Fatores Escolares Associados ao Absenteísmo no Ensino Médio: Um Estudo com Modelos Explicáveis

Abílio Nogueira Barros[1] Markson Rebelo Marcolino[1,3], Débora Barbosa Leite Silva[1], Leonardo Brandão Marques[1], Elthon Oliveira[1], Diego Dermeval[1], Flavia Galvani[4], Anita Gea Martinez Stefani[5], Emanuel Marques Queiroga[1,2], Cristian Cechinel[1,3], Thales Vieira[1]

[1] Center of Excellence for Social Technologies (NEES) – Universidade Federal de Alagoas (UFAL)

[2]Instituto Federal de Educação, Ciência e Tecnologia Sul-Rio-Grandense (IFSul)

[3]Centro de Ciências, Tecnologias e Saúde – Universidade Federal de Santa Catarina (UFSC)

[4]Blavatnik School of Government, University of Oxford

[5]Ministério da Educação

abilionbarros@gmail.com, emanuelmqueiroga@gmail.com, debora.silva@nees.ufal.br, leonardo.marques@nees.ufal.br, thales.vieira@nees.ufal.br, elthon.oliveira@nees.ufal.br, diego.matos@nees.ufal.br markson.marcolino@ufsc.br, cristian.cechinel@ufsc.br flavia.galvani@bsg.ox.ac.uk, anitastefani@mec.gov.br

Resumo. *Este estudo investiga os fatores institucionais escolares que influenciam o absenteísmo dos alunos no Ensino Médio, segmentando as escolas públicas conforme seu nível de complexidade de gestão. Dados de absenteísmo extraídos do Sistema Gestão Presente foram cruzados com os dados do CENSO escolar 2024 para a geração de diferentes modelos de classificação usando Aprendizado de Máquina. Por meio da análise dos valores SHAP destes modelos, foram identificadas as contribuições de diferentes variáveis institucionais na predição da permanência escolar. Os resultados mostram que os fatores mais relevantes variam entre os grupos, reforçando a necessidade de estratégias específicas para cada contexto escolar. De maneira geral, fatores estruturais e de suporte têm importância em todas as faixas, porém são mais relevantes em escolas com baixa complexidade de gestão, enquanto aspectos como composição do corpo docente e métodos pedagógicos são mais relevantes em escolas com complexidade de gestão mais alta.*

1. Introdução

A permanência dos alunos no Ensino Médio tem sido um dos principais desafios enfrentados pelas redes públicas de ensino no Brasil[Tartuce et al. 2018]. As altas taxas de evasão e abandono escolar comprometem não apenas os índices educacionais, mas também ampliam desigualdades sociais, limitando o acesso dos jovens a oportunidades acadêmicas e profissionais futuras[Soares et al. 2015, Tartuce et al. 2018]. Compreender os fatores que influenciam a permanência escolar é, portanto, uma tarefa essencial para o aprimoramento de políticas públicas na área da educação. Tradicionalmente, estudos sobre o tema têm se concentrado em variáveis socioeconômicas ou no desempenho individual dos estudantes [Lopes Filho and Silveira 2021]. No entanto, a escola, enquanto unidade institucional e espaço estruturado de aprendizagem, desempenha papel central nesse processo.

DOI: 10.5753/wapla.2025.15903

Essa relação, no entanto, não é homogênea. As escolas públicas brasileiras operam sob diferentes graus de complexidade de gestão, o que torna inadequado o uso de análises agregadas que desconsideram essa diversidade. Modelos analíticos que não segmentam as escolas por perfil institucional correm o risco de mascarar padrões relevantes ou mesmo reforçar desigualdades já existentes. Nesse contexto, este estudo propõe uma abordagem baseada em aprendizado de máquina para classificar escolas de acordo com o desempenho em permanência dos alunos, dentro de faixas previamente definidas por nível de complexidade. Ao invés de buscar apenas a acurácia preditiva, o foco está na interpretabilidade dos resultados.

2. Trabalhos Relacionados

O estudo de [Krüger et al. 2023] avalia modelos de IA Explicável (XAI) na classificação de desempenho de estudantes. Foram utilizados dados reais e métricas de interpretabilidade com foco em ética e transparência. Os resultados destacam a utilidade da XAI na compreensão de fatores que impactam o desempenho escolar. Ainda, o artigo de [Melo et al. 2022] aplica técnicas de XAI para prever evasão escolar em alunos do IFRN. Foi proposto um índice de explicabilidade com base em *checklist* da literatura. A abordagem mostrou-se eficaz para identificar perfis de evasão e apoiar decisões educacionais mais éticas e transparentes. O estudo de [Bulut et al. 2024] analisou dados do 9º ano usando modelos de árvores e de aprendizagem profunda, destacando o melhor desempenho e explicabilidade dos primeiros. Concluiu que o sentimento de pertencimento do estudante é fator-chave na previsão do abandono escolar.

O trabalho de [Marques Queiroga et al. 2024] utilizou dados abertos do SAEB e do Censo da Educação Básica (INEP) para identificar oito fatores escolares que impactam o desempenho dos alunos e a equidade educacional. Essa análise permitiu a visualização das disparidades entre regiões, estados e escolas, facilitando a avaliação de políticas existentes e a identificação de escolas específicas para investimentos direcionados com base em suas necessidades. As políticas públicas sugeridas por este estudo incluem o estabelecimento de uma infraestrutura robusta de qualidade de dados e o incentivo a iniciativas focadas na medição e promoção da equidade educacional. Por sua vez, o trabalho de [Queiroga et al. 2024] desenvolveu uma metodologia baseada em *machine learning* utilizando dados de alunos do ensino médio do Espírito Santo para prever o risco de reprovação e abandono. Os modelos preditivos gerados servem como um sistema de alerta precoce, permitindo que educadores e administradores escolares implementem intervenções proativas e personalizadas para apoiar os alunos em risco.

3. Metodologia

As bases selecionadas para este estudo tiveram como objetivo permitir o cruzamento entre o absenteísmo real dos alunos no ano de referência e as características estruturais, pedagógicas e quantitativas das unidades escolares. buscou-se também incorporar um critério de segmentação das escolas com base em algum indicador oficial que possibilitasse a formação de subconjuntos mais homogêneos. Para isso, foi utilizado o Índice de Complexidade da Gestão Escolar (ICG), que é um indicador sintético que busca estimar a complexidade da gestão escolar com base em aspectos como porte da escola, número de etapas de ensino ofertadas, número de turnos de funcionamento, entre outros fatores

operacionais e estruturais. O índice é dividido em seis faixas distintas, variando de menor (1) para maior complexidade de gestão (6).

3.1. Criação do conjunto de dados

A construção da base de dados partiu da consolidação de três fontes: (i) o Sistema Gestão Presente (SGP), com registros de frequência de estudantes do ensino médio referentes ao ano de 2024; (ii) o Censo da Educação Básica de 2024, disponibilizado pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP); e (iii) o Índice de Complexidade da Gestão Escolar (ICG) de 2024, também disponibilizado pelo INEP.

Na etapa inicial do pré-processamento, foram consideradas apenas as escolas com dados válidos nas três bases mencionadas. Isso garantiu a consistência e a integridade das análises, resultando em uma amostra composta por pouco mais de 20 mil instituições escolares ativas no Brasil. A base do Censo Escolar foi submetida a uma curadoria manual, por meio da qual foram mantidas apenas variáveis diretamente relacionadas ao ensino médio, já que atualmente os registros do SGP se referem exclusivamente a essa etapa de ensino. Foram descartadas variáveis ligadas ao ensino fundamental, à Educação de Jovens e Adultos (EJA) e à infraestrutura escolar não vinculada diretamente ao ensino médio.

Em seguida, variáveis quantitativas foram padronizadas de acordo com o número de matrículas por escola, de modo a evitar distorções causadas pelo porte institucional. Além disso, variáveis categóricas com múltiplos valores foram binarizadas, permitindo uma modelagem mais eficiente nos algoritmos de classificação. A base do SGP foi utilizada para gerar a classe alvo do problema de classificação. A partir dos registros de frequência dos alunos, foi calculada a proporção de estudantes em cada faixa de absenteísmo proposta por [Kearney 2021]: Pequeno ($> 95\%$) indica alta presença, Moderado (91–95%) sugere poucas faltas, Significativo (81–90%) revela ausência frequente, Alto (71–80%) aponta padrão preocupante e Severo ($\leq 70\%$) representa risco elevado de evasão do aluno. As instituições com mais de 50% de seus alunos classificados nas faixas ”Pequeno”e ”Moderado”foram rotuladas com a classe 1, enquanto as escolas com mais de 50% dos estudantes enquadrados nas faixas ”Significativo”, ”Alto”e ”Severo”foram rotuladas com a classe 0, indicando altos níveis de absenteísmo.

Após a junção com os dados do ICG, a base consolidada foi segmentada em seis subconjuntos distintos, de acordo com a faixa de complexidade da gestão escolar. Cada subconjunto manteve a mesma estrutura de atributos, permitindo análises comparáveis entre diferentes contextos de complexidade. A quantidade final das escolas por faixa de IGC e categoria de Absenteísmo é ilustrada na Tabela 1.

Flag Absenteísmo	**ICG**					
	1	2	3	4	5	6
0	143 (59%)	1372 (43%)	1362 (41%)	5618 (60%)	2335 (54%)	1462 (75%)
1	98 (41%)	1856 (57%)	1939 (59%)	3792 (40%)	2015 (46%)	484 (25%)

Tabela 1. Distribuição de escolas por ICG e flag de absenteísmo

Com a base consolidada contendo inicialmente cerca de 350 variáveis, foi conduzido um processo de redução de dimensionalidade utilizando o cálculo de informação

mútua (com limiar de 0,04 e remoção de valores de correlação maiores que 0,85) para mitigar a redundância e simplificar a modelagem. Foram eliminadas variáveis com baixa informação mútua (inferior a 0,05) e atributos com alta correlação (superior a 0,85), o que resultou na remoção de mais de 50 variáveis. Ao final dessa etapa, a base permaneceu com aproximadamente 294 atributos.

3.2. Teste e Implementação dos modelos de Aprendizado de Máquina

A próxima etapa, consistiu na seleção de variáveis por meio de um algoritmo do tipo *Random Forest*, aplicado individualmente a cada uma das seis bases segmentadas por faixa de complexidade. A importância das variáveis foi utilizada como critério para a retenção de atributos mais relevantes. Em média, cada base passou a conter 35 atributos, permitindo a redução da complexidade computacional e favorecendo uma modelagem mais interpretável. A abordagem foi escolhida visando evidenciar possíveis diferenças nos fatores preditivos mais relevantes conforme a complexidade da gestão escolar.

Na sequência, foram testados quatro algoritmos de classificação baseados em árvores de decisão (LightGBM, XGBoost, CatBoost, LogisticRegression), utilizando validação cruzada com *5 folds*. O melhor desempenho observado entre os classificadores foi para o algoritmo CatBoost. Esse algoritmo foi então utilizado como base para uma etapa adicional de ajuste de hiperparâmetros, com o objetivo de maximizar a performance do modelo final. Os resultados de desempenho para os melhores modelos de cada faixa de ICG são apresentados na Tabela 2.

ICG	Acurácia	Precisão	Revocação	F1-Score
Nível 1	0.8525	0.8077	0.8400	0.8235
Nível 2	0.7447	0.7570	0.8190	0.7867
Nível 3	0.6889	0.6932	0.8433	0.7609
Nível 4	0.7552	0.7128	0.6572	0.6839
Nível 5	0.7849	0.7755	0.7540	0.7646
Nível 6	0.8501	0.8077	0.5207	0.6332

Tabela 2. Desempenho por IGC da escola com CatBoost

Como pode ser observado na Tabela 2, a análise por níveis de complexidade indica que os níveis 1 e 6 alcançaram alta acurácia, embora o nível 6 tenha apresentado baixa capacidade de identificar corretamente os casos positivos (recall). Os níveis 2 e 5 mostraram desempenho mais equilibrado entre precisão e recall, enquanto o nível 3 priorizou a identificação de positivos, com menor precisão. O nível 4 teve o pior F1-Score, sugerindo um desequilíbrio nas métricas, possivelmente causado por uma maior heterogeneidade entre as escolas desse grupo

Como última etapa, foram aplicadas técnicas de interpretabilidade por meio da biblioteca SHAP (SHapley Additive ExPlanations), que permitem mensurar a contribuição de cada atributo na predição do modelo. Para fins de apresentação neste artigo, foram destacados os cinco atributos mais relevantes para cada um dos seis modelos treinados. Entretanto, ressalta-se que a análise completa contemplou os vinte principais atributos, proporcionando uma visão mais abrangente sobre os fatores que influenciam os padrões de frequência escolar no ensino médio.

4. Resultados e discussão

Após a execução dos experimentos, visando uma análise interpretável e considerando as limitações de espaço do artigo, foram selecionadas as cinco variáveis mais relevantes para cada modelo. O gráfico *beeswarm* do SHAP é utilizado para visualizar o impacto de cada atributo de entrada nas previsões feitas por um modelo de aprendizado de máquina, considerando múltiplas instâncias. No gráfico, cada ponto representa uma escola, sua posição horizontal mostra a contribuição de um atributo específico para a predição dessa categoria, e a cor representa o valor real do atributo. Essa visualização permite entender quais características mais influenciam a distinção entre escolas com maior quantidade de alunos com baixos nível de absenteísmo (1) e altos níveis de absenteísmo (0).

4.1. ICG níveis 1 e 2

A Figura 1 apresenta o impacto dos atributos para os modelos voltados para escolas com o ICG 1.

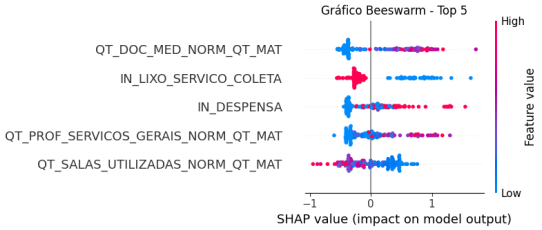

Figura 1. Distribuição SHAP dos preditores para a Faixa 1 de ICG.

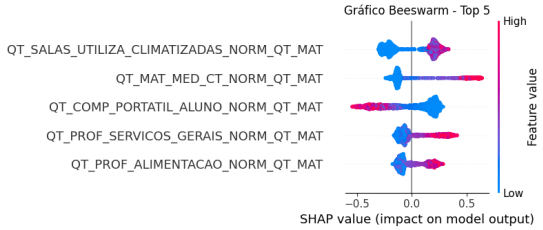

Figura 2. Distribuição SHAP dos preditores para a Faixa 2 de ICG.

Como ilustrado na Figura, a variável QT_DOC_MED_NORM_QT_MAT impacta positivamente o modelo (em rosa), sugerindo que mais docentes no ensino médio estão associados a menor absenteísmo. A ausência de coleta de lixo (IN_LIXO_SERVICO_COLETA, em azul) exerce leve influência negativa, sem relação clara com a presença dos alunos. A presença de dispensa escolar (IN_DESPENSA) está ligada à menor ausência, enquanto mais profissionais de serviços gerais (QT_PROF_SERVICOS_GERAIS_NORM_QT_MAT) também se associam a menores níveis de absenteísmo. Finalmente, escolas com maior número de salas utilizadas (QT_SALAS_UTILIZADAS_NORM_QT_MAT) tendem a apresentar melhor presença estudantil.

Podemos observar na Figura 2 que escolas do nível 2 apresentam salas climatizadas (QT_SALAS_UTILIZA_CLIMATIZADAS_NORM_QT_MAT) que têm impacto positivo para valores altos, sugerindo que o uso dessas salas proporciona um conforto ambiental durante os momentos pedagógicos e influencia positivamente a frequência dos alunos. Uma maior quantidade de matrículas de cursos técnicos integrados à educação profissional durante o ensino médio (QT_MAT_MED_CT_NORM_QT_MAT) tem leve impacto negativo quando altos, fato esse que não indica diretamente algo pejorativo à instituição, podendo ser interpretado como uma maior carga horária presente e, assim, aumento do nível de absenteísmo.

A presença de computadores portáteis por aluno (QT_COMP_PORTATIL_ALUNO_NORM_QT_MAT) tem leve impacto negativo, embora menos expressivo do que variáveis de infraestrutura física descritas nesse

grupo. Assim como no grupo 1, a presença de profissionais de serviços gerais (QT_PROF_SERVICOS_GERAIS_NORM_QT_MAT) é relevante, com impacto positivo nos casos de maior quantidade. O número de profissionais responsáveis por alimentação (QT_PROF_ALIMENTACAO_NORM_QT_MAT) mostra impacto sutilmente positivo, sugerindo que suporte nutricional também contribui para manter a presença dos alunos.

Em comparação de escolas entre a faixa 1 e 2, temos que unidades escolares que pertencem à faixa 1 apresentam maior importância para *features* que tratam da infraestrutura básica, dispensa e utilização de salas. Já para as escolas da faixa 2, temos que o uso de salas climatizadas dá lugar à utilização de salas, sugerindo a percepção da diferença dos modelos entre as faixas. Profissionais de apoio (*serviços gerais*, *alimentação*) aparecem como relevantes nos dois grupos, reforçando a importância do suporte *não-docente* como elemento intra-faixas na permanência dos alunos.

4.2. ICG níveis 3 e 4

Para as escolas classificadas na complexidade nível 3 na Figura 3, temos mais uma vez que valores mais altos no atributo QT_SALAS_UTILIZA_CLIMATIZADAS_NORM_QT_MAT estão fortemente associados a um menor absenteísmo.

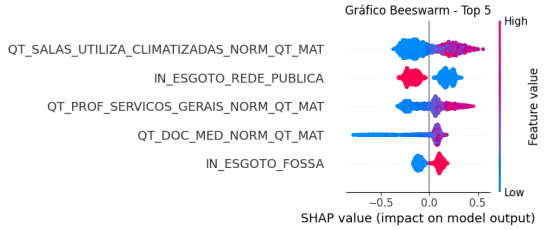

Figura 3. Distribuição SHAP dos preditores para a Faixa 3 de ICG.

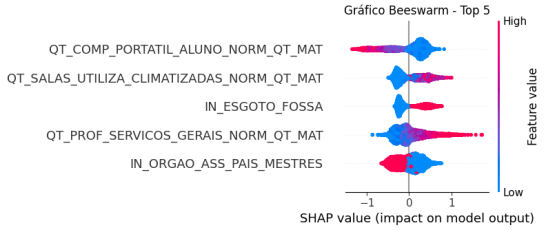

Figura 4. Distribuição SHAP dos preditores para a Faixa 4 de ICG.

A presença de rede pública de esgoto (IN_ESGOTO_REDE_PUBLICA = 1) está associada a piores resultados (classe 0), possivelmente por representar áreas com infraestrutura precária. Em contraste, a presença de fossa (IN_ESGOTO_FOSSA = 1) está ligada a melhores resultados (classe 1), sugerindo condições mais controladas. A maior quantidade de profissionais de serviços gerais (QT_PROF_SERVICOS_GERAIS_NORM_QT_MAT) segue indicando impacto positivo na presença dos alunos. Já menores valores de docentes no ensino médio (QT_DOC_MED_NORM_QT_MAT) relacionam-se a piores resultados no modelo.

Já para escolas de complexidade nível 4, observa-se na Figura 4 que o principal atributo é QT_COMP_PORTATIL_ALUNO_NORM_QT_MAT, indicando que quantidades moderadas ou baixas de computadores portáteis podem favorecer a permanência dos alunos, enquanto valores altos tendem a ter efeito oposto. Em seguida, a variável QT_SALAS_UTILIZA_CLIMATIZADAS_NORM_QT_MAT reforça, mais uma vez, o impacto positivo do uso de salas climatizadas na presença estudantil. Também se destaca o atributo IN_ESGOTO_FOSSA, cuja presença está associada a melhores resultados, sugerindo melhor estrutura escolar.

Apresentando impacto bidirecional, a variável QT_PROF_SERVICOS_GERAIS_NORM_QT_MAT tem papel relevante no modelo:

maior quantidade de profissionais de serviços gerais está associada a melhores resultados, enquanto valores baixos indicam pior desempenho. Por fim, diferentemente de achados em outros estudos, a presença do órgão de pais e mestres (IN_ORGAO_ASS_PAIS_MESTRES) está ligada a piores resultados, sendo que sua ausência tem impacto neutro no modelo.

4.3. ICG níveis 5 e 6

Avançando para escolas da faixa 5, constata-se na Figura 5 que, quando o número de professores com formação em pedagogia (QT_PROF_PEDAGOGIA_NORM_QT_MAT) é alto, o resultado tende a reduzir o desempenho do absenteísmo da escola.

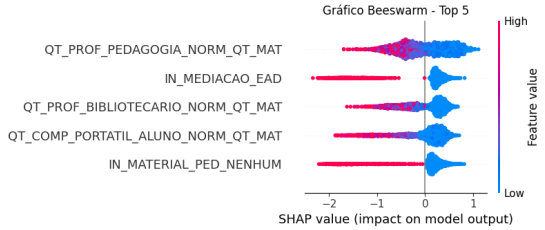

Figura 5. Distribuição SHAP dos preditores para a Faixa 5 de ICG.

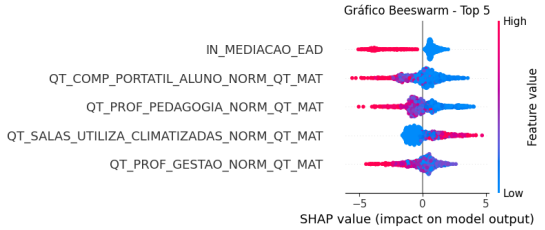

Figura 6. Distribuição SHAP dos preditores para a Faixa 6 de ICG.

Valores mais baixos de professores com formação em pedagogia (QT_PROF_PEDAGOGIA_NORM_QT_MAT) estão associados a melhores resultados, sugerindo que o equilíbrio na composição do corpo docente é mais relevante do que a especialização isolada. A mediação didática pedagógica a distância (IN_MEDIACAO_EAD) apresenta, em geral, valores SHAP negativos, indicando impacto desfavorável no desempenho.

Altas proporções de professores bibliotecários (QT_PROF_BIBLIOTECARIO_NORM_QT_MAT) também não trazem ganhos ao modelo, apresentando valores SHAP predominantemente negativos. Da mesma forma, elevados valores de computadores portáteis por aluno (QT_COMP_PORTATIL_ALUNO_NORM_QT_MAT) continuam associados a resultados inferiores, com raras exceções pontuais. Por fim, a ausência de material pedagógico (IN_MATERIAL_PED_NENHUM) está claramente ligada a desempenho negativo, reforçando a importância de recursos físicos ou digitais para a permanência escolar.

Por fim, para as escolas da faixa 6, como mostrado na Figura 6, a mediação didático-pedagógica a distância (IN_MEDIACAO_EAD) permanece associada a piores resultados, sem apresentar comportamento neutro. A relação entre computadores portáteis por aluno (QT_COMP_PORTATIL_ALUNO_NORM_QT_MAT) segue ambígua, com valores baixos ainda associados a bons resultados, mas altas quantidades não implicando maior presença dos alunos.

Assim como na faixa 5, uma alta proporção de professores com formação em pedagogia (QT_PROF_PEDAGOGIA_NORM_QT_MAT) está relacionada a menor assiduidade, reforçando a necessidade de balanceamento no corpo docente. O uso de salas climatizadas (QT_SALAS_UTILIZA_CLIMATIZADAS_NORM_QT_MAT) continua sendo um fator positivo para a permanência dos alunos. Por fim, a quantidade de professores atuando

na gestão (QT_PROF_GESTAO_NORM_QT_MAT) surge como o quinto atributo mais relevante, indicando que altos valores podem impactar negativamente a presença estudantil. Esse efeito pode estar relacionado à normalização pela quantidade de matrículas, sugerindo que um quadro administrativo excessivo reduz a disponibilidade de professores em sala de aula.

A análise comparativa entre as escolas de grupos 5 e 6 revela padrões convergentes e divergências críticas nos fatores que impactam a presença frequente do aluno. Em ambos os grupos, a mediação EAD está consistentemente associada a piores resultados, sugerindo falhas estruturais na implementação da mediação remota escolar. Além disso, a quantidade de professores pedagogos apresenta um efeito divergente: em excesso, correlaciona-se com redução de desempenho, indicando a necessidade de equilíbrio na composição docente. Contudo, enquanto no Grupo 5 os computadores portáteis não demonstram benefício claro, no Grupo 6 há uma relação ambígua, com casos pontuais de impacto positivo. Outra divergência marcante é o papel da gestão escolar: no Grupo 6, seu excesso está fortemente ligado a piores resultados, possivelmente por desequilibrar a alocação de recursos entre administração e sala de aula.

Por fim, o uso de salas climatizadas mantém-se como um fator positivo em ambos os grupos, reforçando a importância de condições básicas adequadas. Esses achados destacam que, embora desafios como a mediação EAD e o equilíbrio docente sejam comuns, intervenções devem ser adaptadas às particularidades de cada nível, especialmente em relação à composição de docentes na escola.

5. Considerações Finais

O processo de classificação realizado permitiu segmentar as escolas conforme diferentes níveis de complexidade, viabilizando uma análise mais direcionada sobre os fatores que impactam a presença dos alunos de acordo com os modelos. A partir dos valores SHAP, foi possível identificar padrões relevantes em cada faixa, evidenciando que as variáveis de maior influência mudam substancialmente conforme o contexto institucional.

Observou-se que variáveis relacionadas à infraestrutura física, como a climatização das salas, apresentam impacto positivo recorrente em praticamente todas as faixas de complexidade. Por outro lado, fatores como a mediação EAD e a composição do quadro docente — especialmente no que se refere à atuação na gestão e à formação pedagógica — revelaram efeitos mais variados, exigindo equilíbrio e adequação ao perfil de cada escola. A presença de profissionais de apoio demonstrou-se consistentemente relevante, reforçando o papel do suporte não-docente como fator transversal à permanência estudantil. De forma geral, fatores estruturais e de suporte (salas, equipe de apoio, saneamento) são mais significativos em escolas de baixa a média complexidade (faixas 1 a 4), enquanto aspectos relacionados à composição do corpo docente e às práticas pedagógicas (como o uso de EAD) ganham maior importância nas faixas superiores (5 e 6).

Uma das limitações do trabalho é que os modelos identificam apenas uma possível correlação e não necessariamente uma causalidade entre os atributos das escolas e o absenteísmo. Para identificação de causalidade, outras técnicas como Counterfactual Analysis podem ser implementadas em trabalhos futuros.

Este trabalho foi financiado pelo Ministério da Educação (MEC), Brasil, por meio do Termo de Execução Descentralizada TED13914.

Referências

[Bulut et al. 2024] Bulut, O., Wongvorachan, T., He, S., and Lee, S. (2024). Enhancing high-school dropout identification: A collaborative approach integrating human and machine insights. *Discover Education*, 3(1):109.

[Kearney 2021] Kearney, C. A. (2021). Integrating systemic and analytic approaches to school attendance problems: Synergistic frameworks for research and policy directions. In *Child & Youth Care Forum*, volume 50, pages 701–742. Springer.

[Krüger et al. 2023] Krüger, J. G. C., de Souza Britto Jr, A., and Barddal, J. P. (2023). An explainable machine learning approach for student dropout prediction. *Expert Systems with Applications*, 233:120933.

[Lopes Filho and Silveira 2021] Lopes Filho, J. A. B. and Silveira, I. F. (2021). Detecção precoce de estudantes em risco de evasão usando dados administrativos e aprendizagem de máquina. *Revista Ibérica de Sistemas e Tecnologias de Informação*, (E40):480–495.

[Marques Queiroga et al. 2024] Marques Queiroga, E., Sarmanho Siqueira, E., Dos Santos Portela, C., Damasceno Cordeiro, T., Ibert Bittencourt, I., Isotani, S., Ferreira Mello, R., Muñoz, R., and Cechinel, C. (2024). Data-driven strategies for achieving school equity: Insights from brazil and policy recommendations. *IEEE Access*, 12:101646–101659.

[Melo et al. 2022] Melo, E., Silva, I., Costa, D. G., Viegas, C. M., and Barros, T. M. (2022). On the use of explainable artificial intelligence to evaluate school dropout. *Education Sciences*, 12(12):845.

[Queiroga et al. 2024] Queiroga, E. M., Santana, D., da Silva, M., de Aguiar, M., dos Santos, V., Mello, R. F., Bittencourt, I. I., and Cechinel, C. (2024). Anticipating student abandonment and failure: Predictive models in high school settings. In *International Conference on Artificial Intelligence in Education*, pages 351–364. Springer.

[Soares et al. 2015] Soares, T. M., Fernandes, N. d. S., Nóbrega, M. C., and Nicolella, A. C. (2015). Fatores associados ao abandono escolar no ensino médio público de minas gerais. *Educação e Pesquisa*, 41(3):757–772.

[Tartuce et al. 2018] Tartuce, G. L. B., Moriconi, G. M., Davis, C. L., and Nunes, M. M. (2018). Desafios do ensino médio no brasil: iniciativas das secretarias de educação. *Cadernos de Pesquisa*, 48(168):478–504.