ENCITEC
Ensino de Ciências e Tecnologia em Revista

USO APLICADO DE MACHINE LEARNING COM RANDOM FOREST E SELF-ORGANIZING MAP COMO MODELO PREDITIVO PARA EVASÃO DE ALUNOS

APPLIED USE OF MACHINE LEARNING WITH RANDOM FOREST AND SELF-ORGANIZING MAP AS A PREDICTIVE MODEL FOR STUDENT DROPOUT

Anderson Costa Lima[1], Orivaldo Vieira de Santana Junior[2], Efrain Pantaleón Matamoros[3]

Recebido: Julho/2025 - Aprovado: Setembro/2025

Resumo: Este artigo é um recorte de uma dissertação de mestrado desenvolvida no Programa de Pós-Graduação em Ciência, Tecnologia e Inovação da Universidade Federal do Rio Grande do Norte (UFRN). Com o objetivo de investigar fatores relacionados à evasão no ensino superior por meio de técnicas de Mineração de Dados Educacionais (MDE) e Machine Learning (ML). O estudo busca explorar os dados do curso Interdisciplinar em Ciências e Tecnologia (C&T) da própria universidade, considerando os alunos ingressantes entre os anos de 2014 a 2023, com o objetivo de desenvolver modelos analíticos que identifiquem características de intervenção para o desenvolvimento acadêmico dos estudantes. O estudo combina algoritmos de ML, como Random Forest (classificação) e Self-Organizing Maps (clustering), em uma abordagem híbrida, aplicada à MDE. O resultado é um modelo preditivo capaz de identificar atributos que influenciam a evasão, utilizando a explicabilidade da técnica SHapley Additive exPlanations para encontrar um conjunto de características explicáveis e ao mesmo tempo, fornecer uma melhoria significativa no poder preditivo do modelo.

Palavras-chave: Evasão escolar, Mineração de Dados Educacionais, Random Forest, Self-Organizing Maps, Machine Learning, SHapley Additive exPlanations

Abstract: This article is an excerpt from a master's dissertation developed within the Graduate Program in Science, Technology and Innovation at the Federal University of Rio Grande do Norte (UFRN). Its objective is to investigate factors related to dropout in higher education through Educational Data Mining (EDM) and Machine Learning (ML) techniques. The study analyzes data from the Interdisciplinary Science and Technology (C&T) program at the same university, considering students enrolled between 2014 and 2023, with the goal of developing analytical models that identify intervention characteristics to support students' academic development. The study

1 http://orcid.org/0009-0007-8697-391X, Mestre em Ciência Tecnologia e Inovação. Universidade Federal do Rio Grande do Norte (UFRN), Campus Universitário - Lagoa Nova, Natal - RN, 59078-970, Brasil. andersonlimac@gmail.com
2 http://orcid.org/0000-0003-4918-3162, Doutor em Ciências Da Computação. Universidade Federal do Rio Grande do Norte (UFRN), Campus Universitário - Lagoa Nova, Natal - RN, 59078-970, Brasil. orivaldo@gmail.com
3 http://orcid.org/0000-0003-2060-8024, Doutor em Engenharia Mecânica. Universidade Federal do Rio Grande do Norte (UFRN), Campus Universitário - Lagoa Nova, Natal - RN, 59078-970, Brasil. efrain.pantaleon@ufrn.br

ENCITEC - Ensino de Ciências e Tecnologia em Revista | Santo Ângelo | v. 16 | n. 1 | p. 7-29 | jan./abr. 2026
DOI: http://dx.doi.org/10.31512/encitec.v16i1.2260

combines ML algorithms, such as Random Forest (classification) and Self-Organizing Maps (clustering), in a hybrid approach applied to EDM. The result is a predictive model capable of identifying attributes that impact dropout, using the explainability technique SHapley Additive exPlanations (SHAP) to identify a set of interpretable features while simultaneously providing a significant improvement in the model's predictive power.

Keywords: Student dropout, Educational Data Mining, Random Forest, Self-Organizing Maps, Machine Learning, SHapley Additive exPlanations

1 Introdução

A evasão escolar é um problema global que afeta instituições de ensino em diferentes níveis, comprometendo o desenvolvimento de indivíduos e sociedade. No Brasil, as taxas de evasão no ensino superior atingem níveis preocupantes, com mais da metade dos alunos desistindo dos cursos antes da conclusão, conforme relatório apresentado pelo Sindicato das Entidades Mantenedoras de Estabelecimentos de Ensino Superior no Estado de São Paulo (SEMESP, 2023).

Essa realidade acarreta prejuízos econômicos e sociais significativos, além de limitar o potencial de desenvolvimento do país. Nesse contexto, a identificação precoce de estudantes em risco de evasão torna-se fundamental para a implementação de estratégias de intervenção eficazes.

Nesse cenário, a Mineração de Dados Educacionais (MDE) surge como uma ferramenta promissora para a análise de dados acadêmicos e a identificação de padrões relacionados à evasão (BAKER et al., 2011). A MDE busca aprofundar a compreensão dos fatores que influenciam o desempenho dos alunos e, a partir desse conhecimento, apoiar o desenvolvimento de estratégias voltadas à prevenção da evasão.

No entanto, apesar do crescente interesse pela MDE, pesquisas nacionais sobre o tema, especialmente relacionadas à modalidade presencial, ainda não eram amplamente difundidas, principalmente antes da pandemia de Covid-19 (MASCHIO et al., 2018). Além disso, a área enfrenta desafios significativos, como a escassez de conjuntos de dados consistentes e a necessidade de desenvolver métodos mais robustos que integrem diferentes técnicas de mineração de dados (SUKHIJA et al., 2016).

Diante desse contexto, o presente artigo propõe o desenvolvimento de um método híbrido, combinando os algoritmos *Random Forest* (classificação) e *Self-Organizing Maps* (clustering), para analisar os dados de alunos de um curso de graduação presencial da Universidade Federal do Rio Grande do Norte (UFRN) e identificar as principais características associadas à evasão.

A MDE neste contexto busca ampliar a compreensão sobre o fenômeno da evasão em instituições públicas de ensino superior no Brasil, fornecendo subsídios para a criação de ferramentas inteligentes capazes de identificar precocemente alunos em risco de evasão. Com base nisso, os objetivos específicos deste estudo são: (1) desenvolver um método híbrido de MDE, combinando algoritmos de *clustering* e classificação, para analisar dados de evasão escolar em um curso de graduação presencial na UFRN; (2) identificar as principais características relacionadas à evasão dos alunos, explorando a relação entre

os atributos dos estudantes e suas trajetórias acadêmicas; (3) avaliar o desempenho do modelo e sua capacidade de generalização para novos dados; e (4) contribuir para o desenvolvimento de ferramentas inteligentes que auxiliem na detecção precoce de alunos em risco de evasão, possibilitando a implementação de medidas proativas e direcionadas.

2 Revisão da literatura

A evasão no ensino superior tem sido uma preocupação constante das instituições de ensino em todos os níveis. Com a pandemia de Covid-19, notou-se um aumento significativo dessas taxas, sobretudo na rede pública, que se mostrou mais vulnerável aos impactos da suspensão das aulas presenciais (HONORATO; BORGES, 2022, apud KANTORSKI, 2023).

Dados da SEMESP (2023), com base em informações do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), evidenciam a gravidade do cenário da evasão no ensino superior. Entre os estudantes que ingressaram no período de 2017 a 2021, 55,5% abandonaram seus cursos, revelando que mais da metade não concluiu a graduação. O relatório mostra ainda que apenas 26,3% dos estudantes finalizaram a formação dentro do prazo previsto, enquanto 18,1% permaneciam matriculados. Os impactos da evasão são expressivos e se refletem não apenas no âmbito educacional, mas também na economia e no equilíbrio social.

Conforme destacado por Pfeiffer (1999 apud PRESTES; FIALHO, 2018), o fenômeno da evasão está diretamente associado a indicadores de desenvolvimento humano, como pobreza, vulnerabilidade social, desemprego, questões de saúde, expectativa de vida e participação política. Para Oliveira et al. (2019), trata-se de um epicentro de grandes problemas, incluindo a desaceleração do crescimento tecnológico nacional, a persistência de questões sociais e o desperdício econômico.

É inegável que inúmeros fatores contribuem para esse cenário, variando conforme a região do país, o tipo de instituição (pública ou privada), a estrutura acadêmica e, ainda, a modalidade de ensino, seja presencial ou a distância (BARDAGI; HUTZ, 2012).

Para Mesquita et al. (2021), essa compreensão pode trazer luz a diversas respostas que permitam identificar fatores que influenciam o sucesso ou insucesso do estudante na graduação. Um dos principais grandes desafios, segundo os autores, é a aplicação de métodos e técnicas de inteligência computacional, capazes de extrair, de maneira eficiente e precisa, características relevantes a partir das grandes bases de dados disponíveis em ambientes educacionais.

A extração de conhecimento dessas grandes bases de dados é conhecido como KDD (*Knowledge Discovery in Database*s). O KDD é um processo destinado à identificação de padrões, desconhecidos, potencialmente úteis e compreensíveis em grandes repositórios de dados (Fayyad, Piatetsky, & Smyth, 1996). Esse processo envolve uma sequência de etapas, que incluem a coleta, pré-processamento, transformação, mineração e interpretação de dados, com o objetivo de descobrir *insights* relevantes e úteis a partir dos dados.

Em ambientes educacionais, Ramos et al. (2020), conforme citado por Costa et al. (2022) sugerem que se faça adaptações no processo do KDD, iniciando com a coleta dos dados, seguida da realização do pré-processamento, mapeamento das características que estão fortemente relacionadas evasão escolar. Para garantir a qualidade do modelo preditivo, o processo termina com análises da explicabilidade, com o objetivo de encontrar um conjunto de características explicáveis e, ao mesmo tempo, alcançar uma melhoria significativa no poder preditivo dos modelos utilizados.

Nesse contexto, a MDE surge como uma ferramenta promissora para auxiliar na compreensão e na mitigação da evasão escolar. Inserida no processo de KDD, a MDE adapta suas etapas, concentrando-se no desenvolvimento de métodos voltados à exploração de informações em ambientes educacionais (Baker et al., 2011).

Estudos recentes reforçam a relevância da Mineração de Dados Educacionais (MDE). Torres Marques et al. (2022) destacam a eficácia de técnicas de classificação na previsão da permanência estudantil, enquanto Souza e Santos (2021) evidenciam seu papel na identificação de fatores que afetam o desempenho dos alunos. De forma complementar, Albertoni et al. (2024) ressaltam o potencial da Inteligência Artificial para apoiar políticas e práticas educacionais, promovendo um ambiente de aprendizagem mais favorável ao sucesso estudantil.

Diante da relevância da MDE e de seu potencial para apoiar a tomada de decisão no contexto educacional, este estudo utilizou como base dois algoritmos de *Machine Learning (ML)* : *Random Forest* (RF) e *Self-Organizing Maps* (SOM).

O RF é um algoritmo de aprendizado supervisionado amplamente empregado em tarefas de classificação. Sua estrutura baseia-se em um conjunto de Árvores de Decisão totalmente independentes, cada uma construída com características próprias (Breiman, 2001). A combinação dos resultados provenientes dessas árvores contribui para a redução do sobreajuste (*overfitting*), ampliando a capacidade de generalização do modelo para novos dados. Além disso, a utilização de amostras aleatórias na construção das árvores proporciona resultados mais robustos frente a variações no conjunto de treinamento.

O segundo algoritmo utilizado neste estudo foi o SOM, uma Rede Neural de aprendizado não supervisionado. Esse modelo gera representações visuais simplificadas de dados multivariados ao organizar os vetores de entrada de acordo com suas similaridades, normalmente mensuradas por métricas de distância, como a Euclidiana (Deboeck e Kohonen, 2000, citados por Teixeira Almeida, s/d). Dessa forma, o SOM não apenas facilita a identificação de padrões a partir da atividade dos neurônios no mapa, mas também permite a análise da estrutura dos dados por meio das matrizes.

Nesse sentido, para aprofundar a análise proporcionada pelo SOM e compreender a explicabilidade do modelo, quanto à contribuição dos atributos na formação da rede, foi empregado o método SHAP (*SHapley Additive Explanations*), que quantifica a influência de cada variável no resultado final do modelo (Lundberg e Lee, 2017).

ENCITEC - Ensino de Ciências e Tecnologia em Revista | Santo Ângelo | v. 16 | n. 1 | p. 7-29 | jan./abr. 2026
DOI: http://dx.doi.org/10.31512/encitec.v16i1.2260

3 Método

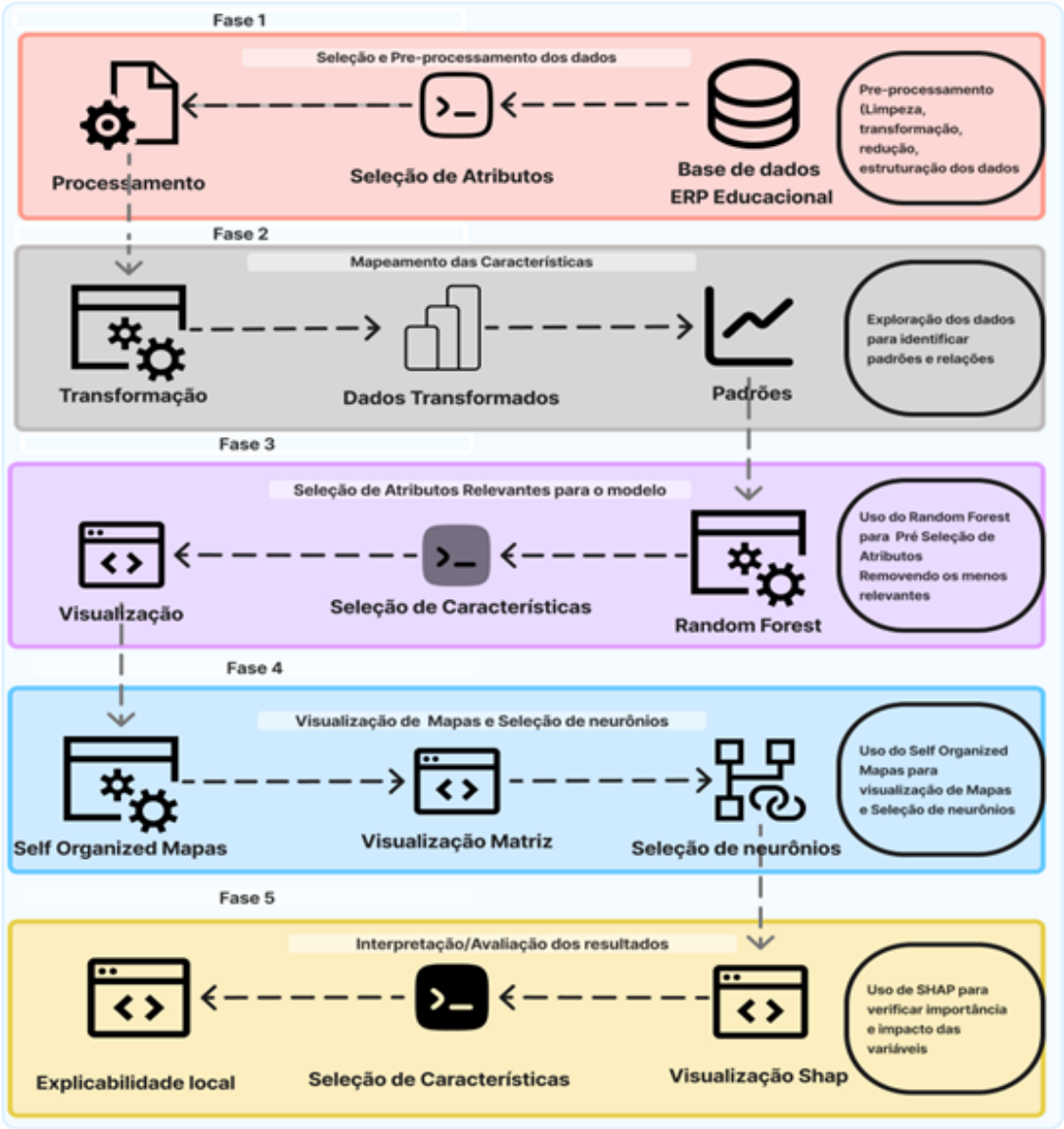
A metodologia proposta utiliza conceitos de mineração de dados educacionais e algoritmos de *Machine Learning* para analisar a evasão escolar no ensino superior, com foco em um curso de exatas, embora possa ser estendida a outros cursos. Adota abordagem quantitativa, pelo uso de dados para investigar causas da evasão, e qualitativa, pela análise documental da situação acadêmica de estudantes do curso Interdisciplinar em Ciência e Tecnologia (2014–2023) da Universidade Federal do Rio Grande do Norte, campus Natal. Além disso, é explicativa, buscando identificar fatores que contribuem para a evasão (Pereira et al., 2018).

O processo de análise foi dividido em cinco etapas, seguindo um modelo de processo KDD adaptado:

1. **Coleta e Pré-processamento dos Dados**: Os dados foram coletados do sistema acadêmico da UFRN (SIGAA) e submetidos a um processo de pré-processamento para limpeza e tratamento, incluindo a remoção de dados duplicados e a transformação de atributos.
2. **Mapeamento das características:** Após a coleta e pré-processamento dos dados, realizou-se o mapeamento das características para identificar padrões e relações associados à evasão escolar.
3. **Seleção de Atributos Relevantes**: Utilizou-se o algoritmo *Random Forest* (RF) para pré-selecionar os atributos mais relevantes na predição da evasão, gerando um conjunto otimizado de variáveis para análise.
4. **Clusterização dos Dados**: O SOM foi utilizado para agrupar os alunos em clusters com características semelhantes, explorando padrões de evasão e identificando grupos de alunos com perfis distintos.
5. **Explicabilidade do Modelo**: O método SHAP foi utilizado para analisar a contribuição de cada atributo para a predição da evasão, revelando os fatores mais determinantes para a ocorrência desse fenômeno.

ENCITEC - Ensino de Ciências e Tecnologia em Revista | Santo Ângelo | v. 16 | n. 1 | p. 7-29 | jan./abr. 2026
DOI: http://dx.doi.org/10.31512/encitec.v16i1.2260

Figura 1 - Visão completa do processo.



Fonte: (O Próprio Autor, 2024).

4 Resultados e Discussão

4.1 Coletas e Pré-processamento dos Dados

Os dados utilizados no estudo foram coletados do sistema acadêmico da UFRN (SIGAA), abrangendo informações sobre os alunos 10.811 ingressantes no curso Interdisciplinar em Ciências e Tecnologia (C&T) entre os anos de 2014 e 2023 com um total de 251.105 amostras. Após a coleta, os dados foram submetidos a um processo de pré-processamento, incluindo a limpeza e tratamento das informações. Essa etapa consistiu em:

ENCITEC - Ensino de Ciências e Tecnologia em Revista | Santo Ângelo | v. 16 | n. 1 | p. 7-29 | jan./abr. 2026
DOI: http://dx.doi.org/10.31512/encitec.v16i1.2260

- **Remoção de dados duplicados**: Foram identificados e removidos registros duplicados da base de dados, garantindo a consistência e integridade das informações.
- **Transformação de atributos**: Alguns atributos não estavam adequados para a análise de dados e precisaram ser transformados para melhor se adequarem à aplicação dos algoritmos. Essa etapa envolveu a conversão de dados categóricos para numéricos e a normalização de atributos com escalas diferentes.

4.2 Mapeamento das Características

O mapeamento das características envolveu a análise de informações dos 10.811 ingressantes, considerando gênero, origem escolar, forma de ingresso, idade e desempenho acadêmico. Os dados foram comparados entre estudantes que concluíram o curso e aqueles que evadiram, com o objetivo de identificar padrões e fatores associados à evasão.

4.3 Seleção de Atributos Relevantes

Após o mapeamento, empregou-se o *Random Forest* (RF) com o objetivo de realizar uma pré-seleção de atributos e identificar aqueles com maior influência na predição da evasão escolar. Essa etapa buscou otimizar o modelo, considerando apenas os atributos mais relevantes para a análise (Dash & Liu, 1997). Atributos com maior relevância para o desempenho do classificador foram classificados como "fortes", enquanto aqueles de baixa influência, cuja remoção não compromete a resultados do modelo, foram considerados "fracos". Essa abordagem contribuiu para aprimorar a eficácia da análise, otimizando a exploração dos dados e permitindo uma compreensão mais clara das variáveis importantes, distinguindo-as daquelas irrelevantes ou sem impacto (Santana Jr., 2015).

4.3.1 Treinamento do Modelo

Na etapa de treinamento, o modelo utilizou dados de estudantes ingressantes em 2015, totalizando 35.197 registros. Para reduzir o risco de *overfitting*, aplicou-se validação cruzada *K-Fold*. Nessa etapa, o conjunto de dados foi dividido em quatro partes, mantendo a proporção de 75% para treinamento e 25% para validação. A cada iteração, uma das partes era utilizada como conjunto de teste, enquanto as três restantes eram utilizadas para treinar o modelo, garantindo maior capacidade de generalização e evitando que o modelo se ajustasse de forma excessiva aos dados de treinamento. Em cada rodada da validação cruzada, aproximadamente 26.398 amostras eram destinadas ao treinamento e 8.799 para teste.

ENCITEC - Ensino de Ciências e Tecnologia em Revista | Santo Ângelo | v. 16 | n. 1 | p. 7-29 | jan./abr. 2026
DOI: http://dx.doi.org/10.31512/encitec.v16i1.2260

Quadro 1 - Métricas de desempenho.

Ano	*Acurácia*	*Precision*	*Recall*	*F1-Score*	*Classe*	*Amostras*
2015	***0.93***	***0.91***	***0.98***	***0.95***	***NÃO EVADIDO***	***35197***
	0.93	***0.97***	***0.86***	***0.91***	***EVADIDO***	***35197***

Fonte: (O Próprio Autor, 2024).

Os resultados apresentados no Quadro 1 demonstram o bom desempenho do modelo RF na predição da evasão escolar para os alunos ingressantes com os dados iniciais de 2015. O modelo alcançou uma acurácia geral de 93%, indicando uma boa taxa de acertos na classificação dos alunos em relação a amostra de dados.

Observando as métricas para cada classe, o modelo se destaca na precisão para a classe "Evadido" (97%), o que significa que, dentre os alunos que foram classificados como evadidos pelo modelo, 97% realmente evadiram do curso. Por outro lado, o modelo apresentou um recall ligeiramente menor para essa mesma classe (86%), indicando que nem todos alunos que efetivamente evadiram foram identificados pelo modelo. Já para a classe "Não Evadido", o modelo também apresentou boa precisão (91%) e um recall de (98%), o que significa que o modelo foi eficaz em identificar corretamente os alunos que permaneceram no curso.

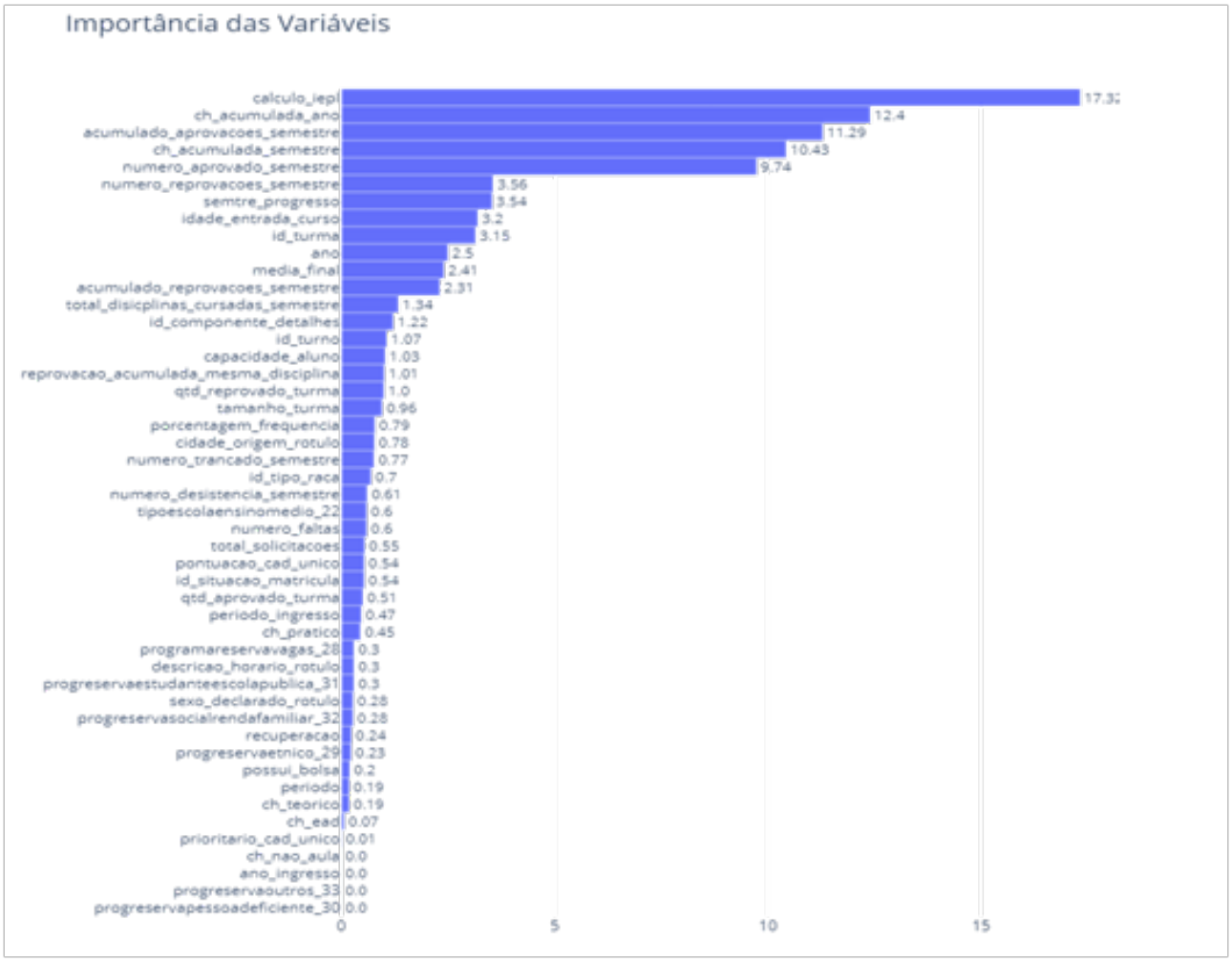
De forma geral, o modelo demonstrou uma boa capacidade de identificar os alunos evadidos (precisão) e a capacidade de prever a maioria dos casos reais de evasão (recall), o que é refletido no valor do F1-Score para ambas as classes. Após a avaliação do modelo, seguimos para a visualização dos atributos candidatos. Este processo é importante de modo que se possa compreender a força dos atributos, escolhendo apenas aqueles que têm maior relevância na previsão, como também na identificação de variáveis que possam afetar a evolução dos estudantes no decorrer do curso.

4.3.2 Avaliação e Seleção

Os testes iniciais foram realizados com a proporção de 25% dos dados de ingressantes de 2015 em cada interação, posteriormente, com dados dos ingressantes de 2016 e 2017, incluindo dados do histórico acadêmico e informações ao longo de sua trajetória acadêmica com a instituição. Isto garantiu a robustez do modelo e sua capacidade de generalização para dados desconhecidos. A análise dos resultados permitiu identificar os atributos com maior influência na predição da evasão escolar (Figura 2). Os atributos com baixa influência foram removidos do modelo, otimizando sua performance e simplificando a estrutura.

ENCITEC - Ensino de Ciências e Tecnologia em Revista | Santo Ângelo | v. 16 | n. 1 | p. 7-29 | jan./abr. 2026
DOI: http://dx.doi.org/10.31512/encitec.v16i1.2260

Figura 2 - Pré-seleção de Atributos.



Fonte: (O Próprio Autor, 2024).

Atributos com pouca influência na capacidade preditiva do modelo, com valores de importância inferiores a 0.01, foram considerados “fracos” e removidos do modelo para evitar o *overfitting*. A remoção de atributos irrelevantes contribuiu para a robustez do modelo, evitando que variáveis com pouca relevância influenciassem nas previsões. Além disso, o torna mais interpretável, facilitando a compreensão dos fatores que realmente impactam as decisões.

Após a remoção dos atributos “fracos”, o modelo RF foi retestado com os 25% dos dados de 2015 para verificar se a remoção dos atributos impactou o desempenho do modelo. Os resultados apresentados no Quadro 2 demonstram que o modelo manteve um desempenho consistente, com métricas de acurácia, precisão, recall e F1-score muito próximas às observadas antes da seleção de atributos. Esta constatação indica que os atributos removidos não contribuem significativamente para a capacidade preditiva do modelo e que a remoção não prejudicou a capacidade de predição.

ENCITEC - Ensino de Ciências e Tecnologia em Revista | Santo Ângelo | v. 16 | n. 1 | p. 7-29 | jan./abr. 2026
DOI: http://dx.doi.org/10.31512/encitec.v16i1.2260

Quadro 2 - Métricas de desempenho do modelo após a remoção de atributos.

Ano	*Acurácia*	*Precision*	*Recall*	*F1-Score*	*Classe*	*Amostras*
2015	***0.92***	***0.90***	***0.98***	***0.94***	***NÃO EVADIDO***	***35197***
	0.92	***0.97***	***0.86***	***0.91***	***EVADIDO***	***35197***

Fonte: (O Próprio Autor, 2024).

Em seguida, buscou-se testar a capacidade de generalização do modelo RF e sua aplicabilidade em cenários distintos, realizou-se testes adicionais com os dados de alunos ingressantes em anos subsequentes àquele utilizado para teste inicial com as amostras de (2015). Para tanto, foram utilizados todos os dados dos ingressantes de 2016 e 2017, constituindo conjuntos de dados desconhecidos para o modelo.

Vale ressaltar que o modelo pode ser treinado novamente a cada semestre para aumentar a qualidade de suas respostas. A escolha dos anos de 2016 e 2017 visa garantir a similaridade na estrutura curricular do curso C&T, minimizando a influência de possíveis mudanças no currículo que poderiam impactar os resultados. Esta estratégia permite analisar o desempenho do modelo em um contexto similar ao do treinamento e testes iniciais, avaliando sua capacidade de previsão com novos dados, coletados em períodos distintos e com dados desconhecidos.

Quadro 3 - Métricas de desempenho do modelo com dados desconhecidos.

Ano	*Acurácia*	*Precision*	*Recall*	*F1-Score*	*Classe*	*Amostras*
2016	***0.91***	***0.91***	***0.97***	***0.94***	***NÃO EVADIDO***	***36509***
	0.91	***0.94***	***0.85***	***0.89***	***EVADIDO***	***36509***
2017	***0.89***	***0.92***	***0.93***	***0.93***	***NÃO EVADIDO***	***35714***
	0.89	***0.84***	***0.82***	***0.83***	***EVADIDO***	***35714***

Fonte: (O Próprio Autor, 2024).

Observando-se o Quadro 3, foi possível analisar o desempenho do modelo para cada classe em cada ano de validação:

- 2016: O modelo apresentou um desempenho consistente em ambas as classes, com valores de precisão, recall e F1-Score acima de 85%. Com o recall para a classe “Não Evadido” (97%), o que significa que o modelo identificou corretamente a grande maioria dos alunos que permaneceram no curso.
- 2017: Embora a acurácia geral tenha sido ligeiramente menor em 2017 (89%), o modelo ainda apresenta bom desempenho, com métricas acima de 82% para ambas as classes. A precisão para “Não Evadido” (92%) e o recall para “Evadido” (82%) indicam um bom equilíbrio na

ENCITEC - Ensino de Ciências e Tecnologia em Revista | Santo Ângelo | v. 16 | n. 1 | p. 7-29 | jan./abr. 2026
DOI: http://dx.doi.org/10.31512/encitec.v16i1.2260

capacidade do modelo de identificar corretamente tanto os alunos que permaneceram quanto os que evadiram do curso.

Comparando os resultados da validação com outros estudos que avaliam modelos preditivos para a evasão escolar, as acurácias obtidas neste estudo (91% para 2016 e 89% para 2017) se mostram alinhadas com a literatura, demonstrando a eficácia do modelo RF na identificação de alunos em risco de evasão. O Quadro 4 apresenta os resultados obtidos com outros estudos na área.

Quadro 4 - Resultado de outros estudos na área com Random Forest.

Ano	*Acurácia*	*Precision*	*Recall*	*F1-Score*	*Classe*	*Autores*
2024	*0.87*	*0.91*	*0.91*	*0.91*	*EVADIDO*	*OLIVEIRA; MEDEIROS, 2024*
2022	*0.86*	*0.83*	*0.89*	*0.86*	*EVADIDO*	*MELO;SOUZA;SANTOS, 2022*
2020	*87,1*	*92,4*	*85,6*	**	*EVADIDO*	*SANTOS;GOYA, 2020*

Fonte: (O Próprio Autor, 2024).

A predição do modelo com dados de 2016 e 2017 reforçou a robustez do algoritmo RF e sua capacidade de generalizar para novos dados. Com resultados de acurácia consistentes e grande capacidade de identificar os principais atributos relacionados à evasão. Entretanto, é importante considerar que a maioria das pesquisas utilizadas como base de comparação apresenta limitações em relação ao tamanho de suas amostras e ao número de variáveis analisadas. A maioria dos estudos utiliza menos de 3 mil alunos e menos de 20 variáveis. No entanto, o presente estudo, ao analisar um conjunto de dados amplo, com um período de tempo de 9 anos, com mais 3 mil alunos e mais de 40 variáveis, se mostra capaz de gerar insights mais profundos e confiáveis sobre os fatores que podem influenciar a evasão escolar.

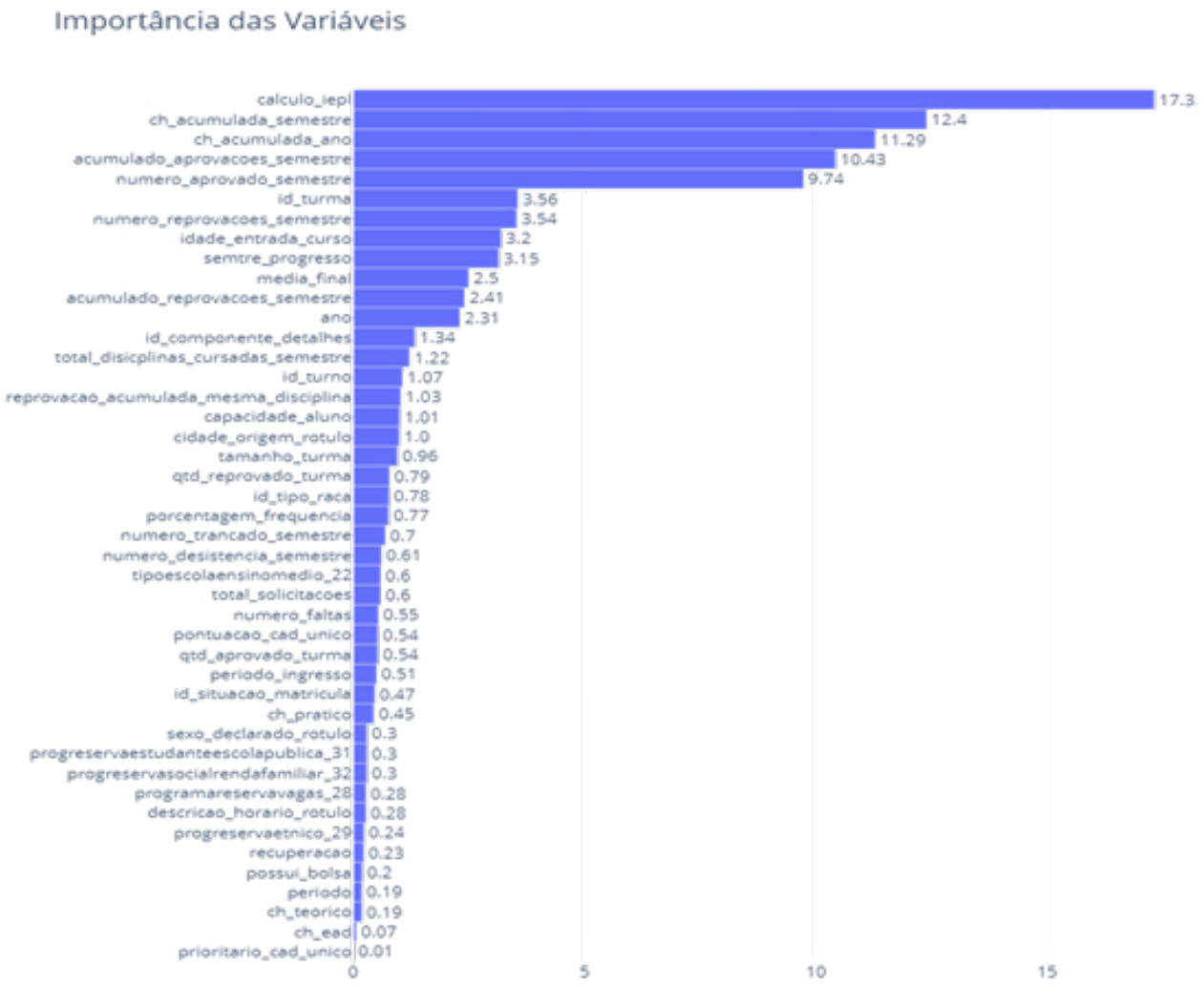
Após a etapa de pré-seleção de atributos com o RF, o presente estudo utilizou o algoritmo SOM para aprofundar a análise dos dados e gerar insights sobre os padrões de evasão no curso C&T.

4.4 Agrupamentos dos Dados

Após a fase de teste e a remoção dos atributos com valores inferiores a 0.01, conforme destacados na (Figura 3). Os atributos foram então selecionados para geração da matriz da rede.

ENCITEC - Ensino de Ciências e Tecnologia em Revista | Santo Ângelo | v. 16 | n. 1 | p. 7-29 | jan./abr. 2026
DOI: http://dx.doi.org/10.31512/encitec.v16i1.2260

Figura 3 - Atributos Selecionados para a Rede SOM.



Fonte: (O Próprio Autor, 2024).

Para a geração da rede SOM, utilizou-se todos os dados disponíveis, abrangendo informações sobre os 10.811 alunos ingressantes no curso Interdisciplinar em Ciências e Tecnologia (C&T) entre os anos de 2014 e 2023. Isto incluiu dados do histórico acadêmico e informações de todo o vínculo dos alunos com a instituição, totalizando 196.517 amostras.

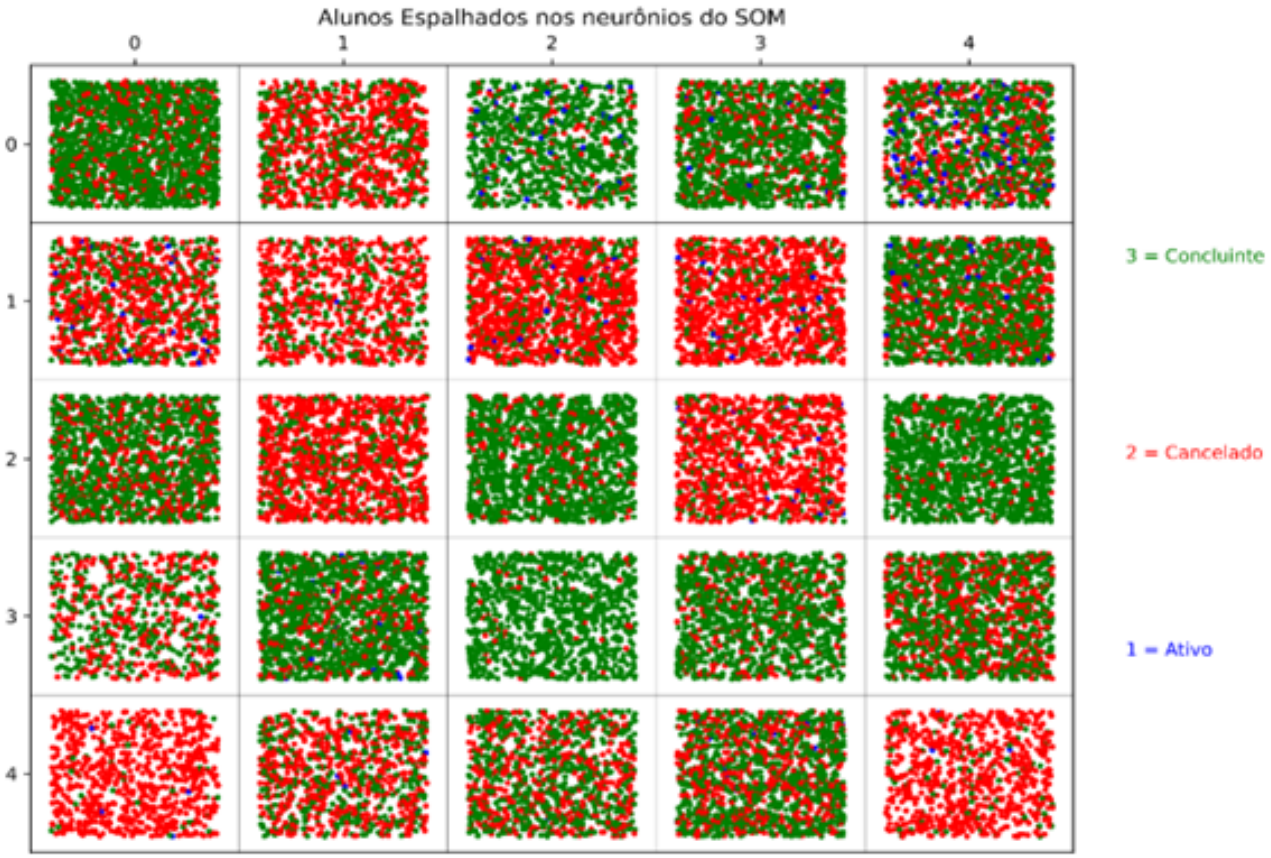
A Rede SOM proporcionou uma análise visual por meio de um mapa de características que ilustra a relação entre diversos atributos e a situação educacional dos alunos. Esta abordagem possibilitou uma compreensão mais aprofundada dos dados, facilitando a identificação de grupos que compartilham características semelhantes e, consequentemente, o processo de compreensão desses alunos.

Na (Figura 4), traz-se a Matriz de rede SOM com agrupamento por similaridade, revelando-se a capacidade do SOM de organizar os dados, agrupando alunos com características semelhantes em neurônios próximos no mapa, o que facilita a identificação de padrões e perfis de evasão. Esta matriz

ENCITEC - Ensino de Ciências e Tecnologia em Revista | Santo Ângelo | v. 16 | n. 1 | p. 7-29 | jan./abr. 2026
DOI: http://dx.doi.org/10.31512/encitec.v16i1.2260

destaca as regiões de maior e menor densidade de alunos em cada neurônio. A cor de cada ponto representa a situação do aluno no curso: verde para “Concluído”, vermelho para “Cancelado” e azul para “Ativo”.

Figura 4 - Matriz de Rede SOM com Agrupamento por similaridade.



Fonte: (O Próprio Autor, 2024).

Observa-se que a distribuição das cores da (Figura 4) não é aleatória. Alguns neurônios são predominantemente compostos por alunos “Concluintes”, enquanto outros concentram alunos “Cancelados”, e estes neurônios com altas concentrações de uma mesma categoria tendem a se agrupar em regiões específicas do mapa. Esta proximidade reforça a ideia de que o SOM identificou clusters de alunos com perfis distintos que influenciam na sua probabilidade de evasão. A presença de alunos “Ativos” (azul) em diferentes neurônios, por sua vez, ilustra a heterogeneidade daqueles que ainda estão cursando, representando diferentes níveis de risco de evasão.

4.5 Explicabilidade do Modelo

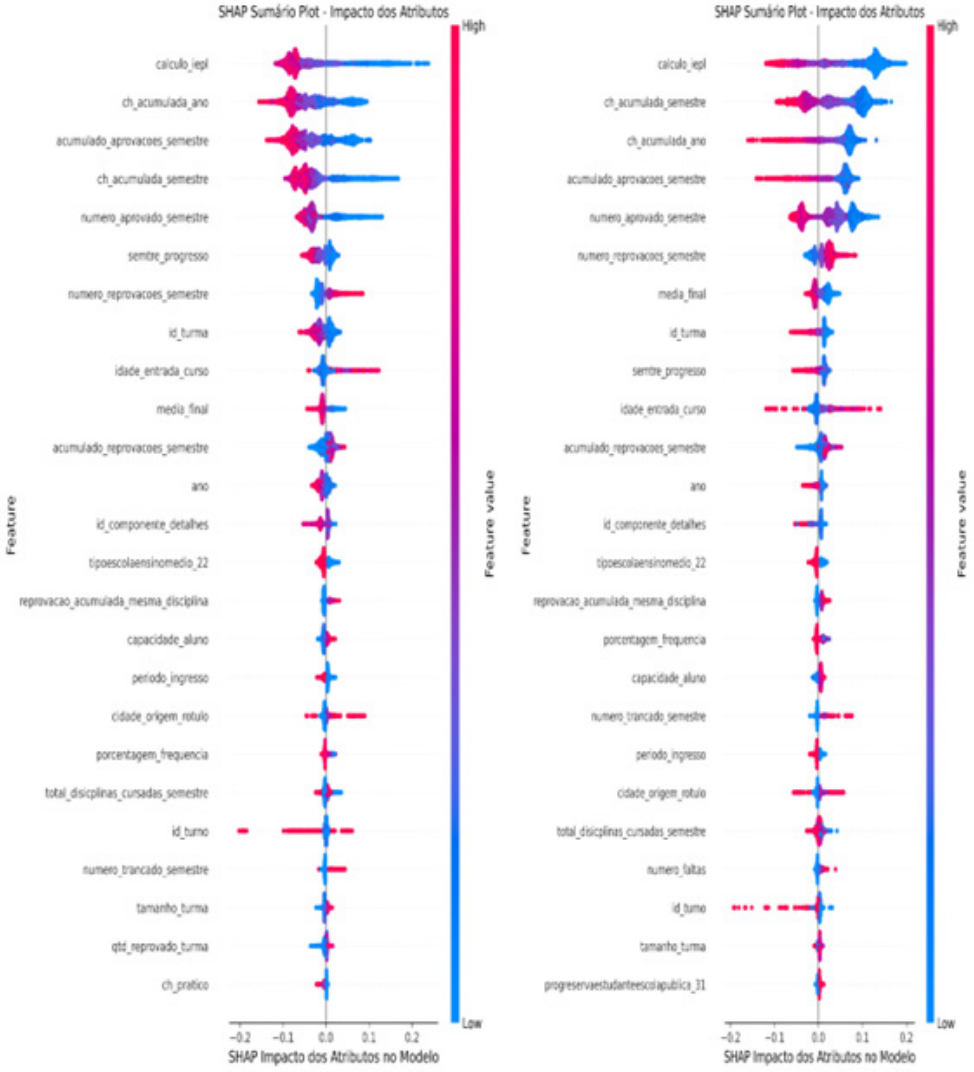
Para aprofundar a análise dos fatores que influenciam o desempenho dos estudantes, especialmente em relação à evasão, utilizou-se o método SHAP que permitiu a explicação global ou individual, calculando a contribuição de cada atributo para a previsão final do modelo, fornecendo insights sobre a influência de cada variável na probabilidade de evasão.

ENCITEC - Ensino de Ciências e Tecnologia em Revista | Santo Ângelo | v. 16 | n. 1 | p. 7-29 | jan./abr. 2026
DOI: http://dx.doi.org/10.31512/encitec.v16i1.2260

Além de analisar o neurônio vencedor, explorou-se, também, os neurônios próximos influenciados por ele. Esta abordagem amplia a capacidade de identificar padrões e perfis de evasão, permitindo uma visão mais completa do comportamento dos alunos. Ao considerar a vizinhança de neurônios, conseguimos identificar nuances e variações nas características dos alunos, o que proporciona insights sobre os fatores que contribuem para a evasão escolar. Tal combinação de SHAP e Rede SOM reforça a qualidade destas técnicas na análise educacional, destacando sua capacidade de revelar relações complexas e padrões nos dados educacionais.

Dessa forma, escolheu-se neurônios específicos da matriz do SOM, que exibem concentrações significativas de alunos "Cancelados" (representados em vermelho) nas posições [1,0], [1,1], [1,2] e [2,1], assim como de "Concluintes" (representados em verde) nas posições [3,1], [3,2], [3,3] e [2,2]. Esse procedimento visa identificar os atributos que mais influenciam a probabilidade de evasão nos grupos.

Figura 5 A e B - Densidade dos dados em relação aos valores Shapley nos grupos analisados.



A) Neurônio com alunos cancelados

B) Neurônio com alunos concluintes

Fonte: (O Próprio Autor, 2024).

ENCITEC - Ensino de Ciências e Tecnologia em Revista | Santo Ângelo | v. 16 | n. 1 | p. 7-29 | jan./abr. 2026
DOI: http://dx.doi.org/10.31512/encitec.v16i1.2260

A avaliação da importância global dos atributos (Figura 5 A e B) mostra uma compreensão abrangente do modelo, conforme a análise dos neurônios selecionados para avaliar os diferentes grupos de alunos.

Na **Figura 5 (A)**, buscou-se analisar os neurônios com maior densidade de alunos cancelados, observando-se que os atributos destes neurônios apresentam uma maior densidade de amostras, influenciando o modelo a responder à classe positiva, ou seja, classificando os alunos como evadidos.

Na **Figura 5 (B)**, passou-se a analisar os neurônios com maior concentração de alunos concluintes, observando-se uma maior densidade de amostras influenciando o modelo a responder à classe negativa, classificando os alunos como não evadidos.

Para uma compreensão mais clara da figura acima, alguns pontos importantes foram destacados:

- **Categorização:** Os valores positivos estão associados aos alunos com risco de evasão. Já os valores negativos estão associados à permanência dos alunos e consequentemente a conclusão.
- **Importância do recurso**: Os atributos são classificados em ordem decrescente de acordo com sua relevância para o modelo. Isso nos permite identificar quais características têm maior impacto na previsão.
- **Impacto:** A localização horizontal indica se o efeito desse valor está associado a uma previsão com impacto positivo, influenciando o modelo a caracterizar um aluno como evadido com base nesse atributo, ou negativo, influenciando a caracterização do aluno como não evadido.
- **Cores**: As cores indicam se o impacto do atributo é alto (em vermelho) ou baixo (em azul) para o modelo. Isso nos ajuda a visualizar como cada atributo contribui para as previsões do modelo e sua influência no resultado final.
- **Escala**: Na linha central, temos o eixo 0, que indica nenhum impacto sobre o modelo. Valores negativos indicam que o atributo contribui para caracterizar o aluno como não evadido, enquanto valores positivos indicam que o atributo contribui para caracterizar o aluno como possível evadido.

A análise conjunta dos *clusters* do SOM e do SHAP permite uma compreensão mais profunda do problema da evasão, complementando as informações e fornecendo *insights* importantes para a criação de estratégias de intervenção mais eficazes.

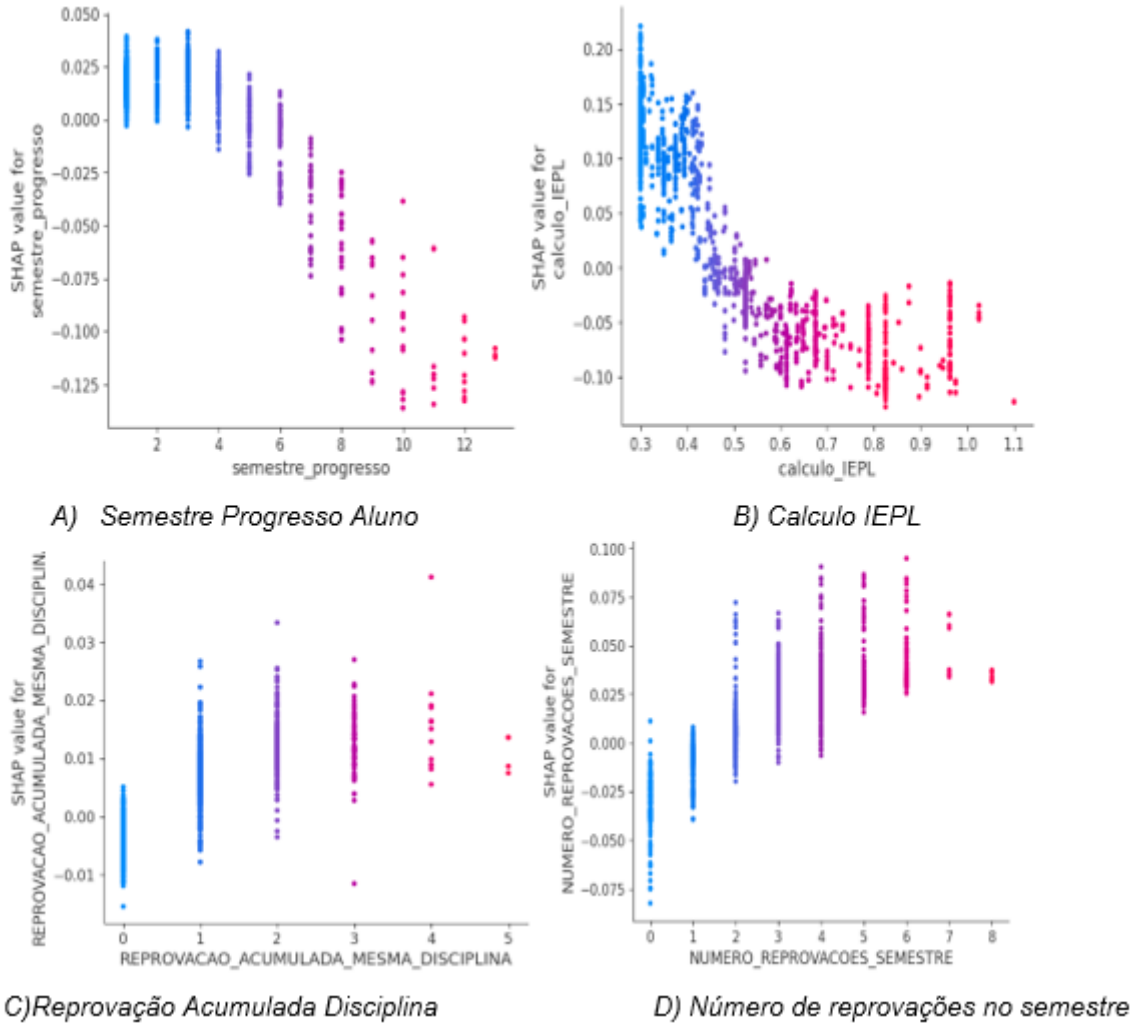
4.5.1 Contribuição dos Atributos na Predição do Modelo

Além da explicabilidade global, que permite o entendimento de forma abrangente a influência dos atributos nas previsões do modelo, também foi explorado o comportamento específico das características de alguns atributos usando os neurônios selecionados na rede SOM, como destacado na Figura 4. Estes neurônios agrupam características similares representadas pelo mesmo vetor na rede SOM, considerando tanto o neurônio vencedor quanto o padrão de entrada. Esta abordagem facilita a identificação de padrões com características semelhantes, o que enriquece a compreensão sobre os fatores que influenciam a evasão.

ENCITEC - Ensino de Ciências e Tecnologia em Revista | Santo Ângelo | v. 16 | n. 1 | p. 7-29 | jan./abr. 2026
DOI: http://dx.doi.org/10.31512/encitec.v16i1.2260

Observando-se a (Figura 6 A, B, C, D), pode-se ver as escalas em que o eixo horizontal (X) representa os valores dos atributos analisados, enquanto o eixo vertical (Y) representa as escalas de predição do modelo, com as pontuações obtidas. A linha central, posicionada no eixo 0, indica nenhum impacto no modelo. Valores negativos indicam que o atributo tende a influenciar o modelo a responder à classe negativa, caracterizando o aluno como não evadido. Por outro lado, valores positivos indicam que o atributo tende a influenciar o modelo a responder à classe positiva, caracterizando o aluno como possível evadido. As cores, em vermelho para impacto alto e em azul para impacto baixo, destacam visualmente como cada atributo contribui para as previsões do modelo e sua influência nos resultados finais dos grupos analisados.

Figura 6 - Gráfico de Dependência 1.



Fonte: (O Próprio Autor, 2024).

A **Figura 6 (A)** está relacionada à permanência do aluno no curso, destacando que os cinco primeiros semestres têm uma importância positiva para a classificação do modelo como evasão. Entretanto, a partir do sexto semestre, esta influência passa a ter um viés negativo, indicando não evasão. Esta mudança

ENCITEC - Ensino de Ciências e Tecnologia em Revista | Santo Ângelo | v. 16 | n. 1 | p. 7-29 | jan./abr. 2026
DOI: http://dx.doi.org/10.31512/encitec.v16i1.2260

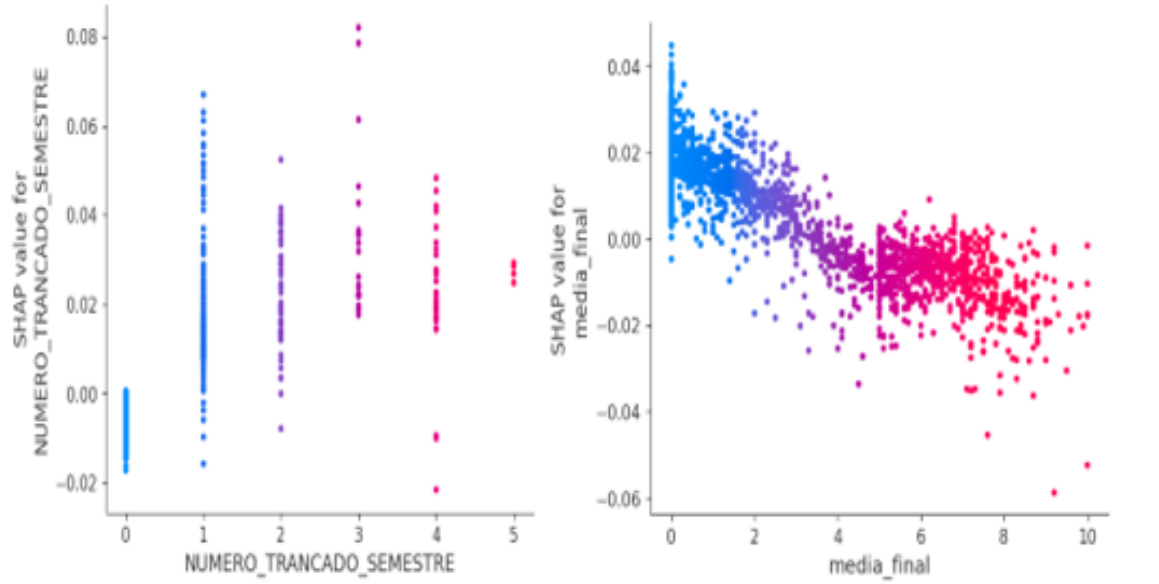
na dinâmica sugere que os primeiros anos são mais críticos em termos de previsão de evasão, enquanto a estabilidade após esse período indica maior probabilidade de continuidade no curso.

A **Figura 6 (B)** está relacionada ao IEPL em relação ao percentual da carga horária realizada em comparação com a carga horária esperada de um aluno em cada período letivo do curso. Ao analisar o gráfico, percebe-se a influência desse atributo na classificação do modelo, mostrando que alunos que mantêm o percentual baixo de 0,5 têm uma influência positiva para classificar o aluno como possível evasão. Por outro lado, valores acima desse limiar indicam um melhor desempenho do aluno, levando essa influência a ter um viés negativo e indicando não evasão. Este padrão sugere que a eficiência do aluno em cumprir sua carga horária esperada é um indicador importante da probabilidade de evasão, um baixo índice de IEPL sendo um alerta precoce para possíveis problemas.

A **Figura 6 (C)** apresenta a relação de reprovações em uma mesma disciplina, mostrando que esse atributo tem uma forte ligação com a evasão. Conforme o número de reprovações aumenta, sua influência proporcional também aumenta, o que evidencia sua grande relevância na classificação do aluno como possível evadido. Este resultado sugere que repetir a mesma disciplina pode ser um indicador importante pois mostra dificuldades acadêmicas persistentes, aumentando significativamente o risco de evasão.

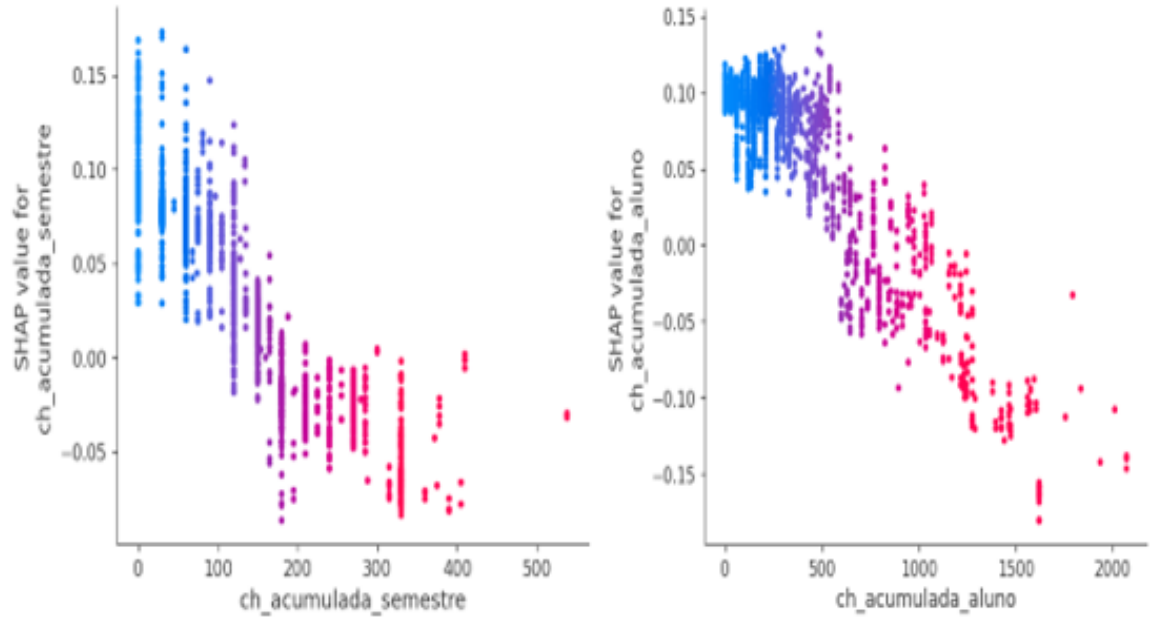
A **Figura 6 (D)** está relacionada à quantidade de reprovações por semestre, e conforme o número de reprovações aumenta, sua relação com a evasão também cresce proporcionalmente. Este atributo se torna relevante, pois pode indicar que o aluno está enfrentando dificuldades no progresso acadêmico ao reprovar em várias disciplinas ao longo dos semestres. Tal padrão reforça a importância de monitorar o desempenho contínuo do aluno ao longo do curso e intervir precocemente quando as reprovações se acumulam.

Figura 7 - Gráfico de Dependência 2.



A) Quantidade de Trancamentos *B) Média do Aluno nas Disciplinas*

ENCITEC - Ensino de Ciências e Tecnologia em Revista | Santo Ângelo | v. 16 | n. 1 | p. 7-29 | jan./abr. 2026
DOI: http://dx.doi.org/10.31512/encitec.v16i1.2260



C)Carga horária acumulada em cada semestre *D) Carga horária Total acumulada*

Fonte: (O Próprio Autor, 2024).

A **Figura 7 (A)** está relacionada ao número de trancamentos dos semestres, e ao analisar o gráfico, é possível perceber sua influência positiva para classificar o aluno como evadido. Isto sugere que o número de trancamentos está correlacionado com a possível evasão, já que os alunos que realizam muitos trancamentos acabam por prolongar sua permanência no curso. De maneira geral, os inúmeros trancamentos podem estar relacionados com uma possível evasão, indicando dificuldades ou falta de comprometimento por parte do aluno.

A **Figura 7 (B)** mostra a relação entre o desempenho do aluno nas disciplinas. Ao observar este atributo, é possível perceber sua relação com a classificação em ambos os casos. Alunos com notas inferiores a 5 tendem a influenciar o modelo de forma positiva para caracterizar o aluno como evadido. Por outro lado, notas mais altas tendem a afetar negativamente o que poderia caracterizar o aluno como não evadido, com base no desempenho nas disciplinas por meio das notas obtidas. Essa análise destaca a importância do acompanhamento do desempenho acadêmico como um indicador-chave na previsão de evasão.

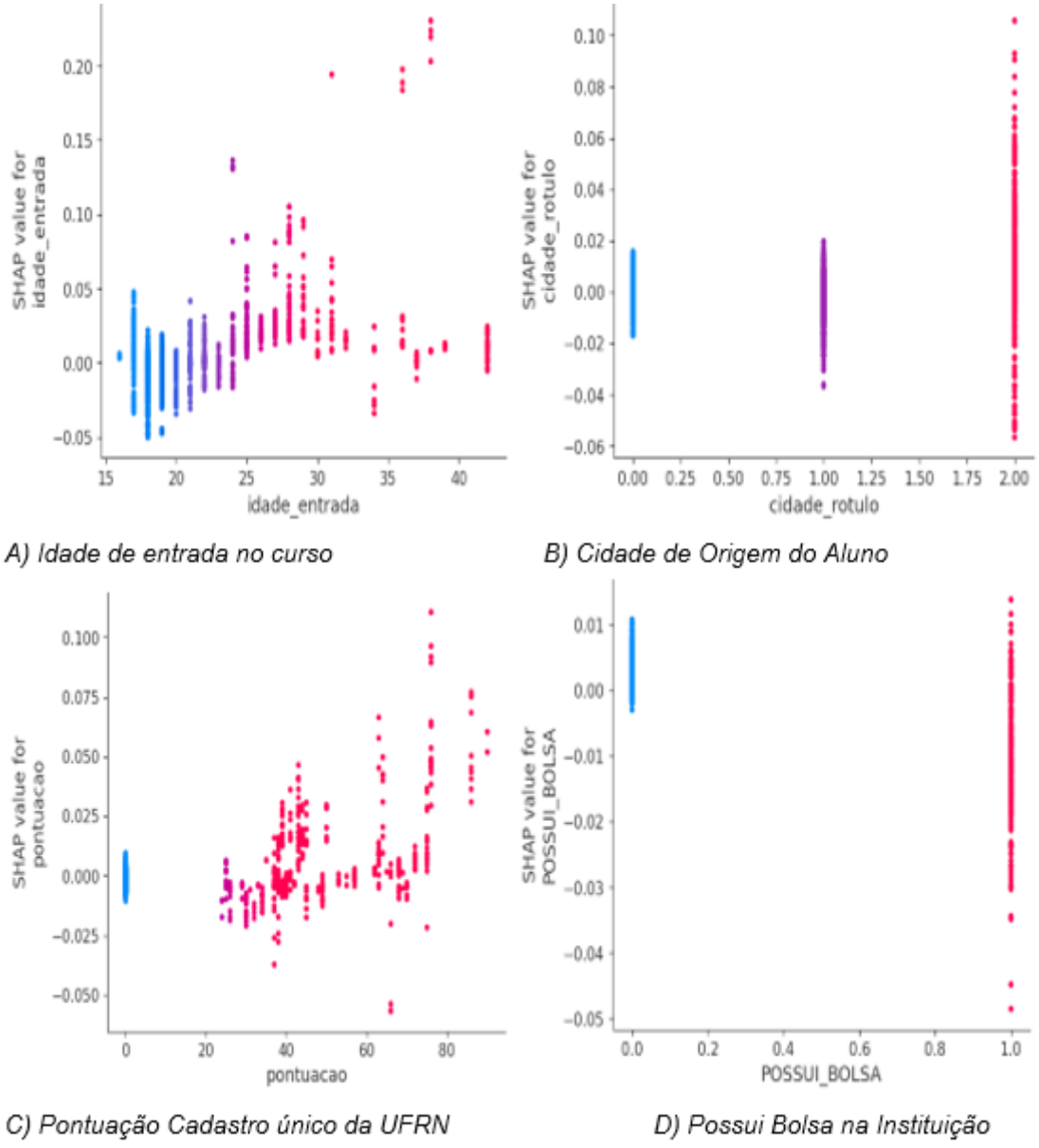
A **Figura 7 (C)** avalia o acúmulo de carga horária das disciplinas cursadas durante um semestre. Percebe-se que um baixo acúmulo de carga horária exerce uma grande influência na caracterização de um aluno como possível evadido. Por outro lado, um maior acúmulo impacta negativamente o modelo na caracterização do aluno como não evadido. Isto sugere que a carga horária das disciplinas cursadas pode ser um indicador significativo de risco de evasão, destacando a importância de analisar adequadamente o acúmulo de carga horária dos alunos.

A **Figura 7 (D)** complementa a análise anterior, mostrando que um baixo acúmulo de carga horária total ao longo dos semestres influencia positivamente a classificação de um aluno como evadido. Por outro

ENCITEC - Ensino de Ciências e Tecnologia em Revista | Santo Ângelo | v. 16 | n. 1 | p. 7-29 | jan./abr. 2026
DOI: http://dx.doi.org/10.31512/encitec.v16i1.2260

lado, alunos que acumulam carga horária de forma consistente mostram progresso e engajamento, sendo classificados como não evadidos. Isto reforça a importância de considerar o padrão de acúmulo de carga horária na identificação de alunos em risco de evasão.

Figura 8 - Gráfico de Dependência 3.



Fonte: (O Próprio Autor, 2024).

A **Figura 8 (A)** destaca a importância da variável idade na entrada dos alunos no curso, mostrando que alunos mais jovens, entre 17 e 24 anos, tendem a influenciar negativamente o modelo como alunos com menos risco de evasão. À medida que a idade aumenta, esta influência passa a ter um forte impacto no modelo para caracterizar um aluno como possível evadido. Este padrão sugere que a idade pode ser um indicador significativo das trajetórias acadêmicas dos alunos.

A **Figura 8 (B)** ressalta a importância da cidade de origem dos alunos, demonstrando que alunos que residem na capital (indicado por 0) têm uma influência limitada no modelo, seja de forma negativa ou positiva. Por outro lado, alunos originários do interior do estado (indicado por 1) exercem uma influência relativa em ambos os casos. Enquanto isso, alunos provenientes de outros estados (indicado por 2) tendem a influenciar positivamente para caracterizar um aluno como possível evadido. Esta análise destaca como a cidade de origem dos alunos pode ser um fator importante na previsão de evasão, revelando padrões distintos de influência com base na localização geográfica dos estudantes. Além disso, fatores como acesso a recursos educacionais, qualidade de ensino e oportunidades de emprego podem variar entre diferentes regiões, afetando assim a trajetória acadêmica dos alunos.

A **Figura 8 (C)** destaca o fator social, representado pela pontuação obtida ao realizar o cadastro único com perfil de estudante prioritário, que possibilita concorrer e ser contemplado com auxílios e bolsas da Assistência Estudantil. De forma geral, esse sistema prioriza os discentes mais carentes, onde pontuações menores refletem uma maior carência. O que chama atenção neste atributo é que mesmo em pontuações baixas, ele tem uma forte influência na classificação do aluno como evadido ou não evadido. É notável também as pontuações mais altas de alunos que, em algum momento, pleitearam auxílios e bolsas. No entanto, por terem pontuações mais altas, deixaram de ser considerados prioritários e, consequentemente, podem influenciar na decisão do aluno. Isto ressalta a importância do suporte financeiro e da assistência social na retenção e no sucesso acadêmico dos alunos, especialmente aqueles em situação de vulnerabilidade socioeconômica.

A **Figura 8 (D)** destaca o atributo "possui bolsa", complementando a análise anterior. Neste aspecto, é possível perceber que discentes sem bolsa têm pouca influência na classificação do aluno como evadido. Por outro lado, os alunos que possuem algum tipo de bolsa exercem uma forte influência para caracterizar o aluno como não evadido. Isso sugere que o acesso a bolsas de estudo e auxílios financeiros pode ser um fator importante na retenção dos alunos e na conclusão de seus estudos. Além disso, evidencia a importância de políticas de inclusão e apoio financeiro para garantir a igualdade de oportunidades no ensino superior.

5 Conclusão

Este estudo demonstrou a eficácia de uma metodologia híbrida, que combina técnicas de Mineração de Dados Educacionais (MDE) e *Machine Learning* (ML), para identificar fatores relacionados à evasão escolar no curso Interdisciplinar em Ciências e Tecnologia (C&T) da Universidade Federal do Rio Grande do Norte (UFRN). A integração de MDE e ML permitiu o desenvolvimento de um modelo preditivo que alcançou uma acurácia de 93% na identificação de alunos em risco de evasão. O modelo *Random Forest* (RF), além dos testes iniciais, foi retestado com dados adicionais para ingressantes de 2016 e 2017 avaliando a capacidade de generalização para dados desconhecidos, apresentando resultados consistentes, com uma acurácia de 91% e 89% para os respectivos anos.

A pesquisa revelou que a evasão escolar é influenciada por uma combinação complexa de fatores, incluindo aspectos curriculares, socioeconômicos e demográficos. Através da análise com *Self-Organizing Maps* (SOM), foram identificados clusters de alunos com características semelhantes, evidenciando a heterogeneidade dos perfis dos estudantes. A técnica *SHapley Additive exPlanations* (SHAP), aplicada aos clusters, destacou a importância de variáveis como o desempenho acadêmico (semestre de progresso, IEPL, reprovações), a idade de ingresso no curso, a cidade de origem e a condição socioeconômica, representada pela pontuação no Cadastro Único da UFRN.

O estudo evidenciou que a concessão de bolsas e outros tipos de apoio financeiro pela instituição se mostra importante na retenção dos alunos, impactando positivamente a permanência e o sucesso dos estudantes.

É importante destacar as limitações da pesquisa. A falta de uma estrutura robusta para a implantação da ferramenta e a concentração do estudo em um único curso o C&T da UFRN, o que limitou a generalização dos resultados para outros cursos da instituição e para outros contextos educacionais. A utilização de uma amostra específica também pode influenciar a representatividade dos resultados, No entanto, essa abordagem pode ser aplicada a outros cursos.

Para superar essas limitações, recomenda-se que pesquisas futuras adotem uma abordagem mais abrangente, incluindo outros cursos da UFRN, outras instituições de ensino superior e diferentes contextos socioeconômicos. A ampliação do escopo da pesquisa permitirá uma melhor compreensão dos fatores que influenciam a evasão escolar e a elaboração de soluções mais eficazes para enfrentá-la.

A pesquisa demonstrou o potencial da combinação de MDE e ML para auxiliar na compreensão e na prevenção da evasão escolar, fornecendo insights valiosos para a criação de estratégias mais eficazes para garantir a permanência e o sucesso dos estudantes. No entanto, a realização de estudos futuros mais amplos é fundamental para fortalecer a base de conhecimento e generalizar os resultados para outros contextos.

Referências

BARDAGI, M. P.; HUTZ, C. S. Rotina acadêmica e relação com colegas e professores: impacto na evasão universitária. ***Psico***, [S. l.], v. 43, n. 2, 2012. Disponível em: https://revistaseletronicas.pucrs.br/ojs/index.php/revistapsico/article/view/7870. Acesso em: 17 maio 2024.

BAKER, R. S. J. de; CARVALHO, A. M. J. B. de; ISOTANI, S. Mineração de dados educacionais: oportunidades para o Brasil. ***Revista Brasileira de Informática na Educação***, v. 19, 2011. Disponível em: https://doi.org/10.5753/RBIE.2011.19.02.03. Acesso em: 23 out. 2023.

BREIMAN, L. Random forests. ***Machine Learning*** *Kluwer Academics*, vol. 45, p. 5–32, 2001.

COSTA, Herbert da Silva; CARDOSO, Anderson Cordeiro; NETTO, Cristiane Mendes; MARTINS-JR, David Correa; SIMÕES, Sérgio Nery. A framework for prediction of dropout in distance learning through XAI techniques in Virtual Learning Environment. In: ***ENCONTRO NACIONAL DE***

INTELIGÊNCIA ARTIFICIAL E COMPUTACIONAL (ENIAC), 19., 2022, Campinas/SP. **Anais** [...]. Porto Alegre: Sociedade Brasileira de Computação, 2022. p. 270–281. ISSN 2763-9061. DOI: https://doi.org/10.5753/eniac.2022.227586.

DASH, M.; LIU, H. Feature selection for classification. ***Intelligent Data Analysis***, v. 1, p. 131–156, 1997.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. Da mineração de dados à descoberta de conhecimento em bancos de dados. ***Revista AI***, [S. l.], v. 3, p. 37, 1996. DOI: 10.1609/aimag.v17i3.1230. Disponível em: https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/1230. Acesso em: 18 jan. 2024.

KANTORSKI, Gustavo Zanini; MARTINS, Ricardo Zimmermann; BALEJO, Arthur; FRICK, Marcio. Mineração de Dados Educacionais para predição da evasão em cursos de graduação presenciais no ensino superior. In: ***SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO (SBIE)***, 34., 2023, Passo Fundo/RS. **Anais** [...]. Porto Alegre: Sociedade Brasileira de Computação, 2023. p. 1133–1142. DOI: https://doi.org/10.5753/sbie.2023.233746.

LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. In: ***Proceedings of the 31st International Conference on Neural Information Processing Systems***. [S. l.]: s.n., 2017. p. 4768–4777.

MASCHIO, P. d. T.; VIEIRA, M. A.; da COSTA, N. T.; MELO, S.; JÚNIOR, C. P. Um panorama acerca da mineração de dados educacionais no Brasil. In: ***Anais do XXIX Simpósio Brasileiro de Informática na Educação (SBIE 2018)***, Porto Alegre, RS, Brasil, 2018. p. 1936–1940.

MELO, Eduardo Cardoso; SOUZA, Fernanda Sumika Hojo de; SANTOS, Edimilson Batista dos. Predição da evasão escolar nos cursos superiores do IFMG – Campus Bambuí com o apoio de técnicas de aprendizado de máquina. ***Revista Eletrônica de Sistemas de Informação e Gestão Tecnológica***, v. 12, n. 1, 2022. Disponível em: http://periodicos.unifacef.com.br/resiget/article/view/2342/1719. Acesso em: 15 abr. 2024.

MESQUITA, J. L. de; SOUSA, R. R. de; NASCIMENTO, S. M.; SOUZA, T. F. de. Academic analytics como apoio ao sucesso na graduação: uma revisão sistemática da literatura. ***Brazilian Journal of Development***, v. 7, n. 10, p. 99882–99897, 2021.

OLIVEIRA, C. H. M.; SANTOS, F. R. T.; LEITINHO, J. L.; FARIAS, L. G. A. T. Busca dos fatores associados à evasão: um estudo de caso no Campus Universitário da UFC em Crateús. ***Revista Internacional de Educação Superior***, Campinas, v. 5, p. 1–23, 2019. Disponível em: https://periodicos.sbu.unicamp.br/ojs/index.php/riesup/article/view/8652897. Acesso em: 12 jun. 2023.

OLIVEIRA, R. dos S.; MEDEIROS, F. P. A. de. Modelo de predição de evasão escolar com base em dados de autoavaliação de cursos de graduação. ***Revista Brasileira de Informática na Educação***, [S. l.], v. 32, p. 1–21, 2024. DOI: 10.5753/rbie.2024.3542. Disponível em: https://journals-sol.sbc.org.br/index.php/rbie/article/view/3542. Acesso em: 2 fev. 2024.

PEREIRA, A. S.; SHITSUKA, D. M.; PARREIRA, F. J.; SHITSUKA, R. **Metodologia da pesquisa científica**. [free e-book]. Santa Maria/RS: Ed. UAB/NTE/UFSM, 2018. Disponível em: https://repositorio.ufsm.br/bitstream/handle/1/15824/Lic_%20%20Computacao_Metodologia-Pesquisa-Cientifica.pdf?sequence=1. Acesso em: 18 jun. 2023.

PRESTES, Emília Maria da Trindade; FIALHO, Maríllia Gabriella Duarte. Evasão na educação superior e gestão institucional: o caso da Universidade Federal da Paraíba. ***Ensaio: Aval. Pol. Publ. Educ.***, Rio de Janeiro, v. 26, n. 100, p. 869–889, 2018. Disponível em: https://www.scielo.br/j/ensaio/a/3yg5dbpbt6SWdKtpVZ8mNsv/. Acesso em: 12 jun. 2023.

ALBERTONI, R. M.; MARCONDES, R.; DE CARVALHO RUTZ DA SILVA, S.; SZESZ JUNIOR, A. Inteligência Artificial na educação inclusiva: um mapeamento sistemático das aplicações e perspectivas. ***Ensino de Ciências e Tecnologia em Revista – ENCITEC***, v. 14, n. 3, p. 453–463, 16 dez. 2024.

SANTANA JR, O. V. **Auto-organização e aprendizagem por demonstração na determinação de marcha robótica**. 2015. Dissertação (Doutorado em Ciência da Computação) — UFPE.

SANTOS, Patricia; GOYA, Denise. Aprendizado de máquina aplicado à análise de evasão em cursos de sistemas de informação. In: ***Fórum de Educação em Sistemas de Informação - Simpósio Brasileiro de Sistemas de Informação (SBSI)***, 16., 2020, Evento Online. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2020. p. 210–213. DOI: https://doi.org/10.5753/sbsi.2020.13145.

SOUZA, V. F. de; SANTOS, T. C. B. dos. Processo de mineração de dados educacionais aplicado na previsão do desempenho de alunos: uma comparação entre as técnicas de aprendizagem de máquina e aprendizagem profunda. ***Revista Brasileira de Informática na Educação***, [S. l.], v. 29, p. 519–546, 2021. DOI: 10.5753/rbie.2021.29.0.519. Disponível em: https://journals-sol.sbc.org.br/index.php/rbie/article/view/2975. Acesso em: 5 maio 2024.

SUKHIJA, Karan; JINDAL, Manish; AGGARWAL, Naveen. The recent state of educational data mining: a survey and future visions. In: ***3rd International Conference on MOOCs, Innovation and Technology in Education (MITE)***, 2015, Amritsar, Índia. Amritsar, Índia: IEEE, 2015. p. 354–359. Disponível em: https://doi.org/10.1109/MITE.2015.7375344. Acesso em: 16 ago. 2023.

SEMESP. **Mapa do Ensino Superior**, 13ª edição, 2023. Disponível em: https://www.semesp.org.br/mapa/edicao-13/. Acesso em: 22 jun. 2023.

TORRES MARQUES, Leonardo; TORRES MARQUES, Bruno; MORAIS SILVA, Carlos Alexandre. A descoberta das causas da retenção acadêmica utilizando mineração de dados: uma revisão sistemática da literatura. ***RENOTE***, Porto Alegre, v. 20, n. 1, p. 263–272, 2022. DOI: 10.22456/1679-1916.126672. Disponível em: https://seer.ufrgs.br/index.php/renote/article/view/126672. Acesso em: 4 set. 2023.