PREDIÇÃO DE DESEMPENHO ACADÊMICO USANDO *MACHINE LEARNING*

Mateus Gabriel Missio[1], Sidnei Renato Silveira[1], Guilherme Bernardino da Cunha[1], Vinicius Gadis Ribeiro[2]

[1]UFSM – Universidade Federal de Santa Maria – Campus Frederico Westphalen/RS - Brasil (mgmissio@gmail.com)
[2]UFRGS – Universidade Federal do Rio Grande do Sul – Campus Litoral Norte, Brasil

Resumo: Este artigo apresenta o desenvolvimento de uma solução de análise preditiva para apoiar a identificação de estudantes com risco de reprovação escolar. Foram utilizados os algoritmos Regressão Logística e *XGBoost*, combinados a estratégias de balanceamento de classes com SMOTE (*Synthetic Minority Over-sampling Technique*), com o objetivo de reduzir o viés em favor da classe majoritária e melhorar a detecção dos casos de reprovação. Os resultados indicaram que a Regressão Logística apresentou o melhor equilíbrio para o propósito do trabalho. Com base nisso, as probabilidades geradas foram adotadas como referência para classificação do nível de risco em um *dashboard* desenvolvido no Power BI.

Palavras-chave: Análise de Dados; Aprendizado de Máquina; Predição de Desempenho Escolar.

INTRODUÇÃO

Este artigo apresenta o estudo e a aplicação de técnicas de *Machine Learning* para prever o desempenho acadêmico de estudantes a partir da análise de dados educacionais históricos. Nos últimos anos, o avanço das tecnologias digitais e a ampliação do uso de plataformas educacionais têm gerado um volume expressivo de dados que, quando adequadamente tratados e analisados, podem revelar padrões importantes sobre o processo de aprendizagem (Penteado et al., 2019). No entanto, grande parte das instituições de ensino ainda utiliza essas informações de forma predominantemente administrativa, sem explorá-las como apoio estratégico para decisões pedagógicas ou para a identificação precoce de alunos em risco de reprovação ou evasão.

Neste contexto, o objetivo geral deste trabalho consistiu em estudar e validar modelos de *Machine Learning* capazes de prever a reprovação escolar com base em variáveis comportamentais, familiares e acadêmicas presentes em uma base educacional pública. Além disso, buscou-se disponibilizar os resultados em um *dashboard* interativo, com o intuito de identificar alunos em situação de risco e subsidiar intervenções pedagógicas mais eficazes.

Durante a evolução do projeto, a estratégia metodológica foi ajustada. Embora a proposta inicial previsse o uso de bases abertas do INEP (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira), optou-se pela utilização do *dataset Student Performance*, disponibilizado no UCI *Machine Learning Repository* (UCI.EDU, 2026), devido à sua estrutura organizada, ampla adoção em estudos acadêmicos e adequação ao objetivo de classificação de risco. A base reúne atributos referentes ao contexto familiar, hábitos de estudo, faltas, apoio escolar, reprovações anteriores e desempenho final do estudante.

Para atingir o objetivo proposto, foram realizadas as seguintes atividades: 1) carregamento e inspeção da base de dados, 2) tratamento e codificação das variáveis categóricas, 3) criação da variável alvo de reprovação, 4) separação em conjuntos de treino e teste, 5) aplicação de balanceamento com SMOTE (*Synthetic Minority Over-sampling Technique*), 6) treinamento dos modelos Regressão Logística e *XGBoost*, 7) comparação de desempenho por meio de métricas tais como precisão, *recall*, *F1-score* e matriz de confusão, 8) análise das variáveis mais influentes e, 9) por fim, construção de um *dashboard* utilizando a ferramenta *Microsoft Power* BI.

A escolha do tema deste trabalho surgiu a partir da problematização que enfrentam muitos docentes e/ou gestores educacionais quanto ao uso de técnicas pedagógicas preventivas voltadas para estudantes com elevado risco de reprovação ou que se configuram como desafios institucionais. O cenário é agravado pelo fato de que a maioria das instituições acumula volumosos dados históricos dos alunos, tais como notas, frequência e registros comportamentais,

os quais, na prática, são raramente usados como subsídio para decisões intervenientes ao longo do ano letivo (Ceneviva et al., 2022).

Essa constatação motivou a buscar soluções para transformar esses dados armazenados em um recurso efetivo de apoio à gestão educacional, antecipando a identificação de situações de risco e permitindo intervenções pedagógicas mais assertivas e oportunas, o projeto pode contribuir para melhorar processos educacionais, mas também reforçar a confiança e a responsabilidade institucional.

Para dar conta desta proposta, o artigo está organizado da seguinte forma: a Seção 2 apresenta o referencial teórico; a Seção 3 descreve trabalhos relacionados; a Seção 4 detalha a solução implementada, com destaque para a base utilizada, os modelos adotados e os resultados obtidos; e, por fim, a Seção 5 reúne as considerações finais.

REFERENCIAL TEÓRICO

Esta seção apresenta um referencial teórico sobre as áreas envolvidas neste trabalho destacando o uso de ferramentas relevantes para realizar análises de dados. Para o desenvolvimento deste trabalho foram aplicados modelos de classificação, os quais permitiram prever situações relevantes ligadas ao desempenho acadêmico dos estudantes.

EDUCAÇÃO E O USO DE DADOS ACADÊMICOS

A educação contemporânea tem sido cada vez mais impactada pela transformação digital. Nesse ponto, muitas instituições de ensino acumulam grandes volumes de informações relevantes ao histórico dos alunos. No campo de *Learning Analytics*, esse conjunto de informações cria oportunidades para medir, analisar e reportar dados relevantes com o objetivo de compreender e otimizar os processos de ensino e aprendizagem (Cardoso et al., 2022; Possato et al., 2025). Estudos apontam que a análise de dados pode contribuir para a identificação de alunos em risco, reduzindo índices de evasão e favorecendo intervenções pedagógicas mais oportunas.

O trabalho de Manzanares et al. (2018), por exemplo, demonstrou que padrões de engajamento em um AVA (Ambiente Virtual de Aprendizagem) podem ser utilizados para prever desempenho acadêmico. Esse tipo de resultado reforça a pertinência de investigar o uso de modelos preditivos no contexto educacional.

MACHINE LEARNING, MODELOS PREDITIVOS E MINERAÇÃO DE DADOS

ML (*Machine Learning*) é um ramo da Inteligência Artificial que permite a construção de algoritmos capazes de aprender padrões a partir de dados e realizar previsões ou classificações sem programação explícita (Alpaydin, 2020; Awad; Khanna, 2015; Mitchell, 1997). ML provê a base técnica para a Mineração de Dados (*Data Mining*), para filtrar bases de dados automaticamente, buscando regularidades ou padrões e realizar uma análise preditiva (Silveira; Vit, 2023).

A Mineração de Dados é uma das etapas do processo de Descoberta de Conhecimento em Bancos de Dados (KDD – *Knowledge Discovery in Databases*). Segundo Carvalho (Carvalho, 2005 citado por Silveira; Vit, 2023), a Mineração de Dados envolve um conjunto de técnicas de Inteligência Artificial, com apoio da Estatística, com o objetivo de descobrir conhecimentos que estão implícitos em grandes massas de dados.

A Mineração de Dados permite que se definam regras de negócio para auxiliar nas tomadas de decisões. Busca-se, com isso, criar um planejamento das atividades que podem ser desenvolvidas e pensadas a médio e longo prazo, tentando fazer assim uma previsão de tendências futuras baseada no passado (Lorenzi; Silveira, 2011).

FERRAMENTAS PARA ANÁLISE PREDITIVA

A aplicação de análises preditivas no contexto educacional e institucional demanda o uso de ferramentas e linguagens específicas, que viabilizem desde a coleta e tratamento dos dados até a modelagem e interpretação dos resultados, conforme mostra a Figura 1.

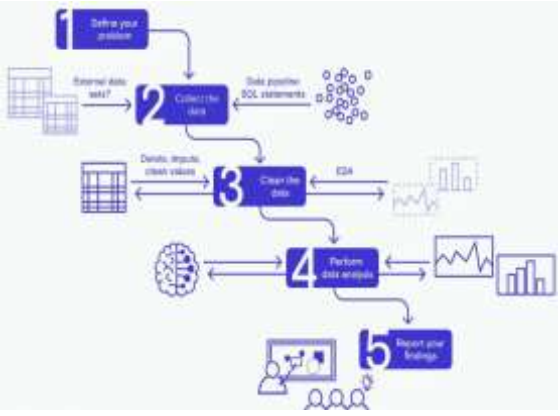

Figura 1: Ilustração do fluxo sequencial da ciência de dados, destacando desde a motivação para coletar informações até a visualização final dos resultados.
Fonte: Rudder Stack.com, 2025

A Figura 1 apresenta o fluxo básico de um processo de análise de dados, dividido em cinco etapas principais: definição do problema, coleta dos dados, limpeza dos dados, análise dos dados e apresentação dos resultados. Esse modelo orienta a estruturação de projetos em ciência de dados, garantindo que as

informações sejam tratadas de forma organizada e eficiente, desde a origem até a geração de insights úteis para a tomada de decisão.

Entre as ferramentas que podem ser empregadas e que possibilita o fluxo ideal apresentado na Figura 1, destacam-se:

- Linguagem de Programação *Python*: destaca-se como uma das linguagens de programação mais utilizadas em ciência de dados, oferecendo bibliotecas consolidadas (KABIR; AHMED, 2024), tais como:
    - *Pandas*: para manipulação e limpeza de dados;
    - *scikit-learn*: biblioteca voltada à construção de modelos preditivos;
    - *XGBoost*: que implementam algoritmos de *boosting* com alta performance;
    - SHAP: utilizada para análise de aplicabilidade dos modelos, permitindo compreender o impacto de cada variável nas previsões.
- *Power BI*: ferramenta de *Business Intelligence* que possibilita a construção de *dashboards* interativos, facilitando a comunicação dos resultados analíticos de forma visual e acessível (Microsoft.com, 2025).

TRABALHOS RELACIONADOS

A presente seção tem, como objetivo, contextualizar e fundamentar o desenvolvimento deste trabalho por meio da análise de estudos anteriores que abordam temáticas semelhantes. Neste contexto, apresentamos artigos e pesquisas que exploram o uso de técnicas de *Machine Learning* na predição de desempenho acadêmico, identificação de alunos em risco de evasão ou reprovação, e aplicação de ferramentas analíticas no contexto educacional.

A análise desses trabalhos permite compreender os avanços já realizados na área, identificar lacunas existentes e destacar os diferenciais da solução apresentada neste estudo. Além disso, contribui para validar a relevância da abordagem adotada, evidenciando como modelos preditivos podem apoiar a tomada de decisão pedagógica e promover intervenções mais eficazes no ambiente escolar.

O trabalho apresentado por Caetano (2016), sob o título "O Uso de Técnicas de Aprendizado de Máquina na Predição de Desempenho Acadêmico de Alunos em Cursos Superiores" investiga o uso de algoritmos de aprendizado de máquina para prever o desempenho de estudantes em cursos superiores. A autora utiliza dados acadêmicos como notas, frequência e participação em atividades para treinar modelos preditivos, com destaque para algoritmos tais como *Decision Trees*, *Naive Bayes e WEKA*. O estudo reforça a importância da análise preditiva como ferramenta de apoio à gestão educacional.

De forma semelhante, Benevento (2024) propõe a aplicação de algoritmos de *Machine Learning* para prever o desempenho acadêmico de alunos, com o objetivo de personalizar o ensino e melhorar a administração escolar. O trabalho explora modelos tais como *Random Forest* e SVM (*Support Vector Machine*) e discute a integração dessas ferramentas com plataformas educacionais.

No artigo de Franqueira *et al*. (2024) são analisados os desafios e oportunidades da aplicação de IA na avaliação de desempenho acadêmico no ensino médio. A abordagem é baseada em revisão bibliográfica e destaca o potencial da IA para identificar padrões de aprendizagem e apoiar intervenções pedagógicas.

Analisando as informações, observamos que os trabalhos pesquisados compartilham o objetivo comum de aplicar técnicas de *Machine Learning* para prever o desempenho acadêmico dos estudantes. No entanto, o trabalho desenvolvido apresenta diferenciais relevantes em relação às abordagens anteriores. Enquanto os trabalhos de Caetano (2016) e Benevento (2024) focam na aplicação de algoritmos preditivos em contextos específicos como cursos superiores ou plataformas institucionais, o estudo desenvolvido propõe uma solução mais abrangente, baseada em dados abertos disponíveis no repositório UCI *Machine Learning Repository* (UCI.EDU, 2026) o que amplia a aplicabilidade do modelo para diferentes instituições e níveis de ensino. Além disso, o uso de múltiplos algoritmos consolidados (tais como *XGBoost* e Regressão Logística) permite uma análise comparativa mais robusta, com validação cruzada e temporal.

Outro diferencial importante está na visualização dos resultados por meio de um *dashboard* interativo, voltado para professores e gestores. Essa funcionalidade não é explorada nos trabalhos anteriores, que se concentram mais na modelagem do problema do que na entrega de ferramentas práticas para tomada de decisão pedagógica.

Por fim, o trabalho de Franqueira *et al*. (2024), embora contribua com reflexões sobre o uso da IA na educação, tem caráter mais teórico e exploratório. Já o trabalho apresentado neste artigo se caracteriza como uma pesquisa aplicada, com foco na implementação de soluções concretas para a identificação precoce de alunos em risco de evasão e/ou reprovação e na promoção de intervenções pedagógicas baseadas em evidências.

SOLUÇÃO IMPLEMENTADA

Este trabalho consistiu na aplicação de técnicas de *Machine Learning* para identificar estudantes em risco de reprovação, com base em uma base pública educacional estruturada. Diferentemente do planejamento inicial, a solução foi implementada a partir do *dataset Student Performance*, disponível no *UCI Machine Learning Repository* (UCI, 2026), e não mais com os microdados do INEP. Essa mudança metodológica ocorreu porque a base da UCI se mostrou mais adequada ao escopo do projeto nesta etapa, facilitando a experimentação, a reprodutibilidade e a comparação entre modelos. O conjunto de dados empregado foi o *student-mat.csv*, correspondente ao desempenho de estudantes do ensino secundário português na disciplina de Matemática. O repositório informa que o *dataset* foi coletado em duas escolas portuguesas e reúne atributos demográficos, sociais, familiares e escolares, além de variáveis de desempenho acadêmico. A base possui 395 registros no arquivo de Matemática, enquanto o repositório completo disponibiliza 649 registros somando também a disciplina de Língua Portuguesa; no total, são 33 atributos, incluindo variáveis como gênero, idade, escolaridade dos pais, tempo de estudo, histórico de reprovações, apoio escolar, faltas e notas parciais e finais. O próprio UCI destaca que as variáveis G1 e G2 possuem forte correlação com a nota final G3, motivo pelo qual sua utilização deve ser cuidadosamente tratada em tarefas preditivas.

É importante destacar que este trabalho não visou ao desenvolvimento de novos modelos matemáticos ou algoritmos inéditos, mas sim à aplicação prática de algoritmos já consolidados na literatura científica e disponíveis em bibliotecas computacionais. A intenção foi avaliar a eficiência dessas técnicas no contexto educacional, explorando seus resultados, limitações e aplicabilidade na identificação precoce de estudantes em risco.

Como parte da solução, desenvolveu-se um *dashboard* interativo para visualização dos resultados, permitindo que docentes e gestores acompanhem indicadores-chave relacionados ao risco de reprovação. O painel foi projetado para traduzir análises complexas em informações acessíveis e de fácil compreensão, com o intuito de apoiar decisões pedagógicas mais assertivas. Conforme destacam Zhuang, Concannon e Manley (2022), *dashboards* analíticos favorecem a identificação ágil de padrões, tendências e anomalias, além de apoiarem o monitoramento contínuo de indicadores de desempenho. No caso deste trabalho, os principais KPIs (*Key Performance Indicators*) contemplam a quantidade de estudantes classificados por nível de risco, o total de reprovações reais e previstas, a distribuição de alunos com baixo, médio e alto risco e a comparação entre o comportamento dos modelos analisados.

A avaliação dos modelos preditivos foi realizada por meio de métricas consolidadas, tais como acurácia, precisão, *recall* e F1-Score, assegurando uma análise rigorosa da performance dos algoritmos utilizados. Essas métricas são amplamente reconhecidas como fundamentais para validar a eficácia de modelos de classificação, especialmente em contextos com dados desbalanceados, como ocorre em problemas de previsão de reprovação escolar. Nesse sentido, a análise das matrizes de confusão também desempenhou papel importante, pois permitiu verificar não apenas os acertos globais, mas principalmente a capacidade dos modelos em identificar corretamente os estudantes reprovados.

Neste contexto, este trabalho caracteriza-se como uma pesquisa aplicada. Segundo Gil (2008), a pesquisa aplicada tem como principal objetivo gerar conhecimento voltado à solução de problemas práticos e específicos, atendendo aos interesses de uma determinada área de atuação. Ela se baseia em conhecimentos já existentes para promover intervenções diretas no mundo real, com foco em resultados concretos.

Na segunda etapa do projeto, referente ao desenvolvimento do *dashboard*, adotou-se a metodologia da Dissertação-Projeto, conforme proposto por Ribeiro e Zabadal (2010). Essa abordagem envolve a caracterização de um problema técnico, a demonstração de sua relevância e, posteriormente, o desenvolvimento de um programa, sistema ou protótipo que ofereça uma solução eficaz.

BASES DE DADOS E PRÉ-PROCESSAMENTO

O conjunto de dados empregado neste trabalho foi o *Student Performance Dataset*, disponibilizado no UCI *Machine Learning Repository* (UCI.EDU, 2026). Essa base reúne informações de estudantes do ensino secundário de Portugal e contempla variáveis relacionadas ao contexto familiar, apoio escolar, hábitos de estudo, número de faltas, relacionamento familiar, histórico de reprovações anteriores e notas obtidas ao longo do período letivo. Para o desenvolvimento do estudo, foi utilizado o arquivo *student-mat.csv*, referente à disciplina de Matemática, por apresentar estrutura adequada ao objetivo de prever a reprovação escolar com base em atributos educacionais, sociais e familiares.

A variável-alvo foi definida a partir da nota final G3, considerando-se como reprovado o estudante com valor inferior a 10. A Figura 2 apresenta um recorte da base original utilizada no trabalho, evidenciando parte dos atributos acadêmicos, sociais e familiares disponíveis no conjunto de dados.

Figura 2 - Recorte da base de dados *Student Performance Dataset* utilizada no desenvolvimento do trabalho, apresentando parte dos atributos acadêmicos, sociais e familiares dos estudantes. Fonte: UCI *Machine Learning Repository* (2026).

Após a obtenção da base, iniciou-se a etapa de pré-processamento dos dados com apoio das bibliotecas Pandas e *NumPy*. Nessa fase, foram realizadas atividades de inspeção inicial da estrutura do conjunto de dados, identificação dos tipos de variáveis e preparação das informações para a modelagem. Em seguida, as variáveis categóricas foram transformadas em atributos numéricos por meio da técnica de codificação binária com a função *get_dummies*, permitindo sua utilização pelos algoritmos de classificação.

Após a definição da variável-alvo, as colunas G1, G2 e G3 foram retiradas do conjunto de entrada dos modelos, concentrando a análise em atributos sociais, familiares e escolares anteriores ao resultado final do estudante. Na sequência, a base tratada foi dividida em dois subconjuntos, sendo 80% destinados ao treinamento e 20% ao teste, com aplicação de estratificação para manter a proporção entre aprovados e reprovados em ambas as partes da amostra. A Figura 3 apresenta um recorte da base após o pré-processamento, evidenciando a seleção de atributos numéricos e a transformação das variáveis categóricas.

Figura 3 – Recorte da base de dados após o pré-processamento, evidenciando a seleção de atributos numéricos e a transformação das variáveis categóricas para uso nos modelos preditivos. Fonte: Os autores (2026).

Como o conjunto de dados apresentava desbalanceamento entre as classes, aplicou-se a técnica SMOTE (*Synthetic Minority Over-sampling Technique*) sobre o conjunto de treinamento, com a finalidade de ampliar a representatividade da classe minoritária e melhorar a capacidade preditiva dos modelos em relação aos estudantes reprovados. Esse procedimento foi importante para tornar a etapa de modelagem mais adequada ao objetivo do trabalho, que é identificar corretamente alunos em situação de risco.

TECNOLOGIAS EMPREGADAS

A implementação da solução proposta utilizou um conjunto de tecnologias modernas e amplamente adotadas na área de ciência de dados e análise preditiva. A Figura 4 apresenta um resumo do fluxo empregado para o desenvolvimento do projeto.

Figura 4: Ilustração do fluxo sequencial para desenvolvimento do projeto, destacando as ações de cada etapa e as devidas tecnologias e ferramentas utilizadas. Fonte: Os autores (2026)

A Figura 4 apresenta o fluxo das etapas envolvidas no desenvolvimento do projeto. Ela mostra desde a coleta das informações até a implantação final em *dashboards* interativos, passando por fases como pré-processamento, modelagem, interpretação dos resultados e visualização. O objetivo é ilustrar de forma clara e organizada como as diferentes ações se conectam para transformar dados brutos em insights úteis e acessíveis.

As tecnologias selecionadas, de acordo com a Figura 4, incluem:

- **Linguagem de Programação *Python*:** com o uso das bibliotecas: *Pandas e* ***NumPy*** para manipulação e limpeza dos dados;
- ***Scikit-learn:*** biblioteca de aprendizado de máquina (Coelho et al., 2021) que oferece uma ampla gama de algoritmos e ferramentas para modelagem, validação e avaliação de desempenho. Os algoritmos utilizados incluem:
    - **Regressão Logística**: modelo estatístico de classificação binária que estima a probabilidade de ocorrência de um evento. É amplamente utilizado por sua interpretabilidade e eficiência em problemas lineares;
    - **XG Boost** (*eXxtreme Gradient Boosting*): técnica de *boosting* que constrói modelos sequenciais otimizando o erro residual. É reconhecida por sua alta performance,

capacidade de lidar com dados desbalanceados e controle refinado de regularização.

- ***SHAP*** para interpretabilidade dos modelos;
- **Ambiente de Desenvolvimento**: *Jupyter Notebook* para prototipagem e desenvolvimento;
- **Ferramenta de *Dashboard***: *Microsoft Power BI* (Microsoft.com, 2025) para criação do painel interativo, permitindo a visualização de indicadores de desempenho e alertas de risco.

Essa *stack* tecnológica foi escolhida por sua escalabilidade, suporte à reprodutibilidade e ampla documentação, facilitando a implementação e a validação dos modelos.

MODELAGEM COM *MACHINE LEARNING* - AVALIAÇÃO DOS MODELOS E ANÁLISE DAS MATRIZES DE CONFUSÃO

Após o treinamento dos modelos, realizou-se a etapa de avaliação de desempenho, com o objetivo de verificar a capacidade dos algoritmos em classificar corretamente os estudantes em situação de aprovação e reprovação. Para isso, foram utilizadas métricas tradicionais de classificação, tais como acurácia, precisão, *recall* e *F1-score*, complementadas pela análise das matrizes de confusão, que permitem observar de forma detalhada os acertos e erros cometidos pelos modelos.

A matriz de confusão é uma ferramenta importante na avaliação de modelos preditivos, pois apresenta a distribuição entre classificações corretas e incorretas, distinguindo os casos de alunos aprovados e reprovados (Scikit-Learn, 2026). No contexto deste trabalho, sua análise foi especialmente relevante, uma vez que o principal interesse não está apenas no acerto global, mas na capacidade de identificar corretamente os estudantes com maior risco de reprovação.

No modelo de Regressão Logística, os resultados mostraram desempenho satisfatório na identificação dos estudantes reprovados, apresentando melhor equilíbrio entre sensibilidade e taxa de erro quando comparado ao *XGBoost*. A matriz de confusão indicou que o modelo conseguiu classificar corretamente 36 estudantes aprovados e 18 estudantes reprovados, enquanto errou 17 casos de aprovados classificados como reprovados e 8 casos de reprovados classificados como aprovados, mostrado na Figura 5. Esse resultado evidencia que a Regressão Logística foi mais eficiente para o objetivo central do estudo, isto é, a identificação preventiva de estudantes em risco de reprovação.

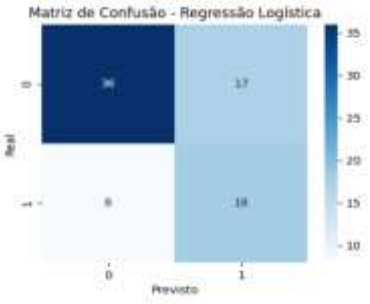

Figura 5 – Matriz de Confusão da Regressão Logística. Fonte: Os autores (2026)

A partir da análise da Figura 5, observa-se que a Regressão Logística apresentou maior capacidade de identificação da classe dos estudantes reprovados, reduzindo a quantidade de falsos negativos em relação ao modelo comparativo. Em aplicações educacionais, esse aspecto é particularmente importante, pois erros desse tipo representam estudantes em situação de risco que deixam de ser sinalizados pelo modelo.

No caso do *XGBoost*, os resultados também foram consistentes, porém ligeiramente inferiores ao modelo anterior quanto ao objetivo de identificar alunos reprovados. A matriz de confusão, apresentada na Figura 6, mostra que o modelo classificou corretamente 33 estudantes aprovados e 16 estudantes reprovados, ao passo que registrou 20 casos de aprovados classificados como reprovados e 10 casos de reprovados classificados como aprovados. Esses valores indicam que, embora o modelo tenha conseguido identificar parte significativa da classe minoritária, seu desempenho foi inferior ao da Regressão Logística no equilíbrio entre detecção de risco e quantidade de erros de classificação.

Figura 6 – Matriz de Confusão do modelo *XGBoost*. Fonte: Os autores (2026)

A análise da Figura 6 evidencia que o *XGBoost* apresentou comportamento mais conservador na distinção entre aprovados e reprovados, porém com menor eficiência na recuperação dos estudantes realmente em risco quando comparado à Regressão Logística. Assim, embora ambos os modelos tenham apresentado resultados relevantes após o

balanceamento da base, a Regressão Logística mostrou-se mais adequada para a finalidade deste trabalho.

De forma geral, os resultados obtidos nas matrizes de confusão mostram que a aplicação de estratégias de balanceamento contribuiu para melhorar a identificação da classe de reprovação, superando os resultados iniciais observados antes do tratamento do desbalanceamento. Além disso, a análise comparativa reforça que, para o contexto estudado, a Regressão Logística apresentou melhor desempenho prático como modelo de apoio à classificação do risco acadêmico.

Em razão disso, a Regressão Logística foi adotada como referência principal para a classificação final do nível de risco apresentada no *dashboard*, enquanto o *XGBoost* foi mantido como modelo comparativo e como apoio à interpretação das variáveis mais influentes.

INTERPRETABILIDADE DOS RESULTADOS

Além da avaliação quantitativa dos modelos, realizou-se a análise das variáveis mais relevantes para a classificação dos estudantes, com o objetivo de compreender quais atributos exerceram maior influência nas previsões geradas. Para isso, utilizou-se o modelo *XGBoost*, cuja estrutura permite extrair medidas de importância das variáveis com base na contribuição de cada atributo para as divisões realizadas ao longo do processo de treinamento.

Essa etapa foi importante porque permitiu ampliar a interpretação dos resultados para além das métricas de desempenho, possibilitando identificar quais fatores acadêmicos, sociais e familiares estiveram mais associados ao risco de reprovação. Em outras palavras, enquanto as matrizes de confusão mostraram o nível de acerto dos modelos, a análise de importância das variáveis permitiu compreender quais características dos estudantes mais impactaram as previsões.

Entre as variáveis de maior relevância identificadas pelo modelo (Figura 7), destacaram-se *Mjob_services* (que indica a atuação da mãe no setor de serviços); *reason_other* e *reason_home* (relacionadas ao motivo de escolha da escola, seja por outros fatores ou pela proximidade da residência); *failures* (representa o número de reprovações anteriores do estudante); *sex_M* (indicativo de estudantes do sexo masculino); *reason_reputation* (variável associada à escolha da escola por sua reputação); *famre* (expressa a qualidade das relações familiares); *Fjob_other* (referente à ocupação do pai em outras atividades não especificadas nas categorias principais); *absences* (correspondente ao número de faltas escolares); *higher_yes* (indica a intenção do estudante de prosseguir os estudos no ensino superior); *Fedu* (variável relacionada ao nível de escolaridade do pai); *Mjob_health* (identifica mães atuantes na área da saúde); s*chool_MS* (representa estudantes vinculados à escola Mousinho da Silveira); *famsize_LE3* (indica famílias com até três membros); e *schoolsup_yes* (referente ao recebimento de apoio escolar extra).

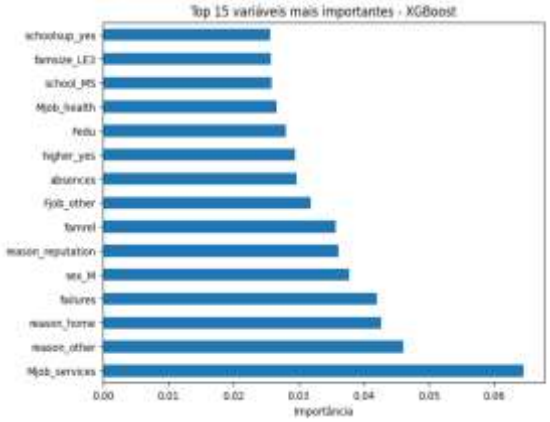

Figura 7 – Variáveis mais importantes identificadas pelo modelo *XGBoost*. Fonte: Os autores (2026)

A análise da Figura 7 permite observar que a variável *failures*, relacionada ao número de reprovações anteriores, aparece entre os atributos de maior importância. Esse resultado é coerente com o contexto educacional, pois indica que o histórico acadêmico do estudante constitui um forte indício de risco futuro. Da mesma forma, a variável *absences*, referente ao número de faltas, também se destacou, reforçando a relação entre frequência escolar e desempenho acadêmico.

Outro conjunto de variáveis relevantes está associado ao contexto familiar e social. A presença de atributos como *famrel* (qualidade do relacionamento familiar), *Fedu* (escolaridade familiar), *famsize_LE3* (tamanho da família) e ocupações dos pais, como *Mjob_services*, *Mjob_health* e *Fjob_other*, sugere que fatores externos ao ambiente estritamente escolar também influenciam o desempenho dos estudantes. Esses elementos podem representar aspectos socioeconômicos, disponibilidade de apoio familiar e condições gerais de acompanhamento da vida escolar.

Também se destacaram variáveis relacionadas à motivação e à escolha da escola, como *higher_yes*, *reason_home*, *reason_reputation* e *reason_other*. A variável *higher_yes*, por exemplo, indica a intenção do estudante de prosseguir os estudos em nível superior, o que pode estar associado a maior engajamento e perspectiva acadêmica. Já as variáveis relacionadas ao motivo de escolha da escola sugerem que fatores de proximidade, reputação institucional ou condições pessoais podem ter impacto indireto sobre o desempenho.

A variável *schoolsup_yes*, referente ao recebimento de apoio escolar adicional, também aparece entre as mais influentes. Sua presença pode ser interpretada de duas maneiras complementares: por um lado, pode indicar uma tentativa de reforço pedagógico; por outro, pode refletir a existência prévia de dificuldades acadêmicas por parte do estudante. Dessa forma, sua importância no modelo reforça a ideia de que a reprovação escolar é um fenômeno multifatorial, associado tanto ao histórico do aluno quanto às estratégias de suporte oferecidas.

De forma geral, os resultados obtidos nesta etapa indicam que o risco de reprovação escolar não pode ser explicado por um único atributo isolado. Ao contrário, ele resulta da interação entre histórico acadêmico, frequência, contexto familiar, aspectos motivacionais e condições de apoio escolar (Silva Júnior; Silva, 2024).

Embora o *XGBoost* não tenha sido o modelo final adotado para a classificação de risco no *dashboard* implementado, sua utilização nesta etapa mostrou-se valiosa para fins de interpretabilidade, permitindo organizar os principais fatores associados às previsões e enriquecer a análise educacional proposta neste trabalho.

VISUALIZAÇÃO DOS RESULTADOS E IMPLEMENTAÇÃO DO *DASHBOARD*

Após a interpretação dos modelos, os dados foram organizados para a etapa de visualização, utilizando a ferramenta *Microsoft Power BI*. Nessa fase, os resultados da modelagem foram estruturados em tabelas, indicadores e gráficos, com o objetivo de tornar as análises mais compreensíveis e acessíveis para professores e gestores.

Para isso, foi gerado um conjunto de dados contendo as variáveis dos estudantes, o resultado real, as previsões dos modelos, as probabilidades de reprovação e a classificação final do nível de risco. Como a Regressão Logística apresentou melhor desempenho na identificação dos alunos reprovados, a classificação final de risco exibida no painel foi baseada na probabilidade gerada por esse modelo.

A organização visual contemplou elementos como gráficos de barras, tabelas, cartões de indicadores e filtros interativos, permitindo observar o total de estudantes analisados, o número de reprovações reais e previstas, a distribuição dos alunos por nível de risco e os fatores associados ao desempenho acadêmico. Essa etapa foi essencial para transformar os resultados estatísticos e computacionais em informações visuais de apoio à tomada de decisão pedagógica.

A etapa final do trabalho consistiu na implantação do *dashboard* por meio da ferramenta *Microsoft Power BI*, com o objetivo de apresentar os resultados da modelagem preditiva de forma visual, interativa e acessível. O painel foi desenvolvido para apoiar professores e gestores na identificação de estudantes com maior probabilidade de reprovação.

No *dashboard* (apresentado na Figura 8), foram organizados filtros por gênero, nível de risco e situação de reprovação, além de indicadores gerais, como quantidade de alunos analisados, total previsto para reprovação e total previsto para aprovação. Também foi incluída a distribuição dos estudantes por nível de risco, permitindo identificar rapidamente os alunos que demandam maior atenção pedagógica.

Outro recurso importante é a tabela detalhada com informações individuais dos estudantes, como idade, faltas, reprovações anteriores, acesso à internet, relação familiar, nível de risco e previsão final. Essa visualização permite relacionar o resultado previsto às características de cada aluno, tornando a análise mais contextualizada.

Figura 8 – *Dashboard* analítico desenvolvido no Power BI para acompanhamento do risco de reprovação. Fonte: Os autores (2026)

De modo geral, o *dashboard* permite que os gestores educacionais e os professores visualizem, de forma rápida, os estudantes com maior risco de reprovação, auxiliando no planejamento de intervenções pedagógicas preventivas. Assim, a ferramenta atua como apoio à tomada de decisão, transformando os resultados dos modelos em informações mais claras e úteis para o contexto educacional.

CONSIDERAÇÕES FINAIS

Os resultados obtidos ao longo do desenvolvimento mostram que os objetivos centrais deste trabalho foram alcançados. Foi possível estruturar um *pipeline* completo de análise de dados educacionais, contemplando desde a coleta e preparação da base até a modelagem preditiva, a interpretação dos resultados e a visualização final em *dashboard*. Essa trajetória permitiu transformar dados educacionais em informações analíticas capazes de apoiar a identificação de estudantes com maior probabilidade de reprovação

A principal mudança em relação às versões anteriores do artigo foi a substituição das bases do INEP pela base *Student Performance*, da *UCI Machine*

Learning Repository (UCI.EDU, 2026), bem como a consolidação da Regressão Logística e do *XGBoost* como modelos principais do experimento. Essa redefinição tornou a execução mais viável no escopo do trabalho e permitiu a obtenção de resultados concretos, comparáveis e passíveis de interpretação.

Entre os modelos avaliados, a Regressão Logística apresentou melhor aderência ao objetivo do estudo, destacando-se na identificação dos estudantes reprovados e servindo como base para a classificação final do nível de risco representada no *dashboard* implementado. O *XGBoost*, por sua vez, mostrou-se relevante como modelo comparativo e, sobretudo, como apoio à análise das variáveis mais influentes, contribuindo para ampliar a compreensão dos fatores associados ao risco de reprovação.

De forma geral, os resultados reforçam que a reprovação escolar não pode ser compreendida a partir de um único fator isolado, mas sim como resultado da interação entre aspectos acadêmicos, familiares, sociais e comportamentais. Nesse sentido, o trabalho evidencia o potencial do uso de técnicas de *Machine Learning* no contexto educacional, especialmente como instrumento de apoio à tomada de decisão pedagógica e à implementação de intervenções preventivas mais direcionadas.

Como trabalhos futuros, recomenda-se ampliar os experimentos com outras bases educacionais, testar novos algoritmos de classificação, incorporar técnicas adicionais de explicabilidade dos modelos e validar o *dashboard* implementado com usuários do contexto escolar. Também se sugere, em uma etapa posterior, aplicar a metodologia em bases brasileiras mais amplas, inclusive retomando os microdados do INEP, a fim de avaliar a robustez da solução em contextos educacionais distintos e de maior escala.

REFERÊNCIAS

Alapydin, E. Introduction to Machine Learning. 4. ed. Cambridge: MIT Press, 2020.

Awad, M.; Khanna, R. Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers. Berkeley: Apress, 2015.

Benevento, M. A. O Uso de Algoritmos de Aprendizagem de Máquina para Prever o Desempenho do Aluno. FGV EAESP. Disponível em: https://repositorio.fgv.br/items/60046f6b-51c8-40b1-9fb7-7b3940b73c4a/full, 2024. Acesso em 20 out. 2025.

Caetano, M. M. O Uso de Técnicas de Aprendizado de Máquina na Predição de Desempenho Acadêmico de Alunos em Cursos Superiores. FACCAMP, 2016. Disponível em: https://cc.unifaccamp.edu.br/Dissertacoes/MaiteMarquesCaetano.pdf. Acesso em: 26 out. 2025.

Cardoso, M. M. R.; Lima, J. V. F. S.; Oliveira, M. H. V.; Paiva, R. O. A. O Uso de Learning Analytics em Ambientes de Aprendizagem Online: um Mapeamento Sistemático da Literatura. Revista Brasileira de Informática na Educação - RBIE, 30, 396-418, 2022. Disponível em: https://journals-sol.sbc.org.br/index.php/rbie/article/view/2664. Acesso em: 14 set. 2025.

Ceneviva, R.; Andrade, F. M.; Koslinski, M. C.; Núñez, C. P. Avaliação escolar e uso de dados e evidências na educação brasileira, 2022. In: Koga et al. (Orgs.). Políticas públicas e usos de evidências no Brasil: conceitos, métodos, contextos e práticas. Brasília: Instituto de Pesquisa Econômica Aplicada – IPEA. Cap. 27. Disponível em: https://portalantigo.ipea.gov.br/agencia/images/stories/PDFs/livros/livros/220412_lv_o_que_informa_miolo_cap27.pdf. Acesso em: 13 set. 2025.

Coelho, F. F.; Aorim, D. P. L.; Camargos, M. A. Analisando métodos de machine learning e avaliação do risco de crédito. Revista Gestão & Tecnologia, v. 21, n. 1, p. 89–116. DOI: 10.20397/2177-6652/2021.v21i1.2089, 2021. Disponível em: https://revistagt.fpl.emnuvens.com.br/get/article/view/2089. Acesso em: 13 nov. 2025.

Franqueira, A. S. *et al*. Inteligência Artificial na Avaliação de Desempenho Acadêmico: Desafios e Oportunidades no Ensino Médio. Revista FT, 2024. Disponível em: https://revistaft.com.br/inteligencia-artificial-na-avaliacao-de-desempenho-academico-desafios-e-oportunidades-no-ensino-medio/. Acesso em 15 out. 2025.

Gil, A. C. Métodos e técnicas de pesquisa social. 6. ed. São Paulo: Atlas, 2008.

Kabir, M. A.; Ahmed, M. R. Python for Data Analytics: A Systematic Literature Review of Tools, Techniques, and Applications. Academic Journal on Science, Technology, Engineering & Mathematics Education, v. 4, n. 04, 2024. Disponível em: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5051680. Acesso em: 13 set. 2025.

Lorenzi, F.; Silveira, S. R. Desenvolvimento de Sistemas de Informação Inteligentes. Porto Alegre: Editora do UniRitter, 2011.

Manzanares, M. C. S.; Sánchez, R. M.; González, A. A.; Llmazares, M. C. E.; Dios, M. A. Q. Detection of at-risk students with Learning Analytics Techniques. European Journal of Investigation in Health, Psychology and Education, Basel, v. 8, n. 3, p. 129-144, 2018. Disponível em: https://www.mdpi.com/2254-9625/8/3/129. Acesso em: 13 set. 2025.

Microsoft.com. Power BI: visualização de dados. Disponível em: https://www.microsoft.com/pt-

br/power-platform/products/power-bi. Acesso em: 10 nov. 2025.

Mitchell, T. M. Machine Learning. New York: McGraw-Hill, 1997.

Penteado, B. L.; Bittencourt, I. I.; Isotani, S. Modelo de Referência para Dados Abertos Educacionais em Nível Macro. Anais do XXX SBIE - Simpósio Brasileiro de Informática na Educação, 2019. Disponível em: https://repositorio.usp.br/directbitstream/392cd836-9b0b-46bd-be6a-a92e9896c980/2975644.pdf. Acesso em: 27 out. 2025.

Possato, A. B.; Ferla, T.; CURTULO, J. P.; GUIMARÃES, J. C.; CAMPOS, V.; SILVA, D. R. (2025). Learning Analytics (LA) e Educação 5.0: Convergências e Desafios. Revista Políticas Públicas & Cidades. Disponível em: https://journalppc.com/RPPC/article/view/1410. Acesso em: 13 set. 2025.

Ribeiro, V. G.; Zabadal, J. R. Pesquisa em Computação. Porto Alegre: UniRitter, 2010.

RudderStack.com. Data Analytics Processes. Disponível em: https://www.rudderstack.com/learn/data-analytics/data-analytics-processes/. Acesso em: 26 set. 2025.

Scikit-Learn. *Confusion_matrix*. Scikit-learn Documentation, [S. l.], 2026. Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html. Acesso em: 4 jun. 2026.

Silva Júnior, T. M.; Silva, J. E. Aplicação de um modelo de machine learning para prever a evasão escolar. Revista E&S Estratégias e Soluções. 2024. Disponível em: https://revistaes.com.br/artigos/aplicacao-de-um-modelo-de-machine-learning-para-prever-a-evasao-escolar. Acesso em: 27 fev. 2026.

Silveira, S. R.; Vit, A. R. D. Inteligência Artificial: os computadores podem ser criativos? In: Inteligência Artificial e Propriedade Intelectual. FERNANDES, M. S.; CALDEIRA, C. M. G. (Org). Rio de Janeiro: GZ Editora, 2023.

UCI.EDU. UCI Machine Learning Repository. Disponível em: https://archive.ics.uci.edu/. Acesso em: 13 maio 2026.

Zhuang, M.; Concannon, D.; Manley, E. A framework for evaluating dashboards in healthcare. IEEE Transactions on Visualization and Computer Graphics, 28 (4). pp. 1715-1731. ISSN 1077-2626, 2022. Disponível em: https://pubmed.ncbi.nlm.nih.gov/35213306/. Acesso em: 27 out. 2025.