

Predição de evasão no curso de Sistemas de Informação da UFSM utilizando dados históricos

Leonardo W. Dallanora¹, Sérgio L. S. Mergen²

¹Centro de Tecnologia – Universidade Federal de Santa Maria (UFSM)
97105-900 — Santa Maria — RS — Brasil

²Departamento de Linguagens e Sistemas de Computação
Universidade Federal de Santa Maria (UFSM)

{lwdallanora, mergen}@inf.ufsm.br

Abstract. *Higher education dropout generates significant social and economic impacts. In the Information Systems program at UFSM, historical data reveals that 54.8% of students do not complete their degree, highlighting the need for studies focused on dropout prediction. Through exploratory analysis, key factors were identified to differentiate dropout students from others, such as performance instability, academic disengagement, insufficient achievement, and vulnerability in critical courses to compose the model's features. This study utilizes Educational Data Mining and the Balanced Random Forest algorithm to classify dropout risk. Experiments across multiple time windows demonstrated an accuracy of 84.93% in the long term and 92.32% in the semester immediately preceding dropout, validating the system's effectiveness for early warning systems.*

Resumo. *A evasão no ensino superior gera impactos sociais e econômicos significativos. No curso de Sistemas de Informação da UFSM, dados históricos revelam que 54,8% dos alunos não concluem a graduação, evidenciando a necessidade de trabalhos voltados a predição da evasão. Através de análise exploratória, foram identificadas questões importante para diferenciar alunos evasores dos demais, analisando instabilidade no desempenho, desengajamento acadêmico, insuficiência de notas e vulnerabilidade em disciplinas críticas para compor as features do modelo. Este trabalho utiliza Mineração de Dados Educacionais e o algoritmo Balanced Random Forest para classificar o risco de evasão. Experimentos em múltiplas janelas temporais demonstraram precisão de 84,93% em longo prazo e 92,32% no semestre iminente à evasão, validando a eficácia do sistema para alertas antecipados.*

1. Introdução

A evasão no ensino superior é um fenômeno persistente que impacta a eficiência das instituições brasileiras. Segundo os dados mais recentes apresentados nas Notas Estatísticas do Censo da Educação Superior 2024, divulgado pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira [INEP 2025], o sistema universitário brasileiro atingiu a marca histórica de mais de 10 milhões de matrículas. No entanto, os indicadores revelam um gargalo preocupante no fluxo de formação. No mesmo ano em que mais de 5 milhões de novos alunos ingressaram no ensino superior, apenas 1,3 milhão

concluíram a graduação. Esse descompasso acentuado entre o alto volume de entrada e o baixo volume de saída, conforme define [Favero 2006], indica o rompimento precoce do vínculo do estudante, gerando prejuízos sociais e econômicos significativos.

Esse cenário nacional se reflete de forma ainda mais acentuada na realidade local, utilizada como base para esse estudo. Ao analisarmos os dados históricos do Bacharelado em Sistemas de Informação da UFSM, disponibilizados pela coordenação do curso para realização desse trabalho, cobrindo todas as turmas de 2009 até 2025, o quadro é desafiador: 54,8% dos alunos que ingressaram no curso não chegaram ao final. Para chegar a esse número, foi considerado como 'evasão' qualquer saída que não fosse a formatura, ou seja, alunos que abandonaram, cancelaram a matrícula ou pediram transferência. O fato de mais da metade dos estudantes ficar pelo caminho reforça que é importante entender os padrões por trás dessas saídas para tentar reverter essa situação.

Diante do volume de dados, este trabalho inicia com uma análise exploratória para identificar padrões que distinguem formados de evadidos. Dessa investigação, extraíram-se indicadores de consistência acadêmica, como desvio padrão de notas e reprovações por nota, frequência e em disciplinas difíceis. Esses indicadores serviram como entrada para o algoritmo de classificação *Balanced Random Forest*, visando detectar alunos com potencial para evasão. Conforme destacado nos experimentos, a estratégia de classificação utilizada permitiu identificar potenciais evasores com elevada precisão mesmo em estágios mais iniciais da graduação.

O artigo está organizado em cinco seções: a Seção 2 revisa trabalhos relacionados; a Seção 3 detalha a metodologia e a estratégia de classificação adotada; a Seção 4 discute os resultados obtidos e a Seção 5 apresenta as considerações finais.

2. Trabalhos Relacionados

A aplicação de técnicas de Mineração de Dados Educacionais (Educational data mining, EDM) consolidou-se como fundamental para a gestão acadêmica. [Romero and Ventura 2024] e [Baker and Yacef 2010] destacam que a área evoluiu significativamente, passando de simples análises descritivas e relatórios estatísticos a implementação de “Sistemas de Alerta Antecipado”, projetados para identificar estudantes em risco antes que a evasão se concretize, possibilitando intervenções proativas.

No que se refere a indicadores preditivos, há um consenso crescente de que o desempenho acadêmico apresenta uma relevância igual ou superior à dos dados socioeconômicos na detecção de alunos em risco. O trabalho de [Dekker et al. 2009] já demonstrava a eficácia de usar apenas notas e frequência. Mais recentemente, [Santos et al. 2021] validaram essa hipótese em cursos de computação brasileiros, mostrando que o histórico escolar inicial é suficiente para gerar modelos com alta acurácia. Esse achado é relevante, pois sugere que as instituições podem detectar problemas sem a necessidade de coletar dados externos ao ambiente acadêmico.

Quanto aos algoritmos, existe uma tendência para uma transição de classificadores únicos para métodos de *Ensemble*. Enquanto estudos iniciais focavam em Árvores de Decisão simples, [Hortêncio Filho et al. 2020] compararam diversos classificadores em dados de institutos federais e concluíram que métodos mais robustos que Árvores de Decisão apresentaram melhores resultados. [Salaberri et al. 2024] confirmam essa superioridade em estudo recente, onde o algoritmo *Random Forest* superou redes neurais e

regressão logística na precisão da identificação de evasores. A arquitetura proposta por [Manhães et al. 2014] também evidencia que combinar múltiplos classificadores tende a mitigar vieses individuais dos modelos.

Além do algoritmo, o tratamento do desbalanceamento é crucial. [Colpo et al. 2021] demonstraram que representar os dados de forma acumulada e progressiva ao longo dos semestres eleva significativamente a acurácia. Por fim, [Fernández et al. 2018] e [Agrusti et al. 2019] alertam que, como existe uma classe minoritária, o uso de técnicas de balanceamento (como as aplicadas neste trabalho) é indispensável para evitar que o modelo apenas maximize a acurácia global ignorando os alunos que pertencem a classe minoritária.

Diferenciando-se das abordagens citadas, este trabalho introduz uma análise baseada em múltiplas janelas temporais de antecedência. Enquanto existe um foco frequente na precisão imediata, este trabalho tem a proposta de simular o desempenho do modelo em diferentes estágios da vida acadêmica, partindo de um cenário iminente até previsões com certa antecedência. Essa abordagem permite não apenas validar o algoritmo Balanced Random Forest diante de históricos menos consolidados, mas também oferece um suporte à decisão pedagógica mais eficaz, identificando o risco com tempo para intervenções preventivas. Assim, o trabalho une a eficácia técnica do tratamento de dados desbalanceados à utilidade prática de um sistema de alerta antecipado de longo prazo.

3. Metodologia de Desenvolvimento

Este trabalho visa prever, de forma antecipada, a evasão no curso de Sistemas de Informação. Para isso, realiza-se uma análise exploratória de dados históricos e a extração de indicadores acadêmicos, utilizadas no treino e teste de um algoritmo de aprendizagem de máquina para identificar alunos em risco e validar a acurácia das previsões.

A arquitetura, exibida na Figura 1, tem os seguintes módulos: (a) Pré-processamento dos dados dos alunos; (b) Análise Exploratória de dados, gerando gráficos para visualização e descoberta de novos indicadores; (c) Segregação dos indicadores por períodos temporais; (d) Aprendizado de máquina para prever possíveis alunos que podem evadir.

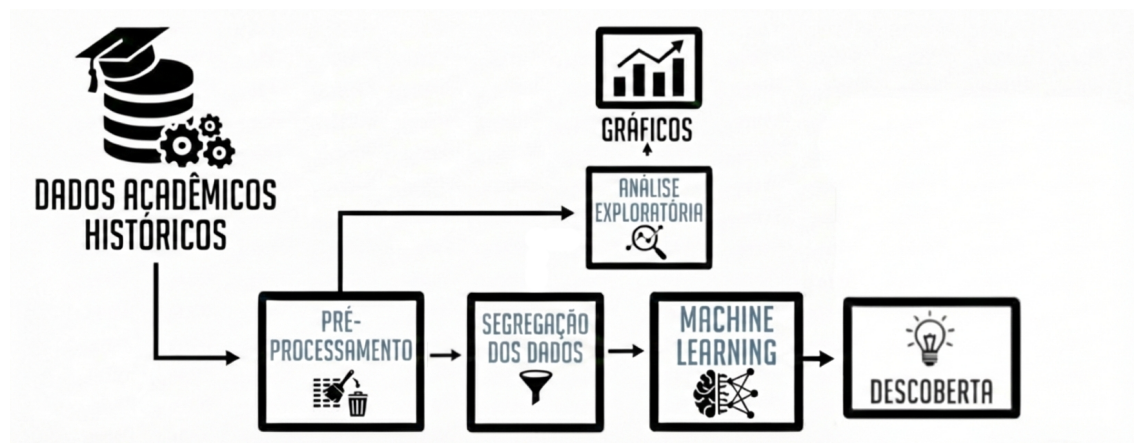


Figura 1. Arquitetura adotada no desenvolvimento do projeto

3.1. Pré-processamento

A etapa de pré-processamento dos dados incluiu a limpeza e a definição do escopo da análise para garantir a consistência do modelo preditivo. Inicialmente, foi realizado um filtro para remover registros de alunos que, embora matriculados, não apresentaram nenhuma atividade acadêmica registrada. Esses casos foram desconsiderados por não fornecerem dados comportamentais ou de desempenho suficientes para a análise de padrões de evasão ao longo do curso.

Em relação à definição da variável alvo, optou-se por uma abordagem de classificação binária. O desfecho “Sucesso” foi atribuído exclusivamente aos alunos com status “Formado”. Todas as outras formas de desligamento definitivo, como abandono, cancelamento de matrícula e transferências internas e externas, foram agrupadas na classe “Evasão”.

Por fim, foram excluídos da amostra os alunos com status ‘Estudante Regular’. Como esses alunos ainda estão cursando a graduação, seus desfechos (conclusão ou evasão) são desconhecidos, o que introduziria ruído e incerteza ao treinamento do modelo supervisionado.

3.2. Análise Exploratória

Após o pré-processamento, a etapa de Análise Exploratória de Dados (AED) buscou identificar padrões comportamentais e de desempenho que distinguem os grupos de alunos “Formados”(Sucesso) e “Evadidos”. A investigação visual e estatística permitiu a descoberta de indicadores relevantes para a caracterização do perfil do estudante, com destaque para a variabilidade de notas e as reprovações.

O primeiro indicador analisado foi a consistência acadêmica, medida através do desvio padrão das notas de cada aluno. A premissa é que a média final, isoladamente, pode mascarar a trajetória do aluno. A hipótese investigada sugere que alunos com desempenho muito oscilante, alternando entre disciplinas com rendimento alto e disciplinas com baixo rendimento, estariam com maior risco de evasão do que aqueles com desempenho constante.

Conforme demonstrado na Figura 2, há uma disparidade clara na distribuição dos dados. O grupo de alunos evadidos apresenta um desvio padrão predominantemente maior em comparação aos alunos formados. Isso indica que o desempenho irregular, alternando entre notas altas e muito baixas é uma característica marcante do grupo de risco, enquanto os formados tendem a manter uma estabilidade acadêmica ao longo do curso.

Além da instabilidade nas notas, as reprovações de diferentes origens foram analisadas. Foi realizada uma análise comparativa considerando a origem da reprovação: por nota (que indica insuficiência de aprendizado ou dificuldade cognitiva com o conteúdo) e por frequência (que indica ausência e quebra de vínculo com a rotina acadêmica).

A Figura 3 revela que, embora as reprovações por nota ocorram em ambos os grupos, a incidência por frequência é drasticamente superior entre os evadidos. Esse padrão sugere que o desengajamento atua como um precursor comportamental da evasão definitiva. Em ambas as categorias, as taxas de reprovação são consistentemente mais elevadas nos alunos que evadiram do que nos formados, consolidando-as como indicadores fundamentais para o modelo preditivo.

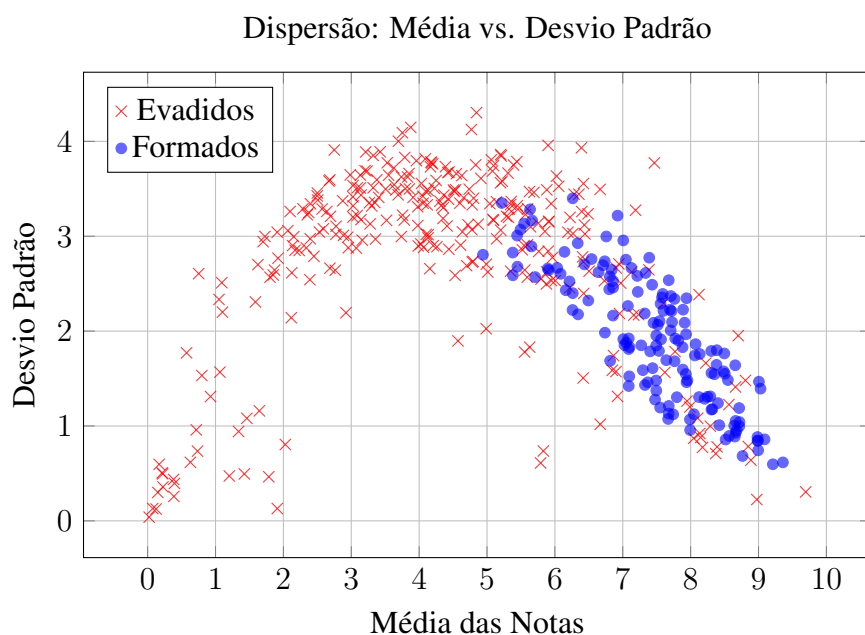


Figura 2. Gráfico de dispersão mostrando a relação entre média e desvio padrão.

Percebe-se também que diferentes formas de evasão (como Abandono e Transferência) apresentam intensidades diferentes desses indicadores, mas invariavelmente superiores às do grupo de formados, mostrando que independente da forma de evasão, as reprovações de ambas as origens são preditores consideráveis para o risco de evasão.

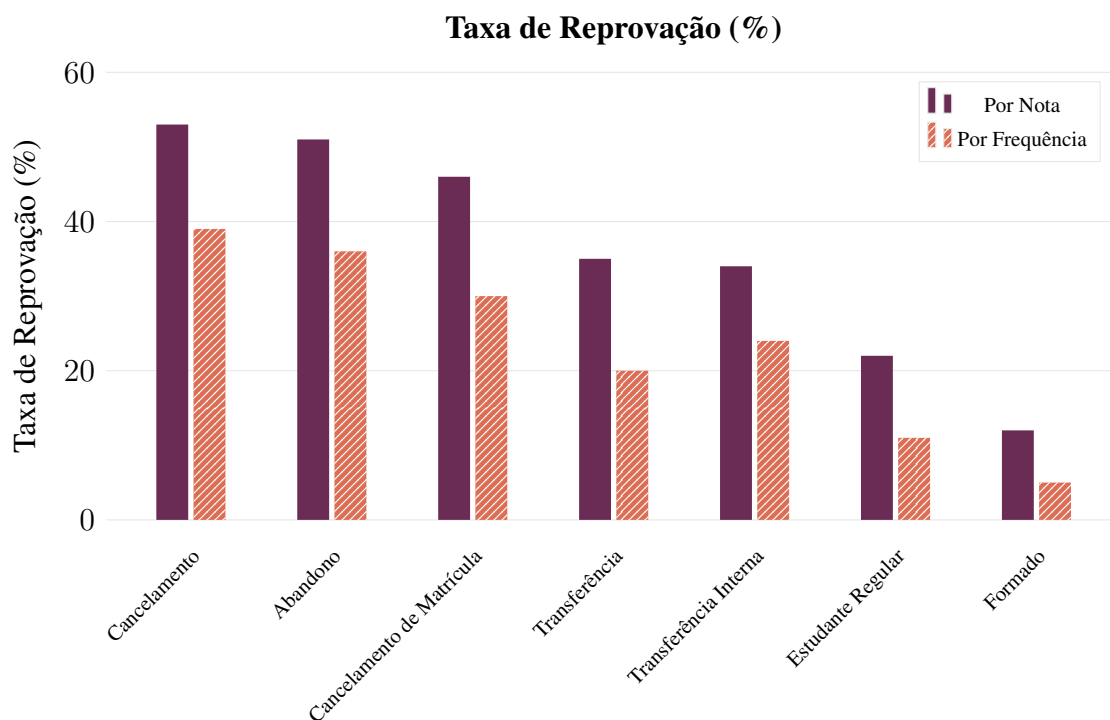


Figura 3. Comparativo das Taxas de Reprovação por Tipo de Evasão.

Para a determinação das taxas de reprovação por nota, adotou-se uma abordagem

focada no desempenho acadêmico efetivo. O cálculo considerou exclusivamente as disciplinas obrigatórias e contabilizou apenas os alunos que concluíram o semestre letivo, ou seja, aqueles que obtiveram como situação final 'Aprovado com nota' ou 'Reprovado com nota'.

Dessa forma, o indicador reflete a dificuldade do conteúdo e os critérios avaliativos, medindo a proporção de insucesso apenas entre os alunos que frequentaram as aulas e se submeteram às avaliações. Embora o desengajamento seja um indicador claro, a análise revela que ele é acompanhado por barreiras técnicas em certas disciplinas que concentram o insucesso acadêmico. Para identificar essas disciplinas no fluxo do curso, a Tabela 1 apresenta as 14 disciplinas que funcionam como os principais obstáculos para a permanência do estudante. A quantidade de reprovações nessas disciplinas também foi levada em consideração como indicador para o modelo, visto que o insucesso precoce nesses componentes pode gerar um efeito cascata no desengajamento do aluno.

Disciplina	Semestre	Reprovação (%)
Organização de Computadores	2	57,92
Cálculo "A"	1	51,79
Circuitos Digitais	1	45,65
Laboratório de Programação I	1	45,34
Matemática Discreta	2	43,40
Estruturas de Dados "A"	2	43,24
Sistemas Operacionais "A"	4	40,15
Pesquisa e Ordenação de Dados "A"	3	39,42
Lógica e Algoritmo	1	37,79
Arquiteturas de Computadores "A"	3	37,69
Laboratório de Programação II	2	36,90
Paradigmas de Programação	3	31,56
Fundamentos de Banco de Dados	4	29,94
Engenharia de Software "A"	3	29,86

Tabela 1. Disciplinas obrigatórias com maiores taxas de reprovação por nota.

A análise das disciplinas listadas mostra que o maior índice de reprovações se concentra nos dois primeiros anos do curso. Disciplinas como Organização de Computadores (57,92%) e Cálculo A (51,79%) atuam como obstáculos críticos que testam a adaptação do ingressante. Para o modelo preditivo, o insucesso nessas matérias iniciais funciona como um sinalizador de risco, pois o acúmulo de reprovações em disciplinas básicas pode gerar um efeito cascata, aumentando o desengajamento e a probabilidade de evasão antes mesmo que o aluno chegue às disciplinas mais avançadas.

3.3. Segregação dos dados

Com o objetivo de avaliar a aplicabilidade do modelo em cenários reais de intervenção pedagógica, os dados foram separados em diferentes janelas temporais de antecedência em relação ao desfecho final do aluno (formatura ou evasão). A metodologia consistiu em restringir o histórico acadêmico disponível para o treinamento, simulando até que ponto o algoritmo conseguiria prever o risco de evasão com determinada antecedência.

A ênfase na janela de um período maior de tempo estabelecida para esse trabalho justifica-se pela necessidade de ações a partir da predição. Identificar um aluno em risco apenas um semestre antes da evasão oferece pouco tempo hábil para que a coordenação do curso ou a instituição implementem medidas de retenção eficazes. Em contrapartida, uma previsão realizada com dois anos de antecedência permite um planejamento pedagógico preventivo.

Além disso, alunos que não possuíam tempo de permanência superior ao corte temporal, foram desconsiderados nesse determinado corte. Por exemplo, um estudante que evadiu no 4º semestre foi excluído da análise de “4 semestres de antecedência”, uma vez que não haveria dados históricos anteriores a esse período para alimentar o modelo. No entanto, esse mesmo aluno foi incluído nos outros cortes temporais, cenários nos quais já existia um histórico de desempenho consolidado passível de análise.

3.4. Aprendizado de máquina

Para a etapa de classificação supervisionada, optou-se pela utilização do algoritmo *Balanced Random Forest Classifier*, uma variação do clássico *Random Forest* adaptada para lidar com conjuntos de dados desbalanceados. A escolha deste algoritmo justifica-se pela natureza dos dados, onde existe uma classe predominante: a de alunos evadidos, que totaliza 298 registros, em contraste com a classe minoritária de formados, composta por 149 registros. Essa desproporção (aproximadamente 2 para 1) poderia induzir classificadores tradicionais a enviesarem as predições a favor da classe majoritária, motivo pelo qual a aplicação de uma estratégia de balanceamento se fez necessária.

O algoritmo *Random Forest* é um método de *ensemble* (conjunto) que constrói múltiplas árvores de decisão durante o treinamento. Para mitigar o problema de balanceamento, o *Balanced Random Forest* introduz uma etapa de subamostragem (*under-sampling*) dinâmica: para cada árvore construída no conjunto, o algoritmo extrai uma amostra que equilibra artificialmente as classes, garantindo que cada árvore de decisão aprenda a distinguir os padrões de evasão com a mesma importância dada aos padrões de sucesso.

Os hiperparâmetros do modelo foram definidos experimentalmente para garantir a generalização e evitar o *overfitting* (sobreajuste). Foram utilizados:

- **Número de Estimadores** ($n_estimators = 300$): Define a criação de 300 árvores de decisão, garantindo robustez estatística e estabilidade nas predições.
- **Profundidade Máxima** ($max_depth = 5$): Limita o crescimento das árvores a 5 níveis. Essa restrição é crucial para evitar que o modelo decore ruídos específicos dos dados de treino, favorecendo a criação de regras de decisão mais genéricas e aplicáveis a novos alunos.
- **Estratégia de Amostragem** ($sampling_strategy = 'all'$): Assegura que todas as classes sejam reamostradas para terem a mesma representatividade durante o treinamento de cada estimador.

Para a validação do modelo, adotou-se a técnica de Validação Cruzada Estratificada (*Stratified K-Fold Cross-Validation*) com $k = 10$. Nesta abordagem, o conjunto de dados é particionado em 10 subconjuntos (*folds*) distintos. O procedimento de avaliação é iterativo e realizado 10 vezes de forma independente. Em cada iteração, um modelo novo é instanciado e treinado utilizando 9 subconjuntos, enquanto o subconjunto restante

é reservado exclusivamente para teste. O modelo é reiniciado a cada ciclo, garantindo que dentro de uma mesma avaliação, um aluno jamais esteja presente simultaneamente nos dados de treino e de teste, eliminando o risco de vazamento de dados. A estratificação assegura que a proporção original de formados e evadidos seja mantida em cada divisão.

4. Resultados Alcançados

A avaliação do desempenho do modelo Balanced Random Forest foi conduzida com foco entre três indicadores: Acurácia, sensibilidade (recall) e precisão (Assertividade). Enquanto a acurácia mede o acerto geral do algoritmo, os outros dois indicadores detalham a qualidade dos alertas sobre os alunos em risco. A sensibilidade indica a capacidade do modelo de perceber os alunos que vão evadir. Um alto sensibilidade significa que o modelo é sensível o suficiente para não deixar quase ninguém passar despercebido, garantindo que o grupo de risco seja identificado. A precisão refere-se à confiabilidade, quando o modelo aponta que um aluno está em risco, a Precisão mede o quanto podemos confiar nesse aviso.

No contexto de evasão, o equilíbrio entre esses dois indicadores é importante. Se houvesse um modelo com alta sensibilidade, mas baixa precisão, ele apontaria quase todos os alunos como possíveis evadidos. Por outro lado, os resultados obtidos demonstram que o modelo proposto não apenas identifica a maioria dos estudantes em risco, mas faz com um elevado índice de confiança (precisão), garantindo que esforços de intervenção sejam direcionados de forma correta.

Para validar o modelo, os dados não foram tratados de maneira estática. Foi realizada uma segregação temporal dos dados, criando quatro janelas de observação distintas baseadas na antecedência em relação ao evento de evasão.

Os recortes definidos foram:

- **4 Semestres de Antecedência (2 Anos):** Cenário de longo prazo, utilizando apenas os dados acumulados até dois anos antes da evasão.
- **3 Semestres de Antecedência (1,5 Ano):** Cenário de médio prazo.
- **2 Semestres de Antecedência (1 Ano):** Cenário de curto prazo.
- **1 Semestre de Antecedência:** Cenário iminente, utilizando todo o histórico até o semestre anterior à saída.

Essa segregação permitiu observar como o comportamento dos indicadores evolui conforme o histórico do aluno se consolida.

A Tabela 2 detalha a evolução dos indicadores. Observa-se que a precisão em relação aos evadidos apresenta os valores mais robustos em todos os cenários.

Tabela 2. Evolução da acurácia, sensibilidade e precisão conforme a janela temporal.

Janela temporal	Acurácia (%)	Sensibilidade (Evasão) (%)	Precisão (%)
4 Semestres	75.94	72.21	84.93
3 Semestres	78.68	77.02	89.98
2 Semestres	79.99	77.62	89.66
1 Semestre	81.90	79.62	92.32

Os dados revelam uma performance interessante da precisão, que inicia em 84,93% no cenário de maior antecedência (4 semestres) e atinge 92,32% no semestre imediatamente anterior à evasão. Esse crescimento consistente indica que, conforme o histórico do aluno se consolida, a assertividade do modelo sobre quem de fato irá evadir torna-se mais precisa.

Mesmo no cenário de dois anos de antecedência, onde a incerteza é naturalmente maior, o índice de 84,93% de precisão é relevante. Isso significa que a grande maioria dos alunos sinalizados como evasores pelo algoritmo realmente apresenta um comportamento de evasão. Essa característica é fundamental para a gestão educacional, pois demonstra confiabilidade na previsão.

Embora o desempenho atinja seu ápice nas janelas de curto prazo, os resultados de longo prazo são promissores para intervenções, mantendo a sensibilidade acima de 72% sem sacrificar a precisão. Esta eficácia precoce deve-se à seleção de indicadores como o desvio padrão, que captura a instabilidade acadêmica de forma mais precisa que a média global, indicando que a evasão é um processo cumulativo e reforçando que intervenções precoces são mais eficientes que ações tardias.

5. Considerações Finais

Este trabalho propôs um modelo preditivo para antecipar a evasão no curso de Sistemas de Informação da UFSM, utilizando a extração de indicadores de desempenho e o algoritmo *Balanced Random Forest*. A investigação confirmou que a evasão não é um evento isolado, mas um processo que deixa "rastros" acadêmicos identificáveis com anos de antecedência. Ficou demonstrado que a instabilidade nas notas (alto desvio padrão) e o desengajamento (reprovações por frequência) são preditores mais robustos do que a simples média aritmética, permitindo uma visão mais profunda da trajetória do estudante.

Os resultados validam a eficácia do modelo como ferramenta de suporte à decisão, evidenciando que não é necessário aguardar o estágio crítico da trajetória do aluno para identificar o risco com segurança. O principal resultado deste estudo é a identificar possíveis evasores com antecedência, alcançando mais de 84% de precisão com dois anos de antecedência oferece à gestão acadêmica uma janela de oportunidade crucial para possíveis intervenções.

Além disso, o baixo impacto de um eventual falso positivo (identificar um aluno como possível evasor que acabe não chegando no momento de quase evasão) justifica uma possível aplicação do modelo. Priorizar a identificação de alunos em risco real, mesmo com uma margem de erro controlada, é fundamental para evitar a perda definitiva do vínculo do estudante com a instituição.

O estudo conclui que o modelo é preciso e confiável, funcionando bem mesmo com poucos dados iniciais e superando as dificuldades de trabalhar com grupos desequilibrados. A análise por janelas de tempo prova que o aprendizado de máquina, quando usa os indicadores certos, transforma o histórico dos alunos em conhecimento. Como perspectiva para trabalhos futuros, propõe-se a aplicação deste modelo em outros cursos da instituição, visando verificar se o desempenho preditivo se mantém consistente em diferentes cursos ou se perfis de alunos distintos exigem novos indicadores.

Referências

- Agrusti, F., Bonavolontà, G., and Mezzini, M. (2019). University dropout prediction through educational data mining techniques: A systematic review. *Journal of e-Learning and Knowledge Society*, 15(3).
- Baker, R. S. and Yacef, K. (2010). Data mining for education. In *International Encyclopedia of Education*, volume 7, pages 112–118. Elsevier, Oxford, UK.
- Colpo, M. P., Primo, T. T., and Aguiar, M. S. (2021). Predição da evasão estudantil: uma análise comparativa de diferentes representações de treino na aprendizagem de modelos genéricos. In *Anais do XXXII Simpósio Brasileiro de Informática na Educação (SBIE)*, Porto Alegre. SBC.
- Dekker, G. W., Pechenizkiy, M., and Vleeshouwers, J. M. (2009). Predicting students drop out: A case study. In *Proceedings of the 2nd International Conference on Educational Data Mining (EDM 2009)*, Cordoba, Espanha.
- Favero, R. V. M. (2006). Dialogar ou evadir: eis a questão. Um estudo sobre a permanência e a evasão na educação a distância. Dissertação (mestrado em educação), Universidade Federal do Rio Grande do Sul, Porto Alegre.
- Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., and Herrera, F. (2018). Smote for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *Journal of Artificial Intelligence Research*, 61.
- Hortêncio Filho, F. W. B., Vinuto, T. S., and Leal, B. C. (2020). Análise de classificadores para predição de evasão dos campi de uma instituição de ensino federal. In *Anais do XXXI Simpósio Brasileiro de Informática na Educação (SBIE)*, Porto Alegre. SBC.
- INEP (2025). Censo da Educação Superior 2024: Notas Estatísticas. Acesso em: 14 dez. 2025.
- Manhães, L. M., da Cruz, S. M., and Zimbrão, G. (2014). Wave: an architecture for predicting dropout in undergraduate courses using edm. In *Proceedings of the 29th Annual ACM Symposium on Applied Computing (SAC)*, Gyeongju, Korea. ACM.
- Romero, C. and Ventura, S. (2024). Educational Data Mining and Learning Analytics: An updated survey. *arXiv preprint arXiv:2402.07956*.
- Salaberri, P., Piovesan, S., and Irala, V. (2024). Mineração de dados educacionais para a predição de evasão: Experiência em uma universidade do sul do brasil. *Administração: Ensino e Pesquisa (RAEP)*, 25(1).
- Santos, C. H. D. C., Martins, S. L., and Plastino, A. (2021). É Possível Prever Evasão com Base Apenas no Desempenho Acadêmico? In *Anais do XXXII Simpósio Brasileiro de Informática na Educação (SBIE)*, Porto Alegre. SBC.