

Plataforma preditiva de combate à evasão escolar para o Ensino Fundamental: construção do modelo preditivo

Gabriel Monteiro Cordeiro de Sousa¹, Lucas Francisco Alcantara Sales Macedo¹
David Carlos Pereira da Cunha¹, Alex Sandro Gomes¹

¹Centro de Informática – Universidade Federal de Pernambuco (UFPE)
Recife – PE – Brasil

{gmcs, lfasm, dcpc, asg}@cin.ufpe.br

Abstract. School dropout in elementary education is often preceded by disengagement, absenteeism, and low performance not always addressed early. This paper presents a predictive model for a platform combating school dropout, using AI and educational data mining for early identification of at-risk students. The solution combines preprocessing, machine learning, risk classification, and dashboards, following CRISP-DM. Five models were compared on a dataset with 15,193 records and 73 columns (84.36% permanence, 15.64% dropout). XGBoost with SMOTE achieved the best balance: accuracy 0.912, F1-score 0.718, precision 0.720, recall 0.716. Results indicate technical feasibility, requiring further validation.

Keywords School dropout, Educational data mining, Machine learning, Early warning systems, Elementary education.

Resumo. A evasão escolar no Ensino Fundamental costuma ser antecedida por desengajamento, infrequência e baixo desempenho nem sempre tratados precocemente. Este artigo apresenta o modelo preditivo de uma plataforma de combate à evasão, baseada em IA e mineração de dados, para identificação precoce de risco. A solução combina pré-processamento, aprendizado de máquina, classificação de risco e dashboards, seguindo o CRISP-DM. Foram comparados cinco modelos em um conjunto de 15.193 registros e 73 colunas (84,36% permanência, 15,64% evasão). O XGBoost com SMOTE apresentou o melhor equilíbrio: acurácia de 0,912, F1-score de 0,718, precisão de 0,720 e recall de 0,716. Os resultados indicam viabilidade técnica, com validação ainda pendente.

Palavras-Chave Evasão escolar, Mineração de dados educacionais, Aprendizado de máquina, Sistemas de alerta precoce, Ensino Fundamental.

1. Introdução

A permanência escolar é uma condição central para a garantia do direito à educação e para a redução de desigualdades sociais. No Brasil, embora a matrícula no Ensino Fundamental tenha alcançado alta cobertura, o abandono e a evasão seguem como problemas relevantes, observados como processos progressivos de afastamento do estudante da escola. Dados recentes da PNAD Contínua Educação indicam que, em 2024, 8,7 milhões de jovens de 14 a 29 anos não haviam concluído o ensino médio por terem abandonado a escola ou por nunca a terem frequentado; o mesmo levantamento registra

sinais de abandono ainda em idades compatíveis com o Ensino Fundamental, com 6,5% dos jovens tendo deixado a escola até os 13 anos e 6,8% aos 14 anos [IBGE 2025].

No Censo Escolar da Educação Básica 2024, o Inep registra o Ensino Fundamental como uma das principais etapas da educação básica brasileira e informa que os anos finais concentravam aproximadamente 11,5 milhões de matrículas, o que reforça a escala do problema quando se pensa em mecanismos de acompanhamento, prevenção e intervenção [INEP 2025]. Nesse cenário, a identificação tardia de estudantes em risco reduz a capacidade de resposta das escolas, pois muitas ações passam a ocorrer apenas quando a infrequência se tornou recorrente ou quando o vínculo escolar já se encontra fragilizado.

A evasão escolar é um fenômeno multifatorial, cujas causas se distribuem em pelo menos três dimensões interligadas: a socioeconômica, que inclui a necessidade de trabalho precoce e as condições materiais da família; a trajetória escolar, marcada por reprovações, distorção idade-série e dificuldades de aprendizagem; e a dimensão relacional e contextual, que abrange baixa vinculação com a escola, problemas familiares, violência, deslocamento e sofrimento socioemocional. Em todos esses casos, a ausência de acompanhamento sistemático amplifica o risco [Bowers et al. 2013, Han et al. 2025, Barker e Siddiqui 2026]. Por isso, abordagens baseadas apenas em um indicador isolado, como faltas acumuladas, tendem a ser insuficientes para apoiar a prevenção. Sistemas de alerta precoce e técnicas de mineração de dados educacionais podem contribuir para organizar sinais dispersos, transformando dados administrativos e acadêmicos em informações acionáveis para equipes escolares.

Este artigo visa apresentar a construção do modelo preditivo a ser inserido em uma plataforma de combate à evasão escolar nos anos finais do Ensino Fundamental, baseada em inteligência artificial e mineração de dados educacionais. O texto está organizado da seguinte forma: a Seção 2 apresenta conceitos relacionados à evasão, abandono e infrequência, além de trabalhos relacionados sobre modelos preditivos e sistemas de alerta precoce; a Seção 3 descreve a plataforma preditiva proposta e sua arquitetura modular; a Seção 4 detalha o método adotado, incluindo o entendimento do problema, a preparação dos dados, as estratégias de balanceamento e modelagem e os cuidados éticos; a Seção 5 apresenta e discute os resultados dos experimentos; a Seção 6 sintetiza as limitações do estudo; e a Seção 7 apresenta as conclusões e perspectivas futuras.

2. Trabalhos Relacionados e Estado da Arte

Esta seção delimita os conceitos educacionais associados à permanência escolar e revisa trabalhos recentes sobre modelos preditivos, sistemas de alerta precoce e indicadores monitoráveis no contexto da evasão escolar.

2.1. Evasão, Abandono e Infrequência

Neste artigo, a infrequência é entendida como a presença de faltas recorrentes com vínculo formal ainda ativo, o abandono como a interrupção da frequência em um período letivo, e a evasão como um processo mais amplo de não permanência no percurso escolar [INEP 2025, Bowers et al. 2013]. Essa distinção orienta a plataforma a atuar antes da consolidação do afastamento, quando ainda há espaço para intervenção pedagógica, social ou administrativa. Nos anos finais do Ensino Fundamental, a transição entre infância e adolescência amplia essa complexidade: reprovações, defasagem idade-série, mudanças

familiares, trabalho precoce e dificuldades de aprendizagem podem se combinar, exigindo que a tomada de decisão preserve o papel da análise humana e da responsabilidade institucional.

2.2. Modelos Preditivos e Sistemas de Alerta Precoce

Em sistemas de alerta precoce, a mineração de dados educacionais e o Learning Analytics combinam capacidade preditiva e visualização analítica para apoiar equipes escolares na identificação antecipada de estudantes em risco [Baker e Yacef 2009, Buckingham Shum e Ferguson 2012]. Estudos sobre predição de abandono destacam que indicadores de desempenho, frequência e trajetória escolar apresentam valor preditivo, mas também apontam limites de precisão e sensibilidade e a necessidade de conectar a predição a intervenções efetivas [Bowers et al. 2013].

Estudos com dados administrativos nacionais confirmam que trajetórias de frequência constituem um dos preditores mais informativos, com potencial para antecipar processos de abandono escolar [Barker e Siddiqui 2026]. Do ponto de vista do horizonte preditivo, [Psyridou et al. 2024] demonstram que modelos de aprendizado de máquina podem identificar risco de abandono do ensino médio utilizando informações disponíveis ainda ao final da educação primária, antecipando sinais de risco antes da consolidação do desfecho. Iniciativas de aplicação de técnicas de inteligência artificial à educação primária também reforçam a viabilidade de incorporar modelos preditivos a contextos de gestão escolar [Hidalgo-Hidalgo et al. 2025].

Quanto aos algoritmos, Random Forest é recorrente na literatura, com resultados expressivos em contextos distintos, como dados administrativos sul-coreanos [Han et al. 2025] e escolas primárias rurais peruanas [Gallardo et al. 2025]; variantes de *gradient boosting* também mostram desempenho competitivo em bases de larga escala, como no Chile [Rodríguez et al. 2023] e no Marrocos [El Jihaoui et al. 2025]. Um achado transversal é que a técnica de balanceamento de classes exerce influência comparável à escolha do algoritmo [Rezk e Selim 2024, Selim e Rezk 2023].

No contexto brasileiro, ainda há desafios relevantes para a aplicação de modelos preditivos à evasão no Ensino Fundamental: acesso restrito a bases individualizadas da educação básica, heterogeneidade entre planilhas, atraso no registro de informações e risco de interpretações deterministas. Na literatura analisada, não foi identificado estudo brasileiro de Ensino Fundamental com modelo preditivo aplicado a alunos individuais, lacuna que motiva o desenvolvimento do sistema proposto.

3. Plataforma Preditiva de Combate à Evasão

Este trabalho apresenta o desenvolvimento do modelo preditivo de uma plataforma que permite traduzir registros acadêmicos e administrativos em indicadores visuais de fácil interpretação para tomada de decisão. A intenção de design é aumentar a capacidade de professores, gestores e equipes pedagógicas de escolas públicas de organizar esforços de acolhimento e acompanhamento de estudantes sob risco de evasão.

A plataforma foi concebida como apoio à decisão para auxiliar gestores, orientadores educacionais, coordenadores pedagógicos e professores na identificação precoce de estudantes dos anos finais do Ensino Fundamental sob risco de abandono.

A solução não automatiza decisões sobre estudantes, mas organiza probabilidades, faixas de risco e indicadores para subsidiar a análise humana.

A solução possui uma arquitetura modular: o *backend* em Python/Flask gerencia o banco PostgreSQL e expõe a API para a interface web. Na prática, quando a escola envia uma planilha, o *backend* higieniza os registros, roda as previsões de evasão e devolve as faixas de risco já calculadas para os painéis de visualização. A Figura 1 ilustra o fluxo do *pipeline*, e a Tabela 1 resume os módulos funcionais do sistema.

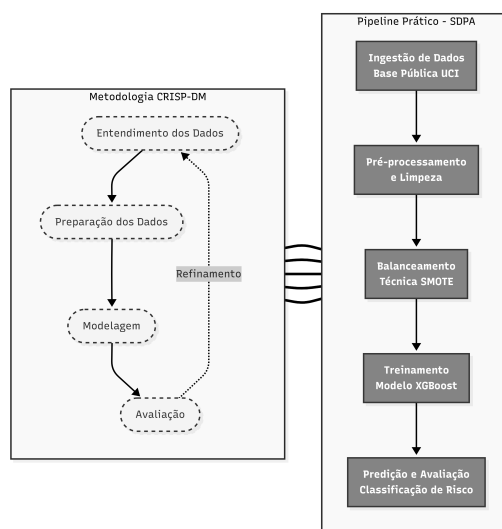


Figura 1. Fluxograma do *pipeline* de aprendizado de máquina da plataforma e sua correspondência com o CRISP-DM.

4. Método

A condução metodológica do projeto foi organizada segundo o CRISP-DM (*Cross-Industry Standard Process for Data Mining*) [Shearer 2000], adaptado ao problema educacional para conectar a compreensão do contexto escolar às etapas de preparação, modelagem, avaliação e implantação da solução.

4.1. Entendimento do Problema

A etapa de entendimento do problema partiu da necessidade de identificar estudantes em risco antes que a evasão se consolidasse. A partir de documentos de requisitos, entrevistas e discussões com atores educacionais, o sistema foi definido como uma ferramenta de apoio à decisão, voltada a priorizar acompanhamento e não apenas produzir relatórios descritivos após a evasão já ter ocorrido. Essa decisão influenciou a escolha das métricas, privilegiando a análise da classe positiva de evasão e, especialmente, o *recall*, pois um falso negativo corresponde a um estudante em risco que não seria sinalizado pelo sistema.

4.2. Entendimento e Preparação dos Dados

Para esta fase experimental, foi utilizado um conjunto com 15.193 registros e 73 colunas, correspondente a um relatório de situação escolar de estudantes do Distrito Federal referente ao ano letivo de 2023 (dados de Censo Escolar/IDEB), obtido diretamente junto à rede de ensino. Por reunir majoritariamente códigos administrativos de situação por

Tabela 1. Módulos principais da plataforma e suas funções.

Módulo	Função no sistema
Autenticação e controle de acesso	Gerenciar o login de usuários autorizados, garantindo a segurança e privacidade dos dados.
Importação de dados	Receber arquivos estruturados (CSV ou XLSX), validar formato e iniciar o processamento.
Pré-processamento	Tratar nulos, padronizar variáveis, codificar categóricas e preparar <i>features</i> .
Modelagem preditiva	Treinar ou carregar modelo para estimar probabilidade de evasão.
Classificação de risco	Converter probabilidade em faixas interpretáveis: baixo, médio e alto risco.
Dashboards e consultas	Apresentar indicadores agregados, listas para consulta de estudantes, filtros e visualizações para apoio à decisão.
Logs e rastreabilidade	Registrar importações, treinamentos, consultas e metadados relevantes.

etapa de ensino – sem indicadores mais ricos de frequência diária ou de comportamento em sala –, esta etapa tem caráter auxiliar e investigativo, servindo para avaliar a viabilidade técnica da *pipeline* antes de uma validação mais ampla. A variável-alvo foi definida a partir do campo `SITUACAO`: a classe de evasão corresponde ao código `SIR`, enquanto a classe de permanência agrega as demais categorias (aprovado, reprovado, abandono formalmente registrado e óbito). Observou-se forte desbalanceamento, detalhado na Tabela 2.

As 73 colunas estruturadas cobrem domínios complementares: identificação administrativa e georreferenciamento da escola (`CO_ENTIDADE`, `NO_ENTIDADE`, `NU_LATITUDE`, `NU_LONGITUDE`), indicadores de qualidade (`NOTA_IDEB`, `IDEB_REAL`), perfil demográfico do estudante (`IDADE`, `TP_SEXO`, `TP_COR_RACA`) e situação de matrícula por etapa de ensino (`SIT_EI_PRE`, `SIT_EI_TOTAL`, `SIT_EF_AI`, `SIT_EF_TOTAL`, `SIT_EJA_TOTAL`, entre outras), dialogando com a literatura que aponta frequência, notas e trajetória escolar como preditores relevantes [Han et al. 2025, Rodríguez et al. 2023, Barker e Siddiqui 2026]. A variável `TP_SITUACAO` foi removida por vazamento de informação, por corresponder diretamente ao desfecho.

Tabela 2. Distribuição da variável-alvo antes do balanceamento.

Classe	Interpretação no experimento	Quantidade	Percentual
0	Permaneceu	12.817	84,36%
1	Evadiu	2.376	15,64%
Total	–	15.193	100%

A preparação dos dados incluiu limpeza de registros, tratamento de valores ausentes, separação entre *features* e alvo, codificação de variáveis categóricas e padronização de variáveis numéricas. Para colunas numéricas, foi adotada imputação por mediana; para colunas categóricas, imputação por moda. Variáveis categóricas foram transformadas por *OneHotEncoder* com estratégia de ignorar categorias não vistas no treino, e variáveis numéricas foram padronizadas com *StandardScaler*.

Um cuidado metodológico central foi a remoção de variáveis com vazamento de informação. No experimento, a variável *TP_SITUACAO* foi removida por estar diretamente relacionada ao desfecho, o que poderia inflar artificialmente os resultados. Foram comparados dois fluxos: o primeiro removia apenas a variável de vazamento; o segundo removia também variáveis com alta correlação com o alvo. A comparação indicou que a remoção excessiva poderia eliminar atributos úteis, e o fluxo com remoção mínima se mostrou mais consistente.

4.3. Balanceamento, Modelos e Métricas

Diante do desbalanceamento, foram comparadas duas técnicas de *oversampling*: *Random OverSampling* e SMOTE. O *Random OverSampling* duplica exemplos da classe minoritária, enquanto o SMOTE cria exemplos sintéticos por interpolação entre instâncias próximas da classe minoritária [Chawla et al. 2002]. A aplicação dessas técnicas foi considerada apenas no conjunto de treinamento, preservando a avaliação em dados não balanceados de teste.

Foram avaliados cinco modelos clássicos de classificação para dados tabulares: Regressão Logística, Random Forest, Gradient Boosting, XGBoost e *Support Vector Machine*. O XGBoost foi escolhido como modelo de destaque por sua robustez em dados tabulares, capacidade de capturar relações não lineares e desempenho observado nos experimentos. A literatura descreve o XGBoost como um sistema escalável de *boosting* de árvores, com otimizações para dados esparsos e aprendizagem eficiente [Chen e Guestrin 2016].

As métricas observadas foram acurácia, F1-score macro, precisão e *recall* da classe 1 – esta última com atenção especial, pois no contexto de prevenção é mais crítico deixar de identificar estudantes em risco do que gerar falsos positivos, medida particularmente informativa para problemas desbalanceados [Saito e Rehmsmeier 2015].

4.4. Cuidados Éticos e Proteção de Dados

Embora os experimentos preliminares descritos neste artigo tenham utilizado uma base auxiliar e ainda pouco detalhada, a plataforma foi concebida para operar futuramente com dados educacionais sensíveis e mais ricos. Por isso, sua evolução deve observar a LGPD, mecanismos de controle de acesso, anonimização em relatórios, rastreabilidade das operações e protocolos institucionais de autorização para uso de dados. Essa preocupação é particularmente relevante porque probabilidades de evasão podem produzir efeitos indevidos caso sejam interpretadas como diagnóstico definitivo sobre estudantes.

5. Resultados e Discussão

Os experimentos indicaram que o fluxo com remoção mínima de variáveis, associado ao SMOTE, preservou informações relevantes e apresentou o melhor equilíbrio geral entre os

cenários avaliados. A Tabela 4 compara os cinco modelos testados nesse cenário. Embora Regressão Logística, Gradient Boosting e SVM tenham alcançado valores elevados de *recall* para a classe de evasão, eles apresentaram menor precisão, aumentando a tendência a falsos positivos. O XGBoost, por sua vez, obteve a maior acurácia, o maior F1-score macro e a maior precisão da classe positiva, mantendo *recall* competitivo.

A comparação evidencia que o XGBoost foi o modelo mais equilibrado para o objetivo do experimento: alcançou acurácia de 0,912, F1-score macro de 0,718 e precisão de 0,720 para a classe de evasão. O SVM apresentou o maior *recall* da classe positiva (0,901), seguido pela Regressão Logística (0,878), mas ambos tiveram precisão inferior (0,510 e 0,531, respectivamente), indicando maior quantidade relativa de falsos positivos. O Random Forest obteve bom desempenho geral, com acurácia de 0,903 e precisão de 0,699, mas *recall* menor para evasão (0,669).

A Tabela 3 ilustra como a probabilidade estimada pelo modelo é traduzida, na interface da plataforma, em uma faixa de risco interpretável pelas equipes escolares.

Tabela 3. Tradução da saída do modelo em faixas de risco no dashboard.

Faixa de risco	Probabilidade estimada	Indicação no dashboard
Baixo	< 30%	Acompanhamento de rotina
Médio	30% – 60%	Atenção e busca ativa preventiva
Alto	> 60%	Prioridade para intervenção imediata

Tabela 4. Comparação dos modelos no cenário SMOTE + Fluxo 1 (classe 1 = evasão).

Modelo	Acurácia	F1-score macro	Precisão classe 1	Recall classe 1
Regressão Logística	0,860	0,662	0,531	0,878
Random Forest	0,903	0,684	0,699	0,669
Gradient Boosting	0,886	0,675	0,611	0,756
XGBoost	0,912	0,718	0,720	0,716
SVM	0,849	0,651	0,510	0,901

Assim, a escolha do XGBoost não decorre de superioridade universal do algoritmo, mas do melhor compromisso observado entre identificar estudantes em risco e evitar excesso de sinalizações incorretas. Em um contexto escolar, modelos com *recall* muito alto e baixa precisão podem gerar listas extensas de estudantes sinalizados, sobrecarregando equipes pedagógicas e reduzindo a viabilidade operacional do acompanhamento. Portanto, a escolha final do modelo e dos limiares de risco precisa considerar não apenas métricas estatísticas, mas também a capacidade das escolas de agir sobre os alertas produzidos.

Os resultados preliminares reforçam três pontos: a evasão pode ser tratada como problema de classificação supervisionada, desde que haja cuidado com a definição do alvo e a temporalidade das variáveis; o desbalanceamento é central, exigindo técnicas e métricas apropriadas para evitar conclusões enganosas; e a utilidade prática do modelo depende da capacidade da plataforma de transformar probabilidades em informações acionáveis para as equipes escolares.

A classificação em baixo, médio e alto risco torna a saída do modelo mais compreensível para usuários não técnicos, mas introduz decisões normativas: limiares fixos, como 30% e 60%, devem ser vistos como parâmetros iniciais, sujeitos a ajuste a partir da validação com dados históricos, curvas *precision-recall* e disponibilidade das equipes. A literatura sobre sistemas de alerta precoce reforça que a predição só tem valor institucional quando conectada a intervenções claras, acompanhamento humano e políticas de permanência [Bowers et al. 2013]; por isso, a evolução do sistema deve incluir também validação de usabilidade, clareza dos *dashboards* e definição de protocolos de ação.

6. Limitações

A principal limitação estrutural desta etapa é a natureza da base pública utilizada, já discutida na Seção 4.2: mesmo comprovando a viabilidade técnica da *pipeline*, esses dados não substituem a validação longitudinal em bases institucionais reais – que frequentemente apresentam ruídos, campos ausentes e inconsistências administrativas – nem foram avaliados com usuários finais. No contexto brasileiro, a aplicação de aprendizado de máquina à evasão ainda encontra desafios relevantes: acesso restrito a bases individualizadas da educação básica, heterogeneidade entre planilhas e risco de interpretações deterministas; por isso, uma abordagem responsável deve apresentar o risco como sinal de apoio à decisão, e não como sentença sobre o estudante.

Do ponto de vista metodológico, o risco de vazamento de informação exige revisão contínua para garantir que as *features* reflitam apenas o conhecimento prévio ao desfecho, simulando um cenário real de detecção precoce. No aspecto sociotécnico, o uso de predições suscita preocupações éticas quanto à estigmatização do estudante; para mitigar esse efeito, a ferramenta deve operar estritamente como apoio probabilístico à decisão, mantendo a responsabilidade sobre o acompanhamento e as intervenções com os profissionais da educação.

7. Conclusão

Este artigo apresentou a construção de um modelo preditivo para uma plataforma de combate à evasão escolar nos anos finais do Ensino Fundamental, integrando arquitetura web, API REST, banco de dados, *pipeline* de aprendizado de máquina e *dashboards* voltados à tomada de decisão educacional, estruturada segundo o CRISP-DM.

A principal contribuição deste artigo é descrever uma abordagem aplicada que combina modelo preditivo, arquitetura computacional e visualizações analíticas para apoiar a identificação precoce de estudantes em risco – uma lacuna na literatura, que ainda carece de iniciativas voltadas ao Ensino Fundamental brasileiro com predição em nível individual.

Os experimentos preliminares indicaram que XGBoost combinado com SMOTE e remoção cuidadosa de vazamento de informação produz resultados promissores (acurácia de 0,912, F1-score macro de 0,718) – o principal valor do estudo, porém, está em demonstrar a viabilidade da *pipeline* e explicitar os cuidados necessários para o contexto escolar, não em afirmar desempenho definitivo. Como trabalhos futuros: ampliar as bases de dados do Ensino Fundamental; validar o modelo em diferentes recortes temporais e territoriais; e avaliar, com gestores e professores, seu impacto na permanência escolar.

Declaração sobre uso de Inteligência Artificial

Os autores declaram que ferramentas de Inteligência Artificial (IA) Generativa foram utilizadas neste trabalho exclusivamente para fins de revisão gramatical e aprimoramento da fluidez textual do manuscrito. Os autores assumem total responsabilidade por todo o conteúdo, resultados, análises e integridade do artigo submetido.

Referências

- Baker, R. S. J. d. e Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, 1(1):3–17.
- Barker, C. G. e Siddiqui, N. (2026). Attendance trajectories as early predictors of school dropout: Evidence from a national cohort in chile. *International Journal of Educational Development*.
- Bowers, A. J., Sprott, R., e Taff, S. A. (2013). Do we know who will drop out? a review of the predictors of dropping out of high school: Precision, sensitivity, and specificity. *The High School Journal*, 96(2):77–100.
- Buckingham Shum, S. e Ferguson, R. (2012). Social learning analytics. *Journal of Educational Technology & Society*, 15(3):3–26.
- Chawla, N. V., Bowyer, K. W., Hall, L. O., e Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.
- Chen, T. e Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, San Francisco, CA. ACM.
- El Jihaoui, M., Abra, O. E. K., e Mansouri, K. (2025). Predicting and explaining middle-school dropout risk on imbalanced data. *E3S Web of Conferences*, 680:00073.
- Gallardo, C. D. C., Chulluncuy, L. V. T., e Quintana, C. A. A. (2025). Web application based on machine learning for the prediction and prevention of low academic performance in primary school students in the district of matara, cajamarca. In *Proceedings of the IEEE International Conference on Big Data*.
- Han, Y. et al. (2025). Distinguishing characteristics of out-of-school adolescents in south korea: A machine learning approach. *International Journal of Social Welfare*. Publicado online em outubro de 2024.
- Hidalgo-Hidalgo, K. J., Huaricanha-Estrella, R., Abarca-Torres, A. R., Huaman-Alegria, M., Espino-Oncebay, P., e Rodriguez-Baca, L. S. (2025). Predicting school dropout rates in primary education using artificial intelligence techniques. In *International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME 2025)*.
- IBGE (2025). Pesquisa nacional por amostra de domicílios contínua: Educação 2024. Technical report, IBGE, Rio de Janeiro. Módulo anual de educação, divulgado em junho de 2025.
- INEP (2025). Censo escolar da educação básica 2024: Resumo técnico. Technical report, INEP, Brasília.

- Psyridou, M., Prezja, F., Torppa, M., Lerkkanen, M.-K., Poikkeus, A.-M., e Vasalampi, K. (2024). Machine learning predicts upper secondary education dropout as early as the end of primary school. *Scientific Reports*, 14.
- Rezk, S. S. e Selim, K. S. (2024). Ensemble pruning for predicting school dropouts: A machine learning approach. *Journal of Computational Social Science*, 7(2):1555–1597.
- Rodríguez, P., Villanueva, A., Dombrovskaja, L., e Valenzuela, J. P. (2023). A methodology to design, develop, and evaluate machine learning models for predicting dropout in school systems: The case of chile. *Education and Information Technologies*, 28:10103–10149.
- Saito, T. e Rehmsmeier, M. (2015). The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3):e0118432.
- Selim, K. S. e Rezk, S. S. (2023). On predicting school dropouts in egypt: A machine learning approach. *Education and Information Technologies*, 28(7):9235–9266.
- Shearer, C. (2000). The CRISP-DM model: The new blueprint for data mining. *Journal of Data Warehousing*, 5(4):13–22.