# Big Data na Educação: Predição de Evasão Estudantil com Aprendizado de Máquina

**Filipe Dwan Pereira[1], Rafael Mello[2,3], Emanuel Marques Queiroga[3], Igor Padilha[1,3], George Zambonin[1,2,3], Renan Vital[3,4], Bruna de Oliveira Cassetari[3,4], Mario Antonio Pessoa Santos[3,4], Tyagi Lima[3,4]**

[1]Universidade Federal de Roraima (UFRR), RR, Brasil

[2]Centro de Estudos Avançados de Recife (CESAR), PE, Brasil

[3]Cogna Educacional S.A., SP, Brasil

[4]Universidade Anhanguera Unopar, PR, Brasil

{filipe.dwan}@ufrr.br

***Abstract.*** *This work presents a methodology for dropout prediction using big data (N = 466,387) from an AWS data lake environment. Model training is conducted automatically using the AWS SageMaker Autopilot framework, which is less dependent on data scientists and AI specialists. As a result, models were generated for different stages of the semester, achieving an accuracy of up to 83%. The methodology offers a replicable framework for educational institutions, filling a literature gap on the practical application of ETL and AutoML in educational contexts and providing a basis for the development of decision support systems that can lead to effective and personalized interventions.*

***Resumo.*** *Este trabalho apresenta uma metodologia para a predição da evasão, utilizando dados de big data (N = 466387) de um ambiente de data lake na AWS. O treinamento dos modelos é realizado de forma automatizada utilizando o framework AWS SageMaker Autopilot, menos dependente de cientista de dados e especialistas em IA. Como resultado, foram gerados modelos para diferentes estágios do semestre, alcançando uma acurácia de até 83%. A metodologia oferece um framework replicável para instituições de ensino, preenchendo uma lacuna na literatura sobre a aplicação prática de ETL e AutoML em contextos educacionais e fornecendo uma base para o desenvolvimento de sistemas de apoio à decisão que podem levar a intervenções eficazes e personalizadas.*

## 1. Introdução

A evasão acomete instituições de ensino em diversos níveis, desde o ensino superior tradicional até plataformas de Cursos Online Abertos e Massivos (MOOCs), acarretando em perdas financeiras e sociais [Nagy and Molontay 2024, Rabelo and Zárate 2025]. Este desafio impõe custos não apenas às instituições, que perdem investimentos e recursos, mas também aos próprios estudantes, que podem enfrentar desmotivação e interrupção de suas trajetórias formativas, e à sociedade, que deixa de se beneficiar do potencial de profissionais qualificados [Alghamdi et al. 2025].

A mitigação da evasão passa pela capacidade de identificar os alunos em risco de abandono antecipadamente [Alamri et al. 2019, Pereira et al. 2019, Talebi et al. 2024].

DOI: 10.5753/sbie.2025.12943 

Tal identificação é fundamental, pois permite a implementação de intervenções pedagógicas e de apoio de forma proativa e direcionada, aumentando consideravelmente as chances de retenção e, consequentemente, de sucesso acadêmico dos discentes [Alamri et al. 2019, Talebi et al. 2024]. Neste contexto, o Aprendizado de Máquina (ML) pode ser empregado para analisar conjuntos de dados educacionais e desenvolver modelos preditivos robustos que auxiliam na detecção desses alunos [Alamri et al. 2019, Pereira et al. 2019, Talebi et al. 2024, Alghamdi et al. 2025].

A literatura existente sobre predição de evasão, embora vasta, frequentemente apresenta lacunas no que tange ao detalhamento de pipelines de extração e transformação de dados de sistemas institucionais complexos do mundo real [Alghamdi et al. 2025, Sosa-Alonso et al. 2025]. Muitos estudos utilizam conjuntos de dados públicos ou simplificados, que não refletem os desafios de integração e a heterogeneidade inerentes aos *data warehouses* educacionais de grandes instituições [Alghamdi et al. 2025, Radovanović et al. 2021]. Adicionalmente, enquanto diversas pesquisas analisam logs de atividades de LMS, poucas oferecem uma análise granular de múltiplos tipos de atividades acadêmicas formais (como quizzes, provas, presenças e tarefas) dentro de disciplinas específicas e para um período letivo definido, o que pode limitar a profundidade dos *insights* sobre o desempenho do aluno [Alghamdi et al. 2025]. Outra lacuna reside na transição da predição para a acionabilidade; embora modelos preditivos sejam desenvolvidos, a criação de *features* que sejam intrinsecamente interpretáveis e que possam informar intervenções pedagógicas específicas ainda é um campo com espaço para desenvolvimento [Nagy and Molontay 2024]. De fato, é fundamental utilizar características comportamentais em modelos preditivos, uma vez que esses comportamentos podem ser modificados pelos próprios alunos com o apoio de incentivos promovidos por partes interessadas, como professores, coordenadores e demais agentes educacionais.

Aplicações de *big data* são caracterizadas por lidarem com conjuntos de dados grandes e complexos, que não podem ser facilmente gerenciados ou analisados com ferramentas tradicionais [Patgiri et al. 2023]. Na educação, essas aplicações envolvem dados com alto volume, velocidade e variedade, podendo incluir dados estruturados, semiestruturados e não estruturados [Stojanov and Daniel 2024]. Além disso, considerações sobre a veracidade (confiabilidade dos dados) e o valor (potencial de geração de *insights*) são fundamentais [Stojanov and Daniel 2024]. Para lidar com esses desafios, são necessárias tecnologias especializadas, como *frameworks* de computação distribuída, sistemas de armazenamento escaláveis e ferramentas analíticas sofisticadas [Sosa-Alonso et al. 2025]. No contexto educacional, o uso de *big data* permite a construção de bases robustas e representativas do comportamento discente, possibilitando análises preditivas precisas e acionáveis [Patgiri et al. 2023, Sosa-Alonso et al. 2025].

Este trabalho propõe uma metodologia para a predição da evasão estudantil. A abordagem inicia-se com a extração sistemática de um conjunto de *big data* (N = 466.387), baseado em comportamentos de alunos de diversas universidades de cursos à distância. Segue-se uma etapa de engenharia de atributos, em que dados comportamentais de diversas atividades acadêmicas são agregados. O objetivo é a construção de um *dataset* com dados comportamentais para o treinamento de modelos de ML, utilizando o *framework* Aprendizado de Máquina Automatizado (AutoML) para *big data*, que realize a seleção e ajuste automatizado de modelos, além de permitir avaliar a importância das variáveis do modelo. A principal contribuição reside no detalhamento transparente e re-

plicável deste pipeline de *big data*, oferecendo uma solução robusta para a transformação de dados brutos em variáveis comportamentais, gerando assim uma base sólida para modelos de predição de evasão acionáveis e prontos para o uso de IA explicável. Adicionalmente, o processo inerentemente constrói e destaca *features* baseadas no engajamento e desempenho do aluno em diversas tarefas, oferecendo *dataset* sobre potenciais preditores chave da evasão que podem ser explorados para o desenvolvimento de intervenções pedagógicas personalizadas.

## 2. Trabalhos Relacionados

A predição da evasão estudantil tem sido um foco de intensa pesquisa na área de Informática na Educação, explorando diversas fontes de dados e técnicas de Aprendizado de Máquina [Alghamdi et al. 2025]. O trabalho de Santos et al. [Santos et al. 2021], por exemplo, investigou se era possível prever a evasão utilizando exclusivamente o desempenho dos alunos nas disciplinas cursadas em uma universidade brasileira. Outros estudos [Carvalho et al. 2024, Silva et al. 2024] corroboram a importância de dados acadêmicos, mas frequentemente os combinam com fatores socioeconômicos e demográficos para enriquecer os modelos. Uma revisão sistemática realizada por Alves et al. [Alves et al. 2024] confirmou que a maioria dos modelos preditivos na literatura se baseia em uma combinação de dados demográficos e de desempenho. Uma lacuna notável, no entanto, é que muitos desses estudos operam com conjuntos de dados de escopo limitado ou não abordam explicitamente os desafios de escalabilidade inerentes ao processamento de grandes volumes de dados de sistemas institucionais complexos [Alghamdi et al. 2025]. Em contraste, nosso trabalho enfrenta este desafio ao aplicar a metodologia em um ambiente de *big data* (N = 466.387), detalhando um pipeline de Extração, Transformação e Carga (ETL) robusto e projetado para data lakes na Amazon Web Services (AWS), um cenário prático e de grande escala. Estudos identificam essa questão como uma lacuna prática ainda pouco explorada na literatura [Alghamdi et al. 2025].

A engenharia de atributos é outra etapa crucial, com abordagens variadas na literatura. Estudos como o de Carmo et al. [Êrica Carmo et al. 2022] focam na identificação de trajetórias de aprendizagem, analisando o percurso dos alunos ao longo de múltiplos semestres para encontrar padrões associados à evasão. Outros, como o de Andrade et al. [Andrade et al. 2024], utilizam dados de interação em Sistemas de Gestão de Aprendizagem (LMS) para alimentar sistemas de recomendação que visam mitigar a evasão. Embora essas abordagens sejam valiosas, elas nem sempre se concentram exclusivamente em variáveis que podem ser diretamente influenciadas por intervenções pedagógicas. Nosso estudo, por outro lado, foca na criação de variáveis comportamentais granulares. Ao contrário de dados demográficos estáticos, os atributos que construímos — como a proporção de quizzes concluídos ou o desempenho agregado em provas — são modificáveis pelos próprios alunos. Isso possibilita que intervenções sejam diretas e compreensíveis, concentrando os esforços em aspectos que os estudantes podem, com apoio institucional, efetivamente melhorar.

No que tange aos modelos de aprendizado de máquina, a literatura explora desde algoritmos clássicos, como Árvores de Decisão [Santos et al. 2021], até *ensembles* robustos, como *Random Forest* e o *XGboost*, que são frequentemente aplicados devido à sua alta performance preditiva [Alves et al. 2024, Nagy and Molontay 2024]. A maioria desses trabalhos, no entanto, pressupõe um processo manual de seleção, treinamento e

otimização de modelos, que exige tempo e conhecimento especializado, representando uma barreira para a adoção em larga escala por muitas instituições de ensino. Nossa metodologia aborda essa lacuna prática ao empregar AutoML através do *framework* AWS SageMaker Autopilot. A utilização de AutoML automatiza o pipeline de modelagem, tornando a criação de modelos preditivos de alta performance mais escalável e acessível, mesmo para equipes menores e sem especialização profunda em ciência de dados. Esta abordagem representa um caminho prático para a operacionalização da predição de evasão em escala institucional.

Finalmente, a acurácia preditiva por si só é insuficiente se as decisões do modelo não forem transparentes e justas. A literatura recente tem enfatizado a importância da IA Explicável (XAI) para garantir a confiança e a justiça algorítmica em sistemas educacionais [Carvalho et al. 2024, Silva et al. 2024]. Estudos como o de Silva et al. (2024) [Silva et al. 2024] comparam métodos de XAI (SHAP, LIME, ANCHOR), enquanto Nagy e Molontay (2024) [Nagy and Molontay 2024] defendem o uso dessas ferramentas para apoiar intervenções personalizadas. Uma barreira para uma XAI eficaz é a complexidade das *features* usadas nos modelos. Nossa metodologia contribui diretamente para essa área ao construir um conjunto de dados com variáveis comportamentais que são inerentemente mais interpretáveis do que *features* latentes de modelos complexos. Ao criar preditores que representam ações e desempenhos, nosso trabalho estabelece uma base sólida para a aplicação futura de técnicas de XAI, permitindo não apenas prever a evasão, mas também compreender suas causas e guiar ações pedagógicas.

## 3. Metodologia

A metodologia para predição de evasão foi dividida em três etapas principais. A primeira etapa consiste na descrição da fonte de dados utilizada. A segunda detalha o processo de engenharia e a definição formal dos atributos. Por fim, a terceira etapa descreve a preparação do *dataset* para a modelagem com AutoML.

### 3.1. Fonte e Descrição dos Dados

Os dados para este estudo foram extraídos de um data lake corporativo, hospedado na nuvem da AWS. As informações foram consolidadas a partir de múltiplas fontes de dados institucionais. Estas fontes incluem registros cadastrais dos alunos e seu histórico de rematrículas para um determinado período letivo. Também foram utilizados dados detalhados sobre o desempenho acadêmico e o engajamento dos estudantes em seus cursos e disciplinas. O conjunto de dados bruto continha 466.387 registros (53% de evasão e 47% de não evasão).

A variável alvo do modelo preditivo foi definida para representar a evasão. Um aluno foi classificado como evadido (valor 1) se não efetuou a rematrícula para o semestre seguinte ao período de análise. Caso contrário, se o aluno realizou a rematrícula, foi classificado como não evadido (valor 0). Os dados comportamentais dos alunos foram coletados no semestre 2024-2. A informação se o aluno evadiu foi consilidade em 2025-1.

### 3.2. Engenharia e Definição de Atributos

As principais definições matemáticas e os atributos gerados são descritos a seguir. A Tabela 1 apresenta uma descrição detalhada dos atributos utilizados na modelagem.

Para uma dada atividade acadêmica $a$ de um aluno, com nota bruta $g_a$ e nota máxima possível $g_{max,a}$, a nota normalizada $G_{norm,a}$ é calculada pela Equação 1. Esta normalização permite a comparação de desempenho entre atividades com diferentes pesos e escalas.

$$G_{norm,a} = \frac{g_a}{g_{max,a}}, \quad \text{se } g_{max,a} > 0 \tag{1}$$

Para um conjunto de atividades $A_T$ de um determinado tipo $T$ (e.g., QUIZ, avaliação) dentro de uma disciplina, a contagem de atividades concluídas ($C_{done,T}$) e a proporção de atividades concluídas ($P_{done,T}$) são definidas pelas Equações 2 e 3, respectivamente.

$$C_{done,T} = \sum_{a \in A_T} \mathbb{I}(g_a > 0) \tag{2}$$

$$P_{done,T} = \frac{C_{done,T}}{|A_T|}, \quad \text{se } |A_T| > 0 \tag{3}$$

onde $\mathbb{I}(\cdot)$ é a função indicadora, que retorna 1 se a condição for verdadeira e 0 caso contrário. O atributo $P_{done,T}$ é um indicador do engajamento e da consistência do aluno.

Adicionalmente, para o conjunto de atividades concluídas $A'_T \subseteq A_T$ onde a nota $g_a > 0$, são calculadas métricas agregadas das notas normalizadas. Estas incluem o valor mínimo, máximo e a soma acumulada ($\sum G_{norm,a}$), fornecendo uma visão multifacetada do desempenho do aluno.

**Tabela 1. Descrição dos principais atributos gerados para a modelagem.**

| Nome do Atributo | Descrição |
|---|---|
| Duração da Disciplina | Duração total da disciplina em dias, calculada como a diferença entre a data de fim e a data de início. |
| Nota Final na Disciplina | Nota final consolidada obtida pelo aluno na disciplina. Principal indicador de desempenho sumativo. |
| Disciplina Eletiva | Variável binária (0/1) que indica se a disciplina é de caráter eletivo. |
| Matrícula Tardia | Variável binária (0/1) que indica se a matrícula do aluno na disciplina foi realizada tardiamente. |
| Contagem Total de Quizzes | Número total de quizzes ou questionários disponíveis na disciplina. |
| Proporção de Quizzes Concluídos | Proporção de quizzes que o aluno efetivamente realizou em relação ao total disponível ($P_{done,T}$). |
| Estatísticas de Notas de Quizzes | Métricas de nota normalizada (mínima, máxima, acumulada) para os quizzes realizados. |
| Contagem Total de Presenças | Número total de atividades de presença ou de acompanhamento de frequência. |
| Proporção de Presenças | Proporção de presenças registradas ou atividades de frequência concluídas. |
| Estatísticas de Presença | Estatísticas das "notas"de presença, que podem representar a porcentagem de frequência. |
| Contagem Total de Provas | Número total de provas ou exames formais na disciplina. |
| Proporção de Provas Concluídas | Proporção de provas que o aluno realizou. |
| Estatísticas de Notas de Provas | Métricas de nota normalizada (mínima, máxima, acumulada) obtidas nas provas realizadas. |

### 3.3. Preparação do *Dataset* para Modelagem

O processamento dos dados resultou em um *dataset* estruturado. Cada linha representa o desempenho de um aluno em uma disciplina específica. Para a modelagem, o *dataset* foi segmentado com base na ordem sequencial em que as disciplinas foram cursadas durante o semestre. Esta abordagem permitiu o treinamento de modelos distintos para diferentes estágios do período letivo.

Para cada segmento de disciplina, um subconjunto de dados foi preparado. A variável alvo, indicadora da evasão, foi posicionada como a primeira coluna do conjunto de dados. Esta formatação é um requisito para a plataforma AWS SageMaker Autopilot, utilizada para a modelagem. As demais colunas continham os atributos preditivos detalhados na Tabela 1. Um tratamento de dados nulos foi aplicado, preenchendo valores numéricos ausentes com zero, para garantir a consistência do *dataset* de entrada para o AutoML.

### 3.4. Treinamento de Modelos com AWS SageMaker Autopilot

A etapa de modelagem foi conduzida utilizando o *framework* AWS SageMaker Autopilot, uma plataforma de AutoML. Para cada ordem sequencial de disciplina, um experimento Autopilot independente foi configurado e executado. A configuração do experimento foi padronizada para garantir a comparabilidade entre os modelos.

O problema foi definido como *BinaryClassification* (Classificação Binária), com a variável alvo sendo o indicador de evasão. A métrica objetivo para a otimização dos modelos foi o *F1-Score*, para lidar com conjuntos de dados potencialmente desbalanceados. Cada experimento do Autopilot foi configurado para explorar um máximo de cinco modelos candidatos. O tempo de execução total para cada experimento foi limitado a duas horas, com um tempo máximo de uma hora para o treinamento de cada modelo candidato individual. A ferramenta automaticamente pré-processa os dados, seleciona algoritmos apropriados (como *XGBoost*, *Random Forest*, entre outros) e otimiza seus hiperparâmetros. O resultado final deste processo é um conjunto de modelos treinados e otimizados, cada um especializado em prever a evasão com base no comportamento do aluno em uma disciplina de uma determinada ordem sequencial.

Para o treinamento, os dados foram separados em treino, validação e teste, sendo 60%, 10% e 30% para cada conjunto, respectivamente.

## 4. Resultados

Para avaliar a metodologia proposta, foram treinados modelos de classificação para cada uma das quatro primeiras ordens sequenciais de disciplina. A Tabela 2 resume o desempenho dos modelos finais selecionados pelo AWS Autopilot para cada ordem, avaliados nos seus respectivos conjuntos de teste. As métricas apresentadas incluem acurácia, precisão, recall e F1-score.

**Tabela 2. Resumo dos Resultados de Classificação por Ordem da Disciplina.**

| Métrica | Ordem 1 | Ordem 2 | Ordem 3 | Ordem 4 |
|---|---|---|---|---|
| **Acurácia** | 0.77 | 0.81 | 0.83 | 0.83 |
| **Precision** | | | | |
| Classe 0 (Não Evadido) | 0.78 | 0.79 | 0.82 | 0.85 |
| Classe 1 (Evadido) | 0.76 | 0.82 | 0.84 | 0.81 |
| **Recall** | | | | |
| Classe 0 (Não Evadido) | 0.71 | 0.82 | 0.85 | 0.85 |
| Classe 1 (Evadido) | 0.82 | 0.79 | 0.80 | 0.81 |
| **F1-Score** | | | | |
| Classe 0 (Não Evadido) | 0.74 | 0.81 | 0.83 | 0.85 |
| Classe 1 (Evadido) | 0.79 | 0.81 | 0.82 | 0.81 |

A análise de importância das *features*, gerada pelos modelos, revelou que a *Nota Final na Disciplina* consistentemente se destacou como o preditor mais influente em todas as ordens de disciplina. Para a disciplina de ordem 1, por exemplo, esta feature teve uma importância relativa de 0.39. Para as disciplinas de ordem 3 e 4, sua importância diminuiu

para aproximadamente 0.20, mas ainda permaneceu como o fator mais relevante. Outras *features*, como as relacionadas à performance agregada em provas e em atividades de presença, também demonstraram relevância considerável, especialmente nas disciplinas de ordem 3 e 4. A *Duração da Disciplina* também se mostrou um preditor consistentemente útil, embora com menor peso. A Tabela 3 detalha a importância relativa das principais *features* para cada modelo. A importância é calculada com a média dos gini-index das árvores de decisões que compõe o modelo ensemble XGboost, que foi o modelo encontrado com maior desempenho pelo AutoPilot da AWS. *Features* com importância menor que 0.01 não foram consideradas na Tabela 3.

**Tabela 3. Importância Relativa das Principais *Features* por Ordem da Disciplina.**

| Atributo | Ordem 1 | Ordem 2 | Ordem 3 | Ordem 4 |
|---|---|---|---|---|
| Nota Final na Disciplina | 0.3905 | 0.3339 | 0.2064 | 0.1990 |
| Soma das Notas Normalizadas de Provas | 0.0761 | 0.0850 | 0.1198 | 0.1148 |
| Maior Nota Normalizada de Provas | 0.0703 | 0.0544 | 0.1186 | 0.0874 |
| Soma das Notas Normalizadas de Presença | 0.0506 | 0.0537 | 0.0764 | 0.1367 |
| Menor Nota Normalizada de Provas | 0.0697 | 0.0581 | 0.1059 | 0.1091 |
| Proporção de Provas Concluídas | 0.0365 | 0.0698 | 0.0652 | 0.0443 |
| Soma das Notas Normalizadas de Quizzes | 0.0519 | 0.0538 | 0.0547 | 0.0331 |
| Duração da Disciplina | 0.0293 | 0.0666 | 0.0298 | 0.0176 |

## 5. Discussões

Os resultados obtidos demonstram a eficácia da metodologia de engenharia de atributos e do treinamento de múltiplos modelos para prever a evasão em diferentes estágios de um semestre letivo. Observa-se uma clara tendência de melhoria na performance preditiva à medida que avançamos na ordem das disciplinas. A acurácia geral aumenta de 77% para a primeira disciplina para 83% na terceira e quarta, indicando que o comportamento do aluno ao longo do semestre fornece sinais preditivos mais fortes. Este achado é consistente com a literatura, que sugere que a precisão da predição de evasão tende a aumentar conforme mais dados sobre o desempenho e engajamento do aluno se tornam disponíveis. Estudos como os de [Vaarma and Li 2024] e [Radovanović et al. 2021] corroboram que a performance dos modelos melhora com o tempo, e [Santos et al. 2021] observou uma tendência similar de aumento de acurácia ao gerar modelos para múltiplos semestres.

A análise de importância das *features* reforça a relevância central do desempenho acadêmico como principal indicador de risco de evasão. A variável *Nota Final na Disciplina* foi, de forma consistente, a mais importante em todos os modelos. Isso é intuitivo, pois a nota final reflete o sucesso consolidado do aluno em uma disciplina. Este resultado alinha-se diretamente com estudos como o de [Vaarma and Li 2024], que identificaram créditos acumulados e número de disciplinas reprovadas como os preditores mais importantes, sendo a nota final uma *proxy* direta desses indicadores. Da mesma forma, o trabalho de [Santos et al. 2021] utilizou o desempenho em disciplinas chave como os principais nós de decisão em suas árvores. A contribuição da nossa abordagem reside na granularidade, mostrando que não apenas o resultado final, mas também as atividades que o compõem (como as métricas agregadas de provas e presenças) são preditores significativos. Um insight particularmente interessante é a mudança na importância relativa

das *features*: a nota final, embora sempre relevante, perde um pouco de sua dominância para as *features* de presença nas disciplinas de ordem 3 e 4. Isso pode indicar que, para alunos que persistem até estágios mais avançados do semestre, o engajamento contínuo (representado pela frequência) torna-se um indicador de resiliência e probabilidade de permanência tão ou mais forte que as notas isoladamente.

A estratégia de treinar um modelo distinto para cada ordem sequencial de disciplina permite uma predição contextualizada e dinâmica. Para um aluno no início do semestre (cursando a disciplina de ordem 1), o modelo correspondente pode ser usado para uma previsão inicial. Conforme o semestre avança e o aluno cursa as disciplinas subsequentes (ordem 2, 3 e 4), modelos mais precisos e informados por um histórico comportamental mais longo podem ser aplicados. Esta abordagem de *predição em fases* é uma vantagem sobre metodologias que utilizam um único modelo estático, pois permite uma avaliação contínua e adaptativa do risco de evasão. É uma refinação da ideia apresentada por [Santos et al. 2021], que propôs um modelo por semestre; aqui, detalhamos a predição em etapas dentro de um mesmo semestre. Isso habilita as instituições a ajustarem suas estratégias de intervenção com base na evolução do aluno ao longo do período letivo, utilizando o modelo mais adequado para cada momento.

## 6. Conclusões, Limitações e Trabalhos Futuros

Este trabalho demonstrou a construção e validação de um *pipeline* para a predição da evasão estudantil, partindo da extração de *big data* de um data lake institucional até a implementação de modelos de aprendizado de máquina com o AWS SageMaker Autopilot. Como resultados, obtive-se modelos com acurácia de até 83%. Identificou-se a nota final na disciplina como o preditor mais significativo do risco de evasão. A principal contribuição deste estudo reside na apresentação de um *framework* transparente, escalável e replicável, que aborda os desafios práticos de se trabalhar com grandes volumes de dados em infraestruturas educacionais complexas. Ao detalhar um processo que integra ETL e AutoML, este trabalho preenche uma lacuna prática na literatura e oferece uma metodologia para instituições que buscam operacionalizar a análise preditiva para apoiar a tomada de decisões e o desenvolvimento de intervenções pedagógicas relacionadas à evasão.

Como limitações, a análise é retrospectiva e baseada em um *snapshot* de um período letivo específico. Embora a abordagem de modelar por ordem de disciplina ofereça uma visão dinâmica, ela não captura a evolução contínua do comportamento do aluno como uma série temporal completa. Além disso, a interpretabilidade dos modelos gerados pelo AutoML pode ser expandida com uso de técnicas de XAI como SHAP e análises contrafactuais, para aprofundar a compreensão das predições individuais. Como aponta a pesquisa de Nagy & Molontay (2024) [Nagy and Molontay 2024], a capacidade de explicar por que um aluno é considerado em risco é fundamental para planejar intervenções pedagógicas personalizadas. As *features* detalhadas criadas em nosso pipeline, como o desempenho em diferentes tipos de avaliação, fornecem uma excelente base para tais análises explicativas. Outra direção futura é a transição de um modelo de snapshot para um modelo longitudinal, que capture as tendências de engajamento e desempenho ao longo de múltiplos semestres, potencialmente utilizando arquiteturas de redes neurais como *Long-term Short-term* (LSTM).

## Referências

Alamri, A., Alshehri, M., Cristea, A., Pereira, F. D., Oliveira, E., Shi, L., and Stewart, C. (2019). Predicting moocs dropout using only two easily obtainable features from the first week's activities. In *Intelligent Tutoring Systems: 15th International Conference, ITS 2019, Kingston, Jamaica, June 3–7, 2019, Proceedings 15*, pages 163–173. Springer.

Alghamdi, S., Soh, B., and Li, A. (2025). A comprehensive review of dropout prediction methods based on multivariate analysed features of mooc platforms. *Multimodal Technologies and Interaction*, 9(1):3.

Alves, A., Inácio, C., Pozzebon, E., and Silva, J. (2024). Aspectos relevantes dos modelos preditivos de inteligência artificial no combate à evasão escolar em cursos de graduação: uma revisão sistemática. In *Anais do XXXV Simpósio Brasileiro de Informática na Educação*, pages 1503–1514, Porto Alegre, RS, Brasil. SBC.

Andrade, T., Almeida, C., Barbosa, J., and Rigo, S. (2024). Análise de desempenho dos alunos após a utilização do sistema de recomendação Éforo-sr para a mitigação de evasão e promoção da retenção. In *Anais do XXXV Simpósio Brasileiro de Informática na Educação*, pages 85–100, Porto Alegre, RS, Brasil. SBC.

Carvalho, C., Mattos, J., and Aguiar, M. (2024). Interpretabilidade e justiça algorítmica: Avançando na transparência de modelos preditivos de evasão escolar. In *Anais do XXXV Simpósio Brasileiro de Informática na Educação*, pages 1658–1673, Porto Alegre, RS, Brasil. SBC.

Nagy, M. and Molontay, R. (2024). Interpretable dropout prediction: towards xai-based personalized intervention. *International Journal of Artificial Intelligence in Education*, 34(2):274–300.

Patgiri, R., Deka, G. C., and Biswas, A. (2023). Principles of big graph: In-depth insight.

Pereira, F. D., Oliveira, E., Cristea, A., Fernandes, D., Silva, L., Aguiar, G., Alamri, A., and Alshehri, M. (2019). Early dropout prediction for programming courses supported by online judges. In *International conference on artificial intelligence in education*, pages 67–72. Springer.

Rabelo, A. M. and Zárate, L. E. (2025). A model for predicting dropout of higher education students. *Data Science and Management*, 8(1):72–85.

Radovanović, S., Delibašić, B., and Suknović, M. (2021). Predicting dropout in online learning environments. *Computer Science and Information Systems/ComSIS*, 18(3):957–978.

Santos, C. H., Martins, S., and Plastino, A. (2021). É possível prever evasão com base apenas no desempenho acadêmico? In *Anais do XXXII Simpósio Brasileiro de Informática na Educação*, pages 792–802, Porto Alegre, RS, Brasil. SBC.

Silva, F., Feitosa, R., Batista, L., and Santana, A. (2024). Análise comparativa de métodos de explicabilidade da inteligência artificial no cenário educacional: um estudo de caso sobre evasão. In *Anais do XXXV Simpósio Brasileiro de Informática na Educação*, pages 2968–2977, Porto Alegre, RS, Brasil. SBC.

Sosa-Alonso, J. J., López-Aguilar, D., Álvarez-Pérez, P. R., and González-Morales, O. (2025). Predicting university dropout: connecting big data and structural models. *Studies in Higher Education*, pages 1–18.

Stojanov, A. and Daniel, B. K. (2024). A decade of research into the application of big data and analytics in higher education: A systematic review of the literature. *Education and information technologies*, 29(5):5807–5831.

Talebi, K., Torabi, Z., and Daneshpour, N. (2024). Ensemble models based on cnn and lstm for dropout prediction in mooc. *Expert Systems with Applications*, 235:121187.

Vaarma, M. and Li, H. (2024). Predicting student dropouts with machine learning: An empirical study in finnish higher education. *Technology in Society*, 76:102474.

Êrica Carmo, Gasparini, I., and Oliveira, E. (2022). Identificação de trajetórias de aprendizagem em um curso de graduação e sua relação com a evasão escolar. In *Anais do XXXIII Simpósio Brasileiro de Informática na Educação*, pages 323–333, Porto Alegre, RS, Brasil. SBC.