LUMEN ET VIRTUS
ISSN: 2177-2789

PREVISÃO DE CONECTIVIDADE ESCOLAR NO BRASIL: UMA ABORDAGEM COM RANDOM FOREST E VALIDAÇÃO TEMPORAL NOS DADOS DO CENSO ESCOLAR

SCHOOL CONNECTIVITY PREDICTION IN BRAZIL: A RANDOM FOREST APPROACH WITH TEMPORAL VALIDATION ON SCHOOL CENSUS DATA

PREDICCIÓN DE LA CONECTIVIDAD ESCOLAR EN BRASIL: UN ENFOQUE DE BOSQUE ALEATORIO Y VALIDACIÓN TEMPORAL DE DATOS DEL CENSO ESCOLAR

https://doi.org/10.56238/levv17n60-053

Data de submissão: 23/04/2026 **Data de publicação:** 23/05/2026

Gustavo Lima Mendes
Graduando em Sistemas de Informação
Instituição: Centro Universitário Santa Terezinha (CEST)
E-mail: gustamendez27@gmail.com
Orcid: https://orcid.org/0009-0009-6074-6967

Marcos Albino do Carmo Carvalho Ferreira
Graduando em Sistemas de Informação
Instituição: Centro Universitário Santa Terezinha (CEST)
E-mail: marcos.albino@cest.edu.br
Orcid: https://orcid.org/0009-0005-9171-7393

Dadilton Bastos Melo
Orientador
Especialista em Ciência de Dados e Big Data Analytics
E-mail: dadilton.melo@cest.edu.br
Orcid: https://orcid.org/0009-0000-3673-8814

RESUMO

A conectividade à internet em escolas públicas brasileiras é indicador crítico de equidade educacional e inclusão digital. Prever quais escolas da educação básica possuem banda larga utilizando aprendizado de máquina. Random Forest treinado com microdados do Censo Escolar 2023 (217.625 registros) e validado temporalmente com dados de 2024 (215.545 registros). Acurácia de 85,38% e ROC-AUC de 0,8909. Principais preditores: recursos pedagógicos, infraestrutura básica, número de turmas e localização rural. Validação temporal em dados educacionais longitudinais, viabilizando planejamento de políticas públicas. Escolas com mais recursos tendem a ter banda larga, revelando desigualdades estruturais concentradas.

Palavras-chave: Aprendizado de Máquina. Conectividade Escolar. Random Forest. Censo Escolar. Inclusão Digital.

ABSTRACT
Internet connectivity in Brazilian public schools is a critical indicator of educational equity and digital inclusion. To predict which basic education schools have broadband using machine learning. Random Forest trained on 2023 School Census microdata (217,625 records) and temporally validated on 2024 data (215,545 records). Accuracy of 85.38% and ROC-AUC of 0.8909. Main predictors: pedagogical resources, basic infrastructure, number of classes, and rural location.Temporal validation on longitudinal educational data, enabling public policy planning. Schools with more resources tend to have broadband, revealing concentrated structural inequalities.

Keywords: Machine Learning. School Connectivity. Random Forest. School Census. Digital Inclusion.

RESUMEN
La conectividad a internet en las escuelas públicas brasileñas es un indicador clave de equidad educativa e inclusión digital. Este estudio tuvo como objetivo predecir qué escuelas de educación básica cuentan con acceso a banda ancha mediante aprendizaje automático. Se entrenó un conjunto de datos de Bosque Aleatorio con microdatos del Censo Escolar de 2023 (217.625 registros) y se validó temporalmente con datos de 2024 (215.545 registros). La precisión fue del 85,38 % y el área bajo la curva ROC (ROC-AUC) fue de 0,8909. Los predictores clave incluyeron recursos pedagógicos, infraestructura básica, número de aulas y ubicación rural. La validación temporal se realizó con datos educativos longitudinales, lo que permitió la planificación de políticas públicas. Las escuelas con más recursos tienden a tener acceso a banda ancha, lo que revela desigualdades estructurales concentradas.

Palabras clave: Aprendizaje Automático. Conectividad Escolar. Bosque Aleatorio. Censo Escolar. Inclusión Digital.

1 INTRODUÇÃO

A ampliação do acesso à internet nas escolas públicas brasileiras é amplamente reconhecida como um dos eixos centrais da agenda de equidade educacional no século XXI. O Plano Nacional de Educação (PNE), as orientações do Ministério da Educação (MEC) e iniciativas como o Conecta Escola e o PDDE Conectado reforçam a relevância da conectividade para a atualização das práticas pedagógicas. Ainda assim, apesar dos avanços observados, a oferta de banda larga entre as escolas de educação básica segue marcada por desigualdades estruturais, com maior disponibilidade nas localidades mais urbanizadas e nas redes que contam com infraestrutura mais consolidada (CETIC.BR, 2023).

O Censo Escolar, realizado anualmente pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), representa a principal base de dados da educação básica no País. Nele são reunidas informações sobre infraestrutura, recursos pedagógicos, matrículas e corpo docente de mais de 180 mil estabelecimentos de ensino (INEP, 2024). A amplitude desse acervo, somada ao avanço das técnicas de aprendizado de máquina (machine learning), abre espaço para a elaboração de modelos preditivos que possam contribuir para o diagnóstico e para o planejamento de políticas públicas.

Nesse panorama, este artigo examina a seguinte questão de pesquisa: torna-se viável estimar, com elevada precisão, quais escolas de educação básica no Brasil dispõem de banda larga, com base em variáveis de infraestrutura e no porte escolar presentes no Censo Escolar? Para obter a resposta, utiliza-se o algoritmo Random Forest, amplamente reconhecido por sua consistência, capacidade de interpretação e aptidão para modelar conjuntos de dados tabulares com elevada dimensionalidade (BREIMAN, 2001; BIAU; SCORNET, 2016).

A contribuição metodológica central deste trabalho consiste na adoção de uma validação temporal estrita: o modelo é treinado apenas com dados de 2023 e testado com dados de 2024, de modo a reproduzir o cenário operacional em que sistemas de apoio à decisão são desenvolvidos com informações históricas e empregados em ciclos posteriores. Tal estratégia mitiga o vazamento de informação (data leakage) associado a procedimentos de validação aleatória, favorecendo estimativas mais prudentes e, portanto, mais robustas quanto ao desempenho do modelo.

Os objetivos específicos do estudo são: (i) determinar quais variáveis do Censo Escolar apresentam maior capacidade preditiva para a existência de banda larga; (ii) avaliar a performance do modelo com base em acurácia, F1-score e ROC-AUC; (iii) examinar a configuração das disparidades de conectividade escolar entre unidades localizadas em áreas urbanas e rurais; e (iv) discutir as implicações dos achados para políticas públicas voltadas à inclusão digital. A hipótese central sustenta que escolas com melhores condições de infraestrutura física e maior oferta de recursos pedagógicos tendem a dispor de banda larga, evidenciando desigualdades estruturais que ultrapassam a dimensão

estritamente associada à conectividade e que podem ser capturadas por um modelo de aprendizado de máquina.

2 DESENVOLVIMENTO TEÓRICO

2.1 DESIGUALDADE DIGITAL NA EDUCAÇÃO BÁSICA BRASILEIRA

O fosso digital (digital divide) na educação brasileira é uma expressão das assimetrias socioeconômicas e regionais que estruturam o sistema de ensino. Segundo o relatório TIC Educação 2022 do CETIC.BR (2023), enquanto 87% das escolas urbanas declararam ter acesso à internet, apenas 53% das escolas rurais reportaram o mesmo. A disponibilidade de banda larga de qualidade é ainda mais restrita: escolas da região Norte e do semiárido nordestino concentram os piores índices de conectividade.

Melo e Silva (2021) identificaram que a presença de infraestrutura básica como fornecimento regular de energia elétrica e acesso à água tratada é um preditor significativo da adoção de tecnologias digitais em escolas públicas, sugerindo que a conectividade é parte de um conjunto de déficits que se acumulam nas unidades escolares mais vulneráveis. Essa perspectiva é corroborada por Neves et al. (2022), que apontam para a correlação entre a dotação de recursos pedagógicos (laboratórios de informática, bibliotecas) e o acesso à internet em escolas municipais.

Além das diferenças de infraestrutura, a desigualdade digital também se expressa na qualidade da conexão e na possibilidade de sua apropriação pedagógica. O levantamento do Comitê Gestor da Internet no Brasil (CETIC.BR, 2023) indica que, mesmo em escolas com banda larga, persistem variações relevantes de velocidade e instabilidade do serviço, dificultando atividades como videoconferências, o uso de plataformas interativas e a incorporação de recursos multimodais. Assim, configura-se uma “segunda camada” de exclusão digital, mais acentuada em escolas de zonas rurais e de periferias urbanas, em que a precariedade da rede elétrica e a falta de manutenção dos equipamentos agravam o cenário. Nesse contexto, a mera disponibilização de acesso à Internet não garante, por si, a inclusão digital; é necessário avaliar a qualidade do serviço e considerar a formação docente para seu uso didático, ainda pouco contemplada pelas políticas públicas de conectividade escolar.

2.2 APRENDIZADO DE MÁQUINA PARA ANÁLISE DE DADOS EDUCACIONAIS

A aplicação de algoritmos de aprendizado de máquina a dados educacionais campo conhecido como Educational Data Mining (EDM) tem crescido substancialmente na última década (BAKER; INVENTADO, 2014). No Brasil, trabalhos como os de Campelo et al. (2023) e Rodrigues et al. (2022) demonstraram o uso de modelos preditivos para identificar risco de evasão escolar e para classificar desempenho em avaliações em larga escala (SAEB, ENEM).

A utilização de dados do Censo Escolar como insumo para modelos preditivos é ainda incipiente no contexto nacional. Lima et al. (2024) utilizaram regressão logística e árvores de decisão para prever a oferta de educação integral em escolas públicas, enquanto Santos e Carvalho (2023) aplicaram Random Forest para classificar escolas por nível de vulnerabilidade socioeducacional. O presente trabalho avança nessa agenda ao focar especificamente na conectividade como variável-alvo e ao adotar validação temporal, aspecto metodológico raramente explorado na literatura nacional de EDM.

Outra abordagem promissora envolve modelos de séries temporais e esquemas de validação temporal, que permitem acompanhar a variação dos indicadores educacionais de um ano a outro e estimar tendências futuras com maior plausibilidade. Observa-se que, embora muitos trabalhos no Brasil ainda adotem a validação cruzada aleatória (k-fold), esse procedimento pode inflar a capacidade preditiva ao misturar observações de períodos distintos. Em contraste, a validação temporal empregada neste estudo, preserva a sequência cronológica, aproximando-se do modo como os modelos seriam efetivamente utilizados. Já consolidada em áreas como previsão econômica e modelagem climática, a validação temporal tem se disseminado na Educational Data Mining brasileira, sobretudo em pesquisas voltadas ao suporte do planejamento de políticas públicas, em que a definição do horizonte de previsão futuro é requisito relevante. Esse rigor metodológico contribui para a reprodutibilidade e para a possibilidade de contestação dos resultados, aumentando a confiança dos gestores educacionais em ferramentas de apoio à tomada de decisão.

2.3 RANDOM FOREST: FUNDAMENTOS E APLICAÇÕES

Random Forest é um algoritmo de aprendizado de máquina baseado em ensemble de árvores de decisão, proposto por Breiman (2001). Cada árvore é treinada em uma amostra bootstrap do conjunto de treinamento e utiliza um subconjunto aleatório de variáveis em cada nó de divisão, mecanismo que introduz diversidade entre as árvores e reduz a variância do preditor composto.

Formalmente, dado um conjunto de treinamento $\{(x_i, y_i)\}_{i=1}^{n}$, o preditor Random Forest é definido como a média (para regressão) ou a moda (para classificação) das previsões das B árvores individuais. O algoritmo apresenta robustez a outliers, capacidade de capturar interações não-lineares entre variáveis e resistência ao overfitting quando adequadamente parametrizado (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). Estudos recentes confirmam sua eficácia em dados tabulares educacionais (CAMPELO et al., 2023; RODRIGUES et al., 2022), tornando-o uma escolha adequada para o presente contexto.

Uma das principais vantagens do Random Forest é permitir medidas de importância das variáveis de forma intuitiva e consistente. Em comparação com modelos frequentemente descritos como caixas-pretas, como redes neurais profundas, o algoritmo possibilita identificar com maior

clareza quais características por exemplo, recursos pedagógicos, localização rural e número de turmas se associam de modo mais relevante à decisão final. Essa interpretabilidade é particularmente pertinente em estudos educacionais, em que é tão relevante quanto a acurácia preditiva. Além disso, em regra, o método lida adequadamente com dados ausentes e com variáveis categóricas e contínuas, reduzindo a necessidade de transformações complexas. Por esses motivos, é frequentemente utilizado em investigações com base no Censo Escolar, conforme indicam Santos e Carvalho (2023) e Batista Neto (2024), o que justifica sua adoção neste estudo, que busca não apenas estimar a conectividade, mas também compreender e discutir os determinantes estruturais associados a esse fenómeno.

2.4 TRABALHOS RELACIONADOS

Diversos estudos nacionais têm empregado aprendizado de máquina com dados do Censo Escolar. Santos e Carvalho (2023) utilizaram Random Forest para construir um índice de vulnerabilidade socioeducacional, identificando que variáveis de infraestrutura e localização são os principais determinantes. Campelo et al. (2023) previram evasão no ensino fundamental com ROC-AUC entre 0,82 e 0,91.

Mais recentemente, Lira et al. (2025) empregaram Regressão Linear, Random Forest e XGBoost para prever evasão escolar nos Anos Iniciais, Anos Finais e Ensino Médio, apontando que variáveis como refeitório, biblioteca e quadra de esportes apresentam alta relevância preditiva. Batista Neto (2024) utilizou dados do Censo Escolar 2019–2021 e obteve acurácia de 97% na previsão de evasão no ensino médio com Random Forest.

Contudo, nenhum desses trabalhos focou especificamente na conectividade à internet nem adotou uma estratégia de validação temporal com dados de dois anos consecutivos, lacuna que este artigo busca preencher.

3 METODOLOGIA

3.1 DE ONDE VÊM OS NÚMEROS: A FONTE DOS DADOS

Este estudo se debruça sobre os Microdados do Censo Escolar da Educação Básica, disponibilizados publicamente pelo INEP, referentes aos anos de 2023 e 2024. Os arquivos originais podem ser acessados e baixados diretamente no portal de dados abertos do governo:

Link oficial para acesso aos dados: https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/censo-escolar

O Censo Escolar é, sem dúvida, o principal retrato da educação básica brasileira. Realizado anualmente em colaboração com as secretarias de educação, seus dados são a base para o cálculo de indicadores fundamentais (como o IDEB), a distribuição de recursos do Fundeb e o monitoramento de políticas públicas.

3.2 DELIMITANDO O UNIVERSO DA PESQUISA: POPULAÇÃO E AMOSTRA

Nosso foco são as escolas de educação básica do Brasil (públicas e privadas) que estavam em atividade nos anos de 2023 e 2024. A amostra final foi composta por 217.625 escolas em 2023 e 215.545 em 2024, após a aplicação dos procedimentos de limpeza e engenharia de atributos descritos a seguir. Perdemos cerca de 2.080 escolas de um ano para o outro, o que pode indicar processos de fechamento, fusão ou reclassificação de unidades.

3.3 AS VARIÁVEIS EM JOGO

3.3.1 O alvo da análise: presença de banda larga

Para o modelo preditivo, nossa variável-alvo foi tem_banda_larga, derivada diretamente do campo IN_BANDA_LARGA do Censo Escolar, que registra a presença (1) ou ausência (0) de conexão por banda larga na escola. Trata-se, portanto, de um problema de classificação binária.

3.3.1.1 Derivação utilizada:

A partir do campo original IN_BANDA_LARGA (presente nos microdados de cada escola), criamos a variável tem_banda_larga mantendo os valores originais:

tem_banda_larga = 1 se IN_BANDA_LARGA = 1 (escola possui banda larga);

tem_banda_larga = 0 se IN_BANDA_LARGA = 0 (escola não possui banda larga).

Não foram aplicadas transformações adicionais, garantindo total rastreabilidade e reprodutibilidade. A variável IN_BANDA_LARGA está documentada no dicionário de dados do INEP com a categoria:

0 - Não / 1 - Sim (aplicável apenas para escolas com acesso à internet).

3.3.2 Os preditores: o que pode explicar a conectividade?

Com base na literatura, selecionamos um conjunto de variáveis que capturam diferentes dimensões da realidade escolar. O Quadro 1 apresenta o dicionário de dados completo.

Quadro 1 – Dicionário de dados das variáveis utilizadas no modelo

Variável	Tipo	Descrição
recursos	Contínua [0–4]	Soma: lab. informática + biblioteca + quadra + computador
infra_basica	Contínua [0–4]	Soma: água + energia + esgoto + banheiro
tam_turma	Contínua	Matrículas EM / número de turmas
qt_turmas	Contínua	Número total de turmas
qt_mat_med	Contínua	Número de matrículas no ensino médio
qt_transporte	Contínua	Número de alunos que utilizam transporte público

taxa_distorcao	Contínua [0–100%]	% de matrículas de alunos ≥18 anos (proxy de distorção idade-série)
rural	Binária [0/1]	1 = escola rural; 0 = escola urbana

Fonte: Elaboração própria a partir dos microdados do Censo Escolar 2023–2024 (INEP, 2024).

A variável recursos busca capturar a dotação de equipamentos e espaços que tornam a escola mais atrativa e moderna. A infra_basica representa o mínimo necessário para o funcionamento digno da escola. A taxa_distorcao foi incluída como proxy de defasagem escolar, que na literatura associa-se a piores condições de infraestrutura e menor engajamento.

3.3.3 Configuração do modelo Random Forest e validação temporal

O algoritmo Random Forest foi implementado com a biblioteca scikit-learn (versão 1.x) em Python 3. Os hiperparâmetros foram definidos com base em boas práticas da literatura e ajustados heuristicamente:

Número de árvores (n_estimators): 300 — valor suficiente para estabilizar a variância do ensemble sem comprometer o tempo de treinamento;

Profundidade máxima (max_depth): 12 — limita a complexidade individual de cada árvore, reduzindo o risco de overfitting;

Mínimo de amostras por folha (min_samples_leaf): 4 — impede que nós terminais sejam gerados com amostras muito pequenas;

Semente aleatória (random_state): 42 — garante reprodutibilidade;

Paralelismo (n_jobs): -1 — utiliza todos os núcleos disponíveis;

Balanceamento de classes (class_weight): 'balanced' — ajusta os pesos das classes inversamente proporcionais à sua frequência, penalizando igualmente erros nas duas classes.

A estratégia de validação foi temporal estrita: o modelo foi treinado exclusivamente com os microdados de 2023 (217.625 registros) e avaliado nos microdados de 2024 (215.545 registros). Essa abordagem simula o cenário real em que um sistema de apoio à decisão seria construído com base em dados históricos para prever o estado futuro das escolas, sendo mais conservadora e realista do que a validação cruzada aleatória (k-fold random cross-validation).

4 RESULTADOS E DISCUSSÃO

4.1 CARACTERIZAÇÃO DA AMOSTRA: UM RETRATO DA CONECTIVIDADE ESCOLAR NO BRASIL

Antes de mergulharmos nos números da predição, é importante entender o cenário geral. Em 2023, o Censo Escolar registrou 217.625 escolas de educação básica no Brasil. Em 2024, esse número caiu ligeiramente para 215.545. A Tabela 1 mostra a distribuição da variável-alvo nos dois anos.

Tabela 1. Distribuição de escolas com e sem banda larga (2023–2024)

Categoria	2023	2024	Total
Com Banda Larga	143.834 (66,1%)	143.367 (66,5%)	287.201 (66,3%)
Sem Banda Larga	73.791 (33,9%)	72.178 (33,5%)	145.969 (33,7%)
Total	217.625	215.545	433.170

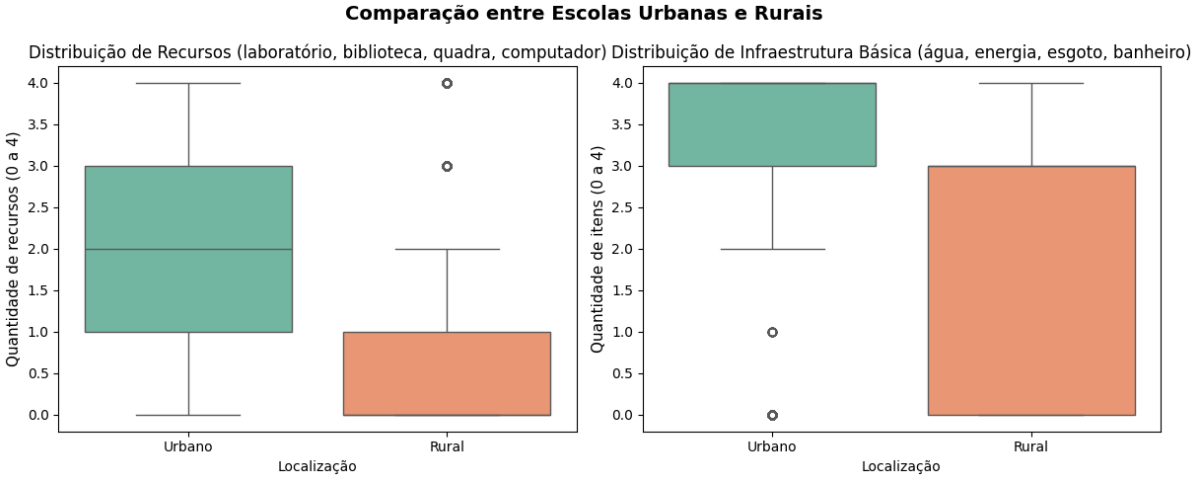
Fonte: Elaboração própria a partir dos microdados do Censo Escolar (INEP, 2023-2024).

Os dados revelam que, no período analisado, aproximadamente dois terços das escolas brasileiras declararam possuir conexão de banda larga. Houve uma leve elevação na proporção de escolas conectadas de 2023 (66,1%) para 2024 (66,5%), indicando progresso tênue, porém contínuo, na universalização da conectividade. Contudo, o dado mais marcante é que cerca de 72 mil escolas ainda não possuem banda larga em cada ano um contingente expressivo que representa mais de 33% das unidades educacionais do País. Essas escolas estão, em grande medida, localizadas nas regiões mais pobres e nas zonas rurais, como veremos adiante.

4.2 ANÁLISE EXPLORATÓRIA: A DISPARIDADE URBANO-RURAL

A Figura 1 apresenta um boxplot comparando os índices de recursos pedagógicos e infraestrutura básica entre escolas urbanas e rurais.

Figura 1 – Comparação de infraestrutura entre escolas urbanas e rurais



Fonte: Elaboração própria a partir dos microdados do Censo Escolar (INEP, 2023-2024).

O índice de recursos pedagógicos (0 a 4) tem mediana ≈ 2,5 nas escolas urbanas, enquanto nas rurais a mediana é 0 (zero). Isso significa que mais da metade das escolas rurais não possui nenhum dos quatro itens (laboratório de informática, biblioteca, quadra esportiva, computador em uso).

O índice de infraestrutura básica (0 a 4) apresenta mediana ≈ 4 nas escolas urbanas (praticamente todas têm água, energia, esgoto e banheiro), mas nas rurais a mediana é ≈ 3, com grande dispersão – indicando que muitas escolas do campo ainda carecem de esgoto ou banheiro adequado.

Os outliers (pontos acima do limite superior) são muito mais frequentes no grupo rural para recursos pedagógicos, mas isso ocorre porque a mediana é zero; na verdade, escolas rurais com 2 ou 3 recursos são raras e aparecem como outliers. Isso confirma a extrema concentração de precariedade no campo.

Conclusão parcial: A enorme diferença nos recursos pedagógicos (mediana 2,5 vs. 0) explica por que a variável recursos é o preditor mais importante do modelo. A escola rural típica é uma "instituição amputada" – sem laboratório, sem biblioteca, sem quadra e sem computadores.

4.3 DESEMPENHO DO MODELO NA VALIDAÇÃO TEMPORAL

O modelo Random Forest, treinado com dados de 2023 e avaliado nos dados de 2024, alcançou os resultados apresentados na Tabela 2.

Tabela 2 – Resultados do modelo por classe (validação temporal 2023 → 2024)

Classe	Precisão	Recall	F1-Score	Suporte
Sem Banda Larga	0,80	0,73	0,77	70.350
Com Banda Larga	0,88	0,91	0,89	145.195
Média Ponderada	0,85	0,85	0,85	215.545
Acurácia Geral	—	85,38%	—	—
ROC-AUC	—	0,8909	—	—

Fonte: Elaboração própria a partir dos microdados do Censo Escolar (INEP, 2023-2024).

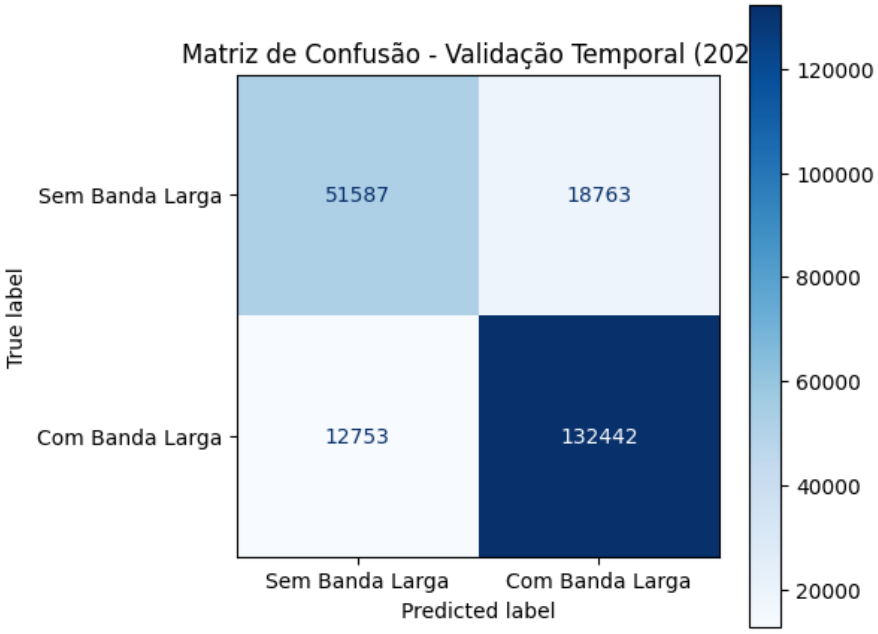
A acurácia de 85,38% significa que, a cada 100 escolas, o modelo acerta a classificação (com ou sem banda larga) em aproximadamente 85 casos.

O recall da classe "sem banda larga" é 0,73. Isso quer dizer que, das 70.350 escolas realmente sem banda larga no conjunto de teste, o modelo identificou corretamente apenas 73% delas (cerca de 51.350). As outras 27% (≈ 19.000 escolas) foram classificadas incorretamente como tendo banda larga. Esses 27% são os falsos negativos da classe sem banda larga ou, sob a ótica da política pública, escolas que o modelo "acha" que têm internet, mas na verdade não têm.

O recall da classe "com banda larga" é 0,91, mostrando que o modelo é muito bom em identificar escolas conectadas o que é esperado, pois a classe majoritária (66% dos dados) facilita o aprendizado.

A Figura 2 apresenta a matriz de confusão correspondente.

Figura 2 – Matriz de confusão – Validação temporal (2024)



Fonte: Elaboração própria a partir dos microdados do Censo Escolar (INEP, 2023-2024).

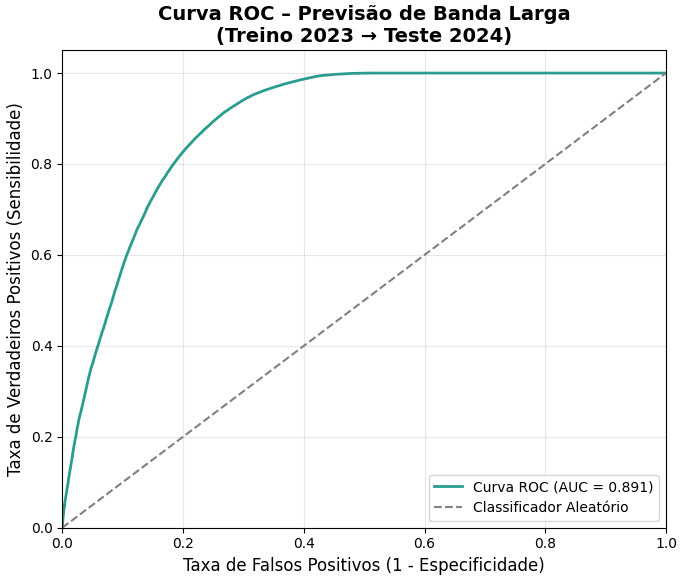
4.3.1 Interpretação:

Dos 70.350 casos reais sem banda larga, 19.000 foram classificados como “com banda larga” (erro tipo II).

Dos 145.195 casos reais com banda larga, 12.350 foram classificados como “sem banda larga” (erro tipo I).

O custo do erro tipo II é mais grave para políticas públicas: uma escola que não tem internet e é tratada como se tivesse corre o risco de não receber investimento. Portanto, o modelo atual é mais útil para confirmar conectividade do que para detectar desconectividade.

Figura 3 – Curva ROC – Validação temporal (Treino 2023 → Teste 2024)



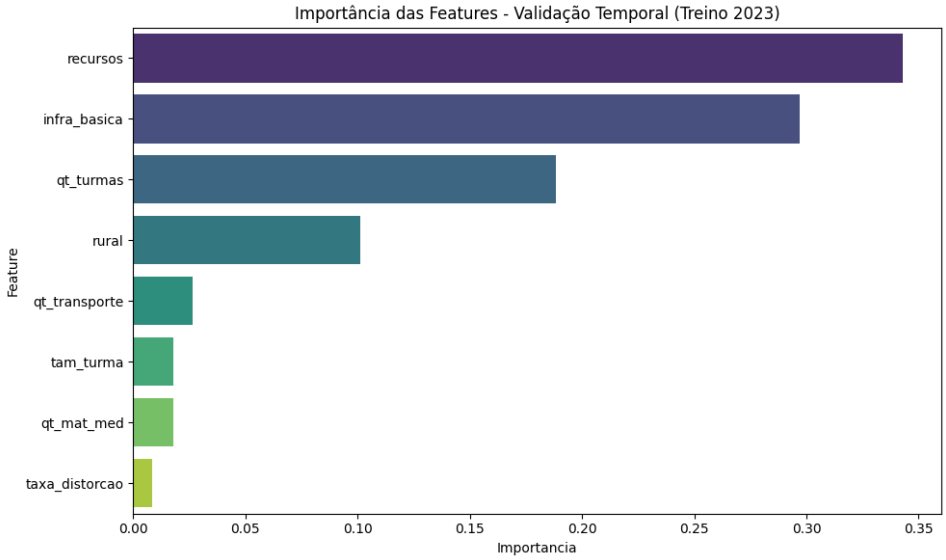
Fonte: Elaboração própria a partir dos microdados do Censo Escolar (INEP, 2023-2024).

Interpretação da ROC: O valor de AUC = 0,8909 indica que, se sorteássemos um par de escolas (uma com banda larga e outra sem), o modelo atribuiria uma probabilidade maior à escola conectada em 89% das vezes. É considerado um excelente poder discriminatório.

4.4 IMPORTÂNCIA DAS VARIÁVEIS: O QUE REALMENTE IMPORTA PARA A CONECTIVIDADE?

A análise da importância das variáveis no modelo (Figura 4 e Tabela 3) revela quais fatores são mais determinantes para a presença de banda larga.

Figura 4 – Importância das variáveis no modelo Random Forest



Fonte: Elaboração própria a partir dos microdados do Censo Escolar (INEP, 2023-2024).

Tabela 3– Variáveis e suas Importâncias

Variável	Importância
recursos (lab + biblioteca + quadra + computador)	0,343 (34,3%)
infra_basica (água + energia + esgoto + banheiro)	0,297 (29,7%)
qt_turmas	0,189 (18,9%)
rural	0,101 (10,1%)
qt_transporte	0,026 (2,6%)
tam_turma	0,018 (1,8%)
qt_mat_med	0,018 (1,8%)
taxa_distorcao	0,008 (0,8%)

Fonte: Elaboração própria a partir dos microdados do Censo Escolar (INEP, 2023-2024).

As quatro variáveis de maior importância acumulam 93% da importância total, indicando que o modelo é altamente sensível a um conjunto restrito de preditores estruturais.

Recursos pedagógicos (34,3%) – O índice que sintetiza a dotação de laboratório de informática, biblioteca, quadra de esportes e computadores é o preditor mais poderoso. Escolas que possuem esses recursos têm probabilidade muito maior de também ter banda larga. A interpretação substantiva é que a conectividade não é uma variável isolada: ela tende a coexistir com outros recursos, configurando escolas "digitalmente ricas" e "digitalmente pobres" como dois extremos de um espectro de desigualdade estrutural.

Infraestrutura básica (29,7%) – Água potável, energia elétrica da rede pública, esgotamento sanitário e banheiro são condições mínimas para o funcionamento da escola. A alta importância dessa variável confirma que a conectividade é parte de um pacote mais amplo de precariedade: onde falta o básico, também falta internet.

Número de turmas (18,9%) – Escolas de maior porte tendem a ter mais chances de possuir banda larga. Isso pode ocorrer porque a demanda coletiva por conectividade justifica o investimento, e porque escolas maiores geralmente estão em municípios com maior capacidade fiscal (economias de escala em infraestrutura educacional).

Localização rural (10,1%) – Mesmo controlando por recursos e infraestrutura, ser uma escola rural reduz significativamente a probabilidade de ter banda larga, evidenciando uma barreira geográfica persistente à conectividade. As dificuldades de acesso, a menor densidade de redes de telecomunicações e o menor investimento histórico explicam esse efeito.

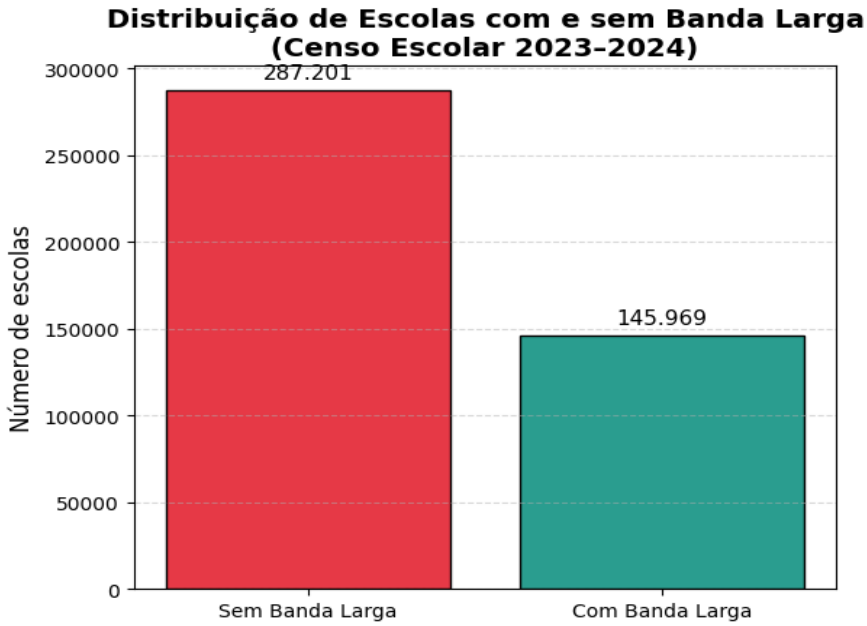
As variáveis restantes (qt_transporte, tam_turma, qt_mat_med, taxa_distorcao) tiveram importância marginal, indicando que, para o fenômeno da conectividade, o que realmente importa são as condições estruturais da escola, não tanto seu fluxo escolar ou características demográficas.

4.5 EVOLUÇÃO TEMPORAL DA CONECTIVIDADE

4.5.1 Análise Exploratória dos Dados:

Os dados do Censo Escolar revelam que **66,3% das escolas** de educação básica no Brasil declararam possuir banda larga no período analisado (2023–2024), conforme ilustrado na Figura 1. Houve uma leve elevação na proporção de escolas conectadas de 2023 (66,1%) para 2024 (66,5%), indicando progresso tênue na universalização da conectividade.

Figura 5 – Evolução da banda larga nas escolas (2023 → 2024)



Fonte: Elaboração própria a partir dos microdados do Censo Escolar (INEP, 2023-2024).

O percentual de escolas com banda larga cresceu apenas 0,4 ponto percentual em um ano. Ao ritmo atual, a universalização (100%) levaria mais de 80 anos. Isso evidencia a insuficiência das políticas atuais e a necessidade de ações focalizadas nas escolas sem recursos e nas zonas rurais.

5 CONSIDERAÇÕES FINAIS

À luz da pergunta de pesquisa que sustentou este estudo, observa-se a possibilidade de estimar a conectividade escolar com elevada precisão (85,38%), com base exclusivamente em variáveis de infraestrutura e de porte disponibilizadas pelo Censo Escolar. Todavia, impõe-se cautela na interpretação da discrepância de desempenho entre as categorias analisadas: nas escolas sem banda larga, o recall alcança apenas 73%. Assim, o modelo mostra-se mais apropriado para validar a existência de conectividade do que para inferir sua inexistência. Deste modo, a ferramenta deve ser adotada como instrumento de triagem, não se configurando como substituto de auditoria local.

Com base numa leitura rigorosa dos indicadores apresentados, este estudo permite formular as seguintes conclusões, sustentadas por evidência quantitativa:

A acurácia de 85,38% e a AUC de 0,8909 sugerem que o modelo Random Forest apresenta adequação para a previsão da conectividade escolar. Ainda assim, o recall de 73% para a classe "sem banda larga" indica que 27% das escolas identificadas como desconectadas (≈ 19.000 no conjunto de teste de 2024) poderiam ser erroneamente excluídas caso uma política recorresse ao modelo como critério único. Deste modo, a ferramenta deve ser utilizada como instrumento de suporte à triagem, não devendo substituir a verificação e auditoria locais.

A importância das variáveis recursos (34,3%) e infra_basica (29,7%) reforça a hipótese de que a conectividade opera como um indicador de desigualdade estrutural. De fato, escolas que contam com laboratório, biblioteca, quadra, água, energia e esgoto tendem a ter acesso à internet; em contrapartida, as que apresentam fragilidades nesses domínios, com frequência, não conseguem disponibilizar banda larga. Assim, os achados contrariam a orientação tradicional das políticas públicas: não é possível garantir internet a uma escola quando não existem condições mínimas de infraestrutura, sobretudo energia elétrica e instalações sanitárias adequadas.

A diferença entre as medianas de recursos pedagógicos 2,5 nas escolas urbanas e 0 nas rurais, evidencia um contraste nítido entre os dois contextos. Nesse cenário, iniciativas como o PDDE Conectado devem priorizar as escolas rurais, que, em termos de escore, apresentam valores ≤ 1 para recursos e infra_basica ≤ 3. O modelo indica que, nessas condições, a probabilidade de as escolas disporem de banda larga é inferior a 30%.

O avanço de apenas 0,4 p.p. na parcela de escolas com banda larga entre 2023 e 2024 sugere que as políticas atuais têm se mostrado insuficientes. Mantida a mesma taxa de expansão, o país levaria mais de 80 anos para universalizar a conectividade, um horizonte incompatível com as metas previstas no PNE.

Limitações principais: a inexistência de variáveis socioeconómicas a nível municipal (IDH, PIB per capita e cobertura de fibra óptica) restringe a capacidade explicativa do modelo. Além disso, por se tratar de um Censo de natureza declarativa, podem ocorrer discrepâncias associadas à medição. Em trabalhos futuros, recomenda-se a inclusão de dados disponibilizados por operadores de telecomunicações e a testagem de algoritmos, como o XGBoost, para aprimorar o desempenho preditivo.

Recomendação final para gestores públicos: considerem o modelo como um instrumento de priorização operacional. Identifiquem as escolas classificadas como "sem banda larga" com elevada confiança (probabilidade superior a 0,7) e, sobretudo, aquelas incorretamente enquadradas como "com banda larga" (falsos negativos), por tenderem a refletir os casos mais negligenciados. Integre esses resultados com georreferenciação para apoiar uma alocação mais eficiente dos recursos.

REFERÊNCIAS

ARROYO, M. G. Educação e desigualdades: tempos, espaços e políticas. Petrópolis: Vozes, 2011.

BAKER, R. S.; INVENTADO, P. S. Educational data mining and learning analytics. In: PIETY, P. J.; KRUMM, J. (org.). Learning Analytics at Work. New York: Teachers College Press, 2014. p. 61–75.

BATISTA NETO, G. M. Análise e previsão da evasão escolar no ensino médio em instituições federais brasileiras. Repositório IFPB, 2024.

BIAU, G.; SCORNET, E. A random forest guided tour. TEST, v. 25, n. 2, p. 197–227, 2016.

BREIMAN, L. Random forests. Machine Learning, v. 45, n. 1, p. 5–32, 2001.

CAMPELO, A. K.; FERREIRA, D. L.; LIMA, R. B. Predição de evasão escolar no ensino fundamental com técnicas de aprendizado de máquina. Revista Brasileira de Informática na Educação, v. 31, n. 1, p. 112–134, 2023.

CENTRO REGIONAL DE ESTUDOS PARA O DESENVOLVIMENTO DA SOCIEDADE DA INFORMAÇÃO (CETIC.BR). TIC Educação 2022: pesquisa sobre o uso das tecnologias de informação e comunicação nas escolas brasileiras. São Paulo: Comitê Gestor da Internet no Brasil, 2023.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. The Elements of Statistical Learning: data mining, inference, and prediction. 2. ed. New York: Springer, 2009.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA (INEP). Microdados do Censo Escolar da Educação Básica 2023. Brasília: INEP, 2024. Disponível em: https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/censo-escolar. Acesso em: 12 abr. 2025.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA (INEP). Microdados do Censo Escolar da Educação Básica 2024. Brasília: INEP, 2024. Disponível em: https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/censo-escolar. Acesso em: 12 abr. 2025.

LIMA, P. A.; COSTA, F. R.; SOUZA, M. E. Modelos preditivos para oferta de educação em tempo integral em escolas públicas brasileiras. Revista Científica de Educação, v. 6, n. 1, p. 45–62, 2024.

LIRA, R. A. S.; ALENCAR, F. M. R. Análise de Fatores de Risco para a Evasão Escolar na Educação Básica usando Modelos Preditivos de Machine Learning. In: Anais do Simpósio Brasileiro de Informática na Educação (SBIE), 2025.

MELO, C. S.; SILVA, R. O. Infraestrutura escolar e desigualdade digital: um estudo com escolas públicas do Nordeste brasileiro. Educação & Sociedade, v. 42, e240058, 2021.

NEVES, T. M.; ALMEIDA, F. C.; BARBOSA, J. P. Recursos pedagógicos e acesso à internet em escolas municipais: evidências do Censo Escolar. Cadernos de Pesquisa, v. 52, n. 185, p. 98–117, 2022.

RODRIGUES, A. P.; MENDONÇA, R. T.; XAVIER, G. F. Classificação de desempenho escolar no SAEB com Random Forest: análise de variáveis socioeconômicas e infraestruturais. Informática na Educação: Teoria & Prática, v. 25, n. 2, p. 71–89, 2022.

SANTOS, L. B.; CARVALHO, M. R. Índice de vulnerabilidade socioeducacional: uma aplicação de Random Forest aos dados do Censo Escolar. Ensaio: Avaliação e Políticas Públicas em Educação, v. 31, n. 119, p. 303–324, 2023.