Análise e Previsão da Evasão Escolar na Educação Infantil: Uma Abordagem de Séries Temporais com Foco em Políticas Públicas

(3ª Etapa)

Ana Cláudia de Lima Aleixo[1], Ana Carolina Martins[2],

Eduardo David[3], Gustavo Santiago[4]

[1]Faculdade de Computação e Informática (FCI)
Universidade Presbiteriana Mackenzie – São Paulo, SP – Brasil

10414992@mackenzie.br,
10174844@mackenzie.br,

10414643@mackenzie.br,

10414643@mackenzie.br.

Resumo: Este projeto tem como objetivo analisar e prever a evasão escolar na Educação Infantil, com foco em creches e pré-escolas da rede pública brasileira. A pesquisa considera recortes por tipo de turno (integral ou parcial), rede de ensino (estadual ou municipal) e localização (urbana ou rural), a fim de compreender padrões regionais e identificar cenários críticos. A partir de dados do Censo Escolar do INEP (2015–2023), o estudo utiliza técnicas de séries temporais combinadas com algoritmos de aprendizado de máquina para realizar previsões e apoiar a formulação de políticas públicas voltadas à permanência escolar. A análise considera que fatores estruturais e socioeconômicos influenciam fortemente a evasão, especialmente em contextos de vulnerabilidade. Como contribuição extensionista, o projeto se alinha ao ODS 4 da ONU, promovendo educação inclusiva e de qualidade. Espera-se que os resultados possam orientar gestores públicos na implementação de ações mais eficazes e baseadas em evidências, além de fomentar o uso de tecnologia e ciência de dados na educação pública brasileira.

Palavras-chave: Evasão Escolar, Educação Infantil, Séries Temporais

Abstract: *This project aims to analyze and forecast school dropout rates in Early Childhood Education, focusing on public daycare and preschool institutions in Brazil. The study segments data by school shift (full-time or part-time), school network (state or municipal),*

and location (urban or rural), in order to understand regional patterns and identify critical dropout scenarios. Using data from the INEP School Census (2015–2023), the project applies time series techniques combined with machine learning algorithms to make predictions and support public policies that promote school retention. The analysis acknowledges that structural and socioeconomic factors strongly influence dropout rates, especially in vulnerable contexts. As a community-oriented initiative, the project aligns with the UN's SDG 4, promoting inclusive and quality education. The expected outcome is to provide actionable insights for public education managers and encourage the adoption of data science tools in educational planning and decision-making.

Keywords: *School Dropout, Early Childhood Education, Time Series*

SUMÁRIO

1. Introdução

A evasão escolar na Educação Infantil representa um desafio estrutural para o sistema educacional brasileiro, pois impacta diretamente o desenvolvimento integral das crianças e compromete sua permanência nas etapas seguintes da educação básica. A Educação Infantil — composta por creches e pré-escolas — é a base para o desenvolvimento cognitivo, afetivo e social. No entanto, estudos mostram que crianças de contextos vulneráveis, especialmente nas redes públicas e áreas rurais, têm maior risco de interrupção precoce da trajetória escolar (Silva et al., 2020).

O Censo Escolar, realizado anualmente pelo INEP, apresenta dados sobre matrículas, fluxo e rendimento dos estudantes da Educação Básica no Brasil. Segundo o relatório de 2023, "embora haja avanços na ampliação do acesso, as taxas de abandono permanecem expressivas em determinados contextos, especialmente na educação infantil em tempo parcial e em áreas rurais" (INEP, 2023, p. 14). Isso revela a urgência de análises segmentadas e preditivas que possam embasar decisões políticas mais precisas e eficazes.

A dificuldade em acompanhar a evasão escolar com granularidade suficiente — como por etapa de ensino, turno e rede — limita a capacidade de intervenção estratégica dos gestores públicos. Além disso, a maior parte dos estudos existentes se concentra no Ensino Fundamental e Médio, deixando a Educação Infantil em segundo plano. No entanto, como destacam Costa e Lima (2021), “os primeiros anos de vida são determinantes para a formação de habilidades cognitivas e emocionais que sustentam o sucesso escolar futuro” (p. 89). A ausência de políticas focalizadas neste estágio compromete não só o direito à educação, mas o desenvolvimento humano em longo prazo.

O uso de séries temporais combinadas com técnicas de aprendizado de máquina tem se mostrado eficaz para entender padrões educacionais e antecipar cenários críticos. Modelos como Prophet e LSTM têm sido empregados em estudos para prever matrículas, evasão e desempenho escolar, permitindo respostas mais rápidas e eficientes por parte dos órgãos responsáveis (Ferreira et al., 2022). A proposta deste projeto, portanto, é não apenas identificar tendências históricas, mas criar um produto analítico que possa ser reutilizado e adaptado por secretarias de educação, ONGs e gestores escolares para promover permanência, equidade e qualidade na Educação Infantil.

Neste contexto, este projeto busca analisar séries temporais relacionadas à evasão escolar na Educação Infantil, considerando a segmentação por tipo de turno (parcial ou

integral), rede de ensino (estadual ou municipal) e localização (urbana ou rural). O objetivo é aplicar modelos de aprendizado de máquina para prever padrões futuros e apoiar políticas públicas baseadas em evidências. A análise contribuirá para os Objetivos de Desenvolvimento Sustentável, em especial o ODS 4, que visa "assegurar a educação inclusiva, equitativa e de qualidade" até 2030 (ONU, 2015).

A evasão escolar na Educação Infantil é um fenômeno silencioso e, muitas vezes, negligenciado pelas políticas públicas educacionais. Embora os dados do Censo Escolar evidenciem avanços no acesso à creche e à pré-escola, a permanência e a continuidade da criança no ambiente educacional ainda são desafiadoras, especialmente em contextos vulneráveis. A ausência de dados consolidados e análises preditivas nesse segmento limita a ação do Estado e fragiliza o direito à educação na primeira infância, fase determinante para o desenvolvimento humano.

Este projeto nasce da percepção de que a evasão escolar, quando diagnosticada precocemente, pode ser combatida com políticas públicas assertivas, baseadas em evidências e orientadas por dados. A aplicação de modelos de séries temporais permite não apenas compreender a trajetória histórica do problema, mas antecipar cenários críticos e agir preventivamente. Essa abordagem oferece uma vantagem estratégica para gestores públicos, ao possibilitar decisões fundamentadas em projeções reais e segmentadas por rede, turno e localização.

A relevância do tema também está relacionada ao seu impacto social e à necessidade de romper ciclos de desigualdade. Crianças que abandonam a escola precocemente tendem a apresentar maiores dificuldades acadêmicas ao longo da vida, menor inserção no mercado de trabalho e menor acesso a oportunidades. Investir em estratégias de permanência desde a Educação Infantil é, portanto, investir em equidade, inclusão e desenvolvimento sustentável, pilares fundamentais do ODS 4 — Educação de Qualidade.

Além disso, a proposta do projeto atende ao caráter extensionista exigido pela formação acadêmica, ao buscar devolver à sociedade um produto analítico acessível, útil e de aplicação prática. A construção de um modelo preditivo com base em dados públicos primários fortalece a transparência, o controle social e o uso da ciência de dados em benefício do coletivo. Dessa forma, o projeto propõe não apenas uma solução técnica, mas uma ferramenta de transformação social que contribui com a formulação de políticas educacionais mais justas e eficazes.

2. Objetivo

O principal objetivo deste projeto é desenvolver um produto analítico capaz de analisar, interpretar e prever a evasão escolar na Educação Infantil, com base nos históricos do Censo Escolar organizados em séries temporais. A proposta envolve a segmentação dos dados por tipo de turno (parcial ou integral), rede de ensino (estadual ou municipal) e localização geográfica (urbana ou rural), de modo a identificar padrões e comportamentos distintos entre os diferentes contextos educacionais. Com isso, espera-se compreender a evolução da evasão ao longo do tempo e antecipar tendências futuras.

Como metas específicas, o projeto visa: (i) estruturar e limpar os dados educacionais públicos de forma adequada para análise temporal; (ii) aplicar modelos estatísticos e algoritmos de aprendizado de máquina — como Prophet ou LSTM — para geração de previsões; (iii) produzir visualizações e indicadores acessíveis para o público interessado, especialmente gestores educacionais; e (iv) disponibilizar publicamente o produto final como contribuição extensionista alinhada aos Objetivos de Desenvolvimento Sustentável, com destaque para o ODS 4. A intenção é que esse projeto sirva como base para decisões mais informadas, políticas públicas mais eficazes e ações concretas de enfrentamento à evasão escolar desde os primeiros anos da educação básica.

3. Descrição da Base de Dados

O conjunto de dados utilizado neste projeto foi extraído do Censo Escolar da Educação Básica, disponível no portal do INEP (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira). Trata-se de uma base oficial e de fonte primária, composta por dados anuais e declaratórios fornecidos pelas instituições de ensino públicas e privadas de todo o país. Para este estudo, foram utilizados os dados referentes ao período de 2020 a 2024, com foco específico nas etapas da Educação Infantil, contemplando Creche e Pré-escola.

A base contém informações detalhadas sobre o número de matrículas por escola, etapa de ensino, turno (parcial ou integral), rede administrativa (estadual ou municipal) e localização (urbana ou rural). Além disso, inclui indicadores relevantes como taxa de abandono, situação de matrícula, permanência, e códigos territoriais que permitem a agregação dos dados por região, município ou estado. A estrutura dos arquivos segue o

formato CSV (Comma Separated Values), com milhares de registros por ano, organizados em campos padronizados e dicionários de variáveis disponibilizados pelo próprio INEP.

A preparação dos dados exigiu etapas de filtragem, categorização e reestruturação para permitir a análise temporal. Foram selecionadas apenas as instituições públicas de ensino, e os dados foram organizados em séries temporais com granularidade anual. Essa organização possibilita a modelagem estatística e a aplicação de algoritmos de previsão, mantendo a rastreabilidade e a integridade das informações. A escolha dessa base garante confiabilidade, atualidade e alinhamento com os objetivos extensionistas do projeto.

4. Referencial Teórico

A evasão escolar é um dos principais desafios da educação pública brasileira e pode ser compreendida como o abandono da trajetória escolar antes da conclusão da etapa educacional correspondente. Segundo Cunha (2020), "o fenômeno da evasão é multifacetado, envolvendo fatores socioeconômicos, culturais e estruturais que ultrapassam os muros da escola" (p. 112). Esses fatores variam conforme o território, o nível de ensino e o perfil da população atendida.

Apesar de sua importância, a Educação Infantil enfrenta sérios desafios em relação ao acesso, permanência e qualidade, especialmente nas redes públicas. Segundo dados do IBGE (2022), ainda há uma considerável desigualdade no atendimento entre áreas urbanas e rurais, bem como entre redes estaduais e municipais. Essa disparidade reflete não apenas a disponibilidade de vagas, mas também as condições estruturais e pedagógicas das instituições de ensino infantil.

Séries temporais são sequências de dados ordenadas no tempo, geralmente coletadas em intervalos regulares, e são amplamente utilizadas para análise de tendências, sazonalidades e previsões em diversas áreas, incluindo a educação. De acordo com Hyndman e Athanasopoulos (2021), “o uso de séries temporais permite a modelagem de padrões históricos para projetar cenários futuros e orientar decisões estratégicas” (p. 15). Esse tipo de análise é particularmente útil para políticas públicas, pois permite antecipar demandas e prevenir problemas estruturais.

Na educação, a aplicação de modelos de séries temporais tem ganhado espaço por sua capacidade de prever fenômenos como matrículas, taxas de evasão, desempenho

escolar e até alocação de recursos. Modelos estatísticos tradicionais como ARIMA vêm sendo amplamente utilizados, mas, mais recentemente, algoritmos de aprendizado de máquina como LSTM e Prophet têm se mostrado eficazes em contextos educacionais (Ferreira et al., 2022). Essas abordagens oferecem maior flexibilidade na captura de padrões não lineares e de múltiplas variáveis.

No presente projeto, o uso de séries temporais permitirá identificar os momentos de maior risco de evasão, a eficácia de políticas passadas e a projeção de cenários futuros. Ao integrar dados segmentados por rede, tipo de turno e localização, o modelo possibilita uma visão mais rica e útil para a gestão pública. Como destacam Souza e Lima (2020), "a análise temporal em educação ainda é subexplorada, mas tem um potencial imenso para apoiar decisões mais inteligentes e eficazes" (p. 102).

5. Metodologia

5.1 Discursão e Diagrama da Solução

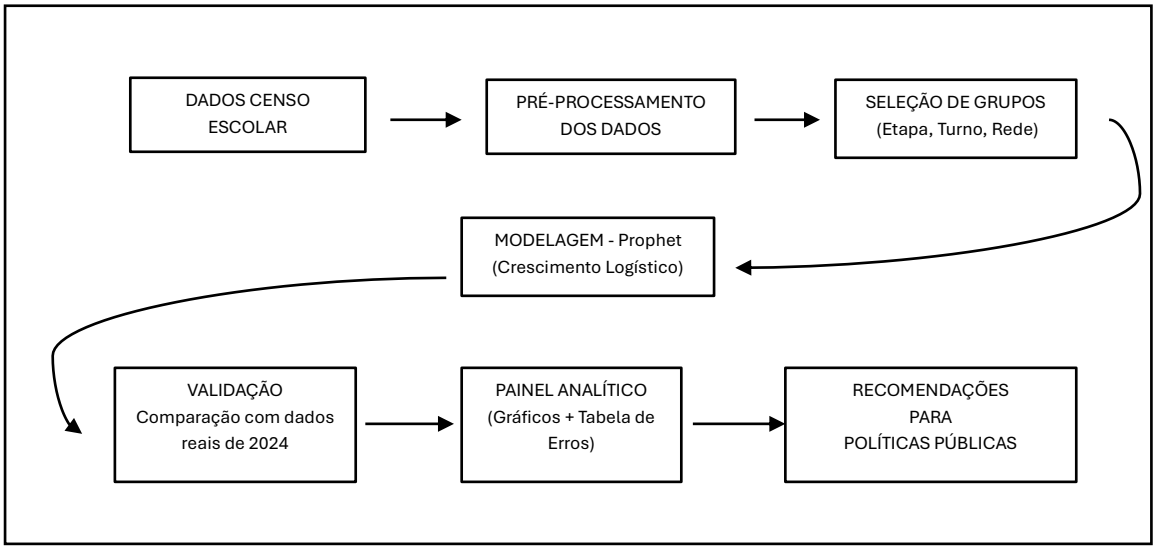
A solução proposta parte da premissa de que a análise de séries temporais pode oferecer suporte valioso à gestão educacional, especialmente quando o objetivo é antecipar padrões de evasão escolar e direcionar políticas públicas de forma preventiva. O fluxo construído, representado no diagrama da solução, estabelece um pipeline completo que vai desde a ingestão de dados públicos (Censo Escolar) até a produção de insights acessíveis a tomadores de decisão.

O processo inicia-se com a organização e padronização dos dados de matrícula, segmentados por etapa de ensino, turno e rede/localização. Essa estrutura permite a análise granular e contextualizada da realidade educacional. A partir da estruturação dos dados, a solução aplica o modelo Prophet, com crescimento logístico e limiares ajustados, adequando-se às particularidades da série (ex: valores baixos, tendência decrescente, ausência de sazonalidade). A validação da previsão com dados reais de 2024 reforça a confiabilidade do modelo e aponta suas limitações, especialmente em cenários sujeitos a intervenções externas não modeladas.

Os resultados são consolidados em um painel analítico de fácil interpretação, composto por gráficos comparativos e tabelas com erros absolutos. A partir disso, são formuladas recomendações estratégicas para os gestores públicos, indicando os segmentos mais críticos e as regiões que demandam atenção prioritária.

A proposta, portanto, extrapola o escopo técnico e posiciona a ciência de dados como instrumento de transformação social, ao articular evidência empírica, visualização clara e inteligência preditiva como suporte à formulação de políticas públicas realistas, intersetoriais e baseadas em inovação. Abaixo fluxo da solução

Figura 1. Fluxo da solução proposta



Fonte: Autores da pesquisa

5.2 Modelo Base

5.3 Pipelines da Solução

A solução proposta para o problema da evasão escolar na Educação Infantil está organizada em um pipeline dividido em sete etapas principais. O Notebook do projeto pode ser acessado no link https://github.com/AnaAleixo/Projeto-Apl-IV/blob/main/Notebook_do_Modelo.ipynb.

Cada fase foi planejada para garantir a qualidade dos dados, a eficácia da análise e a aplicabilidade prática dos resultados. A seguir, detalha-se o pipeline da solução, que foi dividida em duas etapas. A etapa da EDA e da modelagem da Serie Temporais, bem como a respectiva discussão de suas funções e relevância no contexto do projeto.

5.4Análise Exploratória dos Dados (EDA)

Compreensão da fonte de dados

Origem: Censo Escolar (INEP) – 2020 a 2024

Variável observada: Número de alunos matriculados na Educação Infantil

Segmentações:

- Etapa: Creche e Pré-escola
- Turno: Parcial e Integral
- Rede/localização: Estadual/Municipal, Urbana/Rural

Objetivo: Conhecer a estrutura dos dados, suas segmentações e refletir sobre o que pode influenciar a evasão (ex: localização, tipo de turno, etc.).

Leitura e limpeza dos dados

- Importação da planilha em .xlsx com a estrutura padronizada.
- Definição correta dos nomes das colunas (Etapa, Turno, Rede_Localizacao, Ano, Total).
- Conversão das colunas Ano e Total para tipos numéricos (int, float).

Objetivo: Garantir que os dados estejam no formato correto para análises estatísticas e temporais, evitando erros futuros no modelo.

Segmentação por grupo de análise

- Dividimos os dados em 4 subconjuntos:
    1. Educação Infantil – Turno Parcial
    2. Educação Infantil – Turno Integral
    3. Pré-escola – Turno Parcial
    4. Pré-escola – Turno Integral

Objetivo: Analisar padrões individualmente dentro de cada realidade, já que a evasão pode se comportar de forma diferente entre esses grupos.

Análise visual (gráficos de linha por ano)

- Evolução das matrículas ao longo dos anos, separadas por Rede_Localizacao.
- Uso de **escala logarítmica** quando há grande discrepância entre redes.
- Gráficos específicos para grupos pequenos (ex: Estadual Urbana), com visualização isolada.

Objetivo: Identificar tendências temporais, quedas abruptas ou crescimentos incomuns que podem indicar pontos de atenção para políticas públicas.

Estatística descritiva por grupo

- Cálculo de média, desvio padrão, valores mínimo e máximo por grupo.
- Avaliação da dispersão entre redes e turnos.

Objetivo: Ter uma visão numérica da concentração ou variabilidade dos dados, ajudando a entender se o comportamento é homogêneo ou desigual entre as categorias.

Por que seguimos esse caminho?

- Começamos pela estruturação dos dados para garantir consistência na base.
- Optamos por uma EDA segmentada porque **a evasão escolar tem múltiplos determinantes contextuais** (ex: rede municipal rural pode ter comportamento diferente da rede estadual urbana).
- A análise visual serve para **apoiar a modelagem futura**: conseguimos avaliar se as séries têm **tendência, sazonalidade ou rupturas**, o que será essencial para a escolha e calibração dos modelos preditivos (como Prophet ou ARIMA).

5.5 Modelagem de Séries Temporais

Seleção e preparação da série temporal

Para cada combinação de Etapa, Turno e Rede/Localização, é construída uma série temporal com duas colunas obrigatórias para o Prophet:

- ds: representa a data ou o ano no formato yyyy-mm-dd
- y: representa o valor observado (número de matrículas)

Essa estrutura padronizada é requerida pelo Prophet para análise e previsão temporal.

Divisão dos dados para treinamento e validação

Dado que a série possui apenas cinco pontos (2020 a 2024), foi adotada uma estratégia de validação prática e ética para evitar perda excessiva de informação:

- O modelo é inicialmente treinado com os dados de 2020 a 2023.
- Em seguida, realiza-se a validação usando o ano de 2024, com o intuito de avaliar se a tendência projetada se confirma com os dados mais recentes.

Essa abordagem do tipo **"holdout temporal"** é recomendada para séries curtas, pois permite avaliar a capacidade preditiva do modelo sem comprometer sua estabilidade.
Após a validação, o modelo é reentrenado com toda a série histórica (2020–2024) para gerar as previsões para os anos seguintes (ex: 2025 a 2027).

Instanciação e configuração do modelo Prophet

Um modelo Prophet básico é instanciado com tendência linear. A configuração padrão é suficiente para séries curtas com baixa ou nenhuma sazonalidade aparente, como é o caso das séries analisadas. Ajustes adicionais podem ser aplicados conforme o comportamento dos dados.

5.6 Análise da Validação do Modelo

A validação das previsões realizadas pelo modelo Prophet, usamos o MAE – Mean Absolute Error (Erro Absoluto Médio), tomando como referência o ano de 2024, o resultado revelou variações significativas entre os valores previstos e os dados reais de matrícula para os diferentes grupos analisados. De modo geral, o modelo apresentou maior precisão para séries com volume absoluto mais alto e comportamento historicamente mais estável.

No caso da Pré-escola – Turno Parcial – Rede Estadual Rural, que apresenta maior volume de dados e uma curva de tendência mais suave, o erro absoluto foi de 201 matrículas, o menor entre os quatro grupos. Isso indica que o modelo foi capaz de captar razoavelmente bem o padrão da série, mesmo considerando a limitação do número de observações (cinco anos). Já para o grupo Educação Infantil – Turno Integral – Rede Estadual Rural, cujo histórico é mais instável e com valores absolutos mais baixos, o erro absoluto foi de 80 matrículas, o que, proporcionalmente, representa um desvio relevante frente ao volume real observado (104).

As maiores discrepâncias ocorreram nos grupos Educação Infantil – Parcial – Estadual Rural e Pré-escola – Integral – Estadual Rural, com erros absolutos de 548,59 e 307,69 matrículas, respectivamente. Em ambos os casos, o modelo capturou a tendência de queda observada até 2023, mas subestimou a retomada ou manutenção da matrícula registrada em 2024. Isso pode indicar que eventos exógenos (ex: reabertura de turmas, políticas de indução à matrícula, mudanças territoriais) não foram representados diretamente na série histórica, o que limitou a capacidade preditiva do modelo.

Esses resultados reforçam a importância de complementar modelos quantitativos com informações qualitativas ou contextuais, e indicam que o Prophet pode ser uma ferramenta útil para sinalizar tendências gerais e antecipar riscos, mas que sua acurácia é sensível à qualidade, à consistência e à granularidade dos dados históricos disponíveis. A

integração com dados adicionais e a atualização frequente do modelo podem aumentar sua robustez para fins de planejamento educacional

7. Produto e Divulgação Pública

Por fim, todo o código, base de dados processada e documentação serão disponibilizados em repositório público (GitHub), como forma de promover a reprodutibilidade e o caráter extensionista do projeto. A solução será apresentada a partir de um relatório técnico e, se possível, por meio de eventos, seminários ou oficinas com gestores públicos, contribuindo para a implementação de políticas alinhadas ao ODS 4 (Educação de Qualidade).

8. Cronograma do Projeto

Entrega	Data
Definição do projeto e equipe	**28/02**
Referencial Teórico e Cronograma	**28/03**
Implementação Parcial	**25/04**
Implementação e Entrega Final	**30/05**

9. Referências Bibliográficas

CUNHA, Rogério da Silva. Panorama da evasão escolar no Brasil: causas e consequências. Revista Brasileira de Educação, v. 25, n. 81, p. 110-125, 2020.

FERREIRA, Lucas R.; SANTOS, Renata M.; OLIVEIRA, Marcos A. Modelos preditivos aplicados à evasão escolar: uma revisão sistemática. Cadernos de Informática, v. 10, n. 2, p. 45-63, 2022.

HECKMAN, James J. The Economics of Inequality: The Value of Early Childhood Education. American Educator, v. 35, n. 1, p. 31-47, 2011.

HYNDMAN, Rob J.; ATHANASOPOULOS, George. Forecasting: Principles and Practice. 3. ed. Melbourne: OTexts, 2021. Disponível em: https://otexts.com/fpp3/. Acesso em: 26 mar. 2025.

IBGE. Educação 2022: Pesquisa Nacional por Amostra de Domicílios Contínua - PNAD. Rio de Janeiro: IBGE, 2022. Disponível em: https://www.ibge.gov.br/. Acesso em: 26 mar. 2025.

INEP. Censo Escolar da Educação Básica 2023: resumo técnico. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, Brasília: INEP, 2023.

OLIVEIRA, Cláudia M.; MENDES, Patrícia A. Importância da Educação Infantil no desenvolvimento integral da criança. Revista Educação em Foco, v. 22, n. 1, p. 30-41, 2019.

RIBEIRO, Tânia M.; SILVA, Felipe C. Motivos da evasão escolar no Brasil: um olhar sociológico. Revista Educação e Sociedade, v. 42, n. 154, p. 72-84, 2021.

SOUZA, Camila R.; LIMA, Eduardo H. Análises temporais aplicadas à educação pública brasileira: tendências e desafios. Revista Brasileira de Ciências Aplicadas, v. 12, n. 3, p. 95-105, 2020.

Grossi, et al. (2021). Editorial: Greenhouse Gas Emissions and Terrestrial Ecosystems. Frontiers.

Kumar, et al. (2021). Metodologia holística para estimativa de emissões e remoções anuais de GEE em parques naturais. Frontiers in Environmental Science

Grossi, et al. (2021). "Métodos para estimar emissões e remoções de GEE." *Frontiers in Environmental Science*.

Hui, et al. (2021). "Impacto das atividades humanas nas emissões de GEE." *Journal of Environmental Management*.

Kumar, et al. (2021). "Conservação de terra e gestão de carbono." *Science Advances*.