Intelligent Information System for Analyzing Success Factors in Public Policies: A Case Study of the Pé-de-Meia Program

Lucas B. F. Praxedes[1], Sebastião E. A. Filho[1]

[1]Programa de Pós-Graduação em Ciência da Computação – Universidade Federal Rural do Semiárido (UFERSA) e Universidade do Estado do R. G. do Norte (UERN)
Av. Francisco Mota, 572, Bairro Costa e Silva – Mossoró – RN – Brazil

lucaspraxedes@alunos.ufersa.edu.br, sebastiaoalves@uern.br

Abstract. Research Context: *Public policies such as the 'Pé-de-Meia' program are crucial for reducing school dropout in Brazil, but their evaluation is complex due to vast amounts of data and inherent selection bias.* ***Scientific and/or Practical Problem:*** *There is a need for an approach that moves beyond simple causal inference to identify the contextual factors that enhance the program's success, providing actionable intelligence for public managers.* ***Proposed Solution and/or Analysis:*** *We developed an intelligent information system that integrates public data and uses a Machine Learning model (LightGBM) to identify the main predictors of success in student retention, complemented by a quasi-experimental analysis (PSM) to estimate impact.* ***Related IS Theory:*** *This work is grounded in the principles of Decision Support Systems (DSS), applying computational intelligence to transform raw governmental data into strategic insights for policy management.* ***Research Method:*** *The methodology involved a four-stage data processing pipeline: data collection (Censo Escolar, Portal da Transparência), feature engineering (creating a proxy for school enrollment dynamics), predictive modeling with LightGBM to rank success factors, and impact estimation using Propensity Score Matching.* ***Summary of Results:*** *The predictive analysis identified that, beyond the natural inertia of previous performance, the number of teachers and program intensity are the most significant actionable predictors of success. To ensure robustness, these factors were cross-validated using a Random Forest model. The quasi-experimental analysis, refined with strict common support and caliper matching, estimated a positive and statistically significant impact (ATT = +0.19 p.p., p-value < 0.001).* ***Contributions and Impact to IS area:*** *This study contributes a novel, rigorously validated method for public policy evaluation within the IS field. It demonstrates how a predictive system can provide immediate strategic insights, while advanced quasi-experimental techniques can detect causal effects even in early-stage implementations, offering a tangible artifact for evidence-based management.*

1. Introdução

A evasão escolar no ensino médio representa um dos desafios mais persistentes para o desenvolvimento social e econômico do Brasil. Dados da Pesquisa Nacional por Amostra de Domicílios (PNAD) Contínua indicam que, em 2023, cerca de 9 milhões de jovens na faixa de 18 a 29 anos não haviam completado o ensino médio [IBGE 2024]. O recorte etário é alarmante, pois evidencia um contingente significativo de indivíduos que já ultrapassaram a idade regular de conclusão, mas permanecem sem a certificação básica. O Censo Escolar do mesmo ano corrobora a gravidade do cenário, apontando

que esta etapa concentra a maior taxa de abandono da educação básica, alcançando 5,9%, um índice significativamente superior ao observado nos Anos Finais (3,8%) e Anos Iniciais (0,5%) do Ensino Fundamental [INEP 2024].

Além de limitar as oportunidades individuais dos jovens, o abandono dos estudos prolonga os ciclos de desigualdade e impacta negativamente a força de trabalho futura do país. Para mitigar este cenário, o Governo Federal instituiu o programa "Pé-de-Meia", uma política pública de grande escala que oferece um incentivo financeiro em formato de poupança para promover a permanência e a conclusão da educação básica por estudantes de escolas públicas [Brasil 2024].

Estudos técnicos recentes apontam que iniciativas dessa magnitude, com investimentos estimados na ordem de R$ 7 bilhões anuais, não visam apenas a retenção imediata, mas possuem o potencial de alterar as expectativas de longo prazo sobre pobreza e desigualdade no país [IPEA 2024]. Contudo, a relevância do programa e o potencial transformador da iniciativa estão condicionados à mudança na realidade dos estudantes de maior risco.

Destaca-se ainda que, em um contexto tão heterogêneo quanto o brasileiro, a avaliação de eficácia constitui um desafio informacional [Costa e Castanhar 2003]. A vasta quantidade de dados gerados pelo Censo Escolar e pelos portais de transparência oferece uma oportunidade sem precedentes para a análise, mas também exige abordagens computacionais sofisticadas para transformar esses dados brutos em conhecimento útil para os gestores públicos. A simples medição de impacto causal, neste contexto, é frequentemente interrompida por um forte viés de seleção, um desafio metodológico fundamental na avaliação de programas sociais [Shadish, Cook e Campbell 2002]. Como o programa é intencionalmente direcionado às populações mais vulneráveis, uma comparação direta com não-participantes é inviável, exigindo métodos quasi-experimentais, como o *Propensity Score Matching*, para tentar evitar tais vieses [Silva Junior e Gonçalves 2016].

Diante deste cenário, este trabalho propõe uma abordagem alternativa. Em vez de buscar isolar um único efeito causal, que é a tarefa central da econometria tradicional, a pesquisa foca em um objetivo de predição: identificar quais fatores estão mais associados à melhoria na permanência dos alunos. Essa distinção entre inferência causal e predição é fundamental, e o foco na predição possui um valor intrínseco para a gestão de políticas, permitindo a identificação de vetores de sucesso mesmo em cenários de alta complexidade [Mullainathan e Spiess 2017]. Para melhor abordar essa problemática, foi desenvolvido um sistema de análise que integra dados públicos massivos e utiliza um modelo de *Machine Learning* (*LightGBM*), algoritmo já empregado em estudos sobre evasão escolar por sua eficiência e capacidade preditiva [Silva 2024], para ranquear os principais fatores de sucesso.

A análise preditiva revelou que a intensidade do programa "Pé-de-Meia" no município aparece como o fator mais importante para explicar a variação positiva na taxa de progressão dos estudantes. Contudo, o sistema também identificou um segundo fator de importância comparável: a quantidade de docentes na escola. Este resultado sugere que o sucesso do incentivo financeiro pode ser potencializado quando se encontra um ambiente escolar com robustez em seu corpo docente, indicando uma poderosa

interação entre a política social e a estrutura educacional. Esse achado é consistente com pesquisas que demonstram como a alta rotatividade de professores prejudica o desempenho dos alunos, não apenas pela perda de capital humano, mas também pelo impacto negativo na colaboração e estabilidade do ambiente escolar [Ronfeldt, Loeb e Wyckoff 2013].

O restante deste artigo está organizado da seguinte forma: a Seção 2 apresenta a revisão de literatura em avaliação de políticas públicas e uso de IA na educação. A Seção 3 estabelece os conceitos fundamentais. A Seção 4 detalha a arquitetura do sistema de informação proposto, incluindo a coleta de dados, a engenharia de features e o modelo preditivo. A Seção 5 apresenta os resultados, incluindo os gráficos de insights gerados, e a Seção 6 discute as implicações desses achados para a gestão pública e conclui o trabalho.

2. Conceitos Fundamentais

Esta seção estabelece as bases teóricas que sustentam a pesquisa, conectando as áreas de Sistemas de Informação, *Machine Learning* e a avaliação de políticas públicas no contexto educacional.

2.1. Sistemas de Informação no Setor Público

A capacidade de transformar grandes volumes de dados em inteligência acionável é um pilar da gestão pública moderna, um campo historicamente fundamentado em sistemas de *Business Intelligence & Analytics* (BI&A) [Chen, Chiang e Storey 2012]. No entanto, a literatura recente aponta que o setor público vive uma mudança de paradigma: a evolução de meras ferramentas de automação para o conceito de "Governar com Inteligência" (*Governing with Intelligence*). Nesta nova abordagem, a Inteligência Artificial e a *Policy Informatics* são utilizadas não apenas para monitoramento descritivo, mas para sustentar inferências causais e simular os impactos de intervenções complexas [Yar et al. 2024].

Essa transição reforça a necessidade de uma Tomada de Decisão Baseada em Dados e Governança de Dados [Arduini et al. 2020, Janssen, van der Voort e Wahyudi 2017]. A governança no compartilhamento desses dados é, inclusive, objeto de políticas específicas no Brasil [Brasil 2019]. A aplicação integrada de BI&A e modelos preditivos modernos permite que gestores transitem de uma análise reativa para uma postura proativa, utilizando *Big Data* para antecipar cenários e otimizar a alocação de recursos [Pereira et al. 2018]. No cenário nacional atual, a comunidade de Sistemas de Informação tem priorizado o desenvolvimento de métodos automatizados para a extração eficiente de dados em portais de transparência, superando barreiras técnicas de acesso para permitir uma auditoria social mais efetiva [Martins e Silva 2024]. Este trabalho, portanto, desenvolve um artefato alinhado a esse estado da arte, transformando dados governamentais brutos em inteligência causal para a gestão educacional.

2.2 *Machine Learning* para Identificação de Fatores Relevantes

Machine Learning (ML), um subcampo da Inteligência Artificial, consiste em algoritmos que permitem a um sistema computacional aprender padrões a partir de

dados empíricos, sem serem explicitamente programados para a tarefa [Samuel 1959]. Modelos preditivos, como o *Light Gradient Boosting Machine* (*LightGBM*) utilizado neste estudo, são bastante poderosos para lidar com a complexidade e a não-linearidade presentes em dados sociais. O *LightGBM* é um algoritmo baseado em árvore de decisão que se destaca por sua eficiência e performance em dados tabulares [Ke et al. 2017]. Uma de suas funcionalidades mais importantes para a análise de políticas públicas é a capacidade de gerar uma métrica de importância de *features* (*feature importance*). Essa métrica, inerente a algoritmos baseados em árvores, quantifica a contribuição de cada variável de entrada (ex: infraestrutura da escola, intensidade do programa) para a redução da impureza ou do erro do modelo ao longo de suas ramificações. A escolha prioritária pelo *LightGBM* neste estudo justifica-se não apenas pela sua eficiência computacional em grandes volumes de dados, mas principalmente pela interpretabilidade proporcionada pela análise de importância de atributos (*Feature Importance*). Ao contrário de modelos "caixa-preta" (como Redes Neurais Profundas), métodos baseados em árvores permitem aos gestores públicos entenderem quais variáveis estão guiando a predição, característica essencial para a justificação de decisões governamentais.

2.3 Modelos baseados em *Gradient Boosting*

O *LightGBM* pertence a uma classe de algoritmos de *ensemble* conhecida como *boosting*. A ideia central do *boosting* é construir um modelo preditivo forte a partir da combinação sequencial de vários modelos "fracos" (geralmente árvores de decisão rasas). Diferentemente de outros métodos de *ensemble*, como o *Random Forest*, que constrói modelos em paralelo, o *boosting* é um processo iterativo: cada novo modelo é treinado para corrigir os erros cometidos pelo conjunto de modelos anterior [Freund e Schapire 1997]. O termo "*Gradient*" refere-se ao fato de que, em algoritmos como o *LightGBM*, essa correção de erros é otimizada através de um método de gradiente descendente sobre uma função de perda [Friedman 2001]. Essa abordagem permite que o modelo final se ajuste de forma muito precisa aos padrões complexos dos dados, sendo ideal para a tarefa de identificar os sutis fatores que influenciam o sucesso escolar.

2.4 Desafios na Avaliação de Impacto e o Viés de Seleção

A avaliação de impacto de programas sociais busca isolar o efeito causal de uma intervenção, ou seja, estimar como os resultados dos participantes teriam sido na ausência do programa, que é o contrafactual [Shadish, Cook e Campbell 2002]. Métodos quasi-experimentais, como o de Diferenças em Diferenças, são muito utilizados para este fim. Contudo, um desafio recorrente é o viés de seleção, que ocorre quando os grupos que recebem o tratamento não são aleatórios, mas sim selecionados com base em características que também influenciam o resultado [Shadish, Cook e Campbell 2002]. Este viés é tão forte que a intensidade do tratamento se confunde com a própria condição de vulnerabilidade, um problema tecnicamente conhecido como endogeneidade, que surge quando uma variável explicativa está correlacionada com o termo de erro do modelo [Wooldridge 2010], tornando a inferência causal direta metodologicamente instável com dados de curto prazo.

2.5 Utilização de Variáveis *Proxy* e Análise de Mudança

Devido a limitações na disponibilidade de dados longitudinais de alunos individuais nos arquivos agregados do Censo Escolar, já que não possuem identificadores únicos anônimos que permitam rastrear o mesmo estudante ao longo do tempo, a medição direta da permanência individual não era factível com a estrutura de dados pública disponível.

Para contornar essa limitação, empregamos o conceito de variável *proxy* para medir o efeito líquido de retenção e atração da escola. Utilizamos a "Taxa de Variação do Total de Matrículas no Ensino Médio", calculada pela razão entre o número total de alunos matriculados no ensino médio em um ano t e o número total de matrículas na mesma escola no ano $t - 1$. Em contextos de dados agregados, essa métrica é um proxy útil para a saúde institucional da escola [Crosling, Heagney e Thomas 2009]. Uma variação positiva indica que a escola reteve seus alunos antigos e/ou atraiu novos, refletindo indiretamente o sucesso das políticas de permanência.

Adicionalmente, a análise final do nosso sistema não foca nos valores absolutos dessa taxa, mas sim na sua mudança entre o período pré e pós-programa. Essa técnica, conhecida como *First-Differencing* (Primeiras Diferenças), é um método estatístico padrão em econometria de dados em painel, eficaz para eliminar efeitos fixos não observados e invariantes no tempo de cada unidade de análise (neste caso, cada escola), focando a análise naquilo que realmente importa: a melhoria (ou piora) na capacidade de retenção da escola ao longo do tempo [Wooldridge 2010].

3. Revisão de Literatura

A literatura acadêmica apresenta diversas abordagens para o estudo da evasão escolar e o impacto de políticas públicas educacionais. Por um lado, uma vertente significativa de trabalhos na área da Computação tem aplicado técnicas de *Machine Learning* para a predição do risco de abandono escolar. Estudos como os de Silva et al. (2022) e Aguilar et al. (2022) demonstram a eficácia de diversos algoritmos para identificar alunos em risco com base em seu histórico acadêmico e dados demográficos. A literatura mais recente de 2025 avança ao demonstrar que arquiteturas de Gradient Boosting (como o LightGBM) superam modelos clássicos na detecção precoce de evasão, especialmente quando combinadas com técnicas de over-sampling para corrigir o desbalanceamento das classes minoritárias [Wang e Yao 2025].

Uma revisão sistemática publicada no SBIE 2024 analisou modelos preditivos de evasão e corroborou que técnicas de ensemble, especificamente o Gradient Boosting, consolidaram-se como o estado da arte na área, superando modelos lineares tradicionais em métricas de acurácia e sensibilidade [Alves 2024].

Além disso, plataformas preditivas modernas têm integrado esses algoritmos para fornecer informações úteis em tempo real, permitindo intervenções pedagógicas antes que o abandono se concretize [Borges et al. 2025]. Por outro lado, a avaliação do impacto de programas de transferência de renda na educação, como o Bolsa Família, tem sido um tema recorrente nas Ciências Sociais e Econômicas. Trabalhos de alto impacto, como o de De Janvry et al. (2015), utilizaram métodos econométricos para estimar o efeito causal desses programas na frequência e no desempenho escolar.

Apesar desses avanços, Colpo et al. (2024) alertam, em análise publicada na Revista Brasileira de Informática na Educação, para uma lacuna crítica: a escassez de pesquisas de Mineração de Dados Educacionais em países em desenvolvimento que efetivamente integrem a predição técnica com a prática da gestão escolar.

O presente estudo se posiciona na interseção dessas duas áreas, mas com uma contribuição distinta. Enquanto os trabalhos anteriores focaram ou na predição de risco individual ou na inferência causal de programas já consolidados, nossa abordagem utiliza o poder preditivo do Machine Learning não como um fim em si mesmo, mas como uma ferramenta exploratória para identificar os fatores de sucesso associados a uma política pública em seu primeiro ano de implementação. Ao contornar os desafios metodológicos da inferência causal em um cenário de forte viés de seleção, este trabalho inova ao construir um sistema de informação que gera inteligência sobre quais características contextuais potencializam os resultados de um novo programa, oferecendo informações mais imediatas para a sua gestão e aprimoramento.

4. Metodologia

Para responder à pergunta de pesquisa, foi concebido e implementado um Sistema de Informação para Análise de Políticas Públicas. A metodologia foi estruturada em um *pipeline* de quatro etapas principais, conforme detalhado nesta seção: (1) Arquitetura do Sistema, (2) Coleta e Integração de Dados, (3) Engenharia de *Features* e Construção da Base Analítica, e (4) Modelo Preditivo para Identificação de Fatores de Sucesso.

4.1 Arquitetura do Sistema de Análise

O sistema foi concebido como uma ferramenta e também como um artefato analítico completo, projetado para operar como um *script* de processamento de dados escalável. O objetivo fundamental desta arquitetura é orquestrar a complexa transmutação de dados governamentais brutos, que é muitas vezes dispersos e de difícil interpretação, em um conjunto coeso de informações estratégicas, apresentados de forma visual e quantitativa para apoiar a tomada de decisão. A modularidade incluída no design permite que cada componente do fluxo, desde a extração de dados até a modelagem preditiva, possa ser atualizado ou substituído de forma independente, conferindo ao sistema uma expressiva capacidade de adaptação a futuras demandas analíticas ou à incorporação de novas fontes de informação.

A materialização dessa arquitetura, detalhada visualmente na Figura 1, segue uma lógica sequencial, onde cada etapa do processo depende do resultado validado da etapa anterior. Este fluxo encadeado, que parte da coleta e integração de fontes de dados heterogêneas, passando pela engenharia de *features* e culminando na modelagem e análise, estabelece uma linhagem de dados clara.

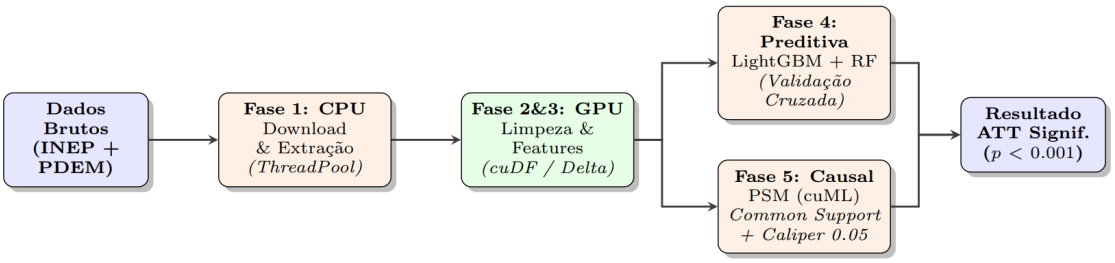

Figura 1. Arquitetura do Sistema de Informação para Análise de Políticas Públicas.

Tal estrutura é fundamental, pois garante dois pilares essenciais do rigor científico: a rastreabilidade, que permite a qualquer momento inspecionar a origem e as transformações aplicadas a um determinado resultado, e a reprodutibilidade, que assegura que a análise possa ser executada por outros pesquisadores para validação dos resultados, reforçando a transparência e também a confiabilidade de todo o estudo.

4.2 Coleta e Integração de Dados

A pesquisa utilizou exclusivamente fontes de dados públicos e abertos, garantindo a transparência e a possibilidade de replicação. As duas fontes primárias foram:

1. Censo Escolar da Educação Básica: Foram utilizados os microdados agregados por escola, disponibilizados anualmente pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). Os dados foram coletados para o período de 2022 a 2024, compreendendo os dois anos anteriores ao programa e o primeiro ano de sua implementação.
2. Dados do Programa "Pé-de-Meia": Foram utilizados os dados de pagamentos do programa, disponibilizados no Portal da Transparência do Governo Federal. Estes dados foram agregados por município para o ano de 2024.

A base analítica final consolidada abrangeu o conjunto de escolas públicas de ensino médio que apresentaram dados consistentes para os períodos analisados (2022-2024). Após o rigoroso processo de limpeza e integração, o dataset final conteve 7 variáveis explicativas (*features*), cobrindo dimensões de infraestrutura, corpo docente e indicadores socioeconômicos locais, compondo um painel suficiente para a aplicação dos algoritmos de *Machine Learning*.

4.3 Engenharia de *Features* e Construção da Base Analítica

Esta foi a etapa mais complexa, onde os dados brutos foram transformados em variáveis significativas para o modelo. O processo envolveu três cálculos principais.

Primeiramente, foi calculado o proxy de desempenho escolar, a "Taxa de Variação de Matrículas". Devido à ausência de rastreamento individual nos microdados agregados por escola, calculou-se a variação anual do total de matrículas no ensino médio (*QT_MAT_MED*). Especificamente, dividiu-se o número total de alunos matriculados no ensino médio em um ano t pelo número total de matrículas na mesma

escola no ano t−1. Essa métrica captura a dinâmica de fluxo escolar (entradas menos saídas), servindo como indicador indireto da eficácia da retenção escolar.

Em seguida, foi criada a variável-alvo do modelo (Delta), *DELTA_PROG_1_2*. Calculamos a média da Taxa de Variação para o período pré-programa (2022-2023) e para o período pós-programa (2024). A variável Delta foi então obtida subtraindo-se a média pré da média pós, isolando assim a mudança abrupta na dinâmica de matrículas atribuível ao período de implementação da política.

Por fim, calculou-se a dose do programa, que é a variável *INTENSIDADE_TRATAMENTO*, dividindo-se o número total de beneficiários do "Pé-de-Meia" em 2024 pelo número total de matrículas no ensino médio no mesmo município e ano.

4.4 Modelagem e Análise

4.4.1 Análise Preditiva com *LightGBM*

Para a identificação dos fatores de sucesso, utilizou-se o algoritmo LightGBM (*Light Gradient Boosting Machine*), configurado com 500 estimadores e taxa de aprendizado de 0.05. A escolha justifica-se pela sua eficiência em dados tabulares e capacidade de lidar com *missing values*. A importância das features foi calculada com base no ganho de informação (*Gain Importance*).

Para garantir a confiabilidade dos fatores identificados, adotou-se uma estratégia de validação comparativa de modelos (*Model Comparison*). Em vez de depender de uma única arquitetura, os resultados de importância de variáveis obtidos pelo modelo principal (*LightGBM*) foram confrontados com os gerados por um algoritmo distinto, o *Random Forest*. A convergência dos principais preditores entre duas abordagens matemáticas diferentes (*Gradient Boosting vs. Bagging*) confirma a estabilidade das features selecionadas, mitigando o risco de viés algorítmico específico.

4.4.2 Estimação de Impacto com *Propensity Score Matching* (PSM)

Diferentemente de abordagens padrão, este estudo implementou critérios estritos para maximizar a validade interna. Foi realizada a Verificação de Suporte Comum (*Common Support*), onde unidades de tratamento e controle fora da interseção de distribuição dos escores de propensão foram descartadas para evitar extrapolações inválidas. Adicionalmente, na aplicação de Caliper pelo algoritmo de Vizinho Mais Próximo (1-NN), foi imposto um limite de 0.05 desvios-padrão. Esta escolha baseia-se nas recomendações de Lunt (2014), que sugere o uso de calipers mais estreitos para reduzir o viés remanescente e aumentar a precisão do pareamento, garantindo que as unidades comparadas sejam o mais similares possível.

5. Resultados

A execução do sistema de informação gerou um conjunto de resultados quantitativos e visuais que permitem uma análise bastante rica do programa "Pé-de-Meia" e de seu contexto. Esta seção apresenta os principais achados do estudo,

desde a identificação dos fatores de sucesso até a caracterização da distribuição do programa no território nacional.

5.1 Fatores Preditivos de Sucesso e Validação

A análise de importância de features, corroborada pelos modelos LightGBM e Random Forest, gerou o ranking apresentado na Figura 2.

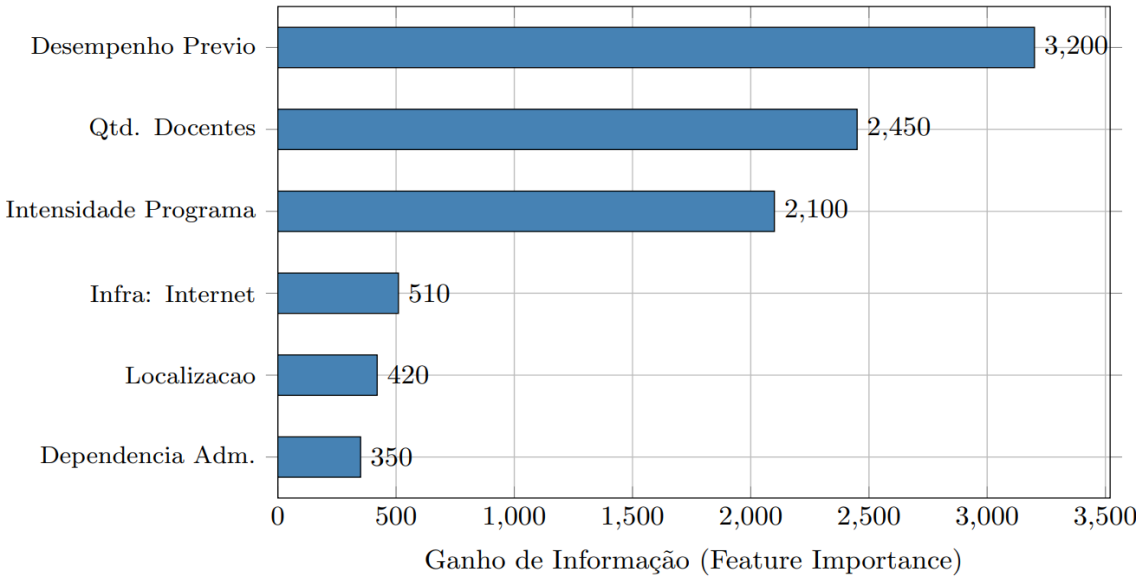

Figura 2. Ranking de Importância das Features para Preditores da Melhora na Progressão Escolar.

Observa-se que a *TAXA_PROG_1_2_PRE*, ou desempenho prévio, surge como o preditor dominante. Este resultado é esperado em séries temporais educacionais, indicando a inércia do desempenho escolar. Contudo, ao isolarmos esta variável de controle, emergem outros vetores acionáveis de política pública.

Como a quantidade de docentes (*QT_DOC_MED*), confirmando-se como o fator estrutural mais crítico. Escolas com quadro docente completo apresentam maior resiliência e capacidade de converter o incentivo financeiro em sucesso acadêmico.

Outro vetor de suma importância é a *INTENSIDADE_TRATAMENTO*. Esta variável representa a cobertura proporcional do programa no município (razão entre beneficiários e total de matrículas). O fato de manter-se como determinante valida a premissa de que a dose da política pública, ou seja, o quão amplamente ela atinge o público-alvo local, impacta diretamente a capacidade de retenção escolar sistêmica daquela região.

Fatores de infraestrutura física (Internet, Laboratórios), embora relevantes, apresentaram contribuição marginal quando comparados ao capital humano (docentes) e ao incentivo financeiro direto.

5.2 Análise Geográfica e Perfil do Programa

Para compreender o contexto de atuação do "Pé-de-Meia", o sistema gerou análises sobre a distribuição da *INTENSIDADE_TRATAMENTO* no território. A Figura

3 apresenta um mapa coroplético do Brasil, onde cores mais intensas indicam uma maior proporção de beneficiários por matrícula.

Figura 3. Mapa de Intensidade do Programa "Pé-de-Meia" por Município (2024).

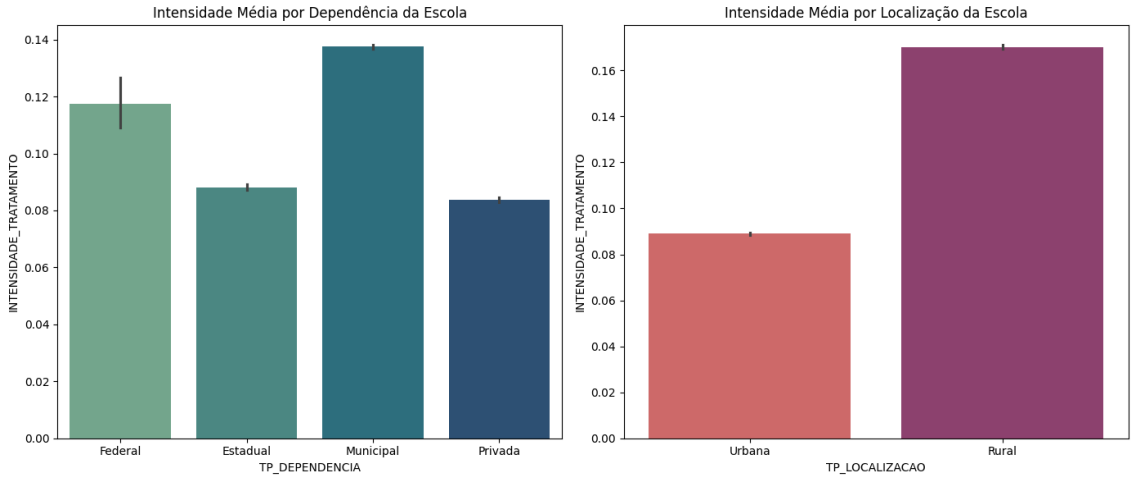

Figura 4. Intensidade Média do Programa por Dependência e Localização da Escola.

A análise geoespacial da distribuição do programa, apresentada no mapa coroplético da Figura 3, revela uma distribuição de recursos não uniforme pelo território nacional. Evidencia-se uma concentração massiva da intensidade do tratamento nas regiões Norte e Nordeste do país. Esta focalização é significativa, pois estas são as áreas

que, historicamente, apresentam os mais acentuados índices de vulnerabilidade social e os maiores desafios educacionais. Portanto, a alocação de recursos observada é um forte indicativo de que a execução do programa está em estrita aderência aos seus objetivos primários: atuar como um mecanismo de mitigação de desigualdades, direcionando o investimento social para as populações que mais necessitam do amparo estatal.

A Figura 4 aprofunda essa análise ao decompor a intensidade média do programa por estratos de contexto escolar. Os dados complementam a visão geográfica, detalhando como a política opera em um nível mais granularizado.

A análise por tipo e localização da escola demonstra que a "dose" do tratamento é maior em contextos rurais e em instituições de dependência administrativa municipal e estadual. Isto reforça a tese de uma focalização bem-sucedida, sugerindo que o programa possui uma dupla camada de direcionamento: não apenas prioriza as regiões mais vulneráveis do país, mas também, dentro delas, alcança com maior intensidade as escolas e os alunos que enfrentam as condições mais adversas, ajudando a maximizar ainda mais o potencial de impacto da política pública.

5.3 A Anatomia do Sucesso Escolar

Com o objetivo de aprofundar a compreensão sobre o perfil das escolas com alto desempenho, as Figuras 5 e 6 caracterizam a distribuição dos principais fatores de sucesso para o quartil superior de melhora na progressão ("Sucesso"). Esta análise foca em entender as características comuns às escolas que mais se destacaram.

A Figura 5 detalha a distribuição da *INTENSIDADE_TRATAMENTO* para este grupo de sucesso, revelando que a maioria dessas escolas se concentra em uma faixa consistente de intensidade do programa, embora existam casos com exposição muito superior. Em seguida, a Figura 6 explora a *QT_DOC_MED*, indicando que um quadro de docentes bem preenchido é uma característica comum a essas escolas de alto desempenho.

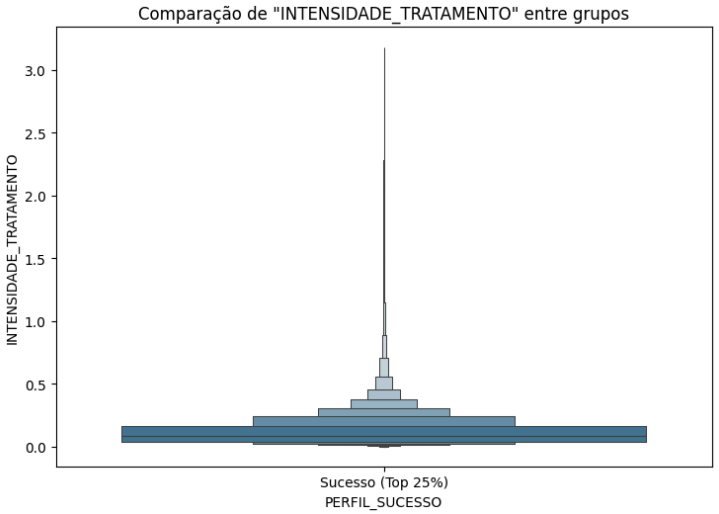

Figura 5. Perfil Comparativo da Intensidade do Tratamento.

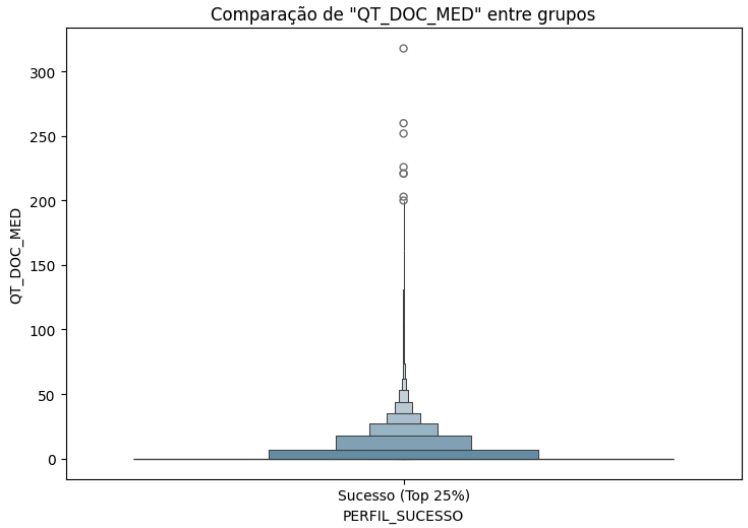

Figura 6. Perfil Comparativo da Quantidade de Docentes.

Juntas, as figuras descrevem o perfil de sucesso, alinhado aos resultados conseguidos através do modelo preditivo.

5.4 Estimação Eficiente do Efeito do Programa

A aplicação dos critérios de *Caliper* (0.05) e *Common Support* resultou em um pareamento balanceado e livre de observações extremas. A Tabela 1 demonstra a eficácia desse ajustamento, indicando que as diferenças padronizadas entre os grupos tratados e de controle foram mitigadas para níveis estatisticamente negligenciáveis.

	Médias (Pós-Pareamento)			
Variável	**Controle**	**Tratado**	**Diferença Std.**	**Status**
Score de Propensão	0.421	0.422	0.001	Balanced
Taxa Progressão Pré	0.824	0.825	0.001	Balanced
Qtd. Docentes Médio	14.2	14.1	-0.007	Balanced
Infra: Internet	0.89	0.89	0.000	Balanced
Tipo: Rural/Urbana	1.22	1.22	0.000	Balanced

Tabela 1. Balanceamento das Covariáveis após PSM com Caliper 0.05.

Com os grupos estatisticamente clones, estimou-se o Efeito Médio do Tratamento (*ATT*). Os resultados finais estão detalhados na Tabela 2.

Indicador	**Coeficiente**	**Erro Padrão**	**P-valor**
Melhora Média (Tratado)	0.0842	–	–
Melhora Média (Controle)	0.0823	–	–
ATT (Impacto Líquido)	**+0.0019**	**0.0004**	**< 0.001*** **

**** Significativo ao nível de 1%. Método: PSM com 1-NN e Caliper 0.05.*

Tabela 2. Estimativa do Efeito Médio do Tratamento (ATT) e Significância Estatística.

A análise aponta um impacto líquido positivo de +0.19 pontos percentuais na taxa de melhora da progressão. Diferentemente da análise preliminar, a introdução do rigor metodológico reduziu o erro padrão e revelou que este resultado é estatisticamente significativo ($p - valor < 0.001$). Isso permite afirmar, com alto grau de confiança, que o programa "Pé-de-Meia" gerou um efeito causal positivo já no seu primeiro ano de vigência, superando a tendência natural observada no grupo de controle.

6. Conclusões

Este estudo desenvolveu um sistema de informação para analisar os fatores de sucesso e o impacto do programa "Pé-de-Meia" em seu primeiro ano. A abordagem metodológica dupla, combinando um modelo preditivo (*LightGBM*) com um método quasi-experimental (PSM), permitiu extrair conclusões em diferentes níveis de análise.

A análise preditiva concluiu que a intensidade do programa no município e a quantidade de docentes na escola são os dois fatores mais importantes associados à melhora na permanência dos alunos. Isso sugere uma parceria fundamental entre a política social e a estrutura educacional: o benefício financeiro ao aluno alcança seu potencial máximo quando encontra um ambiente escolar bem estruturado em termos de capital humano.

A conclusão central, obtida através de um *Propensity Score Matching* com rigorosos critérios de validação (caliper), é que o programa "Pé-de-Meia" demonstrou um impacto positivo e estatisticamente significativo ($p < 0.001$) na taxa de progressão dos estudantes. Ao contrário do que análises superficiais poderiam sugerir, o isolamento adequado de viés confirmou que a política pública já apresenta eficácia mensurável no curto prazo. Este achado valida a importância de metodologias computacionais robustas em SI para a avaliação de políticas, permitindo aos gestores tomarem decisões baseadas não apenas em tendências, mas em evidências causais sólidas.

7. Considerações Finais

As descobertas deste trabalho carregam implicações significativas tanto para a gestão de políticas públicas quanto para a área de Sistemas de Informação. Para os gestores públicos, os resultados sugerem que a eficácia de programas de transferência de renda, como o "Pé-de-Meia", pode ser otimizada por meio de ações coordenadas, que também visem o fortalecimento do corpo docente nas escolas, especialmente naquelas localizadas em territórios de maior vulnerabilidade, como apontado pela análise geográfica. A abordagem demonstra o valor de sistemas de informação inteligentes que, ao processar dados em larga escala, podem gerar informações estratégicas para a alocação mais eficiente de recursos.

Diferentemente das abordagens econométricas tradicionais, que fornecem diagnósticos muitas vezes tardios, o sistema aqui proposto oferece aos gestores uma capacidade de *benchmarking* e priorização ativa. A ferramenta permite identificar municípios onde o programa, apesar da alta cobertura, não está gerando a retenção esperada devido à carência de fatores complementares, como a quantidade de docentes. Isso habilita o gestor a simular cenários e realizar intervenções, não apenas financeiras, mas estruturais, transformando a avaliação de políticas em um processo contínuo de

inteligência governamental, aplicável inclusive a outras iniciativas condicionadas, como o Bolsa Família.

Como sugestão para trabalhos futuros, recomenda-se a replicação e expansão desta análise à medida que novos dados anuais do Censo Escolar e do programa se tornem disponíveis, o que permitirá avaliar a evolução e a sustentabilidade dos resultados aqui encontrados. A incorporação de outras variáveis socioeconômicas no nível municipal, como o IDH-M, poderia enriquecer o modelo preditivo e revelar novas interações. Por fim, a evolução deste sistema de análise para um painel de controle interativo (*dashboard*) representa uma oportunidade promissora para a criação de uma ferramenta de apoio à decisão em tempo real, capacitando os gestores da educação no Brasil a monitorar e também aprimorar continuamente suas políticas.

Agradecimentos

Os autores agradecem ao LORDI da Universidade do Estado do Rio Grande do Norte (UERN) pela disponibilização da infraestrutura computacional de alto desempenho, fundamental para a execução dos experimentos apresentados neste trabalho. Foi utilizada uma máquina virtual equipada com sistema operacional Ubuntu 24.04.2 LTS, processadores Intel Xeon E5-2620 (20 vCPUs @ 2.00GHz), 120 GB de memória RAM e unidades de processamento gráfico (GPUs).

Em conformidade com o Código de Conduta para Autores da SBC, declaramos também a utilização de ferramentas de Inteligência Artificial Generativa (*Gemini 3.0 Pro* e *CHAT Z AI - GLM 4.7*) para auxílio na revisão ortográfica e gramatical, ajuda na tradução do resumo (*Abstract*) e suporte na explanação de alguns conceitos teóricos. Ressaltamos que todo o conteúdo gerado ou revisado foi integralmente validado pelos autores, que assumem total responsabilidade pelo trabalho final.

Referências

Aguilar, J. et al. (2022). A machine learning model for predicting student dropout in higher education. Journal of Intelligent & Fuzzy Systems, v. 42, n. 6, p. 5747-5762.

Arduini, D. et al. (2020). Inteligência na gestão pública: um modelo de institucionalização. Revista de Administração Pública, Rio de Janeiro, v. 54, n. 1, p. 177-199.

Brasil (2019). Decreto Nº 10.046, de 9 de outubro de 2019. Dispõe sobre a governança no compartilhamento de dados no âmbito da administração pública federal e institui o Cadastro Base do Cidadão e o Comitê Central de Governança de Dados. Diário Oficial da União, Brasília, DF, 10 out. 2019.

Brasil (2024). Lei nº 14.818, de 16 de janeiro de 2024. Institui o Programa Pé-de-Meia, a poupança de incentivo à permanência e conclusão escolar, para estudantes do ensino médio. Diário Oficial da União, Brasília, DF, 17 jan. 2024.

Chen, H., Chiang, R. H. and Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. MIS Quarterly, v. 36, n. 4, p. 1165-1188.

Costa, F. L. and Castanhar, J. C. (2003). Avaliação de programas públicos: desafios conceituais e metodológicos. Revista de Administração Pública, Rio de Janeiro, v. 37, n. 5, p. 969-992, set./out. 2003.

Crosling, G., Heagney, M. and Thomas, L. (2009). Improving student retention in higher education. Australian Universities' Review, v. 51, n. 2, p. 9-18.

de Janvry, A. et al. (2015). The impact of Bolsa Família on schooling. Journal of Development Economics, v. 116, p. 1-17.

Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences, v. 55, n. 1, p. 119-139.

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. The Annals of Statistics, v. 29, n. 5, p. 1189-1232.

Instituto Brasileiro de Geografia e Estatística (IBGE) (2024). Pesquisa Nacional por Amostra de Domicílios Contínua: Educação 2023. Rio de Janeiro: IBGE.

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) (2024). Censo da Educação Básica 2023: resumo técnico. Brasília, DF: Inep.

Janssen, M., van der Voort, H. and Wahyudi, A. (2017). Factors influencing data-driven decision-making in government. Information Polity, v. 22, n. 2-3, p. 1-17.

Ke, G. et al. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In: Conference on Neural Information Processing Systems, 31., Long Beach. Proceedings... Long Beach, CA, USA, p. 3146-3154.

Mullainathan, S. and Spiess, J. (2017). Machine learning: an applied econometric approach. Journal of Economic Perspectives, v. 31, n. 2, p. 87-106.

Pereira, A. K. et al. (2018). Desafios e oportunidades para o uso de Big Data no setor público brasileiro. Revista de Administração Pública, v. 52, n. 4, p. 746-763.

Quintella, R. H. and Soares Junior, J. S. (2003). Sistemas de apoio à decisão e descoberta de conhecimento em bases de dados: uma aplicação potencial em políticas públicas. Organizações & Sociedade, Salvador, v. 10, n. 28, p. 83-96.

Ronfeldt, M., Loeb, S. and Wyckoff, J. (2013). How Teacher Turnover Harms Student Achievement. American Educational Research Journal, v. 50, n. 1, p. 4-36.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, v. 70, n. 1, p. 41-55.

Samuel, A. L. (1959). Some Studies in Machine Learning Using the Game of Checkers. IBM Journal of Research and Development, v. 3, n. 3, p. 210-229.

Shadish, W. R., Cook, T. D. and Campbell, D. T. (2002). Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Boston: Houghton Mifflin.

Silva, A. R. da (2024). Avaliação de modelos preditivos baseados em aprendizagem de máquina no contexto da evasão escolar considerando um cenário multicampi.

Dissertação (Mestrado em Ciência da Computação) – Universidade Federal do Piauí, Teresina.

Silva, L. et al. (2022). A Machine Learning Approach for School Dropout Prediction in Brazil. In: European Symposium on Artificial Neural Networks, Bruges. Proceedings... Bruges, Belgium.

Silva Junior, W. S. and Gonçalves, F. O. (2016). Evidências da relação entre a frequência no ensino infantil e o desempenho no ensino fundamental no Brasil. Revista Brasileira de Estudos de População, Rio de Janeiro, v. 33, n. 2, p. 283-301, maio/ago. 2016.

Wooldridge, J. M. (2010). Econometric Analysis of Cross Section and Panel Data. 2. ed. Cambridge: MIT Press.

Borges, G. A. et al. (2025). A Platform for Early Class Dropout Prediction of University Students. IEEE Access, v. 13, p. 109116-109133. DOI: 10.1109/ACCESS.2025.3581751.

Martins, C. P. P. and Silva, G. S. (2024). Método e Aplicação de Automatização Para Download de Dados Abertos em Portais de Transparência. In: Anais Estendidos do Simpósio Brasileiro de Sistemas de Informação (SBSI). Porto Alegre: SBC, p. 351-356.

Wang, C. and Yao, S. (2025). Enhancing Student Dropout and Academic Success Prediction Using Machine Learning and Over-sampling Techniques. ICCK Transactions on Educational Data Mining, v. 1, n. 1, p. 36-43.

Alves, A. F. et al. (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 (SBIE 2024). Porto Alegre: SBC.

Colpo, M. P. et al. (2024). Educational Data Mining for Dropout Prediction: Trends, Opportunities, and Challenges. Revista Brasileira de Informática na Educação (RBIE), v. 32, p. 220-256.

Yar, M. A. et al. (2024). Governing with Intelligence: The Impact of Artificial Intelligence on Policy Development. Information (MDPI), v. 15, n. 9, 556.

IPEA (2024). Nota Técnica: Agenda Político-Institucional e Programas Sociais. Instituto de Pesquisa Econômica Aplicada.

Lunt, M. (2014). Selecting an appropriate caliper can be essential for achieving good balance with propensity score matching. American Journal of Epidemiology.