REVISTA TÓPICOS
ISSN 2965-6672 | Qualis A2

PREDIÇÃO PRECOCE DE EVASÃO DISCENTE EM CURSOS DE IDIOMAS UTILIZANDO APRENDIZADO DE MÁQUINA: UM ESTUDO LONGITUDINAL BASEADO EM FREQUÊNCIA E TRAJETÓRIA ACADÊMICA

EARLY STUDENT DROPOUT PREDICTION IN LANGUAGE COURSES USING MACHINE LEARNING: A LONGITUDINAL STUDY BASED ON ATTENDANCE AND ACADEMIC TRAJECTORIES

Ciências Sociais Aplicadas • 01/09/2026

REGISTRO DOI: 10.70773/revistatopicos/788113184

João Cláudio Nunes Carvalho[1]

Daniel Alencar Barros Tavares[2]

Ronaldo Tadeu Pontes Milfont[3]

Daniel Rodrigues Monyu[4]

Rafael Bianchini Glavam[5]

RESUMO

A evasão estudantil desafia as organizações educacionais porque, quando identificada apenas após o cancelamento, deixa pouca margem para intervenção preventiva. Este estudo apresenta uma abordagem longitudinal de aprendizado de máquina para a predição precoce do risco de evasão em cursos de idiomas, com dados de frequência e trajetória acadêmica da plataforma CCLI. Foram construídos 11.031 snapshots temporais a partir de 1.253 matrículas elegíveis, das quais 343 foram classificadas como evasão, entre janeiro de 2025 e maio de 2026. O desfecho foi definido a partir dos motivos de cancelamento, com exclusão de conclusão de contrato e atingimento do resultado esperado. Regressão Logística, Random Forest, XGBoost e LightGBM foram avaliados em horizontes de antecipação de 30, 60 e 90 dias, com treinamento no passado e teste em período posterior (validação temporal) e com uso exclusivo de variáveis disponíveis no momento da previsão, por isso o LTV (meses de vínculo do estudante com a instituição) e o tempo de matrícula (duração do vínculo específico), definidos somente no encerramento, foram excluídos para evitar vazamento de informação. No horizonte de 60 dias, a Regressão Logística baseada apenas na frequência corrente e na média histórica obteve o maior PR-AUC (0,337), superando Random Forest (0,258), XGBoost (0,199) e LightGBM (0,190), treinados com conjuntos mais amplos de atributos. Na priorização operacional, os 10% de maior risco segundo a Regressão Logística concentraram 31,9% das evasões futuras, com precisão de 45,8%, quase metade dos alertas se confirma, e os demais permanecem úteis para contato preventivo.

Palavras-chave: Aprendizado de Máquina; Evasão Estudantil; Mineração de Dados Educacionais; Learning Analytics; Predição de Risco.

ABSTRACT

Student dropout challenges educational organizations because, when it is identified only after cancellation, little room remains for preventive intervention. This study presents a longitudinal machine-learning approach for early dropout-risk prediction in language courses, using attendance and academic-trajectory data from the CCLI platform. A total of 11,031 temporal snapshots were constructed from 1,253 eligible enrollments, of which 343 were classified as dropout events, between January 2025 and May 2026. The outcome was defined from cancellation reasons, excluding contract completion and achievement of the expected learning outcome. Logistic Regression, Random Forest, XGBoost, and LightGBM were evaluated over 30-, 60-, and 90-day prediction horizons, with training on past data and testing on later periods (temporal validation) and using only variables available at prediction time; for this reason, LTV (months of the student's relationship with the institution) and enrollment tenure (duration of the specific enrollment), which are defined only at closure, were excluded to prevent information leakage. At the 60-day horizon, a Logistic Regression based only on current attendance and the historical mean achieved the highest PR-AUC (0.337), outperforming Random Forest (0.258), XGBoost (0.199), and LightGBM (0.190), which were trained with broader feature sets. In operational prioritization, the top 10% highest-risk enrollments according to the Logistic Regression captured 31.9% of future dropouts with 45.8% precision, nearly half of the alerts are confirmed, and the remainder remain useful for preventive contact.

Keywords: Machine Learning; Student Dropout; Educational Data Mining; Learning Analytics; Risk Prediction.

1. INTRODUÇÃO

A digitalização dos processos educacionais ampliou a capacidade de registrar, organizar e analisar informações relacionadas ao comportamento e à trajetória dos estudantes. Frequência, matrícula, modalidade de ensino, nível, turma, interações e histórico de permanência constituem exemplos de dados que, quando analisados de forma longitudinal, podem contribuir para a identificação de padrões associados ao desengajamento. Nesse cenário, técnicas de Mineração de Dados Educacionais, Learning Analytics e Aprendizado de Máquina passaram a ocupar espaço crescente em sistemas de apoio à decisão voltados à permanência estudantil.

A evasão representa um problema particularmente relevante porque seu reconhecimento tardio reduz a possibilidade de intervenção. Quando o abandono é percebido apenas após a formalização de um cancelamento, a instituição perde a oportunidade de utilizar sinais anteriores para estabelecer contato, compreender dificuldades e oferecer suporte. Em cursos de idiomas, nos quais a continuidade e a regularidade da participação são importantes para a progressão, mudanças no padrão de frequência podem constituir indicadores úteis de redução de engajamento.

Uma análise baseada somente na frequência observada em um único período, entretanto, pode ser insuficiente. Dois estudantes podem apresentar a mesma frequência atual e, ainda assim, possuir trajetórias distintas. Um deles pode manter desempenho estável, enquanto outro pode apresentar queda sucessiva ao longo dos meses. A incorporação da dimensão temporal permite representar esse comportamento por variáveis como média móvel, variação recente, inclinação da tendência, volatilidade e persistência de períodos de baixa frequência.

O CCLI é uma plataforma educacional que mantém informações relacionadas a matrículas, turmas, estudantes, frequência, consultores e outras dimensões operacionais. Esses registros permitem reconstruir a trajetória de frequência ao longo do tempo e examinar se padrões anteriores ao cancelamento estão associados a risco futuro de evasão. Entre os campos cadastrais também figura o LTV (Lifetime Value, ou valor do tempo de vida do cliente): no contexto do CCLI, trata-se de uma medida operacional da permanência do estudante, em geral expressa em meses de vínculo desde o início da jornada até o momento do cálculo ou do encerramento. Embora o LTV seja útil para gestão comercial e retenção, ele não deve alimentar um modelo de predição precoce quando só fica bem definido perto do cancelamento ou resume o tempo até a saída: usá-lo assim seria equivalente a entregar ao algoritmo parte da resposta que se deseja prever. A existência de um histórico longitudinal de presença torna a plataforma um contexto adequado para estudos de predição precoce em cursos de idiomas, desde que se privilegiem variáveis disponíveis no instante da previsão como a frequência e se excluam atributos pós-evento.

O problema científico deste estudo consiste, portanto, em determinar se informações históricas disponíveis antes da ocorrência de um evento de evasão possuem capacidade preditiva suficiente para permitir sua identificação antecipada. A formulação adotada não trata a evasão como uma classificação retrospectiva de estudantes que já abandonaram o curso, mas como um problema de previsão temporal no qual somente informações disponíveis até determinado instante podem ser utilizadas para estimar a ocorrência futura do evento.

Diante desse cenário, estabelece-se a seguinte questão de pesquisa: em que medida modelos supervisionados de aprendizado de máquina conseguem predizer antecipadamente a evasão de estudantes em cursos de idiomas utilizando informações longitudinais de frequência e trajetória acadêmica disponíveis no CCLI? Como questões complementares, investiga-se como o desempenho varia entre Regressão Logística, Random Forest, XGBoost e LightGBM; como o desempenho muda em horizontes de 30, 60 e 90 dias; e quais grupos de atributos da trajetória de frequência adicionam informação à frequência observada isoladamente.

O objetivo geral consiste em criar e avaliar modelos supervisionados de aprendizado de máquina para a predição precoce do risco de evasão a partir de informações longitudinais de frequência e trajetória acadêmica. Especificamente, busca-se: definir operacionalmente a evasão; construir uma base temporal por matrícula; elaborar atributos representativos da dinâmica da frequência; comparar algoritmos supervisionados sob protocolo antivazamento; avaliar diferentes horizontes de antecedência; analisar calibração e desempenho operacional de priorização; e investigar padrões de falsos positivos e falsos negativos.

As principais contribuições são: a modelagem longitudinal do risco de evasão; a comparação entre algoritmos supervisionados com validação temporal; a análise de múltiplos horizontes de predição; o estudo de ablação dos grupos de atributos; e a evidência empírica de que uma Regressão Logística com frequência do mês e média histórica supera ensembles com conjuntos de atributos mais amplos no horizonte de 60 dias (PR-AUC de 0,337), enquanto a complexidade adicional degradou o desempenho sob mudança de

prevalência entre treino e teste. O restante do artigo apresenta a fundamentação teórica, a metodologia, os resultados e a discussão, e as considerações finais.

2. FUNDAMENTAÇÃO TEÓRICA

2.1. Evasão Discente e Permanência em Ambientes Educacionais

A evasão discente pode ser compreendida como a interrupção da trajetória educacional antes do encerramento esperado de determinado percurso formativo. Sua definição operacional, contudo, depende do contexto institucional. Cancelamentos administrativos, transferências, substituições de matrícula e conclusões não devem ser automaticamente agrupados como abandono. Para fins preditivos, a qualidade da variável-alvo depende diretamente da capacidade de distinguir eventos que representam efetiva descontinuidade daqueles relacionados a reorganizações administrativas. Tinto (1975) já destacava a evasão como fenômeno multidimensional, o que reforça a necessidade de uma definição operacional explícita antes da modelagem.

Essa distinção é especialmente importante em estudos observacionais. Um modelo treinado com rótulos inconsistentes pode aprender padrões relacionados a procedimentos internos da instituição em vez de sinais comportamentais de desengajamento. Por essa razão, a construção do alvo neste estudo parte dos motivos de cancelamento e da situação de cada matrícula, estabelecendo critérios explícitos de inclusão e exclusão.

2.2. Mineração de Dados Educacionais e Learning Analytics

A Mineração de Dados Educacionais investiga métodos computacionais para identificar padrões em dados produzidos em contextos de aprendizagem. Learning Analytics, por sua vez, enfatiza a mensuração, coleta, análise e comunicação de dados sobre estudantes e seus contextos com o objetivo de compreender e otimizar processos educacionais. Embora possuam origens e ênfases distintas, ambas as áreas fornecem fundamentos para sistemas capazes de transformar registros operacionais em informações de apoio à decisão. Revisões contemporâneas consolidam esse campo (Romero; Ventura, 2020; Siemens; Baker, 2012).

No problema de permanência, essas abordagens permitem combinar variáveis acadêmicas, comportamentais e temporais. Entretanto, a utilidade prática de um sistema preditivo depende não apenas de sua capacidade discriminativa, mas também da disponibilidade dos dados no momento da decisão, da estabilidade do modelo ao longo do tempo e da tradução de probabilidades em listas de prioridade que possam ser utilizadas por equipes de acompanhamento.

2.3. Aprendizado de Máquina para Predição de Evasão

A predição de evasão pode ser formulada como um problema de classificação binária no qual cada observação recebe uma estimativa de probabilidade de ocorrência futura do evento. Modelos lineares fornecem uma referência interpretável para o efeito conjunto das variáveis, enquanto métodos baseados em árvores conseguem representar relações não lineares e interações sem exigir especificação manual de todas as combinações possíveis. Em dados

tabulares, boosting tem sido competitivo na predição de evasão (Chen; Guestrin, 2016; Ke et al., 2017; Goren; Cohen; Rubinstein, 2024).

Em bases tabulares, Regressão Logística, Random Forest e métodos de gradient boosting constituem alternativas frequentemente empregadas. A comparação entre algoritmos, entretanto, não deve ser tratada como objetivo isolado. O interesse científico deste trabalho está em verificar se a informação temporal disponível antes da evasão aumenta a capacidade de predição e se conjuntos mais amplos de atributos acrescentam ganho efetivo em relação a uma especificação parcimoniosa baseada na frequência. Estudos com janela semestral ou variação temporal ao longo da matrícula (Viana; Santana; Rabêlo, 2022; Glandorf et al., 2024) ilustram a importância do protocolo de validação.

2.4. Modelagem Temporal e Predição Precoce

O caráter temporal do problema exige que as variáveis utilizadas em uma previsão sejam estritamente anteriores ao desfecho. Se uma informação registrada após o cancelamento for incorporada ao conjunto de atributos, o modelo poderá apresentar desempenho artificialmente elevado sem possuir utilidade prospectiva. Esse fenômeno, conhecido como vazamento de informação, é uma das principais ameaças à validade de sistemas de predição aplicados a eventos futuros. Glandorf et al. (2024) mostram que o desempenho e a importância das variáveis mudam conforme o instante da predição, o que justifica snapshots e cortes temporais.

Para representar a evolução do estudante, o estudo adota snapshots temporais. Cada snapshot corresponde ao estado de uma matrícula em um instante t e contém apenas informações disponíveis até

aquele momento. A variável-alvo indica se um evento de evasão ocorrerá em uma janela futura de 30, 60 ou 90 dias. Essa formulação aproxima a avaliação experimental da situação operacional na qual o modelo deverá ser utilizado.

2.5. Lacuna de Pesquisa e Posicionamento do Presente Estudo

Uma contribuição relevante em estudos de predição aplicada consiste em ultrapassar a comparação estática entre algoritmos. O presente trabalho é estruturado para avaliar simultaneamente quatro dimensões: a utilização de dados longitudinais; a antecedência da previsão; a comparação entre modelos supervisionados sob restrições anti-vazamento; e a utilidade da lista de risco para uma equipe com capacidade limitada de intervenção.

A Tabela 1 posiciona o presente estudo em relação a três trabalhos recentes e verificados sobre predição de evasão com ênfase temporal. Goren, Cohen e Rubinstein (2024) avaliam XGBoost e redes neurais em ensino superior com dados administrativos e de AVA. Glandorf et al. (2024) analisam a variação temporal da predição ao longo da matrícula universitária. Viana, Santana e Rabêlo (2022) treinam classificadores por janela semestral em uma IFES brasileira. Nenhum desses estudos combina cursos de idiomas, série mensal de frequência e comparação sistemática entre Regressão Logística parcimoniosa e ensembles sob protocolo anti-vazamento com validação temporal, lacuna que este artigo aborda.

Tabela 1. Comparação dos trabalhos relacionados

Estudo	Contexto	Horizonte	Métodos	Validação temporal	Limita

Goren, Cohen e Rubinstein (2024)	Ensino superior (Israel); dados administrativos e AVA	4 e 8 semanas do semestre	XGBoost e redes neurais	Sim (pontos no semestre)	Context univers o; po ganho compo mento AVA

⚠ Esta tabela possui muitas colunas e foi cortada para impressão. Para visualizá-la completa, acesse o artigo original em: https://revistatopicos.com.br/artigos/predicao-precoce-de-evasao-discente-em-cursos-de-idiomas-utilizando-aprendizado-de-maquina-um-estudo-longitudinal-baseado-em-frequencia-e-trajetoria-academica?noblockage

Fonte: elaboração própria.

3. METODOLOGIA

3.1. Caracterização da Pesquisa

A pesquisa é de natureza aplicada e quantitativa, com delineamento observacional longitudinal e abordagem experimental computacional. O objetivo não é estabelecer relações causais entre frequência e evasão, mas avaliar se padrões registrados antes do evento estão associados a risco futuro e podem ser utilizados para produzir predições com antecedência operacionalmente útil. O estudo foi executado sobre registros históricos anonimizados da plataforma CCLI.

3.2. Fonte de Dados e Unidade de Análise

A base foi construída a partir de registros operacionais do CCLI e do histórico de frequência mantido pelo módulo de Sucesso do Aluno. A unidade analítica principal foi a matrícula, e não simplesmente a

pessoa, porque um mesmo estudante pode estar associado a mais de uma turma ou matrícula. Cada observação do conjunto analítico representa o estado de uma matrícula em determinado instante.

A extração utilizou o dump operacional anonimizado de 28 de junho de 2026 e o histórico do módulo Sucesso do Aluno. Após inclusão e exclusão, permaneceram 1.253 matrículas (910 ativas e 343 evasões), 11.031 snapshots e 15.260 registros de frequência entre janeiro de 2025 e maio de 2026. Foram excluídos 788 registros sem status cadastral, 18 matrículas a iniciar, 12 conclusões de contrato/resultado e 5 casos de trancamento ou não início por problema de venda. Identificadores pessoais (nome, CPF, telefone, e-mail e endereço) não integraram a base analítica.

3.3. Definição Operacional da Evasão

Cancelamento não foi tratado automaticamente como sinônimo de evasão. Foram classificados como evasão os status Cancelado e Evasão pedagógica, incluindo motivos como tempo, financeiro, desligamento, saúde, cancelamento corporativo, metodologia, desmotivação, não renovação, mudança, pessoal, meta não atingida e evasão. Foram excluídos da classe positiva Contrato concluído e Atingiu o resultado, além de trancamento temporário e não início associado a problema de venda. Essa taxonomia foi alinhada aos motivos cadastrados na planilha de Gestão Geral e à tabela enrollment_cancellation_reasons do CCLI.

$Y_i = 1$, se a matrícula i apresentar evento definido como evasão; $Y_i = 0$, caso contrário.

Essa definição foi aplicada de maneira uniforme a toda a base antes do treinamento dos modelos, reduzindo o erro de rotulagem e

aproximando a variável-alvo do fenômeno que se pretendia antecipar.

3.4. Construção do Dataset Temporal

Para cada matrícula i e tempo t foi construído um vetor $X_{i,t}$ contendo somente atributos disponíveis até t. A variável-alvo indicou a ocorrência de evasão dentro de uma janela futura. Foram avaliados três horizontes de predição: 30, 60 e 90 dias. Consequentemente, o mesmo snapshot pôde ser positivo para uma janela e negativo para outra, dependendo da distância temporal até o evento.

$$Y_{i,t}^{(h)} = 1, \text{ se ocorrer evasão entre } t \text{ e } t+h; 0, \text{ caso contrário, com } h \in \{30, 60, 90 \text{ dias}\}.$$

3.5. Engenharia de Atributos

Os atributos foram organizados em quatro grupos: frequência corrente, histórico de frequência, características derivadas da trajetória e informações acadêmicas/operacionais (idioma, nível, modalidade e canal). Foram excluídos o LTV (meses de permanência / lifetime value) e o tempo de matrícula até o encerramento, porque ambos se correlacionam com o fato de a matrícula ter sido (ou estar prestes a ser) concluída ou cancelada e, se usados, inflariam as métricas de forma artificial. A configuração principal da Regressão Logística utilizou apenas frequência corrente (F_t) e média histórica até t, com StandardScaler. Random Forest, XGBoost e LightGBM foram treinados com o conjunto ampliado de engenharia temporal e perfil acadêmico-operacional (configuração D da ablação).

Tabela 2. Conjunto inicial de variáveis candidatas

Grupo	Variável	Descrição
Frequência	F_t	Percentual de frequência no período corrente
Histórico	F_{t-1}, F_{t-2}, F_{t-3}	Frequências dos períodos anteriores
Temporal	MA_3	Média móvel dos três períodos mais recentes
Temporal	ΔF	Variação entre frequência atual e anterior
Temporal	Slope	Inclinação da trajetória recente
Temporal	σF	Volatilidade da frequência
Temporal	F_{min}	Menor frequência em janela recente
Permanência	Tempo de matrícula	Excluído (risco de vazamento / correlacionado ao encerramento)
Acadêmico	Idioma / nível / modalidade	Características da matrícula e do curso
Operacional	Consultor / empresa / canal	Variáveis disponíveis e temporalmente válidas

Fonte: elaboração própria.

A média móvel de três períodos foi calculada pela média aritmética das frequências disponíveis na janela. A variação recente foi dada pela diferença entre a frequência atual e a imediatamente anterior. A tendência foi estimada por regressão linear simples sobre os períodos recentes, utilizando o coeficiente angular como indicador de crescimento ou redução. A volatilidade foi representada pelo desvio-padrão das frequências na janela observada.

$$MA_3 = (F_t + F_{t-1} + F_{t-2}) / 3$$

$$\Delta F_t = F_t - F_{t-1}$$

3.6. Estudo de Ablação das Características

Além da comparação entre algoritmos, foi realizado um estudo de ablação, no qual o mesmo algoritmo é treinado com conjuntos progressivos de atributos para identificar quais informações efetivamente melhoram a predição. O Modelo A utiliza apenas a frequência do mês corrente; o Modelo B incorpora as frequências dos meses anteriores; o Modelo C acrescenta atributos derivados (média móvel, variação, inclinação e volatilidade); e o Modelo D adiciona idioma, nível, modalidade e canal, sempre sem LTV e sem tempo de matrícula.

Tabela 3. Configuração do estudo de ablação

Configuração	Atributos
A – Snapshot	Frequência atual
B – Histórico bruto	Frequência atual + frequências anteriores
C – Engenharia temporal	B + média móvel + variação + slope + volatilidade + mínimo recente
D – Perfil acadêmico-operacional	C + idioma, nível, modalidade e canal (sem LTV nem tempo de matrícula)

Fonte: elaboração própria.

3.7. Modelos Investigados

A Regressão Logística foi adotada como modelo linear interpretável, com ênfase na configuração mínima, composta por frequência corrente e média histórica. Essa escolha permite contrastar um

classificador simples, alinhado ao sinal dominante da frequência, com ensembles não lineares. Como referência mínima de desempenho, utilizou-se a prevalência de evasões no conjunto de teste (14,4% no horizonte de 60 dias), valor próximo do PR-AUC esperado para um ranqueamento sem capacidade discriminativa; resultados acima dessa referência indicam ordenação melhor do que o acaso.

Foram avaliados Random Forest, XGBoost e LightGBM como modelos não lineares para dados tabulares, com o conjunto de atributos da configuração D. Uma rede neural multicamada não foi incluída porque o ganho esperado não justificava a complexidade adicional. Os hiperparâmetros seguiram configurações previamente delimitadas, com seed 42, class_weight balanceado (ou scale_pos_weight, no caso do XGBoost) e limiar de decisão escolhido pela maximização do F1-score no conjunto de validação.

3.8. Divisão Temporal e Protocolo Experimental

A avaliação principal utilizou separação temporal. O treinamento concentrou snapshots até dezembro de 2025 (n = 6.560; 2,3% positivos no horizonte de 60 dias), a validação correspondeu a janeiro de 2026 (n = 868; 14,4% positivos) e o teste a fevereiro de 2026 em diante (n = 1.912; 14,4% positivos). O conjunto de teste permaneceu isolado até a escolha do limiar na validação.

O protocolo reproduz a situação de implantação: dados do passado estimam o comportamento posterior. O corte foi definido após inspecionar a concentração de cancelamentos no primeiro semestre de 2026, de modo a preservar eventos positivos na validação e no teste. Snapshots cuja janela futura ultrapassava 28/06/2026 sem

evento observado foram censurados e não entraram na métrica daquele horizonte.

3.9. Prevenção de Vazamento de Informação

Todas as transformações aprendidas a partir dos dados foram ajustadas exclusivamente no conjunto de treinamento. Padronização, imputação, seleção de atributos, definição de pesos e escolha do limiar não utilizaram informações do conjunto de teste. Além disso, variáveis cuja disponibilidade ocorresse após o instante representado pelo snapshot foram descartadas. Com essa estratégia, buscou-se impedir que os modelos aprendessem com sinais do futuro ou com informações indiretamente associadas ao encerramento da matrícula.

3.10. Tratamento do Desbalanceamento e Configuração dos Modelos

O desbalanceamento foi tratado com pesos de classe e ajuste do limiar de decisão no conjunto de validação, sem geração sintética de amostras (SMOTE), a fim de preservar a ordem temporal. A busca de hiperparâmetros utilizou valores previamente delimitados, Random Forest com 250 árvores e profundidade máxima 8; métodos de boosting com 250 iterações e learning rate de 0,06 e seed 42 em todos os procedimentos estocásticos.

3.11. Métricas de Avaliação

A métrica principal foi a área sob a curva precisão-revocação (PR-AUC), que resume a capacidade do modelo de ordenar as matrículas colocando os casos que evadem mais próximos do topo da lista de risco. Em bases desbalanceadas aqui, poucos evasores em relação

aos que permanecem, essa medida é mais informativa do que a acurácia simples, que poderia parecer elevada apenas por prever a ausência de evasão na maioria dos casos. Também foram apresentadas ROC-AUC, precisão (proporção de alertas que correspondem a evasões), revocação (fração das evasões efetivamente alertada), F1-score e Brier Score. A revocação possui relevância prática porque um falso negativo corresponde a uma matrícula que evade sem ter sido sinalizada como prioritária.

$$Precision = TP / (TP + FP)$$

$$Recall = TP / (TP + FN)$$

$$F1 = 2 \times (Precision \times Recall) / (Precision + Recall)$$

3.12. Calibração e Avaliação Operacional

A calibração das estimativas foi avaliada por Brier Score. A capacidade operacional de priorização foi medida por Precision@K e Recall@K nos percentuais de 5%, 10% e 20% das matrículas de maior risco, utilizando o score da Regressão Logística mínima no horizonte de 60 dias.

3.13. Análise Estatística, Explicabilidade e Erros

A interpretabilidade foi tratada de forma global, utilizando coeficientes da Regressão Logística e importância por permutação nos modelos de árvore. A incerteza das métricas da especificação principal foi quantificada por bootstrap com reamostragem por matrícula (2.000 réplicas, seed 42), reportando-se intervalos de confiança de 95% obtidos pelos percentis 2,5 e 97,5 da distribuição bootstrap.

A análise de erros separou falsos positivos e falsos negativos no conjunto de teste. Nos falsos negativos, foram examinadas situações com pouco histórico mensal, mudanças abruptas imediatamente anteriores ao cancelamento e ausência de queda prolongada de frequência; nos falsos positivos, verificou-se em que medida frequências baixas ou em declínio geravam alertas para estudantes que permaneciam ativos. Os resultados dessa análise são apresentados na seção 4.6.

3.14. Privacidade, Ética e Reprodutibilidade

Antes da modelagem, dados pessoais diretamente identificáveis foram removidos ou substituídos por identificadores técnicos. Nome, CPF, telefone, e-mail e endereço não integraram a base analítica. O tratamento observou a legislação brasileira de proteção de dados. Resultados foram apresentados exclusivamente de forma agregada.

Os experimentos utilizaram Python 3.14, pandas 3.0, scikit-learn 1.9, XGBoost 3.4 e LightGBM 4.7, com seed 42 em todos os procedimentos estocásticos. Como a base original contém dados pessoais, ela não é disponibilizada publicamente; para assegurar rastreabilidade, foram preservados o pipeline anonimizado, o dicionário de variáveis e os artefatos de avaliação em repositório interno.

4. RESULTADOS E DISCUSSÕES

Os resultados a seguir referem-se ao conjunto de teste temporal, salvo indicação em contrário, e provêm da execução do protocolo experimental descrito na seção anterior.

4.1. Caracterização da Base Analítica

Após os critérios de inclusão e exclusão, a base analítica cobriu janeiro de 2025 a maio de 2026. Havia 1.253 matrículas elegíveis e 343 eventos de evasão (27,4%). Foram gerados 11.031 snapshots mensais e utilizados 15.260 registros de frequência. A maior parte dos cancelamentos concentra-se no primeiro semestre de 2026, o que motivou o corte temporal adotado. No teste de 60 dias, a prevalência do evento na janela futura foi 14,4%.

Tabela 4. Caracterização da base analítica

Indicador	Valor
Período analisado	jan./2025 a maio/2026
Matrículas elegíveis	**1.253**
Estudantes únicos	1.253
Snapshots temporais	**11.031**
Eventos classificados como evasão	343
Proporção da classe positiva	**27,4% das matrículas (14,4% dos snapshots de teste em 60 dias)**
Registros de frequência utilizados	15.260

Fonte: elaboração própria.

4.2. Desempenho Comparativo dos Modelos

No horizonte de 60 dias, a Regressão Logística com frequência do mês e média histórica obteve o melhor PR-AUC (0,337; IC 95% por bootstrap: 0,280 a 0,406), com revocação de 0,355 e F1-score de 0,394, desempenho bem acima da referência de um ranqueamento

sem capacidade discriminativa (aproximadamente 14,4%, a prevalência de positivos no teste). Random Forest (0,258), XGBoost (0,199) e LightGBM (0,190), treinados com conjuntos mais amplos de atributos (trajetória temporal e perfil acadêmico-operacional), ficaram abaixo da configuração mínima. O XGBoost apresentou revocação relativamente alta (0,315) com precisão baixa (0,200), ou seja, gerou muitos alertas que não se confirmaram. O padrão sugere que, sob o protocolo adotado e diante da mudança na proporção de evasões entre 2025 (treino) e 2026 (teste), o sinal útil concentra-se na frequência recente, e o acréscimo de variáveis favoreceu o sobreajuste ao cenário antigo em detrimento da generalização.

Tabela 5. Desempenho dos modelos no conjunto de teste (horizonte de 60 dias)

Modelo	PR-AUC	ROC-AUC	Precision	Recall	F1
Regressão Logística (F_1 + média)	0,337	0,694	0,441	0,355	0,394
Random Forest	**0,258**	**0,631**	**0,344**	**0,163**	**0,221**
XGBoost	0,199	0,587	0,200	0,315	0,245
LightGBM	**0,190**	**0,558**	**0,187**	**0,272**	**0,221**

⚠ Esta tabela possui muitas colunas e foi cortada para impressão. Para visualizá-la completa, acesse o artigo original em: https://revistatopicos.com.br/artigos/predicao-precoce-de-evasao-discente-em-cursos-de-idiomas-utilizando-aprendizado-de-maquina-um-estudo-longitudinal-baseado-em-frequencia-e-trajetoria-academica?noblockage

Fonte: elaboração própria.

4.3. Efeito do Horizonte de Predição

A predição a 30 dias permaneceu difícil para todos os modelos (PR-AUC de 0,204 para a Regressão Logística mínima; IC 95%: 0,161 a 0,269). Em 60 dias, a Regressão Logística mínima alcançou 0,337 (IC 95%: 0,280 a 0,406), superando Random Forest (0,258), XGBoost (0,199) e LightGBM (0,190). Em 90 dias, manteve a liderança com 0,541 (IC 95%: 0,480 a 0,605), seguida do Random Forest (0,458). O crescimento do PR-AUC com a ampliação da janela é consistente, os intervalos de 30 e 90 dias não se sobrepõem, e reflete a maior facilidade de capturar a deterioração acumulada da frequência; ainda assim, o modelo linear mínimo permaneceu competitivo ou superior aos ensembles em todas as janelas avaliadas.

Tabela 6. PR-AUC por horizonte de predição

Modelo	30 dias	60 dias	90 dias
Regressão Logística (F_t + média)	0,204	0,337	0,541
Random Forest	**0,126**	**0,258**	**0,458**
XGBoost	0,091	0,199	0,393
LightGBM	**0,086**	**0,190**	**0,386**

Fonte: elaboração própria.

4.4. Contribuição das Características Longitudinais

No estudo de ablação com LightGBM, no horizonte de 60 dias, o melhor resultado veio do Modelo A, que utiliza apenas a frequência do mês corrente (PR-AUC de 0,294). A inclusão de frequências passadas (Modelo B: 0,177), de indicadores derivados (Modelo C:

0,180) e de idioma, nível, modalidade e canal (Modelo D: 0,190) reduziu o desempenho. Com os dados disponíveis, portanto, mais atributos não significaram melhor predição. O Modelo B alcançou revocação elevada (0,609), porém com F1-score de 0,240, o que evidencia excesso de falsos positivos. Esse resultado reforça que o sinal mais estável está na frequência recente, o mesmo tipo de informação com que a Regressão Logística mínima obteve o melhor desempenho na comparação entre algoritmos.

Tabela 7. Estudo de ablação dos grupos de atributos

Configuração	PR-AUC	Recall	F1
A – Snapshot	0,294	0,156	0,245
B – Histórico bruto	**0,177**	**0,609**	**0,240**
C – Engenharia temporal	0,180	0,275	0,201
D – Perfil acadêmico-operacional	**0,190**	**0,272**	**0,221**

Fonte: elaboração própria.

4.5. Calibração e Priorização Operacional

A Regressão Logística mínima apresentou Brier Score de 0,205 no teste; XGBoost (0,148), LightGBM (0,149) e Random Forest (0,172) obtiveram valores menores, o que indica escores globalmente mais próximos dos desfechos observados, ainda que com pior ordenação dos casos de risco. Na priorização operacional da Regressão Logística mínima, os 5% de maior escore tiveram precisão de 56,2% e recall de 19,6%; os 10% recuperaram 31,9% das evasões (IC 95%: 26,8% a 36,9%), com precisão de 45,8%; e os 20% recuperaram 47,1%, com precisão de 33,9%. A lista de prioridade oferece cobertura

operacional moderada, coerente com o desempenho discriminativo do modelo.

Tabela 8. Avaliação da lista de prioridade da Regressão Logística mínima (60 dias)

Faixa priorizada	Precision	Recall
Top 5%	0,562	0,196
Top 10%	0,458	0,319
Top 20%	0,339	0,471

Fonte: elaboração própria.

4.6. Análise dos Erros do Modelo

Os falsos positivos da Regressão Logística mínima correspondem, em geral, a matrículas com frequência baixa ou média histórica reduzida que, ainda assim, permaneceram ativas, alertas que podem ser úteis para contato preventivo. Os falsos negativos concentraram-se em casos com pouco histórico mensal, mudança abrupta imediatamente anterior ao cancelamento ou motivo de tempo/saúde sem queda prolongada de frequência. Ensembles com atributos adicionais não reduziram de forma consistente esses erros no teste temporal.

4.7. Discussão dos Resultados

A expectativa inicial de superioridade dos ensembles (Random Forest, XGBoost e LightGBM) não se confirmou. No horizonte de 60 dias, a Regressão Logística mínima alcançou PR-AUC de 0,337, enquanto os modelos com mais atributos ficaram entre 0,190 e

0,258. A explicação mais plausível está na mudança de prevalência: no treino (2025) havia cerca de 2,3% de evasores no horizonte de 60 dias; no teste (2026), cerca de 14,4%. Modelos mais flexíveis tenderam a se ajustar a padrões do período de baixa evasão que não se repetiram na onda de cancelamentos. Em um teste exploratório, a inclusão de LTV e tempo de matrícula elevou artificialmente o PR-AUC da Regressão Logística para valores acima de 0,97, indício claro de vazamento de informação, motivo pelo qual essas variáveis foram excluídas do protocolo oficial. O estudo de ablação reforçou o mesmo ponto: a frequência recente concentra o sinal útil. Cabe registrar que os intervalos de confiança estimados por bootstrap sobrepõem-se parcialmente entre a especificação mínima e o Random Forest, de modo que a vantagem observada deve ser lida como indicativa, e não como separação estatística definitiva. Não se afirma causalidade, e sim associação preditiva; o achado central é que, sob o protocolo adotado, um modelo linear parcimonioso superou algoritmos mais complexos.

A comparação com Goren, Cohen e Rubinstein (2024), Glandorf et al. (2024) e Viana, Santana e Rabêlo (2022) deve considerar diferenças de população, definição de evasão e protocolo. Métricas obtidas com divisão aleatória ou com janela semestral universitária não são equivalentes ao teste temporal prospectivo aqui empregado. O presente estudo distingue-se pelo domínio de cursos de idiomas, pela série mensal de frequência e pela comparação entre uma especificação linear parcimoniosa e ensembles sob protocolo antivazamento com validação temporal.

4.8. Limitações do Estudo

As limitações incluem dados de uma única organização, mudança no processo de registro entre planilha histórica e banco operacional a partir de 2026, ausência de variáveis socioeconômicas, caráter observacional e inexistência de validação externa. A definição de evasão depende da qualidade dos motivos de cancelamento. A concentração de cancelamentos no primeiro semestre de 2026 pode ter influenciado a distribuição temporal dos eventos. Desempenho preditivo não implica eficácia de intervenção; um estudo prospectivo será necessário para avaliar se ações orientadas pelo score reduzem a evasão.

5. CONSIDERAÇÕES FINAIS

Este trabalho criou e avaliou uma abordagem longitudinal para a predição precoce de evasão em cursos de idiomas com dados de frequência e trajetória acadêmica do CCLI. Foram utilizados snapshots por matrícula e horizontes de 30, 60 e 90 dias, com validação temporal e comparação entre algoritmos supervisionados sob restrições antivazamento.

No horizonte de 60 dias, a Regressão Logística com frequência do mês e média histórica alcançou PR-AUC de 0,337 e revocação de 0,355, superando Random Forest (0,258), XGBoost (0,199) e LightGBM (0,190) e ficando bem acima da referência de um ranqueamento sem capacidade discriminativa (14,4% de positivos no teste). No estudo de ablação, a configuração baseada apenas na frequência do mês corrente superou as combinações com defasagens, engenharia temporal e perfil acadêmico-operacional. Na lista dos 10% de maior risco da Regressão Logística mínima, recuperaram-se 31,9% das evasões futuras, com precisão de 45,8%.

Responde-se à questão de pesquisa que informações históricas de frequência possuem capacidade preditiva mensurável para a evasão precoce. Entre os algoritmos avaliados, a Regressão Logística simples foi a mais eficaz; ensembles com atributos adicionais não melhoraram o teste temporal. Ganhos artificiais só apareceram quando se incluíam LTV (meses de permanência) e tempo de matrícula até o encerramento, variáveis que antecipam ou resumem a saída e, por isso, não devem entrar em uma previsão feita antes do cancelamento. Recomenda-se priorizar modelos interpretáveis baseados em frequência e investigar novas fontes de sinal (engajamento qualitativo, interações, feedback) que ainda não estejam redundantes com a trajetória de presença.

Como continuidade, prevê-se validação externa ou prospectiva, monitoramento de estabilidade entre coortes, exploração de variáveis comportamentais não correlacionadas ao encerramento administrativo e investigação de eficácia de intervenções orientadas pelo score, questão que exige desenho experimental distinto do estudo preditivo aqui reportado.

REFERÊNCIAS BIBLIOGRÁFICAS

CHEN, T.; GUESTRIN, C. XGBoost: a scalable tree boosting system. In: ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 22., 2016, San Francisco. *Proceedings [...]*. New York: ACM, 2016. p. 785-794. DOI: https://doi.org/10.1145/2939672.2939785.

GLANDORF, D. et al. Temporal and between-group variability in college dropout prediction. In: INTERNATIONAL CONFERENCE ON LEARNING ANALYTICS AND KNOWLEDGE, 14., 2024, Kyoto.

Proceedings [...]. New York: ACM, 2024. p. 486-497. DOI: https://doi.org/10.1145/3636555.3636906.

GOREN, O.; COHEN, L.; RUBINSTEIN, A. Early prediction of student dropout in higher education using machine learning models. In: INTERNATIONAL CONFERENCE ON EDUCATIONAL DATA MINING, 17., 2024, Atlanta. *Proceedings [...]*. [S. l.]: International Educational Data Mining Society, 2024. p. 349-359. DOI: https://doi.org/10.5281/zenodo.12729834.

KE, G. et al. LightGBM: a highly efficient gradient boosting decision tree. In: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS, 30., 2017. *Proceedings [...]*. [S. l.]: Curran Associates, 2017.

ROMERO, C.; VENTURA, S. Educational data mining and learning analytics: an updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, [s. l.], v. 10, n. 3, e1355, 2020. DOI: https://doi.org/10.1002/widm.1355.

SIEMENS, G.; BAKER, R. S. J. d. Learning analytics and educational data mining: towards communication and collaboration. In: INTERNATIONAL CONFERENCE ON LEARNING ANALYTICS AND KNOWLEDGE, 2., 2012, Vancouver. *Proceedings [...]*. New York: ACM, 2012. p. 252-254. DOI: https://doi.org/10.1145/2330601.2330661.

TINTO, V. Dropout from higher education: a theoretical synthesis of recent research. *Review of Educational Research*, [s. l.], v. 45, n. 1, p. 89-125, 1975. DOI: https://doi.org/10.3102/00346543045001089.

VIANA, F. S.; SANTANA, A. M.; RABÊLO, R. A. L. Avaliação de classificadores para predição de evasão no ensino superior utilizando

janela semestral. In: SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, 33., 2022. *Anais [...]*. Porto Alegre: SBC, 2022. DOI: https://doi.org/10.5753/sbie.2022.224764.

---

[1] Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE), Campus Maracanaú. E-mail: acesse o artigo original para visualizar o e-mail.

[2] Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE), Campus Maracanaú. E-mail: acesse o artigo original para visualizar o e-mail.

[3] Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE), Campus Canindé. E-mail: acesse o artigo original para visualizar o e-mail.

[4] CCLi Consultoria Linguística. E-mail: acesse o artigo original para visualizar o e-mail.

[5] Doutorado em Desenvolvimento Socioeconômico, UCEFF Educacional. E-mail: acesse o artigo original para visualizar o e-mail.