGBM obteve os melhores valores e AUC e acurácia balanceada, enquanto o SVM apresentou o maior recall. Isso sugere que, apesar do bom desempenho geral do LightGBM, o SVM pode ter sido mais sensível à identificação de casos positivos, possivelmente por sua margem de decisão ampla em dados normalizados. No cenário 2, o SVM teve a maior acurácia balanceada, mas Naive Bayes alcançou o maior recall e Random Forest a maior AUC. Esse comportamento pode estar relacionado à forma

como cada algoritmo lida com amostras balanceadas: Naive Bayes tende a aumentar o recall em detrimento da precisão, enquanto Random Forest, sendo mais robusto, se beneficia do balanceamento para melhorar sua capacidade discriminativa.

No cenário 3, o LightGBM dominou todas as métricas, o que reforça sua flexibilidade e robustez, já que seu desempenho não depende fortemente de pré-processamentos como normalização. Já no cenário 4, observamos outra divisão: SVM com a maior acurácia balanceada, QDA com o maior recall e Random Forest com a maior AUC. Ao olhar para as particularidades de cada modelo: QDA pode ter sido favorecido pelo balanceamento ao identificar melhor os casos positivos; SVM mantém boa acurácia balanceada pela forma como define seus limites de decisão; e Random Forest, que segue performando uma boa capacidade de generalização. Por fim, no cenário 5, o LightGBM novamente se destacou em AUC e acurácia balanceada.

Figura 1 – Curva ROC correspondente ao modelo *LightGBM*, no cenário com normalização, sem balanceamento, janela de evasão de 4 semestres

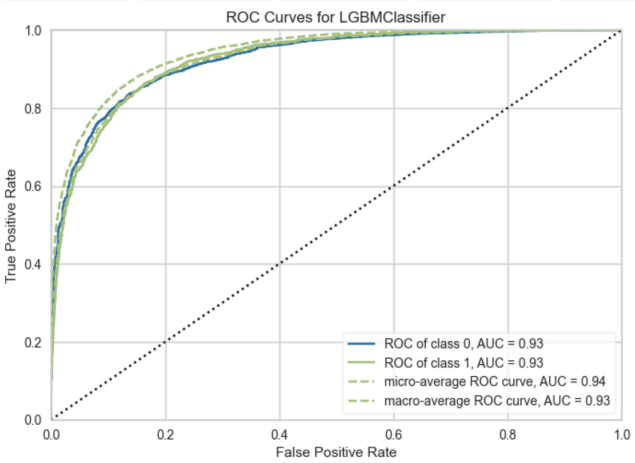
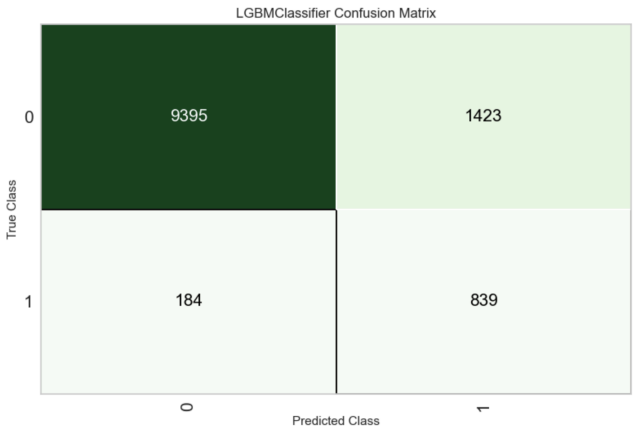


Figura 2 – Matriz de confusão correspondente ao modelo *LightGBM*, no cenário com normalização, sem balanceamento, janela de evasão de 4 semestres



Com base nas figuras 1 e 2, observa-se que o modelo LightGBM, no cenário com normalização, sem balanceamento e janela de evasão de 4 semestres, apresentou uma curva ROC com AUC de 0,93 para ambas as classes, o que confirma sua alta capacidade discriminativa já destacada na análise das métricas tabeladas. A matriz de confusão reforça esse bom desempenho ao mostrar baixa taxa de falsos negativos (184) e um número expressivo de acertos na classe positiva (839), mesmo diante do desbalanceamento dos

dados. A quantidade relativamente maior de falsos positivos (1423) indica uma leve tendência a superestimar a evasão, o que, embora reduza a precisão, pode ser aceitável em contextos onde é preferível errar por excesso. Essas visualizações corroboram a robustez do LightGBM, alinhando-se às conclusões anteriores quanto à sua eficiência em diversos cenários.

6 CONCLUSÕES

De modo geral, o LightGBM apresentou desempenho consistente e superior em quase todos os cenários, destacando-se especialmente nas métricas selecionadas como mais relevantes, como a acurácia balanceada, independente da aplicação de técnicas de normalização ou balanceamento. Essa robustez pode ser atribuída à sua capacidade de modelar relações não lineares complexas, lidar eficientemente com dados desbalanceados e adaptar-se automaticamente a diferentes distribuições e pesos de variáveis. Além de identificar um modelo preditivo com boa performance para o problema de evasão estudantil, este trabalho também desenvolveu e integrou métricas derivadas dos dados educacionais e socioeconômicos, que se mostraram relevantes no processo de predição. A eficácia dessas variáveis pode ser observada nos altos valores de desempenho alcançados, indicando que contribuíram de forma significativa para a capacidade preditiva dos modelos. Assim, podemos concluir que foi possível não apenas selecionar um preditor robusto, mas também construir um conjunto de atributos informativos que ampliam o potencial de aplicação prática deste estudo. Os resultados obtidos contribuem tanto para a literatura acadêmica quanto para a gestão educacional, oferecendo subsídios para ações preventivas mais precisas e fundamentadas no apoio à permanência estudantil.

REFERÊNCIAS

AHMAD, K.; IQBAL, W.; EL-HASSAN, A. et al. Data-driven artificial intelligence in education: A comprehensive review. *IEEE Transactions on Learning Technologies*, 2023. Disponível em: https://ieeexplore.ieee.org/document/10247566

BREIMAN, L.; FRIEDMAN, J.; OLSHEN, R. A.; STONE, C. J. Classification and Regression Trees. 1. ed. *Chapman and Hall/CRC*, 1984. Disponível em: https://doi.org/10.1201/9781315139470

CARRILLO, H.; BRODERSEN, K. H.; CASTELLANOS, J. A. Probabilistic performance evaluation for multiclass classification using the posterior balanced accuracy. In: ARMADA, M.; SANFELIU, A.; FERRE, M. (ed.). *ROBOT2013: First Iberian Robotics Conference.* Cham:

Springer, 2014. (Advances in Intelligent Systems and Computing, v. 252). Disponível em: https://doi.org/10.1007/978-3-319-03413-3_25

CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O.; KEGELMEYER, W. P. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 2002. Disponível em: https://www.jair.org/index.php/jair/article/view/10302

CORTES, C.; VAPNIK, V. Support-vector networks. *Mach Learn, v.* 20, p. 273–297, 1995. Disponível em: https://doi.org/10.1007/BF00994018

CRAMER, J. S. The Origins of Logistic Regression. *Tinbergen Institute Working Paper*, n. 2002-119/4, dez, 2002. Disponível em: https://ssrn.com/abstract=360300

GÉRON, A. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: concepts, tools, and techniques to build intelligent systems*. 2. ed. Sebastopol: O'Reilly Media, 2019.

GEURTS, P.; ERNST, D.; WEHENKEL, L. Extremely randomized trees. *Mach Learn*, v. 63, p. 3–42, 2006. Disponível em: https://doi.org/10.1007/s10994-006-6226-1

HO, T. K. Random Decision Forests. *Murray Hill, NJ: AT&T Bell Laboratories*, 1995. Disponível em: https://web.archive.org/web/20160417030218/http://ect.bell-labs.com/who/tkh/publications/papers/odt.pdf#expand

KE, G. MENG, Q., FINLEY T. et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems*, v. 30, p. 3149-3157, 2017. Disponível em: https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf

KRAWCZYK, B. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence,* v. 5, p. 221–232, 2016. Disponível em: https://doi.org/10.1007/s13748-016-0094-0

ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT (OECD). *Digital Education Outlook*. Paris: OECD Publishing, 2021, isbn: 978-92-64-76216-3. Disponível em: https://doi.org/10.1787/589b283f-en

PYCARET. PyCaret: An open-source, low-code machine learning library in Python. Disponível em: https://pycaret.gitbook.io/docs

SCIKIT-LEARN. Naive Bayes. Disponível em: https://scikit-learn.org/stable/modules/naive_bayes.html

WU, R.; HAO, N. Quadratic discriminant analysis by projection. *Journal of Multivariate Analysis*, v. 190, art. 104987, 2022. ISSN 0047-259X. Disponível em: https://doi.org/10.1016/j.jmva.2022.104987.