UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA APLICADA

ESTÉFANE PEREIRA PINTO DE SOUZA MANHÃES

A EVASÃO ESCOLAR NO ENSINO MÉDIO E A VULNERABILIDADE SOCIOECONÔMICA NO BRASIL: UM ESTUDO BASEADO NA PNAD CONTÍNUA 2019 E 2022

BELO HORIZONTE - MG

2023

ESTÉFANE PEREIRA PINTO DE SOUZA MANHÃES

A EVASÃO ESCOLAR NO ENSINO MÉDIO E A VULNERABILIDADE SOCIOECONÔMICA NO BRASIL: UM ESTUDO BASEADO NA PNAD CONTÍNUA 2019 E 2022

Monografia apresentada ao Programa de Pós-Graduação em Estatística do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais para obtenção do título de Especialista. Área de Concentração: Estatística Aplicada.

Orientador: Prof. Dr. Guilherme Lopes Oliveira.

BELO HORIZONTE - MG

2023

Manhães, Estéfane Pereira Pinto de Souza.

M277e A evasão escolar no ensino médio e a vulnerabilidade socioeconômica no Brasil [recurso eletrônico]: um estudo baseado na PNAD contínua 2019 e 2022 / Estéfane Pereira Pinto de Souza Manhães — 2023.
1 recurso online (55 f. il, color.): pdf.

Orientador: Guilherme Lopes Oliveira.
Monografia (especialização) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística.
Referências: 53-55

1. Estatística. 2. Ensino Médio. 3. Evasão escolar. 4. Curva ROC. 5. Regressão logística I. Oliveira, Guilherme Lopes. II. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística. III. Título.

CDU 519.2 (043)

Ficha catalográfica elaborada pela bibliotecária Belkiz Inez Rezende Costa CRB 6/1510
Universidade Federal de Minas Gerais – ICEx

Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Estatística
Programa de Pós-Graduação / Especialização
Av. Pres. Antônio Carlos, 6627 - Pampulha
31270-901 – Belo Horizonte – MG

E-mail: pgest@ufmg.br
Tel: 3409-5923 – FAX: 3409-5924

ATA DO 303ª. TRABALHO DE FIM DE CURSO DE ESPECIALIZAÇÃO EM ESTATÍSTICA DE ESTÉFANE PEREIRA PINTO DE SOUZA MANHÃES.

Aos dezessete dias do mês de julho de 2023, às 14:00 horas, com utilização de recursos de videoconferência a distância, reuniram-se os professores abaixo relacionados, formando a Comissão Examinadora homologada pela Comissão do Curso de Especialização em Estatística, para julgar a apresentação do trabalho de fim de curso da aluna **Estéfane Pereira Pinto de Souza Manhães**, intitulado: "*A evasão escolar no ensino médio e a vulnerabilidade socioeconômica: um estudo segundo PNAD contínua 2019 e 2022*", como requisito para obtenção do Grau de Especialista em Estatística. Abrindo a sessão, o Presidente da Comissão, Professor Guilherme Lopes de Oliveira – Orientador, após dar conhecimento aos presentes do teor das normas regulamentares, passou a palavra à candidata para apresentação de seu trabalho. Seguiu-se a arguição pelos examinadores com a respectiva defesa da candidata. Após a defesa, os membros da banca examinadora reuniram-se sem a presença da candidata e do público, para julgamento e expedição do resultado final. Foi atribuída a seguinte indicação: a candidata foi considerada Aprovada condicional às modificações sugeridas pela banca examinadora no prazo de 30 dias a partir da data de hoje por unanimidade. O resultado final foi comunicado publicamente à candidata pelo Presidente da Comissão. Nada mais havendo a tratar, o Presidente encerrou a reunião e lavrou a presente Ata, que será assinada por todos os membros participantes da banca examinadora.

Belo Horizonte, 17 de julho de 2023.

gov.br Documento assinado digitalmente
GUILHERME LOPES DE OLIVEIRA
Data: 18/07/2023 09:50:06-0300
Verifique em https://validar.iti.gov.br

Prof. Guilherme Lopes de Oliveira (Orientador)
DECOM/CEFET-MG

gov.br Documento assinado digitalmente
GABRIEL HENRIQUE OLIVEIRA ASSUNCAO
Data: 17/07/2023 15:47:02-0300
Verifique em https://validar.iti.gov.br

M.Sc. Gabriel Henrique Oliveira Assunção
IBGE

gov.br Documento assinado digitalmente
MARCIA MARQUES DE CARVALHO
Data: 17/07/2023 18:46:50-0300
Verifique em https://validar.iti.gov.br

Profa. Dra. Márcia Marques de Carvalho
GET/UFF

Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Estatística
P Programa de Pós-Graduação / Especialização
Av. Pres. Antônio Carlos, 6627 - Pampulha
31270-901 – Belo Horizonte – MG

E-mail: pgest@ufmg.br
Tel: 3409-5923 – FAX: 3409-5924

DECLARAÇÃO DE CUMPRIMENTO DE REQUISITOS PARA CONCLUSÃO DO CURSO DE ESPECIALIZAÇÃO EM ESTATÍSTICA.

Declaro para os devidos fins que Estéfane Pereira Pinto de Souza Manhães, número de registro 2021676972, cumpriu todos os requisitos necessários para conclusão do curso de Especialização em Estatística e que entregou para seu orientador, o professor Guilherme Lopes de Oliveira, o trabalho, que aprovou a versão final. O trabalho foi apresentado no dia 17 de julho de 2023 com o título "A evasão escolar no ensino médio e a vulnerabilidade socioeconômica: um estudo segundo PNAD contínua 2019 e 2022".

Belo Horizonte, 21 de agosto de 2023

Roberto da Costa Quinino:80871291720

Prof. Roberto da Costa Quinino
Coordenador do curso de Especialização em Estatística
Departamento de Estatística / UFMG

Aos meus amores: Stela, Sarah e Sandro

AGRADECIMENTOS

Ao Prof. Dr. Guilherme Lopes Oliveira, pela sua dedicação que foi essencial na elaboração e conclusão do presente estudo. Absorvi desse momento ensinamentos que levarei para vida e para a elaboração de trabalhos futuros, foi incrível.

Aos membros da banca, por aceitarem o convite e pelas contribuições que possibilitaram aprimorar o trabalho.

Aos colegas do curso Dra. Kelly da Silva, Dr. Rafael Romero Nicolino, M.Sc Felipe Augusto Nascimento de Jesus, M.Sc Cristiano Martins Barbosa, pela parceria no desenvolvimento das disciplinas e trabalhos do curso. Em especial, à amiga M.Sc Elis Aparecida Ribeiro de Lima, pela presença constante e parceria durante esses dois anos de especialização e por aceitar mergulhar no mundo da PNADC.

Ao meu querido amigo M.Sc Selmo Eduardo Pires (*in memoriam*), pelo incentivo ao estudo da estatística durante minha formação acadêmica.

Ao meu marido Sandro de Souza Manhães, pelo apoio durante o curso e por entender minha ausência.

As minhas princesas Stela Pereira Manhães e Sarah Pereira Manhães, por estarem sempre ao meu lado estudando ou brincando de trabalhar comigo, a presença de vocês foi fortalecedora.

A minha querida Maria Aparecida Anjos de Fortes, por cuidar com tanto carinho das minhas pequenas durante o período de dedicação à especialização.

Aos meus pais Valéria Quintanilha Pinto e Eduardo José Pereira de Souza, por toda a formação enquanto pessoa.

"*...é preciso ter esperança, mas ter esperança do verbo esperançar; porque tem gente que tem esperança do verbo esperar. E esperança do verbo esperar não é esperança, é espera. Esperançar é se levantar, esperançar é ir atrás, esperançar é construir, esperançar é não desistir! Esperançar é levar adiante, esperançar é juntar-se com outros para fazer de outro modo…".* (Paulo Freire, 1992)

RESUMO

A evasão escolar representa a saída provisória ou definitiva do estudante da escola, que pode ser explicada por questões que abrangem a vulnerabilidade socioeconômica, o contexto sociodemográfico e a necessidade de o indivíduo ingressar no mercado de trabalho para contribuir com o rendimento domiciliar. Este trabalho visa à identificação de fatores sociais, econômicos e demográficos que contribuem para uma maior ou menor chance de indivíduos com idade entre 15 e 17 anos evadirem no Ensino Médio. Aplicou-se o modelo de regressão logística aos microdados do suplemento de educação da PNAD Contínua (PNADc) para os anos de 2019 e 2022, um período prévio e outro posterior à pandemia da Covid-19. Explorou-se o efeito das variáveis sexo, condição de chefe de família, cor/raça, zona de residência, contemplação pelo Programa Bolsa Família (PBF), recebimento de Pensão LOAS, recebimento de Pensão Alimentícia, recebimento de Bolsa de Estudo, macrorregião de residência, condição de trabalho e rendimento domiciliar *per capita* sobre o desfecho. Observou-se que ser chefe de família, ser contemplado pelo PBF, receber Pensão LOAS e exercer alguma atividade laboral aumentam a chance do indivíduo evadir. No entanto, residir em zona urbana e ter maior renda domiciliar *per capita* diminuem a chance do indivíduo evadir. Os achados deste estudo evidenciam que a dedicação em atividades laborais pode influenciar no abandono dos estudos no Ensino Médio. Assim, indivíduos provenientes de domicílios em contexto econômico mais favorecido e que, portanto, não precisam trabalhar para ajudar no sustento do domicílio, têm maior propensão de seguir nos estudos e atingir maior nível de escolaridade. Na subamostra da PNADc selecionada para o estudo, dentre os indivíduos que recebem Bolsa de Estudo, o percentual de não evadidos é de 100%, o que indica que indivíduos com esse suporte tendem a não evadir. Um resultado controverso e que precisa ser melhor estudado diz respeito ao sentido do efeito do PBF, cuja presença foi apontada como fator que favorece a evasão. Em ambos os períodos, 2019 e 2022, os modelos finais tiveram poder discriminatório aceitável, apresentando uma área sob a curva ROC superior a 0,70.

Palavras-chave: Evasão. Ensino Médio. PNAD Contínua. Regressão Logística. Curva ROC.

ABSTRACT

School dropout is the temporary or permanent departure of a student from school. It can be explained by socioeconomic vulnerability, sociodemographic context, and the need for the individual to enter the labor market to contribute to household income. This study aims to identify social, economic and demographic factors that contribute to a greater or lesser chance of individuals aged 15 to 17 years old dropping out of high school. The logistic regression model was applied to the microdata of the educational supplement of the Continuous National Household Survey (PNADc) for the years 2019 and 2022, a period before and after the Covid-19 pandemic. The effect of the variables sex, age, head of household status, race/ethnicity, residential area, coverage by the Bolsa Família Program (PBF), receipt of LOAS Pension, and exercising some labor activity increase the chance of the individual dropping out. However, living in urban areas and having higher per capita household income decrease the chance of the individual dropping out. The findings of this study show that dedication to labor activities can influence the abandonment of studies in high school. Thus, individuals from families in a more favorable economic context and who, therefore, do not need to work to help support the household, are more likely to continue their studies and reach a higher level of education. In the PNADc sub-sample selected for the study, among individuals who receive a scholarship, the percentage of non-dropouts is 100%, which indicates that individuals with this type of support tend not to drop out. A controversial result that needs to be further studied is about the meaning of the effect of the PBF, which presence was pointed out as a factor that favors dropout. In both periods, 2019 and 2022, the final models had acceptable discriminatory power, presenting an area under the ROC curve greater than 0.70.

Keywords: Dropout. High School. Continuous National Household Survey. Logistic regression. ROC curve.

SUMÁRIO

1 INTRODUÇÃO

No Brasil, o direito à educação está pautado no artigo 208° da Constituição Federal, que promulga: O dever do Estado com a educação será efetivado mediante a garantia de: educação básica obrigatória e gratuita dos 4 (quatro) aos 17 (dezessete) anos de idade, assegurada inclusive sua oferta gratuita para todos os que a ela não tiveram acesso na idade própria (BRASIL, 1988). Atualmente, a Resolução nº 3 do Conselho Nacional de Educação, de 3 de agosto de 2005, organiza os níveis de ensino de acordo com a idade e agrupa a Educação Infantil para pessoas de 0 a 5 anos de idade; o Ensino Fundamental, que abrange nove anos de estudo, sendo subdividido em anos iniciais para pessoas de 6 a 10 anos de idade e em anos finais para pessoas de 11 a 14 anos de idade; e estabelece o Ensino Médio para pessoas de 15 a 17 anos de idade (MEC,2009).

O Ensino Médio, segundo a LDB (Lei de Diretrizes e Bases da Educação Nacional) n° 9.394/96, é a etapa final da Educação Básica, cujas finalidades são: "a consolidação e o aprofundamento dos conhecimentos adquiridos no Ensino Fundamental, possibilitando o prosseguimento de estudos; a preparação básica para o trabalho e a cidadania do educando, para continuar aprendendo, de modo a ser capaz de se adaptar com flexibilidade a novas condições de ocupação ou aperfeiçoamento posteriores". A LDB coloca o Ensino Médio como o nível de ensino que prepara o aluno para o ingresso no mercado de trabalho ou na Universidade.

O Censo Escolar da Educação Básica mostra que em 2018 havia 7,71 milhões de alunos matriculados no Ensino Médio. No ano de 2019, período marcado pela pandemia da Covid-19, houve um decréscimo de 3,17% no número de alunos matriculados, correspondendo a 7,47 milhões de matrículas. As matrículas no Ensino Médio começam a se recuperarem no ano de 2020, passando para 7,55 milhões de matrículas, depois para 7,77 milhões em 2021, superando o total observado no ano de 2018, e em 2022 foram registradas 7,87 milhões de matrículas. (BRASIL,2023)

A distribuição do número de matrículas concentrou em 2018 uma participação de 87,9% na rede pública de ensino (municipal, estadual e federal), enquanto na rede privada concentrou 12,1% (conveniada e não conveniada). Em 2019 observou-se uma leve redução no percentual de matrículas na rede pública, que representou 87,4% e a rede privada 12,5%. No ano de 2020, a rede pública concentrou 87,7% e a rede privada 12,3%. Em 2021 e 2022, respectivamente, a rede pública concentrou 88,0% e 87,8% das matrículas. (BRASIL,2023)

Segundo o UNICEF (2018), no Brasil, mais de 7 milhões de estudantes que cursam a educação básica apresentam dois ou mais anos de atraso escolar, a chamada distorção idade-série. O grupo de estudantes com tal distorção é caracterizado por adolescentes que foram reprovados

ou evadiram e retornaram à escola em uma série não compatível com sua idade, sendo a maioria deles pertencentes às camadas mais vulneráveis da população. A distorção idade-série começa nos anos iniciais do ensino fundamental. Devido às dificuldades de aprendizagem, condições sociais, econômicas e culturais, uma parcela de meninos e meninas vão sendo retidos nas séries ou evadem e retornam em séries atrasadas ou não retornam, deixando de frequentar a escola.

No Ensino Médio, os dados do Censo escolar do INEP (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira) mostram uma distorção idade-série descendente. Em 2018, 28,2% dos alunos matriculados no Ensino Médio não estavam na série adequada e em 2022 esse percentual era de 22,2%. Quando se analisa a zona de localização geográfica da residência do estudante, em 2018, na zona urbana, 27,7% dos alunos matriculados estavam com distorção idade-série, mas, em 2022, o percentual diminui para 21,5%. Para a zona rural, em 2018, o cenário foi ainda mais alarmante, pois 39,0% dos estudantes do Ensino Médio estavam idade-série distorcidos, e, em 2022, o percentual reduziu para 34,4%.

Em relação à distorção idade-série no Ensino Médio para 2022, 24,2% dos estudantes da 1ª série estavam atrasados, ao passo que, na 2ª série e 3ª série, as taxas foram de, respectivamente, 21,3% e 20,4%. Observa-se que a taxa é descendente à medida que a série avança. Este fenômeno não significa que o problema da distorção idade-série está sendo corrigida. Isto indica que há um fluxo de alunos evadindo e outro migrando para a Educação de Jovens e Adultos (UNICEF, 2018).

A taxa de escolarização líquida para o Ensino Médio mede o percentual de alunos com idade entre 15 e 17 anos de idade que estão cursando a faixa de ensino ideal para sua idade, ou seja, ela fornece o percentual deste estrato da população que está com idade-série adequada. Segundo dados da PNAD Contínua (Pesquisa Nacional por Amostra de Domicílios Contínua) para o ano de 2022, no segundo trimestre, 75,2% dos brasileiros de 15 a 17 anos de idade estavam cursando o ensino médio e os outros 24,8% evadiram ou estavam cursando o Ensino Fundamental.

Na perceptiva da educação, a evasão, segundo Riffel e Malacarne (2010), é definida como "[...] ato de evadir-se, fugir, abandonar; sair, desistir; não permanecer em algum lugar. Quando se trata de evasão escolar, entende-se a fuga ou abandono da escola em função da realização de outra atividade." No presente estudo, adota-se o termo evasão como o ato do não prosseguimento até o final do ano letivo escolar em vigor, seja qual for o motivo atrelado ao abandono.

Em concordância com Dore *et al.* (2014) "[...] a evasão é um fenômeno complexo, multifacetado e multicausal, atrelado a fatores pessoais, sociais e institucionais, que podem resultar na saída provisória do aluno da escola ou na sua saída definitiva do sistema de ensino." Em muitos casos, o abandono pode ser causado por questões de vulnerabilidade socioeconômica

e necessidade do indivíduo se dedicar às atividades laborais para sua sobrevivência, sendo difícil a conciliação com os estudos.

Segundo Tafner (2005), o conceito de exclusão social permeia, além do critério de renda, fatores econômicos e não econômicos que impõem uma restrição à mobilidade social, como a posição do indivíduo no mercado de trabalho, escolaridade, cor, sexo e origem socioeconômica. Aliada à exclusão social, temos o conceito de vulnerabilidade social que se caracteriza pela marginalização de determinados segmentos populacionais que abrangem indivíduos com baixa escolaridade, negros e mulheres. Para Boff (2023), o conceito de vulnerabilidade socioeconômica é a limitação do acesso ao saneamento básico, moradia, educação, saúde, trabalho, alimentação, segurança entre outros fatores que afetam o bem-estar pessoal e social, que reduz os níveis de qualidade de vida e bem-estar da população.

Neste contexto, este trabalho visa compreender o fenômeno da evasão no Ensino Médio e identificar possíveis fatores socioeconômicos e demográficos atrelados a uma maior ou menor chance de indivíduos com idade entre 15 e 17 anos de idade evadir no Ensino Médio no Brasil, com foco principal em relação ao indivíduo desempenhar alguma atividade remunerada e/ou estar em situação de vulnerabilidade socioeconômica. Para isto, o modelo de regressão logística será aplicado aos microdados do suplemento de educação da PNADC para os anos de 2019 e 2022.

O trabalho está organizado em três seções, além desta introdução. A Seção 2 apresenta uma visão da PNADC usada como fonte de dados, uma descrição das bases de dados e uma revisão sobre o modelo de regressão logística, o qual será utilizado com o intuito de responder aos objetivos do estudo. A Seção 3 apresenta a análise das bases de dados e os resultados obtidos com a aplicação do modelo de regressão logística aos microdados da PNADC. A Seção 4 finaliza o estudo com algumas discussões e considerações finais.

2. MATERIAIS E MÉTODOS

2.1 A Pesquisa Nacional por Amostragem de Domicílio Contínua – PNAD Contínua

O Instituto Brasileiro de Geografia e Estatística (IBGE) implantou progressivamente a partir de 2006 o Sistema Integrado de Pesquisas Domiciliares (SIPD) com o objetivo de reformular as pesquisas domiciliares e criar indicadores que abordassem o mercado de trabalho e possibilitassem produzir informações sobre o desenvolvimento socioeconômico do país. A implantação do SIPD possibilitou um modelo de produção de pesquisas amostrais domiciliares coordenado desde o planejamento, execução, análise e a disseminação dos resultados (IBGE, 2023).

Segundo o IBGE (2023), atualmente a estrutura amostral do SIPD atende a PNADC, a Pesquisa Nacional de Saúde (PNS) e a Pesquisa de Orçamentos Familiares (POF). Devido à amostra mestra, construída a partir de um conjunto de unidades de área selecionadas probabilisticamente do cadastro mestre baseado no Censo Demográfico de 2010, o plano amostral empregado pelo IBGE é conglomerado em dois estágios de seleção, com estratificação das unidades primárias de amostragem. A amostragem é planejada com o objetivo de que a cada trimestre sejam visitadas 15 096 unidades primárias de amostragem distribuída pelo Território Nacional. Em cada unidade primária, são visitados 14 domicílios, totalizando 211 344 domicílios por trimestre.

A PNADC foi implantada pelo IBGE em janeiro de 2012, abrangendo todo o território nacional, com o objetivo de acompanhar continuamente as flutuações e a evolução, no curto, médio e longo prazos, da força de trabalho, além das características demográficas, de educação, de saúde e outras informações necessárias para se estudar o desenvolvimento econômico do país. Conforme mencionado anteriormente, a unidade de investigação é o domicílio e, considerando a abrangência territorial e os pesos amostrais empregados, pode-se dizer que as estatísticas geradas retratam a realidade da população brasileira.

Conforme destaca Braga e Assunção (2023), para atingir os objetivos supracitados, a PNADC produz indicadores trimestrais sobre a força de trabalho e indicadores anuais sobre temas suplementares permanentes (tais como as diferentes formas de trabalho, cuidados de pessoas e tarefas domésticos, tecnologia da informação e da comunicação, etc.), os quais são investigados em um trimestre específico ou aplicados em uma parte da amostra a cada trimestre e acumulados para gerar resultados anuais, sendo produzidos, também, com periodicidade variável, indicadores sobre outros temas suplementares. A PNADC contém tópicos suplementares voltados ao estudo

de temas específicos, como educação. O presente estudo está baseado na parte suplementar sobre educação e atenção primária à saúde que é divulgada anualmente no 2º trimestre da PNADC.

2.2 Descrição da Base de Dados

Os dados da PNADC são desagregados por indivíduo-domicílio e apresentam um tamanho amostral considerável, sendo dispostos em arquivos de microdados que exigem cuidados na sua importação e análise. Braga e Assunção (2023) mostra o passo a passo do uso dos pacotes *PNADcIBGE* e *survey* do *software* R para a análise dos microdados da PNADC. Estes pacotes visam facilitar o download, importação e análise dos dados amostrais complexos da pesquisa.

As análises realizadas neste trabalho tiveram como base os microdados do módulo suplementar da PNADC referente às informações sobre educação no segundo trimestre de 2019 e de 2022. Para fins de definição da população alvo do estudo (critérios de inclusão e exclusão), foram selecionados os indivíduos com idade entre 15 e 17 anos de idade e que se encaixam em algumas das seguintes caraterísticas de escolaridade na data da entrevista:

(i) Frequentam o Ensino Médio regular ou;
(ii) Frequentam a educação de jovens e adultos do Ensino Médio (EJA) ou;
(iii) O curso mais elevado que frequentou foi o regular do Ensino Fundamental ou do 1º grau e que concluíram com aprovação o nono ano ou;
(iv) O curso mais elevado que frequentou foi a EJA ou o supletivo do 1º grau e que concluíram com aprovação o nono ano ou;
(v) O curso mais elevado que frequentou foi o antigo científico, clássico, etc. (médio 2º ciclo) ou;
(vi) O curso mais elevado que frequentou foi o Ensino Médio ou o 2º grau regular ou;
(vii) O curso mais elevado que frequentou foi a EJA ou supletivo do 2º grau ou;
(viii) O curso mais elevado que frequentou foi o Ensino Superior (graduação).

As características de escolaridade descritas nos itens (i)-(viii) acima foram baseadas nas variáveis V3009A e V3003A da PNADC, as quais descrevem, respectivamente, o curso mais elevado que o indivíduo frequentou anteriormente e qual é o curso mais elevado que o indivíduo frequenta na data da entrevista. De modo resumido, estes critérios de inclusão e exclusão definem a seleção de uma subamostra da PNADC abrangendo todos os indivíduos com idade entre 15 e 17 anos de idade do Ensino Médio Regular ou da Educação de Jovens e Adultos com as seguintes características educacionais: concluíram o 9º ano do Ensino Fundamental e não chegaram a ingressar o Ensino Médio; concluíram o 9º ano do Ensino Fundamental e estão cursando o Ensino

Médio; estavam cursando o Ensino Médio, mas não frequentam a escola; ou os que podem já ter concluído o Ensino Médio.

Uma vez definida a população alvo, procedeu-se à definição da variável resposta Y de interesse: evasão no ensino médio, doravante denominada apenas como Evasão. Foram considerados evadidos, para os quais $Y = 1$, os indivíduos selecionados que:

(a) Frequentaram o antigo científico, clássico, etc. (médio 2º ciclo), com aprovação na primeira (o) e segunda (o) série e que não frequentam mais a escola ou;

(b) Frequentaram o Ensino Médio regular ou o 2º grau regular, com aprovação na primeira e segunda série e que não frequentam mais a escola ou;

(c) Frequentaram a EJA ou supletivo do 2º grau, com aprovação na primeira e segunda série e que não frequentam mais a escola ou;

(d) Frequentaram o Ensino Fundamental regular ou o 1º grau regular, com aprovação no Nona (o) ano e que não frequentam mais a escola ou;

(e) Frequentaram a EJA ou supletivo do 1º grau, com aprovação no Nona (o) ano e que não frequentam mais a escola.

Por outro lado, foram considerados como não evadidos, para os quais $Y = 0$, os indivíduos selecionados que:

(I) Frequentam o Ensino Médio regular ou;

(II) Frequentam a EJA do Ensino Médio ou;

(III) Frequentaram o antigo científico, clássico, etc. (médio 2º ciclo), com aprovação na terceira série, na quarta série e que não mais frequentam a escola ou;

(IV) Frequentaram o Ensino Médio regular ou do 2º grau regular com aprovação na terceira série, na quarta série e que não frequentam mais a escola ou;

(V) Frequentaram a EJA ou supletivo do 2º grau, com aprovação na terceira série, na quarta série e que não frequentam mais a escola ou;

(VI) Frequentam o Ensino Superior (graduação).

Na criação da variável resposta Y, seguindo a descrição dos itens (a)-(e) e (I)-(VI) acima, foram utilizadas as seguintes variáveis da PNAD contínua: V3009 (representa o curso mais elevado que o indivíduo frequentou anteriormente); V3013 (representa o ano/série que concluiu com aprovação); e V3003 (representa o curso que o indivíduo frequenta). Cabe ressaltar que as variáveis V3009 e V3013 são destinadas a caracterizar os indivíduos que não frequentam atualmente a escola, enquanto a V3003 caracteriza os indivíduos que frequentam a escola na data da entrevista.

Para o estudo dos fatores associados à probabilidade do indivíduo evadir, foram selecionadas 11 variáveis explicativas que compõem o questionário da PNAD contínua: *Sexo*, condição de chefe (*Chefe*), cor/raça (*Cor_Raça*), zona de residência (*Urb_rur*), recebimento de *Bolsa Família*, recebimento de *Pensão LOAS* (Lei Orgânica de Assistência Social), recebimento de *Pensão Alimentícia*, recebimento de *Bolsa de Estudo*, macrorregião de residência (*GR*), condição de trabalho (*Trabalho*) e rendimento domiciliar *per capita* (*Renda_per_capita*). Tais variáveis estão descritas no QUADRO 1, seguindo a codificação destacada aqui em itálico.

QUADRO 1 - Descrição das variáveis e suas características

Codificação da Variável	Código da variável na PNADC	Descrição	Tipo	Categorias/Escala
Evasão	V3003 V3009 V3013	Caracteriza o indivíduo que evadiu a escola	Categórica	1 – evadido 0 – não evadido
Sexo	V2007	Sexo do indivíduo	Categórica	1 – homem 0 – mulher
Chefe	V2005	Condição no domicílio	Categórica	1 – chefe 0 – não chefe
Cor_Raça	V2010	Classifica a cor do indivíduo	Categórica	1 – branco 0 – não branco
Urb_rur	V1022	Situação do domicílio	Categórica	1 – urbana 0 – rural
Bolsa Família	VI5002A	Recebe rendimentos de Programa Bolsa Família	Categórica	1 – recebe 0 – não recebe
Pensão LOAS	VI5001A	Alguém do domicílio recebe rendimentos de Benefício Assistencial de Prestação Continuada – BPC-LOAS	Categórica	1 – recebe 0 – não recebe
Pensão Alimentícia	VI5006A	Recebe rendimentos de pensão alimentícia, doação ou mesada em dinheiro de pessoa que não morava no domicílio	Categórica	1 – recebe 0 – não recebe
Bolsa de Estudo	VI5008A	Recebe outros rendimentos (bolsa de estudos, rendimento de caderneta de poupança, aplicações financeiras, etc.	Categórica	1 – recebe 0 – não recebe
Trabalho	V4001 V4002 V4003 V4004 V4005	Indica se na semana de referência o indivíduo trabalhou, estagiou, fez algum bico, ou trabalhou em alguma atividade ocasional, ajudou no trabalho remunerado de algum morador do domicílio ou de parente durante pelo menos 1 hora ou tinha algum trabalho remunerado do qual estava temporariamente afastado.	Categórica	1 – trabalha 0 – não trabalha
GR		Grandes regiões geográficas	Categórica	Sudeste(referência), Norte, Nordeste Centro-Oeste e Sul

Renda_per_capita	VDI5008	Rendimento domiciliar per capita	Contínua	Medida em reais

Fonte: Dicionário da PNAD contínua, suplemento de educação, 2019 e 2022.

A análise estatística descritiva permite conhecer o comportamento das variáveis em estudo, sendo importante principalmente para o entendimento do comportamento da variável resposta de interesse. Sendo assim, na Seção 3.1, as frequências estimadas para as categorias da variável Evasão serão apresentadas em tabelas cruzadas com relação às variáveis explicativas descritas na TABELA 1. Tal análise descritiva precederá o ajuste de modelo de regressão logística definido na próxima seção.

2.3 Modelo de Regressão Logística

O modelo de regressão logística binária é empregado quando a variável explicativa estudada assume característica de uma variável categórica com valores 0 ou 1, representado a ocorrência ou não ocorrência do evento, respectivamente (HAIR *et al.*, 2009). Para Fávero (2017), a técnica de regressão logística pode ser empregada quando o fenômeno se apresentar como uma variável qualitativa com uma ou mais categorias. No entanto, a regressão logística binária é representada pelo estudo de um fenômeno caracterizado por uma variável *dummy* com a primeira categoria sendo a referência do não evento tendo Y assumido valor zero ($Y = 0$) e a segunda categoria a ocorrência do evento de interesse na qual Y assumirá valor um ($Y = 1$). Neste estudo, a variável resposta Y é uma *dummy* tal que para o indivíduo evadido no Ensino Médio $Y = 1$ e para o indivíduo não evadido no Ensino Médio $Y = 0$.

Deseja-se prever a chance de um indivíduo com idade entre 15 e 17 anos de idade e que tenha concluído o ensino fundamental evadir no Ensino Médio, verificando possíveis fatores que influenciam numa maior ou menor chance deste evento ocorrer. Para isto, tomando como referência um conjunto de k variáveis explicativas $X_1, X_2, \ldots, X_k$, define-se

$$Z_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki},$$

onde Z_i representa o logit da probabilidade p_i do i-ésimo indivíduo evadir no Ensino Médio; α a constante; $\beta_j (j = 1, 2, \ldots, k)$ os parâmetros estimados para cada variável explicativa X_j; e i representa cada observação da amostra com $i = 1, 2, \ldots, n$, sendo n o tamanho da amostra.

Para determinar a probabilidade p_i de ocorrência do evento, $Y_i = 1$, em função dos parâmetros estimados para cada variável explicativa, definiremos o conceito de chance de ocorrência, conhecido por *odds*, da seguinte forma:

$$Chance(odds)_{Y_i=1} = \frac{p_i}{1-p_i}.$$

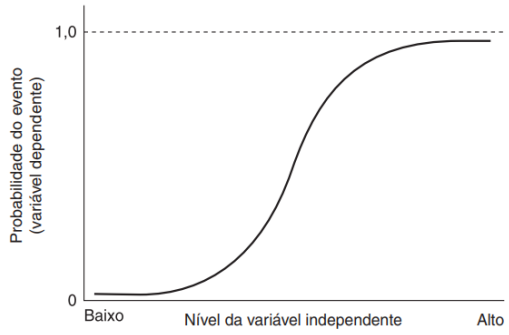
Segundo Fávero (2017), na regressão logística binária o logit Z_i é definido como o logaritmo natural (ln) da chance, sendo:

$$\ln\left(Chance_{Y_i=1}\right) = \ln\left(\frac{p_i}{1-p_i}\right) = logito(p_i) = Z_i,$$

onde, por fim, a expressão para a probabilidade de ocorrência do evento é dada por

$$p_i = \frac{e^{\alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki}}}{1 + e^{\alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki}}}.$$

O modelo de regressão logística é representado pela curva logística que apresenta o formato de S, cujo domínio está no intervalo $(0,1)$. Logo a regressão logística binária, o modelo não fornece uma estimativa direta para os valores da variável dependente, mas sim a probabilidade de ocorrência do evento em estudo. Para níveis baixos das variáveis independentes, a probabilidade tende a ser próxima de zero, enquanto para níveis altos da variável independente, a probabilidade tende a ser próxima de um, como exemplificado na FIGURA 1.



.

FIGURA 1 - Forma da relação logística entre variáveis dependente e independentes.

Fonte: HAIR *et al.*, 2009**.**

A estimação dos parâmetros do modelo de regressão logística faz uso da função de verossimilhança (*likelihood function*) que para uma amostra com n observações é definida como

$$L(p|y) = \prod_{i=1}^{n} \left[p_i^{Y_i}.(1-p_i)^{1-Y_i}\right].$$

Considerando a probabilidade de ocorrência do evento $p_i = \frac{e^{Z_i}}{1+e^{Z_i}}$ e a não ocorrência do evento dado por $1 - p_i = \frac{1}{1+e^{Z_i}}$, a função de verossimilhança pode ser expressa por

$$L(p|y) = \prod_{i=1}^{n} \left[\left(\frac{e^{Z_i}}{1+e^{Z_i}} \right)^{Y_i} . \left(\frac{1}{1+e^{Z_i}} \right)^{1-Y_i} \right],$$

de modo que o logaritmo da função de verossimilhança pode ser expresso por

$$LL(p|y) = \sum_{i=1}^{n} \left\{ \left[(Y_i). \ln \left(\frac{e^{Z_i}}{1+e^{Z_i}} \right) \right] + \left[(1-Y_i). \ln \left(\frac{1}{1+e^{Z_i}} \right) \right] \right\}.$$

Definida a função logarítmica de verossimilhança, um método usual de estimação consiste em determinar o valor dos parâmetros do logit Z_i que maximizam a expressão $LL(p|y)$. A maximização pode ser realizada utilizando ferramentas de programação linear, com objetivo de estimar os parâmetros $\alpha, \beta_1, \beta_2, \dots, \beta_k$. Detalhes teóricos sobre o processo de estimação dos parâmetros pelo método de maximização da função de log-verossimilhança podem ser encontrados em HAIR *et al.*, 2009 e McCullagh e Nelder (1989).

De posse das estimativas para os parâmetros do modelo, é preciso avaliar sua significância estatística, ou seja, analisar a evidência de que a variável explicativa associada a cada um deles é estatisticamente significante para influenciar a probabilidade de ocorrência do evento Evasão ou não. Para isso, pode-se utilizar um procedimento de teste estatístico relacionado às hipóteses

$$\begin{cases} H_0: \beta_j = 0 \\ H_1: \beta_j \neq 0 \end{cases},$$

em que a rejeição da hipótese nula H_0 indica que a j-ésima variável explicativa deve ser mantida no modelo por ter relevância estatística. Um teste comumente utilizado é o teste de Wald, cuja estatística de teste pode ser expressa como $t_{\beta_j} = \frac{\widehat{\beta_j}}{SE(\widehat{\beta_j})}$, sendo $\widehat{\beta_j}$ o estimador de máxima verossimilhança para o parâmetro β_j, para $j = \{1,2,3, \dots, k\}$, e $SE(\widehat{\beta_j})$ representa o erro-padrão (*standard error)* do estimador $\widehat{\beta_j}$. Sob a hipótese nula, a estatística de teste t_{β_j} tem distribuição aproximadamente t-Student.

É importante ressaltar que, no caso da análise baseada em dados amostrais complexos como os provenientes da PNADC, as definições teóricas do método de estimação via maximização da função de log-verossimilhança e estimativas dos erros-padrão dos estimadores são modificadas de modo a considerar o *design* e os pesos amostrais apropriadamente. Com isto em mente, utilizou-se a função *svyglm()* do pacote *survey* do *software* R, a qual permite o ajuste

de modelos lineares generalizados (no caso, a regressão logística) aos dados de pesquisas com *design* amostral complexo, com ponderação de probabilidade inversa e erros-padrão baseados no delineamento amostral do estudo. Detalhes podem ser encontrados em Lumley e Scott (2017) e na documentação do pacote descrito em Lumley (2004).

A interpretação dos parâmetros $\beta_1, \beta_2, \ldots, \beta_k$ do modelo de regressão logística não é feita de forma direta como no modelo de regressão linear usual. A interpretação é feita pela comparação da probabilidade de ocorrência do evento p_i com a probabilidade de fracasso $1 - p_i$, usando a chamada *odds Ratio* (OR, ou razão de chances em português). Essa função é obtida a partir da $Chance_{Y_i=1}$. Por simplicidade, considerando um modelo com apenas uma variável explicativa X, tem-se que, para um indivíduo particular, a chance de sucesso do evento, dado que o valor da variável preditora para esse indivíduo é $\mathrm{X} = \mathrm{x_0}$, é dada por

$$\mathrm{Chance_{Y=1}(x_0)} = \frac{\mathrm{p(x_0)}}{1 - \mathrm{p(x_0)}} = \frac{\mathrm{e^{\beta_0 + \beta_1 x_0}}}{\frac{1 + \mathrm{e^{\beta_0 + \beta_1 x_0}}}{1 + \mathrm{e^{\beta_0 + \beta_1 x_0}}}} = \mathrm{e^{\beta_0 + \beta_1 x_0}}.$$

Similarmente, para um outro indivíduo tal que $\mathrm{X} = \mathrm{x_0} + 1$, tem – se

$$\mathrm{Chance_{Y=1}(x_0 + 1)} = \frac{\mathrm{p(x_0+1)}}{1 - \mathrm{p(x_0+1)}} = \mathrm{e^{\beta_0 + \beta_1 (x_0+1)}}.$$

A comparação destes dois indivíduos através da razão entre as suas chances dadas acima será

$$OR(x_0 + 1, x_0) = \frac{e^{\beta_0 + \beta_1 (x_0+1)}}{e^{\beta_0 + \beta_1 x_0}} = e^{\beta_1}.$$

Desta forma, a razão de chances $OR(x_0 + 1, x_0)$ pode ser interpretada como sendo a magnitude com que a chance de $Y = 1$ se modifica pelo acréscimo de 1 unidade na covariável X. A chance de ocorrência do evento será aumentada se $\beta_1 > 0$, diminuirá se $\beta_1 < 0$ e será considerada igual se $\beta_1 = 0$. No caso de um modelo com múltiplas variáveis explicativas, a interpretação é feita separadamente para cada variável explicativa na forma descrita acima, considerando que as demais variáveis explicativas sejam mantidas fixadas.

2.3.1 Seleção de Variáveis

Assim como em outros contextos de modelagem estatística, a definição do modelo a ser interpretado e utilizado para predições se inicia com a escolha do conjunto de k variáveis explicativas $X_1, X_2, \ldots, X_k$ que, potencialmente, estariam atreladas ao evento de interesse. No caso

deste estudo, este passo do processo de análise está ligado à escolha dos fatores que poderiam influenciar uma maior ou menor chance de ocorrência da evasão no Ensino Médio para a população alvo.

Tal escolha é inicialmente guiada por estudos disponíveis na literatura, onde dados sobre fatores comumente mencionados à evasão são coletados, mas também é definida com base nos objetivos específicos do estudo. Visando a análise da influência de fatores ligados à carga de trabalho e vulnerabilidade social com o desfecho Evasão no Ensino Médio, o processo de modelagem das probabilidades $p_i, i = 1, \dots, n$, foi iniciado com 11 variáveis explicativas descritas na TABELA 1 (*Sexo*, *Chefe*, C*or_Raça*, *Urb_rur*, *Bolsa Família*, *Pensão LOAS*, *Pensão Alimentícia*, *Bolsa de Estudo*, *GR*, *Trabalho* e *Renda_per_capita*).

O processo de definição do modelo final, ou seja, do conjunto de variáveis explicativas que tem significância estatística para explicar a chance de evasão no Ensino Médio, pode ser feito através de diferentes métodos. Dentre os métodos de seleção de variáveis mais comuns, se destacam os métodos *forward e backward*.

No método *forward*, supõe-se inicialmente a não existência de variáveis explicativas no modelo e ele será considerado apenas contendo a constante. Na execução, serão seguidos os seguintes passos:

(i) É estabelecido individualmente um modelo de regressão logístico para cada variável que possivelmente irá compor o modelo;

(ii) Observado o valor obtido para o p-valor de cada variável no teste de significância estatística, as variáveis são dispostas em modo crescente segundo o p-valor obtido;

(iii) As variáveis são incluídas ao modelo uma a uma e observa-se a regressão obtida. Se a variável incluída por último for significativa, então passe-se à inclusão da próxima variável. Caso ela não seja significativa, ela é removida e a próxima variável é analisada;

(iv) O modelo final é estabelecido até que a inclusão da última variável disponível seja analisada. O modelo final conterá apenas variáveis significativas.

No método *backward* a suposição inicial é de que todas as variáveis explicativas fazem parte do modelo e, a partir daí, são seguidos os seguintes passos para determinar quais variáveis serão mantidas no modelo, de modo a identificar o modelo que melhor representa o problema estudado. Os passos para identificação serão:

(i) É estabelecido um modelo de regressão com todas as possíveis variáveis do modelo;

(ii) As variáveis são ordenadas de modo decrescente do p-valor obtido no teste de significância estatística;

(iii) As variáveis são excluídas do modelo uma a uma e observa-se a regressão obtida;

(iv) O modelo final é estabelecido quando, após uma das variáveis retiradas, todas as que restarem no modelo representam variáveis significativas.

Outro método bastante difundido é o chamado método *stepwise*. Ele é um processo automático de seleção de variáveis que realiza o *forward* e *backward* ao mesmo tempo. Em cada etapa, uma variável é considerada para adição ou subtração do conjunto de variáveis explicativas atualmente presentes no modelo, tomando como base algum critério pré-especificado. Normalmente, se assume a forma de uma sequência de testes de significância, mas outras técnicas são possíveis, como o critério de informação de *Akaike* (AIC) ou o critério de informação Bayesiano (BIC). Tomando apenas critérios numéricos como referência, os modelos criados pela aplicação do *stepwise* podem ser simplificações excessivas dos modelos reais dos dados, devendo-se, portanto, mesclar o uso da técnica com a observação prática do sentido das variáveis explicativas dentro do contexto em estudo.

Segundo Paula (2010), a aplicação dos métodos *forward*, *backward* e *stepwise* possibilita a redução no número de variáveis explicativas, pois auxilia na identificação da combinação que mantenha apenas as variáveis significativas ao modelo. No entanto, o autor destaca que os métodos exigem muitas estimativas por máximo verossimilhança, o que exige muito esforço computacional quando há muitas variáveis disponíveis.

Os métodos *forward* e *backward* serão explorados na análise dos dados selecionados da PNADC, considerando o nível de 10% de significância para a decisão da remoção/inclusão ou não das variáveis de acordo com o p-valor do teste de Wald.

2.3.2 Análise da Capacidade Preditiva do Modelo

Após a definição de um modelo final estatisticamente significativo, procede-se à análise da sua capacidade preditiva. Isto é feito a partir da comparação das classificações (ocorrência ou não do evento) estimadas pelo modelo $\hat{Y}$ com as classificações reais observadas Y para a variável resposta para cada elemento da amostra. De forma geral, a matriz de confusão/classificação trazendo o contraste das comparações tem a forma apresentada na FIGURA 2.

		Valor Observado	
		$Y = 1$	$Y = 0$
Valor Estimado	$\hat{Y}$=1	VP	FP
	$\hat{Y}$=0	FN	VN

FIGURA 2 - Matriz de confusão. Fonte: Adaptado de Fawcett (2006)

Na FIGURA 2, os termos VP (verdadeiro positivo) e VN (verdadeiro negativo) indicam que o valor previsto coincide com o valor observado correspondente a um sucesso (Y=1) ou a um fracasso (Y=0), respectivamente; FP (falso positivo) indica o Erro do Tipo I – quando um fracasso observado é classificado como sucesso; e FN (falso negativo) indica o Erro do Tipo II – quando um sucesso observado é classificado como sendo um fracasso.

Uma vez que os valores preditos pelo modelo estão na escala das probabilidades de sucesso, para realizar a classificação das unidades amostrais com base no modelo ajustado, é necessária a definição de uma probabilidade de referência chamada de ponto de corte. A probabilidade estimada para cada indivíduo é comparada com o ponto de corte pré-estabelecido. Se a probabilidade estimada exceder o ponto de corte, então assume-se que o resultado predito para a variável resposta deve ser igual a 1 (sucesso); caso contrário, deve ser igual a 0 (fracasso).

Define-se como Sensibilidade a porcentagem de sucessos corretamente previstos pelo modelo (VP) dentre o total de sucessos observados (VP+FN) e como Especificidade a porcentagem de fracassos corretamente previstos pelo modelo (VN) dentre o total de fracassos observados no banco de dados (VN+FP). No contexto da análise de evasão no Ensino Médio, a Sensibilidade pode ser entendida como a capacidade de identificar corretamente os indivíduos evadidos, enquanto a Especificidade pode ser entendida como a capacidade de identificar os indivíduos que não evadiram no Ensino Médio.

O ponto de corte geralmente definido nos *softwares* estatísticos é 0,50, o que equivale ao ponto central do intervalo de probabilidades (0,1) estimadas pelo modelo logístico. Em geral, quanto maior (mais próximo de 1) é o ponto de corte, maior é a Especificidade do modelo, mas menor é a sua Sensibilidade. Assim, na escolha do ponto de corte, levamos em consideração a intenção do modelo como critério de classificação. Além disso, a escolha do ponto de corte deve ser feita de forma apropriada em dados desbalanceados (quando há grande diferença no percentual de sucessos e fracassos na variável resposta), como é o caso dos dados em análise neste estudo. Conforme destaca Prates *et al.* (2023), quando o evento em estudo é raro, isto é, apresenta baixo percentual de ocorrência na amostra, a opção por pontos de corte menores que 0,50 pode levar a

modelos com valores satisfatórios para Sensibilidade e Especificidade, sem a necessidade de se balancear previamente o banco de dados antes da análise. Tais autores destacam ainda que o balanceamento pode levar a estimativas viesadas da capacidade preditiva do modelo, propondo que a modificação do ponto de corte na definição da classificação é uma estratégia menos danosa.

Neste trabalho, a análise da capacidade preditiva do modelo será feita considerando-se diferentes pontos de cortes. Como salientado por Paiva (2015), se o objetivo for escolher um ponto de corte ideal para efeitos de classificação, pode-se selecionar um ponto de corte que maximiza a Sensibilidade e Especificidade ao mesmo tempo. Neste contexto, a curva ROC (*Receiver Operating Characteristc*) é uma ferramenta gráfica que permite avaliar o desempenho de um modelo de regressão binária tendo como base uma comparação entre a Sensibilidade e a Especificidade. Ela se trata de um gráfico onde são plotadas a Sensibilidade no eixo vertical e (1-Especificidade) no eixo horizontal obtidas a partir da utilização de vários pontos de corte.

A capacidade do modelo em discriminar entre aqueles indivíduos nos quais o evento de interesse ocorreu ($Y = 1$) *versus* aqueles em o evento não ocorreu ($Y = 0$) pode ser avaliada através da área sob a curva ROC, a qual varia de 0,5 a 1,0. Quanto mais área sob a curva ROC (AUC, do inglês *area under the curve*), maior será o poder discriminante do modelo, ou seja, maior será a capacidade do teste em distinguir entre indivíduos evadidos e não evadidos. Não há um consenso na literatura a respeito de um valor de referência para AUC que indicaria uma boa discriminação. Hosmer, Lemeschow e Sturdivant (2013) propõem que: 0,5 < AUC < 0,7 indica baixo poder de discriminação; 0,7 <= AUC < 0,8 indica aceitável poder de discriminação; 0,8 <= AUC < 0,9 indica excelente poder de discriminação; e AUC >= 0,9 indica poder de discriminação acima do normal.

2.4 *Software*

Para o manuseio dos dados foi utilizado o *software* R (R Core Team, 2022), licença *Part of R 4.2.2*. Para acessar aos microdados, de domínio público, da PNADC, utilizou-se o pacote PNADcIBGE (BRAGA, 2023) intitulado *Downloading, Reading and Analyzing PNADC Microdata* na versão 0.7.2 a partir da função *get_pnadc()*. Na implementação das estatísticas descritivas e análise de regressão foi empregado o pacote s*urvey* (LUMLEY, 2023) intitulado *Analysis of Complex Survey Samples*, na versão 4.2-1 que possibilita, a partir do uso da função *svymean( ), svytotal( )* e *svyquantile( ),* calcular as principais medidas descritivas, *svyboxplot( )* traça o boxplot e *svyglm()* ajusta os modelos lineares generalizados aos dados de um *design* de pesquisa complexo, com ponderação de probabilidade inversa e erros padrão baseados no desenho amostral. O pacote WeightedROC foi usado para calcular a Sensibilidade e Especificidade.

3. RESULTADOS

3.1 Análise Descritiva da Base de Dados

Após a aplicação dos critérios de inclusão e exclusão definidos na Seção 2.2, a subamostra da PNADC a ser utilizada neste estudo para o ano de 2019 foi composta por n=17.993 indivíduos, sendo que 2,57% deles evadiram no Ensino Médio e 97,43 % não evadiram no Ensino Médio. No ano de 2022, a subamostra foi composta por n=16.204 indivíduos, sendo que 2,72% deles evadiram no Ensino Médio e 97,67% não evadiram no Ensino Médio. Nota-se, portanto, que o evento de interesse é de certa forma raro e tende a atingir uma pequena parcela da população alvo.

Com o uso da função *svymean()* do pacote *survey* é possível obter uma estimativa para o total de indivíduos evadidos, levando em conta o desenho amostral da PNADC. Para o ano de 2019, estimou-se que 6 571 804 (97,91%) não evadiram no Ensino Médio, havendo, portanto, 2,09% de evasão no Ensino Médio, o que foi correspondente a um total estimado de 140 063 indivíduos evadidos. Para 2022, o percentual de evasão foi um pouco maior (2,33%), sendo equivalente a um total de 162 288 indivíduos evadidos e 6 806 320 (97,67%) não evadidos.

A distribuição proporcional de evadidos e não evadidos por sexo e por ano é apresentada na TABELA 2. Ela mostra que, no ano de 2019, dos indivíduos não evadidos, 47,91% eram do sexo masculino e 52,09% eram do sexo feminino, ao passo que dos indivíduos evadidos, 54,71% eram do sexo masculino e 45,29% eram do sexo feminino. Para o ano de 2022, dos indivíduos não evadidos, 48,44% são sexo masculino e 51,56% do sexo feminino, para os evadidos, 50,60% são do sexo masculino, enquanto 49,40% são do sexo feminino. De modo geral, percebe-se que os indivíduos do sexo feminino são maioria e evadem menos que os indivíduos do sexo masculino. O coeficiente de variação (CV) da estimativa é mais elevado na categoria de evadidos, o que é explicado pelo tamanho amostral menor nesse subgrupo. Esse padrão também será observado em todas as análises bivariadas realizadas na sequência.

TABELA 2 - Distribuição dos indivíduos de 15 a 17 anos de idade em função da Evasão conforme variável sexo em 2019 e 2022

Sexo		Evasão			
		2019		2022	
		Não	Sim	Não	Sim
Masculino	N	3.149.009	76.626	3.296.683	82.121
	CV%	1,17	9,08	1,28	11,43
	% na categoria	47,91	54,71	48,44	50,60
Feminino	N	3.423.801	63.437	3.509.637	80.167
	CV%	1,14	9,88	1,20	9,10
	% na categoria	52,09	45,29	51,56	49,40
Total	N	6.572.810	140.063	6.806.320	162.288
	CV%	0,77	6,54	0,88	7,54
	% na categoria	100,00	100,00	100,00	100,00

Fonte: Dados extraídos pelo autor baseado nos microdados da PNAD contínua, suplemento de educação, 2019 e 2022.
Legenda: N=estimativa do total populacional, CV = coeficiente de variação

O percentual de indivíduos não evadidos caracterizados como não chefe foi de apenas 1,25% em 2019 e de 2,00% em 2022 (TABELA 3). Para o grupo de evadidos, este percentual foi de 21,00% e 16,24%, respectivamente, para 2019 e 2022. Observa-se, então, que a posição de chefe pode levar a uma maior ocorrência da evasão no Ensino Médio por indivíduos de 15 a 17 anos de idade.

TABELA 3 - Distribuição dos indivíduos de 15 a 17 anos de idade em função da Evasão conforme variável Chefe em 2019 e 2022.

Chefe		Evasão			
		2019		2022	
		Não	Sim	Não	Sim
Chefe	N	81.910	29.411	135.944	26.532
	CV%	10,63	16,86	7,56	15,63
	%na categoria	1,25	21,00	2,00	16,24
Não Chefe	N	6.490.900	110.651	6.675.325	136.819
	CV%	0,78	6,64	0,90	8,37
	%na categoria	98,75	79,00	98,00	83,76
Total	N	6.572.810	140.063	6.811.269	163.351
	CV%	0,77	6,54	0,88	7,51
	%na categoria	100,00	100,00	100,00	100,00

Fonte: Dados extraídos pelo autor baseado nos microdados da PNAD contínua, suplemento de educação, 2019 e 2022.
Legenda: N=estimativa do total populacional, CV = coeficiente de variação

Na TABELA 4 é apresentado o percentual de indivíduos classificados em branco e não brancos de acordo com a variável Cor/Raça. Em 2019, entre os indivíduos não evadidos 41,13% eram brancos e 58,87% eram não brancos; para os evadidos, 29,68% eram brancos e 70,32% eram não brancos. Para o ano de 2022, entre os indivíduos que evadiram, 41,25% eram brancos e 58,75% não brancos; para os indivíduos evadidos, 33,61% eram brancos e 66,39% eram não brancos. A tabela evidencia que a evasão ocorre em proporção maior para indivíduos não brancos em ambos os períodos analisados. Destaca-se que a estimativa do total de indivíduos foi menor no ano de 2022 devido à presença de alguns valores faltantes na variável Cor/Raça neste ano.

TABELA 4 - Distribuição dos indivíduos de 15 a 17 anos de idade em função da Evasão conforme variável Cor_Raça em 2019 e 2022

		Evasão			
		2019		2022	
Cor_Raça		Não	Sim	Não	Sim
	N (CV)	2.703.386	41.573	2.808.508	54.640
	CV%	1,49	12,09	1,65	13,29
Branco	%na categoria	41,13%	29,68%	41,25%	33,61%
	N (CV)	3.869.279	98.490	3.996.220	107.466
	CV%	1,14	7,85	1,23	8,85
Não Branco	%na categoria	58,87%	70,32%	58,75%	66,39%
	N (CV)	6.572.810	140.063	6.803.438	161.515
	CV%	0,77	6,54		
Total	%na categoria	100,00%	100,00%	100,00%	100,00%

Fonte: Dados extraídos pelo autor baseado nos microdados da PNAD contínua, suplemento de educação, 2019 e 2022.
Legenda: N=estimativa do total populacional, CV = coeficiente de variação

A distribuição percentual para os indivíduos segundo a zona de residência, área urbana ou rural, é apresentada na TABELA 5. Para o ano de 2019, 83,57% dos indivíduos não evadidos residiam na área urbana, enquanto para os indivíduos evadidos, este percentual foi de 72,57%. No ano de 2022, dos indivíduos não evadidos, 86,03% residiam em área urbana, o que para o grupo de indivíduos evadidos foi de 73,43%. Parece não haver diferença no comportamento da variável entre os dois períodos, mas em ambos é possível notar uma maior proporção de evasão em residentes de áreas rurais do que em áreas urbanas.

TABELA 5 - Distribuição dos indivíduos de 15 a 17 anos de idade em função da Evasão conforme variável Urb_rur em 2019 e 2022.

		Evasão			
		2019		2022	
Urb_rur		Não	Sim	Não	Sim
	N	5.609.937	101.645	5.855.795	119.162
	CV%	0,89	8,37	1,05	9,46
Urbana	%na categoria	85,35%	72,57%	86,03%	73,43%
	N	962.873	38.418	950.524	43.126
	CV%	2,05	8,98	2,34	10,20
Rural	%na categoria	14,65%	27,43%	13,97%	26,57%
	N	6.572.810	140.063	6.806.320	162.288
	CV%	0,77	6,54	0,88	7,54
Total	%na categoria	100,00%	100,00%	100,00%	100,00%

Fonte: Dados extraídos pelo autor baseado nos microdados da PNAD contínua, suplemento de educação, 2019 e 2022.
Legenda: N=estimativa do total populacional, CV = coeficiente de variação

Da distribuição percentual dos indivíduos quanto ao recebimento do auxílio do Programa Bolsa Família (PBF) do Governo Federal, TABELA 6, nota-se para 2019 que os indivíduos não evadidos que recebem a bolsa família representam 0,59%, enquanto 99,41% não recebem; para os evadidos, 4,44% recebem a bolsa família e 95,56% não recebem. Para o ano de 2022, do grupo de indivíduos que não evadiram, 0,35% recebem o auxílio e 99,65% não recebem; para os evadidos, 2,63% recebem o auxílio, enquanto 97,37% não recebem. Portanto, em ambos os períodos, tem-se uma maior ocorrência proporcional de evasão entre indivíduos de domicílios em maior vulnerabilidade socioeconômica por serem comtempladas pelo PBF. A distribuição proporcional é bastante similar entre os dois períodos.

TABELA 6 - Distribuição dos indivíduos de 15 a 17 anos de idade em função da Evasão conforme variável Bolsa Família em 2019 e 2022

		Evasão			
		2019		2022	
Bolsa Família		Não	Sim	Não	Sim
	N	38.506	6.223	23.850	4.276
	CV%	12,71	27,33	16,52	37,49
Recebe	%na categoria	0,59%	4,44%	0,35%	2,63%
	N	6.534.303	133.840	6.782.469	158.012
	CV%	0,78	6,75	0,89	7,56
Não recebe	%na categoria	99,41%	95,56%	99,65%	97,37%
	N	6.572.810	140.063	6.806.320	162.288
	CV%	0,77	6,54	0,88	7,54
Total	%na categoria	100,00%	100,00%	100,00%	100,00%

Fonte: Dados extraídos pelo autor baseado nos microdados da PNAD contínua, suplemento de educação, 2019 e 2022.
Legenda: N=estimativa do total populacional, CV = coeficiente de variação

As frequências quanto à variável pensão LOAS são retratadas na TABELA 7. Para o ano de 2019, do grupo de indivíduos que não evadiram, 0,26% recebem pensão LOAS e 99,74% não recebem; para os indivíduos evadidos, 2,42% recebem e 97,58% não recebem. Para o ano de 2022, dos não evadidos, 0,29% recebem pensão LOAS e 99,71% não recebem; para os evadidos, 2,93% recebem pensão LOAS e 97,07% não recebem o auxílio. O comportamento da distribuição das proporções é similar ao observado para a contemplação ou não do PBF (TABELA 6).

TABELA 7 - Distribuição dos indivíduos de 15 a 17 anos de idade em função da Evasão conforme variável pensão LOAS em 2019 e 2022

		Evasão			
		2019		2022	
Pensão LOAS		Não	Sim	Não	Sim
	N	16.984	812	19.421	4.751
	CV%	18,15	76,82	19,33	53,86
Recebe	%na categoria	0,26%	2,42%	0,29%	2,93%
	N	6.555.825	139.250	6.786.898	157.537
	CV%	0,78	6,52	0,89	7,55
Não recebe	%na categoria	99,74%	97,58%	99,71%	97,07%
	N	6.572.810	140.063	6.806.320	162.288
	CV%	0,77	6,54	0,88	7,54
Total	%na categoria	100,00%	100,00%	100,00%	100,00%

Fonte: Dados extraídos pelo autor baseado nos microdados da PNAD contínua, suplemento de educação, 2019 e 2022.
Legenda: N=estimativa do total populacional, CV = coeficiente de variação

Para a variável pensão alimentícia, TABELA 8, a distribuição percentual no ano de 2019 do grupo de indivíduos que não evadiram indica que 4,47% recebem pensão alimentícia e 95,53% não recebem; para o grupo dos indivíduos que evadiram, 2,42% recebem a pensão alimentícia e 97,58% não recebem. No ano de 2022, das pessoas não evadidas, 4,04% recebem pensão alimentícia, enquanto 95,96% não recebem; dos evadidos, 2,77% recebem a pensão e 97,23% não recebem. Nota-se, então, maior proporcionalidade de recebimento de pensão alimentícia entre os indivíduos não evadidos. Vale destacar que esta variável não está necessariamente relacionada a uma maior ou menor vulnerabilidade social do indivíduo.

TABELA 8 - Distribuição dos indivíduos de 15 a 17 anos de idade em função da Evasão conforme variável pensão alimentícia em 2019 e 2022

		Evasão			
		2019		2022	
Pensão Alimentícia		Não	Sim	Não	Sim
	N	294125	3396	275119	4496
	CV%	5,77	42,37	6,42	36,46
Recebe	%na categoria	4,47%	2,42%	4,04%	2,77%
	N	6278685	136666	6531200	157792
	CV%	0,78	6,60	0,91	7,51
Não recebe	%na categoria	95,53%	97,58%	95,96%	97,23%
	N	6572810	140063	6806320	162288
	CV%	0,77	6,54	0,88	7,54
Total	%na categoria	100,00%	100,00%	100,00%	100,00%

Fonte: Dados extraídos pelo autor baseado nos microdados da PNAD contínua, suplemento de educação, 2019 e 2022.
Legenda: N=estimativa do total populacional, CV = coeficiente de variação

A TABELA 9 retrata a distribuição percentual para a variável bolsa de estudo. Para o ano de 2019, dos indivíduos que não evadiram, 0,61% recebem bolsa de estudo e 99,39% não recebem, ao passo que no grupo de indivíduos evadidos não houve registro de indivíduos que recebem bolsa de estudo, ou seja, 100% dos evadidos não recebiam bolsa de estudo. Em 2022, no grupo dos indivíduos que não evadiram, 0,70% recebem bolsa de estudo e 99,30% não recebem bolsa de estudo. Assim como ocorrido em 2019, não houve registro de indivíduos evadidos que recebem bolsa de estudo. Este é um resultado muito relevante que pode indicar o forte impacto do recebimento de bolsa de estudo sobre o desfecho analisado, evasão no Ensino Médio.

TABELA 9 - Distribuição dos indivíduos de 15 a 17 anos de idade em função da Evasão conforme variável bolsa de estudo em 2019 e 2022

		Evasão			
		2019		2022	
Bolsa de Estudo		Não	Sim	Não	Sim
	N	39865	0	47384	0
	CV	13,69	0	15,10	0
Recebe	%na categoria	0,61%	0,00%	0,70%	0,00%
	N	6532945	140063	6758935	162288
Não	CV	0,77	6,54	0,89	7,54
recebe	%na categoria	99,39%	100,00%	99,30%	100,00%
	N	6572810	140063	6806320	162288
	CV	0,77	6,54	0,88	7,54
Total	%na categoria	100,00%	100,00%	100,00%	100,00%

Fonte: Dados extraídos pelo autor baseado nos microdados da PNAD contínua, suplemento de educação, 2019 e 2022.
Legenda: N=estimativa do total populacional, CV = coeficiente de variação

A TABELA 10 retrata a distribuição percentual para a variável relacionada à condição de trabalho. Para o ano de 2019, no grupo de indivíduos não evadidos, 14,50% trabalham e 85,50% não trabalham; para os indivíduos evadidos, 22,60% trabalham, enquanto 77,40% não trabalham. No ano de 2022, para os não evadidos, 15,49% trabalham e 84,51% não trabalham e para os evadidos, 39,33% trabalham, enquanto 60,67% não trabalham. Estes resultados podem indicar que, independentemente do período analisado, o fato de o indivíduo despender um tempo diário em atividade laboral pode levar a uma maior chance de evadir no Ensino Médio.

TABELA 10 - Distribuição dos indivíduos de 15 a 17 anos de idade em função da Evasão conforme variável Trabalho em 2019 e 2022

		Evasão			
		2019		2022	
Trabalho		Não	Sim	Não	Sim
Trabalha	N	953.217	31.652	1.054.203	63.830
	CV%	3,17	12,34	3,02	13,65
	%na categoria	14,50%	22,60%	15,49%	39,33%
Não trabalha	N	5.619.592	108.411	5.752.116	98.458
	CV%	0,88	7,58	1,01	8,25
	%na categoria	85,50%	77,40%	84,51%	60,67%
Total	N	6.572.810	140.063	6.806.320	162.288
	CV%	0,77	6,54	0,88	7,54
	%na categoria	100,00%	100,00%	100,00%	100,00%

Fonte: Dados extraídos pelo autor baseado nos microdados da PNAD contínua, suplemento de educação, 2019 e 2022.
Legenda: N=estimativa do total populacional, CV = coeficiente de variação

Enfim, a TABELA 11 apresenta as frequências cruzadas sobre a evasão escolar dentre as grandes regiões geográficas brasileiras. Observa-se, no ano de 2019, dentre os não evadidos, que: 42,74% são na região Sudeste; 26,95% na região Nordeste; 12,81% na região Sul; 9,43% na região Norte; 8,07% na região Centro-Oeste. No grupo dos evadidos, o Nordeste apresenta 31,81%; o Sudeste 30,80%; o Sul 12,75%; o Norte 12,62%; o Centro-Oeste 12,02%. Para o ano de 2022, no grupo de indivíduos não evadidos, temos: 40,74% na região Sudeste; 28,03% na região Nordeste; 12,51% na região Sul; 9,98% na região Norte; 8,43% na região Centro-Oeste. No grupo dos evadidos, o Nordeste apresenta 32,63%; o Sudeste 29,89%; o Sul 16,78%; o Norte 13,10%; o Centro-Oeste 8,60%. Dentre as principais diferenças entre a distribuição proporcional dos dois períodos, destaca-se o crescimento do percentual de evadidos provenientes da região Sul em 2022, acompanhado de uma redução deste percentual dentre os residentes da região Centro-Oeste.

TABELA 11 - Distribuição dos indivíduos de 15 a 17 anos de idade em função da Evasão conforme variável grandes regiões geográficas – GR em 2019 e 2022

GR		Evasão			
		2019		2022	
		Não	Sim	Não	Sim
Nordeste	N	1.771.170	44.554	1.908.078	51.331
	CV%	1,65	9,18	1,79	10,09
	%na categoria	26,95%	31,81%	28,03%	32,63%
Sudeste	N	2.809.030	43.137	2.773.742	48.504
	CV%	1,71	14,33	1,80	17,62
	%na categoria	42,74%	30,80%	40,75%	29,89%
Norte	N	619.922	17.680	679.201	21.259
	CV%	2,37	14,42	2,59	15,60
	%na categoria	9,43%	12,62%	9,98%	13,10%
Centro-oeste	N	530.517	16.830	573.614	13.960
	CV%	2,54	19,55	3,05	17,94
	%na categoria	8,07%	12,02%	8,43%	8,60%
Sul	N	842.171	17.862	871.684	27.235
	CV%	2,52	16,28	2,37	14,54
	%na categoria	12,81%	12,75%	12,81%	16,78%
Total	N	6.572.810	140.063	6.806.320	162.288
	CV%	0,77	6,54	0,88	7,54
	%na categoria	100,00%	100,00%	100,00%	100,00%

Fonte: Dados extraídos pelo autor baseado nos microdados da PNAD contínua, suplemento de educação, 2019 e 2022.
Legenda: N=estimativa do total populacional, CV = coeficiente de variação

A TABELA 12 mostra a distribuição proporcional da variável Evasão segundo características sociais e econômicas. A análise difere daquelas mostradas nas TABELAS 2 a 11 pelo fato de as proporções serem calculadas com relação às categorias das variáveis explicativas e não com relação à Evasão.

Os resultados corroboram as análises prévias com base nas TABELAS 2 a 11. Em 2019, o sexo feminino apresentou maior proporção de evasão; 2,19%; enquanto o sexo masculino apresentou 1,97%. Para 2022, houve aumento da evasão em ambos os sexos, sendo que indivíduos do sexo masculino apresentaram maior evasão (2,44% *versus* 2,25%).

Para indivíduos que são chefe, houve uma evasão de 26,42% em 2019 e 16,33% em 2022. Dentro os que não chefe, esse percentual é menor, indicando que indivíduos chefes são mais propensos à evasão.

Indivíduos não brancos apresentam uma frequência de evasão maior do que os brancos. Em 2019, o percentual de evadidos não brancos foi de 2,48% e, em 2022, foi de 2,63%. Em

relação aos indivíduos residentes na área rural, a proporção de evadidos foi maior em ambos os anos em estudo. Em 2019, a taxa era de 2,09% e, em 2022, a taxa era de 4,34%.

Para os indivíduos que recebem Bolsa Família e Pensão LOAS, as frequências de evasão são maiores se comparados aos que não recebem esses benefícios financeiros governamentais. O recebimento de Bolsa Família em 2019 mostra uma evasão de 13,9% e, em 2022, de 15,20%. Quanto ao recebimento de Pensão LOAS, em 2019 o percentual de evadidos era de 4,57% e, em 2022, subiu para 19,65%. Para os indivíduos que recebem Bolsa de Estudo, o percentual de não evadidos é de 100%, indicando que indivíduos com esse suporte tendem a não evadir.

Os indivíduos que participam de algum tipo de atividade de trabalho evadem mais proporcionalmente se comparados aos que não exercem atividade laboral (3,21% *versus* 1,89% em 2019 e 5,71% *versus* 1,70% em 2022), os quais, supostamente, deveriam estar se dedicando exclusivamente aos estudos. Em 2022, o percentual de evasão entre os indivíduos que trabalham quase dobrou se comparado ao observado em 2019.

Em relação às grandes regiões político-administrativas do Brasil, em 2019, para pessoas residentes na região Centro-Oeste, a frequência de evasão foi maior em relação às demais e representa 3,07%. Em 2022, as regiões Norte e Sul apresentam maiores proporções de evadidos, ambas superiores a 3,00%.

A TABELA 13 apresenta as medidas descritivas da variável renda *per capita* para o grupo de indivíduos evadidos e não evadidos em 2019 e 2022. Esta é a única variável explicativa quantitativa dentre aquelas selecionadas para o estudo (TABELA 1). As medidas do grupo de evadidos para o ano de 2019 mostram uma renda domiciliar per capita média de R$ 506,97 e uma mediana de R$ 405,00, evidenciando que há uma assimetria positiva, sendo o desvio-padrão de 37,87 (CV=7,47%). O quartil 1 concentra 25% dos indivíduos com renda domiciliar *per capita* abaixo de R$ 201,00 e o quartil 3 concentra 25% dos indivíduos com renda domiciliar *per capita* acima de R$ 665,00.

TABELA 12 – Evasão segundo as características sociais e econômicas das pessoas de 15 a 17 anos de idade – Brasil – 2019 e 2022

		Ensino Médio				
		2019		2022		
Variável	Categorias	Não evadiu	Evadiu	Não evadiu	Evadiu	Total
Sexo	Masculino	98,03%	1,97%	97,56%	2,44%	100,00%
	Feminino	97,81%	2,19%	97,75%	2,25%	100,00%
Chefe	Chefe	73,58%	26,42%	83,67%	16,33%	100,00%
	Não Chefe	98,32%	1,68%	97,99%	2,01%	100,00%
Cor_Raça	Branco	98,49%	1,51%	98,09%	1,91%	100,00%
	Não Branco	97,52%	2,48%	97,37%	2,63%	100,00%
Urb_rur	Urbano	98,22%	1,78%	97,99%	2,01%	100,00%
	Rural	97,91%	2,09%	95,66%	4,34%	100,00%
Bolsa Família	Recebe	86,1%	13,9%	84,80%	15,20%	100,00%
	Não Recebe	98,0%	2,01%	97,72%	2,28%	100,00%
Pensão LOAS	Recebe	95,43%	4,57%	80,35%	19,65%	100,00%
	Não Recebe	97,92%	2,08%	97,73%	2,27%	100,00%
Pensão Alimentícia	Recebe	98,86%	1,61%	98,39%	1,61%	100,00%
	Não Recebe	97,87%	2,13%	97,64%	2,36%	100,00%
Bolsa de Estudo	Recebe	100,00%	0,00%	100,00%	0,00%	100,00%
	Não Recebe	97,90%	2,10%	97,66%	2,34%	100,00%
Trabalho	Trabalha	96,79%	3,21%	94,29%	5,71%	100,00%
	Não Trabalha	98,11%	1,89%	98,30%	1,70%	100,00%
GR	Nordeste	97,55%	2,45%	97,33%	2,67%	100,00%
	Sudeste	98,49%	1,51%	98,28%	1,72%	100,00%
	Norte	97,23%	2,77%	96,97%	3,03%	100,00%
	Centro-oeste	96,93%	3,07%	97,63%	2,37%	100,00%
	Sul	97,92%	2,08%	96,97%	3,03%	100,00%

Fonte: Dados extraídos pelo autor baseado nos microdados da PNAD contínua, suplemento de educação, 2019 e 2022.

Ainda em 2019, observa-se da TABELA 13 que, para os indivíduos não evadidos, a renda domiciliar *per capita* média é de R$ 940,68 e a mediana é de R$625, também demonstrando uma assimetria positiva na distribuição da renda neste grupo, sendo o desvio-padrão de 15,05 (CV=1,60%). O quartil 1 concentra 25% dos indivíduos com renda domiciliar *per capita* abaixo de R$ 334,00 e o quartil 3 concentra 25% dos indivíduos com renda domiciliar *per capita* acima de R$ 1075,00. Assim, nota-se que a distribuição da renda domiciliar *per capita* é diferente entre os evadidos e não evadidos. Os não evadidos tendem a estar em domicílios com rendas maiores, sendo a distribuição da renda domiciliar *per capita* destas pessoas mais concentrada em torno da média do que nos domicílios dos evadidos (CV de 1,60% *versus* 7,47%, respectivamente). Um resultado similar é observado para o ano de 2022.

TABELA 13 – Medidas descritivas do rendimento domiciliar *per capita* para os indivíduos de 15 a 17 anos de idade em função da Evasão para o Brasil em 2019 e 2022

Ano	Evasão	Mínimo	1° quartil	Mediana	Média	3° quartil	Máximo	Desvio-padrão	CV
2019	Evadido	0	201,00	405,00	506,97	665,00	2900,00	37,87	7,47
	Não evadido	0	334,00	625,00	940,68	1075,00	50022,00	15,05	1,60
2022	Evadido	0	302,00	563,00	714,88	933,00	11000,00	55,56	7,77
	Não evadido	0	426,00	783,00	1193,70	1333,00	49363,00	18,58	1,56

Fonte: Dados extraídos pelo autor baseado nos microdados da PNAD Contínua, suplemento de educação, 2019 e 2022

Os *box-plots* exibidos nas FIGURAS 2 e 3 mostram, respectivamente, a distribuição da renda domiciliar *per capita* em 2019 e 2022 para evadidos e não evadidos. Como a presença de valores discrepantes altos dificulta a visualização de uma possível diferença entre os grupos de indivíduos evadidos e não evadidos, para melhorar a visualização do *box-plot*, construiu-se o gráfico para todos os indivíduos da amostra (lado esquerdo) e o gráfico restringindo a renda domiciliar *per capita* em até no máximo 3 salários-mínimos brasileiros (lado direito). Para os dois períodos, a figura evidencia a análise feita com base na TABELA 13 no sentido de que a distribuição da renda domiciliar *per capita* para os indivíduos evadidos é mais concentrada em valores baixos de renda do que para os não evadidos.

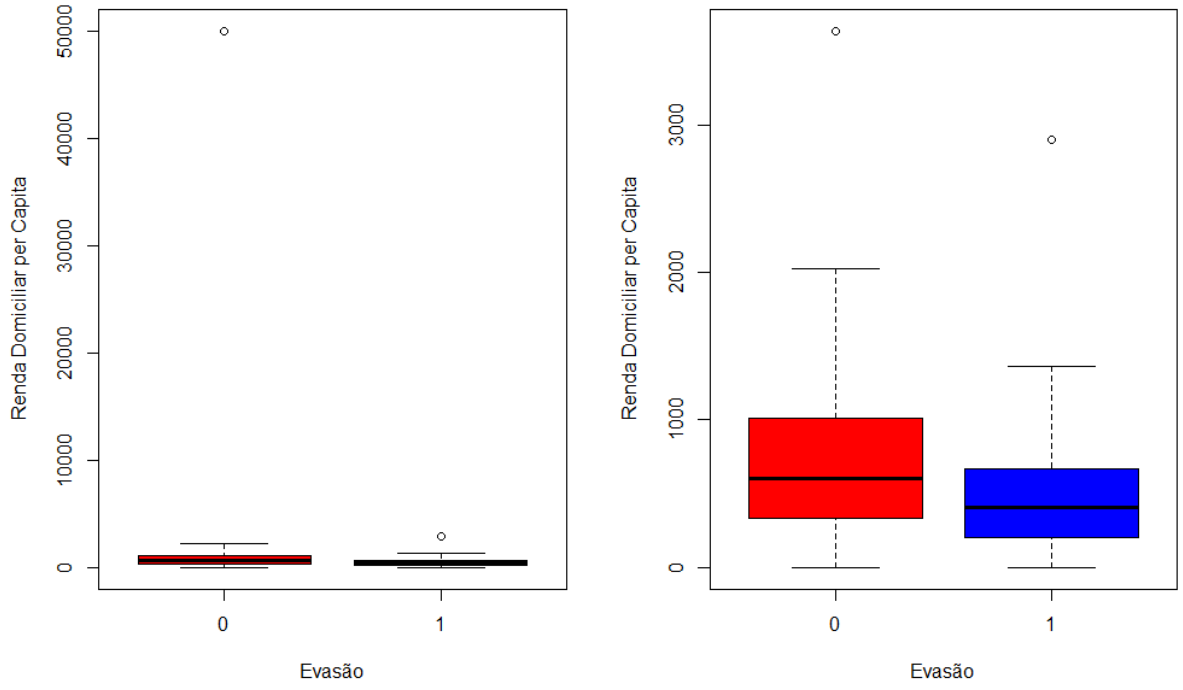


FIGURA 2 - Gráfico *box-plot* da renda domiciliar per capita para evadidos e não evadidos (0 – não evadido, 1 – evadido), 2019. Fonte: Elaboração própria segundo microdados da PNAD Contínua, suplemento de educação, 2019 e 2022.

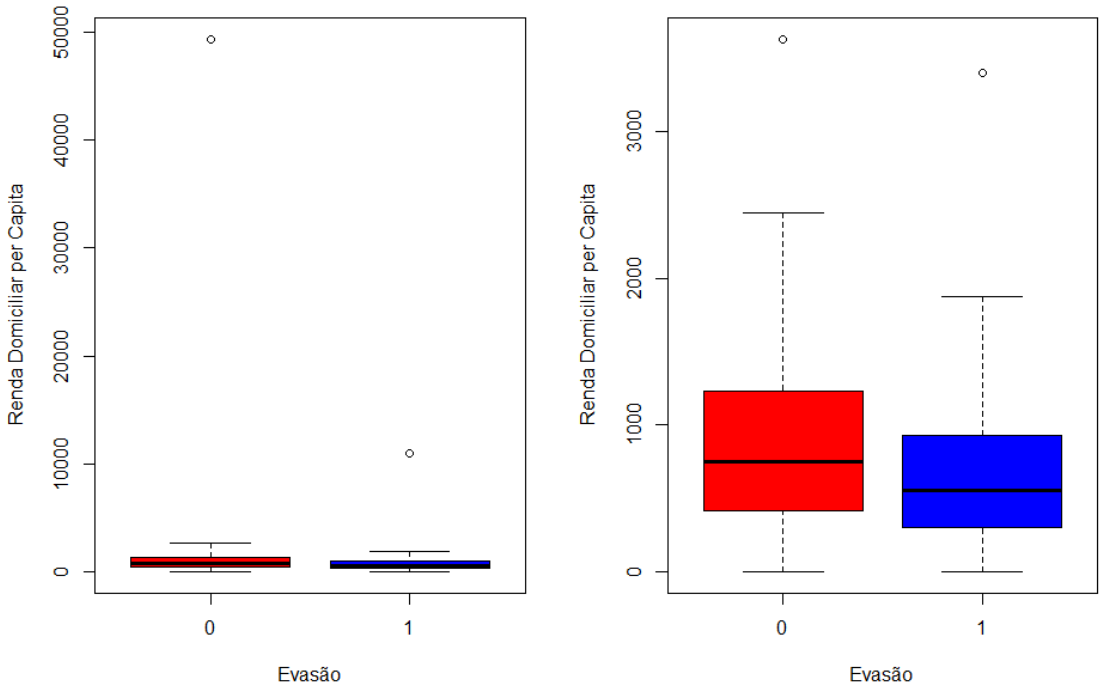


FIGURA 3 - Gráfico *box-plot* da renda domiciliar per capita para evadidos e não evadidos (0 – não evadido, 1 – evadido), 2022. Fonte: Elaboração própria segundo microdados da PNAD Contínua, suplemento de educação, 2019 e 2022.

3.2 Ajuste do Modelo de Regressão Logística

Na construção do modelo de regressão logística com base nos dados da PNADC, utilizou-se os métodos de seleção de variáveis *forward* e *backward* separadamente. Conforme descrito na Seção 2.3.1, em todos os casos, foram analisadas as 11 variáveis explicativas descritas na TABELA 1 (Sexo, Chefe, Cor_Raça, Urb_rur, Bolsa Família, Pensão LOAS, Pensão Alimentícia, Bolsa de Estudo, Trabalho, GR e Renda_per_capita). Além disso, no *forward* e no *backward* a significância dos coeficientes foi analisada com o nível de 10% para decisão da remoção/inclusão ou não das variáveis de acordo com o p-valor do teste de Wald.

3.2.1 Análise para os dados da PNAD Contínua em 2019

Para o ano de 2019, os dois métodos de seleção de variáveis resultaram no mesmo modelo final, isto é, incluíram as mesmas variáveis explicativas em sua versão final. A título de informação, a TABELA A1 do Apêndice traz os resultados dos ajustes do modelo para cada variável individualmente. Os p-valores presentes nesta tabela foram usados como critério de inclusão das variáveis no passo a passo do método *forward*. Similarmente, a TABELA A2 mostra

o resultado do modelo de regressão logística ajustado com todas as variáveis explicativas para o ano de 2019, usada como critério de exclusão dos termos no passo a passo do método *backward*.

Ambas as TABELAS A1 e A2 permitem a identificação de que o coeficiente associado à variável Bolsa de Estudo tem um valor negativo de alta magnitude numérica (abaixo de -12,00 nos dois casos), levando a uma razão de chances da ordem de $OR \approx e^{-12,0} \approx 0,00$. Este resultado indica que a chance de evasão no Ensino Médio para indivíduos que recebem alguma bolsa de estudo é milhares de vezes menor que a chance para aqueles que não recebem nenhuma bolsa de estudo. Este é um resultado esperado, já que não houve nenhuma evasão entre indivíduos que recebem alguma bolsa de estudo (TABELA 9). Logo, de certa forma, os dados indicam que a variável Bolsa de Estudo é determinística na não ocorrência de evasão. Por causa disso e tentando evitar instabilidades no processo de estimação devido ao desbalanceamento extremo dentre as categorias da variável em relação às classes da variável resposta Evasão (zero observações dentre os indivíduos não evadidos), esta variável foi removida do processo de seleção do modelo final e sua relevância para o desfecho Evasão será discutida à parte.

Os sinais dos coeficientes mostram que ser chefe, ser contemplado pelo Programa Bolsa Família e ter alguma atividade de trabalho aumentam a chance do indivíduo evadir, pois os coeficientes betas são positivos. No entanto, residir em zona urbana e ter maior renda familiar *per capita* diminuem a chance do indivíduo evadir, já que os coeficientes são negativos.

Um indivíduo chefe apresenta uma chance de evadir 17 vezes maior do que os que não são chefe. Aqueles atendidos pelo Programa Bolsa Família tem chance de evadir 2,37 vezes maior em relação a um indivíduo que não é contemplado pelo Programa. Para os que realizam algum tipo de atividade laboral, a chance de evadir será 83,4% maior do que para indivíduos que não realizam nenhum tipo de trabalho. Para os indivíduos que vivem na zona urbana, a chance de evasão diminui em 37,28% em relação aos indivíduos que residem na zona rural. Quanto a relação à renda domiciliar *per capita*, o aumento de R$100,00 reduz em 9,7% ($1 - e^{-0,00102*100}$) a chance de o indivíduo evadir.

Para a variável GR, apenas a região Centro-Oeste apresenta diferença significativa ao nível de 10% em relação à região Sudeste, a qual foi tomada como região de referência. Segundo a OR estimada, se um indivíduo reside na região Centro-Oeste, ele apresenta uma chance aproximadamente 1,8 vezes maior de evadir do que um indivíduo da região Sudeste.

TABELA 14 – Modelo final de regressão logística para a Evasão no Ensino Médio no ano de 2019, sugerido pelos métodos *forward e backward.*

Variável	Estimativa do Coeficiente	Erro Padrão	Wald	p-valor	Razão de Chance
Intercepto	-3,250	0,200	-16,279	<0,0001	0,039
Chefe	2,838	0,223	12,718	<0,0001	17,086
Trabalho	0,606	0,143	4,236	<0,0001	1,834
Renda_per_Capita	-0,001	0,000	-4,065	<0,0001	0,999
Urb_rur	-0,466	0,136	-3,427	0,001	0,627
Bolsa Família	0,863	0,419	2,061	0,041	2,370
Nordeste	-0,018	0,181	-0,099	0,921	0,982
Norte	0,157	0,206	0,764	0,446	1,170
Centro-Oeste	0,587	0,248	2,364	0,019	1,798
Sul	0,230	0,238	0,964	0,336	1,258

Fonte: Dados extraídos pelo autor baseado nos microdados da PNAD Contínua, suplemento de educação, 2019

Para a análise preditiva do modelo, foi feito um estudo com diferentes pontos de cortes para a definição das classificações entre evadidos e não evadidos. Para o modelo de regressão apresentado na TABELA 14, observou-se que o limiar de corte (*threshold*) de 0,036 maximiza a área sob a curva ROC (AUC). Os valores de Sensibilidade e Especificidade para cada ponto de corte analisado são apresentados na FIGURA A1 do Apêndice. A FIGURA 4 apresenta a curva da Sensibilidade e a curva equivalente a (1-Especificidade), a área abaixo da curva foi de 0,7767, logo a discriminação do modelo é aceitável. Para este ponto de corte, a taxa de verdadeiros positivos (True Positive Rate - TPR), conhecida como Sensibilidade, foi de 77,63% e representa a probabilidade de o modelo identificar corretamente os indivíduos que evadiram no Ensino Médio. A taxa de falso positivos (True Negative Rate - TNR), conhecida como Especificidade, foi de 77,72% e representa a probabilidade de o modelo identificar corretamente os indivíduos que não evadiram no Ensino Médio. Estes são valores de acerto consideráveis em ambos os sentidos.

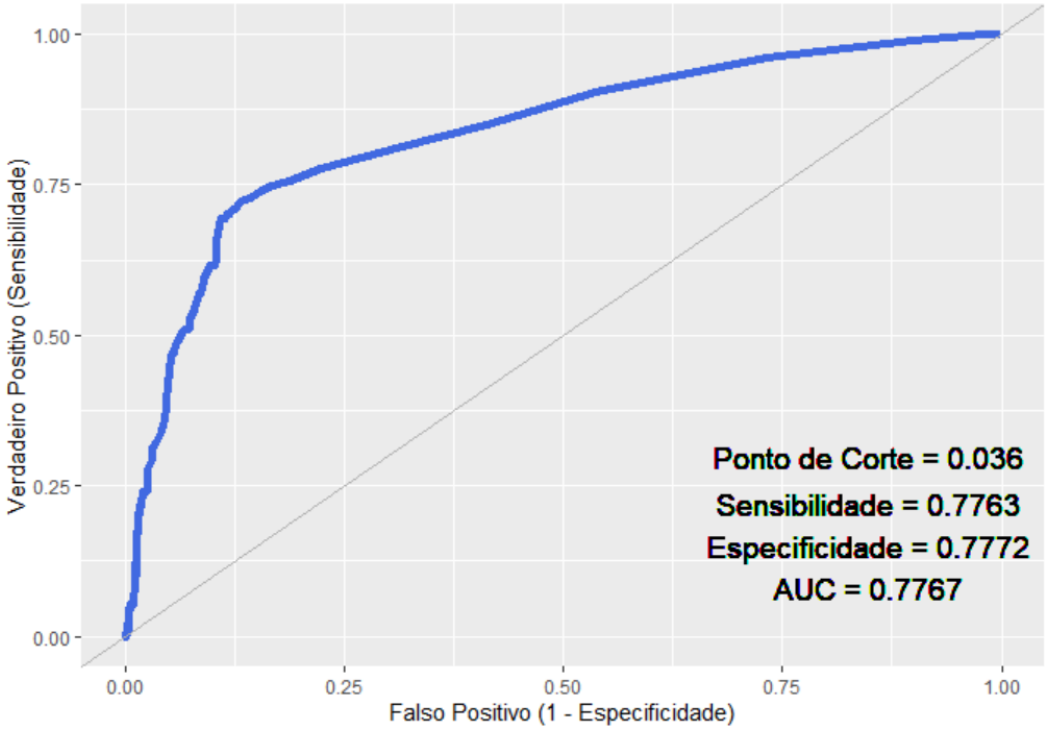


FIGURA 4 – Gráfico da curva ROC do modelo de regressão logística ajustado para dados de 2019. Fonte: Elaboração própria segundo microdados da PNAD Contínua, suplemento de educação, 2019.

3.2.2 Análise para os dados da PNAD Contínua em 2022

Para o ano de 2022, os dois métodos de seleção de variáveis também resultaram no mesmo modelo final. A TABELA A3 do Apêndice traz os resultados dos ajustes do modelo para cada variável individualmente, a qual foi usada como referência no processo de inclusão das variáveis no passo a passo do método *forward*. Similarmente, a TABELA A4 mostra o resultado do modelo de regressão logística ajustado com todas as variáveis explicativas, usada como critério de exclusão dos termos no passo a passo do método *backward*.

Assim como nos dados de 2019, o coeficiente associado à variável Bolsa de Estudo tem um valor negativo de alta magnitude numérica (abaixo de -12,00 nos dois casos) pelo fato de que não houve nenhuma evasão entre indivíduos que recebem alguma bolsa de estudo (TABELA 9). Dessa forma, esta variável foi removida do processo de seleção do modelo final e sua relevância para o desfecho Evasão será discutida à parte.

A TABELA 15 traz o modelo final para os dados de 2022. Observa-se que existe uma relação estatisticamente significativa entre a probabilidade de Evasão no Ensino Médio e as variáveis Chefe, Trabalho, Renda_per_capita, Urb_rur, Bolsa Família, Pensão LOAS e GR, nesta ordem, da maior para a menor significância estatística. A diferença com relação aos dados de 2019 foi a inclusão da variável Pensão LOAS no modelo final. Os sinais dos coeficientes mostram

que ser chefe, ser contemplado pelo Programa Bolsa Família, receber Pensão LOAS e ter alguma atividade de trabalho aumentam a chance do indivíduo evadir, pois os coeficientes são positivos. No entanto, residir em zona urbana e ter maior renda familiar *per capita* diminuem a chance do indivíduo evadir, já que os coeficientes são negativos.

Um o indivíduo chefe apresenta uma chance de evadir 7,7 vezes maior do que os que não são chefe. Aqueles atendidos pelo Programa Bolsa Família tem a chance de evadir 4,5 vezes maior em relação a um indivíduo que não é contemplado pelo Programa. Para os que realizam algum tipo de atividade laboral, a chance de evadir será 288,6% maior do que indivíduos que não realizam nenhum tipo de trabalho. Para os indivíduos que vivem na zona urbana, a chance de evasão diminui em 41,3% em relação aos indivíduos que residem na zona rural. Quanto à renda domiciliar *per capita*, o aumento de R$100,00 na renda reduz em 7,20% ($1 - e^{-0,0007469*100}$) a chance de o indivíduo evadir. Quem recebe Pensão LOAS tem chance de evadir que é mais de 15 vezes maior do que os que não recebem esse benefício governamental.

Para a variável GR, apenas a região Sul apresenta diferença significativa no patamar de 10% em relação à região Sudeste, a qual foi tomada como região de referência. Segundo a OR estimada, se um indivíduo reside na região Sul, ele apresenta uma chance aproximadamente 1,6 vezes maior de evadir do que um indivíduo da região Sudeste. Nos dados de 2019, a diferença foi observada entre as regiões Sudeste e Centro-Oeste. Essa mudança entre Centro-Oeste e Sul, de certa forma, é corroborada pelo que foi observado na análise descritiva dos dados, já que de 2019 para 2022, as regiões Centro-Oeste e Sul apresentaram as maiores modificações nas proporções de evadidos.

TABELA 15 – Modelo final de regressão logística para a Evasão no Ensino Médio no ano de 2022, sugerido pelos métodos *forward e backward*.

Variável	Estimativa do Coeficiente	Erro Padrão	Wald	p-valor	Razão de Chance
Intercepto	-3,392	0,233	-14,529	<0,0001	0,034
Chefe	2,040	0,195	10,470	<0,0001	7,687
Bolsa Família	1,514	0,526	2,876	0,004	4,544
Urb_rur	-0,533	0,160	-3,323	0,001	0,587
Renda_per_Capita	-0,001	0,000	-3,597	<0,0001	0,999
Trabalho	1,357	0,172	7,902	<0,0001	3,886
Pensão LOAS	2,738	1,315	2,082	0,039	15,462
Nordeste	0,142	0,198	0,714	0,476	1,152
Norte	0,311	0,232	1,340	0,182	1,364
Centro-Oeste	0,240	0,267	0,901	0,369	1,272
Sul	0,493	0,244	2,021	0,045	1,638

Fonte: Dados extraídos pelo autor baseado nos microdados da PNAD Contínua, suplemento de educação, 2022

Para a análise preditiva do modelo, foi feito um estudo com diferentes pontos de cortes para a definição das classificações entre evadidos e não evadidos. Para o modelo de regressão apresentado na TABELA 15, observou-se que o limiar de corte (*threshold*) de 0,056 maximiza a área sob a curva ROC (AUC). A FIGURA 5 apresenta a curva da Sensibilidade e a curva equivalente a (1-Especificidade), a área abaixo da curva foi de 0,7116, logo a discriminação do modelo é aceitável. Para este ponto de corte, a taxa de verdadeiros positivos (True Positive Rate - TPR), conhecida como Sensibilidade, foi de 71,07% e representa a probabilidade de o modelo identificar corretamente os indivíduos que evadiram no Ensino Médio. A taxa de falso positivos (True Negative Rate - TNR), conhecida como Especificidade, foi de 71,26% e representa a probabilidade de o modelo identificar corretamente os indivíduos que não evadiram no Ensino Médio. Estes são valores de acerto consideráveis em ambos os sentidos.

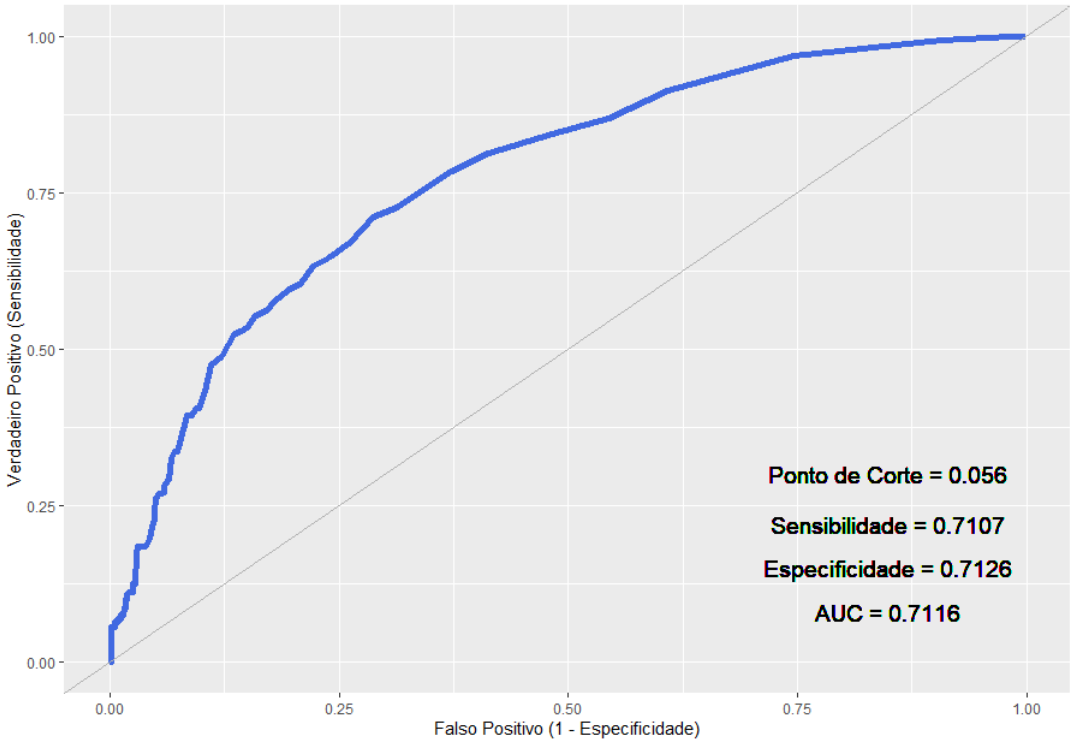


FIGURA 5 – Gráfico da curva ROC do modelo de regressão logística ajustado para dados de 2022. Fonte: Elaboração própria segundo microdados da PNAD Contínua, suplemento de educação, 2022.

Para facilitar a comparação entre os resultados para os dois períodos analisados, os resultados obtidos para os coeficientes da regressão para os anos de 2019 e 2022 podem ser vistos na TABELA 16. Destaca-se que os sinais dos coeficientes foram os mesmos para 2019 e 2022 em

todas as variáveis comuns em ambos. A variável Pensão LOAS não fez parte do modelo final em 2019. O coeficiente da variável Chefe e Urb_rur foi menor em 2022, enquanto para as variáveis Trabalho e Bolsa Família os valores dos coeficientes e, consequentemente, as razões de chance, aumentaram.

TABELA 16 – Modelo final de regressão logística para a Evasão no Ensino Médio nos anos de 2019 e 2022, sugerido pelos métodos *forward e backward.*

Variável	2019		2022	
	Estimativa do Coeficiente	Razão de Chance	Estimativa do Coeficiente	Razão de Chance
Intercepto	-3,250	0,039	-3,392	0,034
Chefe	2,838	17,086	2,040	7,687
Trabalho	0,606	1,834	1,357	3,886
Renda_per_Capita	-0,001	0,999	-0,001	0,999
Urb_rur	-0,466	0,627	-0,533	0,587
Bolsa Família	0,863	2,370	1,514	4,544
Pensão LOAS	-------	--------	2,738	15,462
Nordeste	-0,018	0,982	0,142	1,152
Norte	0,157	1,170	0,311	1,364
Centro-Oeste	0,587	1,798	0,240	1,272
Sul	0,230	1,258	0,493	1,638

Fonte: Dados extraídos pelo autor baseado nos microdados da PNAD Contínua, suplemento de educação, 2022

4 DISCUSSÃO E CONSIDERAÇÕES FINAIS

Este trabalho propôs o estudo do fenômeno da evasão escolar no Ensino Médio visando identificar possíveis fatores sociodemográficos atrelados a uma maior ou menor chance de indivíduos com idades entre 15 e 17 anos evadirem no Ensino Médio no Brasil. Para isto, aplicou-se o modelo de regressão logística aos microdados do suplemento de educação da PNAD Contínua para os anos de 2019 e 2022, um período prévio e outro posterior à pandemia da Covid-19. Vale mencionar que durante o período da pandemia, anos de 2020 e 2021, o suplemento de educação da PNADC não foi aplicado.

Nos dois períodos, 2019 e 2022, explorou-se o efeito das variáveis sexo, condição de chefe, cor/raça, zona de residência, contemplação pelo Programa Bolsa Família, recebimento de Pensão LOAS, recebimento de Pensão Alimentícia, recebimento de Bolsa de Estudo, macrorregião de residência, condição de trabalho e rendimento domiciliar *per capita* sobre o desfecho Evasão no Ensino Médio. Em ambos os períodos os modelos finais tiveram poder discriminatório aceitável, apresentando uma AUC superior a 0,70 e sendo um pouco melhor para o ano de 2019.

Na subamostra da PNADC selecionada para o estudo, dentre os indivíduos que recebem Bolsa de Estudo o percentual de não evadidos é de 100%, indicando que indivíduos com esse suporte tendem a não evadir. Este é um resultado muito relevante que pode indicar o forte impacto do recebimento de bolsa de estudo sobre o desfecho analisado, evasão no Ensino Médio. Uma maior exploração desta característica pode ser feita em outros anos da PNADC, sobretudo em anos em que se observam indivíduos com bolsa de estudo dentre os evadidos e os não evadidos. Desta forma, os modelos estatísticos poderiam estimar de forma mais robusta o real efeito desse atributo.

As variáveis relacionadas ao sexo, cor/raça e recebimento de Pensão Alimentícia não se mostraram significativas em nenhum dos períodos, embora na análise descritiva tenha-se notado uma maior ocorrência de evasão entre indivíduos do sexo masculino.

Carvalho (2016) no estudo da evasão no Ensino Superior encontra as variáveis sexo e cor/raça significativas e, em relação à renda domiciliar *per capita*, conclui que "Á medida que se aumenta a renda domiciliar *per capita*, diminui a probabilidade de evasão." Para os indivíduos chefe ou cônjuge, o autor indica que as chances de evasão aumentam.

Neste estudo, a distribuição da renda domiciliar *per capita* também se mostrou diferente entre os evadidos e não evadidos. Os não evadidos tendem a estar em domicílios com rendas maiores, sendo a distribuição da renda mais concentrada em torno da média do que nos domicílios

dos evadidos. Tanto em 2019 quanto em 2022, o modelo de regressão logística estimou um efeito protetivo para a evasão (OR=0,999 para ambos os anos).

Tanto para 2019 quanto para 2022, observou-se que a posição de chefe pode levar a uma maior ocorrência da evasão no Ensino Médio por indivíduos de 15 a 17 anos de idade (OR=17,1 para 2019 e OR=7,7 para 2022). Em ambos os períodos foi possível notar uma menor proporção de evasão em residentes de áreas urbanas do que aqueles que residem em áreas rurais (OR=0,63 para 2019 e OR=0,59 para 2022).

Considerando a região Sudeste como referência, destaca-se o crescimento do percentual de evadidos provenientes da região Centro-Oeste em 2019 e da região Sul em 2022, acompanhado de uma redução deste percentual dentre os residentes. As regiões Centro-Oeste e Sul experimentaram uma mudança de perfil de 2019 para 2022, com destaque para a região Sul cuja participação entre as evasões aumentou consideravelmente do primeiro para o último período.

Observou-se uma maior ocorrência de evasão entre indivíduos de domicílios em maior vulnerabilidade socioeconômica por serem contempladas pelo PBF (OR=2,4 para 2019 e OR=4,5 para 2022). Embora na análise descritiva, o comportamento da distribuição das amostrais quanto às proporções de recebimento de Pensão LOAS tenha sido similar ao observado para a contemplação ou não do PBF, a variável associada à Pensão LOAS só foi significativa no ano de 2022. Para estas duas variáveis, que de certa forma estão ligadas à vulnerabilidade social, estimou-se um efeito que pode ser contrário ao esperado na prática, sobretudo para o PBF. A estimativa positiva do efeito indica que a chance de evasão no Ensino Médio é maior para os indivíduos atendidos pelos programas sociais se comparados aos que não são contemplados. O recebimento do auxílio do Programa Bolsa Família está atrelado à frequência escolar das crianças e adolescentes do domicílio. Portanto, é esperado que a presença desse atributo diminuiria a chance de evasão no Ensino Médio.

É sabido que a inversão no sinal do efeito de alguma variável explicativa pode ser causado por multicolinearidade (alta correlação) entre as variáveis explicativas. No entanto, mesmo no caso do modelo contendo apenas cada uma destas variáveis individualmente o sinal do coeficiente foi positivo. Estes achados precisam ser melhores explorados e estudos são escassos na literatura. Para o contexto deste estudo, uma possível justificativa pode estar atrelado ao fato de que a Evasão foi mais frequente, conjuntamente, nas regiões Sudeste, Centro-Oeste e Sul do que nas regiões Norte e Nordeste ao passo que a cobertura do PBF é muito maior nestas do que naquelas. Uma hipótese a ser investigada é o efeito do PBF interna e separadamente em cada uma das grandes regiões, onde o efeito regional da variável poderia indicar um resultado diferente.

Por fim, os resultados indicam que, em ambos os períodos analisados, o fato de o indivíduo despender um tempo diário em atividade laboral pode levar a uma maior chance de evadir no Ensino Médio. Os indivíduos que participam de algum tipo de atividade de trabalho evadem mais proporcionalmente se comparados aos que não exercem atividade de trabalho (3,21% *versus* 1,89% em 2019 e 5,71% *versus* 1,70% em 2022), os quais, supostamente, deveriam estar se dedicando exclusivamente aos estudos.

Em 2022, o percentual de evasão entre os indivíduos que trabalham quase dobrou se comparado ao observado em 2019 ao analisar as frequências relativas nas amostras. No ajuste do modelo de regressão logística para os dados de 2019, estimou-se que a chance de evasão é 1,8 maior para os indivíduos que trabalham, para o ano de 2022 este efeito foi de 3,9 sendo o dobro do estimado em 2019.

Com base nos resultados expressos anteriormente, não é possível dizer que após a pandemia da Covid-19 o efeito do trabalho sobre a Evasão foi maior e há ausência de estudos na literatura para viabilizar uma conclusão mais precisa a esse respeito. De todo modo, os achados deste estudo evidenciam que a dedicação em atividades laborais pode influenciar no abandono dos estudos no Ensino Médio. Portanto, indivíduos provenientes de domicílios em contexto econômico mais favorecido têm maior propensão de seguir nos estudos e atingir maior nível de escolaridade.

Na construção da variável Evasão, buscou-se abranger os indivíduos de 15 a 17 anos de idade que concluíram o 9º ano do Ensino Fundamental, pois, pelo fluxo escolar usual, esses indivíduos deveriam ingressar no Ensino Médio. No entanto, observaram-se muitos valores faltantes na base de dados ao aplicar os critérios de inclusão e exclusão que definiram a população alvo do estudo, sendo necessária uma maior investigação a esse respeito.

Para trabalhos e discussões futuras, pode-se investigar o uso de outros métodos de classificação para o estudo da evasão como, por exemplo, o método de Árvore de Decisão ou *Random Forest*. No entanto, pode haver uma inviabilidade devido à limitação computacional no contexto de dados de amostras complexas. A ampliação da faixa de idade, abrangendo os indivíduos com idade até 21 anos de idade, poderia possibilitar a análise da variável idade como possível variável explicativa da evasão e incluiria indivíduos que foram excluídos do presente estudo. A seleção dos indivíduos com idade entre 15 e 17 abrangeu aqueles com idade esperada para cursar o Ensino Médio, no entanto excluiu os indivíduos com idade-série distorcida. Além disso, um estudo interessante seria a análise longitudinal de indivíduos desde a 1ª série do Ensino Fundamental à 3ª série do Ensino Médio, ou em séries sequenciais dentre deste contexto, de modo a analisar os efeitos dos fatores determinantes na conclusão da educação básica e as

probabilidades de conclusão do Ensino Médio na idade prevista de 17 anos. Para isso, seria necessário o uso de técnicas que objetivam identificar o fluxo de respostas de indivíduos ao longo de pesquisas sequenciais da PNADC, cuja viabilidade não foi investigada.

APÊNDICE

TABELA A1 – Resultado do modelo de regressão logística ajustado individualmente para cada variável explicativa para o ano de 2019, usado como critério de inclusão dos termos no passo a passo do método *forward* aplicado para seleção de variáveis.

Variável	Estimativa do Coeficiente	Erro Padrão	Wald	P-valor	Razão de Chance
Chefe	3,048	0,205	14,860	<0,0001	21,063
Bolsa Família	2,066	0,326	6,334	<0,0001	7,889
Urb_rur	-0,789	0,125	-6,292	<0,0001	0,454
Bolsa de Estudo	-12,845	2,693	-4,769	<0,0001	$2,6x10^{-6}$
Renda_per_capita	-0,001	0,000	-4,683	<0,0001	0,999
Trabalho	0,543	0,148	3,678	<0,0001	1,721
Cor_Raça	0,504	0,149	3,374	<0,0001	1,655
Pensão Alimentícia	-0,634	0,467	-1,356	0,177	0,531
Sexo	-0,105	0,142	-0,740	0,460	0,900
Pensão LOAS	0,812	4,537	0,179	0,858	2,252
Nordeste	0,4935	0,1744	2,830	0,005	1,638
Norte	0,6191	0,2048	3,022	0,003	1,857
Centro-Oeste	0,7255	0,254	2,856	0,005	2,066
Sul	0,3229	0,2204	1,465	0,145	1,381

Fonte: Dados extraídos pelo autor baseado nos microdados da PNAD Contínua, suplemento de educação, 2019.

TABELA A2 – Resultado do modelo de regressão logística ajustado com todas as variáveis explicativas para o ano de 2019, usado como critério de exclusão dos termos no passo a passo do método *backward* aplicado para seleção de variáveis.

Variável	Estimativa do Coeficiente	Erro Padrão	Wald	p-valor	Razão de Chance
Pensão LOAS	0,585	4,464	0,131	0,896	1,794
Sexo	0,176	0,147	1,194	0,234	1,192
Pensão Alimentícia	-0,715	0,502	-1,425	0,156	0,489
Cor_Raça	-0,262	0,179	-1,462	0,145	0,769
Bolsa Família	0,890	0,420	2,117	0,036	2,435
Bolsa de Estudo	-12,310	5,727	-2,149	0,033	$4,5x10^{-6}$
Urb_rur	-0,453	0,134	-3,373	0,001	0,636
Renda_per_Capita	-0,001	0,000	-3,743	<0,0001	0,999
Trabalho	0,579	0,139	4,155	<0,0001	1,783
Intercepto	-3,250	0,193	-16,802	<0,0001	0,039
Chefe	2,922	0,229	12,757	<0,0001	18,579
Nordeste	-0,018	0,181	-0,099	0,921	0,982
Norte	0,157	0,206	0,764	0,446	1,170
Centro-Oeste	0,587	0,248	2,364	0,019	1,798
Sul	0,230	0,238	0,964	0,336	1,258

Fonte: Dados extraídos pelo autor baseado nos microdados da PNAD Contínua, suplemento de educação, 2019.

TABELA A3 – Resultado do modelo de regressão logística ajustado individualmente para cada variável explicativa para o ano de 2022, usado como critério de inclusão dos termos no passo a passo do método *forward* aplicado para seleção de variáveis.

Variável	Estimativa do Coeficiente	Erro Padrão	Wald	p-valor	Razão de Chance
Chefe	2,239	0,195	11,510	<0,0001	9,387
Trabalho	1,263	0,166	7,596	<0,0001	3,537
Urb_rur	-0,802	0,139	-5,778	<0,0001	0,449
Bolsa de estudo	-13,005	2,779	-4,679	<0,0001	$2,2x10^{-6}$
Bolsa Família	2,041	0,487	4,188	<0,0001	7,696
Renda_per_Capita	-0,001	0,000	-3,893	0,000	0,999
Cor_Raça	-0,334	0,157	-2,129	0,035	0,716
Pensão LOAS	2,355	1,264	1,864	0,064	10,538
Pensão Alimentícia	-0,391	0,403	-0,971	0,333	0,676
Sexo	0,087	0,145	0,596	0,552	1,091
Nordeste	0,431	0,195	2,214	0,028	1,538
Norte	0,582	0,232	2,510	0,013	1,790
Centro-Oeste	0,331	0,256	1,292	0,198	1,392
Sul	0,580	0,228	2,551	0,012	1,787

Fonte: Dados extraídos pelo autor baseado nos microdados da PNAD Contínua, suplemento de educação, 2022

TABELA A4 – Resultado do modelo de regressão logística ajustado com todas as variáveis explicativas para o ano de 2022, usado como critério de exclusão dos termos no passo a passo do método *backward* aplicado para seleção de variáveis.

Variável	Estimativa do Coeficiente	Erro Padrão	Wald	p-valor	Razão de Chance
Cor_Raça	-0,111	0,188	-0,590	0,556	0,895
Trabalho	1,347	0,165	8,173	0,462	3,844
Sexo	0,125	0,149	0,839	0,403	1,133
Pensão Alimentícia	-0,498	0,391	-1,276	0,204	0,607
Pensão LOAS	2,725	1,316	2,071	0,040	15,261
Bolsa Família	1,527	0,540	2,828	0,005	4,603
Bolsa de Estudo	-12,452	4,139	-3,008	0,003	$3,9x10^{-6}$
Urb_rur	-0,501	0,158	-3,174	0,002	0,606
Renda_per_Capita	-0,001	0,000	-3,390	0,001	0,999
Intercepto	-3,446	0,259	-13,325	<0,0001	0,032
Chefe	2,108	0,197	10,703	<0,0001	8,232
Nordeste	0,154	0,199	0,773	0,440	1,167
Norte	0,303	0,236	1,283	0,201	1,354
Centro-Oeste	0,251	0,271	0,927	0,355	1,285
Sul	0,546	0,244	2,241	0,026	1,727

Fonte: Dados extraídos pelo autor baseado nos microdados da PNAD Contínua, suplemento de educação, 2022

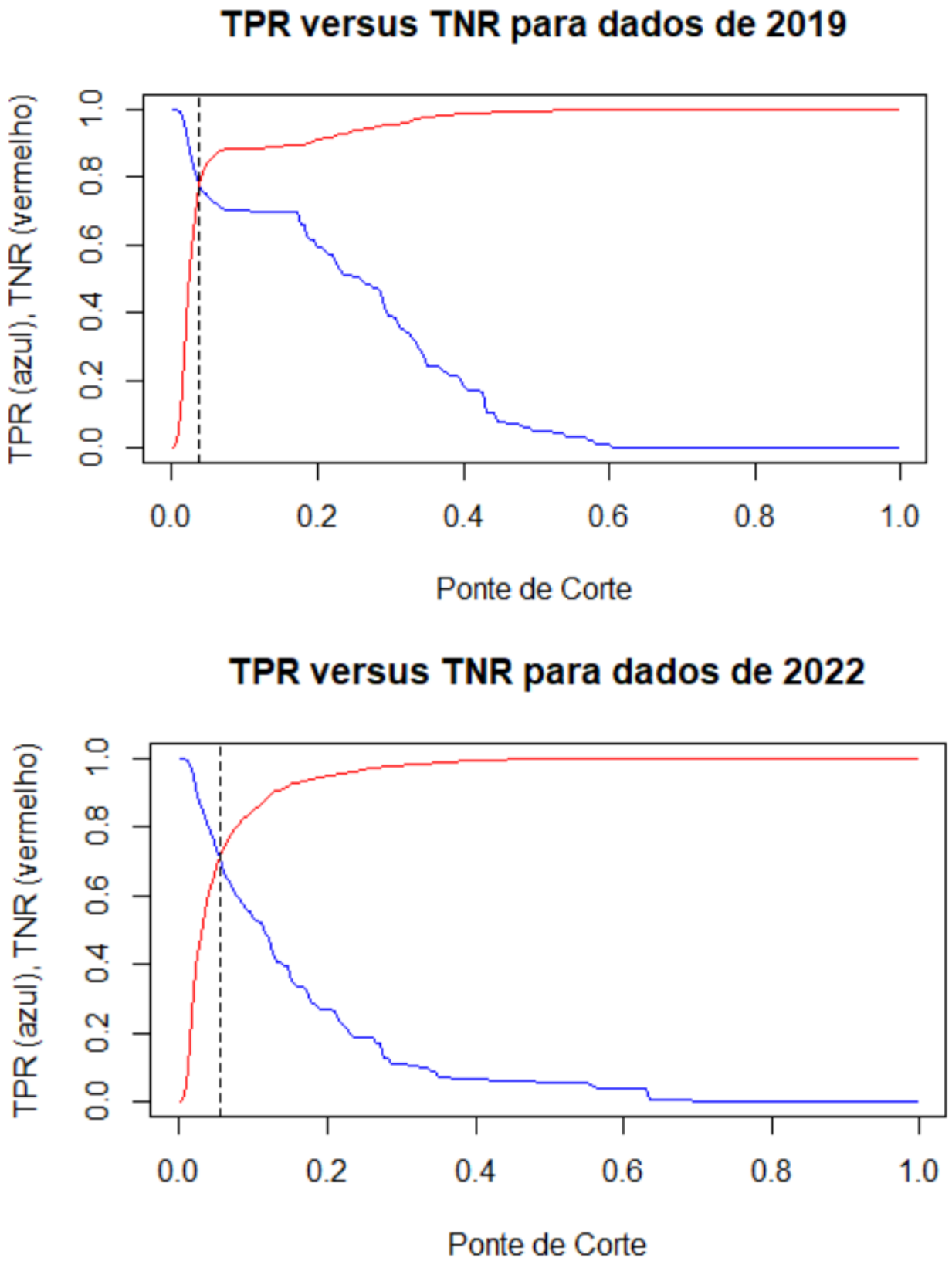


FIGURA A1 – Escolha do ponto de corte que maximiza a área sob a curva ROC (AUC) no modelo de regressão logística final para o ano de 2019 (acima) e 2022 (abaixo).
Fonte: Elaboração própria segundo microdados da PNAD Contínua, suplemento de educação, 2019 e 2022.

REFERÊNCIAS

BRASIL. [Constituição (1988)]. Constituição da República Federativa do Brasil de 1988. Brasília, DF: Presidência da República, [2016]. Disponível em: http://www.planalto.gov.br/ccivil_03/Constituicao/ Constituiçao.htm. Acesso em: 1 jan. 2017.

BRASIL. Lei de Diretrizes e Bases da Educação Nacional, LDB. 9394/1996.

BRASIL. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep). Censo Escolar da Educação Básica 2022: Resumo Técnico. Brasília, 2023.

BRAGA D, ASSUNÇÃO G (2023). _PNADcIBGE: Downloading, Reading and Analyzing PNA DC Microdata_.R package version 0.7.2, <https://CRAN.R-project.org/package=PNADcIBGE> .

BRASIL: o estado de uma nação; edição resumida / Fernando Rezende e Paulo Tafner, editores. Rio de Janeiro: IPEA, 2005. 97 p. : il.

BOFF, R. A.; CABRAL, S. M. VULNERABILIDADE SOCIOECONÔMICA: DESIGUALDA DE SOCIAL, EXCLUSÃO E POBREZA NO BRASIL. **Boletim de Conjuntura (BOCA)**, Boa Vista, v. 13, n. 38, p. 71–88, 2023. DOI: 10.5281/zenodo.7648187. Disponível em: https://revist a.ioles.com.br/boca/index.php/revista/article/view/848. Acesso em: 7 jul. 2023.

BRASIL. Ministério da Educação. Secretaria de Educação Básica. **Ensino fundamental de nove anos: passo a passo do processo de implantação**. Brasília, 2009. 27 p. Disponível em: http://portal.mec.gov.br/dmdocuments/passo_a_passo_versao_atual_16_setembro.pdf. Acesso em: 10 nov. 2022.

BRASIL. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep). Censo Escolar da Educação Básica 2022: Resumo Técnico. Brasília, 2023.

CARVALHO, M. M.; Paulo Tafner. Ensino Superior Brasileiro: a evasão dos alunos e a relação entre formação e profissão. In: XXX Encontro Anual da ANPOCS, 2006, Caxambu. XXX Encontro Anual da ANPOCS. Rio de Janeiro: anpocs, 2006. v. 30.

DORE, Rosemary. Evasão nos cursos técnicos de nível médio da rede federal de educação profissional de Minas Gerais. In: DORE, Rosemary. Evasão na educação: estudos, políticas e propostas de enfrentamento. Brasília: Ifb/Ceprotec/Rimepes, 2014. p. 379-413.

HOCKING TD (2020). _WeightedROC: Fast, Weighted ROC Curves_. R package version 2020.1.31, <https://CRAN.R-project.org/package=WeightedROC>.

FAWCETT, Tom. 2006. «An introduction to ROC analysis». Pattern Recognition Letters 27: 861–74. https://doi.org/10.1016/j.irbm.2014.09.001.

FÁVARO, Luiz Paulo Manual de análise de dados / Luiz Paulo Fávero, Patrícia Belfiore. -1. ed. - Rio de Janeiro: Elsevier, 2017.

HAIR et al. Análise multivariada de dados. 6. ed. São Paulo: Bookman, 2009, 284 – 285 p.
Hastie, T. e Tibshirani, R.; Friedman, J. The Elements of Statistical Learning. New York: Springer, 2001, 731 p.

HOSMER, D. W. e Lemeshow, S. Goodness of fit tests for the multiple logistic regression model, Communications in Statistics - Theory and Methods, v. 9, n. 10, p. 1043- 1069, 1980. DOI: 10.1080 / 03610928008827941. Disponível em: https://www.tandfonline.com/doi/abs/10.1080/03610928008827941. Acesso em: 02 jul. 2020.

HOSMER, D. W., Lemeshow, S. e Sturdivant, R. X. Applied Logistic Regression. 3. ed. Hoboken: John Wiley & Sons, 2013, 500 p

HAIR, Joseph F., William C. Black, Barry J. Babin, e Ronald L. Tatham. 2009. Análise Multivariada de Dados. 6a ed. São Paulo: Bookman.

HOSMER, David W., e Stanley Lemeschow. 2000. Applied Logistic Regression. 2 ed. New York: Wiley.

HOSMER, D. W., Lemeshow, S. e Sturdivant, R. X. Applied Logistic Regression. 3. ed. Hoboken: John Wiley & Sons, 2013, 500 p.

HOCKING, Toby Dylan. **Weighted ROC analysis**. 2020. Disponível em: https://rdrr.io/cran/WeightedROC/f/inst/doc/Definition.pdf. Acesso em: 15 jun. 2023.

HOCKING TD (2020). _WeightedROC: Fast, Weighted ROC Curves_. R package version 2020.1.31, <https://CRAN.R-project.org/package=WeightedROC>.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA - IBGE. **SBN 978-85-240-4561-5**: Pesquisa Nacional por Amostra de Domicílios Contínua. Versão 1.12 ed. Rio de Janeiro: Ibge, 2023. 124 p

IFFEL, S. M.; MALACARNE, V. Evasão escolar no ensino médio: o caso do colégio estadual S anto Agostinho no município de Palotina. Curitiba, PR: Secretaria da Educação e do Esporte, 20 10. Disponível em:< EVASÃO ESCOLAR NO ENSINO MÉDIO: O CASO DO COLÉGIO ES TADUAL SANTO AGOSTINHO NO MUNICÍPIO DE PALOTINA - PR (diaadiaeducacao.pr. gov.br)>. Acesso em: 16 JUN.2023.

JOSEPH F Hair Jr ... [et al.] ; tradução Adonai Schlup Sant'Anna. – 6. ed. – Dados eletrônicos. – Porto Alegre : Bookman, 2009.

LUMLEY T, Scott A (2017) "Fitting Regression Models to Survey Data" Statistical Science 32: 265-278

LUMLEY, T. (2020). survey: análise de amostras complexas de inquéritos.

MCCULLAH P, Nelder J (1989). *Generalized Linear Models*. Chapman & Hall/CRC, London.

McCullagh P, Nelder J (1989). Generalized Linear Models. Chapman & Hall/CRC, London.

OLIVEIRA, Lyncoln Sousa de. **Estudo sobre anomalia congênita no Brasil utilizando dados do SINASC 2017 e 2018 comparando os modelos logit, probit e complemento log-log com apoio de aprendizado de máquina.** 2021. 95 f. TCC (Graduação) - Curso de Estatística, Instituto de Matemática e Estatística, Universidade Federal Fluminense, Niterói, 2021. Disponível em: http://estatistica.uff.br/wp-content/uploads/sites/33/2021/05/tcc_20202_LyncolnSousaDeOliveira_216054055.pdf. Acesso em: 05 maio 2022.

PAIVA, C. C. V. **Previsão da Inadimplência através da Regressão Logística**. 2015. Monografia (Especialização em Estatística) - Departamento de Estatística, Universidade Federal de Minas Gerais, Belo Horizonte, 2015.

Posit team (2022). RStudio: Integrated Development Environment for R. Posit Software, PBC, B oston, MA. URL http://www.posit.co/.

R Core Team (2023). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Disponível em: https://cran.r-project.org/bin/windows/base/old/4.2.3/ Acesso: 10 setembro 2022.

T. LUMLEY (2023) "survey: analysis of complex survey samples". R package version 4.2.

T. LUMLEY (2004) Analysis of complex survey samples. Journal of Statistical Software 9(1):1-19

T. LUMLEY (2010) Complex Surveys: A Guide to Analysis Using R. John Wiley and Sons.

TROVÃO, Cassiano José Bezerra Marques. **Por dentro da PNAD contínua [recurso eletrônico]: uma introdução ao tratamento de dados usando R**. Natal, Rn: Edufrn, 2022. 339 p. Disponível em: http://repositorio.ufrn.br. Acesso em: 16 jun. 2022.

TAFNER, Paulo. Educação no Brasil: Atrasos, Conquistas e Desafios. In: TAFNER, Paulo (ed.). **Brasil: o estado de uma nação – mercado de trabalho, emprego e informalidade**. – Rio de Janeiro: Ipea, 2006. 533: Ipea, 2006. p. 533. Disponível em: https://portalantigo.ipea.gov.br/agencia/images/stories/PDFs/livros/livro_brasil_desenv_en_2006.pdf. Acesso em: 17 out. 2022.

UNICEF (Brasil) (org.). **Enfrentamento da cultura do fracasso escolar**: reprovação, abandono e distorção idade-série. Brasília: Unicef no Brasil, 2021. 65 p. Disponível em: https://www.unicef.org/brazil/media/12566/file/enfrentamento-da-cultura-do-fracasso-escolar.pdf. Acesso em: 5 jun. 2022.

UNICEF (Brasil) (org.). **REPROVAÇÃO, DISTORÇÃO IDADE-SÉRIE E ABANDONO ESCOLAR**. Brasília: Unicef no Brasil, 2019. 12 p. Disponível em: https://www.unicef.org/brazil/relatorios/reprovacao-distorcao-idade-serie-e-abandono-escolar. Acesso em: 6 jun. 2022.

WICKMAM et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686