



UNIVERSIDADE FEDERAL DO CEARÁ
FACULDADE DE ECONOMIA, ADMINISTRAÇÃO, ATUÁRIA E
CONTABILIDADE – FEAAC
PROGRAMA DE ECONOMIA PROFISSIONAL – PEP

FLÁVIA FERNANDA FRANÇA DE LIMA

ANÁLISE DOS PERFIS DE RISCO DO TRABALHO INFANTIL NO BRASIL COM
TÉCNICAS DE MACHINE LEARNING

FORTALEZA

2026

FLÁVIA FERNANDA FRANÇA DE LIMA

ANÁLISE DOS PERFIS DE RISCO DO TRABALHO INFANTIL NO BRASIL COM
TÉCNICAS DE MACHINE LEARNING

Dissertação submetida à Coordenação do Programa de Economia Profissional – PEP, da Universidade Federal do Ceará - UFC, como requisito parcial para a obtenção do grau de Mestre em Economia. Área de Concentração: Economia do Setor Público.

Orientador: Prof. Dr. João Mário Santos de França.
Coorientadora: Profa. Dra. Guaracyane Lima Campelo.

FORTALEZA

2026

FLÁVIA FERNANDA FRANÇA DE LIMA

ANÁLISE DOS PERFIS DE RISCO DO TRABALHO INFANTIL NO BRASIL COM
TÉCNICAS DE MACHINE LEARNING

Dissertação submetida à Coordenação do Programa de Economia Profissional – PEP, da Universidade Federal do Ceará - UFC, como requisito parcial para a obtenção do grau de Mestre em Economia. Área de Concentração: Economia do Setor Público.

Aprovada em: 27/03/2026

BANCA EXAMINADORA

Prof. Dr. João Mário Santos de França (Orientador)
Universidade Federal do Ceará (UFC)

Prof.^a Dra. Guaracyane Lima Campelo (Coorientadora)
Universidade Federal do Ceará (UFC)

Dra. Natália Cecília de França
Controladoria e Ouvidoria Geral do Estado (CGE-CE)

AGRADECIMENTOS

À Universidade Federal do Ceará, em especial ao Programa de Economia Profissional (PEP).

Ao Prof. Dr. João Mário Santos de França pela excelente orientação.

À Prof.^a Dra. Guaracyane Lima Campelo pela dedicação e atenção ao longo de todo o desenvolvimento deste trabalho.

À Dra. Natália Cecília de França, minha irmã de alma, por todos os momentos de troca, apoio emocional e intelectual e, sobretudo, por sua amizade inestimável durante os anos de mestrado.

À Victória e a Henrique, minha família e meu abrigo, por todo o amor, apoio e incentivo constantes.

RESUMO

O presente estudo analisa os perfis de risco associados à incidência do trabalho infantil no Brasil, com o intuito de identificar os fatores socioeconômicos, demográficos e domiciliares relacionados ao risco de inserção antecipada no mercado de trabalho. Levando em consideração uma amostra com crianças de 5 a 13 anos de idade, a análise faz uso de microdados oriundos da Pesquisa Nacional por Amostra de Domicílios Contínua (PNADC) referentes ao ano de 2024. A metodologia baseia-se na aplicação do modelo de *machine learning Extreme Gradient Boosting* (XGBoost), algoritmo de aprendizado supervisionado fundamentado em árvores de decisão, aliado a técnicas de interpretabilidade por meio dos valores *SHapley Additive exPlanations* (SHAP), permitindo identificar a contribuição individual das variáveis explicativas sobre as previsões do modelo. Os resultados revelam que idade da criança, sexo, local de residência, renda domiciliar per capita, escolaridade e ocupação do responsável, especialmente quando vinculada ao setor agrícola, são elementos que diferenciam substancialmente os perfis de risco. Crianças mais velhas, meninos, residentes em áreas rurais e pertencentes a domicílios com menor acúmulo de renda e capital humano apresentaram maior propensão estimada. A abordagem adotada demonstra que ferramentas de aprendizado de máquina interpretável permitem identificar padrões complexos, oferecendo suporte adicional à formulação de políticas voltadas à prevenção e enfrentamento do trabalho infantil no país.

Palavras-chave: trabalho infantil; *machine learning*; vulnerabilidade socioeconômica.

ABSTRACT

This study examines the risk profiles associated with the incidence of child labor in Brazil, aiming to identify the socioeconomic, demographic, and household factors linked to the likelihood of early entry into the labor market. Using a sample of children aged 5 to 13 years, the analysis relies on microdata from the 2024 edition of the Continuous National Household Sample Survey (PNADC). The methodological approach applies the Extreme Gradient Boosting (XGBoost) supervised learning algorithm, combined with interpretability techniques based on SHAP (SHapley Additive exPlanations) values, which make it possible to assess the individual contribution of each explanatory variable to the model's predictions. The results indicate that the child's age and sex, household location, per capita income, and the educational level and occupational sector of the household head—particularly when associated with agricultural work—play a decisive role in shaping risk profiles. Older children, boys, those living in rural areas, and those in households with lower income and human capital exhibit higher predicted probabilities of child labor. The findings demonstrate the potential of interpretable machine learning tools to uncover complex patterns, offering valuable support for the design of policies aimed at preventing and reducing child labor in the country.

Keywords: child labor; machine learning; socioeconomic vulnerability.

LISTA DE QUADROS

Quadro 1 - Definição das variáveis	21
Quadro 2 - Matriz de confusão para a previsão do trabalho infantil	25
Quadro 3 - Matriz de confusão do modelo XGBoost	30
Quadro 4 - Matriz de confusão do modelo Logit	30

LISTA DE FIGURAS

Figura 1 - Incidência do trabalho infantil	27
Figura 2 - Importância global das variáveis na previsão do trabalho infantil (SHAP médio)	32
Figura 3 - Valores SHAP para algumas variáveis explicativas	34

LISTA DE TABELAS

Tabela 1 - Estatísticas descritivas das variáveis independentes	28
---	----

LISTA DE ABREVIATURAS E SIGLAS

ABNT	Associação Brasileira de Normas Técnicas
AUC	Area Under the Curve
CRISP-DM	Cross-Industry Standard Process for Data Mining
ECA	Estatuto da Criança e do Adolescente
ECV	Encuesta de Calidad de Vida
EFHS	Egyptian Family Health Survey
ENAHO	Encuesta Nacional de Hogares
FN	Falso negativo
FP	Falso positivo
IBGE	Instituto Brasileiro de Geografia e Estatística
ILO	International Labour Organization
KDD	Knowledge Discovery in Databases
lrdpc	Logaritmo da renda domiciliar per capita
MTI	Módulo de Trabalho Infantil
NBR	Norma Brasileira
NIR	No Information Rate
OIT	Organização Internacional do Trabalho
PDFOS-ENN	Probability Density Function Over-Sampling and Edited Nearest Neighbor
PNAD	Pesquisa Nacional por Amostra de Domicílios
PNADC	Pesquisa Nacional por Amostra de Domicílios Contínua
PPV	Pesquisa de Padrões de Vida
PSPP	Package for Social Science
ROC	Receiver Operating Characteristic
SHAP	SHapley Additive exPlanations
SMOTE	Synthetic Minority Over-sampling Technique
TURKSTAT	Turkish Statistical Institute
UNICEF	United Nations Children's Fund
VN	Verdadeiro negativo
VP	Verdadeiro positivo
XGBoost	Extreme Gradient Boosting

LISTA DE SÍMBOLOS

$\%$	Porcentagem
$=$	Igualdade
$ $	Módulo ou cardinalidade
D	Conjunto de dados
E	Especificidade
ℓ	Função de perda
\mathcal{F}	Espaço funcional das árvores de decisão
$F1$	Medida F1
f_k	k-ésima árvore de decisão
i	Índice da observação
J	Critério de Youden
K	Número de iterações (árvores)
L	Função objetivo
m	Número de características ou variáveis explicativas
n	Número total de observações
Ω	Termo de regularização de complexidade
ϕ	Função de predição final
R\$	Real (moeda brasileira)
S	Sensibilidade
θ	Vetor de parâmetros
x	Vetor de características
y	Valor observado da variável dependente
\hat{y}	Valor previsto da variável dependente
\in	Pertence a
\mathbb{R}	Conjunto dos números reais
\sum	Operador de somatório

SUMÁRIO

1 INTRODUÇÃO	13
2 REVISÃO DE LITERATURA	15
2.1 Conceito e características do trabalho infantil	15
2.2 Estudos empíricos sobre trabalho infantil	17
3 METODOLOGIA	20
3.1 Base de dados	20
3.2 Modelo XGBoost	22
3.3 Avaliação do modelo preditivo	25
4 RESULTADOS	27
4.1 Estatísticas descritivas	27
4.2 Modelagem e resultados	29
4.3 Valores SHAP	31
5 CONSIDERAÇÕES FINAIS	35
REFERÊNCIAS	36
APÊNDICE A - Razões de chance estimadas no modelo Logit	41

1 INTRODUÇÃO

O desenvolvimento de análises econômicas sobre o trabalho infantil ganhou novo impulso a partir da década de 1990, quando esse fenômeno passou a ser reconhecido como um obstáculo ao desenvolvimento econômico, especialmente em razão da maior ênfase atribuída às políticas de redução da pobreza e à acumulação de capital humano como elementos centrais do processo de desenvolvimento (Kassouf, 2007).

No Brasil, a prevalência caiu de forma relevante, mas segue preocupante. Em 2024, estimou-se 1,650 milhão de crianças e adolescentes em situação de trabalho infantil (4,3% do total), após taxas de 5,2% em 2016 e 4,9% em 2022. A intensidade do problema aumenta com a idade: entre 5 e 13 anos a incidência foi de 1,4%, subindo para 6,2% aos 14-15 e atingindo 15,3% entre 16-17 anos. Mais da metade (55,5%) dos casos concentrou-se na faixa de 16-17 anos, cerca de 915 mil, com 19,5% de jornadas semanais de 40 horas ou mais, e quase metade acima de 25 horas no grupo de 16-17 anos, sinalizando risco de sobrecarga e de prejuízo escolar (IBGE, 2025a).

O trabalho precoce é compreendido na literatura como um fenômeno que, simultaneamente, decorre da vulnerabilidade socioeconômica de determinados grupos sociais e contribui para sua perpetuação, constituindo mecanismo de reprodução da pobreza intergeracional e, em escala mais ampla, impondo restrições à produtividade e ao crescimento econômico das economias nacionais (Arruda; Duailibe, 2023).

O marco constitucional brasileiro adota o princípio da proteção integral a crianças e adolescentes, atribuindo a família, à sociedade e ao Estado o dever de garantir direitos e resguardar contra a exploração (Brasil, 1988). Na prática, entretanto, o risco de inserção precoce resulta da combinação de fatores, como condições econômicas do domicílio, escolaridade e ocupação dos responsáveis, composição familiar, gênero e território, que interagem de modo heterogêneo no espaço (Souza, 2025).

Mapear quem são as crianças mais expostas e em que condições esse risco se materializa é condição prática para aperfeiçoar políticas públicas: além de orientar a focalização e o desenho de estratégias de prevenção, essa leitura permite integrar o enfrentamento do trabalho infantil a agendas mais amplas de qualidade de vida e trabalho decente voltadas a grupos em maior vulnerabilidade econômica. Para sustentar intervenções consistentes, a literatura recomenda evidências produzidas em base empírica robusta, com estatísticas confiáveis e desagregadas por gênero, idade, etnia e classe, de modo a capturar desigualdades

e heterogeneidades do fenômeno (Save the Children, 2013; Brasil; OIT, 2014).

No plano internacional, a evidência acumulada indica que restrição de renda é necessária, mas não suficiente para explicar a inserção precoce no trabalho. Estudos realizados em México, Peru, Turquia, Colômbia e Egito mostram que o risco varia de acordo com capital humano dos responsáveis, arranjos familiares, gênero e contexto territorial (urbano/rural), refletindo arranjos institucionais e socioculturais específicos de cada país. Em síntese, os resultados convergem para um arcabouço multifatorial, no qual fatores familiares, educacionais, culturais e regionais modulam a probabilidade de ocorrência de trabalho infantil para além da renda (Cummings, 2016; Quiñones Fabian *et al.*, 2022; Bingöl; Çilingirtürk, 2023; Romero-Luna; Ordoñez-Erazo; Cobos-Lozada, 2024; Zeini; Fikry, 2025).

No plano nacional, a literatura converge quanto ao papel da pobreza, ao mesmo tempo em que destaca o papel do capital humano dos responsáveis, da frequência escolar e dos arranjos familiares na determinação do trabalho infantil (Kassouf, 2002; Cacciamali; Tatei, 2008; Ramalho; Mesquita, 2013; Jardim *et al.*, 2021; Souza, 2025). De forma geral, os estudos sugerem que o trabalho infantil decorre de um conjunto de vulnerabilidades sobrepostas, nas quais fatores econômicos, demográficos e institucionais atuam de maneira conjunta e heterogênea, modulando o risco conforme o território e a estrutura social analisados.

Partindo dos microdados da PNADC 2024 e considerando crianças de 5 a 13 anos, este estudo adota uma abordagem preditiva com XGBoost, modelo de aprendizado supervisionado baseado em árvores, e medidas de interpretabilidade via *SHapley Additive exPlanations* (SHAP) para mapear perfis de risco de trabalho infantil no Brasil. O foco recai em identificar padrões e fatores associados à inserção precoce, explorando relações não lineares e interações entre atributos individuais, familiares e características do domicílio.

A pesquisa avança a literatura nacional ao empregar *machine learning* interpretável para classificação de risco, ainda pouco explorada no tema. Ao combinar alto poder preditivo com transparência sobre a contribuição de cada variável, a estratégia revela dependências não lineares e efeitos heterogêneos entre as covariáveis, oferecendo insumos acionáveis para monitoramento e focalização de políticas.

O texto organiza-se da seguinte forma: Seção 2, com a revisão de conceitos e determinantes do trabalho infantil em âmbito nacional e internacional; Seção 3, com a descrição dos dados, variáveis e procedimentos empíricos, com destaque para o XGBoost e as métricas de avaliação; Seção 4, contendo a apresentação e discussão dos resultados, incluindo estatísticas descritivas, comparação com o Logit e análises via SHAP; e Seção 5, com as conclusões, limitações e implicações de política pública.

2 REVISÃO DE LITERATURA

2.1 Conceito e características do trabalho infantil

A legislação brasileira estabelece faixas etárias e salvaguardas específicas para crianças e adolescentes, proibindo o trabalho antes dos 14 anos, admitindo aprendizagem a partir dessa idade, e regulando o trabalho apenas a partir dos 16 anos em condições compatíveis com o desenvolvimento integral (Brasil, 1990). Em síntese, o marco legal delimita a idade mínima e prioriza a escolarização, expressando um padrão de proteção que orienta políticas e estatísticas oficiais.

O III Plano Nacional de Prevenção e Erradicação do Trabalho Infantil define o fenômeno do trabalho infantil como qualquer atividade econômica ou de sobrevivência realizada abaixo da idade mínima, ainda que sem remuneração, além de vedar tarefas que coloquem em risco o desenvolvimento físico, psicológico, social ou moral dos adolescentes (Brasil, 2018). Trata-se, portanto, de um conceito centrado na proteção integral e na prevenção de danos.

Nas estatísticas oficiais brasileiras, a PNADC adota um escopo abrangente de atividades, distinguindo ocupação (remunerada ou não), produção para autoconsumo, trabalho voluntário, cuidados de pessoas e afazeres domésticos. Essa distinção alinha-se às recomendações internacionais e permite separar o conjunto de atividades realizadas por crianças e adolescentes daquelas que, de fato, configuram trabalho infantil segundo idade, natureza da tarefa e intensidade de jornada (IBGE, 2025b)

Na perspectiva da OIT, trabalho infantil é mais restrito que o universo de “crianças economicamente ativas”: tarefas leves autorizadas, ou atividades não perigosas acima da idade mínima, não se enquadram nessa definição (OIT, 2006; OIT, 2010). Em geral, os ordenamentos nacionais e os padrões internacionais elencam três grupos proibidos: piores formas (incluindo trabalho forçado e exploração sexual), atividades abaixo da idade mínima e trabalhos perigosos, quando a natureza ou as condições impõem risco ao desenvolvimento (OIT, 2010).

A literatura especializada destaca que crianças e adolescentes ocupados tendem a enfrentar jornadas extensas, ambientes insalubres e baixos rendimentos, além de baixa cobertura por políticas sociais, obstaculizando o seu desenvolvimento físico e mental, bem como sua continuidade na escola (Kassouf, 2002). Em linha com esse diagnóstico, estudos de políticas públicas apontam que a organização precoce da rotina de trabalho pode reduzir as oportunidades de brincar, de experimentar e de expressar interesses, empobrecendo

competências cognitivas e socioemocionais e, associada ao cansaço físico, contribuindo para baixo desempenho ou dificuldades de aprendizagem (Brasil, 2005).

Organismos internacionais enfatizam que o enfrentamento efetivo requer uma abordagem sistêmica, que combine reformas legais, conscientização social, proteção social, produção e uso de dados e o engajamento do setor privado em ações de prevenção e erradicação (UNICEF, 2018). Ao mesmo tempo, a complexidade causal, envolvendo pobreza, escolarização insuficiente, práticas e tradições locais, limita os efeitos de instrumentos legais isolados, exigindo políticas integradas e sensíveis às especificidades territoriais (Macedo, 2012).

Apesar das vedações normativas, o trabalho infantil permanece como uma realidade social relevante no Brasil, o que suscita questionamentos centrais acerca de seus determinantes, tais como o papel da pobreza, da estrutura familiar e da escolaridade de crianças e de seus responsáveis na configuração dos riscos associados à inserção precoce no mercado de trabalho (Kassouf, 2002). Em resposta a essa problemática, consolidou-se um amplo corpo de estudos voltados à identificação dos fatores associados à elevada participação de crianças e adolescentes em atividades laborais, sendo recorrente, nessa literatura, a atribuição de papel central à pobreza como condicionante do fenômeno, ainda que não exclusiva (Barros *et al.*, 1994).

Nesse sentido, famílias em situação de vulnerabilidade econômica tendem a apresentar menor capacidade de postergar a entrada das crianças no trabalho e de investir em sua educação, o que eleva a probabilidade de recorrer ao trabalho infantil para suprir necessidades básicas. Evidências que controlam características familiares associadas à escassez de renda indicam que a relação entre pobreza e trabalho infantil permanece robusta (Save the Children, 2013). Ademais, a fragilidade dos sistemas educacionais e a aceitação social do trabalho precoce contribuem para a reprodução intergeracional desse padrão, ao reduzir a demanda por educação e dificultar a aplicação das normas sobre idade mínima e ensino obrigatório (Macedo, 2012; OIT, 2006).

Ainda assim, a pobreza, isoladamente, não é suficiente para explicar o trabalho infantil, especialmente suas piores formas (OIT, 2006), o que torna necessária uma revisão mais abrangente da literatura, que vá além do fator renda e incorpore múltiplas determinantes do trabalho infantil, considerando diferentes modelos econométricos, bases de dados, recortes geográficos e períodos de análise.

2.2 Estudos empíricos sobre trabalho infantil

A evidência empírica converge para um conjunto multidimensional de fatores: atributos econômicos do domicílio, capital humano e participação laboral dos responsáveis, estrutura familiar, sexo e idade da criança, além de localização e contexto institucional. Esses elementos interagem e se sobrepõem, elevando ou reduzindo a probabilidade de ocorrência do fenômeno dependendo do território e do ciclo de vida considerado (Kassouf, 2007; OIT, 2006).

Com base em um modelo microeconômico de oferta de trabalho infantil inspirado em Rosenzweig (1981) e nos microdados da PNAD 1995 e 1999, Kassouf (2002) observa que crianças que não trabalham apresentam frequência escolar mais elevada; que o número de irmãos mais novos se associa a maior probabilidade de inserção laboral; que a idade da criança eleva o risco; e que escolaridade e salário dos pais reduzem a chance de trabalho infantil.

Examinando a PNAD 2006 para o meio rural e estimando um probit por máxima verossimilhança, Kassouf e Justus (2010) mostram que a probabilidade de trabalho infantil cresce com o tamanho da propriedade agrícola até um pico próximo de 76 hectares, a partir do qual declina, resultado que persiste mesmo com controles para idade, gênero, raça, região e características familiares.

Cacciamali e Tatei (2008), utilizando um modelo probit e dados da PNAD de 2002, concluíram que famílias cujo chefe atua como trabalhador por conta própria apresentam maior probabilidade de incidência de trabalho infantil em comparação àquelas cujo chefe é trabalhador assalariado formal. Em consonância com Kassouf (2002), os autores ressaltam que a elevação da escolaridade dos pais exerce impacto mais relevante na redução do trabalho infantil do que incrementos de renda.

A partir de logit com a PNAD 2001, Maciel (2018) evidencia que anos de estudo do responsável estão negativamente associados à inserção precoce no trabalho em todas as regiões, reforçando o papel do capital humano do chefe do domicílio. Em perspectiva dinâmica em modelos com painel 2001-2009 e heterogeneidade não observada, Ramalho e Mesquita (2013) identificam persistência temporal e intergeracional do trabalho infantil, agravada por pobreza e informalidade. A baixa escolaridade dos pais deprime a renda e eleva a chance de inserção precoce, com efeitos duradouros sobre a acumulação de capital humano.

Com PNADC 2016 e probit bivariado, Jardim *et al.* (2021) reportam maior probabilidade de trabalho infantil em domicílios chefiados por mulheres. Em contraste, Utilizando PNAD 2019 e análise via Package for Social Science (PSPP), Itaoui (2023) encontra incidência mais alta em famílias monoparentais masculinas, sobretudo de baixa renda e em

áreas rurais, apontando a complexidade das relações entre arranjos familiares e trabalho infantil.

Souza (2025), utilizando um modelo de regressão logística (logit) e dados da PNAD de 2023, investigou a relação entre a estrutura familiar e trabalho infantil, apontando que este decorre não apenas da pobreza, mas também dos arranjos familiares, analisados isoladamente ou em interação com fatores sociais e demográficos. Em linhas gerais, o estudo indica menores taxas quando há chefia feminina, maior escolaridade do responsável e arranjos compostos ou estendidos, ao passo que residência na região Norte e sexo masculino da criança elevam o risco de inserção precoce.

Com recorte regional, PPV 1996-1997 e logit multinomial, Menezes (2007) identifica, para Nordeste e Sudeste, maior probabilidade de trabalho infantil entre crianças mais velhas, do sexo masculino, de famílias numerosas e de baixa renda, especialmente em áreas rurais do Nordeste. Para a região Sul, De Souza e Rocha (2023), usando microdados do Censo Demográfico 2010 do IBGE e logit multinomial, mostram a relevância de idade, raça, gênero, ocupação do chefe, localização do domicílio e participação em programas públicos; encontram maior probabilidade para meninos, mais velhos e de raça não branca, além de maior incidência em famílias numerosas e com chefe ocupado, ao passo que transferências de renda contribuem para reduzir evasão escolar.

No recorte de Minas Gerais com Censo 2010, Rocha (2018) registra maior propensão à inserção laboral entre meninos, crianças pardas e negras, em lares chefiados por mulheres, domicílios agrícolas, sem ocupação e em zona rural, com impactos negativos sobre a frequência escolar.

Aplicando o processo Descoberta de Conhecimento em Bancos de Dados e algoritmo J48 aos dados do CadÚnico, Rodrigues, Prata e Silva (2015) mostram que renda isolada não é conclusiva, enquanto frequência escolar se relaciona diretamente à ocorrência do fenômeno; o recebimento de transferências, como Bolsa Família, associa-se à redução do trabalho infantil entre beneficiários.

Uma breve análise da literatura estrangeira, voltada a países como México, Peru, Turquia, Colômbia e Egito permite ampliar a compreensão sobre o perfil e os determinantes do trabalho infantil em diferentes contextos socioeconômicos. Esses estudos evidenciam que, embora a pobreza seja um fator relevante, ela não atua de forma isolada na explicação desse fenômeno. Nesse sentido, destacam-se os trabalhos a seguir.

Cummings (2016) apresentou uma estimativa dos determinantes do trabalho infantil para crianças entre 5 e 17 anos no México. Utilizando modelos de regressão logística e os dados do Módulo de Trabalho Infantil (MTI) de 2009, 2011 e 2013, constatou que o ingresso precoce

no mercado de trabalho no México não é explicado apenas pela pobreza, mas também pela percepção do custo de estudar, nível de escolaridade dos pais, estrutura familiar, gênero e pelo contexto urbano/rural.

Quiñones Fabian *et al.* (2022) desenvolveram um algoritmo explicativo para identificar os fatores associados ao abandono escolar e ao trabalho precoce de crianças e adolescentes na região central do Peru, utilizando dados da Pesquisa Nacional de Famílias (ENAHO) de 2017. A metodologia combinou o uso do Lasso para seleção de variáveis e da Regressão Logística para classificação do risco. Os resultados indicaram que o baixo nível educacional do chefe do domicílio, o status social, o uso de línguas nativas e o sexo masculino elevam a probabilidade de trabalho infantil, além de apontarem maior incidência entre crianças residentes em áreas urbanas em comparação às rurais.

Bingöl e Çilingirtürk (2023) compararam o desempenho dos algoritmos de *machine learning* Regressão Logística, Naïve Bayes e Random Forest na previsão do trabalho infantil na Turquia a partir de técnicas de aprendizado supervisionado aplicadas aos microdados da Pesquisa de Força de Trabalho Infantil de 2019 do Instituto de Estatística da Turquia (TURKSTAT). Os resultados indicaram desempenho superior do algoritmo Random Forest, com acurácia de 97%, destacando o gênero, o nível educacional e a ocupação dos pais como os principais fatores associados ao trabalho infantil no contexto turco.

Romero-Luna, Ordoñez-Erazo e Cobos-Lozada (2024) propuseram a identificação de características associadas à ocupação de menores na Colômbia. A metodologia baseou-se no padrão Cross-Industry Standard Process for Data Mining (CRISP-DM), com a aplicação dos algoritmos DecisionTreeClassifier, Random Forest e XGBoost, utilizando validação cruzada estratificada e dados da Pesquisa de Qualidade de Vida (ECV) de 2022 do Departamento Administrativo Nacional de Estatística da Colômbia. Os resultados indicaram o DecisionTreeClassifier como o modelo mais adequado, alcançando acurácia equilibrada de 80,02%, e evidenciaram que a ocupação dos menores é fortemente influenciada por fatores econômicos, pela composição do domicílio e pela ausência parental.

Zeini e Fikry (2025) investigaram os fatores associados ao trabalho infantil no Egito empregando um classificador de Regressão Logística combinado com a técnica de amostragem PDFOS-ENN, voltada ao tratamento da sobreposição de dados da Pesquisa de Saúde da Família do Egito (EFHS) de 2021. Os resultados indicaram que a idade da criança (especialmente entre 5 e 11 anos), a região geográfica, a pobreza familiar, o status de emprego da mãe e a posse de terras pela família constituem os principais fatores de risco para a inserção precoce no trabalho.

3 METODOLOGIA

3.1 Base de dados

A base empírica decorre dos microdados da PNADC 2024, combinando informações da 1ª e da 5ª visita realizadas pelo IBGE ao longo do ano. O plano amostral permite expandir as respostas para representar a população, e a pesquisa foi concebida para acompanhar, de forma continuada, flutuações trimestrais e a evolução no curto, médio e longo prazos da força de trabalho e de variáveis associadas ao desenvolvimento socioeconômico do país (IBGE, 2026).

A cobertura de crianças e adolescentes nas pesquisas domiciliares do IBGE remonta à década de 1970, quando a PNAD passou a investigar a população a partir de 10 anos. Nos anos 1990, foi incorporado um suplemento específico para 5 a 9 anos, ampliando a visibilidade estatística do tema. Com a PNADC, a partir de 2013, o assunto integra o conjunto de temas complementares, assegurando a continuidade do monitoramento na agenda nacional de enfrentamento ao trabalho infantil (Brasil; OIT, 2014).

Desde 2016, a PNADC inclui um módulo dedicado às atividades de 5 a 17 anos, cobrindo produção para autoconsumo, tarefas domésticas e de cuidado e atividades econômicas. O inquérito produz indicadores mensais e trimestrais da força de trabalho e indicadores anuais suplementares. Para mensuração estatística, o estudo segue a definição operacional alinhada a OIT e UNICEF (2024) adotada pelo IBGE, segundo a qual trabalho infantil envolve atividades que subtraem a infância e a dignidade, com potencial de dano ao desenvolvimento físico e mental e que privam oportunidades adequadas às crianças.

O tratamento dos dados foi realizado no *software* Stata 17, com importação via Data Zoom (2023). Excluíram-se casos cuja condição no domicílio fosse empregado(a) doméstico(a), parente de empregado(a) doméstico(a) e pensionista. A amostra analítica considera crianças de 5 a 13 anos, totalizando 57.221 observações.

A partir dos dados da PNADC, construiu-se a variável de resposta que representa a situação de trabalho infantil da criança e as variáveis explanatórias relacionadas aos atributos do chefe do domicílio, idade das crianças, insegurança alimentar, sexo, características demográficas e de saneamento, dentre outras. Tais variáveis foram selecionadas com base em sua relevância na literatura teórica e empírica (Zeini; Fikry, 2025; Webbink; Smits; De Jong, 2013; Cardoso; Souza, 2004).

Para a construção da variável dependente (y_i), utilizou-se uma *dummy* construída da seguinte forma:

$$y_i = \begin{cases} 1, & \text{se a criança está em situação de trabalho infantil} \\ 0, & \text{caso contrário} \end{cases} \quad (1)$$

O Quadro 1 mostra a definição das variáveis explicativas utilizadas nos modelos implementados.

Quadro 1 – Definição das variáveis

Variável	Definição
menina	<i>Dummy</i> que assume valor 1 se a criança é do sexo feminino e 0, caso contrário
cor_branca	<i>Dummy</i> que assume valor 1 se a criança é da cor branca e 0, caso contrário
idade	Idade da criança, medida em anos
escola	<i>Dummy</i> que assume valor 1 se a criança frequenta a escola e 0, caso contrário
inseg_alim	<i>Dummy</i> que assume valor 1 se a criança vive em situação de insegurança alimentar e 0, caso contrário
aluguel	<i>Dummy</i> que assume valor 1 se a criança reside em domicílio alugado e 0, caso contrário
falta_energia	<i>Dummy</i> que assume valor 1 se a criança reside em domicílio sem acesso à energia elétrica e 0, caso contrário
água_canaliz	<i>Dummy</i> que assume valor 1 se a criança reside em domicílio com acesso à água encanada e 0, caso contrário
esgoto_inad	<i>Dummy</i> que assume valor 1 se o esgoto do banheiro do domicílio não vai para a rede geral/pluvial e 0, caso contrário
lixo_inad	<i>Dummy</i> que assume valor 1 se o lixo do domicílio não é coletado por serviço de limpeza e 0, caso contrário
norte	<i>Dummy</i> que assume valor 1 se a criança reside na região Norte e 0, caso contrário
nordeste	<i>Dummy</i> que assume valor 1 se a criança reside na região Nordeste e 0, caso contrário
sudeste	<i>Dummy</i> que assume valor 1 se a criança reside na região Sudeste e 0, caso contrário
centro_oeste	<i>Dummy</i> que assume valor 1 se a criança reside na região Centro Oeste e 0, caso contrário
urbana	<i>Dummy</i> que assume valor 1 se a criança reside na região na zona urbana e 0, caso contrário
lrdpc	Logaritmo da renda domiciliar per capita (rdpc) construída como $\text{lrdpc} = \log(\text{rdpc}+1)$
dom_benef	<i>Dummy</i> que assume valor 1 se alguém no domicílio recebe benefício social e 0, caso contrário
dom_aposent	<i>Dummy</i> que assume valor 1 se alguém no domicílio recebe aposentadoria e 0, caso contrário
fam_estend	<i>Dummy</i> que assume valor 1 se a espécie da unidade doméstica é estendida e 0, caso contrário
fam_comp	<i>Dummy</i> que assume valor 1 se a espécie da unidade doméstica é composta e 0, caso contrário
dom_criança	Número de crianças no domicílio

Variável	Definição
dom_idoso	Número de idosos no domicílio
dom_nemnem	Número de jovens nem-nem no domicílio
chefe_fem	<i>Dummy</i> que assume valor 1 se o chefe do domicílio é do sexo feminino e 0, caso contrário
chefe_branc	<i>Dummy</i> que assume valor 1 se o chefe do domicílio é branco e 0, caso contrário
chefe_idad	Idade do chefe do domicílio medida em anos
chefe_ocupad	<i>Dummy</i> que assume valor 1 se o chefe do domicílio está ocupado no mercado de trabalho e 0, caso contrário
chefe_previd	<i>Dummy</i> que assume valor 1 se o chefe do domicílio contribui para previdência e 0, caso contrário
chefe_estudo	Anos de estudo do chefe do domicílio
chefe_agric	<i>Dummy</i> que assume valor 1 se o chefe do domicílio trabalha no setor agrícola e 0, caso contrário

Fonte: elaboração própria a partir dos dados da PNADC 2024.

3.2 Modelo XGBoost

Para estimar a probabilidade de ocorrência de trabalho infantil, emprega-se o ***Extreme Gradient Boosting (XGBoost)***, um algoritmo de **aprendizado supervisionado** proposto por Chen e Guestrin (2016) que constrói **árvores de decisão em sequência**. A cada iteração, uma nova árvore é ajustada **para corrigir os resíduos** do conjunto anterior, compondo um **ensemble do tipo boosting** que acumula pequenas melhorias sucessivas até formar o preditor final.

A adoção do XGBoost neste estudo decorre de sua capacidade de capturar relações complexas e não lineares entre variáveis, característica relevante em fenômenos sociais multidimensionais, como o trabalho infantil. Além disso, o algoritmo incorpora mecanismos de regularização que contribuem para maior estabilidade das estimativas e mitigam problemas de sobreajuste, favorecendo a generalização dos resultados.

O XGBoost foi desenvolvido para oferecer alta eficiência computacional e desempenho robusto mesmo em cenários desafiadores, o que o torna especialmente adequado para aplicações preditivas em políticas públicas, onde a identificação precisa de grupos vulneráveis é fundamental (Chen; Guestrin, 2016).

Embora outros modelos amplamente utilizados em *machine learning*, como Random Forest e Support Vector Machines, também representem alternativas metodológicas consistentes, eles apresentam limitações frente às especificidades desta base de dados, especialmente no que se refere à capacidade de capturar interações complexas de maneira tão eficiente quanto os métodos de *boosting*. Diante disso, o XGBoost oferece desempenho mais

alinhado às exigências preditivas do estudo.

Adicionalmente, destaca-se a compatibilidade do modelo com abordagens contemporâneas de interpretabilidade, em especial os valores SHAP. A metodologia SHAP fornece uma interpretação unificada e consistente para modelos complexos de *machine learning*, permitindo que os resultados sejam comunicados e utilizados de forma mais segura e fundamentada em processos decisórios (Lundberg; Lee, 2017).

Considere um conjunto de dados com n observações e m características, $D = \{(x_i, y_i)\}$, com $(|D| = n, x_i \in \mathbb{R}^m, y_i \in \mathbb{R})$, a predição final é expressa após K iterações, e representada como a soma das contribuições de cada árvore $f_k(x_i)$:

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^K f_k(x_i), f_k \in \mathcal{F} \quad (2)$$

Onde x_i representa o vetor de características e y_i a variável resposta. Em que $\mathcal{F} = \{f(x) = w_{q(x)}\}$ ($q: \mathbb{R}^p \rightarrow T, w \in \mathbb{R}^T$) representa o espaço das árvores de regressões. A função q especifica a estrutura de particionamento de cada árvore, isto é, para qual folha cada observação é encaminhada; o parâmetro T representa o número total de folhas. Em cada iteração, f_k é caracterizada por uma estrutura q própria e por um vetor de pesos w .

A estimação do modelo baseia-se na função objetivo composta por dois principais componentes:

- i) Função de perda denotada por $l(y_i, \hat{y}_i)$, que mensura a discrepância entre as previsões \hat{y}_i e os valores observados y_i ;
- ii) Termo de regularização representado por $\Omega(f_k)$ que penaliza a complexidade das árvores de decisão, controlando parâmetros como profundidade e o número de nós terminais visando mitigar o risco sobreajuste (*overfitting*).

Essa função objetivo é expressa por:

$$L(\theta) = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k) \quad (3)$$

Onde $\Omega(f_t) = \gamma T + \frac{1}{2} \lambda \|w\|^2$ em que l é uma função de perda convexa e diferenciável que mede a diferença entre a predição \hat{y}_i e o valor observado y_i , γ e λ são parâmetros de regularização. Dado o funcionamento sequencial do algoritmo, essa otimização é resolvida recursivamente. Considerando o caráter sequencial do algoritmo, essa função é minimizada de forma aditiva e iterativa. A cada iteração, uma nova árvore é ajustada para aproximar o gradiente negativo da função de perda em relação as previsões acumuladas, visando reduzir progressivamente o valor da função objetivo. Tal procedimento permite a correção de erros residuais das iterações anteriores.

A performance do XGBoost é condicionada a um conjunto de hiperparâmetros, que regulam a forma como as árvores de decisões serão construídas e agregadas ao longo do processo de estimação. Em face ao caráter desbalanceado da variável dependente (caracterizado pela baixa prevalência de observações positivas associadas ao trabalho infantil), aplicou-se o hiperparâmetro *scale_pos_weight*¹. Tal parâmetro é responsável pelo reescalonamento ponderado das observações pertencentes à classe positiva da função perda, incrementando a sensibilidade do modelo à detecção de casos de trabalho infantil, preservando a distribuição original dos dados. Adicionalmente, optou-se por não aplicar as técnicas de *oversampling* como SMOTE (*Synthetic Minority Over-sampling Technique*), que produz observações sintéticas da classe minoritária com a finalidade de evitar a expansão artificial do tamanho da amostra e preservar a estrutura de dependência original dos dados e evitar possíveis distorções no processo de aprendizado do algoritmo. Além disso, o objetivo central do estudo não visa obter inferências de parâmetros populacionais cujos pesos de expansão são imprescindíveis. A implementação computacional foi realizada no ambiente estatístico R versão 4.5.2 para Windows.

A compreensão das predições de um modelo é tão essencial quanto a acurácia preditiva do algoritmo. Todavia, diversos modelos de aprendizado de máquina agem como “caixas-pretas”, e apresentando uma estrutura complexa que dificulta a interpretabilidade direta dos seus resultados. Com o intuito de contornar essa limitação, emprega-se o arcabouço de Lundberg e Lee (2017) denominado SHAP que possibilita a decomposição de cada previsão nas contribuições marginais das variáveis explicativas. Os valores SHAP são computados individualmente e quantificam a magnitude e a direção de cada contribuição marginal de cada variável explicativa sobre a predição específica. A análise da distribuição desses valores ao longo do conjunto de dados possibilita identificar a heterogeneidade nos efeitos das covariáveis e revelar padrões não lineares nas relações preditivas

Para sintetizar a importância global das variáveis, utiliza-se o SHAP médio, sendo este construído mediante a agregação dos valores SHAP computados para todas as observações da amostra. Essa métrica captura a intensidade média da contribuição preditiva da variável, independentemente do sinal (positivo ou negativo) do efeito, viabilizando a comparação da importância relativa entre as variáveis do modelo.

Ressalta-se que as variáveis com maior SHAP médio não correspondem necessariamente aquelas que exercem os maiores impactos individuais ao longo de toda a

¹ $scale_pos_weight = \frac{total\ de\ crianças\ fora\ do\ trabalho\ infantil}{total\ de\ crianças\ em\ trabalho\ infantil}$

distribuição amostral. Determinadas variáveis podem apresentar efeitos substantivos apenas em subconjuntos específicos da amostra, enquanto outras exibem contribuições mais modestas, porém consistentes e generalizadas ao longo de todo o conjunto de dados.

3.3 Avaliação do modelo preditivo

Após a etapa do treinamento, executou-se a avaliação do performance da modelagem preditiva, etapa essencial para averiguar sua capacidade de generalização a dados não observados e conseqüentemente sua adequação ao objetivo analítico do estudo. A literatura de aprendizado de máquina recomenda que o desempenho seja consistente tanto no conjunto de treino quanto no de teste para mitigar sobreajuste e sustentar a validade externa dos resultados (James *et al.*, 2013).

A matriz de confusão, estrutura bidimensional 2x2, configura-se como um instrumento analítico que confronta as predições do modelo com os valores efetivamente observados da variável resposta. A diagonal principal representa as classificações corretas (VN e VP) enquanto na diagonal secundária indica os erros de classificação (FP e FN), conforme apresentado no Quadro 2.

Quadro 2 – Matriz de confusão para a previsão do trabalho infantil

		Previsto	
		Não trabalho infantil ($\hat{y} = 0$)	Trabalho infantil ($\hat{y} = 1$)
Observado	Não trabalho infantil ($y = 0$)	Verdadeiro negativo (VN)	Falso positivo (FP)
	Trabalho infantil ($y = 1$)	Falso negativo (FN)	Verdadeiro positivo (VP)

Fonte: elaboração própria.

A partir dessa matriz, foram computadas diversas métricas de desempenho preditivo do modelo:

- *Acurácia*: representa o desempenho global, mensurando a proporção total de acertos (positivos e negativos) sobre o total de casos.

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN}$$

- *Precisão*: quantifica, do total de previsões positivas, a proporção de previsões corretas realizadas.

$$Precisão = \frac{VP}{VP + FP}$$

- *Sensibilidade (Recall)*: expressa a capacidade do modelo de identificar corretamente os casos positivos dentre todas as observações verdadeiramente positivas.

$$Sensibilidade = \frac{VP}{VP + FN}$$

- *F1-Score*: Integra a precisão e a sensibilidade em uma média harmônica. Ela pondera simultaneamente a correta classificação da criança em situação de trabalho infantil (precisão) e sua capacidade de cobertura da totalidade de casos reais na amostra desse fenômeno (sensibilidade). Esse indicador tem relevância pois garante a efetividade do modelo de detectar as crianças em vulnerabilidade.

$$F1 = \frac{2(precisão)(sensibilidade)}{(precisão + sensibilidade)}$$

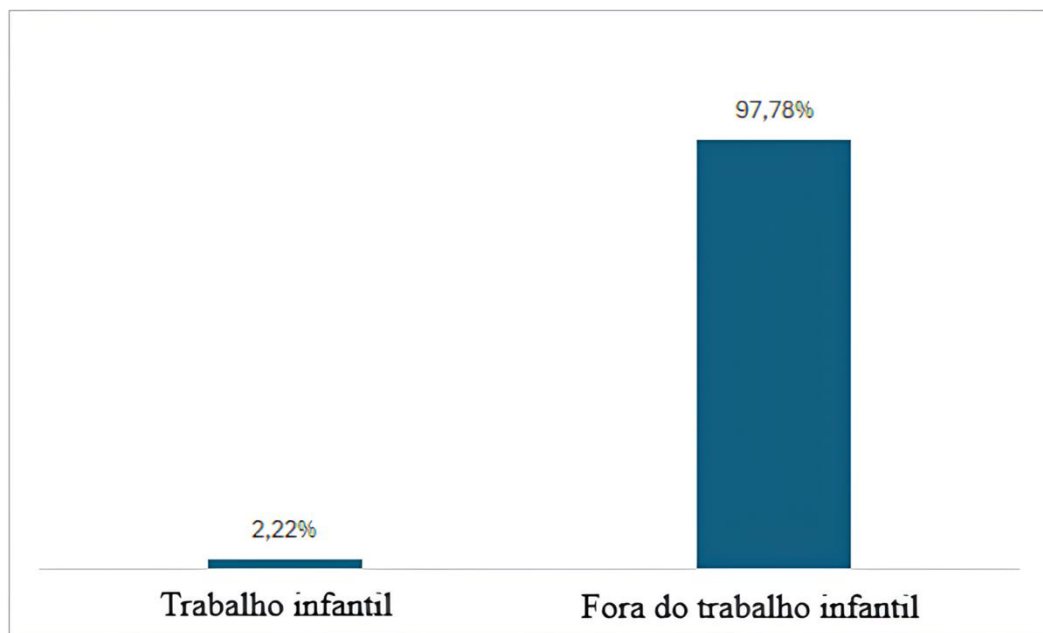
O desempenho preditivo do modelo selecionado foi avaliado no conjunto de teste a partir das probabilidades estimadas. A curva ROC foi construída e definição do ponto de corte ótimo baseou-se no critério de Youden (J), que consiste na maximização da soma entre a sensibilidade (S) e a especificidade (E), menos uma unidade ($J = S + E - 1$), identificando o ponto de corte que oferece o melhor equilíbrio na correta detecção dos casos negativos e positivos. Esse procedimento mostra-se apropriado em contextos caracterizados por desbalanceamento entre classes, como observado neste estudo. A partir desse critério, as probabilidades foram convertidas em classificações binárias e utilizadas para a construção da matriz de confusão conforme descrita anteriormente no Quadro 3 (Fluss; Faraggi; Reiser, 2005).

4 RESULTADOS

4.1 Estatísticas descritivas

A Figura 1 mostra a incidência do trabalho infantil na base de dados, indicando o elevado desbalanceamento entre as duas classes. Apenas 2,22% das crianças consideradas estão em situação de trabalho infantil, enquanto 97,78% referem-se a crianças fora dessa condição.

Figura 1 – Incidência do trabalho infantil



Fonte: elaboração própria com base nos dados da PNADC 2024.

A Tabela 1 relata a análise descritiva entre os grupos, evidenciando as discrepâncias entre crianças afetadas ou não pelo trabalho precoce. O trabalho infantil é mais frequente entre crianças do sexo masculino, não brancas e mais velhas, sugerindo que esses fatores podem estar associados a maior exposição ao risco de inserção precoce no trabalho. A frequência escolar apresenta-se elevada em ambos os grupos, sinalizando ampla cobertura do sistema educacional, todavia não foi suficiente para impedir o trabalho infantil.

Com relação às características do domicílio, crianças em situação de trabalho infantil residem com maior frequência em domicílios com insegurança alimentar, piores condições de saneamento (esgoto e coleta de lixo inadequados), localização rural e uma maior concentração nas regiões Norte e Nordeste. No aspecto econômico, esses domicílios apresentam menor renda domiciliar *per capita* e maior dependência de benefícios sociais, configurando um contexto de vulnerabilidade socioeconômica. Tais evidências corroboram

com os achados encontrados por Kassouf (2002, 2007), Cacciamali e Tatei (2008), Cummings (2016) e Zeini e Fikry (2025).

As características da pessoa responsável pelo domicílio também são diferentes entre os dois grupos de crianças. Nas casas com trabalho infantil, o chefe, em geral, tem menos escolaridade, menor acesso à previdência e maior inserção na agricultura, podendo indicar ocupações mais precárias de baixa produtividade e informalidade. Tais resultados estão alinhados com Kassouf e Justus (2010) e Bingöl e Çilingirtürk (2023).

Tabela 1 – Estatísticas descritivas das variáveis independentes

Variável	Trabalho infantil		Sem trabalho infantil		Total	
	Média	Desvio padrão	Média	Desvio padrão	Média	Desvio padrão
menina	0,3640	0,4814	0,4908	0,4999	0,4880	0,4999
cor_branca	0,2529	0,4349	0,3634	0,4810	0,3610	0,4803
idade	10,4531	2,2829	9,0341	2,5840	9,0656	2,5862
escola	0,9976	0,0486	0,9888	0,1050	0,9890	0,1041
inseg_alim	0,3136	0,4642	0,2899	0,4537	0,2905	0,4540
aluguel	0,1615	0,3682	0,1781	0,3826	0,1777	0,3823
falta_energia	0,0023	0,0486	0,0034	0,0579	0,0033	0,0577
água_canaliz	0,9432	0,2314	0,9631	0,1886	0,9626	0,1897
esgoto_inad	0,5933	0,4914	0,5533	0,4972	0,5542	0,4971
lixo_inad	0,2190	0,4138	0,1768	0,3815	0,1777	0,3823
norte	0,2332	0,4231	0,1724	0,3778	0,1738	0,3789
nordeste	0,4594	0,4985	0,3604	0,4801	0,3626	0,4807
sudeste	0,1300	0,3365	0,2333	0,4229	0,2310	0,4215
centro_oeste	0,0748	0,2633	0,0838	0,2770	0,0836	0,2767
urbana	0,2033	0,4026	0,7211	0,4485	0,7096	0,4540
lrdpc	6,3629	0,8577	6,5576	1,1482	6,5533	1,1429
dom_benef	0,6296	0,4831	0,4646	0,4988	0,4683	0,4990
dom_aposent	0,1725	0,3780	0,1243	0,3299	0,1253	0,3311
fam_estend	0,1820	0,3860	0,2013	0,4010	0,2009	0,4007
fam_comp	0,0118	0,1081	0,0075	0,0864	0,0076	0,0870
dom_criança	1,8329	1,0666	1,7233	0,9368	1,7257	0,9400
dom_idoso	0,1678	0,4522	0,1544	0,4392	0,1547	0,4394
dom_nemnem	0,1607	0,4214	0,1631	0,4174	0,1630	0,4175
chefe_fem	0,4822	0,4999	0,6074	0,4883	0,6046	0,4889
chefe_branc	0,2537	0,4353	0,3096	0,4623	0,3083	0,4618
chefe_idad	43,2490	11,6218	41,4077	11,5622	41,4485	11,5666
chefe_ocupad	0,6611	0,4735	0,6418	0,4795	0,6422	0,4793
chefe_previd	0,2403	0,4275	0,3761	0,4844	0,3731	0,4836

Variável	Trabalho infantil		Sem trabalho infantil		Total	
	Média	Desvio padrão	Média	Desvio padrão	Média	Desvio padrão
chefe_estudo	7,2545	4,4018	9,8913	4,3922	9,8328	4,4095
chefe_agric	0,3672	0,4822	0,0928	0,2901	0,0989	0,2985

Fonte: elaboração própria com base nos dados da PNADC 2024.

4.2 Modelagem e resultados

A base de dados foi particionada aleatoriamente em conjunto de treinamento (80%) e conjunto de teste (20%). Essa separação foi feita para garantir que a avaliação do desempenho dos modelos fosse realizada em dados não utilizados na etapa de estimação.

A calibração do modelo XGBoost foi realizada por meio de ajuste de hiperparâmetros, em que para cada combinação de parâmetros, foi executada validação cruzada com cinco separações, e os modelos foram comparados com base no valor médio da área sob a curva ROC (AUC). A configuração que apresentou o melhor desempenho médio foi selecionada para o treinamento do modelo final. Os hiperparâmetros do modelo final foram:

- Subamostra de observações utilizada em cada árvore: 0,8;
- Número máximo de árvores no processo de *boosting*: 2.000;
- Profundidade máxima das árvores: 6;
- Taxa de aprendizado (*learning rate*): 0,05;
- Proporção de variáveis selecionadas em cada árvore: 0,8;
- Peso mínimo para divisão de nós internos: 1;
- Parâmetro de regularização gama: 0.

O desempenho preditivo do modelo selecionado foi avaliado no conjunto de teste a partir das probabilidades estimadas. A curva ROC foi construída e definição do ponto de corte ótimo baseou-se no critério de Youden (J), que consiste na maximização da soma entre a sensibilidade (S) e a especificidade (E), menos uma unidade ($J = S + E - 1$), identificando o ponto de corte que oferece o melhor equilíbrio na correta detecção dos casos negativos e positivos. Esse procedimento mostra-se apropriado em contextos caracterizados por desbalanceamento entre classes, como observado neste estudo. A partir desse critério, as probabilidades foram convertidas em classificações binárias e utilizadas para a construção da matriz de confusão conforme descrita anteriormente no Quadro 3. (Fluss; Faraggi; Reiser, 2005).

Quadro 3 – Matriz de confusão do modelo XGBoost

		Previsto	
		Não trabalho infantil ($\hat{y} = 0$)	Trabalho infantil ($\hat{y} = 1$)
Observado	Não trabalho infantil ($y = 0$)	9.445	1.777
	Trabalho infantil ($y = 1$)	52	205

Acurácia: 0,8407	Valor preditivo positivo: 0,1034
Intervalo de confiança de 95%: (0,8338; 0,8473)	Valor preditivo negativo: 0,9945
Taxa sem informação (No Information Rate): 0,9776	Prevalência: 0,0224
Valor-p [Acurácia > NIR]: 1	Taxa de detecção: 0,0179
Índice Kappa: 0,1494	Prevalência de detecção: 0,1727
Valor-p do teste de McNemar: < 2e-16	Acurácia balanceada: 0,8197
Sensibilidade: 0,7977	
Especificidade: 0,8417	

Fonte: elaboração própria com base nos dados da PNADC 2024.

Para fins de comparação, estimou-se um modelo logístico de regressão com as mesmas variáveis e a mesma amostra de treinamento. As probabilidades previstas pelo Logit foram avaliadas no conjunto de teste seguindo o mesmo procedimento, com a matriz de confusão apresentada no Quadro 4.

Quadro 4 – Matriz de confusão do modelo Logit

		Previsto	
		Não trabalho infantil ($\hat{y} = 0$)	Trabalho infantil ($\hat{y} = 1$)
Observado	Não trabalho infantil ($y = 0$)	7.400	3.822
	Trabalho infantil ($y = 1$)	33	224

Acurácia: 0,6642	Valor preditivo positivo: 0,0554
Intervalo de confiança de 95%: (0,6554; 0,6728)	Valor preditivo negativo: 0,9956
Taxa sem informação (No Information Rate): 0,9776	Prevalência: 0,0224
Valor-p [Acurácia > NIR]: 1	Taxa de detecção: 0,0195
Índice Kappa: 0,0647	Prevalência de detecção: 0,3525
Valor-p do teste de McNemar: < 2e-16	Acurácia balanceada: 0,7655
Sensibilidade: 0,8716	

Especificidade: 0,6594

Fonte: elaboração própria com base nos dados da PNADC 2024.

A análise comparativa das matrizes de confusão ilustradas no Quadro 3 e no Quadro 4 revelam um desempenho preditivo superior do XGBoost em comparação com o Logit, principalmente no equilíbrio entre identificação adequada das crianças em trabalho infantil e daquelas que não trabalham. O XGBoost apresenta acurácia de 84,07%, maior que a do Logit (66,42%), além de acurácia balanceada mais elevada (0,82 contra 0,77).

Quanto à sensibilidade, o Logit identifica uma proporção ligeiramente maior dos casos de trabalho infantil (87,16%) do que o XGBoost (79,77%). No entanto, essa vantagem ocorre à custa de uma especificidade muito menor (65,94% no Logit contra 84,17% no XGBoost), resultando em um número muito maior de falsos positivos no modelo Logit. O Logit classifica erroneamente 3.822 crianças em situação de trabalho infantil, enquanto o XGBoost reduz esse erro para 1.777 observações (uma redução de 53,5%).

Em síntese, os resultados indicam que o X-GBoost apresenta capacidade discriminatória superior e melhor equilíbrio entre os erros tipo I (falsos positivos) e tipo II (falsos negativos), característica essencial em aplicações de política pública direcionadas ao enfrentamento do trabalho infantil. Nesse cenário, tanto a subidentificação de crianças vulneráveis (que evita o fornecimento de proteção e assistência apropriadas) quanto a superidentificação (sugere a alocação ineficiente de recursos escassos) podem causar custos sociais e econômicos reforçando a relevância de modelos preditivos com alta precisão e sensibilidade balanceada. Tais indícios estão em consonância com estudos recentes de Romero-Luna, Ordoñez-Erazo e Cobos-Lozada (2024) e Bingöl e Çilingirtürk (2023).

4.3 Valores SHAP

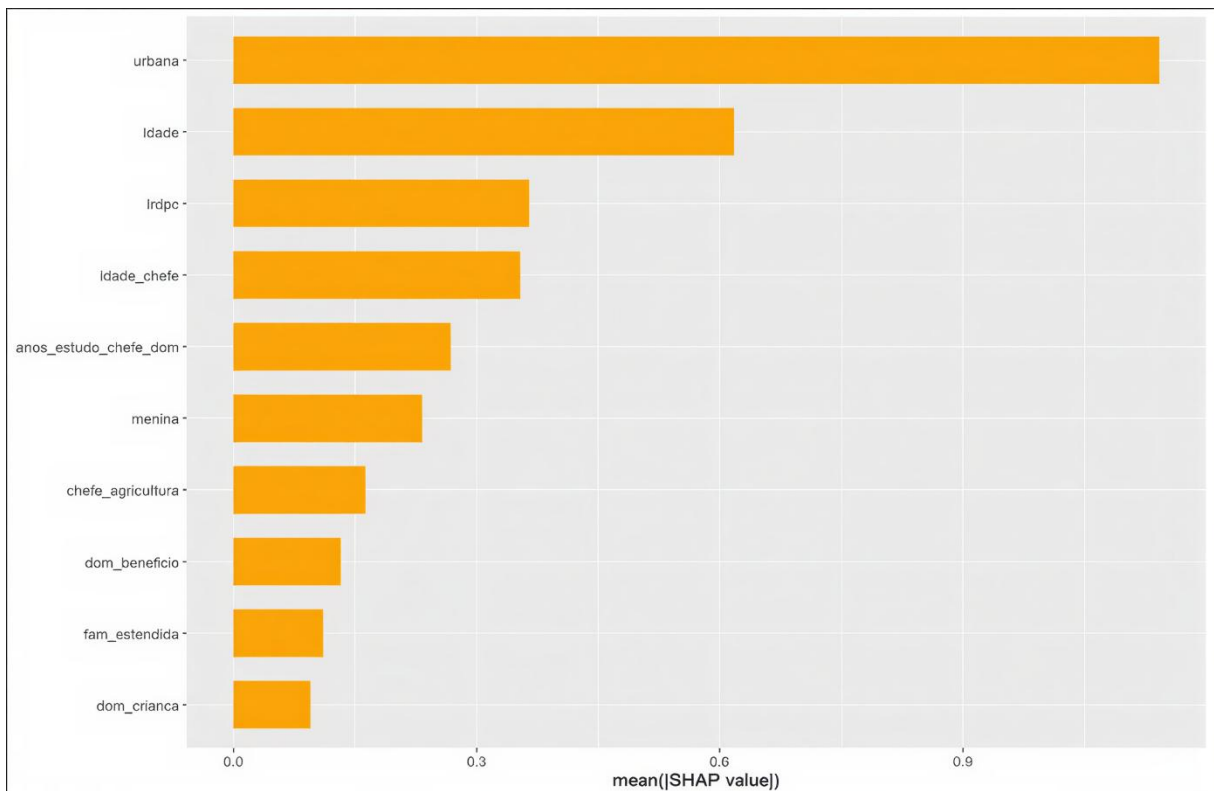
Com o intuito de ampliar a interpretação do modelo XGBoost, os valores SHAP foram calculados com base no conjunto de dados de treinamento, seguindo o procedimento proposto por Lundberg e Lee (2017). Esse arcabouço permite decompor cada predição do modelo em contribuições incrementais das variáveis explicativas, mesurando a intensidade e o sentido (positivo ou negativo) da influência de cada atributo sobre a probabilidade estimada de incidência do trabalho infantil. Nesse sentido, o exame transcende a avaliação do desempenho preditivo, possibilitando o confronto direto entre o aprendizado de máquina e a econometria tradicional, ao viabilizar padrões não lineares e heterogeneidades nos determinantes do

fenômeno.

A Figura 2 mostra a importância global das 10 variáveis mais importantes no modelo de previsão de trabalho infantil, medida pelo valor médio absoluto dos valores SHAP. A variável “urbana” se destaca como a variável de maior relevância preditiva no modelo, indicando que a localização do domicílio (urbano *versus* rural) é o fator que, em média, mais contribui para deslocar as previsões do modelo, ainda que o gráfico não informe a direção desse efeito. Em seguida, idade da criança aparece como a segunda variável mais importante, refletindo que diferenças etárias estão fortemente associadas à probabilidade prevista de trabalho infantil.

Entre os determinantes ligados às condições socioeconômicas do domicílio, destacam-se a renda domiciliar per capita (lrdpc) e os atributos do responsável, em especial idade e escolaridade, corroborando que a estrutura e o capital humano familiar exercem ação fundamental na explicação do fenômeno. Variáveis como sexo da criança, inserção do chefe na agricultura, recebimento de benefícios sociais, arranjos familiares (família estendida) e presença de outras crianças no domicílio embora com contribuições médias inferiores, complementam o perfil de risco mapeado pelo modelo.

Figura 2 – Importância global das variáveis na previsão de trabalho infantil (SHAP médio)



Fonte: elaboração própria com base nos dados da PNADC 2024.

A Figura 3 mostra gráficos de valores SHAP ao longo da distribuição de dados para algumas variáveis explicativas expõem padrões não lineares e heterogêneos que não seriam adequadamente capturados por modelos lineares convencionais. No eixo horizontal constam os valores assumidos por cada variável; enquanto no eixo vertical aparece o valor SHAP, sendo que valores positivos indicam uma contribuição para o aumento no risco de trabalho infantil e valores negativos revelam um efeito protetivo da variável em questão.

O primeiro diagrama da Figura 3 mostra que níveis mais baixos de escolaridade do chefe do domicílio estão associados a valores positivos de SHAP, elevando o risco de trabalho infantil, enquanto o aumento dos anos de estudo reduz progressivamente esse risco. Comparando esse resultado com o modelo Logit estimado, o *odds ratio* é inferior a 1 e altamente significativo, confirmando que maior escolaridade do chefe do domicílio atua como fator de proteção. Uma vantagem é que o SHAP acrescenta que esse efeito não é linear, tornando-se mais forte após determinados níveis de escolaridade.

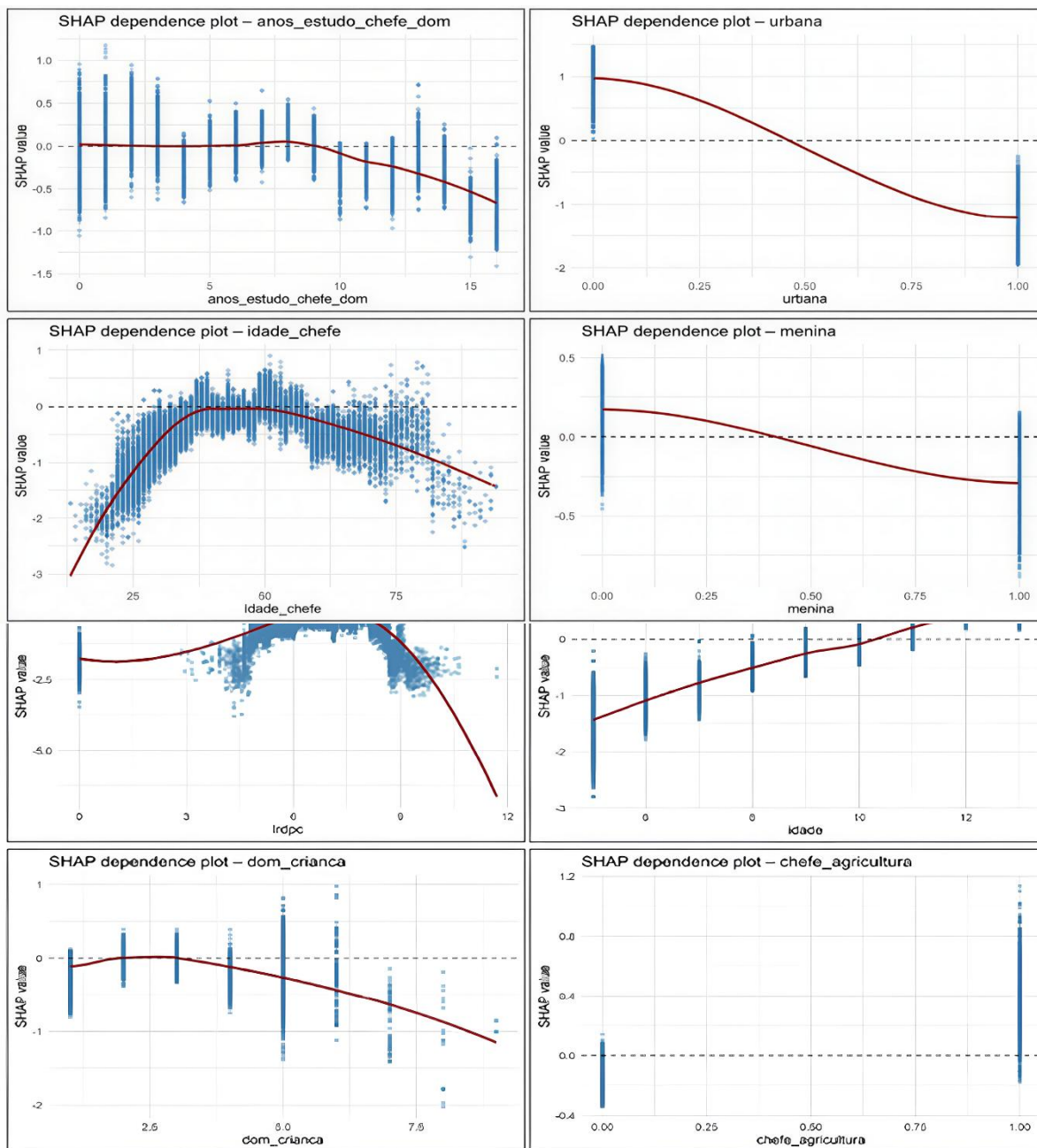
O valor SHAP mostra um efeito fortemente negativo para domicílios urbanos, reduzindo de forma expressiva o risco previsto de trabalho infantil. Esse resultado está de acordo com o Logit, em que a variável “urbana” apresentou razão de chance muito inferior a 1 e altamente significativa. A análise via SHAP sugere uma relação não linear entre a idade do chefe do domicílio e o risco de trabalho infantil. Há um aumento no risco quando o chefe é muito jovem, atingindo um máximo em idades intermediárias e passa a diminuir conforme a idade vai avançando. No modelo logit, a idade do chefe não aparece estatisticamente significativa, o que indica que o efeito médio é fraco, mas o SHAP mostra que há heterogeneidade ao longo da distribuição.

Quanto ao sexo da criança, nota-se que ser menina está associado a valores SHAP negativos, indicando menor risco relativo de trabalho infantil em comparação aos meninos. O logit confirma esse resultado com razão de chance significativamente menor que 1, sinalizando menor risco entre meninas. A renda domiciliar *per capita* mostra uma relação monotônica inversa e não linear: níveis muito baixos de renda elevam o risco de trabalho infantil, enquanto o aumento da renda exerce efeito protetivo com retornos marginais decrescentes. O SHAP indica aumento quase monotônico do risco ao trabalho infantil à medida que a idade da criança cresce. Esse padrão está totalmente alinhado ao logit, que apresenta razão de chance significativamente maior que 1, indicando que crianças mais velhas têm maior probabilidade de trabalhar. Os referidos achados estão em concordância com as pesquisas de Souza (2025), Jardim *et al.* (2021), Ramalho e Mesquita (2013) e Kassouf (2002).

O SHAP sugere um efeito moderado e não linear, com leve aumento do risco em

domicílios com poucas crianças e redução quando o número cresce. No modelo logit, a variável não é estatisticamente significativa, o que é consistente com o efeito médio fraco observado. Por fim, o SHAP mostra valores claramente positivos quando o chefe atua na agricultura, aumentando o risco previsto de trabalho infantil. O logit confirma fortemente esse resultado, com razão de chance elevada e altamente significativa, indicando a inserção do chefe no setor agrícola como fator relevante de aumento do risco de trabalho infantil.

Figura 3 – Valores SHAP para algumas variáveis explicativas



Fonte: elaboração própria com base nos dados da PNADC 2024.

5 CONSIDERAÇÕES FINAIS

O presente estudo verificou os perfis de risco associados ao trabalho infantil no Brasil usando os microdados da PNADC 2024, aplicando um modelo de *machine learning* XGBoost com recorte etário para crianças de 5 a 13 anos de idade. Os achados enfatizam que o trabalho infantil está ligado a várias dificuldades estruturais como menor renda *per capita* familiar, baixa escolaridade do chefe do domicílio, residências em áreas rurais, inserção ocupacional no setor agrícola do chefe do domicílio e crianças com maior idade.

A modelagem via XGBoost apresentou desempenho superior ao Logit, reforçando o potencial de algoritmos de *machine learning* para investigar fenômenos sociais complexos, marcados por não linearidades e múltiplas interações. A aplicação dos valores SHAP complementou a análise ao destacar padrões não lineares como retornos marginais crescentes dos anos de estudo do chefe do domicílio, relação não monotônica da sua idade e diversidade nos efeitos marginais das variáveis ao longo da distribuição.

No âmbito das políticas públicas, as evidências sugerem o fortalecimento de estratégias direcionadas a prevenir e eliminar o trabalho precoce de crianças e adolescentes, como o fortalecimento dos programas sociais com ênfase nos domicílios de baixa renda e na área rural cuja pessoa de referência esteja inserida em atividades agrícolas, e políticas educacionais que evitem a evasão escolar para crianças em faixas etárias elevadas. O uso de algoritmos de *machine learning* permite que os gestores se antecipem ao identificar e proteger as crianças do ciclo de exploração.

Dentre as limitações do estudo, tem-se a transversalidade dos dados que impossibilita análises dinâmicas ao longo do tempo e inferências causais. Trabalhos futuros podem incluir mais anos da PNADC para monitorar a persistência temporal do trabalho infantil através de uma modelagem com dados em painel e comparar com outros algoritmos de aprendizado de máquina.

REFERÊNCIAS

- ARRUDA, K. M.; DUAILIBE, M. D. Resgate das políticas públicas de combate ao trabalho infantil no Brasil. **Revista de Informação Legislativa**, Brasília, v. 60, n. 237, p. 35-58, abr./jun. 2023. Disponível em: https://www12.senado.leg.br/ril/edicoes/60/237/ril_v60_n237_p35. Acesso em: 19 jan. 2026.
- BARROS, R. P.; MENDONÇA, R.; VELAZCO, T. **Is poverty the main cause of child work in urban Brazil?** Rio de Janeiro: IPEA, 1994. (Texto para discussão, n. 131). Disponível em: <https://repositorio.ipea.gov.br/entities/publication/4d2e5bc4-4cbe-408e-b2fd-6c136017a5e3/full>. Acesso em: 19 jan. 2026.
- BASU, Kaushik; PHAM, Van H. The economics of child labor. **American Economic Review**, [s. l.], v. 88, n. 3, p. 412-427, jun. 1998. Disponível em: http://qed.econ.queensu.ca/pub/faculty/sumon/basu_childlabor.pdf. Acesso em: 20 jan. 2026.
- BINGÖL, Mehmet Rohat; ÇILINGIRTÜRK, Ahmet Mete. Comparison and analysis of logistic regression, Naïve Bayes and random forest machine learning algorithms for predicting child labor in Turkey. In: INTERNATIONAL APPLIED STATISTICS CONGRESS, 4., 2023, Sarajevo. **Proceedings** [...]. Sarajevo: UYIK, 2023. p. 281-294.
- BRASIL. [Constituição (1988)]. **Constituição da República Federativa do Brasil**. Brasília, DF: Centro Gráfico, 1988. Disponível em: https://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm. Acesso em: 19 jan. 2026.
- BRASIL. [Estatuto da Criança e do Adolescente (1990)]. **Lei nº 8.069, de 13 de julho de 1990**. Dispõe sobre o Estatuto da Criança e do Adolescente e dá outras providências. Brasília, DF: Presidência da República, 1990. Disponível em: https://www.planalto.gov.br/ccivil_03/leis/18069.htm. Acesso em: 19 jan. 2026.
- BRASIL. Ministério do Desenvolvimento Social e Combate à Fome; **ORGANIZAÇÃO INTERNACIONAL DO TRABALHO (OIT). III Conferência Global sobre Trabalho Infantil: relatório final**. Brasília, DF: Secretaria de Avaliação e Gestão da Informação, 2014. Disponível em: https://www.ilo.org/sites/default/files/wcmsp5/groups/public/@americas/@ro-lima/@ilo-brasilia/documents/publication/wcms_398475.pdf. Acesso em: 20 jan. 2026.
- BRASIL. Ministério do Trabalho e Emprego. Secretaria de Inspeção do Trabalho. **Mapa de indicativos do trabalho da criança e do adolescente**. 3. ed. Brasília, DF: MTE, SIT, 2005. Disponível em: https://acesso.mte.gov.br/data/files/FF8080812BCB2790012BD51D297640E0/pub_7746.pdf. Acesso em: 20 jan. 2026.
- BRASIL. Ministério dos Direitos Humanos. **III Plano Nacional de Prevenção e Erradicação do Trabalho Infantil e Proteção ao Adolescente Trabalhador (2019-2022)**. Brasília, DF: MDH, 2018. Disponível em: <https://bibliotecadigital.mdh.gov.br/jspui/handle/192/9036>. Acesso em: 19 jan. 2026.
- CACCIAMALI, M. C.; TATEI, F. Trabalho infantil e o status ocupacional dos pais. **Revista de Economia Política**, São Paulo, v. 28, n. 2, p. 269-290, abr./jun. 2008. Disponível em:

<https://www.scielo.br/j/rep/a/jhq6GLhStLmbnQTy4BNVgdG/?format=html&lang=pt>. Acesso em: 19 jan. 2026.

CARDOSO, E.; SOUZA, A. P. **The impact of cash transfers on child labor and school attendance in Brazil**. USA: Vanderbilt University, 2004. (Working Paper, n. 0407).

CHEN, T. *et al.* **Package ‘xgboost’**: R version. [S. l.]: CRAN, 2026.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. In: ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 22., 2016, [s. l.]. **Proceedings [...]**. [S. l.]: ACM, 2016. p. 785-794.

CUMMINGS, P. M. Child labor and household composition: determinants of child labor in Mexico. **Asian Journal of Latin American Studies**, [s. l.], v. 29, n. 3, p. 29-54, 2016. Disponível em: <https://bettercarenetwork.org/sites/default/files/ChildLab.pdf>. Acesso em: 20 jan. 2026.

DATA ZOOM. **Data Zoom**: simplifying access to Brazilian microdata. Rio de Janeiro: PUC-Rio, 2023.

DE SOUZA, G. B.; ROCHA, L. E. de V. Determinantes socioeconômicos do trabalho infantil: uma análise logit multinomial para a região Sul do Brasil. **Revista de Gestão e Secretariado**, [s. l.], v. 14, n. 8, p. 14110-14131, 2023. Disponível em: <https://ojs.revistagesec.org.br/secretariado/article/view/2657>. Acesso em: 20 jan. 2026.

FAWCETT, T. An introduction to ROC analysis. **Pattern Recognition Letters**, [s. l.], v. 27, n. 8, p. 861-874, 2006.

FRANÇA, João Italo Dias. **Avaliação de índices para a escolha do ponto de corte ótimo em testes de diagnóstico**. 2024. Tese (Doutorado em Epidemiologia) – Universidade de São Paulo, São Paulo, 2024.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning**: data mining, inference, and prediction. New York: Springer, 2013.

INTERNATIONAL LABOUR ORGANIZATION (ILO). **The worst forms of child labour, education and violent conflict: background paper**. [S.l.: s.n.], 2010. Disponível em: <<https://unesdoc.unesco.org/ark:/48223/pf0000190779>>. Acesso em: 20 jan. 2026.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA (IBGE). **País tinha 1,650 milhão de crianças e adolescentes em situação de trabalho infantil em 2024**. Rio de Janeiro: Agência de Notícias, 19 set. 2025. Disponível em: <https://agenciadenoticias.ibge.gov.br/agencia-noticias/2012-agencia-de-noticias/noticias/44541-pais-tinha-1-650-milhao-de-criancas-e-adolescentes-em-situacao-de-trabalho-infantil-em-2024>. Acesso em: 31 de mar. 2026.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA (IBGE). **Pesquisa Nacional por Amostra de Domicílios Contínua – PNAD Contínua**. Rio de Janeiro: IBGE, [s. d.]. Disponível em: <https://www.ibge.gov.br/estatisticas/sociais/trabalho/17270-pnad-continua.html>. Acesso em: 22 jan. 2026.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA (IBGE). **Pesquisa Nacional por Amostra de Domicílios Contínua: Notas técnicas**. Versão 1.19. Rio de Janeiro: IBGE, 2025. 1 PDF. ISBN 978-85-240-4651-3. Disponível em: https://biblioteca.ibge.gov.br/visualizacao/livros/liv102172_notas_tecnicas.pdf. Acesso em: 31 de mar. 2026.

INTERNATIONAL LABOUR ORGANIZATION; UNITED NATIONS CHILDREN'S FUND. **Child labour: global estimates 2024, trends and the road forward**. Geneva: ILO; New York: UNICEF, 2025.

ITAOU, B. F. **Trabalho infantil e arranjos familiares: ênfase nas famílias monoparentais brasileiras no ano de 2019**. 2023. Dissertação (Mestrado em Ciências Sociais Aplicadas) – Universidade Estadual de Ponta Grossa, Ponta Grossa, 2023. Disponível em: <https://tede2.uepg.br/jspui/handle/prefix/4072>. Acesso em: 19 jan. 2026.

FLUSS, Ronen; FARAGGI, David; REISER, Benjamin. Estimation of the Youden Index and its associated cutoff point. **Biometrical Journal: Journal of Mathematical Methods in Biosciences**, [s. l.], v. 47, n. 4, p. 458-472, 2005.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **An introduction to statistical learning with applications in R**. New York: Springer, 2013.

JARDIM, S. C.; OLIVEIRA, P. R.; TEIXEIRA, E. C.; LIRIO, V. S. Gênero do chefe de família e incidência de trabalho infantil no Brasil. **Revista de Desenvolvimento Econômico – RDE**, Salvador, v. 3, n. 50, p. 59-86, dez. 2021. Disponível em: <https://revistas.unifacs.br/index.php/rde/article/viewFile/7136/4536>. Acesso em: 19 jan. 2026.

KASSOUF, A. L. **Aspectos sócio-econômicos do trabalho infantil no Brasil**. Brasília, DF: Secretaria dos Direitos Humanos, 2002.

KASSOUF, A. L. O que conhecemos sobre o trabalho infantil. **Nova Economia**, Belo Horizonte, v. 17, n. 2, p. 323-350, mai./ago. 2007.

KASSOUF, A. L.; JUSTUS, M. Trabalho infantil no meio rural brasileiro: evidências sobre o “paradoxo da riqueza”. **Economia Aplicada**, Ribeirão Preto, v. 14, n. 3, p. 339–353, 2010. Disponível em: <https://www.scielo.br/j/ecoa/a/7Snp4CvzmVQ7bkwRdZNhqZB/>. Acesso em: 19 jan. 2026.

LUNDBERG, Scott M.; LEE, Su-In. A unified approach to interpreting model predictions. **Advances in Neural Information Processing Systems**, [s. l.], v. 30, 2017.

MACEDO, Joana de Negrier Almeida e. **Trabalho infantil: representações sociais nos media**. Lisboa: DGERT, 2012. (Cadernos de Emprego e Relações de Trabalho, n. 9). Disponível em: <https://www.dgert.gov.pt/wp-content/uploads/2019/04/CERT-09.pdf>. Acesso em: 21 jan. 2026.

MACIEL, Glenda Pereira. **Trabalho infantil no Brasil: uma análise do impacto da escolaridade do chefe de família sobre a inserção da criança no mercado de trabalho**. 2018. Monografia (Graduação em Ciências Econômicas) – Universidade Federal de Ouro Preto, Mariana, 2018. Disponível em: <https://www.monografias.ufop.br/handle/35400000/1413>. Acesso em: 19 jan. 2026.

MENEZES, Aline Maria Souza. **Determinantes do trabalho infantil no Brasil**: uma análise a partir da pesquisa de padrões de vida do IBGE. 2007. Trabalho de Conclusão de Curso (Graduação em Ciências Econômicas) – Universidade Federal do Ceará, Fortaleza, 2007. Disponível em: <https://repositorio.ufc.br/handle/riufc/70342>. Acesso em: 19 jan. 2026.

ORGANIZAÇÃO INTERNACIONAL DO TRABALHO (OIT). **Dimensões do trabalho infantil no Brasil**: desafios persistentes e caminhos para acelerar a erradicação. Brasília, DF: OIT, 2025. Disponível em: <https://www.ilo.org/pt-pt/publications/dimensoes-do-trabalho-infantil-no-brasil-desafios-persistentes-e-caminhos>. Acesso em: 19 jan. 2026.

ORGANIZAÇÃO INTERNACIONAL DO TRABALHO (OIT). **O fim do trabalho infantil: um objetivo ao nosso alcance** = *The End of Child Labour: Within Reach: global report under the follow-up to the ILO Declaration on Fundamental Principles and Rights at Work*. Genebra: International Labour Office, 2006. Disponível em: <https://www.ilo.org/pt-pt/publications/o-fim-do-trabalho-infantil-um-objectivo-ao-nosso-alcance>. Acesso em: 21 jan. 2026.

QUIÑONES FABIAN, Ricky Bryan *et al.* **An algorithm to estimate the risk of child labor**. *Decision Science Letters*, [s. l.], v. 11, p. 1–12, 2022.

RAMALHO, Hilton Martins de Brito; MESQUITA, Shirley Pereira de. Determinantes do trabalho infantil no Brasil urbano: uma análise por dados em painel 2001-2009. **Economia Aplicada**, São Paulo, v. 17, n. 2, p. 193-225, 2013. Disponível em: <https://www.scielo.br/j/ecoa/a/qBJVfnyZCB6QbLGbtKtFRsK/>. Acesso em: 19 jan. 2026.

ROCHA, Luiz Eduardo. Determinantes da alocação domiciliar do tempo das crianças e adolescentes: uma análise logit multinomial para os territórios rural e urbano do estado de Minas Gerais. **Revista Econômica do Nordeste**, [s. l.], v. 49, n. 4, p. 35–51, 2018. Disponível em: <https://www.bnb.gov.br/revista/ren/article/view/561>. Acesso em: 19 jan. 2026.

RODRIGUES, Diego C.; PRATA, David N.; SILVA, Michel A. **Exploring social data to understand child labor**. *International Journal of Social Science and Humanity*, v. 5, n. 1, p. 29–33, jan. 2015. DOI: 10.7763/IJSSH.2015.V5.416.

ROMERO-LUNA, Roxana-María; ORDOÑEZ-ERAZO, Hugo-Armando; COBOS-LOZADA, Carlos-Alberto. IA aplicada a la identificación de características de ocupación de menores en hogares colombianos para detectar posible trabalho infantil. **Revista Científica**, Bogotá, v. 50, n. 2, p. 55–70, maio/ago. 2024.

Rosenzweig, M. L., 1981. “**Household and Non-Household Activities of Youths: Issues of Modeling, Data and Estimation Strategies**” in *Child Work*. Poverty and Underdevelopment. Ed. Gerry Rodgers and Guy Standing. ILO, Geneva.

SAVE THE CHILDREN. **Relatório mundial sobre trabalho infantil e proteção social 2013**. Londres: Save the Children, 2013. Disponível em: https://resourcecentre.savethechildren.net/pdf/2013_world_report_on_cl_and_social_protection_pt_web.pdf. Acesso em: 19 jan. 2026.

SILVA, Vitor Hugo Miro Couto; DE FRANÇA, João Mário Santos. **MODELOS DE**

MACHINE LEARNING NA CLASSIFICAÇÃO DE POBREZA: UMA APLICAÇÃO PARA O ESTADO DO CEARÁ. 2021.

SOUZA, Gabrielle Klen Basilio. **A relação entre os arranjos familiares e a incidência do trabalho infantil.** 2025. Monografia (Bacharelado em Economia) – Universidade Federal de Ouro Preto, Mariana, 2025. Disponível em: https://www.monografias.ufop.br/bitstream/35400000/7740/1/MONOGRRAFIA_Relac3%a7%3%a3oArranjosFamiliares.pdf. Acesso em: 19 jan. 2026.

UNICEF. *Action Against Child Labour: Selected case studies from UNICEF Programmes.* [S.l.: s.n.], 2018. Disponível em: <https://www.unicef.org/media/155001/file/Child%20Labour%20Case%20Studies.pdf>. Acesso em: 20 jan. 2026.

ZEINI, N. T.; FIKRY, P. M. Leveraging machine learning and resampling techniques to analyze contributing factors to child labor in Egypt. **Journal of Computational Social Science**, [s. l.], v. 8, n. 4, p. 87, 2025.

WEBBINK, E.; SMITS, J.; DE JONG, E. Household and context determinants of child labor in 221 districts of 18 developing countries. **Social Indicators Research**, [s. l.], v. 110, n. 2, p. 819-836, 2013.

APÊNDICE A – Razões de chance estimadas no modelo Logit

Coefficiente	Razão de chance	Desvio padrão	z	P-valor	Intervalo 95% confiança	
menina	0,5836	0,0354	-8,87	0,0000	0,5181	0,6573
cor_branca	0,7603	0,0644	-3,24	0,0010	0,6440	0,8974
idade	1,2499	0,0161	17,29	0,0000	1,2186	1,2818
escola	4,3604	2,5521	2,52	0,0120	1,3845	13,7318
inseg_alim	1,0650	0,0682	0,98	0,3250	0,9394	1,2073
aluguel	0,9910	0,0819	-0,11	0,9130	0,8427	1,1652
falta_energia	0,4288	0,2576	-1,41	0,1590	0,1320	1,3921
água_canaliz	0,8115	0,1120	-1,51	0,1300	0,6191	1,0636
esgoto_inad	0,9127	0,0628	-1,33	0,1840	0,7976	1,0443
lixo_inad	1,1193	0,0929	1,36	0,1740	0,9512	1,3170
norte	1,2382	0,1566	1,69	0,0910	0,9662	1,5865
nordeste	1,1583	0,1374	1,24	0,2150	0,9179	1,4614
sudeste	0,9108	0,1146	-0,74	0,4570	0,7117	1,1653
centro_oeste	1,2821	0,1870	1,70	0,0880	0,9632	1,7063
urbana	0,1431	0,0116	-24,03	0,0000	0,1221	0,16768
lrdpc	1,1433	0,0497	3,08	0,0020	1,0499	1,2449
dom_benef	1,0697	0,0810	0,89	0,3740	0,9221	1,2408
dom_aposent	1,4676	0,1663	3,39	0,0010	1,1753	1,8325
fam_estend	0,7236	0,0732	-3,20	0,0010	0,5934	0,8823
fam_comp	1,4712	0,4116	1,38	0,1680	0,8502	2,5457
dom_criança	1,0230	0,0314	0,74	0,4590	0,9632	1,0864
dom_idoso	0,8417	0,0839	-1,73	0,0840	0,6923	1,0233
dom_nemnem	0,9506	0,0696	-0,69	0,4890	0,8234	1,0972
chefe_fem	1,0629	0,0756	0,86	0,3910	0,9247	1,2218
chefe_branco	1,1314	0,0963	1,45	0,1470	0,9575	1,3368
chefe_idad	1,0035	0,0038	0,92	0,3560	0,9960	1,0110
chefe_ocupad	1,1375	0,1023	1,43	0,1520	0,9535	1,3568
chefe_previd	0,7901	0,0682	-2,73	0,0060	0,6671	0,9356
chefe_estudo	0,9595	0,0079	-5,05	0,0000	0,9442	0,9749
chefe_agric	1,9725	0,1680	7,98	0,0000	1,6692	2,3308
Constante	0,0009	0,0006	-9,96	0,0000	0,0002	0,0034

Fonte: Elaboração própria com base nos dados da PNADC 2024.