

Quando o Risco Acadêmico se Torna Visível? O Papel da Persistência Temporal

Tobias Vieira Francisco^{1,2}, Abílio Nogueira Barros², Felipe Vieira Roque²,
Tiago Paulino², Augusto Schmidt², Leonardo Brandão Marques²,
Emanuel Marques Queiroga^{1,2}, Elthon Oliveira², Ig Ibert Bittencourt²,
Cristian Cechinel^{2,4*}, Thales Vieira²

¹ Instituto Federal de Educação, Ciência e Tecnologia Sul-rio-grandense (IFSul)
Brazil

²Núcleo de Excelência em Tecnologias Sociais (NEES) Universidade Federal
de Alagoas (UFAL), Brasil

³Universidade Tecnológica Federal do Paraná (UTFPR), Brasil

⁴Centro de Ciências, Tecnologias e Saúde, Universidade Federal de Santa Catarina (UFSC),
Brasil

tobiasfrancisco@ifsul.edu.br

{abilio.barros, felipe.roque, tiago.paulino}@nees.ufal.br

{augusto.schmidt, leonardo.marques, emanuel.queiroga}@nees.ufal.br

{elthon.oliveira, ig.ibert, thales.vieira}@nees.ufal.br

cristian.cechinel@ufsc.br

Abstract. *Early identification of students at risk of adverse academic outcomes is essential for effective educational intervention. In this paper, we investigate the temporal dynamics of academic risk using longitudinal data from Brazilian upper secondary education. We analyze month-by-month risk trajectories generated by national and state-level prediction approaches and characterize students according to temporal risk profiles. Mixed-effects logistic regression showed that persistent profiles increase the odds of future risk continuity by more than three times relative to newly signaled students ($OR > 3.3$). These findings suggest that the visibility of academic risk is not solely determined by the presence of a warning signal, but also by its persistence over time, providing additional information for prioritizing educational interventions.*

Resumo. *O monitoramento precoce de estudantes em risco de resultados acadêmicos adversos é essencial para a realização de intervenções educacionais eficazes. Neste artigo, investigamos a dinâmica temporal do risco acadêmico utilizando dados longitudinais do ensino médio brasileiro. Analisamos trajetórias mensais de risco geradas por abordagens de predição em nível nacional e estadual e caracterizamos os estudantes de acordo com perfis*

*Corresponding author

temporais de risco. A regressão logística de efeitos mistos mostrou que perfis persistentes aumentam em mais de três vezes as chances de continuidade futura do risco em comparação com estudantes recentemente sinalizados ($OR > 3.3$). Esses resultados sugerem que a visibilidade do risco acadêmico não é determinada apenas pela presença de um sinal de alerta, mas também por sua persistência ao longo do tempo, fornecendo informações adicionais para a priorização de intervenções educacionais.

1. Introdução

A não conclusão do ensino médio, incluindo evasão escolar, repetência e absenteísmo crônico, permanece um desafio persistente para os sistemas educacionais em todo o mundo, com consequências sociais e econômicas de longo prazo para indivíduos e sociedades [UIS - UNESCO Institute for Statistics 2012]. No Brasil, esse fenômeno é particularmente acentuado, refletindo desigualdades estruturais, disparidades regionais e limitações institucionais que reduzem a eficácia de políticas educacionais universais [Queiroga et al. 2024b]. Pesquisas anteriores demonstram que o desligamento escolar não é um evento repentino, mas sim um processo cumulativo e dinâmico, moldado por fatores acadêmicos, comportamentais, familiares e contextuais que emergem progressivamente ao longo da trajetória educacional dos estudantes [Parr and Bonitz 2015]. Nesse contexto, a evolução temporal dos sinais de engajamento estudantil torna-se especialmente relevante, uma vez que mudanças na frequência, participação e comportamento acadêmico ao longo do tempo podem revelar padrões associados ao aumento da vulnerabilidade educacional.

Em resposta, as áreas de Learning Analytics (LA) e Educational Data Mining (EDM) ganharam destaque como abordagens voltadas ao apoio à tomada de decisão baseada em dados na educação [Romero and Ventura 2020]. Ao explorar dados administrativos, comportamentais e acadêmicos, modelos de aprendizado de máquina têm demonstrado grande potencial para a identificação precoce de estudantes em risco de desfechos acadêmicos adversos, incluindo evasão e reprovação [Lakkaraju et al. 2015]. Estudos em larga escala realizados em contextos nacionais, como Dinamarca, Uruguai e Brasil, demonstraram que esses modelos podem ser treinados utilizando dados coletados rotineiramente, alcançando desempenho preditivo significativo mesmo sob forte desbalanceamento de classes [Queiroga et al. 2022]. Entretanto, a maior parte dos estudos existentes formula a predição como um problema estático de classificação, normalmente avaliando se um estudante irá eventualmente evadir ou reprovar em um determinado momento fixo no tempo. Essa perspectiva limita a compreensão de como os sinais preditivos evoluem ao longo do ano letivo e de como o risco se intensifica ou se estabiliza à medida que novas evidências comportamentais se tornam disponíveis. Consequentemente, analisar a intensidade e a progressão dos sinais preditivos ao longo do tempo torna-se essencial para equilibrar antecipação e confiabilidade na tomada de decisão educacional [Cechinel et al. 2026].

No Brasil, infraestruturas educacionais de dados em larga escala, como o Sistema Gestão Presente (SGP) [Barros et al. 2025, Barros et al. 2026], têm possibilitado a integração de dados administrativos e comportamentais em escala nacional. Essa infraestrutura pode apoiar políticas públicas como o programa *Pé-de-Meia*, uma iniciativa de transferência condicionada de renda destinada à redução da evasão escolar, que

utiliza esses dados para monitorar a participação dos estudantes e verificar o cumprimento dos critérios de elegibilidade. Além do suporte à implementação dessas políticas, a disponibilidade de registros estudantis continuamente atualizados cria oportunidades para o desenvolvimento de modelos preditivos temporais capazes de monitorar como o risco acadêmico evolui ao longo do ano letivo. Tais infraestruturas permitem não apenas a identificação de estudantes em situação de vulnerabilidade, mas também a análise de como a certeza preditiva e o comportamento dos modelos se modificam à medida que novas informações mensais são incorporadas. Com base nisso, o presente estudo investiga duas questões centrais de pesquisa:

- **RQ1:** Quais configurações distintas de risco acadêmico temporal emergem ao monitorar as trajetórias de predição dos estudantes ao longo de um ano letivo?
- **RQ2:** Esse sinal preditivo permanece robusto em escopos de modelagem centralizados (nacional) e descentralizados (estadual)?

Neste artigo, investigamos a evolução temporal da predição de risco acadêmico ao longo do ano letivo. Analisamos como o desempenho preditivo e a intensidade do risco evoluem mês a mês à medida que a predição de risco dos estudantes se desenvolve. Adotamos uma estratégia de avaliação na qual modelos de aprendizado de máquina são treinados e avaliados utilizando informações temporais progressivamente disponíveis, permitindo observar como as predições realizadas nas etapas iniciais diferem daquelas produzidas em fases mais avançadas em termos de confiança, estabilidade e desempenho preditivo. Esse framework possibilita uma análise detalhada da dinâmica da detecção do risco educacional, identificando períodos críticos nos quais os sinais preditivos se tornam mais confiáveis e examinando o equilíbrio entre a capacidade de intervenção precoce e a precisão preditiva.

O restante deste artigo está organizado da seguinte forma. A Seção 2 apresenta uma revisão dos trabalhos relacionados sobre modelagem preditiva na educação e abordagens de learning analytics temporal. A Seção 3 descreve o delineamento experimental e o framework de avaliação longitudinal. A Seção 4 apresenta e discute os resultados empíricos, a Seção 5 discute os principais achados, e a Seção 6 encerra o artigo com as considerações finais e direções para trabalhos futuros.

2. Trabalhos Relacionados

O uso de LA e EDM para predição de risco acadêmico e evasão escolar cresceu substancialmente ao longo da última década. As primeiras contribuições estabeleceram o papel da análise de dados no apoio à tomada de decisão baseada em evidências e centrada na aprendizagem na educação [Ferguson 2012], enquanto estudos de revisão posteriores documentaram a ampla adoção de técnicas de aprendizado de máquina para prever desempenho acadêmico, reprovação e evasão em contextos de ensino secundário e superior [Balaji et al. 2021].

Diversos estudos em larga escala demonstraram a viabilidade da aplicação de modelos preditivos em nível nacional ou envolvendo múltiplos distritos educacionais por meio do uso de dados administrativos da educação. No Uruguai, Queiroga et al. [Queiroga et al. 2022] relataram uma iniciativa nacional de learning analytics na educação básica (K–12), empregando modelos de Random Forest para prever repetência e evasão utilizando dados de mais de 258,000 estudantes. Mais recentemente, Queiroga et

al. [Queiroga et al. 2024a] propuseram modelos de aprendizado de máquina explicáveis e sensíveis à equidade para antecipar reprovação e evasão escolar em um estado brasileiro, enfatizando interpretabilidade e mitigação de vieses em sistemas de alerta precoce educacionais. De forma semelhante, Sara et al. [Sara et al. 2015] conduziram um estudo nacional na Dinamarca utilizando dados de registros de mais de 36,000 estudantes do ensino médio, alcançando elevado desempenho preditivo com modelos Random Forest. Nos Estados Unidos, Christie et al. [Christie et al. 2019] descreveram um sistema de alerta precoce baseado em aprendizado de máquina implantado em 32 estados, abrangendo milhões de registros educacionais e integrando modelos XGBoost diretamente às plataformas de gestão escolar. Na Coreia do Sul, Lee e Chung [Lee and Chung 2019] analisaram dados nacionais de mais de 165,000 estudantes, demonstrando que métodos baseados em árvores de decisão e boosting superaram abordagens alternativas para a predição de evasão, especialmente em cenários de forte desbalanceamento de classes.

Além da viabilidade técnica, trabalhos anteriores destacaram a importância de alinhar a modelagem preditiva aos processos de tomada de decisão educacional. Lakkaraju et al. [Lakkaraju et al. 2015], em colaboração com grandes distritos escolares dos Estados Unidos, destacaram que métricas de desempenho amplamente utilizadas, como AUC e acurácia, podem não capturar plenamente as necessidades operacionais dos educadores, ressaltando a relevância da seleção de critérios de avaliação que reflitam as prioridades das intervenções.

Lacunas importantes ainda permanecem na literatura. Em particular, análises sistemáticas comparando estratégias de modelagem centralizadas e descentralizadas ainda são limitadas, especialmente em contextos caracterizados por forte heterogeneidade territorial [Coleman et al. 2019]. Os estudos existentes também tendem a enfatizar o desempenho preditivo, oferecendo discussão limitada sobre como o comportamento dos modelos varia entre diferentes territórios e contextos administrativos. Nessa direção, nosso trabalho anterior [Francisco et al. 2027] amplia essa discussão ao comparar estratégias preditivas em nível nacional e estadual utilizando dados educacionais brasileiros em larga escala. Entretanto, investigações mais abrangentes sobre monitoramento temporal e estratégias de generalização regional ainda são escassas. A Tabela 1 resume os principais estudos relacionados discutidos nesta seção.

Table 1. Resumo dos trabalhos relacionados sobre predição de risco educacional e evasão escolar

Estudo	Metodologia	Dados
[Queiroga et al. 2022]	Framework de LA utilizando RF em dados nacionais	298,939 registros de estudantes
[Sara et al. 2015]	Modelos RF aplicados a dados nacionais	36,299 matrículas de estudantes
[Christie et al. 2019]	Integração de modelos em plataformas de gestão escolar	6.4 milhões de registros de estudantes em 32 estados
[Lee and Chung 2019]	Avaliação comparativa de métodos baseados em árvores e boosting	165,715 matrículas de estudantes do ensino médio
[Lakkaraju et al. 2015]	Avaliação de modelos preditivos para apoio à tomada de decisão	200,000 matrículas de estudantes
[Queiroga et al. 2024a]	Modelos de aprendizado de máquina explicáveis e sensíveis à equidade	79,165 registros de estudantes
[Coleman et al. 2019]	Estratégias de modelagem educacional centralizadas versus descentralizadas	>3.5 milhões de estudantes
[Francisco et al. 2027]	Abordagem de modelos nacionais versus estaduais	>6 milhões de estudantes
Este Trabalho	Predições mensais de risco e perfilamento temporal do risco	616,433 registros únicos de estudantes

3. Metodologia

O fluxo metodológico adotado neste estudo está resumido na Figura 1. O processo está organizado em três etapas principais. Primeiramente, a etapa de *Coleta de Dados* consiste na obtenção das predições geradas no estudo anterior de predição de risco educacional em larga escala. Em seguida, a etapa de *Pré-processamento* envolve a integração, o alinhamento temporal e a consolidação das predições mensais em uma base de dados unificada, bem como a construção de perfis temporais de risco e indicadores de transição utilizados para caracterizar a evolução do risco acadêmico ao longo do ano letivo. Por fim, a etapa de *Análise Estatística* compreende a aplicação de métodos estatísticos descritivos e inferenciais, incluindo testes de hipóteses para investigar associações entre os perfis temporais de risco e os desfechos futuros de risco acadêmico. A principal contribuição metodológica deste estudo consiste em transformar predições mensais de risco em trajetórias temporais de risco e, posteriormente, derivar perfis temporais capazes de capturar a persistência e a recorrência dos sinais de risco acadêmico.

3.1. Desenho do Estudo e Pipeline de Modelagem Anterior

Este estudo estende uma investigação anterior em larga escala que analisou o comportamento de equidade e generalização de modelos de predição de risco acadêmico treinados em diferentes escalas territoriais no Brasil [Francisco et al. 2027]. O estudo anterior concentrou-se na comparação entre estratégias preditivas centralizadas e descentralizadas por meio da avaliação de: (i) um modelo nacional treinado utilizando dados de todos os

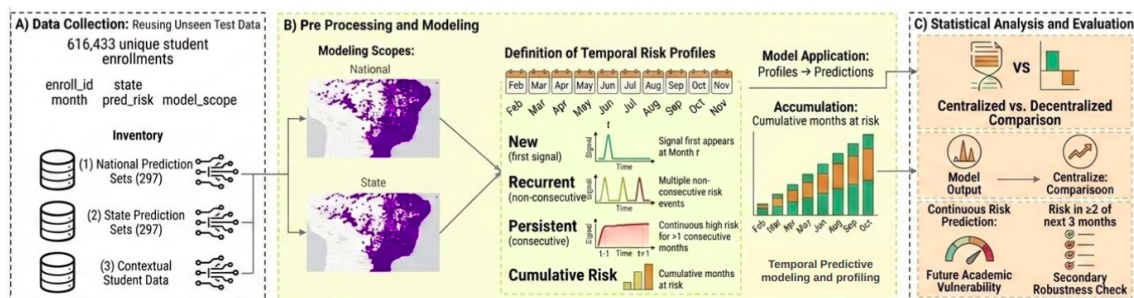


Figure 1. Fluxo metodológico do estudo.

estados brasileiros e (ii) modelos estaduais independentes, treinados separadamente para cada unidade da federação. O objetivo foi investigar se modelos nacionais ou modelos estaduais localizados produziam desempenho preditivo mais confiável e equitativo em contextos educacionais heterogêneos.

O presente trabalho reutiliza as previsões geradas pelo estudo anterior para investigar a dinâmica temporal do risco acadêmico ao longo do ano letivo. Mais especificamente, este estudo analisa como os sinais de risco evoluem, persistem, reaparecem e transitam ao longo do tempo após o processo original de previsão. Um total de 616,433 matrículas únicas de estudantes foi incluído na análise. É importante destacar que apenas as partições de teste geradas no estudo anterior foram utilizadas neste trabalho, garantindo que todas as análises fossem conduzidas exclusivamente sobre dados não vistos e evitando qualquer contaminação proveniente das amostras de treinamento ou validação. Para cada estudante e para cada mês, estavam disponíveis as previsões provenientes de ambos os escopos de modelagem (nível nacional / nível estadual), possibilitando comparações temporais e territoriais diretas entre as estratégias preditivas centralizadas e descentralizadas.

3.2. Construção da Base de Dados Analítica Longitudinal

Inicialmente, foi gerado um inventário de todos os arquivos de previsão mensais para assegurar a consistência entre os pipelines de modelagem nacional e estadual. O inventário continha 594 arquivos de previsão, distribuídos igualmente entre os dois escopos de modelagem, sendo 297 arquivos de previsão mensal produzidos pelo modelo nacional e 297 arquivos gerados pelos modelos estaduais.

Ambas as abordagens abrangiam exatamente o mesmo conjunto de 616,433 matrículas únicas de estudantes, garantindo total comparabilidade entre as estratégias de previsão. As previsões cobrem o ano letivo de fevereiro a dezembro e incluem previsões binárias mensais de risco acadêmico para cada estudante. Posteriormente, todos os arquivos mensais de previsão foram consolidados em uma única base de dados analítica longitudinal, integrando as previsões provenientes de ambos os escopos de modelagem. A base de dados final continha 13,561,526 observações longitudinais, igualmente distribuídas entre as duas estratégias de modelagem, com 6,780,763 observações para cada escopo.

A Figura 2 apresenta a evolução mensal dos estudantes identificados como em risco considerando ambas as abordagens de modelagem. A figura compara a proporção de estudantes classificados como em risco pelo modelo nacional e pelos modelos estaduais ao longo do ano letivo. Como pode ser observado, ambas as estratégias exibem com-

portamento temporal altamente semelhante, sugerindo uma dinâmica de predição consistente em larga escala, apesar das diferenças na granularidade territorial da modelagem. A Tabela 2 resume as principais variáveis incluídas na base de dados analítica longitudinal.

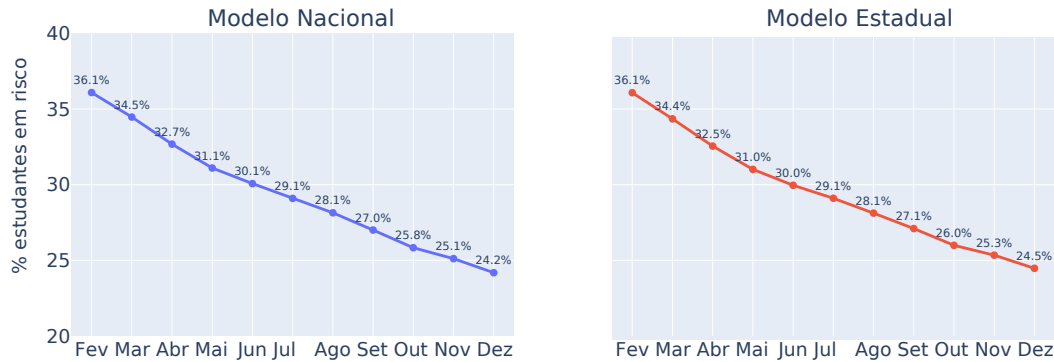


Figure 2. Estudantes identificados como em risco por mês considerando os modelos nacionais e estaduais

Table 2. Variáveis incluídas na base de dados analítica longitudinal

Variável	Descrição
mat_id	Identificador único da matrícula do estudante
estado	Sigla da unidade federativa brasileira
mes, mes_idx	Mês do calendário e índice temporal
pred_risco	Predição binária de risco acadêmico
modelo_escopo	Escopo da modelagem (nacional ou estadual)
flag_ant	Indicador de predição anterior de risco antes do mês t
flag_t_menos_1	Predição de risco no mês imediatamente anterior
risco_profile	Perfil temporal de risco (novo, recorrente, persistente ou não sinalizado)
consecutivo_risco	Número de meses consecutivos em risco
cumulativo_risco	Número acumulado de meses em risco até o mês t
transicao_tipo	Padrão de transição temporal entre meses
y_t_mais_1	Desfecho de risco no mês subsequente
y_proximo3	Continuidade do risco ao longo dos três meses subsequentes
IN_RISCO_SGP	Desfecho oficial final de risco do SGP

A unidade analítica adotada neste estudo é a combinação entre matrícula do estudante, mês e escopo de modelagem. Para analisar a dimensão temporal do risco acadêmico, as predições binárias mensais ($pred_risco_t \in \{0, 1\}$) foram ordenadas temporalmente para cada estudante, gerando sequências longitudinais de estados de risco acadêmico de fevereiro a dezembro.

Para cada estudante i , as predições foram organizadas da seguinte forma:

$$(i, mes_1), (i, mes_2), \dots, (i, mes_T)$$

Uma restrição temporal rigorosa foi mantida para evitar *data leakage*. Todos os preditores disponíveis no mês t foram calculados exclusivamente utilizando informações dos meses 1 até t , enquanto desfechos como Y_{t+1} foram derivados apenas de observações futuras. Essa organização permitiu a análise da persistência, recorrência, transições temporais e padrões de continuidade do risco ao longo do ano letivo.

3.3. Definição dos Perfis Temporais de Risco

Para investigar se os padrões de persistência e recorrência fornecem informação preditiva adicional, os estudantes foram classificados em perfis temporais de risco mutuamente exclusivos em cada mês t , de acordo com seu histórico de sinalização.

Foram definidos os seguintes perfis:

- **Novo:** O estudante é sinalizado como em risco no mês corrente ($pred_risco_t = 1$), mas não apresentou sinais anteriores de risco durante o ano letivo ($flag_ant_t = 0$).
- **Recorrente:** O estudante é sinalizado como em risco no mês corrente ($pred_risco_t = 1$), possui pelo menos um sinal anterior de risco ($flag_ant_t = 1$), mas não foi sinalizado no mês imediatamente anterior ($pred_risco_{t-1} = 0$).
- **Persistente:** O estudante é sinalizado como em risco no mês corrente ($pred_risco_t = 1$) e também foi sinalizado no mês imediatamente anterior ($pred_risco_{t-1} = 1$).
- **Não sinalizado:** O estudante não é sinalizado como em risco no mês corrente ($pred_risco_t = 0$) e nunca foi sinalizado como em risco durante o ano letivo ($flag_ant_t = 0$).

O principal desfecho utilizado para avaliar esses perfis temporais foi a continuidade do risco acadêmico no mês subsequente (Y_{t+1}). Como análise secundária de robustez, também avaliamos se os estudantes permaneceram em risco em pelo menos dois dos três meses subsequentes.

3.4. Análise Estatística

A associação entre os perfis temporais de risco e os desfechos futuros de risco acadêmico foi inicialmente avaliada por meio do teste do qui-quadrado de independência (*Chi-square test of independence*) [McHugh 2013]. Para identificar diferenças específicas entre os perfis, foram realizados testes de comparação de proporções (*pairwise proportion tests*) entre: Novo vs. Recorrente; Novo vs. Persistente; Recorrente vs. Persistente. A correção de Holm [Holm 1979] foi aplicada para controlar a taxa de erro familiar (*family-wise error rate*) nas comparações múltiplas. Os tamanhos de efeito foram reportados utilizando o V de Cramér [Cramér 1999].

Para analisar de forma mais aprofundada a estrutura longitudinal dos dados, empregamos um Modelo Linear Generalizado Misto (GLMM, *Generalized Linear Mixed Model*) [Breslow and Clayton 1993] com função de ligação logística (*logistic link function*). Como o conjunto de dados contém observações repetidas dos mesmos estudantes ao longo dos meses, foi incluído um intercepto aleatório em nível individual (u_i) para considerar a correlação temporal intraestudante. O modelo é definido como:

$$\text{logit}(P(Y_{t+1} = 1)) = \beta_0 + \beta_1 \text{Recorrente} + \beta_2 \text{Persistente} + \beta_3 \text{Estado} + \beta_4 \text{Model} + u_i \quad (1)$$

Em que o perfil "Novo" é utilizado como categoria de referência. Os resultados são apresentados como (*Odds Ratios* – OR) com intervalos de confiança de 95%.

4. Resultados

4.1. Perfis Temporais de Risco entre os Escopos de Modelagem

A Tabela 3 apresenta a distribuição dos perfis temporais de risco para ambos os escopos de modelagem. Tanto nos modelos nacionais quanto nos modelos estaduais, a maioria das observações correspondeu a estudantes que não foram sinalizados como em risco em nenhum mês. Entre os estudantes que receberam predições de risco, o perfil persistente representou a categoria predominante, correspondendo a mais de 77% de todos os casos sinalizados. Em contraste, os perfis recorrentes representaram menos de 6% dos estudantes sinalizados, indicando que o reaparecimento descontínuo do risco foi substancialmente menos frequente do que a persistência temporal contínua.

A Figura 3 apresenta a distribuição temporal dos estudantes identificados como em risco de acordo com os perfis de risco definidos. Como esperado, em fevereiro todos os estudantes classificados como em risco pertencem exclusivamente ao perfil *Novo* (verde). A partir de março, emerge o perfil *Persistente* (vermelho), representando estudantes identificados consecutivamente como em risco, bem como novas ocorrências do perfil *Novo* (representando estudantes que não haviam sido identificados como em risco no mês anterior). A partir de abril, o perfil *Recorrente* (laranja) também passa a ser observado, caracterizando estudantes que retornam à condição de risco após uma interrupção temporária na sinalização. Em ambas as abordagens de modelagem (nacional e estadual), o perfil *Persistente* torna-se o padrão temporal predominante durante a maior parte do ano letivo. Esse resultado sugere que, uma vez identificados como em risco, os estudantes tendem a permanecer continuamente expostos à vulnerabilidade acadêmica nos meses subsequentes.

Table 3. Distribuição dos perfis temporais de risco por escopo de modelagem

Perfil	Nacional		Estadual	
	Quantidade	%	Quantidade	%
Novo	340,345	5.0%	339,985	5.0%
Recorrente	115,717	1.7%	113,579	1.7%
Persistente	1,539,642	22.7%	1,544,127	22.8%
Não sinalizado	4,785,059	70.6%	4,783,072	70.5%
Total	6,780,763	100%	6,780,763	100%

O diagrama de Sankey (Figura 4) ilustra a transição mês a mês entre os diferentes perfis de risco dos estudantes e sua situação acadêmica no período subsequente ($t + 1$). O lado esquerdo representa o perfil de entrada: Novo, Recorrente ou Persistente, enquanto o lado direito apresenta o estado resultante (Com Risco ou Sem Risco). Observa-se que, como os padrões de transição entre os modelos Nacional e Estadual mostraram-se praticamente idênticos (com variação máxima de 0.5%), é apresentado um único fluxo representativo para ambos os escopos de modelagem. Os fluxos são coloridos de acordo com seu perfil de origem: fluxos vermelhos indicam que sinais persistentes apresentam a maior probabilidade de permanecer em estado de risco, enquanto os fluxos verdes (Novo) e azuis (Recorrente) apresentam maiores taxas de reversão do risco ou volatilidade.

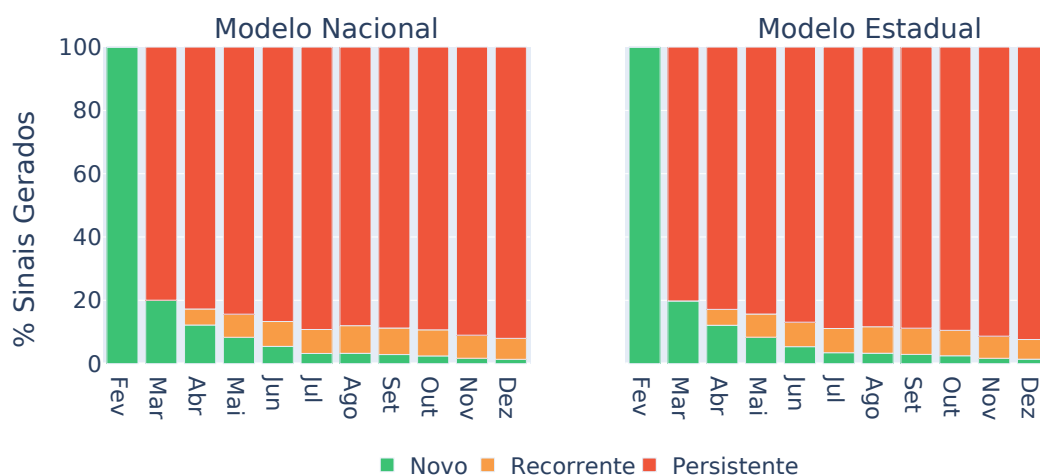


Figure 3. Distribuição dos perfis de sinalização por mês

4.2. Continuidade Temporal do Risco Acadêmico

Analizamos se os perfis temporais de risco estavam associados a diferentes probabilidades de continuidade futura do risco acadêmico. A Tabela 4 apresenta as taxas de continuidade para o desfecho Y_{t+1} nos três perfis temporais.

Table 4. Taxas de continuidade do risco no mês subsequente (Y_{t+1})

Perfil	Nacional		Estadual	
	Taxa	N	Taxa	N
Novo	69.8%	338,323	69.8%	337,874
Recorrente	60.1%	105,847	60.6%	104,205
Persistente	88.4%	1,402,466	88.6%	1,404,767

Os resultados revelam diferenças na continuidade futura do risco entre os perfis temporais. O perfil Persistente estudante-mês apresentou a maior probabilidade de permanecer em risco no mês subsequente, com taxas de continuidade superiores a 88% em ambos os escopos de modelagem. Em contraste, os estudantes classificados como Recorrentes apresentaram as menores taxas de continuidade, em torno de 60%, enquanto os estudantes recentemente sinalizados exibiram níveis intermediários de continuidade, próximos de 70%.

4.3. Associação entre Perfis Temporais e Risco Futuro

A associação entre os perfis temporais e a continuidade futura do risco foi inicialmente avaliada por meio do teste do qui-quadrado de independência. A Tabela 5 resume os resultados estatísticos. Os testes revelaram associações altamente significativas entre os perfis temporais e a continuidade futura do risco acadêmico em ambos os escopos de modelagem.

As comparações pareadas pós-hoc utilizando testes de proporção com correção de Holm demonstraram ainda que todos os pares de perfis diferiram significativamente

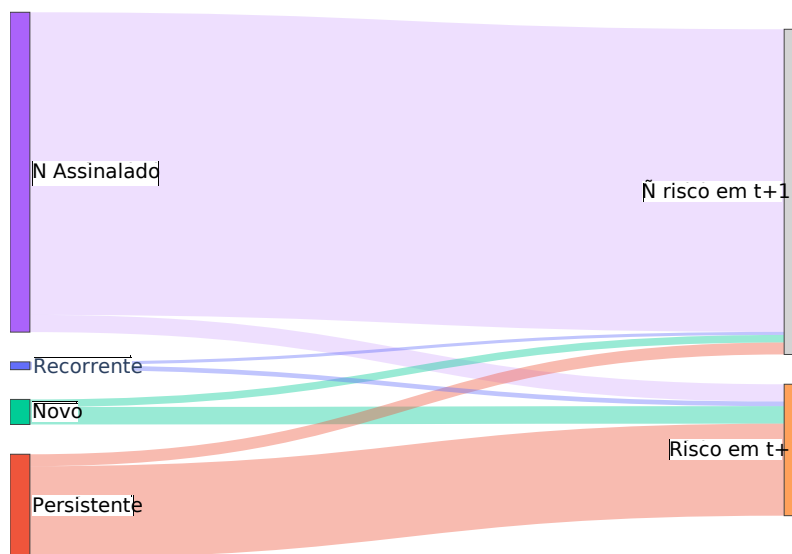


Figure 4. Transições mês a mês entre os perfis de risco dos estudantes e sua situação acadêmica subsequente ($t + 1$).

Table 5. Testes do qui-quadrado para perfis temporais e continuidade futura do risco

Escopo	χ^2	df	p-value
Nacional	112,004.9	2	< 0.001
Estadual	113,047.1	2	< 0.001

entre si ($p < 0.001$). Os perfis Persistentes apresentaram consistentemente taxas de continuidade significativamente superiores às dos perfis Novo e Recorrente, enquanto os perfis Recorrentes exibiram as menores probabilidades de continuidade futura.

4.4. Regressão Logística de Efeitos Mistos

Para considerar a estrutura temporal dos dados e as observações repetidas de cada estudante, foi estimado um modelo de regressão logística de efeitos mistos utilizando a matrícula do estudante como intercepto aleatório. Os resultados da regressão de efeitos mistos são consistentes com as análises descritivas. Estudantes classificados no perfil *Persistente* apresentaram chances substancialmente maiores de serem identificados como em risco no mês subsequente quando comparados aos estudantes classificados no perfil *Novo*. As OR estimadas foram de 3.30 (IC 95%: 3.27–3.33) para a análise nacional e 3.37 (IC 95%: 3.34–3.40) para a análise estadual, indicando que estudantes persistentes apresentaram mais de três vezes a probabilidade de permanecer em risco no ciclo de predição seguinte. A Tabela 6 apresenta as Razões de Chances obtidas para os perfis temporais, utilizando o perfil Novo como categoria de referência.

De modo geral, os resultados demonstram uma distinção clara entre os perfis temporais. Enquanto os perfis *Novo* e *Recorrente* estiveram associados a menores probabilidades de continuidade futura do risco, o perfil *Persistente* apresentou uma associação significativamente mais forte com a continuidade do risco, com chances superiores a três vezes aquelas observadas para estudantes recém-sinalizados em ambos os escopos

analíticos.

Table 6. OR para a continuidade futura do risco acadêmico

Escopo	Comparação	OR	Limite Inferior IC 95%	Limite Superior IC 95%
Nacional	Recorrente vs. Novo	0.65	0.64	0.66
Nacional	Persistente vs. Novo	3.30	3.27	3.33
Estado	Recorrente vs. Novo	0.66	0.65	0.67
Estado	Persistente vs. Novo	3.37	3.34	3.40

De maneira geral, os resultados indicam que o perfil *Persistente* é o mais fortemente associado à continuidade futura do risco. Enquanto estudantes classificados como *Novo* e *Recorrente* apresentaram probabilidades comparáveis ou menores de permanecer em risco no mês subsequente, estudantes classificados como *Persistente* apresentaram mais de três vezes a probabilidade de serem identificados como em risco no ciclo de predição seguinte. Esse padrão mostrou-se altamente consistente tanto nas análises nacionais quanto nas estaduais.

4.5. Análise do Tamanho de Efeito

Para avaliar a força da associação entre os perfis temporais e a continuidade futura do risco acadêmico de forma independente do tamanho da amostra, foi calculado o V de Cramér para ambos os escopos de modelagem. Os tamanhos de efeito obtidos indicam uma associação de magnitude moderada entre os perfis temporais de risco e a continuidade futura do risco acadêmico. Embora os tamanhos de efeito não sejam extremos, eles são substanciais considerando a natureza heterogênea e em larga escala do sistema educacional analisado.

Table 7. Força da associação entre perfis temporais e continuidade futura do risco

Escopo	χ^2	V de Cramér	Força da Associação
Nacional	112,004.9	0.246	Moderada
Estadual	113,047.1	0.247	Moderada

5. Discussão

Este estudo investigou a dinâmica temporal de sistemas de alerta precoce preditivos em larga escala, deslocando a perspectiva analítica de predições estáticas realizadas em um único momento para trajetórias temporais de risco. Ao analisar fluxos mensais de predição de mais de 616.000 estudantes, examinamos como a persistência e a recorrência dos sinais de risco se relacionam com a vulnerabilidade acadêmica futura. Essa perspectiva temporal também pode apoiar a etapa de priorização em sistemas de alerta precoce [Cechinel et al. 2026], uma vez que os perfis resultantes fornecem um critério adicional para hierarquizar estudantes além de uma única predição de risco. A discussão dos resultados está organizada em torno das duas questões de pesquisa que orientaram este estudo.

Nossa primeira questão de pesquisa (RQ1) buscou caracterizar o surgimento e a distribuição das configurações temporais derivadas das trajetórias mensais de predição.

Os resultados empíricos revelam uma distribuição altamente assimétrica entre os estudantes sinalizados. Em ambas as arquiteturas de modelagem (nacional/estadual), quando um estudante é identificado como em risco, esse sinal é predominantemente crônico, e não episódico: o perfil persistente corresponde a aproximadamente 77% de todas as ocorrências sinalizadas. Em contrapartida, os perfis recorrentes, nos quais um estudante deixa de estar em risco e posteriormente retorna a essa condição, são notavelmente raros, representando menos de 6% das observações sinalizadas.

Nossa segunda questão de pesquisa (RQ2) investigou em que medida esses perfis possuem poder preditivo em relação à continuidade do risco no curto prazo (Y_{t+1}), bem como se esse sinal varia entre abordagens de modelagem centralizadas (nacional) e descentralizadas (estadual). Os modelos de regressão logística de efeitos mistos e as taxas empíricas de continuidade fornecem evidências de que as trajetórias temporais contêm informações preditivas relevantes. Estudantes que apresentam um perfil de risco persistente demonstram uma taxa de continuidade extremamente elevada ($> 88.4\%$), correspondente a uma OR superior a três ($OR > 3.3$) em comparação com estudantes recém-sinalizados. Esse resultado reforça que a persistência temporal atua como um fator intensificador da vulnerabilidade futura ao risco.

Embora estruturas nacionais em larga escala, como as implementadas no Uruguai por [Queiroga et al. 2022], na Dinamarca por [Sara et al. 2015] e na Coreia do Sul por [Lee and Chung 2019], tenham demonstrado de forma convincente a viabilidade técnica da utilização de registros administrativos para prever desfechos educacionais, esses estudos normalmente formulam a predição como um evento transversal realizado em um único momento. Em geral, tais abordagens avaliam se um estudante irá eventualmente evadir ou reprovar com base em uma fotografia estática dos dados disponíveis. Em contraste, este trabalho amplia a literatura ao propor uma metodologia sensível à intensidade para predições contínuas ao longo do tempo. Em vez de tratar as saídas mensais dos modelos de aprendizado de máquina como pontos isolados de informação, nosso mecanismo de perfilamento (*Novo*, *Recorrente* e *Persistente*) considera a sequência e o acúmulo desses sinais como uma característica preditiva em si. Essa abordagem permite quantificar a "intensidade" temporal do risco acadêmico, capturando como um sinal se estabiliza ou oscila à medida que o ano letivo avança.

Uma perspectiva complementar é apresentada por [Mosia 2025], que concebem o risco acadêmico como um processo temporal latente utilizando um modelo Bayesiano hierárquico de espaço de estados com dinâmica autorregressiva. Um coeficiente autorregressivo de aproximadamente 0.82 indica que, uma vez que o nível latente de risco de um estudante aumenta, ele tende a permanecer elevado ao longo do tempo, salvo mudanças significativas nas condições subjacentes. Um padrão semelhante é observado no presente estudo para estudantes classificados no perfil *Persistente* (ver Tabela 6). De forma semelhante, [Howard et al. 2018] investigaram quando as intervenções devem ocorrer, enquanto os perfis temporais propostos neste trabalho contribuem para indicar quais estudantes devem ser priorizados. Essa perspectiva questiona o tratamento do risco acadêmico como uma sequência de eventos independentes de predição e, em vez disso, sustenta que o risco deve ser compreendido como uma trajetória dinâmica, cuja evolução ao longo do tempo constitui, por si só, uma informação relevante.

Esses resultados sugerem que a questão não é apenas se um estudante é iden-

tificado como em risco, mas também quão consistentemente esse risco é observado ao longo do tempo. Sob essa perspectiva, a visibilidade do risco acadêmico emerge por meio da persistência temporal: sinais repetidos fornecem evidências mais robustas da continuidade futura do risco do que predições isoladas. Essa dimensão temporal pode, portanto, auxiliar gestores e demais atores educacionais a avançar além da simples detecção binária do risco, permitindo uma compreensão mais refinada sobre quais estudantes necessitam de atenção mais imediata.

6. Conclusão

Este estudo investigou a evolução temporal do risco acadêmico ao deslocar a avaliação das predições de aprendizado de máquina de instantâneos estáticos para o perfilamento das trajetórias dos estudantes. Ao acompanhar fluxos mensais de predição em uma infraestrutura pública de educação em larga escala, demonstramos que a configuração temporal dos sinais de risco contém informações operacionais relevantes para a tomada de decisão. Uma limitação deste trabalho reside nos limites macrorregionais da comparação realizada, que se concentrou exclusivamente em um modelo nacional centralizado e em modelos estaduais independentes. Embora essa abordagem em duas perspectivas tenha permitido revelar macrodinâmicas regionais, ela não considera variações socioeconômicas intraestaduais nem diferenças decorrentes de políticas administrativas em nível municipal.

Além da granularidade da arquitetura de modelagem, a principal direção para pesquisas futuras consiste na incorporação prática desse framework contínuo e sensível à intensidade em sistemas operacionais reais de alerta precoce. Em vez de funcionar apenas como uma ferramenta diagnóstica *offline*, esse framework de perfilamento contínuo deve ser integrado diretamente a painéis de monitoramento utilizados em infraestruturas ativas de gestão educacional, apoiando a implementação de políticas públicas. A evolução dos sistemas de alerta precoce, deixando de operar apenas com notificações binárias para incorporar sistemas capazes de acompanhar trajetórias de risco em tempo real, permitirá que formuladores de políticas públicas e gestores escolares substituam intervenções rígidas e uniformes por estratégias de acompanhamento mais dinâmicas, contextualizadas e sensíveis às necessidades individuais dos estudantes.

Agradecimentos

Este trabalho foi apoiado pelo Ministério da Educação (MEC), por meio dos Termos de Execução Descentralizada (TEDs) nº 13914 e nº 11476; pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), sob o processo nº 445016/2024-8; pelo projeto Ia.Edu – Instituto Nacional de Inteligência Artificial na Educação Desplugada, sob o processo nº 4084883/2024-5; e pela Fundação de Amparo à Pesquisa do Estado de Alagoas (FAPEAL), sob o processo nº FAPEAL60030.0000001481/2025.

References

- Balaji, P., Alelyani, S., Qahmash, A., and Mohana, M. (2021). Contributions of machine learning models towards student academic performance prediction: a systematic review. *Applied Sciences*, 11(21):10007.
- Barros, A. N., Queiroga, E. M., Marcolino, M. R., Silva, D. B. L., Dermeval, D., Lima, A., Marques, L. B., Cechinel, C., and Vieira, T. (2025). Ensuring data quality in na-

- tional educational databases: Insights from brazil’s centralized database of high school students’ data. In *Conference on Digital Government Research*, volume 1.
- Barros, A. N., Roque, F. V., Paulino, T., Schmidt, A., Galvani, F., Oliveira, R., Marques, L. B., Dermeval, D., Stefani, A. G. M., Costa, M. D. S. D., Queiroga, E. M., Oliveira, E., Cechinel, C., and Vieira, T. (2026). Interpretable machine learning for predicting school absenteeism at-scale in Brazil: evidence from 22,000 schools. *Scientific Reports*.
- Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American statistical Association*, 88(421):9–25.
- Cechinel, C., Arantes, F., Pessoa, M., Santana, D. S., Madeira, R. d. A., Cunha, S. F. d., de Oliveira, A. B., Roque, F., Freires, L., de Aguiar, M. A., Vieira, T., Bittencourt, I. I., Isotani, S., and Queiroga, E. M. (2026). The sews framework: Bridging the prediction-to-action gap in educational early warning systems. In Blanchard, E. G., Chen, G., Chi, M., and Isotani, S., editors, *Artificial Intelligence in Education. Late Breaking Results, WideAIED, Practitioners, Industry and Policies, Blue Sky, Doctoral Consortium, FoL Workshops and Tutorials, FoL Invited Papers*, pages 442–456, Cham. Springer Nature Switzerland.
- Christie, S. T., Shi, A., Gobert, L., and Rosen, C. (2019). Machine-learned school dropout early warning at scale. In *Proceedings of the 12th International Conference on Educational Data Mining*.
- Coleman, C., Baker, T., and Stephenson, E. (2019). A better cold-start for early prediction of student at-risk status in new school districts. In *Proceedings of the 12th International Conference on Educational Data Mining*.
- Cramér, H. (1999). *Mathematical methods of statistics*, volume 9. Princeton university press.
- Ferguson, R. (2012). Learning analytics: drivers, developments and challenges. *International Journal of Technology Enhanced Learning*, 4(5/6):304–317.
- Francisco, T. V., Barros, A. N., Roque, F. V., Paulino, T., Schmidt, A., Galvani, F., Oliveira, R., Marques, L. B., Dermeval, D., Barreto, P., Stefani, A. G. M., de Santana da Costa, M., Queiroga, E. M., Oliveira, E., Bittencourt, I. I., Cechinel, C., and Vieira, T. (2027). Territorial fairness in large-scale academic risk prediction: Comparing national and state-level machine learning models in brazil. In Blanchard, E. G., Chen, G., Chi, M., and Isotani, S., editors, *Artificial Intelligence in Education*, pages 225–239, Cham. Springer Nature Switzerland.
- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70.
- Howard, E., Meehan, M., and Parnell, A. (2018). Contrasting prediction methods for early warning systems at undergraduate level. *The Internet and Higher Education*, 37:66–75.
- Lakkaraju, H., Aguiar, E., Shan, C., Miller, D., Bhanpuri, N., Ghani, R., and Addison, K. L. (2015). A machine learning framework to identify students at risk of adverse academic outcomes. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1909–1918.

- Lee, S. and Chung, J. Y. (2019). The machine learning-based dropout early warning system for improving the performance of dropout prediction. *Applied Sciences*, 9(15):3093.
- McHugh, M. L. (2013). The chi-square test of independence. *Biochemia medica*, 23(2):143–149.
- Mosia, M. (2025). A bayesian state-space approach to dynamic hierarchical logistic regression for evolving student risk in educational analytics. *Data*, 10(2):23.
- Parr, A. K. and Bonitz, V. S. (2015). Role of family background, student behaviors, and school-related beliefs in predicting high school dropout. *The Journal of Educational Research*, 108(6):504–514.
- Queiroga, E. M., Paragarino, V. R., Casas, A. P., Primo, T. T., Munoz, R., Ramos, V. C., and Cechinel, C. (2022). Early prediction of at-risk students in secondary education: A countrywide k-12 learning analytics initiative in uruguay. *Information*, 13(9):401.
- Queiroga, E. M., Santana, D., da Silva, M., de Aguiar, M., dos Santos, V., Mello, R. F., Bittencourt, I. I., and Cechinel, C. (2024a). Anticipating student abandonment and failure: Predictive models in high school settings. In *International Conference on Artificial Intelligence in Education*, pages 351–364. Springer.
- Queiroga, E. M., Siqueira, E. S., Portela, C. D. S., Cordeiro, T. D., Bittencourt, I. I., Isotani, S., Mello, R. F., Muñoz, R., and Cechinel, C. (2024b). Data-driven strategies for achieving school equity: Insights from brazil and policy recommendations. *IEEE Access*, 12:101646–101659.
- Romero, C. and Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *WIREs Data Mining and Knowledge Discovery*, 10(3):e1355.
- Sara, N.-B., Halland, R., Igel, I. G., and Alstrup, S. (2015). High-school dropout prediction using machine learning: A danish large-scale study. In *Proceedings of the 23rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*.
- UIS - UNESCO Institute for Statistics (2012). Global education digest 2012—opportunities lost: The impact of grade repetition and early school leaving.