

UNIVERSIDADE DE SÃO PAULO

Instituto de Ciências Matemáticas e de Computação

Análise dos Fatores Socioeconômicos no Desempenho do ENEM: Uma Abordagem com *Machine Learning*

Ramon de Castro Ramos

Monografia - MBA em Ciência de Dados (CeMEAI)

Ramon de Castro Ramos

Análise dos Fatores Socioeconômicos no Desempenho do ENEM: Uma Abordagem com *Machine Learning*

Monografia apresentada ao Centro de Ciências Matemáticas Aplicadas à Indústria do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Ciências de Dados.

Área de concentração: Ciências de Dados

Orientador: Prof. Dr. Adriano Kamimura Suzuki

Versão original

São Carlos
2026

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

d278a de Castro Ramos, Ramon
Análise dos Fatores Socioeconômicos no Desempenho
do ENEM: Uma Abordagem com Machine Learning / Ramon
de Castro Ramos; orientador Adriano Kamimura
Suzuki. -- São Carlos, 2026.
106 p.

Trabalho de conclusão de curso (MBA em Ciência de Dados) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2026.

1. ENEM. 2. Ciéncia de Dados. 3. Machine Learning. 4. Desigualdade Educacional. 5. Capital Cultural. I. Kamimura Suzuki, Adriano, orient. II. Título.

Ramon de Castro Ramos

Analysis of Socioeconomic Factors on ENEM Performance: A Machine Learning Approach

Monograph presented to the Centro de Ciências Matemáticas Aplicadas à Indústria do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, as part of the requirements for obtaining the title of Specialist in Data Science.

Concentration area: Data Science

Advisor: Prof. Dr. Adriano Kamimura Suzuki

Original version

**São Carlos
2026**

Folha de aprovação em conformidade
com o padrão definido
pela Unidade.

No presente modelo consta como
folhadeaprovacao.pdf

*Dedico este trabalho aos meus pais,
por todo o amor, apoio, incentivos e sacrifícios
que me impulsionaram a trilhar o caminho que trilhei.*

AGRADECIMENTOS

Agradeço aos meus pais, pelo amor incondicional e pelo incentivo constante aos estudos.

Ao meu orientador, Prof. Prof. Dr. Adriano Kamimura Suzuki, pelas correções precisas e por me ter guiado na estruturação deste trabalho.

Aos professores e colegas do MBA em Ciência de Dados do ICMC-USP, pela troca de experiências e pelo ambiente de aprendizado estimulante que contribuiu imensamente para o meu crescimento profissional.

Aos meus amigos Gêmeos, que me acompanharam nesta jornada desde o começo.

A P., por tudo.

Por fim, agradeço à comunidade de código aberto (*open source*), cujas ferramentas e bibliotecas tornaram este trabalho tecnicamente possível, democratizando o acesso a tecnologias de ponta.

“Be yourself, everyone else is already taken.”

Oscar Wilde

RESUMO

RAMOS, R. C. Análise dos Fatores Socioeconômicos no Desempenho do ENEM: Uma Abordagem com *Machine Learning*. 2026. 121 p. Monografia (MBA em Ciências de Dados) - Centro de Ciências Matemáticas Aplicadas à Indústria, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2026.

O Exame Nacional do Ensino Médio (ENEM) consolidou-se como a principal porta de entrada para o ensino superior no Brasil, transcendendo seu papel avaliativo para tornar-se um mecanismo central de mobilidade social. No entanto, o desempenho no exame é historicamente marcado por profundas disparidades associadas à origem social dos candidatos. Este trabalho tem como objetivo investigar e quantificar a influência dos fatores socioeconômicos nas notas do ENEM, utilizando técnicas avançadas de Ciência de Dados e *Machine Learning*. Adotando a metodologia CRISP-DM, foram processados e integrados microdados das edições de 2020 a 2023. Foram treinados e avaliados modelos de regressão baseados em árvores de decisão (*Random Forest*, *XGBoost* e *LightGBM*), culminando na construção de um modelo de *ensemble* que apresentou desempenho superior, com erro percentual (MAPE) próximo a 10% para áreas como Linguagem e Código. A aplicação de técnicas de interpretabilidade, incluindo *Permutation Importance* e curvas de sensibilidade, revelou que a Renda Familiar, a Escolaridade da Mãe e a Quantidade de Computadores são os preditores mais determinantes para o desempenho acadêmico, superando variáveis demográficas isoladas. Os resultados corroboram estatisticamente a teoria do Capital Cultural de Pierre Bourdieu e destacam a exclusão digital como uma barreira contemporânea crítica para o acesso ao ensino superior. Conclui-se que o perfil socioeconômico é um preditor robusto do sucesso escolar no Brasil, evidenciando que o ENEM, embora padronizado, reflete e reproduz as desigualdades estruturais da sociedade.

Palavras-chave: Palavras-chave: ENEM. Ciência de Dados. Machine Learning. Desigualdade Educacional. Capital Cultural.

ABSTRACT

RAMOS, R. C. **Analysis of Socioeconomic Factors on ENEM Performance: A Machine Learning Approach.** 2026. 121 p. Monograph (MBA in Data Sciences) - Centro de Ciências Matemáticas Aplicadas à Indústria, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2026.

The *Exame Nacional do Ensino Médio* - ENEM (National High School Exam) has established itself as the primary gateway to higher education in Brazil, transcending its evaluative role to become a central mechanism of social mobility. However, performance in the exam is historically marked by deep disparities associated with the candidates' social background. This work aims to investigate and quantify the influence of socioeconomic factors on ENEM scores using advanced Data Science and Machine Learning techniques. Adopting the CRISP-DM methodology, microdata from the 2020 to 2023 editions were processed and integrated. Decision tree-based regression models (*Random Forest*, *XG-Boost*, and *LightGBM*) were trained and evaluated, culminating in the construction of an *ensemble* model that achieved superior performance, with a Mean Absolute Percentage Error (MAPE) close to 10% for areas such as Languages and Codes. The application of interpretability techniques, including *Permutation Importance* and sensitivity curves, revealed that Family Income, Mother's Education, and Number of Computers are the most determinant predictors of academic performance, surpassing isolated demographic variables. The results statistically corroborate Pierre Bourdieu's theory of Cultural Capital and highlight the digital divide as a critical contemporary barrier to accessing higher education. It is concluded that the socioeconomic profile is a robust predictor of educational success in Brazil, evidencing that the ENEM, although standardized, reflects and reproduces society's structural inequalities.

Keywords: ENEM. Data Science. Machine Learning. Educational Inequality. Cultural Capital.

LISTA DE FIGURAS

Figura 1 – Exemplo de uma Regressão Linear simples com dados fictícios	36
Figura 2 – Exemplo de uma Árvore de Decisão com o <i>dataset Iris</i>	37
Figura 3 – Exemplo de uma <i>Random Forest</i> com o <i>dataset Iris</i>	38
Figura 4 – Modelo CRISP-DM	39
Figura 5 – Histograma das notas - Humanas	52
Figura 6 – Histograma das notas - Natureza	53
Figura 7 – Histograma das notas - Linguagem	53
Figura 8 – Histograma das notas - Matemática	54
Figura 9 – Histograma das notas - Redação	54
Figura 10 – Boxplot das notas por edição - Humanas	56
Figura 11 – Boxplot das notas - Natureza	57
Figura 12 – Boxplot das notas - Linguagem	57
Figura 13 – Boxplot das notas - Matemática	57
Figura 14 – Boxplot das notas - Redação	58
Figura 15 – Dez maiores <i>Permutation Importance</i> - Humanas	63
Figura 16 – Dez maiores <i>Permutation Importance</i> - Natureza	63
Figura 17 – Dez maiores <i>Permutation Importance</i> - Linguagem	64
Figura 18 – Dez maiores <i>Permutation Importance</i> - Matemática	64
Figura 19 – Dez maiores <i>Permutation Importance</i> - Redação	64
Figura 20 – Erro do <i>Grid Search - XGBoost</i>	69
Figura 21 – Erro do <i>Grid Search - LightGBM</i>	69
Figura 22 – Erro do <i>Grid Search - Random Forest</i>	70
Figura 23 – Erro MAPE - <i>XGBoost</i>	72
Figura 24 – Erro MAPE - <i>LightGBM</i>	72
Figura 25 – Erro MAPE - <i>Ensemble (XGBoost + LightGBM)</i>	74
Figura 26 – Erro MAPE - <i>Ensemble (XGBoost + LightGBM + Random Forest)</i> . .	74
Figura 27 – Rank de Importância - Humanas	77
Figura 28 – Rank de Importância - Natureza	77
Figura 29 – Rank de Importância - Linguagem	78
Figura 30 – Rank de Importância - Matemática	78
Figura 31 – Rank de Importância - Redação	78
Figura 32 – Curva de Sensibilidade - Faixa Etária	80
Figura 33 – Curva de Sensibilidade - Sexo	80
Figura 34 – Curva de Sensibilidade - Cor/Raça	81
Figura 35 – Curva de Sensibilidade - Escolaridade do Pai	82
Figura 36 – Curva de Sensibilidade - Escolaridade da Mãe	83

Figura 37 – Curva de Sensibilidade - Ocupação do Pai	84
Figura 38 – Curva de Sensibilidade - Ocupação da Mãe	85
Figura 39 – Curva de Sensibilidade - Renda Familiar	86

LISTA DE TABELAS

Tabela 1 – Variáveis socioeconômicas e suas referências	45
Tabela 2 – Observações e variáveis por edição do ENEM	48
Tabela 3 – Percentual de valores nulos por variável	49
Tabela 4 – Observações e variáveis por conjunto de dados	51
Tabela 5 – Estatísticas descritivas por conjunto de dados	52
Tabela 6 – Assimetria, Curtose e Notas zeradas	55
Tabela 7 – Teste ANOVA das médias das notas por edição	56
Tabela 8 – Quantidade e percentual de outliers nas notas	58
Tabela 9 – Cinco maiores concentrações - Humanas	59
Tabela 10 – Cinco maiores concentrações - Natureza	59
Tabela 11 – Cinco maiores concentrações - Linguagem	60
Tabela 12 – Cinco maiores concentrações - Matemática	60
Tabela 13 – Cinco maiores concentrações - Redação	60
Tabela 14 – Cinco maiores correlações <i>Phik</i> - Humanas	61
Tabela 15 – Cinco maiores correlações <i>Phik</i> - Natureza	61
Tabela 16 – Cinco maiores correlações <i>Phik</i> - Linguagem	61
Tabela 17 – Cinco maiores correlações <i>Phik</i> - Matemática	62
Tabela 18 – Cinco maiores correlações <i>Phik</i> - Redação	62
Tabela 19 – Cinco maiores correlações <i>Phik</i> - Redação	65
Tabela 20 – <i>Grid Search - XGBoost</i>	67
Tabela 21 – <i>Grid Search - LightGBM</i>	67
Tabela 22 – <i>Grid Search - Random Forest</i>	68
Tabela 23 – Hiperparâmetros Ajustados - <i>XGBoost</i>	71
Tabela 24 – Hiperparâmetros Ajustados - <i>LightGBM</i>	71
Tabela 25 – Hiperparâmetros Ajustados - <i>Random Forest</i>	71
Tabela 26 – Erro MAPE - <i>Random Forest</i>	73
Tabela 27 – Cinco melhores modelos - Humanas	75
Tabela 28 – Cinco melhores modelos - Natureza	75
Tabela 29 – Cinco melhores modelos - Linguagem	75
Tabela 30 – Cinco melhores modelos - Matemática	76
Tabela 31 – Cinco melhores modelos - Redação	76

LISTA DE ABREVIATURAS E SIGLAS

AdaBoost	<i>Adaptive Boosting</i>
ANOVA	<i>Analysis of Variance</i> - Análise de Variância
COVID-19	<i>Coronavirus Disease 2019</i> - Doença do Coronavírus 2019
CPU	<i>Central Processing Unit</i> - Unidade Central de Processamento
CRISP-DM	<i>Cross-Industry Standard Process for Data Mining</i> - Processo de Mineração de Dados Padrão entre Indústrias
CSV	<i>Comma-Separated Values</i> - Valores Separados por Vírgula
CUDA	<i>Compute Unified Device Architecture</i> - Arquitetura Unificada de Dispositivos de Computação
ENEM	Exame Nacional do Ensino Médio
Fies	Fundo de Financiamento Estudantil
GPU	<i>Graphics Processing Unit</i> - Unidade de Processamento Gráfico
IBGE	Instituto Brasileiro de Geografia e Estatística
IDH	Índice de Desenvolvimento Humano
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
IQR	<i>Interquartile Range</i> - Intervalo Interquartil
LGPD	Lei Geral de Proteção de Dados Pessoais
LightGBM	<i>Light Gradient Boosting Machine</i>
LTS	<i>Long Term Support</i> - Suporte de Longo Prazo
MAPE	<i>Mean Absolute Percentage Error</i> - Erro Percentual Absoluto Médio
MSE	<i>Mean Squared Error</i> - Erro Quadrático Médio
PIB	Produto Interno Bruto
PLN	Processamento de Linguagem Natural
ProUni	Programa Universidade Para Todos

RAM	<i>Random Access Memory</i> - Memória de Acesso Aleatório
RMSE	<i>Root Mean Squared Error</i> - Raiz do Erro Quadrático Médio
RF	<i>Random Forest</i> - Floresta Aleatória
SISU	Sistema de Seleção Unificada
SMD	<i>Standardized Mean Difference</i> - Diferença de Médias Padronizada
UF	Unidade da Federação
UFABC	Universidade Federal do ABC
XGBoost	<i>Extreme Gradient Boosting</i>
XLSX	Extensão de arquivo de planilha do Microsoft Excel
YAML	<i>YAML Ain't Markup Language</i> - Linguagem de Serialização de Dados

LISTA DE SÍMBOLOS

α	Letra Grega alfa minúscula; nível de significância estatística.
β	Letra Grega beta minúscula; coeficiente de regressão (inclinação da reta/hiperplano).
β_0	Intercepto da regressão linear.
ϵ	Letra Grega epsilon minúscula; termo de erro aleatório na regressão.
μ	Letra Grega mu minúscula; média.
σ	Letra Grega sigma minúscula; desvio padrão.
\leq	Menor ou igual
Md	Mediana
F	Estatística de teste da Análise de Variância (ANOVA)
H_0	Hipótese Nula
n	Tamanho da amostra ou número de observações
p -valor	Probabilidade de significância (valor p)
$Q1$	Primeiro Quartil (25%)
$Q3$	Terceiro Quartil (75%)
R^2	Coeficiente de determinação

SUMÁRIO

1	INTRODUÇÃO	29
2	FUNDAMENTAÇÃO TEÓRICA	31
2.1	O ENEM no Cenário Educacional Brasileiro	31
2.2	Teorias sobre Desigualdades Educacionais: O Capital Cultural de Bourdieu	32
2.3	Fatores Socioeconômicos e Desempenho no ENEM	32
2.4	Características escolares e o “Efeito Escola”	33
2.5	Disparidades Regionais e a Participação no ENEM	34
2.6	Aplicações de Ciência de Dados na Análise do ENEM e resultados obtidos	34
2.7	Métodos de <i>Machine Learning</i>	35
2.7.1	Regressão Linear	35
2.7.2	Árvore de Decisão	36
2.7.3	<i>Random Forest</i>	37
2.7.4	<i>Boosting</i>	38
3	METODOLOGIA	39
3.1	Entendimento de Negócio	39
3.2	Entendimento dos dados	40
3.3	Preparação dos dados	40
3.4	Modelagem	41
3.4.1	Análise Exploratória dos Dados	41
3.5	Treinamento dos Modelos	42
3.6	Avaliação dos Modelos	43
3.7	Influência das Variáveis Preditoras	43
4	RESULTADOS	45
4.1	Entendimento de Negócio	45
4.2	Entendimento dos dados	46
4.2.1	Escolha e Coleta dos Dados	46
4.2.2	Compreensão Inicial dos Dados	46
4.2.2.1	Edição de 2024 do ENEM e LGPD	46
4.2.3	Análise dos Dicionários de Dados	47
4.2.4	Definição da Variável Resposta	47
4.3	Preparação dos dados	47

4.3.1	Preparação do Ambiente Tecnológico e Analítico	47
4.3.2	Leitura dos Dados	48
4.3.3	Integração dos Dados	48
4.3.4	Tratamento de Valores Nulos	49
4.3.5	Separação dos Conjuntos de Dados por Variável Resposta	51
4.4	Modelagem	51
4.4.1	Análise Exploratória dos Dados - Variáveis Resposta	51
4.4.1.1	Distribuições	51
4.4.1.2	Teste de Hipótese	55
4.4.1.3	Análise de Outliers	56
4.4.2	Análise Exploratória - Variáveis Preditoras	59
4.4.2.1	Concentração	59
4.4.2.2	Correlação	60
4.4.2.3	<i>Permutation Importance</i>	63
4.4.2.4	Seleção de Variáveis	65
4.5	Treinamento dos Modelos	66
4.5.1	Ajuste dos Hiperparâmetros	66
4.5.2	Treinamento final dos modelos	71
4.5.3	Construção dos modelos de <i>ensemble</i>	73
4.5.4	Avaliação dos modelos	74
4.6	Influência das Variáveis Preditoras	77
4.6.1	Importância	77
4.6.2	Sensibilidade das Variáveis Respostas	79
4.7	Discussão dos Resultados	86
4.7.1	O Capital Cultural e a Reprodução de Desigualdades	86
4.7.2	A Renda e o Acesso a Recursos	87
4.7.3	Fatores Demográficos	87
4.7.4	Desempenho dos Modelos de Machine Learning	88
5	CONCLUSÃO	89
5.1	Síntese dos Resultados	89
5.2	Limitações do Estudo	89
5.3	Trabalhos Futuros	90
	REFERÊNCIAS	91

APÊNDICES	95
APÊNDICE A – DICIONÁRIO DE DADOS DOS MICRODADOS DO ENEM	97
APÊNDICE B – DICIONÁRIO DE DADOS DO CENSO ESCOLAR	105
APÊNDICE C – CONFIGURAÇÃO DO AMBIENTE VIRTUAL	115
APÊNDICE D – LINK PARA O GITHUB	121

1 INTRODUÇÃO

O Exame Nacional do Ensino Médio (ENEM) consolidou-se, na última década, como a principal avaliação educacional do Ensino Médio no Brasil, transcendendo seu papel inicial de termômetro da qualidade da educação básica para se tornar a porta de entrada para o ensino superior em instituições públicas e privadas, através de programas como o Sistema de Seleção Unificada (SISU), o Programa Universidade Para Todos (ProUni) e o Fundo de Financiamento Estudantil (Fies). Sua relevância reside na capacidade de fornecer um panorama detalhado do desempenho dos estudantes, bem como de aspectos socioeconômicos e contextuais que permeiam o ambiente escolar e familiar dos participantes.

Apesar dos esforços contínuos para aprimorar a qualidade da educação no Brasil, persistem desafios significativos, evidenciados pelas variações no desempenho dos estudantes em avaliações de larga escala como o ENEM. A literatura acadêmica aponta para a influência de múltiplos fatores nesse desempenho, que vão desde as condições socioeconômicas das famílias até as características estruturais e pedagógicas das escolas, além das peculiaridades regionais (1). A análise estatística de microdados do ENEM entre 2021 e 2023, por exemplo, revela desigualdades estruturais marcantes entre estudantes de escolas públicas e privadas (2). A persistência dessas disparidades indica que as desigualdades educacionais no Brasil não são meramente aleatórias, mas profundamente associadas às desigualdades sociais (3).

A análise aprofundada dos microdados do ENEM, portanto, constitui uma oportunidade ímpar para desvendar a complexa interação entre os fatores socioeconômicos, as características do ambiente escolar e as peculiaridades regionais que moldam o desempenho dos estudantes. Isso permite ir além da simples constatação das disparidades, oferecendo um panorama mais claro de como um instrumento concebido para democratizar o acesso ao ensino superior pode, na prática, atuar como um espelho das desigualdades sociais estruturais e, em certos contextos, até mesmo contribuir para a sua perpetuação, um fenômeno consistentemente observado em análises de dados históricos (2). A compreensão desses mecanismos é vital para a formulação de políticas públicas que não apenas mitiguem as lacunas, mas que atuem nas causas-raiz das iniquidades educacionais.

Nesse contexto, este Trabalho de Conclusão de Curso propõe investigar e quantificar a influência dos principais fatores socioeconômicos no desempenho dos estudantes no ENEM. A pergunta central que guia esta pesquisa é: “Quais são os principais fatores socioeconômicos que influenciam o desempenho dos estudantes no ENEM e qual a magnitude da influência de cada um desses conjuntos de fatores nas notas dos participantes?”. O objetivo geral é utilizar os microdados do exame para fornecer *insights* robustos sobre a qualidade da educação básica no Brasil, contribuindo para a identificação de áreas que necessitam de

maior atenção e investimento. A quantificação da influência dos fatores, por meio de modelos preditivos e análise de importância de variáveis (2), é um diferencial crucial. Não se trata apenas de identificar a existência de correlações, mas de medir o grau de impacto, o que é fundamental para a formulação de políticas públicas eficazes e direcionadas.

Para tanto, foram estabelecidos os seguintes objetivos específicos: i) Coletar, pré-processar e realizar uma análise exploratória dos microdados do ENEM (4) selecionando as variáveis relevantes; ii) Identificar padrões, tendências e correlações entre as variáveis selecionadas e o desempenho dos estudantes; iii) Aplicar técnicas de Ciência de Dados para construir modelos preditivos e determinar a importância relativa de cada grupo de fatores; e iv) Discutir os resultados obtidos, correlacionando-os com a literatura existente e extraiendo dados práticos.

A relevância desta pesquisa reside na sua capacidade de oferecer uma análise quantitativa detalhada das correlações entre múltiplos fatores e o desempenho educacional, utilizando uma vasta base de dados. Os dados gerados podem servir como subsídio para educadores, formuladores de políticas públicas e pesquisadores, auxiliando na compreensão das raízes das desigualdades educacionais e na elaboração de estratégias direcionadas para a melhoria do ensino médio no país. A pesquisa não se limita a um exercício acadêmico; ela tem um potencial transformador social ao fornecer dados concretos para subsidiar políticas públicas mais justas e fortalecer a rede pública de ensino (2).

Os próximos capítulos apresentam a metodologia adotada neste trabalho, os resultados obtidos e a respectiva discussão, culminando nas conclusões e recomendações para trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo estabelece o contexto teórico e empírico para o estudo, fundamentando a análise no conhecimento acadêmico existente.

2.1 O ENEM no Cenário Educacional Brasileiro

O Exame Nacional do Ensino Médio (ENEM) teve sua primeira edição em 1998, contando com a participação de aproximadamente 115 mil participantes. Na época, suas notas só eram utilizadas por 2 instituições de ensino superior, número que saltou para 93 instituições no ano seguinte. A importância do ENEM cresceu com o passar dos anos, alcançando a marca de mais de 1 milhão de participantes na sua quarta edição e tornando-se uma das principais formas de acesso ao ensino superior, com a criação do Programa Universidade Para Todos (ProUni) em 2005 (5). Em 2009, com a criação do Sistema de Seleção Unificada (SISU), o ENEM foi reformulado e assumiu o formato que tem hoje: 180 questões objetivas divididas em 4 áreas do conhecimento ((i) Ciências Humanas, (ii) Ciências da Natureza, (iii) Linguagem e Código e (iv) Matemática) e uma Redação. No ano seguinte, os resultados do ENEM passaram a ser adotados pelo Fundo de Financiamento Estudantil (Fies) e, em 2013, quase todas as instituições federais adotaram o ENEM como critério de seleção. Duas universidades portuguesas, a Universidade de Coimbra e a Universidade do Algarve, passaram a usar o ENEM como critério de seleção em 2014, número que chegou a 35 instituições portuguesas em 2018 (5).

É evidente que o ENEM deixou de ser apenas uma ferramenta de avaliação e transformou-se em um instrumento multifacetado que desempenha um papel central na trajetória educacional dos jovens brasileiros. Além de aferir o desempenho dos estudantes ao final do ensino médio, o ENEM serve como a principal porta de acesso ao ensino superior, sendo a base para o SISU, o ProUni e o Fies (6). Essa centralidade significa que qualquer fator que influencie o desempenho no exame tem um impacto direto e significativo nas oportunidades de acesso ao ensino superior e, consequentemente, na mobilidade social dos indivíduos.

Os microdados do ENEM, disponibilizados anualmente pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), representam uma fonte de informação rica e valiosa para pesquisas educacionais (2). Esses dados detalhados permitem uma compreensão aprofundada dos padrões de desempenho, das características socioeconômicas dos participantes e dos contextos escolares, possibilitando análises complexas sobre as desigualdades educacionais no país.

2.2 Teorias sobre Desigualdades Educacionais: O Capital Cultural de Bourdieu

Para compreender a reprodução das desigualdades sociais no sistema educacional, a teoria do capital cultural de Pierre Bourdieu oferece um arcabouço teórico fundamental. Este argumenta que o sucesso escolar não depende apenas do mérito individual ou da capacidade cognitiva, mas também da posse de diferentes formas de capital: o econômico (posses que o indivíduo tem), o social (relacionamentos que podem ser benéficos ao indivíduo), o simbólico (prestígio/honra) e o cultural (conhecimentos reconhecidos por diplomas e títulos) (7).

O capital cultural ainda se divide em três estados: (i) o capital cultural incorporado, composto por elementos pessoais como gostos (musicais, artísticos etc.) e domínio de línguas; (ii) o capital cultural objetivado, composto por posses de livros e obras de arte ou acesso a museus, cinema etc.; (iii) o capital cultural institucionalizado, caracterizado por diplomas e títulos de conhecimento (7).

O acúmulo de capital cultural é o que influenciará o desempenho escolar do indivíduo e futuramente seu posicionamento no mercado de trabalho. Se os dados do ENEM confirmarem a forte influência de variáveis socioeconômicas e de escolaridade parental, isso reforçará a tese da reprodução escolar das desigualdades, sugerindo que o sistema educacional, em vez de ser um equalizador, pode perpetuar as hierarquias sociais. Isso se manifesta, por exemplo, na forma como a escolaridade da mãe e a renda familiar são fatores relevantes para o desempenho e a dispersão das notas do ENEM (1). Oliveira e Cruz (2014) argumentam que a escola, ao reconhecer os alunos mais inteligentes ou aplicados, na verdade está selecionando os alunos com o capital cultural mais diverso e amplo, o que propaga a desigualdade social ao criar os “mitos de aluno inteligente-brilhante / aluno fracassado-invisível”, fazendo com que “o próprio oprimido passe a acreditar que não é capaz de ter sucesso por características pessoais e não do sistema.”

2.3 Fatores Socioeconômicos e Desempenho no ENEM

A literatura é vasta ao associar variáveis socioeconômicas ao desempenho em avaliações de larga escala e o ENEM não é exceção. As persistentes e quantificáveis desigualdades de desempenho ligadas a fatores socioeconômicos (1) indicam que o acesso a “experiências educacionais muito mais ricas” (8) fora do ambiente escolar formal é um preditor poderoso do sucesso no ENEM. Isso sugere que a escola, por si só, pode não ser capaz de compensar totalmente essas desvantagens de origem e que o campo educacional não é nivelado desde o início. Estudos sobre o ENEM consistentemente apontam o impacto de diversos fatores:

- **Renda Familiar:** Uma correlação positiva e significativa é observada entre a renda familiar e as notas do ENEM (1). Análises indicam que a diferença na nota de

Redação pode ser de até 40% entre os grupos de menor e maior renda (8).

- **Raça/Cor:** O desempenho de alunos brancos consistentemente supera o de outros grupos raciais, mesmo quando outras variáveis são controladas (1). Em média, o desempenho de alunos brancos superou o dos demais em menos de 10 pontos nas quatro provas em 2018, controlando outras variáveis (9).
- **Escolaridade dos Pais/Nível Instrucional da Mãe:** Este é um fator relevante para o desempenho e a dispersão das notas dos estudantes (1). MÃes com escolaridade a partir do ensino médio e famílias de renda alta têm um impacto positivo no desempenho (10).
- **Sexo:** Diferenças de desempenho por sexo são notadas, especialmente na prova de Matemática, com vantagem para os homens (até 36 pontos a mais) (10).
- **Idade/Atraso Escolar:** O atraso escolar associa-se negativamente ao desempenho. Alunos com pelo menos um ano de atraso escolar tiveram, em média, de 16,7 a 29,0 pontos a menos nas provas (9).

2.4 Características escolares e o “Efeito Escola”

As características das escolas também exercem influência no desempenho dos estudantes e o conceito de “Efeito Escola” busca mensurar a contribuição da instituição de ensino para o desempenho do aluno, além dos fatores individuais e familiares (10). Achados relevantes incluem:

- **Dependência Administrativa (Pública vs. Privada):** Alunos de escolas privadas consistentemente superam os de escolas públicas (10). Em Matemática, a diferença pode ser de aproximadamente 83,9 pontos entre alunos de escolas privadas e estaduais (9). Um estudo da UFABC, por exemplo, mostrou que em Matemática, apenas 2,9% dos estudantes da rede pública atingiram 720 pontos, contra 20% da rede privada (2).
- **Atributos Escolares:** Fatores como complexidade de gestão, média de horas-aula, número de alunos por turma, qualidade dos professores (esforço e adequação docente) e o nível socioeconômico médio da escola são importantes (10). O nível socioeconômico médio da escola e a regularidade docente destacam-se como os mais significativos, aumentando a nota em 22,7 pontos para cada nível socioeconômico e em 14,6 para cada nível de regularidade docente em escolas privadas (10).

Embora o “Efeito Escola” seja um fator, a literatura sugere que uma grande parte da explicação das notas do ENEM reside em fatores externos ao controle escolar

(10). Isso significa que, embora a qualidade da escola seja importante, as disparidades socioeconômicas dos alunos e o ambiente familiar podem ter um peso ainda maior. Isso desafia a ideia de que a escola, por si só, pode reverter completamente as desigualdades de origem, apontando para a necessidade de políticas holísticas que abordem tanto os fatores intraescolares quanto os extraescolares.

2.5 Disparidades Regionais e a Participação no ENEM

O desempenho no ENEM também exibe variações significativas entre diferentes regiões e unidades da federação (1). As disparidades regionais não são apenas geográficas, mas refletem a heterogeneidade socioeconômica e a capacidade de resposta dos sistemas educacionais locais a crises, como a pandemia de COVID-19 (11). O período pós-pandemia, em particular, evidenciou um agravamento das desigualdades regionais na participação e no desempenho, com quedas não homogêneas nas taxas de inscrição (12). A maior queda proporcional na taxa de inscrição ocorreu na região Sudeste, que de um pico de 63% em 2016, chegou a apenas 26% em 2021, tornando-se a região com o menor indicador naquele ano (11).

2.6 Aplicações de Ciência de Dados na Análise do ENEM e resultados obtidos

A aplicação de técnicas de Ciência de Dados e *Machine Learning* na análise dos microdados do ENEM tem se mostrado uma abordagem poderosa para aprofundar a compreensão dos fatores que influenciam o desempenho (1). Estudos têm utilizado Regressão Linear, Árvores de Decisão, *Random Forest*, *Boosting*, entre outras técnicas, para predição de notas e identificação de fatores relevantes (1, 3, 9, 10, 13–15).

Em seu trabalho, Melo *et al.* (1) utilizaram o método de Regressão Linear Múltipla para modelar a média da prova objetiva, média da Redação e as respectivas variâncias. Seus resultados indicam fortemente que o nível de escolaridade e profissionalização da mãe, a raça do estudante e a renda média da família são relevantes para o desempenho na prova objetiva. Ao adicionar uma componente espacial, os modelos apresentaram uma melhora, indicando que fatores regionais também influenciam o desempenho do estudante.

Moraes *et al.* (10) também aplicaram o método de Regressão Linear Múltipla para analisar o efeito escola no desempenho em Matemática, considerando variáveis como a quantidade média de alunos por turma, a média de horas-aula por dia e mais algumas variáveis que caracterizam a escola. Em sua análise exploratória, os autores identificaram as diferenças e semelhanças entre as escolas públicas e privadas, a exemplo do nível socioeconômico médio dos alunos da escola, onde “87% das escolas privadas estão nos níveis 5 e 6, enquanto 90% das escolas públicas possui nível socioeconômico entre os níveis 3 ou 4. Assim, as escolas públicas lidam [...] com alunos com níveis socioeconômicos menores.”

O nível socioeconômico médio dos alunos da escola chega “a aumentar a nota em 22,7 pontos para cada nível socioeconômico [...] nas escolas privadas e 12,3 pontos [...] nas escolas públicas.” Essa variável foi construída pelos autores e separada em 6 grupos, onde o grupo 6 reúne as escolas com os alunos de maior nível socioeconômico e o grupo 1 reúne as escolas com os alunos de menor nível socioeconômico.

Os Trabalhos de Conclusão de Curso de Amanda Ferraz (14) e Mayra Romero (13), para este mesmo MBA, aplicaram técnicas mais robustas. Ferraz utilizou *Random Forest* e *Boosting* para prever a aprovação de participantes do ENEM no SISU para o curso de Medicina, obtendo resultados satisfatórios com Coeficiente de Correlação de Matthews superior a 0,9. Já Romero desenvolveu e comparou modelos de classificação, incluindo o *Random Forest*, para identificar características socioeconômicas que indicam maior chance de o candidato atingir uma pontuação média acima de 500 pontos no ENEM. Ela concluiu que o *Random Forest* teve o melhor desempenho e que a renda familiar e o número de computadores são informações que impactam a previsibilidade do modelo.

2.7 Métodos de *Machine Learning*

Esta seção apresenta, de forma não exaustiva, alguns dos métodos de *Machine Learning* utilizados em trabalhos anteriores relacionados ao tema deste trabalho. Para isso, foram usadas as referências (16–19) como base para a descrição dos métodos.

2.7.1 Regressão Linear

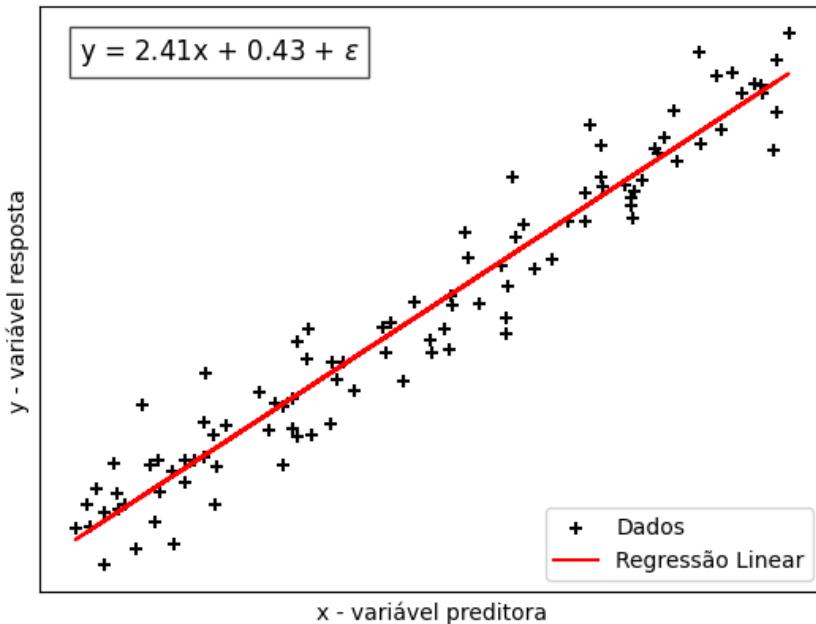
A Regressão Linear é um dos pilares do *Machine Learning*, sendo um método fundamental para a modelagem preditiva. Trata-se de um método paramétrico de aprendizado supervisionado que busca definir um modelo para uma relação linear entre a variável resposta e uma ou mais variáveis preditoras, tendo como objetivo central encontrar a melhor reta (ou hiperplano), em termos de erro na previsão, que descreva essa relação. A implementação mais básica é expressa pela equação

$$Y = \beta_0 + \beta_1 \times X + \epsilon \quad (2.1)$$

onde Y denota a variável resposta, X a variável preditora, β_0 o intercepto (o valor de Y quando $X = 0$), β_1 o coeficiente angular (indicando o impacto de X sobre Y) e ϵ o termo de erro. Em uma Regressão Linear Múltipla, diversas variáveis independentes são consideradas, cada uma com o seu β_i correspondente. Por trás da Regressão Linear, há algumas premissas adotadas, como a linearidade da relação entre X e Y , a independência dos erros, a homocedasticidade e a normalidade dos resíduos. Essas premissas podem ser interpretadas como desvantagens do modelo de Regressão Linear, por restringir ou até mesmo inviabilizar a sua aplicação. Já a fácil interpretação, simplicidade e eficiência

computacional são algumas das vantagens desse método, que também é muito utilizado como *benchmark* de métodos mais complexos.

Figura 1 – Exemplo de uma Regressão Linear simples com dados fictícios



Fonte: elaborado pelo autor.

2.7.2 Árvore de Decisão

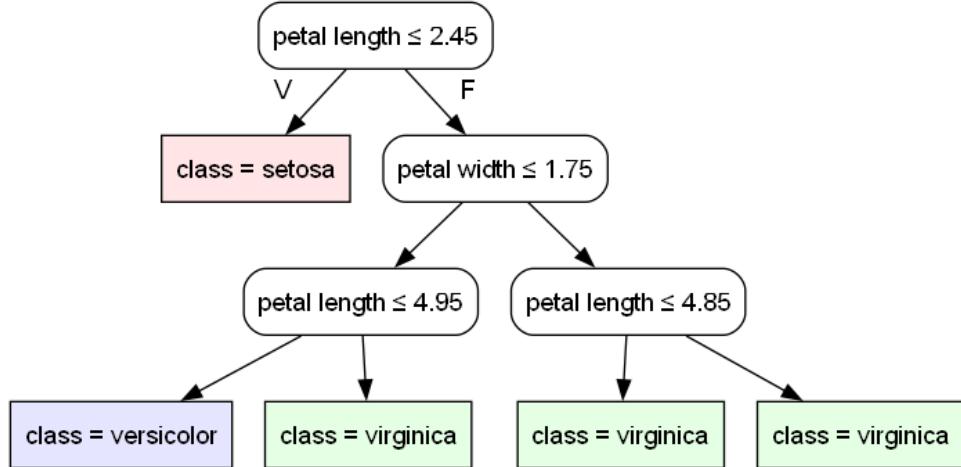
A Árvore de Decisão é um método paramétrico de aprendizado supervisionado que utiliza uma abordagem intuitiva de separação dos dados em grupos semelhantes, através de regras hierárquicas simples e de forma recursiva. Pode ser utilizado para resolver problemas de regressão, com a média da variável resposta em cada folha, ou de classificação, com a classe mais frequente em cada folha.

O processo de divisão segue uma lógica de “se-então”: se o dado de entrada tem o valor de uma variável preditora menor ou igual a um limite, então este segue pelo caminho à esquerda; se não, este segue pelo caminho à direita. É dessa lógica que surge a analogia com árvore, já que as regras usadas para definir o modelo podem ser representadas em um gráfico de árvore binária. A seleção das melhores divisões é baseada, para os problemas de classificação, em alguma medida de impureza, como a Entropia ou o Índice de Gini. Já para os problemas de regressão, as divisões são baseadas na redução de alguma medida de erro, como o Erro Quadrático Médio (*Mean Squared Error* - MSE).

Assim como a Regressão Linear, a Árvore de Decisão é um modelo de fácil interpretação, já que as regras de decisão são explícitas e podem ser visualizadas graficamente. É capaz de lidar com variáveis categóricas e contínuas, o que a torna versátil, não requer normalização dos dados e é robusta a *outliers*. No entanto, ela é propensa ao *overfitting*,

se não aplicadas técnicas de poda, e são instáveis, já que pequenas variações nos dados podem levar a grandes mudanças na estrutura da árvore.

Figura 2 – Exemplo de uma Árvore de Decisão com o *dataset Iris*



Fonte: elaborado pelo autor.

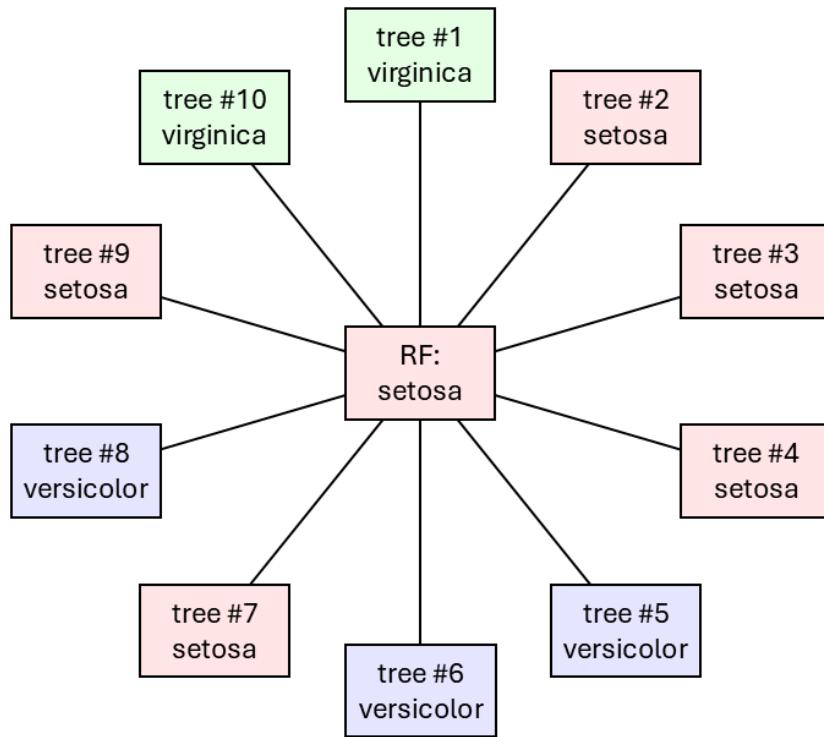
2.7.3 Random Forest

O *Random Forest* é um método derivado da Árvore de Decisão, sendo um dos algoritmos mais populares e eficazes em *Machine Learning*. Ele adota uma abordagem de *ensemble*, ou seja, combina múltiplos modelos para melhorar a precisão e a robustez das previsões. A ideia central é criar uma “floresta” de Árvores de Decisão, onde a decisão final é feita pela média/mediana das previsões para um problema de regressão ou pela classe mais frequente entre todas as árvores no caso de um problema de classificação.

O seu processo de construção envolve duas etapas principais: (i) a amostragem aleatória dos dados, onde cada árvore é treinada em um subconjunto diferente dos dados originais, e (ii) a seleção aleatória de variáveis em cada divisão, o que reduz a correlação entre as árvores e melhora a generalização do modelo. Essa aleatoriedade é crucial para evitar o *overfitting* e aumentar a diversidade entre as árvores.

O *Random Forest* é conhecido por sua alta precisão, capacidade de lidar com grandes conjuntos de dados e variáveis de diferentes tipos, resistência a *outliers* e facilidade de interpretação através da análise da importância das variáveis. No entanto, ele pode ser computacionalmente intensivo e menos interpretável do que uma única árvore de decisão, já que a combinação de múltiplas árvores torna mais difícil entender as regras subjacentes.

Figura 3 – Exemplo de uma *Random Forest* com o dataset *Iris*



Fonte: elaborado pelo autor.

2.7.4 Boosting

O *Boosting* é uma técnica de *ensemble*, combinando múltiplos modelos fracos para criar um modelo forte. A ideia central é treinar sequencialmente uma série de modelos, onde cada novo modelo foca em corrigir os erros cometidos pelos modelos anteriores. Alguns algoritmos populares de *Boosting* incluem o *AdaBoost*, *Gradient Boosting* e *XGBoost*.

O *AdaBoost* (*Adaptive Boosting*) foi um dos primeiros algoritmos de *Boosting* e funciona aumentando o peso dos dados de treinamento que foram classificados incorretamente pelos modelos anteriores. Ao final, as previsões de todos os modelos são combinadas, ponderadas pela precisão de cada modelo.

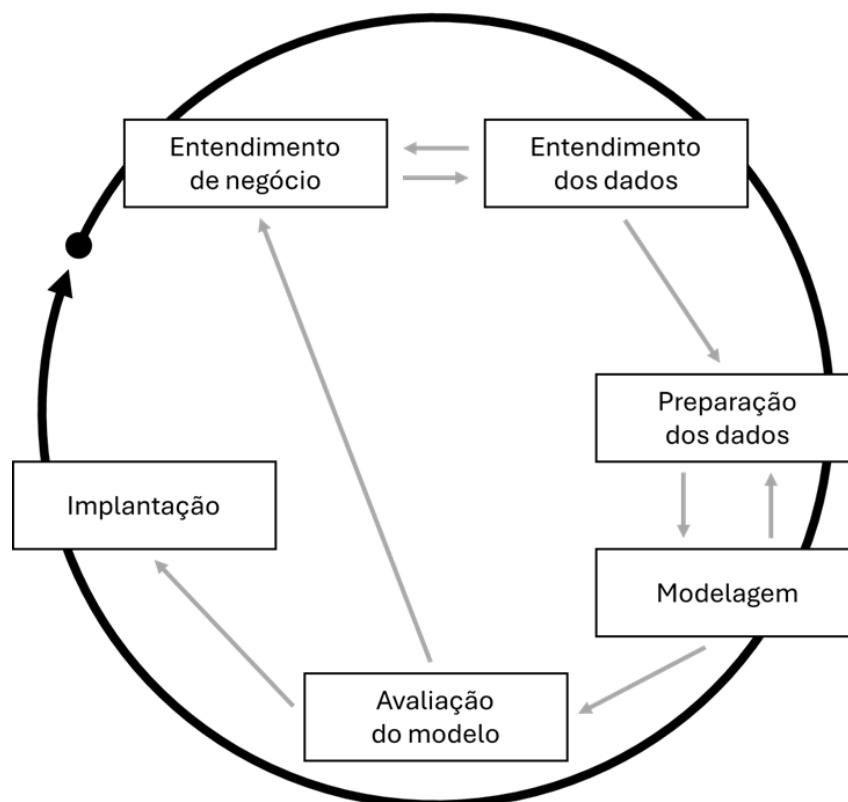
O *Gradient Boosting* usa uma abordagem de otimização, onde cada novo modelo é treinado especificamente nos resíduos do modelo anterior, buscando minimizá-los. Os novos aprendizes são adicionados de forma iterativa e geralmente são árvores de decisão de pequeno porte.

O *XGBoost* (*Extreme Gradient Boosting*) é uma implementação otimizada do *Gradient Boosting*, que oferece melhorias significativas em termos de velocidade e desempenho, implementando técnicas de regularização (L1 e L2), tratamento de valores ausentes, paralelização e outras otimizações.

3 METODOLOGIA

Este capítulo detalha a metodologia de trabalho utilizada, apresentando o delineamento da pesquisa, da coleta e processamento dos dados e as técnicas analíticas empregadas para responder às perguntas de pesquisa. A estrutura metodológica adotada baseou-se no modelo CRISP-DM (*Cross-Industry Standard Process for Data Mining*) (20), contendo as etapas de (i) entendimento de negócio, (ii) entendimento dos dados, (iii) preparação dos dados, (iv) modelagem, (v) avaliação e (vi) implantação.

Figura 4 – Modelo CRISP-DM



Fonte: modificado de Chapman *et al.* (20).

3.1 Entendimento de Negócio

A etapa de entendimento de negócio envolve a definição clara dos objetivos do projeto, a compreensão do contexto em que a pesquisa está inserida, a identificação das partes interessadas, a formulação das perguntas de pesquisa que guiarão a análise dos dados e os resultados que se espera alcançar.

Como o foco principal deste trabalho é a formulação de hipóteses relacionadas aos fatores que influenciam o desempenho dos estudantes no ENEM, através da análise dos dados de performance dos participantes e suas características socioeconômicas, foi

necessário formular perguntas de pesquisa específicas que pudessem ser respondidas através da análise dos dados disponíveis.

3.2 Entendimento dos dados

Com as perguntas de pesquisa definidas, a etapa seguinte consistiu em encontrar dados que fossem adequados para responder a essas perguntas e estabelecer uma forma consistente de coleta e armazenamento desses dados para, em seguida, realizar uma compreensão de sua estrutura e significado.

Após essa etapa, foi possível identificar quais arquivos eram relevantes para a análise e quais variáveis dentro desses arquivos seriam utilizadas como variáveis preditoras e como variáveis resposta. A depender dos tipos de dados a serem utilizados, é necessário submeter o projeto a um comitê de ética em pesquisa para aprovação, garantindo que todos os aspectos éticos relacionados ao uso dos dados sejam devidamente considerados.

3.3 Preparação dos dados

Com os arquivos relevantes selecionados, passou-se para a etapa de preparação dos dados, que envolve a leitura, limpeza, transformação e integração dos dados para torná-los adequados para a modelagem. Essa etapa é crucial, pois a qualidade dos dados impacta diretamente na eficácia dos modelos preditivos construídos posteriormente. Para a execução dessa e das etapas posteriores, foi necessário preparar um ambiente tecnológico e analítico adequado que permitisse a manipulação eficiente dos dados e a construção dos modelos preditivos.

Após a leitura dos dados, estes foram integrados em um único conjunto de dados para facilitar a análise e a modelagem. Foram realizados os ajustes necessários no esquema dos dados para garantir a consistência e a integridade das informações.

Em seguida, as variáveis foram renomeadas para nomes mais intuitivos e de fácil compreensão e foi analisada a necessidade de transformar os valores das variáveis originais em valores mais comprehensíveis, por exemplo, transformar os códigos numéricos de variáveis categóricas em rótulos textuais.

Posteriormente, foi realizada uma análise para identificar e tratar valores nulos, removendo ou imputando valores conforme apropriado. Após o tratamento dos valores nulos, os dados foram separados em diferentes conjuntos de dados, cada um correspondente a uma variável resposta específica, garantindo que cada conjunto contivesse apenas as observações relevantes para a análise daquela variável e sem valores nulos.

3.4 Modelagem

A etapa de modelagem envolveu a análise exploratória dos dados para entender suas características, avaliação de correlações entre as variáveis preditoras, seleção e treinamento dos modelos preditivos e otimização dos hiperparâmetros.

3.4.1 Análise Exploratória dos Dados

A análise exploratória foi iniciada com o entendimento das variáveis respostas, buscando entender a distribuição das notas do ENEM. Em seguida, foi feito um teste de hipótese para avaliar se as médias das notas variam significativamente entre as edições selecionadas para este trabalho. Essa análise é necessária uma vez que as edições do ENEM podem apresentar variações no nível de dificuldade das provas, impactando as notas dos estudantes.

Para isso, foi utilizada a Estatística F de ANOVA (*Analysis of Variance*) (21) para comparar as médias das notas entre as diferentes edições do ENEM e o cálculo do tamanho do efeito para quantificar a magnitude das diferenças encontradas, utilizando a métrica SMD (*Standardized Mean Difference*) com os intervalos definidos por Jacob Cohen (22).

A próxima etapa da análise exploratória foi a identificação de *outliers* nas variáveis respostas, utilizando o critério de 1,5 vezes o intervalo interquartil (IQR) (21). A identificação dos *outliers* é importante para entender a distribuição das notas e avaliar se esses valores extremos podem influenciar os resultados dos modelos preditivos.

Finalizadas as análises das variáveis respostas, foram realizadas três análises sobre as variáveis preditoras: (i) análise de concentração de categorias, (ii) análise de correlação entre as variáveis preditoras e (iii) cálculo do *Permutation Importance* para cada variável preditora.

Utilizando essas três informações, foi feita uma análise qualitativa para identificar possíveis variáveis preditoras a serem removidas do conjunto de dados, seja por apresentarem baixa correlação com as variáveis respostas, por apresentarem alta correlação com outras variáveis preditoras (multicolinearidade) ou por apresentarem baixa importância na previsão das notas do ENEM.

Para a correlação, foi utilizada a biblioteca `phik` (23) (24), que permite calcular a correlação entre variáveis categóricas e numéricas, além de fornecer métricas para avaliar a força da correlação.

Para o cálculo do *Permutation Importance*, foi utilizado o modelo de *Random Forest Regressor* do `cuml` (25) e o método de *permutation importance* da biblioteca `sklearn` (26). Foi necessário separar os conjuntos de dados em treino e teste previamente, a fim de evitar vazamento de dados e garantir que a avaliação da importância das variáveis fosse feita de

forma justa e realista. Para essa separação, foi utilizado o método `train_test_split` da biblioteca `sklearn`, com uma proporção de 80% dos dados para treino e 20% para teste.

A partir desse momento, não se utilizou mais o conjunto de dados completo, mas sim apenas o conjunto de treino para as análises das variáveis preditoras.

3.5 Treinamento dos Modelos

Nesta seção são detalhadas as técnicas de modelagem empregadas e os critérios utilizados para a seleção dos modelos. Como as variáveis respostas são numéricas e contínuas, trata-se de um problema de regressão de aprendizado supervisionado. Assim, foi preciso selecionar modelos de regressão, que são os adequados para prever variáveis contínuas.

Foram utilizados três modelos de regressão: (i) *Random Forest Regressor*, (ii) *XGBoost Regressor* e (iii) *LightGBM Regressor*. Esses modelos são amplamente utilizados devido à sua eficácia e interpretabilidade, além de serem capazes de lidar com grandes volumes de dados e variáveis preditoras.

Para otimizar o desempenho dos modelos selecionados, foi realizada uma busca em grade (*Grid Search*) para identificar os melhores hiperparâmetros. A busca não foi realizada utilizando validação cruzada, devido ao alto volume de dados, mas sim um subconjunto do conjunto de treino equivalente a 10% dos dados originais, garantindo que os resultados fossem robustos e generalizáveis.

A implementação da busca em grade foi feita manualmente, utilizando *loops* para iterar sobre os diferentes valores dos hiperparâmetros e avaliando o desempenho dos modelos utilizando o conjunto de validação. Isso se deu para evitar o alto custo computacional associado à utilização de bibliotecas como `GridSearchCV` da `sklearn`, que realizam a busca em grade utilizando validação cruzada, o que pode ser inviável para grandes volumes de dados e modelos complexos.

Definidos os hiperparâmetros ótimos, os modelos foram treinados utilizando o conjunto de dados preparado na Seção 3.3, com uma quantidade maior de estimadores fracos (através do hiperparâmetro `n_estimators`) para garantir um melhor desempenho dos modelos.

O treinamento foi realizado utilizando o conjunto de treino completo, garantindo que os modelos fossem treinados com a maior quantidade possível de dados para melhorar sua capacidade de generalização.

Finalizando a etapa de treinamento, foram criados modelos de *ensemble* utilizando a técnica de *bagging*, onde os modelos de regressão individuais foram combinados para criar um modelo mais robusto e preciso. A combinação foi feita utilizando a média das previsões dos modelos individuais, garantindo que o modelo de *ensemble* aproveitasse as

forças de cada modelo individual para melhorar a precisão das previsões.

3.6 Avaliação dos Modelos

Após o treinamento final dos modelos, estes foram avaliados utilizando o conjunto de teste e duas métricas de desempenho apropriadas para problemas de regressão: a Raiz do Erro Quadrático Médio (*Root Mean Squared Error* - RMSE) e o Erro Percentual Absoluto Médio (*Mean Absolute Percentage Error* - MAPE). Essas métricas fornecem uma visão clara da precisão das previsões dos modelos em relação aos valores reais das notas do ENEM.

Para a nota da Redação, como esta apresenta uma quantidade limitada de valores possíveis (de 0 a 1000, com incrementos de 20 pontos), foi realizado um tratamento adicional para arredondar as previsões dos modelos para os valores possíveis antes do cálculo das métricas, garantindo que as previsões fossem coerentes com a escala de notas do ENEM, conforme o código abaixo:

```
def arredonda_redacao(nota: float) -> int:
    """
    Arredonda a nota de redação para o múltiplo de 20 mais próximo
    """
    passo = 20
    return np.round(nota / passo) * passo
```

Após o cálculo das métricas de desempenho, duas análises foram realizadas: (i) uma análise entre o erro de treinamento e o erro de teste para avaliar se os modelos estavam sofrendo de *overfitting* e (ii) uma análise entre os modelos para identificar qual modelo apresentava o melhor desempenho na previsão das notas do ENEM.

Na análise de *overfitting*, foi adotado um critério sobre a razão do erro RMSE de teste em relação ao erro RMSE de treinamento. Uma razão maior que 15% foi considerada um indicativo de *overfitting*, sugerindo que o modelo estava se ajustando demais aos dados de treinamento e não generalizando bem para os dados de teste.

Na análise comparativa entre os modelos, para se escolher o melhor modelo para cada variável resposta, foi considerado o modelo que apresentou o menor erro MAPE no conjunto de teste, garantindo que o modelo selecionado fosse aquele com a melhor capacidade de prever as notas do ENEM com precisão.

3.7 Influência das Variáveis Preditoras

Nesta seção, buscou-se responder às perguntas de pesquisa que foram formuladas na seção 3.1, utilizando técnicas de interpretação de modelos e análise de sensibilidade

para medir a magnitude da influência das variáveis preditoras nas variáveis resposta.

Primeiramente, foi realizada uma análise de importância das variáveis preditoras do modelo final selecionado para cada variável resposta. Essa análise permitiu identificar quais variáveis preditoras têm a maior influência na previsão das notas do ENEM, fornecendo *insights* sobre os fatores socioeconômicos que mais impactam o desempenho dos estudantes.

Em seguida, foi realizada uma análise de sensibilidade para avaliar como as variações nas variáveis preditoras afetam as previsões dos modelos. Foi construída uma base sintética de dados, onde algumas variáveis preditoras de interesse foram selecionadas e as demais variáveis preditoras foram preenchidas com o valor mais frequente do conjunto de treino. A ideia foi buscar entender como as variações nas variáveis de interesse impactam as previsões dos modelos, mantendo as demais variáveis constantes.

4 RESULTADOS

Este capítulo apresenta os resultados obtidos a partir da aplicação da metodologia descrita no Capítulo 3 - Metodologia. Os resultados serão apresentados na mesma ordem das etapas descritas na metodologia.

4.1 Entendimento de Negócio

Conforme mencionado no Capítulo 2 - Fundamentação Teórica, o ENEM é um exame de grande relevância no contexto educacional brasileiro e compreender os fatores que impactam o desempenho dos estudantes é crucial para a formulação de políticas educacionais eficazes.

Trabalhos anteriores citam algumas variáveis socioeconômicas como discriminadores de performance no ENEM. A Tabela 1 apresenta essas variáveis identificadas na literatura, juntamente com suas respectivas referências.

Tabela 1 – Variáveis socioeconômicas e suas referências

Variável socioeconômica	Referência
Renda Familiar	Melo <i>et al.</i> (1), Vasconcellos (8)
Raça/Cor	Melo <i>et al.</i> (1), Moraes <i>et al.</i> (10)
Sexo	Moraes <i>et al.</i> (10)
Idade/Atraso Escolar	Jaloto e Primi (9), Moraes <i>et al.</i> (10)
Administração: Pública vs. Privada	Ortega <i>et al.</i> (2), Jaloto e Primi (9) e Moraes <i>et al.</i> (10)
Atributos Escolares	Moraes <i>et al.</i> (10)

Fonte: elaborado pelo autor.

Ao avaliarmos os trabalhos anteriores disponíveis, concluiu-se que há uma variedade de fatores socioeconômicos que podem influenciar o desempenho dos estudantes no ENEM. Com base nisso, foram formuladas as seguintes perguntas de pesquisa:

- **Pergunta 1:** Quais são os principais fatores socioeconômicos que influenciam o desempenho dos estudantes no ENEM?
- **Pergunta 2:** Qual é a magnitude da influência de cada um desses conjuntos de fatores nas notas dos participantes?

4.2 Entendimento dos dados

4.2.1 Escolha e Coleta dos Dados

Como descrito na Seção 3.2, foi necessário identificar dados que fossem relevantes para responder às perguntas de pesquisa formuladas. Foi realizada uma busca por bases de dados públicas que contivessem informações detalhadas sobre os participantes do ENEM, incluindo suas características socioeconômicas e desempenho no exame.

Os microdados do ENEM, disponibilizados anualmente pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), foram escolhidos como a principal fonte de dados para este trabalho e podem ser acessados através do portal do INEP (4).

No mesmo portal, também estão disponíveis os dados do Censo Escolar, que fornecem informações adicionais sobre as escolas de todo o território nacional (15). Esses foram escolhidos como fonte complementar por fornecerem um contexto mais amplo sobre o ambiente educacional.

Foram então selecionadas as edições de 2020 a 2024 (as últimas cinco edições disponíveis) de ambos os conjuntos de dados e os arquivos disponibilizados foram baixados através de *download* simples e armazenados localmente para posterior leitura e manipulação.

Como os dados escolhidos são públicos e anonimizados por quem os distribui, entendeu-se que não há limitações éticas para o uso desses dados neste trabalho e não foi necessário submeter o projeto a um comitê de ética em pesquisa.

4.2.2 Compreensão Inicial dos Dados

Os arquivos de microdados do ENEM e do Censo Escolar são disponibilizados em formato compactado (.zip), separados pelo ano de aplicação do exame/censo.

Dentre os arquivos existentes nos arquivos compactados dos microdados do ENEM, foram selecionados os arquivos .csv (*Comma-Separated Values*) que contêm as informações dos participantes e suas notas e os dicionários de dados de cada edição em formato .xlsx (Formato nativo do Microsoft Excel), que foram utilizados para interpretar os valores categóricos e identificar variáveis importantes.

Para os arquivos compactados do Censo Escolar, foram selecionados os arquivos .csv que contêm as informações das escolas e os dicionários de dados em formato .xlsx.

4.2.2.1 Edição de 2024 do ENEM e LGPD

Na edição de 2024 dos microdados do ENEM, foi feita uma alteração no formato de disponibilização dos dados dos participantes e das notas, que passaram a ser disponibilizados em arquivos separados.

Isso se deu “Devido à vigência da Lei Geral de Proteção de Dados (LGPD), incorporada ao ordenamento jurídico brasileiro por meio da Lei nº 13.709, de 14 de agosto de 2018” (27), conforme descrito no arquivo auxiliar “Leia-Me” (27) disponibilizado junto com os microdados do ENEM 2024.

Ou seja, o formato dos arquivos de microdados do ENEM 2024 difere das edições anteriores, por mais que as informações contidas permanecem as mesmas. Houve a separação dos dados dos participantes e das notas em arquivos distintos e sem uma chave primária que permita a junção dos dois conjuntos de dados. Dessa forma, os dados da edição de 2024 do ENEM não puderam ser utilizados para este trabalho.

4.2.3 Análise dos Dicionários de Dados

Foram analisados os dicionários de dados dos microdados do ENEM e do Censo Escolar para identificar as variáveis disponíveis em cada conjunto de dados. Os dicionários completos estão disponíveis no Apêndice A e B. A partir dessa análise, foi possível identificar as variáveis que seriam relevantes para responder às perguntas de pesquisa formuladas na Seção 4.1.

Não foi possível localizar uma variável que permitisse a identificação única das escolas dos participantes do ENEM nos microdados do exame, o que impossibilitou correlacionar diretamente os dados dos participantes do ENEM com os dados das escolas do Censo Escolar para agregar informações das escolas aos dados dos participantes. Dessa forma, optou-se por utilizar apenas os microdados do ENEM para a realização deste trabalho.

4.2.4 Definição da Variável Resposta

Como este trabalho pretende avaliar o desempenho dos estudantes no ENEM e os fatores que influenciam esse desempenho, a variável resposta deve refletir esse objetivo. Assim, foram utilizadas como variáveis resposta as notas obtidas pelos estudantes nas quatro provas objetivas e na redação do ENEM.

Ou seja, foram usadas cinco variáveis resposta distintas para análise: (i) Nota da prova de Ciências Humanas; (ii) Nota da prova de Ciências da Natureza; (iii) Nota da prova de Linguagem e Código; (iv) Nota da prova de Matemática; e (v) Nota da Redação.

4.3 Preparação dos dados

4.3.1 Preparação do Ambiente Tecnológico e Analítico

Para a execução deste trabalho, foi utilizado um ambiente baseado em Python versão 3.11 através do gerenciador de ambientes virtuais Miniconda3 (28). O computador utilizado possui uma CPU AMD Ryzen 7 9800X3D, 32 GB de memória RAM e uma GPU

NVIDIA GeForce RTX 4070 Ti Super, com 16 GB de memória dedicada com sistemas operacionais Ubuntu 24.04 LTS e Windows 11 Pro.

O ambiente foi especificamente configurado com o ecossistema NVIDIA CUDA-X (29) para possibilitar a execução utilizando a GPU do equipamento, visando acelerar o processamento dos dados e a modelagem. Esta suíte de bibliotecas de software permite executar *pipelines* de Ciência de Dados e análises inteiramente na GPU, minimizando a transferência de dados entre a CPU e a GPU.

Foram utilizados seus principais componentes: `cudf` (30), uma biblioteca para manipulação de `DataFrames` na GPU, análoga ao `pandas` (31), e `cuml` (25), que fornece implementações de algoritmos de *Machine Learning* acelerados por GPU, análoga ao `scikit-learn` (26). Todo o ambiente foi construído sobre a plataforma CUDA 13.1, com as bibliotecas e dependências gerenciadas diretamente pelo Conda.

O arquivo YML de configuração do ambiente virtual utilizado está disponível no Apêndice C.

4.3.2 Leitura dos Dados

Os arquivos `.csv` dos microdados do ENEM foram lidos utilizando o método `read_csv` da biblioteca `pandas`, especificando o separador como ponto e vírgula (`sep=';'`).

Foram utilizados os dados das edições de 2020 a 2023, que possuem a quantidade de observações e variáveis descritas na Tabela 2. As tabelas foram carregadas já desconsiderando colunas que não agregam ao modelo, como o número de inscrição do participante, por exemplo.

Tabela 2 – Observações e variáveis por edição do ENEM

Edição	Observações	Variáveis
2020	5.783.109	52
2021	3.389.832	52
2022	3.476.105	52
2023	3.933.955	52

Fonte: elaborado pelo autor.

4.3.3 Integração dos Dados

Analizando o dicionário de dados de cada edição, foi possível observar que todas as edições possuem o mesmo esquema, ou seja, as mesmas variáveis com os mesmos nomes estão presentes em todas as edições selecionadas. Assim, a integração entre edições foi realizada por meio da concatenação vertical dos quatro conjuntos de dados, utilizando o método `concat` da biblioteca `pandas`.

Em seguida, foi feita uma modificação no nome das variáveis para nomes que fossem mais intuitivos e de compreensão rápida do conteúdo. Essa modificação foi realizada utilizando o método `rename`, a partir de um dicionário que mapeava os nomes originais para os novos nomes desejados.

Com o dicionário de dados analisado, foi diagnosticado que algumas variáveis categóricas estavam codificadas com valores numéricos que não eram intuitivos. Assim, seus valores foram transformados, substituindo os códigos numéricos por descrições textuais mais comprehensíveis através do método `map` da biblioteca `pandas`, utilizando dicionários de mapeamento construídos especificamente para cada variável categórica que necessitava de transformação.

4.3.4 Tratamento de Valores Nulos

Inicialmente, foi feito o cálculo do percentual de valores nulos por variável. A Tabela 3 apresenta estes valores do conjunto de dados integrados, antes dos tratamentos.

Tabela 3 – Percentual de valores nulos por variável

Variável	Percentual de nulos
sigla_uf_escola	78,1%
cod_municipio_escola	78,1%
tp_adm_escola	78,1%
funcionamento_escola	78,1%
tp_local_escola	78,1%
tp_ensino	69,8%
tp_escola	68,2%
ano_conclusao_ensino_medio	51,6%
nota_ciencias_natureza	40,4%
nota_matematica	40,4%
nota_ciencias_humanas	37,0%
nota_redacao	37,0%
nota_linguagem_codigos	37,0%
03_ocupacao_pai	12,5%
01_escolaridade_pai	9,8%
04_ocupacao_mae	9,0%
estado_civil	4,2%
02_escolaridade_mae	3,5%
cor_raca	1,8%
10_qtde_carro	0,6%
05_qtde_moradores	0,6%

Continua na próxima página...

Variável	Percentual de nulos
06_renda_familiar	0,6%
07_dias_trabalhador_domestico	0,6%
08_qtde_banheiro	0,6%
09_qtde_quarto	0,6%
18_flag_aspirador_po	0,6%
11_qtde_motocicleta	0,6%
12_qtde_geladeira	0,6%
13_qtde_freezer	0,6%
14_qtde_maq_lavar_roupa	0,6%
15_qtde_maq_secar_roupa	0,6%
16_qtde_micro_ondas	0,6%
17_qtde_maq_lavar_louca	0,6%
22_qtde_celular	0,6%
19_qtde_tv	0,6%
20_flag_aparelho_dvd	0,6%
21_flag_tv_assinatura	0,6%
24_qtde_computadores	0,6%
23_flag_telefone_fixo	0,6%
25_flag_internet	0,6%
nacionalidade	0,05%

Fonte: elaborado pelo autor.

O percentual de valores nulos varia significativamente, com algumas variáveis apresentando mais de 70% de valores nulos, enquanto outras possuem menos de 1%. Variáveis com uma alta proporção de valores nulos podem comprometer a análise se for realizada alguma imputação de valores. Sendo assim, foi decidido remover as variáveis que apresentavam mais de 50% de valores nulos, resultando na remoção de nove variáveis do conjunto de dados.

Para as variáveis das notas, por serem as variáveis resposta deste trabalho, foi realizada uma análise mais detalhada. Primeiro, verificou-se a existência de valores zerados nessas variáveis e se possuem significados diferentes de valores nulos. Para isso, foi feita uma análise com a presença nas provas e o status da redação.

Foi identificado que a nota zerada significa que o participante esteve presente na prova, mas obteve nota zero, enquanto o valor nulo indica que o participante ou não realizou a prova, ou foi eliminado, ou teve sua redação anulada. Dessa forma, optou-se por manter as observações com notas zeradas no conjunto de dados, removendo apenas as observações com notas nulas. A decisão de incorporar ou não as notas zero na análise é

discutida na Seção 4.4.1.3.

Para as demais variáveis com valores nulos, foi realizada uma análise consolidada, ou seja, foram retiradas as observações que possuíam valor nulo em qualquer uma das variáveis restantes, o que resultou na retirada de 4.341.559 observações do conjunto de dados.

Assim, restaram 12.241.442 observações e 45 variáveis no conjunto de dados após os tratamentos.

4.3.5 Separação dos Conjuntos de Dados por Variável Resposta

Após a separação dos conjuntos de dados por variável resposta, conforme descrito na Seção 3.3, foram criados cinco conjuntos de dados distintos. A Tabela 4 apresenta a quantidade de observações e variáveis em cada conjunto de dados.

Tabela 4 – Observações e variáveis por conjunto de dados

Conjunto de Dados	Observações	Variáveis
Ciências Humanas	7.895.093	35
Ciências da Natureza	7.500.050	35
Linguagem e Código	7.895.093	35
Matemática	7.500.050	35
Redação	7.895.093	35

Fonte: elaborado pelo autor.

A estrutura de dicionários foi utilizada para manter o controle dos conjuntos de dados, suas respectivas variáveis resposta e variáveis preditoras ao longo do trabalho.

4.4 Modelagem

4.4.1 Análise Exploratória dos Dados - Variáveis Resposta

4.4.1.1 Distribuições

O primeiro passo da análise exploratória foi entender o domínio das variáveis respostas e constatou-se que, para as notas das provas objetivas (Ciências Humanas, Ciências da Natureza, Linguagem e Código e Matemática), havia mais de cinco mil notas distintas, com variações pequenas entre elas (décimos de pontos). Já para a nota da Redação, o número de notas distintas era significativamente menor (apenas 50 notas) com variação de 20 em 20 pontos.

A tabela 5 apresenta as estatísticas descritivas das notas de cada prova, obtido através do método `describe` da biblioteca `pandas`.

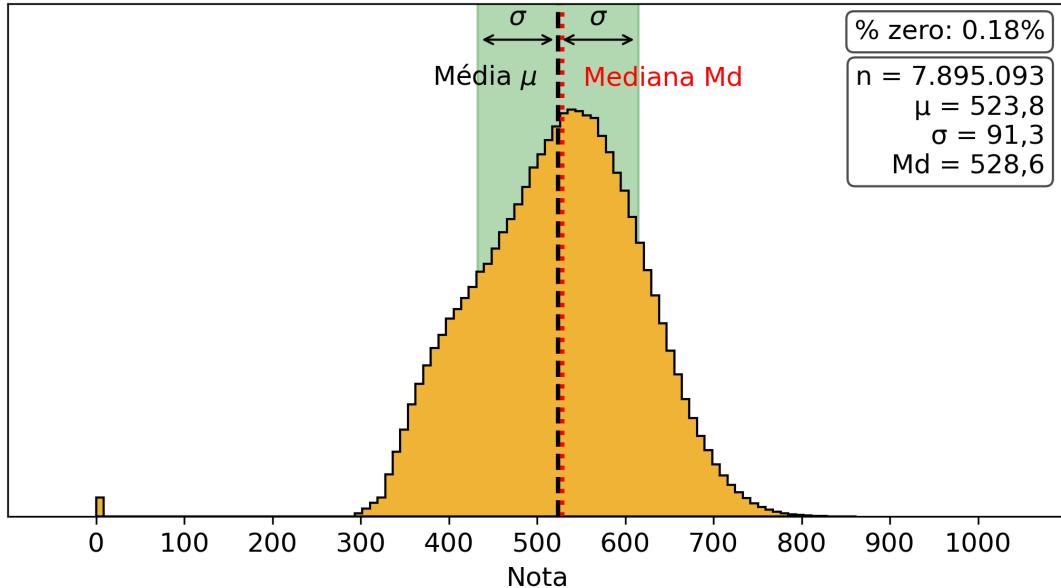
Tabela 5 – Estatísticas descritivas por conjunto de dados

Estatística	Humanas	Natureza	Linguagem	Matemática	Redação
Contagem	7.895.093	7.500.050	7.895.093	7.500.050	7.895.093
Média	523,8	496,9	519,0	538,7	616,6
Desvio Padrão	91,3	81,4	91,3	121,5	204,7
Mínimo	0,0	0,0	0,0	0,0	0,0
25º Percentil	460,1	437,3	460,1	441,6	520
50º Percentil	528,6	490,3	528,6	526	620
75º Percentil	588,2	551,9	588,2	624,9	760
Máximo	862,6	875,3	826,1	985,7	1000

Fonte: elaborado pelo autor.

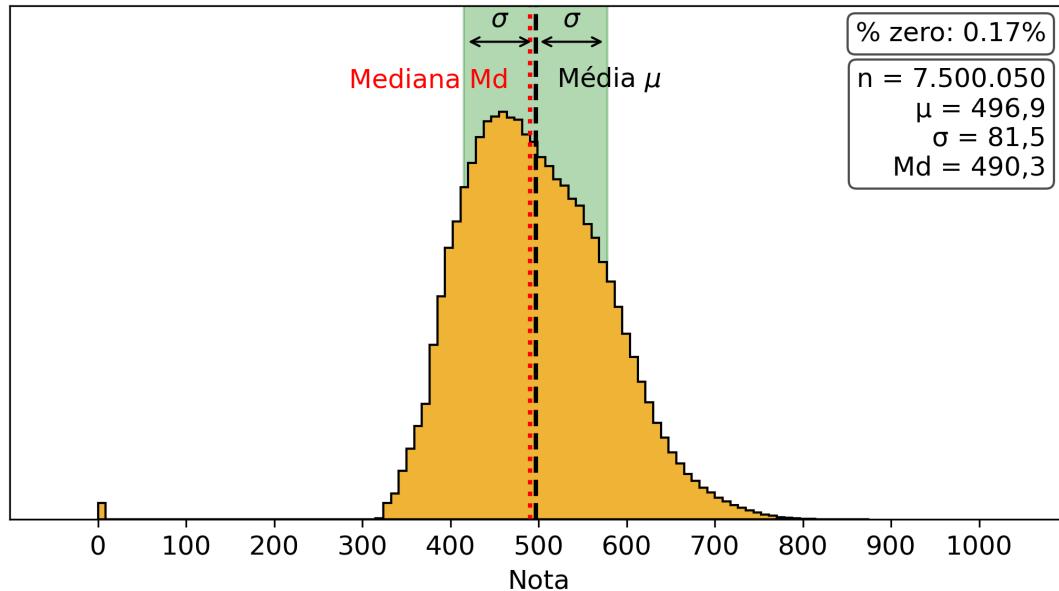
Em seguida, foram construídos histogramas de cada nota para entender a distribuição. As Figuras 5 a 9 apresentam os histogramas das notas de cada prova.

Figura 5 – Histograma das notas - Humanas



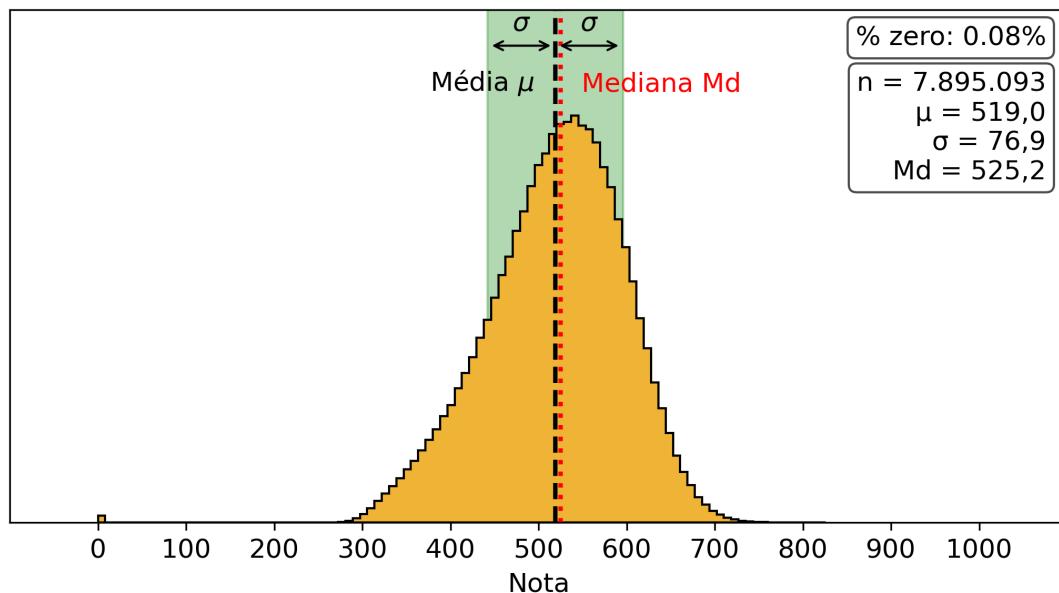
Fonte: elaborado pelo autor.

Figura 6 – Histograma das notas - Natureza



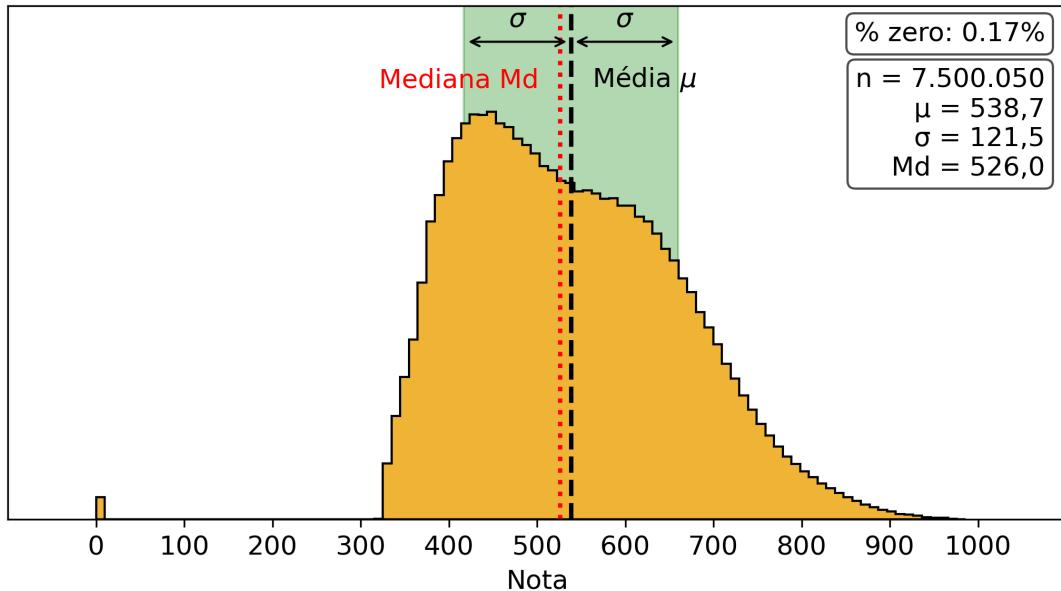
Fonte: elaborado pelo autor.

Figura 7 – Histograma das notas - Linguagem



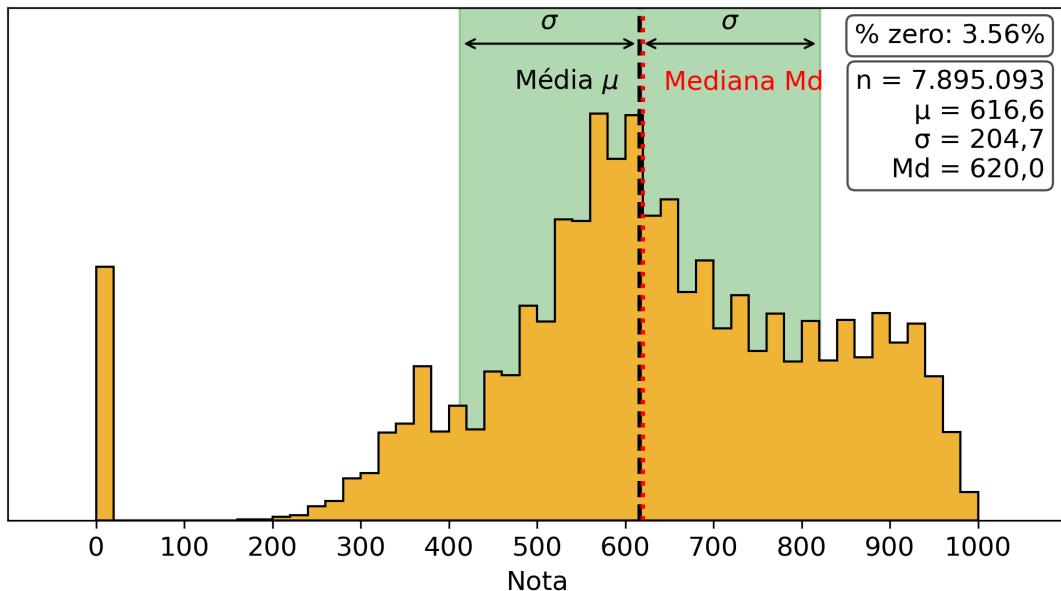
Fonte: elaborado pelo autor.

Figura 8 – Histograma das notas - Matemática



Fonte: elaborado pelo autor.

Figura 9 – Histograma das notas - Redação



Fonte: elaborado pelo autor.

A tabela 6 apresenta os valores de assimetria e curtose das notas de cada prova, obtidos através dos métodos `skew` e `kurtosis` da biblioteca `pandas`, e o percentual de notas zero em cada conjunto de dados.

Tabela 6 – Assimetria, Curtose e Notas zeradas

Variável	Assimetria	Curtose	Notas zeradas
Humanas	-0,3408	1,1269	0,18%
Natureza	0,0321	1,8372	0,17%
Linguagem	-0,5113	1,1780	0,08%
Matemática	0,3138	0,0850	0,17%
Redação	-0,7457	1,0488	3,56%

Fonte: elaborado pelo autor.

Analizando os valores de assimetria e curtose, é possível observar que as distribuições das notas possuem diferentes características.

A assimetria da nota de Redação é a mais negativa, o que indica que os alunos tiveram, em geral, o melhor desempenho, assim como nas provas de Linguagem e Código e Ciências Humanas, que também apresentam assimetria negativa, porém com valores menores.

Na prova de Ciências da Natureza, a assimetria é praticamente nula, indicando uma distribuição mais simétrica das notas, enquanto a nota de Matemática apresenta a assimetria mais positiva, indicando um desempenho relativamente pior dos alunos nessa prova.

Analizando os valores da curtose, a prova de Matemática foi a única a apresentar uma curtose próxima de zero, indicando uma distribuição mais próxima da normalidade. As outras provas apresentaram valores maiores que 1, indicando distribuições com caudas mais pesadas e picos mais acentuados.

A quantidade de notas zeradas é relativamente baixa, variando entre 0,08% e 3,56%, sendo a nota de Redação a que possui o maior percentual de notas zeradas.

4.4.1.2 Teste de Hipótese

Foi realizado o teste de hipótese ANOVA com nível de significância de 0,1% para comparar as médias das notas por edição do ENEM em cada conjunto de dados, onde a hipótese nula H_0 é de que as médias são iguais entre as edições. A Tabela 7 apresenta os valores de F, p-valor e a métrica SMD (*Standardized Mean Difference*) obtidos para cada conjunto de dados.

Tabela 7 – Teste ANOVA das médias das notas por edição

Variável	Valor F	Rejeita-se H_0 ?	p-valor	SMD	Tamanho do efeito
Humanas	13.161	Sim	0,0000	0,185	Insignificante
Natureza	3.431	Sim	0,0000	0,087	Insignificante
Linguagem	26.506	Sim	0,0000	0,269	Pequeno
Matemática	13.041	Sim	0,0000	0,197	Insignificante
Redação	27.498	Sim	0,0000	0,242	Pequeno

Fonte: elaborado pelo autor.

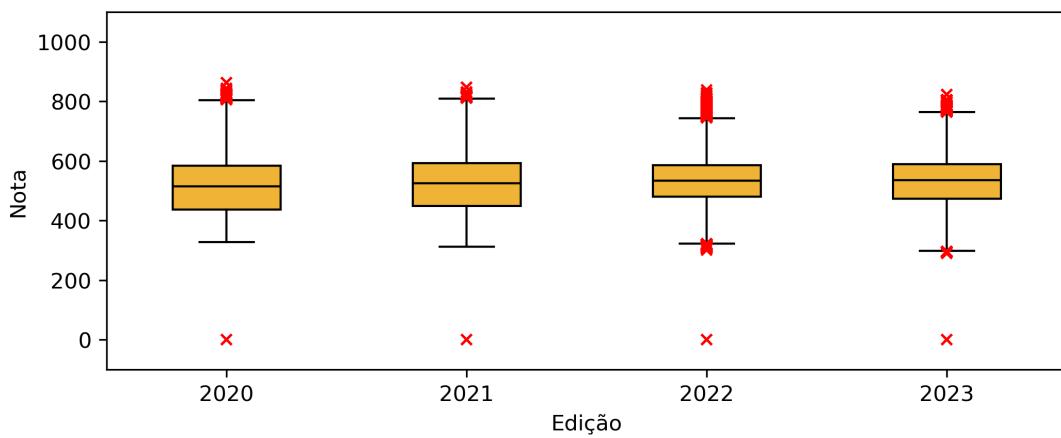
O teste de hipótese ANOVA indicou que há diferenças estatisticamente significativas entre as médias das notas por edição do ENEM em todos os conjuntos de dados, uma vez que o p-valor é menor que o nível de significância de 0,1%. Porém, ao analisar o tamanho do efeito através da métrica SMD, foi possível constatar que as diferenças entre as edições são insignificantes ou pequenas, o que indica que as edições do ENEM não tiveram um impacto relevante nas notas dos participantes.

Dessa forma, foi decidido manter todas as edições do ENEM no conjunto de dados para a modelagem preditiva e sem a necessidade de segmentação por edição.

4.4.1.3 Análise de Outliers

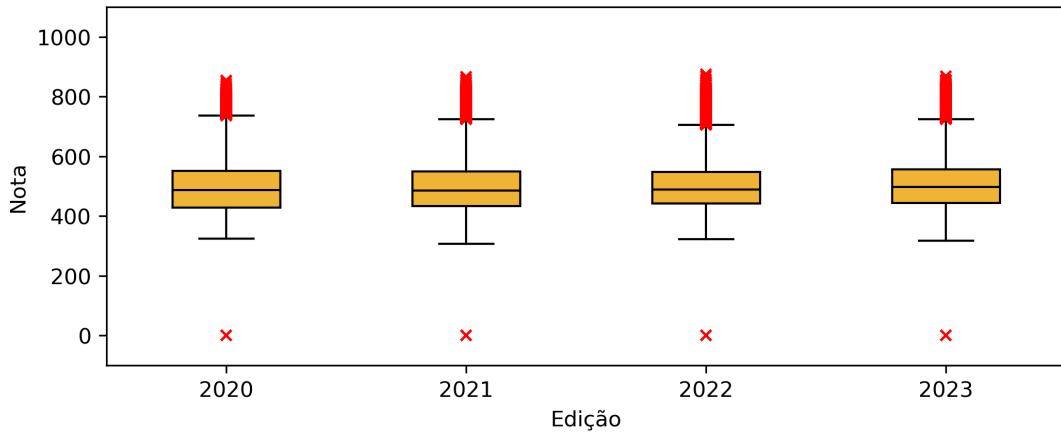
Para a análise dos outliers, foram utilizados os boxplots das notas de cada prova, apresentados nas Figuras 10 a 14.

Figura 10 – Boxplot das notas por edição - Humanas



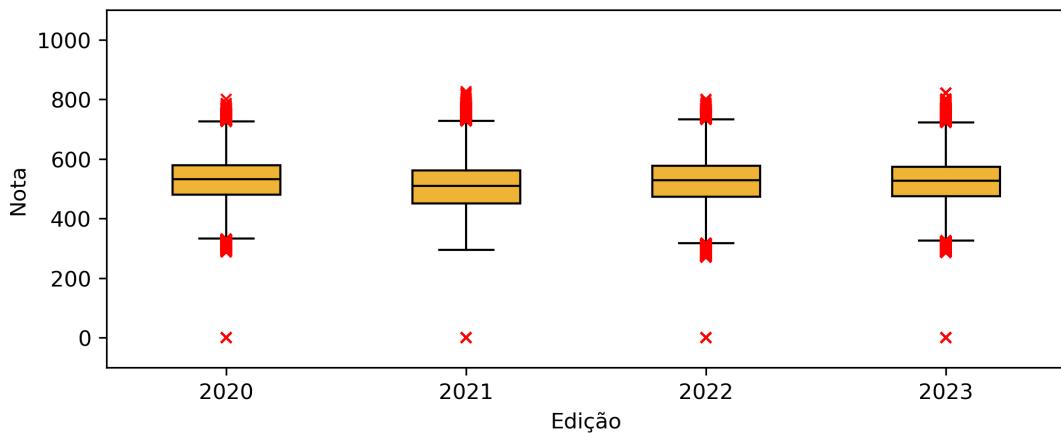
Fonte: elaborado pelo autor.

Figura 11 – Boxplot das notas - Natureza



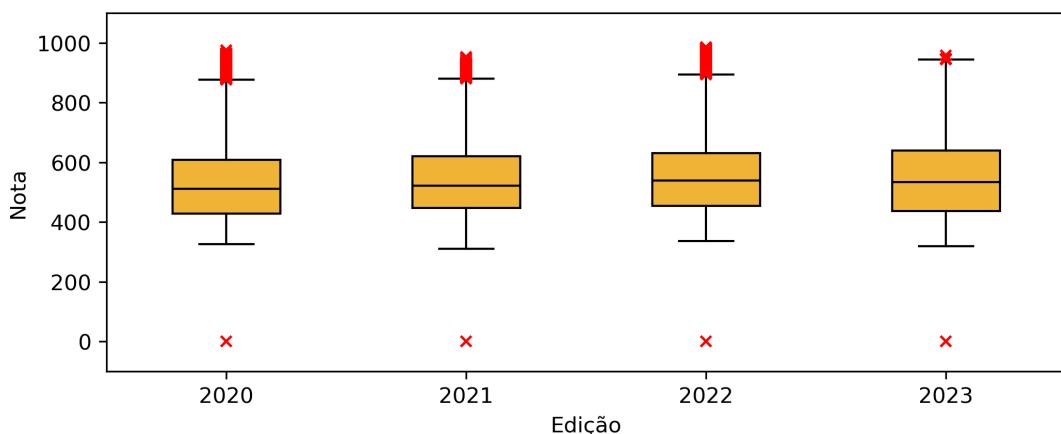
Fonte: elaborado pelo autor.

Figura 12 – Boxplot das notas - Linguagem



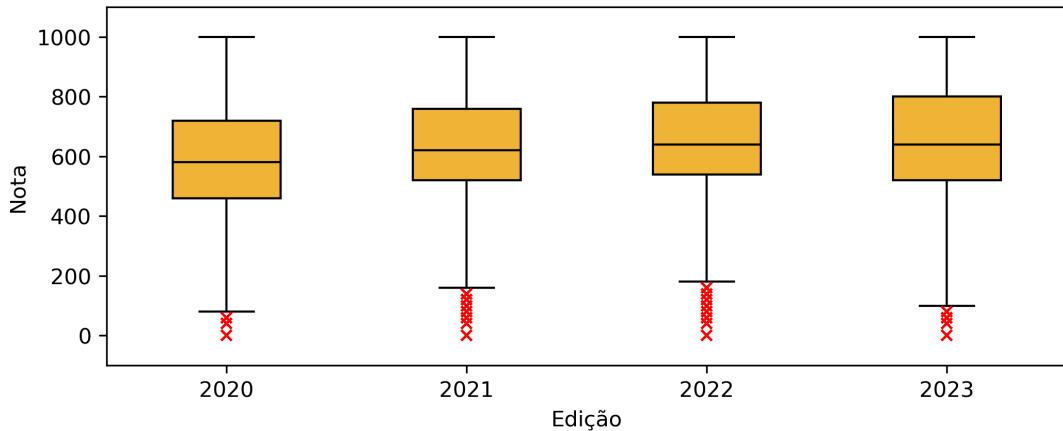
Fonte: elaborado pelo autor.

Figura 13 – Boxplot das notas - Matemática



Fonte: elaborado pelo autor.

Figura 14 – Boxplot das notas - Redação



Fonte: elaborado pelo autor.

Foi utilizado o critério do intervalo interquartil (*Interquartile Range - IQR*) para identificar os *outliers* nas notas de cada prova. Foram considerados *outliers* os valores que estavam abaixo de $Q1 - 1,5 \times IQR$ ou acima de $Q3 + 1,5 \times IQR$, onde $Q1$ é o primeiro quartil, $Q3$ é o terceiro quartil e $IQR = Q3 - Q1$. A Tabela 8 apresenta o limite inferior, o limite superior, a quantidade e o percentual de *outliers* identificados em cada conjunto de dados.

Tabela 8 – Quantidade e percentual de outliers nas notas

Variável	Limite Inferior	Limite Superior	Outliers
Humanas	267,95	780,35	19.596 (0,25%)
Natureza	265,40	723,80	49.017 (0,65%)
Linguagem	314,85	728,85	35.998 (0,46%)
Matemática	166,65	899,85	27.944 (0,37%)
Redação	160,00	1000,00	282.438 (3,58%)

Fonte: elaborado pelo autor.

Os intervalos dos *outliers* removeu as notas zero de todas as provas, o que é consistente com a análise realizada na Seção 4.4.1.1, onde vemos nos histogramas que as notas zero são valores atípicos.

Para a nota da Redação, o intervalo dos *outliers* manteve a nota máxima de 1.000 pontos. Porém, ao analisar a distribuição das notas de Redação e tendo o contexto do ENEM em mente, é razoável considerar as notas máximas como *outliers*, uma vez que são notas extremamente raras. Dessa forma, optou-se por retirar as observações com nota máxima do conjunto de dados da nota de Redação, o que resultou na remoção de mais 116 observações desse conjunto de dados.

4.4.2 Análise Exploratória - Variáveis Preditoras

A análise exploratória das variáveis preditoras seguiu três etapas: (i) concentração, (ii) correlação e (iii) *Permutation Importance*, conforme descrito na Seção 3.4.1.

4.4.2.1 Concentração

Ao calcular a proporção de observações para cada categoria das variáveis categóricas, foi possível identificar três variáveis que apresentavam uma concentração acima de 93%. Devido a essa alta concentração, foi decidido remover essas variáveis do conjunto de dados por não fornecerem informações relevantes para a modelagem preditiva. As Tabelas 9 a 13 apresentam as cinco variáveis categóricas com maior concentração e suas respectivas proporções da categoria de maior concentração para cada conjunto de dados.

Tabela 9 – Cinco maiores concentrações - Humanas

Variável	Maior Concentração
12_qtde_geladeira	92.8%
07_dias_trabalhador_domestico	90.2%
25_flag_internet	89.9%
15_qtde_maq_secar_roupa	86.5%
23_flag_telefone_fixo	84.1%

Fonte: elaborado pelo autor.

Tabela 10 – Cinco maiores concentrações - Natureza

Variável	Maior Concentração
12_qtde_geladeira	92.9%
07_dias_trabalhador_domestico	90.3%
25_flag_internet	89.9%
15_qtde_maq_secar_roupa	86.6%
23_flag_telefone_fixo	84.0%

Fonte: elaborado pelo autor.

Tabela 11 – Cinco maiores concentrações - Linguagem

Variável	Maior Concentração
12_qtde_geladeira	92.9%
07_dias_trabalhador_domestico	90.2%
25_flag_internet	89.9%
15_qtde_maq_secar_roupa	86.5%
23_flag_telefone_fixo	84.1%

Fonte: elaborado pelo autor.

Tabela 12 – Cinco maiores concentrações - Matemática

Variável	Maior Concentração
12_qtde_geladeira	92.8%
07_dias_trabalhador_domestico	90.2%
25_flag_internet	90.0%
15_qtde_maq_secar_roupa	86.5%
23_flag_telefone_fixo	84.0%

Fonte: elaborado pelo autor.

Tabela 13 – Cinco maiores concentrações - Redação

Variável	Maior Concentração
12_qtde_geladeira	92.8%
07_dias_trabalhador_domestico	90.1%
25_flag_internet	90.1%
15_qtde_maq_secar_roupa	86.4%
23_flag_telefone_fixo	83.9%

Fonte: elaborado pelo autor.

4.4.2.2 Correlação

A próxima etapa da análise exploratória das variáveis preditoras foi a análise de correlação utilizando a métrica *Phik* (23) (24). As Tabelas 14 a 18 apresentam as cinco variáveis com maior correlação *Phik* com a variável resposta em cada conjunto de dados.

Tabela 14 – Cinco maiores correlações *Phik* - Humanas

Variável	Correlação <i>Phik</i>
24_qtde_computadores	44,5%
03_ocupacao_pai	39,7%
04_ocupacao_mae	37,6%
08_qtde_banheiro	35,7%
18_flag_aspirador_po	35,4%

Fonte: elaborado pelo autor.

Tabela 15 – Cinco maiores correlações *Phik* - Natureza

Variável	Correlação <i>Phik</i>
24_qtde_computadores	44,6%
03_ocupacao_pai	40,2%
04_ocupacao_mae	38,0%
08_qtde_banheiro	37,4%
18_flag_aspirador_po	36,4%

Fonte: elaborado pelo autor.

Tabela 16 – Cinco maiores correlações *Phik* - Linguagem

Variável	Correlação <i>Phik</i>
24_qtde_computadores	44,1%
03_ocupacao_pai	41,7%
04_ocupacao_mae	39,9%
08_qtde_banheiro	36,8%
18_flag_aspirador_po	36,1%

Fonte: elaborado pelo autor.

Tabela 17 – Cinco maiores correlações *Phik* - Matemática

Variável	Correlação <i>Phik</i>
24_qtde_computadores	47,9%
03_ocupacao_pai	44,7%
04_ocupacao_mae	42,5%
08_qtde_banheiro	41,9%
18_flag_aspirador_po	40,5%

Fonte: elaborado pelo autor.

Tabela 18 – Cinco maiores correlações *Phik* - Redação

Variável	Correlação <i>Phik</i>
03_ocupacao_pai	35,9%
24_qtde_computadores	35,6%
04_ocupacao_mae	34,9%
08_qtde_banheiro	33,2%
10_qtde_carro	29,3%

Fonte: elaborado pelo autor.

Entre as cinco variáveis de maior correlação *Phik* com a variável resposta em cada conjunto de dados, foi observado que três variáveis estavam presentes em todos os conjuntos de dados: (i) quantidade de computadores, (ii) ocupação do pai e (iii) ocupação da mãe. Essas variáveis estão relacionadas a aspectos socioeconômicos dos participantes, o que é consistente com a literatura sobre o ENEM (1,8).

Ao analisar as matrizes de correlação *Phik* completas para cada conjunto de dados, foi possível identificar um par de variáveis que apresentavam uma correlação perfeita (*Phik* = 1.0): *flag* de treineiro e *status* da conclusão do ensino médio.

Devido a essa correlação perfeita, analisou-se a distribuição cruzada das categorias dessas duas variáveis, onde foi possível observar que 100% das observações da categoria "Treineiro" da variável *flag* de treineiro estavam associadas à categoria "Termina o ensino médio após o ano da prova" da variável *status* da conclusão do ensino médio.

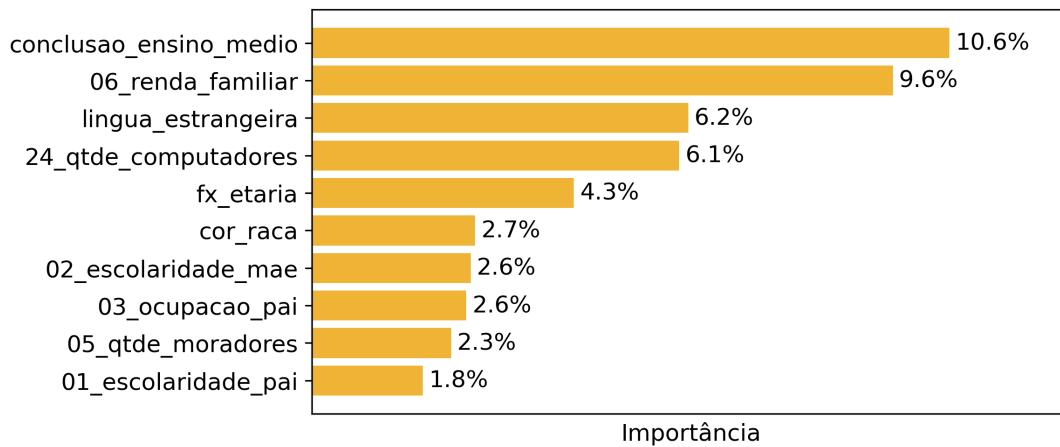
Diante disso, dado que a variável de *status* da conclusão do ensino médio apresenta mais categorias e, portanto, mais informações, foi decidido manter essa variável no conjunto de dados e remover a variável *flag* de treineiro.

4.4.2.3 Permutation Importance

Conforme descrito na Seção 3.4.1, a última etapa da análise exploratória das variáveis preditoras foi a análise de importância utilizando a métrica *Permutation Importance*.

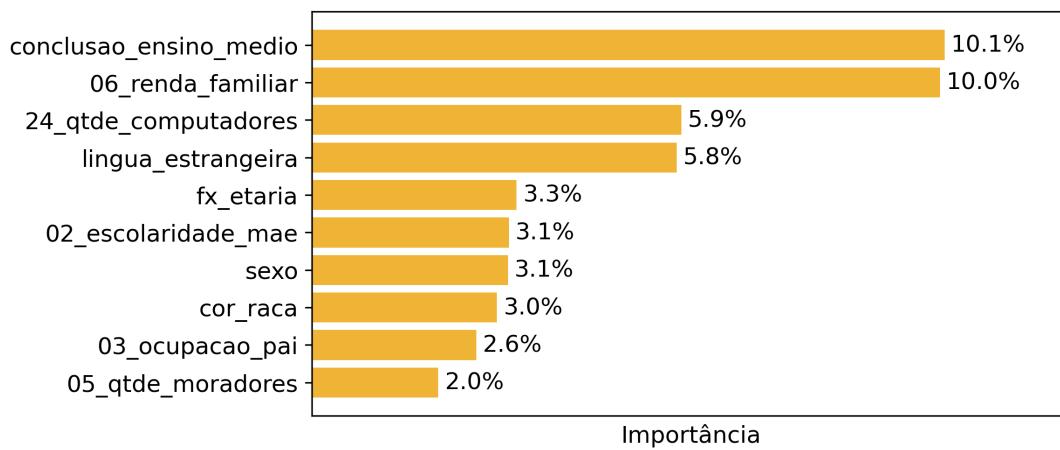
Realizadas as separações dos conjuntos de dados em treino e teste, foi treinado um modelo de *Random Forest Regressor* em cada conjunto de dados para em seguida calcular o *Permutation Importance* de cada variável preditora. As Figuras 15 a 19 apresentam os gráficos de importância das dez variáveis mais importantes para cada conjunto de dados.

Figura 15 – Dez maiores *Permutation Importance* - Humanas



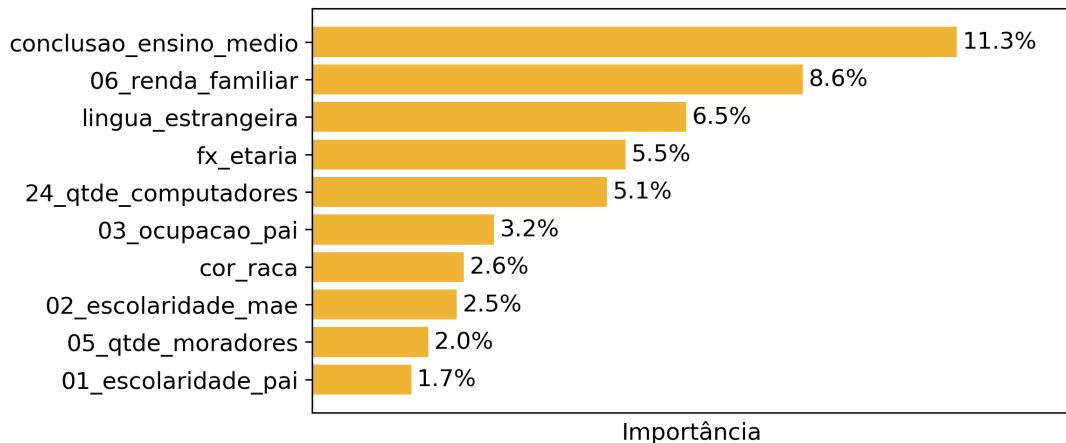
Fonte: elaborado pelo autor.

Figura 16 – Dez maiores *Permutation Importance* - Natureza



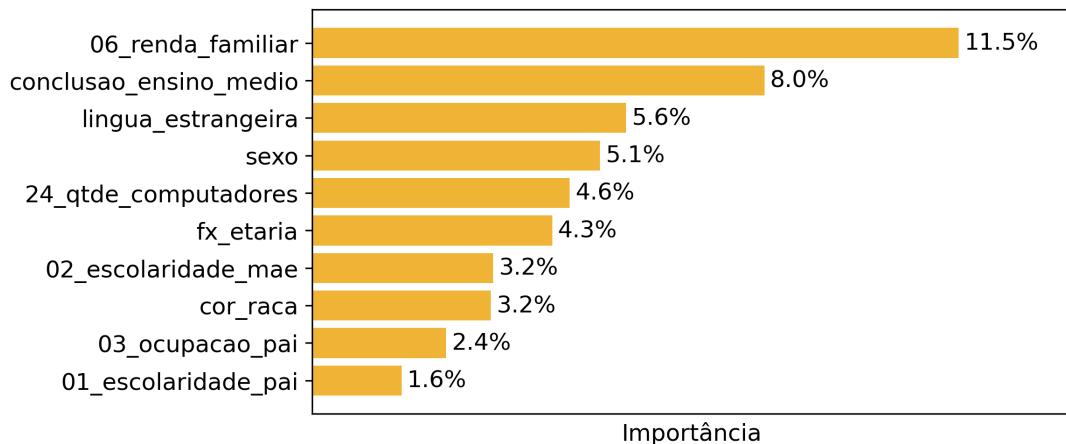
Fonte: elaborado pelo autor.

Figura 17 – Dez maiores *Permutation Importance* - Linguagem



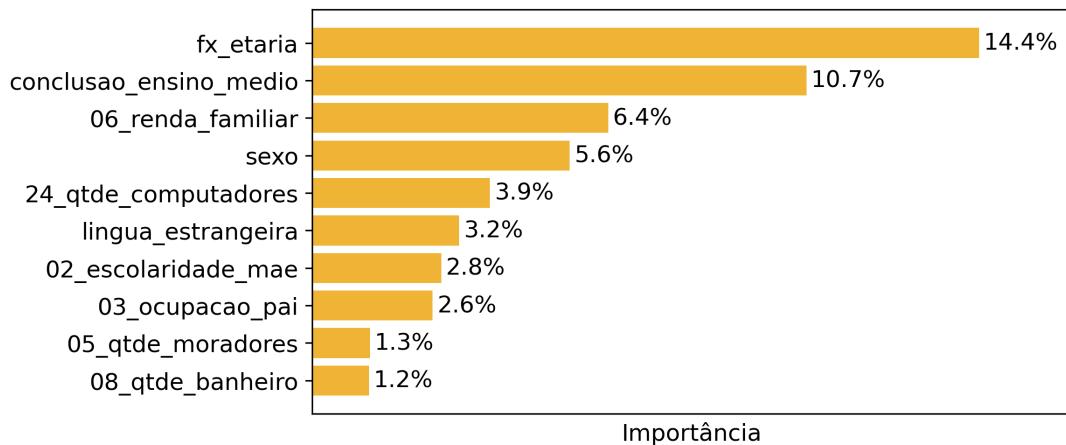
Fonte: elaborado pelo autor.

Figura 18 – Dez maiores *Permutation Importance* - Matemática



Fonte: elaborado pelo autor.

Figura 19 – Dez maiores *Permutation Importance* - Redação



Fonte: elaborado pelo autor.

4.4.2.4 Seleção de Variáveis

Até o momento já foram aplicados dois critérios para a seleção das variáveis preditoras: (i) concentração, removendo variáveis com concentração de categoria maior que 93% e (ii) correlação perfeita com outra variável preditora, resultando em quatro variáveis removidas do conjunto de dados: nacionalidade, quantidade de máquina de lavar louça, estado civil e *flag* de treineiro.

Com as informações de correlação *Phik* e *Permutation Importance*, foram aplicados mais dois critérios para a seleção das variáveis preditoras: (i) correlação baixa com a variável resposta ($Phik < 0,05\%$) e (ii) um critério duplo de correlação alta ($Phik > 85\%$) com outras variáveis preditoras e menor *Permutation Importance* entre as variáveis correlacionadas.

No critério de baixa correlação com a variável resposta, nenhuma variável preditora foi removida, uma vez que todas apresentaram correlação *Phik* maior que 0,05%, sendo o menor valor, entre todos os conjuntos de dados, de 3,64% da variável de quantidade de motocicletas com a nota da Redação.

No critério duplo, foi apresentado apenas uma dupla de variáveis com correlação alta ($Phik > 85\%$) em cada conjunto de dados: *status* da conclusão do ensino médio e a faixa etária do participante. Apenas no conjunto de dados da Redação, a variável de faixa etária apresentou maior *Permutation Importance* e então seria mantida no conjunto de dados, enquanto a variável de *status* da conclusão do ensino médio seria removida.

Porém, considerando a concentração cruzada das duas variáveis (apresentada na Tabela 19 para o conjunto de dados da Redação), foi decidido manter ambas as variáveis.

No critério duplo, foi apresentado apenas uma dupla de variáveis com correlação alta ($Phik > 85\%$) em cada conjunto de dados: status da conclusão do ensino médio e a faixa etária do participante. Apenas no conjunto de dados da Redação, a variável de faixa etária apresentou maior *Permutation Importance* e então seria mantida no conjunto de dados, enquanto a variável de *status* da conclusão do ensino médio seria removida.

Tabela 19 – Cinco maiores correlações *Phik* - Redação

Faixa etária	Concluiu	Conclui agora	Vai concluir	Não / Nem
menor de 17 anos	-	0.4%	10.1%	0.2%
17 anos	0.3%	15.9%	6.6%	-
18 anos	6.3%	16.0%	0.6%	-
19 anos	8.8%	2.4%	0.1%	-
20 anos	6.2%	0.6%	-	-
21 anos	4.3%	0.2%	-	-

Continua na próxima página...

Faixa etária	Concluiu	Conclui agora	Vai concluir	Não / Nem
22 anos	3.1%	0.1%	-	-
23 anos	2.4%	-	-	-
24 anos	1.9%	-	-	-
25 anos	1.5%	-	-	-
26 a 30 anos	4.5%	0.1%	-	-
31 a 35 anos	2.4%	-	-	-
36 a 40 anos	1.7%	-	-	-
41 a 45 anos	1.2%	-	-	-
46 a 50 anos	0.7%	-	-	-
51 a 55 anos	0.4%	-	-	-
56 a 60 anos	0.2%	-	-	-
61 a 65 anos	0.1%	-	-	-
66 a 70 anos	-	-	-	-
maior de 70 anos	-	-	-	-

Fonte: elaborado pelo autor.

Na tabela acima, as categorias da variável de status da conclusão do ensino médio foram renomeadas da seguinte maneira: (i) Já concluiu o ensino médio: “Concluiu”; (ii) Vai concluir o ensino médio no ano da prova: “Conclui agora”; (iii) Vai concluir no ano seguinte à prova: “Vai concluir”; e (iv) Não cursa e nem concluiu o ensino médio: “Não / Nem”.

4.5 Treinamento dos Modelos

4.5.1 Ajuste dos Hiperparâmetros

A primeira etapa do treinamento dos modelos foi o ajuste dos hiperparâmetros utilizando a técnica de *Grid Search*. Inicialmente, foi utilizado o método `GridSearchCV` da biblioteca `scikit-learn` (26) para realizar o ajuste dos hiperparâmetros dos modelos. Porém, a execução do código foi interrompida subitamente algumas vezes, possivelmente devido a alto consumo de memória da GPU ou de memória RAM.

Assim, para contornar esse problema, o *Grid Search* foi implementado manualmente. Foi estabelecido um dicionário com os hiperparâmetros e seus respectivos valores a serem testados para cada modelo e gerada uma lista com todas as combinações possíveis desses hiperparâmetros.

Em seguida, através de um *loop*, cada combinação de hiperparâmetros foi utilizada para instanciar cada modelo, treinar o modelo com os dados de treino e avaliar o desempenho do modelo com os dados de validação.

Para gerar o conjunto de dados de validação, foi feita uma separação no conjunto de dados de treino de forma que o conjunto de validação tenha 10% dos dados originais, considerando o conjunto de teste que já foi separado. O desempenho do modelo foi avaliado utilizando a métrica da Raiz Quadrada do Erro Quadrático Médio (*Root Mean Squared Error* - RMSE).

Para cada algoritmo, foi estabelecido um conjunto de hiperparâmetros e seus respectivos valores a serem testados, assim como hiperparâmetros fixos em valores pré-definidos. As Tabelas 20 a 22 apresentam os hiperparâmetros e seus respectivos valores a serem testados para cada modelo, bem como os hiperparâmetros fixados em valores pré-definidos.

Tabela 20 – *Grid Search - XGBoost*

Hiperparâmetro	Valores
learning_rate	[0.05, 0.10, 0.20]
max_depth	[6, 8, 10]
min_child_weight	[1, 5, 10]
colsample_bytree	[0.70, 0.85, 1.0]
subsample	[0.70, 0.85, 1.0]
n_estimators	100
objective	"reg:squarederror"
tree_method	"hist"
device	"cuda"
eval_metric	"rmse"

Fonte: elaborado pelo autor.

Tabela 21 – *Grid Search - LightGBM*

Hiperparâmetro	Valores
num_leaves	[31, 63, 127]
learning_rate	[0.05, 0.10, 0.20]
min_child_samples	[20, 50, 100]
colsample_bytree	[0.70, 0.85, 1.0]
subsample	[0.70, 0.85, 1.0]
n_estimators	100
objective	"regression"
metric	"rmse"
device	"cpu"

Continua na próxima página...

Hiperparâmetro	Valores
n_jobs	-1

Fonte: elaborado pelo autor.

Tabela 22 – *Grid Search - Random Forest*

Hiperparâmetro	Valores
max_depth	[10, 15, 20]
max_features	[0.7, 0.9, 1.0]
max_samples	[0.8, 0.9, 1.0]
split_criterion	"mse"
bootstrap	True
n_bins	256
min_samples_leaf	15
n_streams	4
n_estimators	100

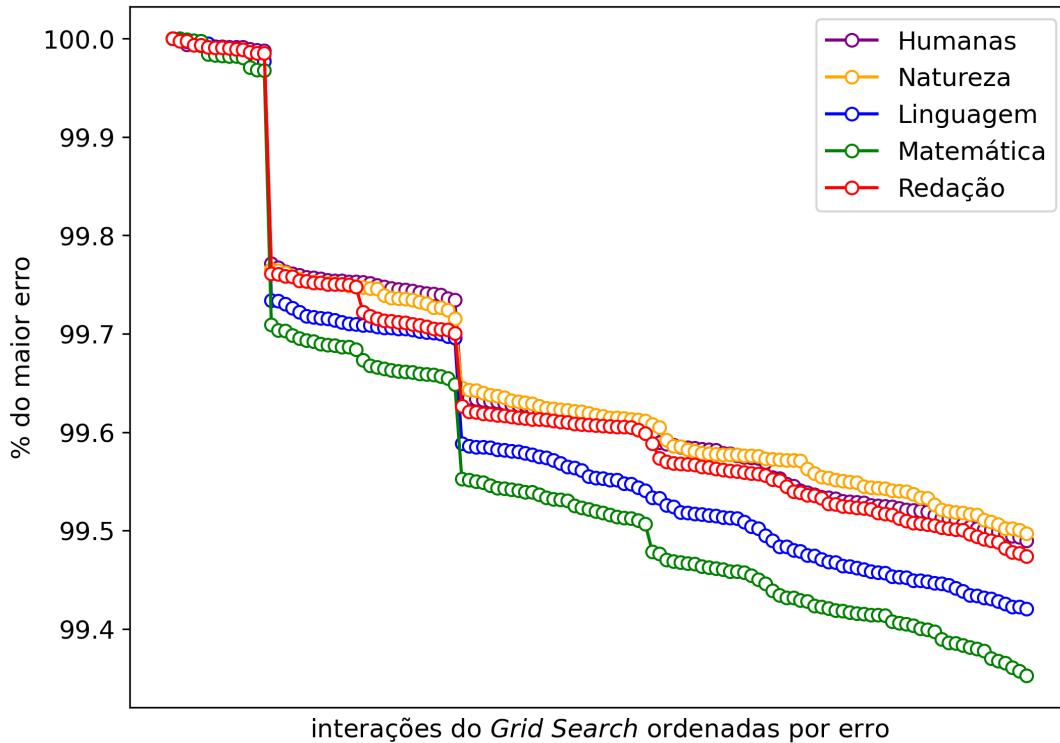
Fonte: elaborado pelo autor.

Foram treinados 243 combinações de hiperparâmetros para os algoritmos de *XG-Boost* e *LightGBM* (5 hiperparâmetros com 3 valores cada) e 27 combinações de hiperparâmetros para o algoritmo de *Random Forest* (3 hiperparâmetros com 3 valores cada), totalizando 270 modelos treinados para cada conjunto de dados (Ciências Humanas, Ciências da Natureza, Linguagem e Código, Matemática e Redação) e 1.350 modelos treinados no total, o que levou aproximadamente 5 horas para ser executado.

As Figuras 20 a 22 apresentam o erro RMSE para as combinações de hiperparâmetros testadas para cada modelo e conjunto de dados, com as iterações do *Grid Search* ordenadas em ordem decrescente de erro.

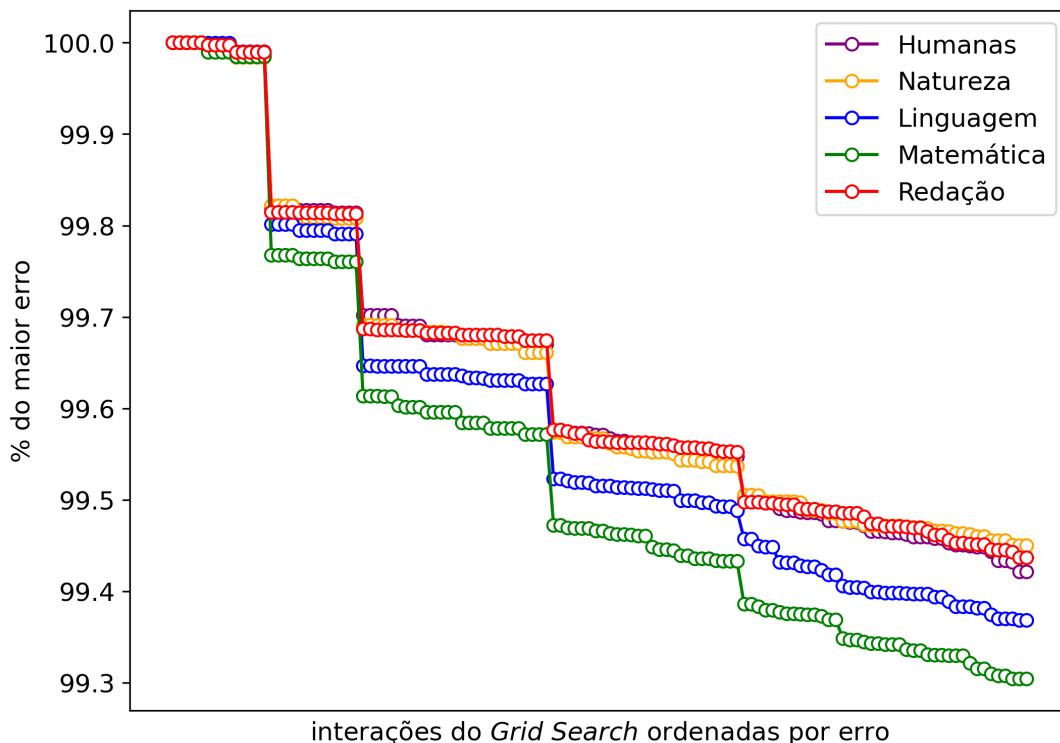
Para melhor visualização, os valores do erro foram relativizados ao maior erro RMSE encontrado para cada modelo e conjunto de dados, ou seja, o maior erro RMSE encontrado para cada modelo e conjunto de dados foi considerado como 100% e os demais erros foram relativizados em relação a esse valor.

Figura 20 – Erro do *Grid Search - XGBoost*



Fonte: elaborado pelo autor. Obs.: apenas metade dos dados foi plotada para melhor visualização.

Figura 21 – Erro do *Grid Search - LightGBM*



Fonte: elaborado pelo autor. Obs.: apenas metade dos dados foi plotada para melhor visualização.

Figura 22 – Erro do *Grid Search - Random Forest*

Fonte: elaborado pelo autor.

A redução no erro RMSE para as melhores combinações de hiperparâmetros acabou não sendo tão significativa para os modelos de *XGBoost* e *LightGBM*, apresentando uma redução menor que 1% em relação ao maior erro RMSE encontrado na *Grid Search*. Já para o modelo de *Random Forest*, a redução no erro RMSE foi mais significativa, apresentando uma redução de aproximadamente 2,2% em relação ao maior erro RMSE encontrado na *Grid Search*.

Neste ajuste, não foi selecionado diferentes valores para o hiperparâmetro `n_estimators` (número de estimadores), uma vez que a execução do código para as combinações com esse parâmetro maior que 100 foi interrompida subitamente algumas vezes, também possivelmente devido ao alto consumo de memória da GPU/RAM. Assim, o ajuste desse hiperparâmetro foi realizado na próxima etapa, juntamente com o treinamento final dos modelos.

Foram escolhidos os melhores hiperparâmetros para cada modelo e conjunto de dados baseados no menor erro RMSE encontrado na *Grid Search*. As Tabelas 23 a 25 apresentam os melhores hiperparâmetros encontrados para cada modelo em cada conjunto de dados, bem como o erro RMSE correspondente a cada combinação de hiperparâmetros.

Tabela 23 – Hiperparâmetros Ajustados - *XGBoost*

Hiperparâmetro	Humanas	Natureza	Linguagem	Matemática	Redação
learning_rate	0.1	0.1	0.1	0.1	0.1
max_depth	10.0	10.0	10.0	10.0	10.0
min_child_weight	10.0	10.0	1.0	5.0	5.0
colsample_bytree	0.7	0.7	0.7	0.7	0.7
subsample	1.0	1.0	1.0	1.0	1.0
RMSE	75,4	65,1	63,1	96,0	148,1

Fonte: elaborado pelo autor.

Tabela 24 – Hiperparâmetros Ajustados - *LightGBM*

Hiperparâmetro	Humanas	Natureza	Linguagem	Matemática	Redação
num_leaves	127.0	127.0	127.0	127.0	127.0
learning_rate	0.2	0.2	0.2	0.2	0.2
min_child_samples	100.0	50.0	50.0	100.0	100.0
colsample_bytree	0.7	0.7	0.85	0.7	0.7
subsample	1.0	0.7	1.0	1.0	1.0
RMSE	75,4	65,1	63,1	96,0	148,2

Fonte: elaborado pelo autor.

Tabela 25 – Hiperparâmetros Ajustados - *Random Forest*

Hiperparâmetro	Humanas	Natureza	Linguagem	Matemática	Redação
max_depth	20.0	20.0	20.0	20.0	20.0
max_features	0.7	0.7	0.7	0.7	0.7
max_samples	0.8	0.8	0.8	0.8	0.8
RMSE	75,6	65,4	63,3	96,4	148,6

Fonte: elaborado pelo autor.

4.5.2 Treinamento final dos modelos

Conforme descrito na Seção 3.5, o ajuste do hiperparâmetro `n_estimators` (número de estimadores) foi realizado nesta etapa, juntamente com o treinamento final dos modelos.

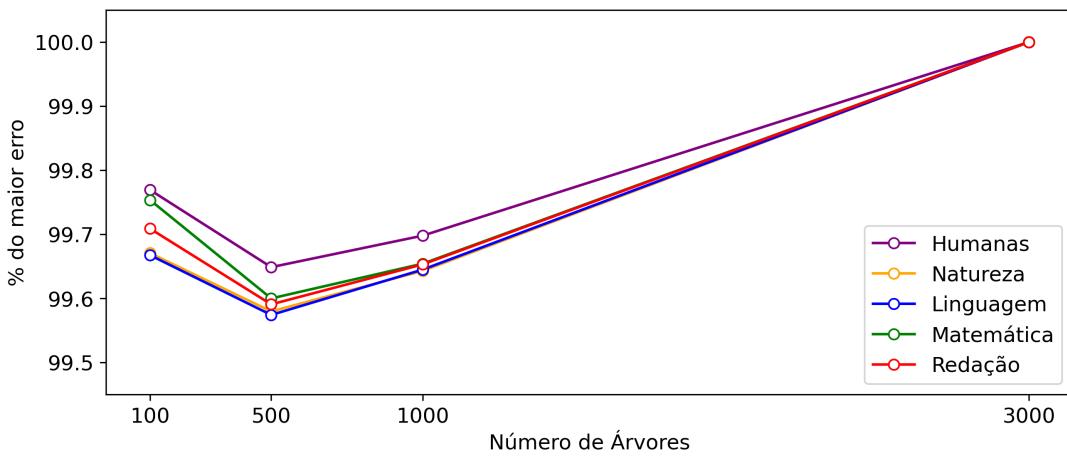
Para os modelos de *XGBoost* e *LightGBM*, foram testados os valores de 100, 500, 1.000 e 3.000 estimadores. Para o modelo de *Random Forest*, só foi possível treinar o modelo com 100 e 500 estimadores, uma vez que a execução do código para as combinações

com 1.000 e 3.000 estimadores foi interrompida subitamente algumas vezes, possivelmente devido ao alto consumo de memória da GPU ou memória RAM.

As Figuras 23 e 24 apresentam o erro MAPE do conjunto de teste, relativo ao maior erro MAPE encontrado para cada modelo e conjunto de dados, para os modelos de *XGBoost* e *LightGBM* com os melhores hiperparâmetros encontrados na etapa de ajuste dos hiperparâmetros, considerando o número de estimadores como 100, 500, 1.000 e 3.000.

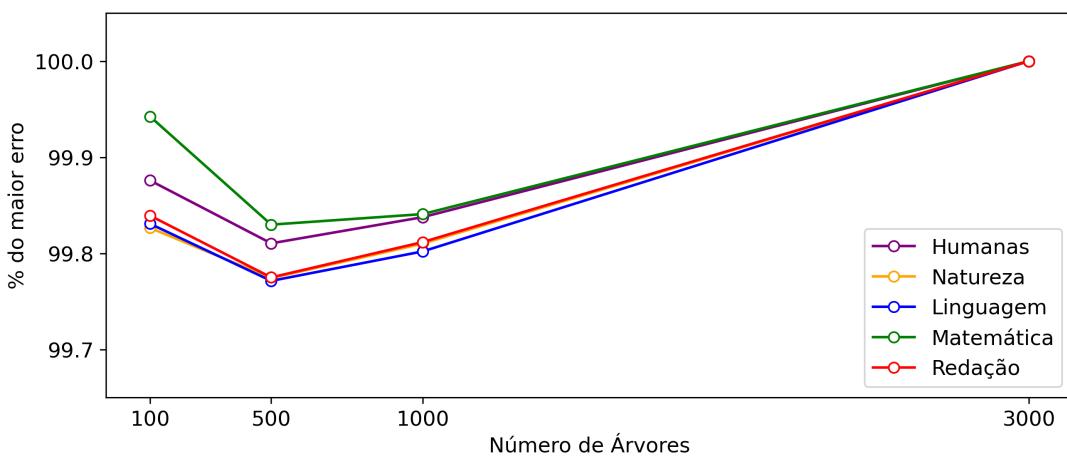
A Tabela 26 apresenta o erro MAPE do conjunto de teste para o modelo de *Random Forest* com os melhores hiperparâmetros encontrados na etapa de ajuste dos hiperparâmetros, considerando o número de estimadores como 100 e 500.

Figura 23 – Erro MAPE - *XGBoost*



Fonte: elaborado pelo autor.

Figura 24 – Erro MAPE - *LightGBM*



Fonte: elaborado pelo autor.

Tabela 26 – Erro MAPE - *Random Forest*

Área	100 estimadores	500 estimadores
Ciências Humanas	12.369%	12.364%
Ciências Natureza	10.924%	10.920%
Linguagem e Código	10.264%	10.260%
Matemática	15.160%	15.154%
Redação	21.327%	21.320%

Fonte: elaborado pelo autor.

Assim como aconteceu no ajuste dos demais hiperparâmetros, a redução no erro MAPE para diferentes números de estimadores acabou não sendo tão significativa para os três modelos. Para o modelo de *XGBoost*, a maior redução no erro MAPE foi de 0,4258% na prova de Linguagem e Código. Para o modelo de *LightGBM*, a maior redução no erro MAPE foi de 0,2286% na prova de Linguagem e Código. Para o modelo de *Random Forest*, a maior redução no erro MAPE foi de 0,0068% na prova de Redação.

4.5.3 Construção dos modelos de *ensemble*

A partir dos três modelos treinados, foram construídos dois modelos de *ensemble* usando a técnica de *bagging*: (i) um modelo de *ensemble* utilizando os modelos de *XGBoost* e *LightGBM* e (ii) um modelo de *ensemble* utilizando os modelos de *XGBoost*, *LightGBM* e *Random Forest*.

Para ambos os modelos de *ensemble*, foi utilizada a média aritmética das previsões dos modelos individuais para obter a previsão final do modelo de *ensemble*.

O *ensemble* utilizando o modelo de *Random Forest* foi construído considerando a média das previsões do modelo de acordo com a quantidade de modelos disponíveis. Ou seja, como o modelo de *Random Forest* não foi treinado com 1.000 e 3.000 estimadores, a média do *ensemble* foi calculada apenas com os modelos de *XGBoost* e *LightGBM* para esses números de estimadores, e com os três modelos para os números de estimadores de 100 e 500.

As Figuras 25 e 26 apresentam o erro MAPE do conjunto de teste, relativo ao maior erro MAPE encontrado para cada modelo e conjunto de dados, para os modelos de *ensemble* construídos.

Figura 25 – Erro MAPE - *Ensemble (XGBoost + LightGBM)*

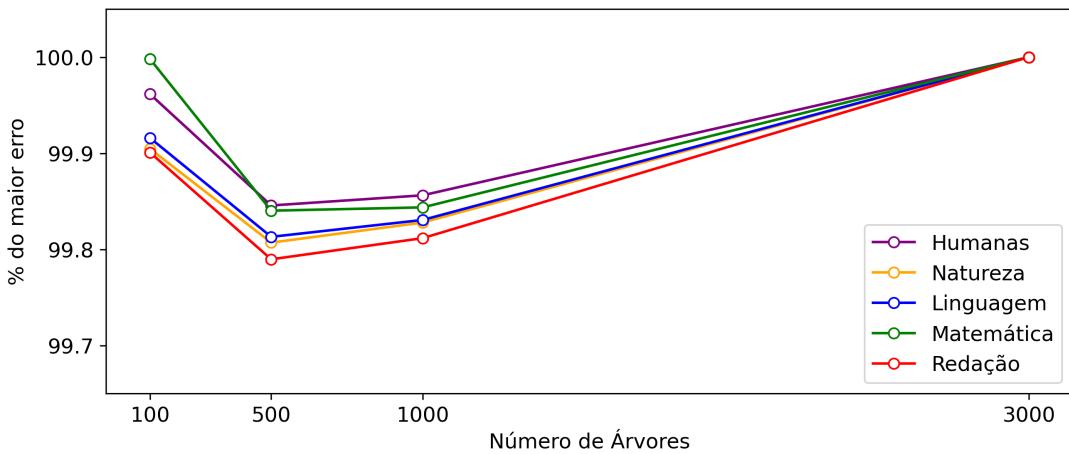
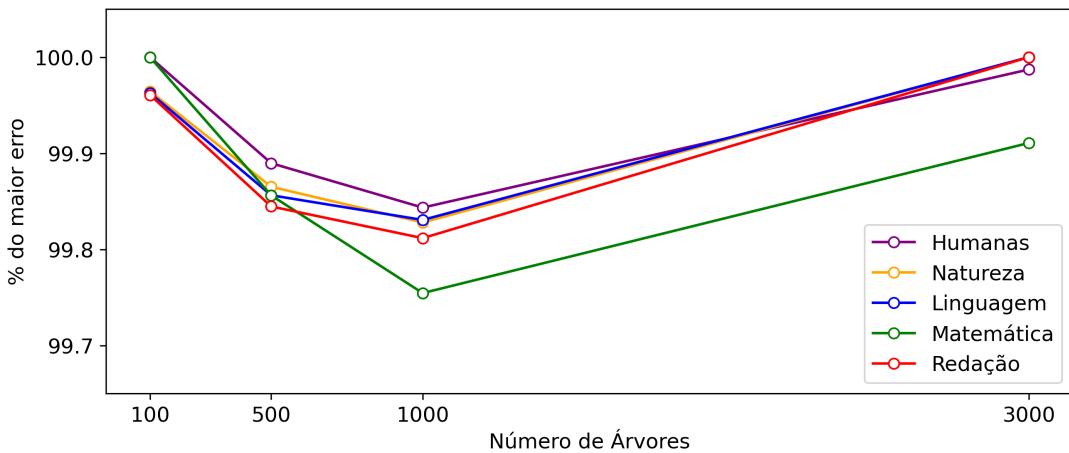


Figura 26 – Erro MAPE - *Ensemble (XGBoost + LightGBM + Random Forest)*



Fonte: elaborado pelo autor.

Novamente, a redução no erro MAPE para os modelos de *ensemble* acabou não sendo tão significativa para os diferentes números de estimadores. Para o modelo de *ensemble* utilizando os modelos de *XGBoost* e *LightGBM*, a maior redução no erro MAPE foi de 0,2103% na prova de Redação. Para o modelo de *ensemble* utilizando os modelos de *XGBoost*, *LightGBM* e *Random Forest*, a maior redução no erro MAPE foi de 0,1883% também na prova de Redação.

4.5.4 Avaliação dos modelos

Nessa etapa, a primeira análise realizada foi sobre um possível *overfitting* dos modelos. Conforme o critério adotado e explicado na Seção 3.6, nenhum dos modelos apresentou *overfitting* segundo esse critério, ou seja, nenhum dos modelos apresentou um erro do conjunto de teste 15% maior que o erro do conjunto de treino.

Em seguida, foi feita a comparação dos modelos entre si para cada conjunto de dados, a fim de se encontrar o melhor modelo para cada conjunto de dados. Para isso, foi adotado o critério do menor erro MAPE do conjunto de teste.

As Tabelas 27 a 31 apresentam os cinco modelos que apresentaram os menores erros MAPE do conjunto de teste para cada conjunto de dados, ordenados do menor para o maior erro MAPE.

Tabela 27 – Cinco melhores modelos - Humanas

Modelo	Qtd. estimadores	MAPE teste
<i>XGB + LGBM</i>	500	12,302290%
<i>XGB + LGBM</i>	1.000	12,303594%
<i>XGB</i>	500	12,306396%
<i>XGB + LGBM + RF</i>	500	12,309288%
<i>XGB</i>	1.000	12,312470%

Fonte: elaborado pelo autor.

Tabela 28 – Cinco melhores modelos - Natureza

Modelo	Qtd. estimadores	MAPE teste
<i>XGB + LGBM</i>	500	10.862886%
<i>XGB + LGBM</i>	1.000	10.865162%
<i>XGB</i>	500	10.867356%
<i>XGB + LGBM + RF</i>	500	10.869213%
<i>LGBM</i>	500	10.872770%

Fonte: elaborado pelo autor.

Tabela 29 – Cinco melhores modelos - Linguagem

Modelo	Qtd. estimadores	MAPE teste
<i>XGB + LGBM</i>	500	10.205704%
<i>XGB + LGBM</i>	1.000	10.207500%
<i>XGBoost</i>	500	10.209998%
<i>XGB + LGBM + RF</i>	500	10.210127%
<i>XGB + LGBM</i>	100	10.216225%

Fonte: elaborado pelo autor.

Tabela 30 – Cinco melhores modelos - Matemática

Modelo	Qtd. estimadores	MAPE teste
<i>XGB + LGBM</i>	500	15.045981%
<i>XGB + LGBM</i>	1.000	15.046495%
<i>XGBoost</i>	500	15.052611%
<i>LightGBM</i>	500	15.059630%
<i>XGBoost</i>	1.000	15.060808%

Fonte: elaborado pelo autor.

Tabela 31 – Cinco melhores modelos - Redação

Modelo	Qtd. estimadores	MAPE teste
<i>XGB + LGBM</i>	500	21.208775%
<i>XGB + LGBM</i>	1.000	21.213456%
<i>XGBoost</i>	500	21.214043%
<i>LightGBM</i>	500	21.219782%
<i>XGB + LGBM + RF</i>	500	21.220464%

Fonte: elaborado pelo autor.

Com base nessas tabelas, os modelos finais escolhidos para cada conjunto de dados foram:

- **Humanas:** modelo de *ensemble* utilizando os modelos de *XGBoost* e *LightGBM* com 500 estimadores;
- **Natureza:** modelo de *ensemble* utilizando os modelos de *XGBoost* e *LightGBM* com 500 estimadores;
- **Linguagem e Código:** modelo de *ensemble* utilizando os modelos de *XGBoost* e *LightGBM* com 500 estimadores;
- **Matemática:** modelo de *ensemble* utilizando os modelos de *XGBoost* e *LightGBM* com 500 estimadores;
- **Redação:** modelo de *ensemble* utilizando os modelos de *XGBoost* e *LightGBM* com 500 estimadores.

Essa escolha vai em linha com o apresentado até o momento, onde os modelos individuais apresentaram o menor erro MAPE com 500 estimadores e os modelos de *ensemble* apresentaram o menor erro MAPE em relação aos modelos individuais.

4.6 Influência das Variáveis Preditoras

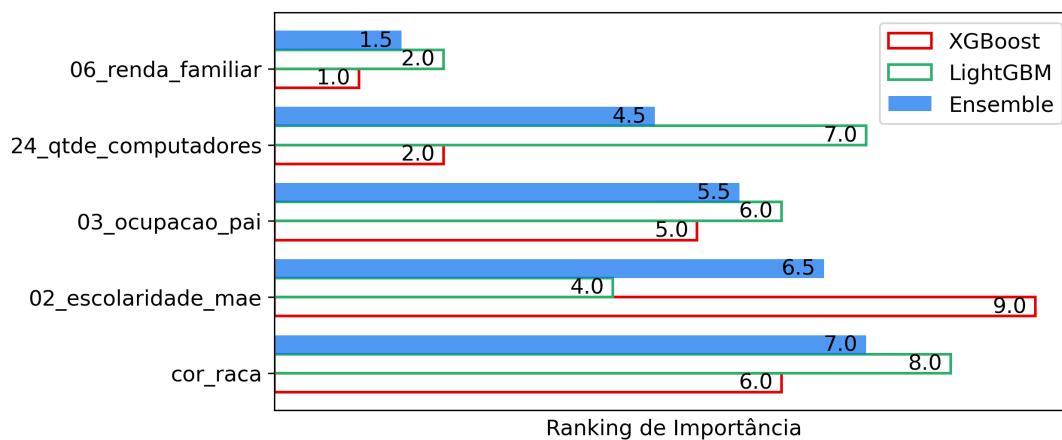
4.6.1 Importância

Com o modelo final em mãos, foi possível analisar a importância das variáveis preditoras para cada modelo. Para isso, foi utilizado o método `feature_importances_` dos algoritmos para se extrair a importância de cada variável preditora para cada modelo.

Como o modelo final é um modelo de *ensemble*, a importância de cada variável preditora para o modelo de *ensemble* foi calculada usando a média aritmética do ranking da importância de cada variável preditora para os modelos de *XGBoost* e *LightGBM*.

As Figuras 27 a 31 apresentam a importância das variáveis preditoras para cada modelo final, ordenadas da maior para a menor importância.

Figura 27 – Rank de Importância - Humanas



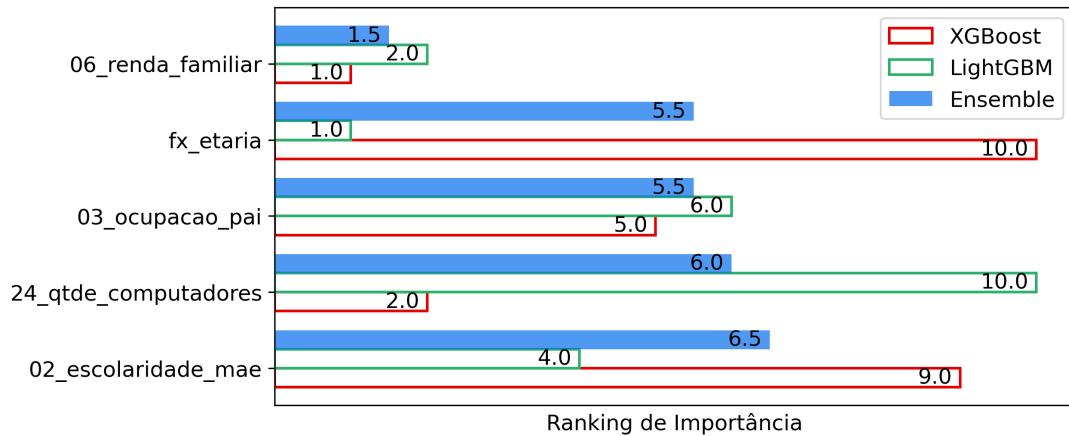
Fonte: elaborado pelo autor.

Figura 28 – Rank de Importância - Natureza



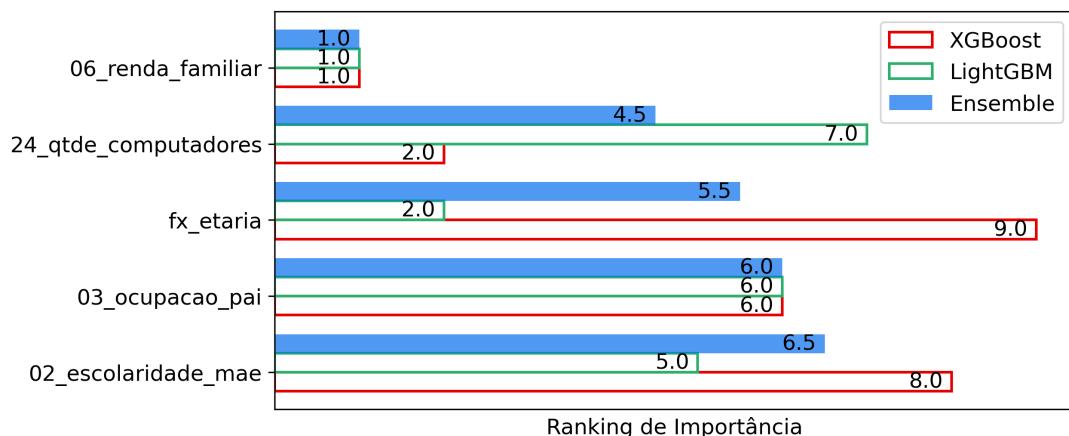
Fonte: elaborado pelo autor.

Figura 29 – Rank de Importância - Linguagem



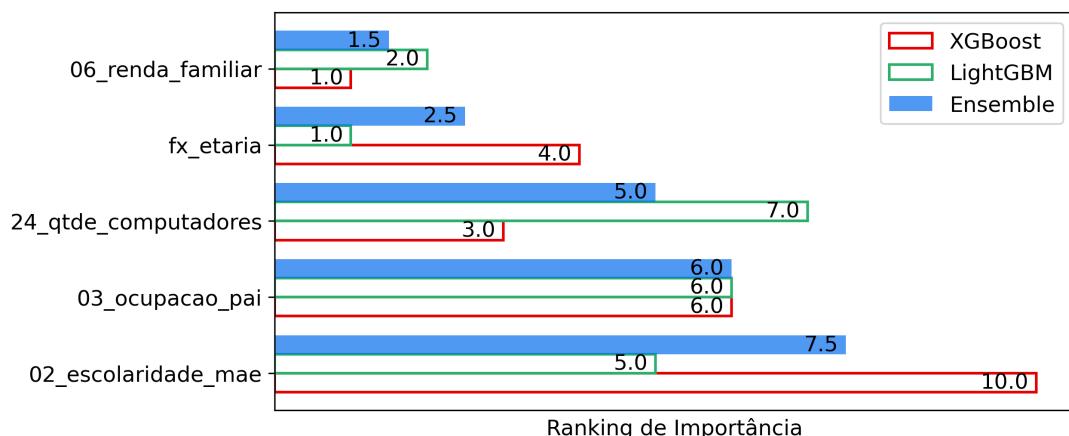
Fonte: elaborado pelo autor.

Figura 30 – Rank de Importância - Matemática



Fonte: elaborado pelo autor.

Figura 31 – Rank de Importância - Redação



Fonte: elaborado pelo autor.

As variáveis preditoras de renda familiar, quantidade de computadores em casa, ocupação do pai e escolaridade da mãe apareceram entre as 5 variáveis mais importantes para os cinco modelos finais. A variável da faixa etária do estudante apareceu para os modelos finais de Linguagem e Código, Matemática e Redação, mas não apareceu para os demais. A quinta variável para o modelo final de Ciências da Natureza foi a variável de escolaridade do pai e para o modelo final de Ciências Humanas foi a variável de cor/raça do estudante.

4.6.2 Sensibilidade das Variáveis Respostas

Para se analisar a sensibilidade das variáveis respostas em relação às alterações nas variáveis preditoras, foi criada uma base sintética de dados, onde oito variáveis preditoras foram preenchidas com todos os seus possíveis valores e as demais variáveis preditoras foram preenchidas com os seus valores mais frequentes do conjunto de treino. As variáveis preditoras selecionadas foram:

- `fx_etaria`: faixa etária do estudante;
- `sexo`: sexo do estudante;
- `cor_raca`: cor/raça do estudante;
- `01_escolaridade_pai`: escolaridade do pai do estudante;
- `02_escolaridade_mae`: escolaridade da mãe do estudante;
- `03_ocupacao_pai`: ocupação do pai do estudante;
- `04_ocupacao_mae`: ocupação da mãe do estudante;
- `06_renda_familiar`: renda familiar do estudante.

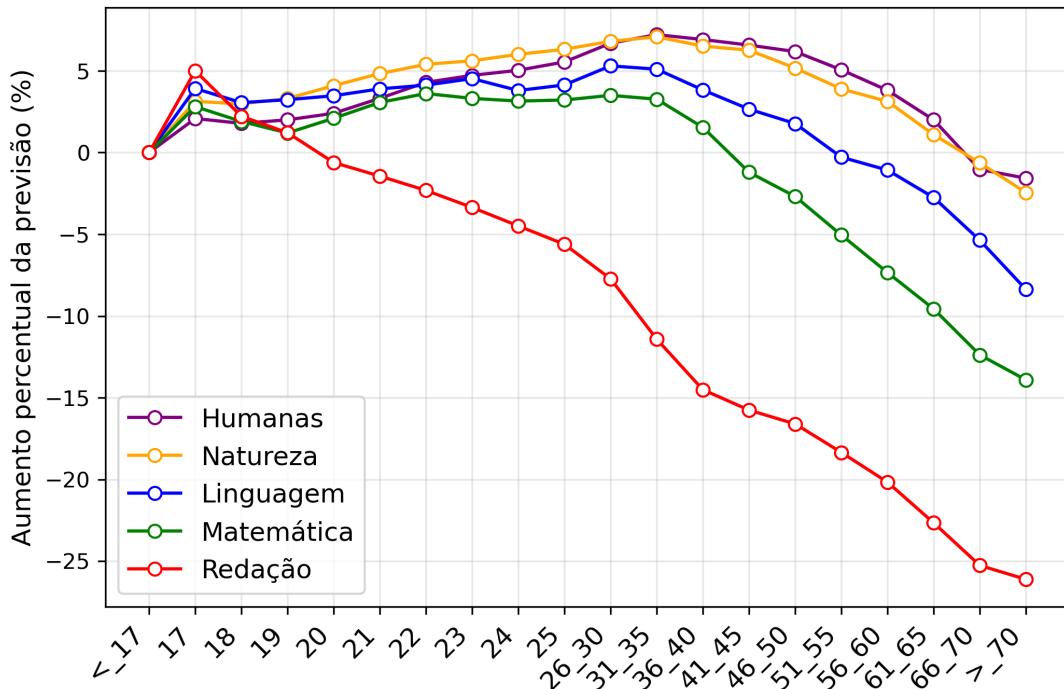
Foram selecionadas essas oito variáveis preditoras para a análise de sensibilidade por serem de interesse para a análise e por apresentarem uma quantidade razoável de valores distintos, o que possibilita uma análise mais detalhada da sensibilidade das variáveis respostas em relação às alterações nessas variáveis preditoras. A base sintética criada possui 4.165.000 registros,

As Figuras 32 a 39 apresentam as curvas de sensibilidade de cada variável resposta em relação às alterações nas variáveis preditoras selecionadas, onde cada curva representa a média das previsões do modelo final para cada valor da variável preditora, mantendo as demais variáveis preditoras fixas.

As curvas de sensibilidade foram normalizadas ao primeiro ponto da curva, ou seja, o valor da curva para o primeiro ponto da variável preditora foi definido como 0 e os

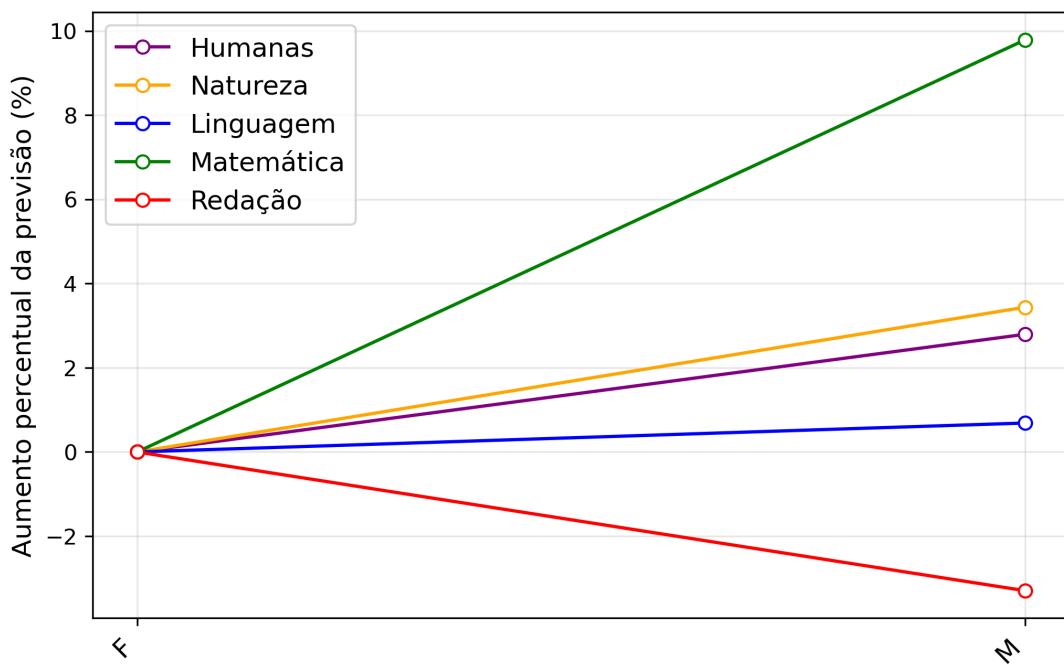
demais pontos foram calculados como o aumento percentual da curva para cada ponto em relação ao primeiro ponto.

Figura 32 – Curva de Sensibilidade - Faixa Etária



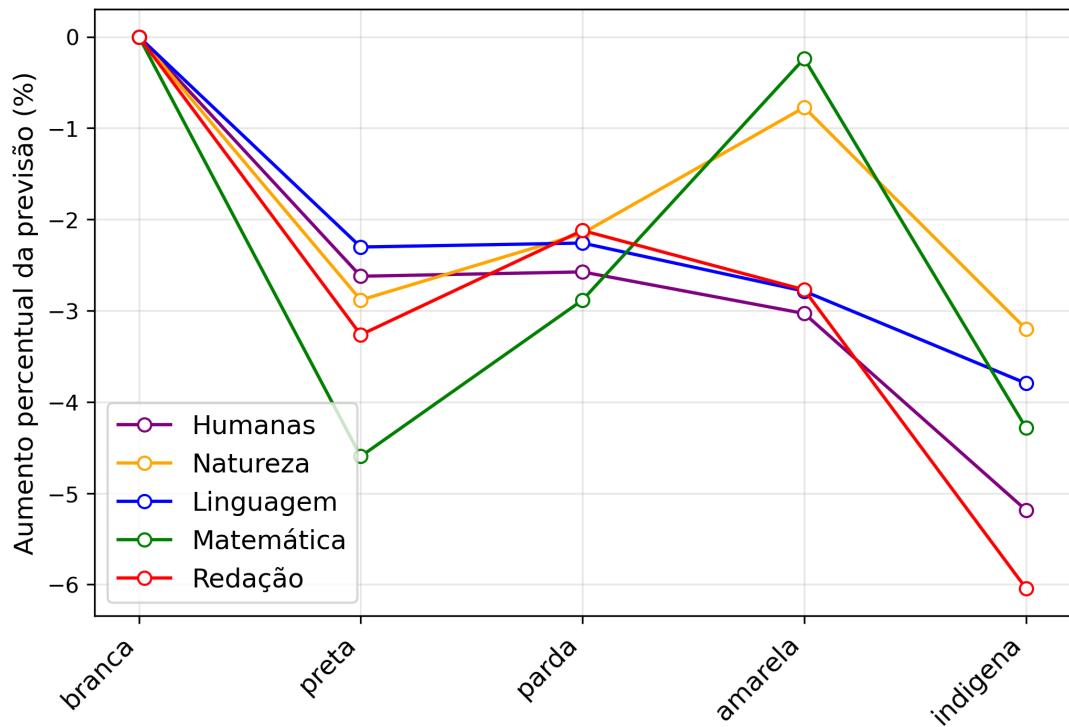
Fonte: elaborado pelo autor.

Figura 33 – Curva de Sensibilidade - Sexo



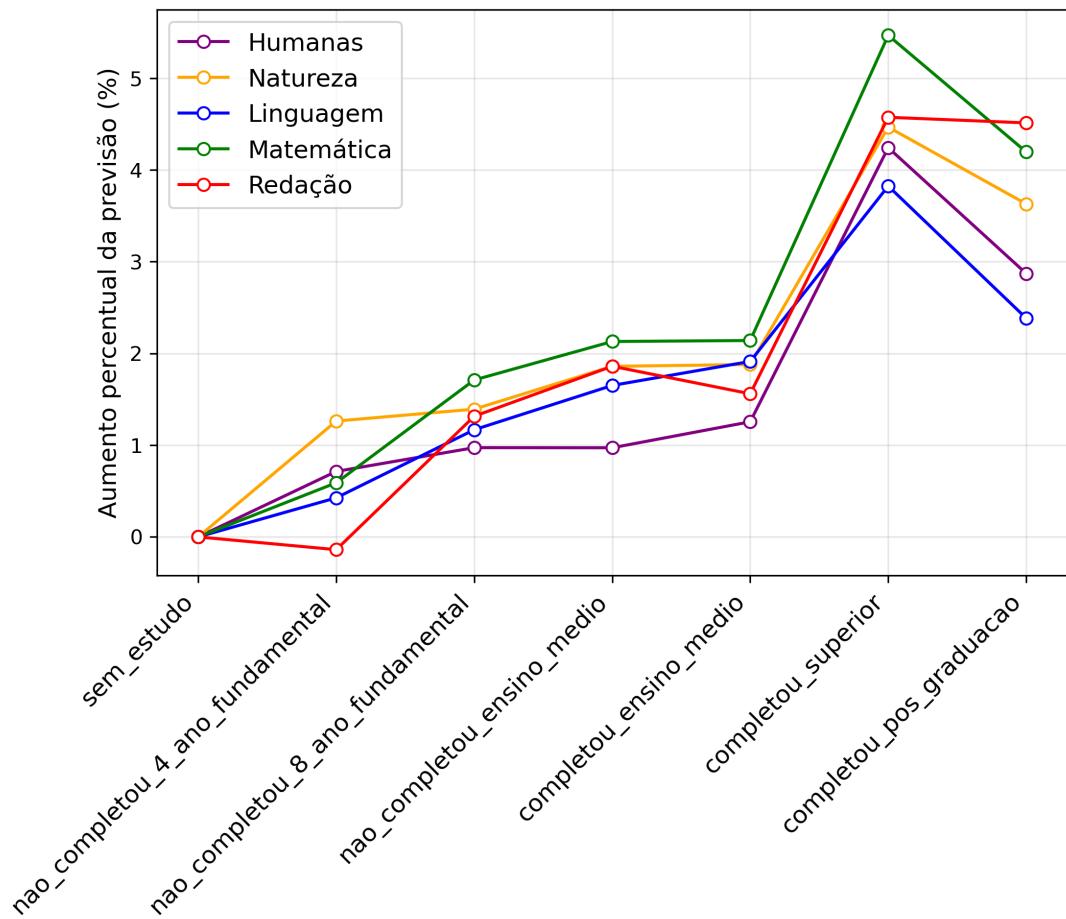
Fonte: elaborado pelo autor.

Figura 34 – Curva de Sensibilidade - Cor/Raça



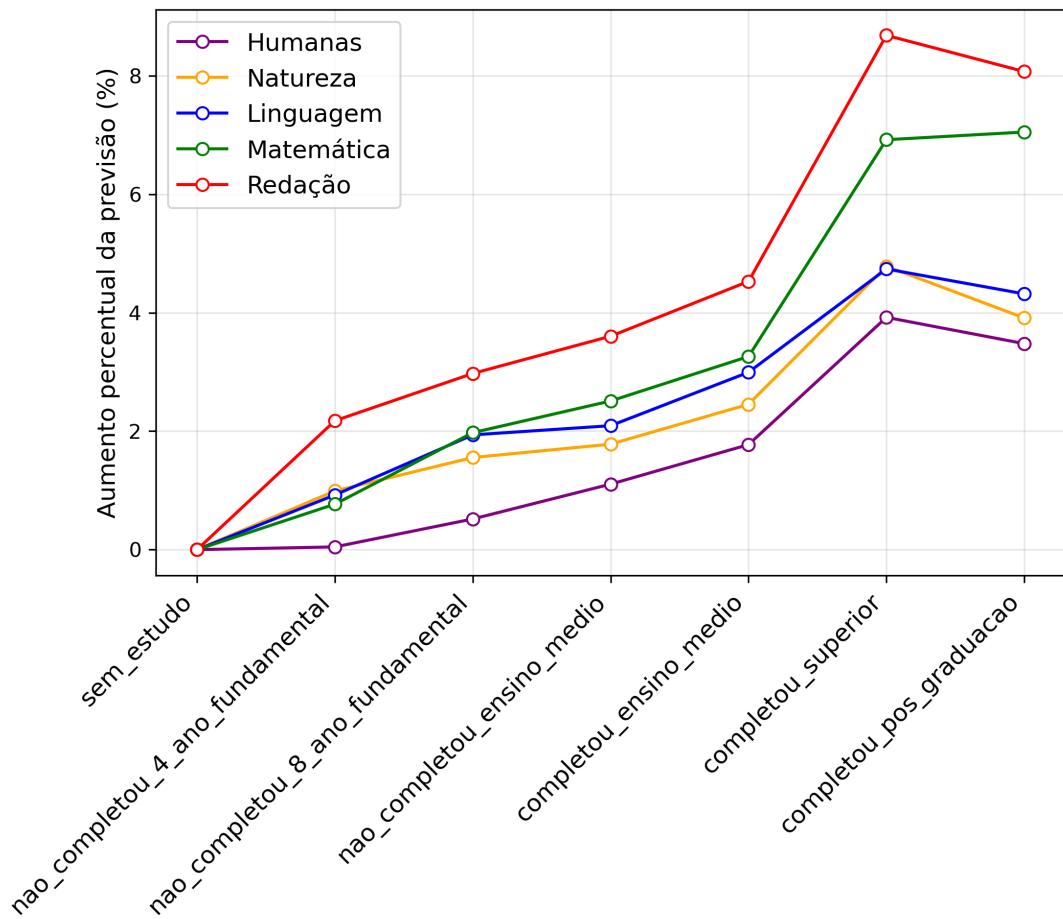
Fonte: elaborado pelo autor.

Figura 35 – Curva de Sensibilidade - Escolaridade do Pai



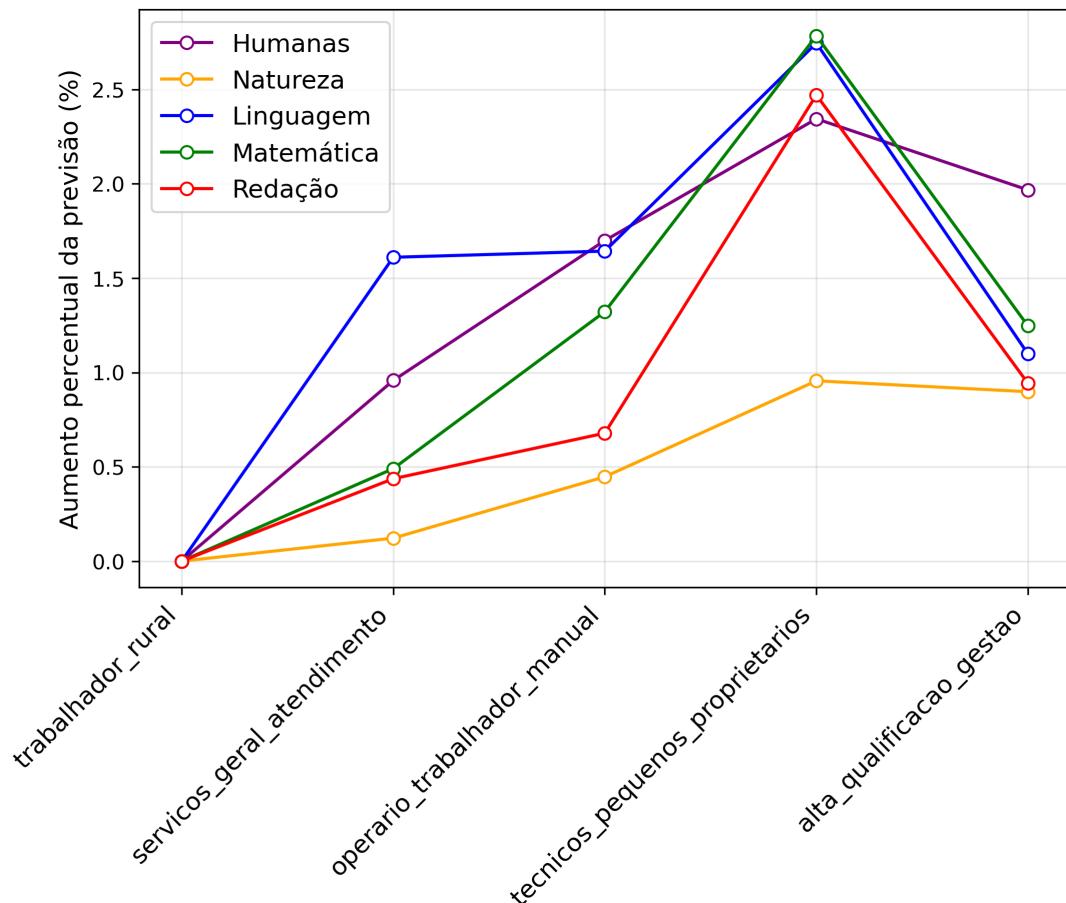
Fonte: elaborado pelo autor.

Figura 36 – Curva de Sensibilidade - Escolaridade da M  e



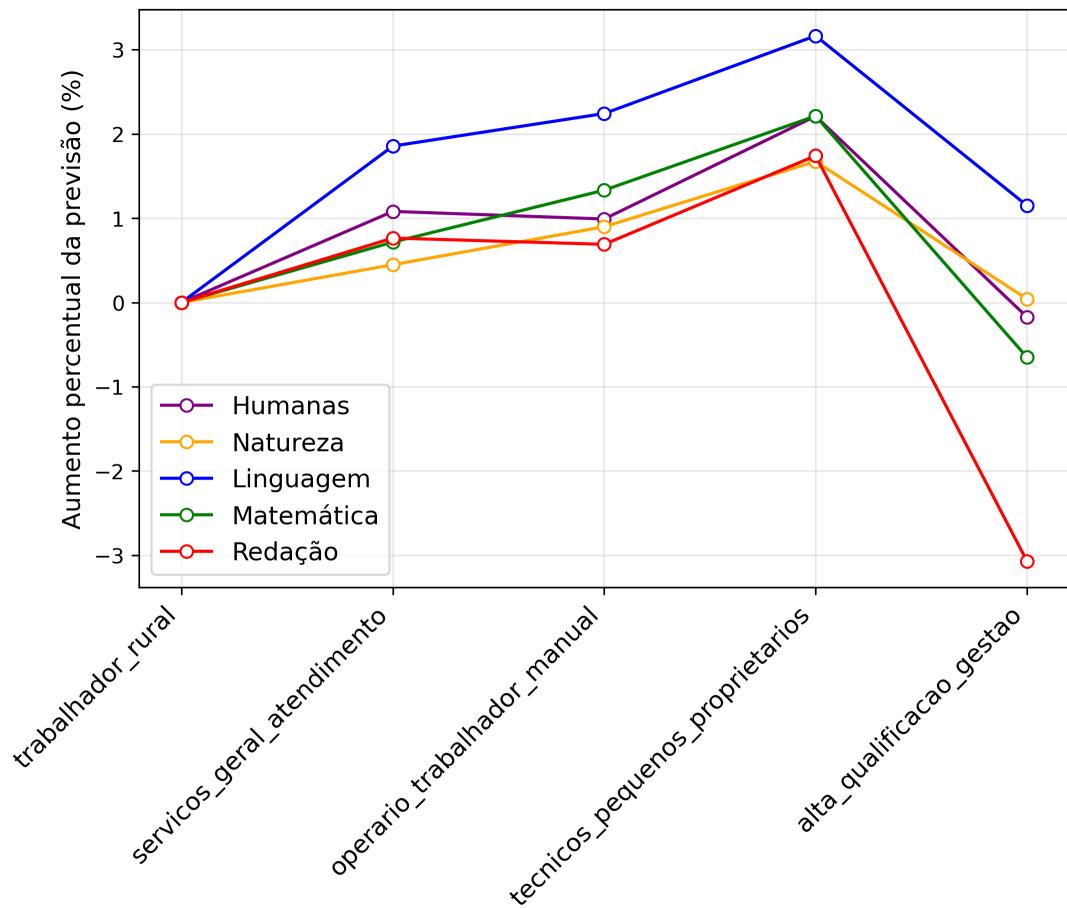
Fonte: elaborado pelo autor.

Figura 37 – Curva de Sensibilidade - Ocupação do Pai



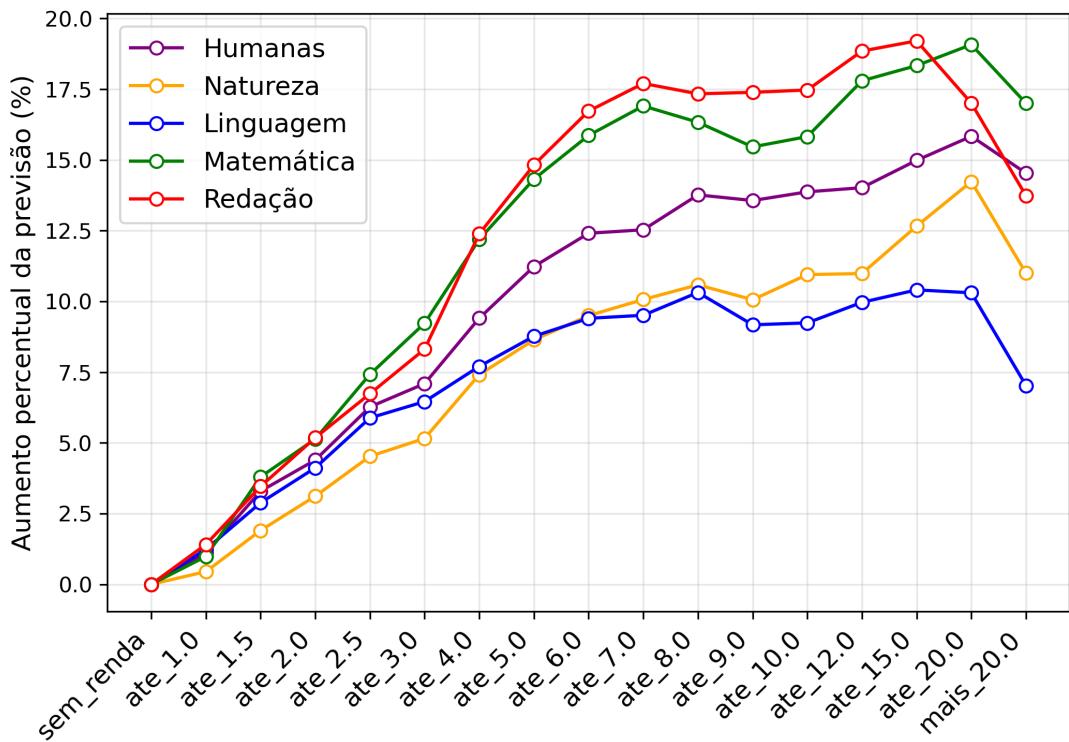
Fonte: elaborado pelo autor.

Figura 38 – Curva de Sensibilidade - Ocupação da Mãe



Fonte: elaborado pelo autor.

Figura 39 – Curva de Sensibilidade - Renda Familiar



Fonte: elaborado pelo autor.

4.7 Discussão dos Resultados

A análise dos modelos preditivos e, em especial, das curvas de sensibilidade e dos *rankings* de importância de variáveis, permite ir além das métricas de erro (RMSE e MAPE) e compreender os fenômenos sociais subjacentes ao desempenho no ENEM. Nesta seção, os resultados quantitativos são interpretados à luz da fundamentação teórica apresentada no Capítulo 2, estabelecendo conexões entre os dados e a realidade educacional brasileira.

4.7.1 O Capital Cultural e a Reprodução de Desigualdades

Os resultados obtidos corroboram fortemente a teoria do Capital Cultural de Pierre Bourdieu, discutida na Seção 2.2. Ao observar os gráficos de importância das variáveis (Figuras 27 a 31), nota-se que a **Escolaridade da Mãe** e a **Ocupação do Pai** aparecem consistentemente entre os preditores mais influentes em todas as áreas do conhecimento.

Além da influência parental direta, o fato de a escolaridade da mãe e a ocupação do pai serem variáveis tão relevantes reforça a percepção dos papéis de gênero tradicionais e da divisão sexual do trabalho. Nesse contexto, a mãe é frequentemente vista como a principal responsável pelo cuidado e educação dos filhos, enquanto o pai é associado ao papel de provedor financeiro da família (32).

Ambas as curvas de sensibilidade para a escolaridade dos pais (Figuras 35 e 36)

apresentam uma relação monotônica e crescente, indicando que o aumento do nível de instrução dos genitores está associado a um aumento na nota prevista do candidato. O ajuste de uma regressão linear simples para cada uma dessas curvas resulta em um Coeficiente de Determinação (R^2) de 91% para a escolaridade da mãe e de 86% para a escolaridade do pai.

Esses comportamentos validam a hipótese de que o capital cultural familiar, institucionalizado na forma de diplomas, atua como um facilitador do desempenho acadêmico. A inclinação acentuada dessas curvas sugere que o sistema educacional, refletido no ENEM, valoriza e recompensa o repertório cultural herdado, confirmando a tese de que a escola tende a transformar diferenças sociais em distinções escolares.

4.7.2 A Renda e o Acesso a Recursos

A variável **Renda Familiar** apresentou-se como um dos discriminadores mais fortes de desempenho. Esta variável representa múltiplos do salário mínimo da época do exame, tornando-se um indicador direto do poder aquisitivo da família. O fato de esta ser a variável de maior importância para os modelos de *ensemble* corrobora a premissa de que o acesso a recursos financeiros é um fator decisivo para o sucesso escolar.

A curva de sensibilidade associada (Figura 39) exibe um crescimento rápido nas faixas iniciais de renda, tendendo a uma estabilização nas faixas mais altas. Até a faixa de 8 salários mínimos, o aumento é acentuado, com cada mudança de faixa representando em torno de 10 pontos na nota do estudante. Isso indica que a carência de recursos básicos tem um impacto devastador na nota, enquanto o acúmulo de riqueza, após certo ponto, oferece retornos marginais decrescentes.

Entretanto, a importância destacada da variável **Quantidade de Computadores**, que figura entre os principais preditores, merece atenção especial. Mais do que um simples bem de consumo, o computador tornou-se, especialmente no contexto pós-pandemia, a ferramenta primordial de acesso ao conhecimento (33).

Sua alta relevância no modelo sugere que a exclusão digital é, hoje, uma das faces mais perversas da desigualdade educacional. A posse de computadores não reflete apenas poder econômico, mas a capacidade de estudar de forma autônoma, acessar videoaulas e materiais complementares, o que é decisivo em um exame conteudista como o ENEM (33).

4.7.3 Fatores Demográficos

A análise da sensibilidade da variável **Faixa Etária** (Figura 32) revela uma tendência preocupante: o desempenho tende a decrescer conforme a idade do participante avança além da idade regular de conclusão do Ensino Médio (17-18 anos). Participantes mais velhos frequentemente enfrentam a dupla jornada de trabalho e estudo (34, 35), dispendo

de menos tempo para preparação, o que se reflete em notas inferiores, perpetuando um ciclo de dificuldade de acesso ao ensino superior.

Em relação à **Cor/Raça**, a curva de sensibilidade (Figura 34) reforça a existência de disparidades raciais estruturais. Mesmo quando isolada pelo modelo (mantendo-se as demais variáveis constantes na análise de sensibilidade), observa-se uma variação no desempenho predito entre candidatos brancos e pretos/pardos. Isso sugere que o racismo estrutural opera através de mecanismos que não são capturados apenas pelas variáveis de renda ou escolaridade parental (36).

4.7.4 Desempenho dos Modelos de Machine Learning

Do ponto de vista técnico, a superioridade do modelo de *ensemble* (*XGBoost + LightGBM*), conforme demonstrado nas tabelas da Seção 4.5, justifica a utilização de técnicas de *Machine Learning* mais complexas em detrimento de regressões lineares simples.

A capacidade desses modelos de capturar relações não lineares é fundamental, visto que a relação entre fatores socioeconômicos e desempenho educacional não é linear, como se observa pelas curvas de sensibilidade que apresentam patamares e saturações. Os erros percentuais (MAPE) obtidos indicam que o perfil socioeconômico é, infelizmente, um forte preditor do sucesso escolar no Brasil.

Isso leva à conclusão de que o ENEM, embora desenhado para ser uma ferramenta de acesso democrático, ainda reflete de maneira fiel as profundas desigualdades da sociedade brasileira.

5 CONCLUSÃO

Este trabalho dedicou-se a investigar e quantificar a influência de fatores socioeconômicos no desempenho dos estudantes no Exame Nacional do Ensino Médio (ENEM), utilizando técnicas avançadas de Ciência de Dados e *Machine Learning*. A partir da integração de microdados de múltiplas edições e da aplicação do processo CRISP-DM, foi possível não apenas prever as notas, mas, principalmente, interpretar os modelos para compreender as dinâmicas de desigualdade educacional no Brasil.

5.1 Síntese dos Resultados

Em resposta às perguntas de pesquisa formuladas, conclui-se que o desempenho no ENEM é fortemente determinado pelo contexto socioeconômico do participante. As análises de *Permutation Importance* e as curvas de sensibilidade demonstraram que a **Renda Familiar**, a **Quantidade de Computadores** e a **Escolaridade dos Pais** (especialmente a materna) são os preditores mais influentes nas notas, superando variáveis demográficas isoladas.

A validação da teoria do Capital Cultural de Pierre Bourdieu foi evidenciada pela relação monotônica crescente entre a escolaridade dos pais e o desempenho dos filhos. Os dados confirmam que o sistema de avaliação, embora padronizado, reflete disparidades de origem: filhos de pais com ensino superior e maior renda partem de um patamar significativamente mais elevado, perpetuando o ciclo de reprodução social.

Um achado particular deste estudo foi a magnitude da influência da variável **Quantidade de Computadores**. Identificada como um dos principais discriminadores de desempenho, essa variável aponta para a exclusão digital como uma barreira crítica moderna. No contexto pós-pandemia, o acesso a equipamentos de tecnologia da informação deixou de ser um diferencial para se tornar um pré-requisito para a competitividade no exame.

Do ponto de vista técnico, a abordagem de *ensemble* (combinando *XGBoost* e *LightGBM*) mostrou-se superior aos modelos individuais, atingindo erros percentuais (MAPE) na casa dos 10%. Isso demonstra que a relação entre fatores sociais e desempenho educacional é complexa e não-linear, exigindo modelos robustos capazes de capturar saturações (como o teto de influência da renda) e interações entre variáveis.

5.2 Limitações do Estudo

Apesar dos resultados robustos, este trabalho encontrou limitações, principalmente relacionadas à disponibilidade e formato dos dados:

- **Dados de 2024 e LGPD:** A alteração na estrutura dos microdados de 2024, em adequação à Lei Geral de Proteção de Dados (LGPD), impediu a junção direta entre as informações socioeconômicas e as notas dos participantes, impossibilitando o uso da edição mais recente do exame neste estudo.
- **Identificação das Escolas:** A ausência de uma chave estrangeira nos microdados públicos do ENEM que permitisse o vínculo direto com o Censo Escolar limitou a análise do “Efeito Escola”. Não foi possível incorporar variáveis estruturais das instituições (como infraestrutura predial ou formação docente) aos modelos preditivos dos alunos.
- **Hardware:** Embora o ambiente com GPU tenha acelerado o processamento, o volume massivo de dados exigiu adaptações, como a implementação manual do *Grid Search* e a limitação de estimadores em certas etapas para evitar estouro de memória.

5.3 Trabalhos Futuros

Para a continuidade desta pesquisa e aprofundamento no tema, sugerem-se as seguintes abordagens:

- **Análise Espacial Georreferenciada:** Incorporar dados geográficos para analisar como as desigualdades se distribuem espacialmente entre municípios e regiões, cruzando as notas com o IDH ou PIB local.
- **Processamento de Linguagem Natural (PLN):** Aplicar técnicas de PLN nos temas das redações e, se disponíveis, nos espelhos das redações, para investigar se determinados temas favorecem grupos socioeconômicos específicos.
- **Análise Longitudinal:** Caso o INEP restabeleça o vínculo dos dados sob a LGPD, realizar um estudo longitudinal acompanhando coortes de alunos para verificar a evolução da desigualdade ao longo de uma década completa.
- **Políticas Públicas:** Utilizar os modelos preditivos para simular o impacto de políticas de inclusão, como a distribuição de computadores ou programas de reforço escolar focados em grupos demográficos específicos identificados como vulneráveis pelas curvas de sensibilidade.

Por fim, este trabalho reafirma que a Ciência de Dados é uma ferramenta poderosa para as Ciências Sociais. Ao quantificar o peso das desigualdades, oferece-se não apenas um diagnóstico técnico, mas um argumento estatístico sólido para a defesa de políticas públicas que visem democratizar, de fato, o acesso ao ensino superior no Brasil.

REFERÊNCIAS

- 1 MELO, R. O. *et al.* Impacto das variáveis socioeconômicas no desempenho do ENEM: uma análise espacial e sociológica. **Revista de Administração Pública**, v. 55, n. 6, p. 1271–1294, nov./dez. 2021.
- 2 ORTEGA, A. *et al.* Análise comparativa: Escola pública x escola privada no ENEM. In: **Primeiro Hackthon de Dados pela Universidade Federal do ABC**. São Paulo: [S.l.: s.n.], 2025. Relatório.
- 3 NASCIMENTO, M. M. *et al.* Análise estatística e pluriescalar das desigualdades educacionais: aspirações científicas e desempenho de estudantes no ENEM. **Sociologias**, v. 27, 2025. Disponível em: <https://doi.org/10.1590/1807-0337/e130399>.
- 4 INEP. **Microdados ENEM**. Disponível em: <https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem>.
- 5 INEP. **Histórico do ENEM**. Disponível em: <https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enem/historico>.
- 6 INEP. **ENEM**. Disponível em: <https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enem>.
- 7 OLIVEIRA, L. K. S.; CRUZ, R. C. Capital cultural e educação: uma análise da obra de bordieu. In: **XIII Encontro Cearense de Historiadores da Educação - ECHE, III Encontro Nacional do Núcleo de História e Memória da Educação - ENHIME, III Simpósio Nacional de Estudos Culturais e Geoeducacionais - SINECGEO**. [S.l.: s.n.], 2014. p. 1247–1255. ISBN: 978-85-8126-065-5. Documento de evento, sem data e local de publicação explícitos.
- 8 VASCONCELLOS, F. **Resultados do ENEM refletem desigualdades comuns no país**. 2013. Disponível em: <https://oglobo.globo.com/brasil/educacao/resultados-do-enem-refletem-desigualdades-comuns-no-pais-10445682>.
- 9 JALOTO, A.; PRIMI, R. Fatores socioeconômicos associados ao desempenho no ENEM. **Em Aberto**, v. 34, n. 112, p. 125–141, dec 2021. Disponível em: <https://www.researchgate.net/publication/357656960>.
- 10 MORAES, C. P. d. *et al.* Efeito escola a partir de indicadores educacionais: análise entre escolas públicas e privadas no ENEM. **REVISTA META: AVALIAÇÃO**, v. 14, n. 42, p. 67–93, mar 2022.
- 11 BARTHOLO, T. *et al.* **Oportunidades educacionais de estudantes concluintes do Ensino Médio: Relatório 1-Inscrição e Participação no ENEM entre 2013 e 2021**. Rio de Janeiro, 2023.
- 12 HIROMI, F. ENEM mais desigual requer atenção dos gestores. **Aprendizagem em Foco**, n. 92, oct 2023. Disponível em: <https://www.institutounibanco.org.br/boletim/enem-mais-desigual-requer-atencao-dos-gestores/>.

- 13 ROMERO, M. C. **Aplicando técnicas de Machine Learning para avaliar resultados do ENEM**. 2021. 72 p. Dissertação (Trabalho de Conclusão de Curso (MBA em Ciências de Dados)) — Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2021.
- 14 FERRAZ, A. P. **Prevendo a aprovação de um participante do ENEM no SISU para o curso de Medicina**. 2020. 70 p. Dissertação (Trabalho de Conclusão de Curso (MBA em Ciências de Dados)) — Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2020.
- 15 INEP. **Microdados Censo Escolar**. Local: Brasília, DF. [s.d.]. Disponível em: <https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/censo-escolar>.
- 16 JAMES, G. *et al.* **An Introduction to Statistical Learning: with Applications in Python**. Boca Raton: CRC Press, 2023.
- 17 GRUS, J. **Data Science from Scratch: First Principles with Python**. 2. ed. Sebastopol, CA: O'Reilly Media, Inc., 2019.
- 18 LINDHOLM, A. *et al.* **Machine Learning: A First Course for Engineers and Scientists**. Cambridge, UK; New York, NY: Cambridge University Press, 2022.
- 19 BREIMAN, L. Random forests. **Machine Learning**, v. 45, n. 1, p. 5–32, oct 2001.
- 20 CHAPMAN, P. *et al.* **CRISP-DM 1.0: Step-by-step Data Mining Guide**. [S.l.], 2000. Disponível em: <https://mineracaodedados.wordpress.com/wp-content/uploads/2012/12/crisp-dm-1-0.pdf>.
- 21 BUSSAB, W. d. O.; MORETTIN, P. A. **Estatística Básica**. 9. ed. São Paulo: Saraiva, 2017.
- 22 COHEN, J. **Statistical Power Analysis for the Behavioral Sciences**. 2. ed. Hillsdale: Lawrence Erlbaum Associates, 1988.
- 23 KPMG Advisory N.V. **Phi_K Correlation Constant**. 2024. Disponível em: <https://phik.readthedocs.io/en/latest/>.
- 24 BAAK, M. *et al.* A new correlation coefficient between categorical, ordinal and interval variables with Pearson characteristics. **Computational Statistics and Data Analysis**, v. 152, p. 107043, 2020. ISSN 0167-9473.
- 25 NVIDIA. **Welcome to cuML's documentation!** 2023. Disponível em: <https://docs.rapids.ai/api/cuml/stable/>.
- 26 PEDREGOSA, F. *et al.* Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.
- 27 INEP. **Microdados do Enem 2024: Leia-Me**. Brasília, 2025. Disponível em: <https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem>.
- 28 Anaconda. **Getting started with Miniconda**. n.d. Disponível em: <https://www.anaconda.com/docs/getting-started/miniconda/main>.

-
- 29 NVIDIA. **CUDA-X Data Science Libraries**. n.d. Disponível em: <https://developer.nvidia.com/topics/ai/data-science/cuda-x-data-science-libraries>.
- 30 NVIDIA. **Welcome to the cuDF documentation!** n.d. Disponível em: <https://docs.rapids.ai/api/cudf/stable/>.
- 31 The pandas development team. **pandas documentation**. 2025. Disponível em: <https://pandas.pydata.org/docs/>.
- 32 BRASIL. **De um lado, os papéis de gênero tradicionais e a divisão sexual do trabalho**. s.d. Disponível em: <https://www.gov.br/mdh/pt-br/navegue-por-temas/politicas-para-mulheres/arquivo/assuntos/poder-e-participacao-politica/programas-acoes/de-um-lado-os-papeis-de-genero-tradicionais-e-a-divisao-sexual-do-trabalho>. Acesso em: 15 fev. 2026.
- 33 IDOETA, P. A. '**Aluno dividia celular com dois irmãos': 51% na rede pública ainda não têm acesso a computador com internet**'. 2021. Disponível em: <https://g1.globo.com/educacao/noticia/2021/11/08/aluno-dividia-celular-com-dois-irmaos-51-na-rede-publica-ainda-nao-tem-acesso-a-computador-com-internet.html>. Acesso em: 15 fev. 2026.
- 34 IBGE. **Síntese de indicadores sociais: uma análise das condições de vida da população brasileira: 2023**. Rio de Janeiro: IBGE, 2023. Coordenação de População e Indicadores Sociais. Disponível em: <https://biblioteca.ibge.gov.br/visualizacao/livros/liv102008.pdf>.
- 35 CORROCHANO, M. C.; ABRAMO, H. W. O trabalho e a escolarização de jovens no ensino médio: desafios e perspectivas. **Ensino Médio em Revista**, v. 1, p. 3–16, 2018. Disponível em: <http://www.revista.ensinomedio.org.br>.
- 36 ALMEIDA, S. L. d. **Racismo estrutural**. São Paulo: Suéli Carneiro; Pólen, 2019. (Feminismos Plurais).

APÊNDICES

APÊNDICE A – DICIONÁRIO DE DADOS DOS MICRODADOS DO ENEM

DICIONÁRIO DE VARIÁVEIS - ENEM 2023					
NOME DA VARIÁVEL	Descrição	Variáveis Categóricas		Tamanho	Tipo
		Categoria	Descrição		
DADOS DO PARTICIPANTE					
NU_INSCRICAO	Número de inscrição ¹			12	Numérica
NU_ANO	Ano do Enem			4	Numérica
TP_FAIXA_ETARIA	Faixa etária ²	1	Menor de 17 anos	2	Numérica
		2	17 anos		
		3	18 anos		
		4	19 anos		
		5	20 anos		
		6	21 anos		
		7	22 anos		
		8	23 anos		
		9	24 anos		
		10	25 anos		
		11	Entre 26 e 30 anos		
		12	Entre 31 e 35 anos		
		13	Entre 36 e 40 anos		
		14	Entre 41 e 45 anos		
		15	Entre 46 e 50 anos		
		16	Entre 51 e 55 anos		
		17	Entre 56 e 60 anos		
		18	Entre 61 e 65 anos		
		19	Entre 66 e 70 anos		
		20	Maior de 70 anos		
TP_SEXO	Sexo	M	Masculino	1	Alfanumérica
		F	Feminino		
TP_ESTADO_CIVIL	Estado Civil	0	Não informado	1	Numérica
		1	Solteiro(a)		
		2	Casado(a)/Mora com companheiro(a)		
		3	Divorciado(a)/Desquitado(a)/Separado(a)		
TP_COR_RACA	Cor/raça	4	Viúvo(a)	1	Numérica
		0	Não declarado		
		1	Branca		
		2	Preta		
		3	Parda		
		4	Amarela		
TP_NACIONALIDADE	Nacionalidade	5	Indígena	1	Numérica
		6	Não dispõe da informação		
		0	Não informado		
		1	Brasileiro(a)		
		2	Brasileiro(a) Naturalizado(a)		
TP_ST_CONCLUSAO	Situação de conclusão do Ensino Médio	3	Estrangeiro(a)	1	Numérica
		4	Brasileiro(a) Nato(a), nascido(a) no exterior		
		1	Já conclui o Ensino Médio	1	Numérica
		2	Estou cursando e concluiréi o Ensino Médio em 2023		
TP_ANO_CONCLUIU	Ano de Conclusão do Ensino Médio	3	Estou cursando e concluiréi o Ensino Médio após 2023		
		4	Não conclui e não estou cursando o Ensino Médio		
		0	Não informado	1	Numérica
		1	2022		
		2	2021		
		3	2020		
		4	2019		
		5	2018		
		6	2017		
		7	2016		
		8	2015		
		9	2014		
		10	2013		
		11	2012		
		12	2011		
		13	2010		
		14	2009		
TP_ESCOLA	Tipo de escola do Ensino Médio	15	2008	1	Numérica
		16	2007		
		17	Antes de 2007		
TP_ENSINO	Tipo de instituição que concluiu ou concluirá o Ensino Médio	1	Não Respondeu	1	Numérica
		2	Pública		
		3	Privada		
IN_TREINEIRO	Indica se o inscrito fez a prova com intuito de apenas treinar seus conhecimentos ³	1	Ensino Regular	1	Numérica
		2	Educação Especial - Modalidade Substitutiva		
CO_MUNICIPIO_ESC	Código do município da escola	1	Sim	1	Numérica
		0	Não		
		DADOS DA ESCOLA			
		1º dígito: Região		7	Numérica
		1º e 2º dígitos: UF			
		3º, 4º, 5º e 6º dígitos: Município			
		7º dígito: dígito verificador			

NO_MUNICIPIO_ESC	Nome do município da escola		150	Alfanumérica
CO_UF_ESC	Código da Unidade da Federação da escola		2	Numérica
SG_UF_ESC	Sigla da Unidade da Federação da escola		2	Alfanumérica
TP_DEPENDENCIA_ADMIN_ESC	Dependência administrativa (Escola)	1 Federal	1	Numérica
		2 Estadual		
		3 Municipal		
		4 Privada		
TP_LOCALIZACAO_ESC	Localização (Escola)	1 Urbana	1	Numérica
TP_SIT_FUNC_ESC	Situação de funcionamento (Escola)	2 Rural	1	Numérica
		1 Em atividade		
		2 Paralisada		
		3 Extinta		
		4 Escola extinta em anos anteriores.		
DADOS DO LOCAL DE APLICAÇÃO DA PROVA				
CO_MUNICIPIO_PROVA	Código do município da aplicação da prova 1º dígito: Região 1º e 2º dígitos: UF 3º, 4º, 5º e 6º dígitos: Município 7º dígito: dígito verificador		7	Numérica
NO_MUNICIPIO_PROVA	Nome do município da aplicação da prova			
CO_UF_PROVA	Código da Unidade da Federação da aplicação da prova			
SG_UF_PROVA	Sigla da Unidade da Federação da aplicação da prova			
DADOS DA PROVA OBJETIVA				
TP_PRESENCA_CN	Presença na prova objetiva de Ciências da Natureza	0 Faltou à prova	1	Numérica
		1 Presente na prova		
		2 Eliminado na prova		
TP_PRESENCA_CH	Presença na prova objetiva de Ciências Humanas	0 Faltou à prova	1	Numérica
		1 Presente na prova		
		2 Eliminado na prova		
TP_PRESENCA_LC	Presença na prova objetiva de Linguagens e Códigos	0 Faltou à prova	1	Numérica
		1 Presente na prova		
		2 Eliminado na prova		
TP_PRESENCA_MT	Presença na prova objetiva de Matemática	0 Faltou à prova	1	Numérica
		1 Presente na prova		
		2 Eliminado na prova		
CO_PROVA_CN	Código do tipo de prova de Ciências da Natureza	1221 Azul	4	Numérica
		1222 Amarela		
		1223 Rosa		
		1224 Cinza		
		1225 Rosa - Ampliada		
		1226 Rosa - Superampliada		
		1227 Laranja - Braile		
		1228 Laranja - Adaptada Ledor		
		1229 Verde - Videoprosa - Libras		
		1301 Azul (Reaplicação)		
		1302 Amarela (Reaplicação)		
		1303 Cinza (Reaplicação)		
		1304 Rosa (Reaplicação)		
		1191 Azul		
		1192 Amarela		
CO_PROVA_CH	Código do tipo de prova de Ciências Humanas	1193 Branca	4	Numérica
		1194 Rosa		
		1195 Rosa - Ampliada		
		1196 Rosa - Superampliada		
		1197 Laranja - Braile		
		1198 Laranja - Adaptada Ledor		
		1199 Verde - Videoprosa - Libras		
		1271 Azul (Reaplicação)		
		1272 Amarela (Reaplicação)		
		1273 Branca (Reaplicação)		
		1274 Rosa (Reaplicação)		
CO_PROVA_LC	Código do tipo de prova de Linguagens e Códigos	1201 Azul	4	Numérica
		1202 Amarela		
		1203 Rosa		
		1204 Branca		
		1205 Rosa - Ampliada		
		1206 Rosa - Superampliada		
		1207 Laranja - Braile		
		1208 Laranja - Adaptada Ledor		
		1209 Verde - Videoprosa - Libras		
		1281 Azul (Reaplicação)		
CO_PROVA_MT	Código do tipo de prova de Matemática	1282 Amarela (Reaplicação)	4	Numérica
		1283 Rosa (Reaplicação)		
		1284 Branca (Reaplicação)		
		1211 Azul		
		1212 Amarela		

		1291	Azul (Reaplicação)		
		1292	Amarela (Reaplicação)		
		1293	Rosa (Reaplicação)		
		1294	Cinza (Reaplicação)		
NU_NOTA_CN	Nota da prova de Ciências da Natureza			9	Numérica
NU_NOTA_CH	Nota da prova de Ciências Humanas			9	Numérica
NU_NOTA_LC	Nota da prova de Linguagens e Códigos			9	Numérica
NU_NOTA_MT	Nota da prova de Matemática			9	Numérica
TX_RESPOSTAS_CN	Vetor com as respostas da parte objetiva da prova de Ciências da Natureza ⁴		A,B,C,D, E, * (dupla marcação), . (em branco)	45	Alfanumérica
TX_RESPOSTAS_CH	Vetor com as respostas da parte objetiva da prova de Ciências Humanas ⁴		A,B,C,D, E, * (dupla marcação), . (em branco)	45	Alfanumérica
TX_RESPOSTAS_LC	Vetor com as respostas da parte objetiva da prova de Linguagens e Códigos ⁵		A,B,C,D, E, * (dupla marcação), . (em branco), 9 (Item não apresentado)	45	Alfanumérica
TX_RESPOSTAS_MT	Vetor com as respostas da parte objetiva da prova de Matemática ⁴		A,B,C,D, E, * (dupla marcação), . (em branco)	45	Alfanumérica
TP_LINGUA	Língua Estrangeira	0	Inglês	1	Numérica
		1	Espanhol		
TX_GABARITO_CN	Vetor com o gabarito da parte objetiva da prova de Ciências da Natureza ⁶			45	Alfanumérica
TX_GABARITO_CH	Vetor com o gabarito da parte objetiva da prova de Ciências Humanas ⁶			45	Alfanumérica
TX_GABARITO_LC	Vetor com o gabarito da parte objetiva da prova de Linguagens e Códigos ⁷			50	Alfanumérica
TX_GABARITO_MT	Vetor com o gabarito da parte objetiva da prova de Matemática ⁶			45	Alfanumérica
DADOS DA REDAÇÃO					
TP_STATUS_REDACAO	Situação da redação do participante	1	Sem problemas	1	Numérica
		2	Anulada		
		3	Cópia Texto Motivador		
		4	Em Branco		
		6	Fuga ao tema		
		7	Não atendimento ao tipo textual		
		8	Texto insuficiente		
		9	Parte desconectada		
NU_NOTA_COMP1	Nota da competência 1 - Demonstrar domínio da modalidade escrita formal da Língua Portuguesa.			9	Numérica
NU_NOTA_COMP2	Nota da competência 2 - Compreender a proposta de redação e aplicar conceitos das várias áreas de conhecimento para desenvolver o tema, dentro dos limites estruturais do texto dissertativo-argumentativo em prosa.			9	Numérica
NU_NOTA_COMP3	Nota da competência 3 - Selecionar, relacionar, organizar e interpretar informações, fatos, opiniões e argumentos em defesa de um ponto de vista.			9	Numérica
NU_NOTA_COMP4	Nota da competência 4 - Demonstrar conhecimento dos mecanismos linguísticos necessários para a construção da argumentação.			9	Numérica
NU_NOTA_COMP5	Nota da competência 5 - Elaborar proposta de intervenção para o problema abordado, respeitando os direitos humanos.			9	Numérica
NU_NOTA_REDACAO	Nota da prova de redação			9	Numérica
DADOS DO QUESTIONÁRIO SOCIOECONÔMICO					
Q001	Até que série seu pai, ou o homem responsável por você, estudou?	A	Nunca estudou.	1	Alfanumérica
		B	Não completou a 4ª série/5º ano do Ensino Fundamental.		
		C	Completo a 4ª série/5º ano, mas não completou a 8ª série/9º ano do Ensino Fundamental.		
		D	Completo a 8ª série/9º ano do Ensino Fundamental, mas não completou o Ensino Médio.		
		E	Completo o Ensino Médio, mas não completou a Faculdade.		
		F	Completo a Faculdade, mas não completou a Pós-graduação.		
		G	Completo a Pós-graduação.		
		H	Não sei.		
		A	Nunca estudou.		
Q002	Até que série sua mãe, ou a mulher responsável por você, estudou?	B	Não completou a 4ª série/5º ano do Ensino Fundamental.	1	Alfanumérica
		C	Completo a 4ª série/5º ano, mas não completou a 8ª série/9º ano do Ensino Fundamental.		
		D	Completo a 8ª série/9º ano do Ensino Fundamental, mas não completou o Ensino Médio.		
		E	Completo o Ensino Médio, mas não completou a Faculdade.		
		F	Completo a Faculdade, mas não completou a Pós-graduação.		
		G	Completo a Pós-graduação.		
		H	Não sei.		

Q003	<p>A partir da apresentação de algumas ocupações divididas em grupos ordenados, indique o grupo que contempla a ocupação mais próxima da ocupação do seu pai ou do homem responsável por você. (Se ele não estiver trabalhando, escolha uma ocupação pensando no último trabalho dele).</p>	A	Grupo 1: Lavrador, agricultor sem empregados, bôia fria, criador de animais (gado, porcos, galinhas, ovelhas, cavalos etc.), apicultor, pescador, lenhador, seringueiro, extrativista.	1	Alfanumérica
		B	Grupo 2: Diarista, empregado doméstico, cuidador de idosos, babá, cozinheiro (em casas particulares), motorista particular, jardineiro, faxineiro de empresas e prédios, vigilante, porteiro, carteiro, office-boy, vendedor, caixa, atendente de loja, auxiliar administrativo, recepcionista, servente de pedreiro, repositor de mercadoria.		
		C	Grupo 3: Padeiro, cozinheiro industrial ou em restaurantes, sapateiro, costureiro, joalheiro, torneiro mecânico, operador de máquinas, soldador, operário de fábrica, trabalhador da mineração, pedreiro, pintor, eletricista, encanador, motorista, caminhoneiro, taxista.		
		D	Grupo 4: Professor (de ensino fundamental ou médio, idioma, música, artes etc.), técnico (de enfermagem, contabilidade, eletrônica etc.), policial, militar de baixa patente (soldado, cabo, sargento), corretor de imóveis, supervisor, gerente, mestre de obras, pastor, microempresário (proprietário de empresa com menos de 10 empregados), pequeno comerciante, pequeno proprietário de terras, trabalhador autônomo ou por conta própria.		
		E	Grupo 5: Médico, engenheiro, dentista, psicólogo, economista, advogado, juiz, promotor, defensor, delegado, tenente, capitão, coronel, professor universitário, diretor em empresas públicas ou privadas, político, proprietário de empresas com mais de 10 empregados.		
		F	Não sei.		
Q004	<p>A partir da apresentação de algumas ocupações divididas em grupos ordenados, indique o grupo que contempla a ocupação mais próxima da ocupação da sua mãe ou da mulher responsável por você. (Se ela não estiver trabalhando, escolha uma ocupação pensando no último trabalho dela).</p>	A	Grupo 1: Lavradora, agricultora sem empregados, bôia fria, criadora de animais (gado, porcos, galinhas, ovelhas, cavalos etc.), apicultora, pescadora, lenhadora, seringueira, extrativista.	2	Numérica
		B	Grupo 2: Diarista, empregada doméstica, cuidadora de idosos, babá, cozinheira (em casas particulares), motorista particular, jardineira, faxineira de empresas e prédios, vigilante, porteira, carteira, office-boy, vendedora, caixa, atendente de loja, auxiliar administrativa, recepcionista, servente de pedreiro, repositora de mercadoria.		
		C	Grupo 3: Padeira, cozinheira industrial ou em restaurantes, sapateira, costureira, joalheira, torneira mecânica, operadora de máquinas, soldadora, operária de fábrica, trabalhadora da mineração, pedreira, pintora, eletricista, encanadora, motorista, caminhoneira, taxista.		
		D	Grupo 4: Professora (de ensino fundamental ou médio, idioma, música, artes etc.), técnica (de enfermagem, contabilidade, eletrônica etc.), policial, militar de baixa patente (soldado, cabo, sargento), corretora de imóveis, supervisora, gerente, mestre de obras, pastora, microempresária (proprietária de empresa com menos de 10 empregados), pequena comerciante, pequena proprietária de terras, trabalhadora autônoma ou por conta própria.		
		E	Grupo 5: Médica, engenheira, dentista, psicóloga, economista, advogada, juíza, promotora, defensora, delegada, tenente, capitão, coronel, professora universitária, diretora em empresas públicas ou privadas, política, proprietária de empresas com mais de 10 empregados.		
		F	Não sei.		
Q005	<p>Incluindo você, quantas pessoas moram atualmente em sua residência?</p>	1	1, pois moro sozinho(a).	2	Numérica
		2	2		
		3	3		
		4	4		
		5	5		
		6	6		
		7	7		
		8	8		
		9	9		
		10	10		
		11	11		
		12	12		
		13	13		

		<table border="1"> <tr><td>14</td><td>14</td></tr> <tr><td>15</td><td>15</td></tr> <tr><td>16</td><td>16</td></tr> <tr><td>17</td><td>17</td></tr> <tr><td>18</td><td>18</td></tr> <tr><td>19</td><td>19</td></tr> <tr><td>20</td><td>20</td></tr> </table>	14	14	15	15	16	16	17	17	18	18	19	19	20	20		
14	14																	
15	15																	
16	16																	
17	17																	
18	18																	
19	19																	
20	20																	
Q006	Qual é a renda mensal de sua família? (Some a sua renda com a dos seus familiares.)	<p>A Nenhuma Renda B Até R\$ 1.320,00 C De R\$ 1.320,01 até R\$ 1.980,00. D De R\$ 1.980,01 até R\$ 2.640,00. E De R\$ 2.640,01 até R\$ 3.300,00. F De R\$ 3.300,01 até R\$ 3.960,00. G De R\$ 3.960,01 até R\$ 5.280,00. H De R\$ 5.280,01 até R\$ 6.600,00. I De R\$ 6.600,01 até R\$ 7.920,00. J De R\$ 7.920,01 até R\$ 9240,00. K De R\$ 9.240,01 até R\$ 10.560,00. L De R\$ 10.560,01 até R\$ 11.880,00. M De R\$ 11.880,01 até R\$ 13.200,00. N De R\$ 13.200,01 até R\$ 15.840,00. O De R\$ 15.840,01 até R\$ 19.800,00. P De R\$ 19.800,01 até R\$ 26.400,00. Q Acima de R\$ 26.400,00.</p>	1	Alfanumérica														
Q007	Em sua residência trabalha empregado(a) doméstico(a)?	<p>A Não. B Sim, um ou dois dias por semana. C Sim, três ou quatro dias por semana. D Sim, pelo menos cinco dias por semana.</p>	1	Alfanumérica														
Q008	Na sua residência tem banheiro?	<p>A Não. B Sim, um. C Sim, dois. D Sim, três. E Sim, quatro ou mais.</p>	1	Alfanumérica														
Q009	Na sua residência tem quartos para dormir?	<p>A Não. B Sim, um. C Sim, dois. D Sim, três. E Sim, quatro ou mais.</p>	1	Alfanumérica														
Q010	Na sua residência tem carro?	<p>A Não. B Sim, um. C Sim, dois. D Sim, três. E Sim, quatro ou mais.</p>	1	Alfanumérica														
Q011	Na sua residência tem motocicleta?	<p>A Não. B Sim, uma. C Sim, duas. D Sim, três. E Sim, quatro ou mais.</p>	1	Alfanumérica														
Q012	Na sua residência tem geladeira?	<p>A Não. B Sim, uma. C Sim, duas. D Sim, três. E Sim, quatro ou mais.</p>	1	Alfanumérica														
Q013	Na sua residência tem freezer (independente ou segunda porta da geladeira)?	<p>A Não. B Sim, um. C Sim, dois. D Sim, três. E Sim, quatro ou mais.</p>	1	Alfanumérica														
Q014	Na sua residência tem máquina de lavar roupa? (o tanquinho NÃO deve ser considerado)	<p>A Não. B Sim, uma. C Sim, duas. D Sim, três. E Sim, quatro ou mais.</p>	1	Alfanumérica														
Q015	Na sua residência tem máquina de secar roupa (independente ou em conjunto com a máquina de lavar roupa)?	<p>A Não. B Sim, uma. C Sim, duas. D Sim, três. E Sim, quatro ou mais.</p>	1	Alfanumérica														
Q016	Na sua residência tem forno micro-ondas?	<p>A Não. B Sim, um. C Sim, dois. D Sim, três. E Sim, quatro ou mais.</p>	1	Alfanumérica														
Q017	Na sua residência tem máquina de lavar louça?	<p>A Não. B Sim, uma. C Sim, duas. D Sim, três. E Sim, quatro ou mais.</p>	1	Alfanumérica														
Q018	Na sua residência tem aspirador de pó?	<p>A Não. B Sim.</p>	1	Alfanumérica														
Q019	Na sua residência tem televisão em cores?	<p>A Não. B Sim, uma. C Sim, duas. D Sim, três. E Sim, quatro ou mais.</p>	1	Alfanumérica														
Q020	Na sua residência tem aparelho de DVD?	<p>A Não.</p>	1	Alfanumérica														

Questão	Pergunta		
Q021	Na sua residência tem aparelho de DVD?		
		B Sim.	
		A Não.	
		B Sim.	
		A Não.	
		B Sim, um.	
		C Sim, dois.	
		D Sim, três.	
		E Sim, quatro ou mais.	
Q022	Na sua residência tem telefone celular?		
Q023	Na sua residência tem telefone fixo?		
		A Não.	
		B Sim.	
		A Não.	
		B Sim, um.	
		C Sim, dois.	
		D Sim, três.	
		E Sim, quatro ou mais.	
Q024	Na sua residência tem computador?		
Q025	Na sua residência tem acesso à Internet?		
		A Não.	
		B Sim.	

APÊNDICE B – DICIONÁRIO DE DADOS DO CENSO ESCOLAR

Alteração de estrutura ou metodologia de cálculo
Variável nova
Descontinuidade

N	Nome da Variável	Descrição da Variável	Tipo	Tam. ⁽¹⁾	Categoria	Notas Importantes
1	NU_ANO_CENSO	Ano do Censo	Num	4		
2	NO_REGIAO	Nome da Região Geográfica	Char	20		
3	CO_REGIAO	Código da Região Geográfica	Num	1		
4	NO_UF	Nome da Unidade da Federação	Char	50		
5	SG_UF	Sigla da Unidade da Federação	Char	2		
6	CO_UF	Código da Unidade da Federação	Num	2		
7	NO_MUNICIPIO	Nome do Município	Char	150		
8	CO_MUNICIPIO	Código do Município	Num	7		
9	NO_REGIAO_GEOG_INTERM	Nome da Região Geográfica Intermediária	Char	100		A Fundação Instituto Brasileiro de Geografia e Estatísticas (IBGE), propôs em 2017 uma nova divisão geográfica regional
10	CO_REGIAO_GEOG_INTERM	Código da Região Geográfica Intermediária	Num	4		do Brasil que corresponde às antigas Mesorregiões e Microrregiões. Assim, a partir de 2019, as regiões geográficas regionais são classificadas em Regiões Geográficas Intermediárias e Regiões Imediatas. No Censo Escolar, para fins de comparação ao longo dos anos, foram mantidas as variáveis de Mesorregiões e Microrregiões.
11	NO_REGIAO_GEOG_IMED	Nome da Região Geográfica Imediata	Char	100		
12	CO_REGIAO_GEOG_IMED	Código da Região Geográfica Imediata	Num	6		
13	NO_MESORREGIAO	Nome da Mesorregião	Char	100		
14	CO_MESORREGIAO	Código da Mesorregião	Num	4		
15	NO_MICRORREGIAO	Nome da Microrregião	Char	100		
16	CO_MICRORREGIAO	Código da Microrregião	Num	5		
17	NO_DISTRITO	Divisão Intramunicipal - Nome do Distrito	Char	100		
18	CO_DISTRITO	Divisão Intramunicipal - Código do Distrito	Num	9		
19	NO_ENTIDADE	Nome da Escola	Char	100		
20	CO_ENTIDADE	Código da Escola	Num	8		
21	TP_DEPENDENCIA	Dependência Administrativa	Num	1	1 - Federal 2 - Estadual 3 - Municipal 4 - Privada	
22	TP_CATEGORIA_ESCOLA_PRIVADA	Categoría da escola privada	Num	1	1 - Particular 2 - Comunitária 3 - Confessional 4 - Filantrópica - Não aplicável para escolas públicas	
23	TP_LOCALIZACAO	Localização	Num	1	1 - Urbana 2 - Rural	
24	TP_LOCALIZACAO_DIFERENCIADA	Localização diferenciada da escola	Num	1	0 - A escola não está em área de localização diferenciada 1 - Área de assentamento 2 - Terra indígena 3 - Comunidade quilombola 8 - Área onde se localizam povos e comunidades tradicionais	<p>De 2007 a 2011: 0 - Não aplicável 1 - Área de assentamento 2 - Terra indígena 3 - Área remanescente de quilombos</p> <p>De 2012 a 2018: 0 - A escola não está em área de localização diferenciada 1 - Área de assentamento 2 - Terra indígena 3 - Área remanescente de quilombos 4 - Unidade de uso sustentável 5 - Unidade de uso sustentável em terra indígena 6 - Unidade de uso sustentável em área remanescente de quilombos</p> <p>De 2019 a 2022: 0 - A escola não está em área de localização diferenciada 1 - Área de assentamento 2 - Terra indígena 3 - Área onde se localiza comunidade remanescente de quilombos</p> <p>Em 2023 foi adicionada a categoria 8: 8 - Área onde se localizam povos e comunidades tradicionais</p>
25	DS_ENDERECHO	Endereço	Char	100		
26	NU_ENDERECHO	Número	Char	10		
27	DS_COMPLEMENTO	Complemento	Char	20		
28	NO_BAIRRO	Bairro	Char	50		
29	CO_CEP	CEP	Char	8		
30	NU_DDD	DDD	Num	8		
31	NU_TELEFONE	Telefone	Num	8		
32	TP_SITUACAO_FUNCIONAMENTO	Situação de funcionamento	Num	1	1 - Em Atividade 2 - Paralisada 3 - Extinta (ano do Censo) 4 - Extinta em Anos Anteriores	
33	CO_ORGAO REGIONAL	Código do Órgão Regional de Ensino	Char	5		
34	DT_ANO LETIVO_INICIO	Início do ano letivo	Data	20		
35	DT_ANO LETIVO TERMINO	Término (previsão) do ano letivo	Data	20		
36	IN_VINCULO_SECRETARIA EDUCACAO	Órgão ao qual a escola pública está vinculada - Secretaria de Educação/Ministério da Educação	Num	1	0 - Não 1 - Sim - Não aplicável para escolas privadas	
37	IN_VINCULO_SEGURANCA PUBLICA	Órgão ao qual a escola pública está vinculada - Secretaria de Segurança Pública/Forças Armadas/Militar	Num	1	0 - Não 1 - Sim - Não aplicável para escolas privadas	
38	IN_VINCULO_SECRETARIA SAUDE	Órgão ao qual a escola pública está vinculada - Secretaria de Saúde/Ministério da Saúde	Num	1	0 - Não 1 - Sim - Não aplicável para escolas privadas	
39	IN_VINCULO_OUTRO_ORGAO	Órgão ao qual a escola pública está vinculada - Outro órgão da administração pública	Num	1	0 - Não 1 - Sim - Não aplicável para escolas privadas	
40	IN_PODER_PUBLICO_PARCERIA	Parceria ou convênio com o poder público (parceria ou convênio firmado entre a Administração Pública e instituições privadas ou instituições públicas de ensino, autarquias e fundações da administração indireta e demais instituições de educação profissional técnica de nível médio dos serviços sociais autônomos que integram o sistema federal de ensino, para financiamento do atendimento educacional ou para a oferta do itinerário de formação técnica e profissional do ensino médio)	Num	1	0 - Não 1 - Sim	<p>Entre 2007 e 2021: A variável IN_CONVENIADA_PP foi renomeada para: IN_PODER_PUBLICO_PARCERIA</p>
41	TP_PODER_PUBLICO_PARCERIA	Poder público responsável pela parceria ou convênio entre a Administração Pública e outras instituições	Num	1	1 - Municipal 2 - Estadual 3 - Estadual e Municipal - Não aplicável para escolas sem parceria ou convênio com o poder público	<p>Entre 2007 e 2021: A variável TP_CONVENIO_PODER_PUBLICO foi renomeada para: TP_PODER_PUBLICO_PARCERIA</p>
42	IN_CONVENIADA_PP	Conveniada com o poder público	Num	1	0 - Não 1 - Sim - Não aplicável para escolas públicas	
43	TP_CONVENIO_PODER_PUBLICO	Dependência do convênio com o poder público	Num	1	1 - Municipal 2 - Estadual 3 - Estadual e Municipal - Não aplicável para escolas públicas ou privadas não conveniadas	
44	IN_FORMA_CONT_TERMO_COLABORA	Forma de contratação entre a Administração Pública e outras instituições - Termo de colaboração (Lei nº 13.019/2014)	Num	1	0 - Não 1 - Sim - Não aplicável para escolas sem parceria/convênio	
45	IN_FORMA_CONT_TERMO_FOMENTO	Forma de contratação entre a Administração Pública e outras instituições - Termo de fomento (Lei nº 13.019/2014)	Num	1	0 - Não 1 - Sim - Não aplicável para escolas sem parceria/convênio	
46	IN_FORMA_CONT_ACORDO_COOP	Forma de contratação entre a Administração Pública e outras instituições - Acordo de cooperação (Lei nº 13.019/2014)	Num	1	0 - Não 1 - Sim - Não aplicável para escolas sem parceria/convênio	
47	IN_FORMA_CONT_PRESTACAO_SERV	Forma de contratação entre a Administração Pública e outras instituições - Contrato de prestação de serviço	Num	1	0 - Não 1 - Sim - Não aplicável para escolas sem parceria/convênio	
48	IN_FORMA_CONT_COOP_TEC_FIN	Forma de contratação entre a Administração Pública e outras instituições - Termo de cooperação técnica e financeira	Num	1	0 - Não 1 - Sim - Não aplicável para escolas sem parceria/convênio	
49	IN_FORMA_CONT_CONSORCIO_PUB	Forma de contratação entre a Administração Pública e outras instituições - Contrato de consórcio público/Convênio de cooperação	Num	1	0 - Não 1 - Sim - Não aplicável para escolas sem parceria/convênio	
50	IN_FORMA_CONT_MU_TERMO_COLAB	Forma(s) de contratação da parceria ou convênio entre a escola e a Secretaria municipal de educação - Termo de colaboração (Lei nº 13.019/2014)	Num	1	0 - Não 1 - Sim - Não aplicável para sem parceria/convênio	
51	IN_FORMA_CONT_MU_TERMO_FOMENTO	Forma(s) de contratação da parceria ou convênio entre a escola e a Secretaria municipal de educação - Termo de fomento (Lei nº 13.019/2014)	Num	1	0 - Não 1 - Sim - Não aplicável para sem parceria/convênio	

52	IN_FORMA_CONT_MU_ACORDO_COOP	Forma(s) de contratação da parceria ou convênio entre a escola e a Secretaria municipal de educação - Acordo de cooperação (Lei nº 13.019/2014)	Num	1	0 - Não 1 - Sim - Não aplicável para sem parceria/convênio	
53	IN_FORMA_CONT_MU_PREST_SERV	Forma(s) de contratação da parceria ou convênio entre a escola e a Secretaria municipal de educação - Contrato de prestação de serviço	Num	1	0 - Não 1 - Sim - Não aplicável para sem parceria/convênio	
54	IN_FORMA_CONT_MU_COOP_TEC_FIN	Forma(s) de contratação da parceria ou convênio entre a escola e a Secretaria municipal de educação - Termo de cooperação técnica e financeira	Num	1	0 - Não 1 - Sim - Não aplicável para sem parceria/convênio	
55	IN_FORMA_CONT_MU_CONSORCIO_PUB	Forma(s) de contratação da parceria ou convênio entre a escola e a Secretaria municipal de educação - Contrato de consórcio público/Convênio de cooperação	Num	1	0 - Não 1 - Sim - Não aplicável para sem parceria/convênio	
56	IN_FORMA_CONT_ES_TERMO_COLAB	Forma(s) de contratação da parceria ou convênio entre a escola e a Secretaria estadual de educação - Termo de colaboração (Lei nº 13.019/2014)	Num	1	0 - Não 1 - Sim - Não aplicável para sem parceria/convênio	
57	IN_FORMA_CONT_ES_TERMO_FOMENTO	Forma(s) de contratação da parceria ou convênio entre a escola e a Secretaria estadual de educação - Termo de fomento (Lei nº 13.019/2014)	Num	1	0 - Não 1 - Sim - Não aplicável para sem parceria/convênio	
58	IN_FORMA_CONT_ES_ACORDO_COOP	Forma(s) de contratação da parceria ou convênio entre a escola e a Secretaria estadual de educação - Acordo de cooperação (Lei nº 13.019/2014)	Num	1	0 - Não 1 - Sim - Não aplicável para sem parceria/convênio	
59	IN_FORMA_CONT_ES_PREST_SERV	Forma(s) de contratação da parceria ou convênio entre a escola e a Secretaria estadual de educação - Contrato de prestação de serviço	Num	1	0 - Não 1 - Sim - Não aplicável para sem parceria/convênio	
60	IN_FORMA_CONT_ES_COOP_TEC_FIN	Forma(s) de contratação da parceria ou convênio entre a escola e a Secretaria estadual de educação - Termo de cooperação técnica e financeira	Num	1	0 - Não 1 - Sim - Não aplicável para sem parceria/convênio	
61	IN_FORMA_CONT_ES_CONSORCIO_PUB	Forma(s) de contratação da parceria ou convênio entre a escola e a Secretaria estadual de educação - Contrato de consórcio público/Convênio de cooperação	Num	1	0 - Não 1 - Sim - Não aplicável para sem parceria/convênio	
62	IN_TIPO_ATEND_ESCOLARIZACAO	Tipo de atendimento oferecido por meio da parceria ou convênio - Escolarização	Num	1	0 - Não 1 - Sim - Não aplicável para escolas sem parceria/convênio	
63	IN_TIPO_ATEND_AC	Tipo de atendimento oferecido por meio da parceria ou convênio - Atividade Complementar	Num	1	0 - Não 1 - Sim - Não aplicável para escolas sem parceria/convênio	
64	IN_TIPO_ATEND_AEE	Tipo de atendimento oferecido por meio da parceria ou convênio - Atendimento Educacional Especializado (AEE)	Num	1	0 - Não 1 - Sim - Não aplicável para escolas sem parceria/convênio	
65	IN_MANT_ESCOLA_PRIVADA_EMP	Mantenedora da escola privada - Empresa ou grupo empresarial do setor privado ou pessoa física	Num	1	0 - Não 1 - Sim - Não aplicável para escolas públicas	
66	IN_MANT_ESCOLA_PRIVADA_ONG	Mantenedora da escola privada - Organização Não Governamental (ONG) - internacional ou nacional	Num	1	0 - Não 1 - Sim - Não aplicável para escolas públicas	
67	IN_MANT_ESCOLA_PRIVADA_OSCIP	Mantenedora da escola privada - Organização da Sociedade Civil de Interesse Público (Oscip)	Num	1	0 - Não 1 - Sim - Não aplicável para escolas públicas	
68	IN_MANT_ESCOLA_PRIV_ONG_OSCIP	Mantenedora da escola privada - Organização Não Governamental (ONG) - internacional ou nacional - Organização da Sociedade Civil de Interesse Público (Oscip)	Num	1	0 - Não 1 - Sim - Não aplicável para escolas públicas	A partir de 2019: Variável derivada juntando IN_MANT_ESCOLA_PRIVADA_ONG e IN_MANT_ESCOLA_PRIVADA_OSCIP
69	IN_MANT_ESCOLA_PRIVADA_SIND	Mantenedora da escola privada - Sindicatos de trabalhadores ou patronais, associações e cooperativas	Num	1	0 - Não 1 - Sim - Não aplicável para escolas públicas	
70	IN_MANT_ESCOLA_PRIVADA_SIST_S	Mantenedora da escola privada - Sistema S (Sesi, Senai, Sesc, outros)	Num	1	0 - Não 1 - Sim - Não aplicável para escolas públicas	
71	IN_MANT_ESCOLA_PRIVADA_S_FINS	Mantenedora da escola privada - Instituições sem fins lucrativos	Num	1	0 - Não 1 - Sim - Não aplicável para escolas públicas	
72	NU_CNPJ_ESCOLA_PRIVADA	Número do CNPJ da escola privada	Char	14	9999999999999 - Sem declaração	
73	NU_CNPJ_MANTENEDORA	Número do CNPJ da mantenedora principal da escola privada	Char	14	9999999999999 - Sem declaração	
74	TP_REGULAMENTACAO	Regulamentação/Autorização no conselho ou órgão municipal, estadual ou federal de educação	Num	1	0 - Não 1 - Sim 2 - Em tramitação	
75	TP_RESPONSAVEL_REGULAMENTACAO	Esfra administrativa do conselho ou órgão responsável pela Regulamentação/Autorização	Num	1	1 - Federal 2 - Estadual 3 - Municipal 4 - Estadual e Municipal 5 - Federal e Estadual 6 - Federal, Estadual e Municipal 9 - Não informado - Não aplicável para escolas sem regulamentação	
76	CO_ESCOLA_SEDE_VINCULADA	Código da escola sede	Num	8	- Não aplicável para escolas que não possuem vínculo	
77	CO_IES_OFERTANTE	Código da IES vinculada à escola	Num	14	- Não aplicável para escolas que não possuem vínculo	
78	IN_LOCAL_FUNC_PREDIO_ESCOLAR	Local de funcionamento da escola - Prédio Escolar	Num	1	0 - Não 1 - Sim	
79	TP_OCUUPACAO_PREDIO_ESCOLAR	Forma de ocupação do Prédio escolar	Num	1	1 - Próprio 2 - Alugado 3 - Cedido - Não aplicável para escolas que não ocupam prédio escolar	
80	IN_LOCAL_FUNC_SALAS_EMPRESA	Local de funcionamento da escola - Salas de empresa	Num	1	0 - Não 1 - Sim	
81	IN_LOCAL_FUNC_SOCIOEDUCATIVO	Local de funcionamento da escola - Unidade de Atendimento socioeducativo	Num	1	0 - Não 1 - Sim	
82	IN_LOCAL_FUNC_UNID_PRISIONAL	Local de funcionamento da escola - Unidade Prisional	Num	1	0 - Não 1 - Sim	
83	IN_LOCAL_FUNC_PRISIONAL_SOCIO	Local de funcionamento da escola - Unidade Prisional ou Unidade de atendimento socioeducativo	Num	1	0 - Não 1 - Sim	
84	IN_LOCAL_FUNC_TEMPLO_IGREJA	Local de funcionamento da escola - Templo/Igreja	Num	1	0 - Não 1 - Sim	
85	IN_LOCAL_FUNC_CASA_PROFESSOR	Local de funcionamento da escola - Casa do professor	Num	1	0 - Não 1 - Sim	
86	IN_LOCAL_FUNC_GALPAO	Local de funcionamento da escola - Galpão/Rancho/Paiol/Barracão	Num	1	0 - Não 1 - Sim	
87	TP_OCUUPACAO_GALPAO	Forma de ocupação do Galpão/Rancho/Paiol/Barracão	Num	1	1 - Próprio 2 - Alugado 3 - Cedido 9 - Não informado - Não aplicável para escolas que não ocupam Galpão/Rancho/Paiol/Barracão	
88	IN_LOCAL_FUNC_SALAS_OUTRA_ESC	Local de funcionamento da escola - Salas em outra escola	Num	1	0 - Não 1 - Sim	
89	IN_LOCAL_FUNC_OUTROS	Local de funcionamento da escola - Outros	Num	1	0 - Não 1 - Sim	
90	IN_PREDIO_COMPARTILHADO	Prédio compartilhado com outra escola	Num	1	0 - Não 1 - Sim - Não aplicável para escolas que não ocupam prédio escolar	
91	IN_AGUA_FILTERADA	Água consumida pelos alunos	Num	1	0 - Não 1 - Sim	
92	IN_AGUA_POTAVEL	Fornecê água potável para o consumo humano	Num	1	0 - Não 1 - Sim	
93	IN_AGUA_REDE_PUBLICA	Abastecimento de água - Rede pública	Num	1	0 - Não 1 - Sim	
94	IN_AGUA_POCO_ARTESIANO	Abastecimento de água - Poço artesiano	Num	1	0 - Não 1 - Sim	

95	IN_AGUA_CACIMBA	Abastecimento de água - Cacimba/Cisterna/Poço	Num	1	0 - Não 1 - Sim	
96	IN_AGUA_FONTE_RIO	Abastecimento de água - Fonte/Rio/Igarapé/Riacho/Corregos	Num	1	0 - Não 1 - Sim	
97	IN_AGUA_INEXISTENTE	Abastecimento de água - Não há abastecimento de água	Num	1	0 - Não 1 - Sim	
98	IN_AGUA_CARRO_PIPA	Abastecimento de água - Carro-pipa	Num	1	0 - Não 1 - Sim	
99	IN_ENERGIA_REDE_PUBLICA	Abastecimento de energia elétrica - Rede pública	Num	1	0 - Não 1 - Sim	
100	IN_ENERGIA_GERADOR	Abastecimento de energia elétrica - Gerador	Num	1	0 - Não 1 - Sim	
101	IN_ENERGIA_GERADOR_FOSSIL	Abastecimento de energia elétrica - Gerador movido a combustível fóssil	Num	1	0 - Não 1 - Sim	
102	IN_ENERGIA_OUTROS	Abastecimento de energia elétrica - Outros (Energia alternativa)	Num	1	0 - Não 1 - Sim	
103	IN_ENERGIA_RENOVAVEL	Abastecimento de energia elétrica - Fontes de energia renováveis ou alternativas (gerador a biocombustível e/ou biodigestores, eólica, solar, outras)	Num	1	0 - Não 1 - Sim	
104	IN_ENERGIA_INEXISTENTE	Abastecimento de energia elétrica - Não há energia elétrica	Num	1	0 - Não 1 - Sim	
105	IN_ESGOTO_REDE_PUBLICA	Esgoto sanitário - Rede pública	Num	1	0 - Não 1 - Sim	
106	IN_ESGOTO_FOSSA_SEPTICA	Esgoto sanitário - Fossa Séptica	Num	1	0 - Não 1 - Sim	
107	IN_ESGOTO_FOSSA_COMUM	Esgoto sanitário - Fossa rudimentar/comum	Num	1	0 - Não 1 - Sim	
108	IN_ESGOTO_FOSFA	Esgoto sanitário - Fossa	Num	1	0 - Não 1 - Sim	A partir de 2019: Variável derivada juntando IN_ESGOTO_FOSFA_SEPTICA e IN_ESGOTO_FOSFA_COMUM
109	IN_ESGOTO_INEXISTENTE	Esgoto sanitário - Não há esgotamento sanitário	Num	1	0 - Não 1 - Sim	
110	IN_LIXO_SERVICO_COLETA	Destinação do lixo - Serviço de coleta	Num	1	0 - Não 1 - Sim	
111	IN_LIXO_QUEIMA	Destinação do lixo - Queima	Num	1	0 - Não 1 - Sim	
112	IN_LIXO_ENTERRA	Destinação do lixo - Enterra	Num	1	0 - Não 1 - Sim	
113	IN_LIXO_DESTINO_FINAL_PUBLICO	Destinação do lixo - Leva a uma destinação final financiada pelo poder público	Num	1	0 - Não 1 - Sim	
114	IN_LIXO_DESCARTA_OUTRAAREA	Destinação do lixo - Descarta em outra área	Num	1	0 - Não 1 - Sim	
115	IN_LIXO_JOGA_OUTRAAREA	Destinação do lixo - Joga em outra área	Num	1	0 - Não 1 - Sim	
116	IN_LIXO_OUTROS	Destinação do lixo - Outros	Num	1	0 - Não 1 - Sim	
117	IN_LIXO_RECICLIA	Tratamento do lixo/resíduos que a escola realiza - Reciclagem	Num	1	0 - Não 1 - Sim	
118	IN_TRATAMENTO_LIXO_SEPARACAO	Tratamento do lixo/resíduos que a escola realiza - Separação do lixo/resíduos	Num	1	0 - Não 1 - Sim 9 - Não informado	
119	IN_TRATAMENTO_LIXO_REUSEITIZA	Tratamento do lixo/resíduos que a escola realiza - Reproveitamento/reutilização	Num	1	0 - Não 1 - Sim 9 - Não informado	
120	IN_TRATAMENTO_LIXO_RECICLAGEM	Tratamento do lixo/resíduos que a escola realiza - Reciclagem	Num	1	0 - Não 1 - Sim 9 - Não informado	
121	IN_TRATAMENTO_LIXO_INEXISTENTE	Tratamento do lixo/resíduos que a escola realiza - Não faz tratamento	Num	1	0 - Não 1 - Sim 9 - Não informado	
122	IN_ALMOXARIFADO	Dependências físicas existentes e utilizadas na escola - Almoxarifado	Num	1	0 - Não 1 - Sim	
123	IN_AREA_VERDE	Dependências físicas existentes e utilizadas na escola - Área de vegetação ou gramado	Num	1	0 - Não 1 - Sim	
124	IN_AREA_PLANTIO	Dependências físicas existentes e utilizadas na escola - Área de horta, plantio e/ou produção agrícola	Num	1	0 - Não 1 - Sim	
125	IN_AUDITORIO	Dependências físicas existentes e utilizadas na escola - Auditório	Num	1	0 - Não 1 - Sim	
126	IN_BANHEIRO_FORA_PREDIO	Dependências físicas existentes e utilizadas na escola - Banheiro fora do prédio	Num	1	0 - Não 1 - Sim	
127	IN_BANHEIRO_DENTRO_PREDIO	Dependências físicas existentes e utilizadas na escola - Banheiro dentro do prédio	Num	1	0 - Não 1 - Sim	
128	IN_BANHEIRO	Dependências físicas existentes e utilizadas na escola - Banheiro	Num	1	0 - Não 1 - Sim	
129	IN_BANHEIRO_EI	Dependências físicas existentes e utilizadas na escola - Banheiro adequado a educação infantil	Num	1	0 - Não 1 - Sim	
130	IN_BANHEIRO_PNE	Dependências físicas existentes e utilizadas na escola - Banheiro acessível, adequado ao uso de pessoas com deficiência ou mobilidade reduzida	Num	1	0 - Não 1 - Sim	
131	IN_BANHEIRO_FUNCIONARIOS	Dependências físicas existentes e utilizadas na escola - Banheiro exclusivo para os funcionários	Num	1	0 - Não 1 - Sim	
132	IN_BANHEIRO_CHUVEIRO	Dependências físicas existentes e utilizadas na escola - Banheiro ou vestuário com chuveiro	Num	1	0 - Não 1 - Sim	De 2012 a 2018: "Banheiro com chuveiro" A partir de 2019: "Banheiro ou vestuário com chuveiro"
133	IN_BERCARIO	Dependências físicas existentes e utilizadas na escola - Berçário	Num	1	0 - Não 1 - Sim	
134	IN_BIBLIOTECA	Dependências físicas existentes e utilizadas na escola - Biblioteca	Num	1	0 - Não 1 - Sim	
135	IN_BIBLIOTECA_SALA_LEITURA	Dependências físicas existentes e utilizadas na escola - Biblioteca e/ou Sala de leitura	Num	1	0 - Não 1 - Sim	
136	IN_COZINHA	Dependências físicas existentes e utilizadas na escola - Cozinha	Num	1	0 - Não 1 - Sim	
137	IN_DESPENSA	Dependências físicas existentes e utilizadas na escola - Despensa	Num	1	0 - Não 1 - Sim	
138	IN_DORMITORIO_ALUNO	Dependências físicas existentes e utilizadas na escola - Dormitório de aluno(a)	Num	1	0 - Não 1 - Sim	De 2012 a 2018: "Alojamento de aluno" A partir de 2019: "Dormitório de aluno(a)"
139	IN_DORMITORIO_PROFESSOR	Dependências físicas existentes e utilizadas na escola - Dormitório de professor(a)	Num	1	0 - Não 1 - Sim	De 2012 a 2018: "Alojamento de professor" A partir de 2019: "Dormitório de professor(a)"
140	IN_LABORATORIO_CIENCIAS	Dependências físicas existentes e utilizadas na escola - Laboratório de ciências	Num	1	0 - Não 1 - Sim	
141	IN_LABORATORIO_INFORMATICA	Dependências físicas existentes e utilizadas na escola - Laboratório de informática	Num	1	0 - Não 1 - Sim	
142	IN_LABORATORIO_EDUC_PROF	Dependências físicas existentes e utilizadas na escola - Laboratório específico para a Educação Profissional	Num	1	0 - Não 1 - Sim	
143	IN_PATIO_COBERTO	Dependências físicas existentes e utilizadas na escola - Pátio coberto	Num	1	0 - Não 1 - Sim	
144	IN_PATIO_DESCOBERTO	Dependências físicas existentes e utilizadas na escola - Pátio descoberto	Num	1	0 - Não 1 - Sim	
145	IN_PARQUE_INFANTIL	Dependências físicas existentes e utilizadas na escola - Parque infantil	Num	1	0 - Não 1 - Sim	
146	IN_PISCINA	Dependências físicas existentes e utilizadas na escola - Piscina	Num	1	0 - Não 1 - Sim	
147	IN_QUADRA_ESPORTES	Dependências físicas existentes e utilizadas na escola - Quadra de esportes coberta ou descoberta	Num	1	0 - Não 1 - Sim	
148	IN_QUADRA_ESPORTES_COBERTA	Dependências físicas existentes e utilizadas na escola - Quadra de esportes coberta	Num	1	0 - Não 1 - Sim	
149	IN_QUADRA_ESPORTES_DESCOBERTA	Dependências físicas existentes e utilizadas na escola - Quadra de esportes descoberta	Num	1	0 - Não 1 - Sim	
150	IN_REFETORIO	Dependências físicas existentes e utilizadas na escola - Refeitório	Num	1	0 - Não 1 - Sim	
151	IN_SALA_ATELIE_ARTES	Dependências físicas existentes e utilizadas na escola - Sala/Ateliê de artes	Num	1	0 - Não 1 - Sim	
152	IN_SALA_MUSICA_CORAL	Dependências físicas existentes e utilizadas na escola - Sala de música/coral	Num	1	0 - Não 1 - Sim	
153	IN_SALA_ESTUDIO_DANCA	Dependências físicas existentes e utilizadas na escola - Sala/estúdio de dança	Num	1	0 - Não 1 - Sim	
154	IN_SALA_MULTIUSO	Dependências físicas existentes e utilizadas na escola - Sala multiuso (música, dança e artes)	Num	1	0 - Não 1 - Sim	
155	IN_SALA_ESTUDIO_GRAVACAO	Dependências físicas existentes e utilizadas na escola - Estúdio de gravação e edição	Num	1	0 - Não 1 - Sim	
156	IN_SALA_OFICINAS EDUC PROF	Dependências físicas existentes e utilizadas na escola - Salas de oficinas da Educação Profissional	Num	1	0 - Não 1 - Sim	
157	IN_SALA_DIRETORIA	Dependências físicas existentes e utilizadas na escola - Sala de Diretoria	Num	1	0 - Não 1 - Sim	
158	IN_SALA_LEITURA	Dependências físicas existentes e utilizadas na escola - Sala de leitura	Num	1	0 - Não 1 - Sim	
159	IN_SALA_PROFESSOR	Dependências físicas existentes e utilizadas na escola - Sala de professores	Num	1	0 - Não 1 - Sim	
160	IN_SALA_REPOUSO_ALUNO	Dependências físicas existentes e utilizadas na escola - Sala de repouso para alun(a)	Num	1	0 - Não 1 - Sim	
161	IN_SECRETARIA	Dependências físicas existentes e utilizadas na escola - Sala de Secretaria	Num	1	0 - Não 1 - Sim	

162	IN_SALA_ATENDIMENTO_ESPECIAL	Dependências físicas existentes e utilizadas na escola - Sala de Recursos Multifuncionais para Atendimento Educacional Especializado (AEE)	Num	1	0 - Não 1 - Sim	
163	IN_TERREIRAO	Dependências físicas existentes e utilizadas na escola - Terreirão (área para prática desportiva e recreação sem cobertura, sem piso e sem edificações)	Num	1	0 - Não 1 - Sim	
164	IN_VIVEIRO	Dependências físicas existentes e utilizadas na escola - Viveiro/criação de animais	Num	1	0 - Não 1 - Sim	
165	IN_DEPENDENCIAS_PNE	Dependências físicas existentes e utilizadas na escola - Dependências e vias adequadas a alunos com deficiência ou mobilidade reduzida	Num	1	0 - Não 1 - Sim	
166	IN_LAVANDERIA	Dependências físicas existentes e utilizadas na escola - Lavanderia	Num	1	0 - Não 1 - Sim	
167	IN_DEPENDENCIAS_OUTRAS	Dependências existentes na escola - Nenhuma das dependências relacionadas	Num	1	0 - Não 1 - Sim	
168	IN_ACESSIBILIDADE_CORRIMAO	Recursos de acessibilidade para pessoas com deficiência ou mobilidade reduzida nas vias de circulação interna na escola - Corrimão e guarda corpos	Num	1	0 - Não 1 - Sim	
169	IN_ACESSIBILIDADE_ELEVADOR	Recursos de acessibilidade para pessoas com deficiência ou mobilidade reduzida nas vias de circulação interna na escola - Elevador	Num	1	0 - Não 1 - Sim	
170	IN_ACESSIBILIDADE_PISOS_TATEIS	Recursos de acessibilidade para pessoas com deficiência ou mobilidade reduzida nas vias de circulação interna na escola - Pisos táticos	Num	1	0 - Não 1 - Sim	
171	IN_ACESSIBILIDADE_VAO_LIVRE	Recursos de acessibilidade para pessoas com deficiência ou mobilidade reduzida nas vias de circulação interna na escola - Portas com vão livre de, no mínimo, 80 cm	Num	1	0 - Não 1 - Sim	
172	IN_ACESSIBILIDADE_RAMPAS	Recursos de acessibilidade para pessoas com deficiência ou mobilidade reduzida nas vias de circulação interna na escola - Rampa	Num	1	0 - Não 1 - Sim	
173	IN_ACESSIBILIDADE_SINAL SONORO	Recursos de acessibilidade para pessoas com deficiência ou mobilidade reduzida nas vias de circulação interna na escola - Sinalização sonora	Num	1	0 - Não 1 - Sim	
174	IN_ACESSIBILIDADE_SINAL_TATIL	Recursos de acessibilidade para pessoas com deficiência ou mobilidade reduzida nas vias de circulação interna na escola - Sinalização tátil (piso/paredes)	Num	1	0 - Não 1 - Sim	
175	IN_ACESSIBILIDADE_SINAL_VISUAL	Recursos de acessibilidade para pessoas com deficiência ou mobilidade reduzida nas vias de circulação interna na escola - Sinalização visual (piso/paredes)	Num	1	0 - Não 1 - Sim	
176	IN_ACESSIBILIDADE_INEXISTENTE	Recursos de acessibilidade para pessoas com deficiência ou mobilidade reduzida nas vias de circulação interna na escola - Nenhum dos recursos de acessibilidade listados	Num	1	0 - Não 1 - Sim	
177	IN_ACESSIBILIDADE_SINALIZACAO	Recursos de acessibilidade para pessoas com deficiência ou mobilidade reduzida nas vias de circulação interna na escola - Sinalização/alarme luminoso	Num	1	0 - Não 1 - Sim	
178	QT_SALAS_EXISTENTES	Número de salas de aula existentes na escola	Num	4		
179	QT_SALAS_UTILIZADAS_DENTRO	Número de salas de aula utilizadas na escola - Dentro do prédio	Num	4		
180	QT_SALAS_UTILIZADAS_FORA	Número de salas de aula utilizadas na escola - Fora do prédio	Num	4		
181	QT_SALAS_UTILIZADAS	Número de salas de aula utilizadas na escola (dentro e fora do prédio)	Num	4		A partir de 2019: Variável derivada juntando QT_SALAS_UTILIZADAS_DENTRO e QT_SALAS_UTILIZADAS_FORA
182	QT_SALAS_UTILIZA_CLIMATIZADAS	Condições das salas de aula utilizadas na escola (dentro e fora do prédio escolar) - Número de salas de aula climatizadas	Num	4		
183	QT_SALAS_UTILIZADAS_ACESSEVIS	Condições das salas de aula utilizadas na escola (dentro e fora do prédio escolar) - Número de salas de aula com acessibilidade para pessoas com deficiência ou mobilidade reduzida	Num	4		
184	IN_EQUIP_PARABOLICA	Equipamentos existentes na escola para uso técnico e administrativo - Antena parabólica	Num	1	0 - Não 1 - Sim	
185	IN_COMPUTADOR	Equipamentos existentes na escola para uso técnico e administrativo - Computador	Num	1	0 - Não 1 - Sim	
186	IN_EQUIP_COPIADORA	Equipamentos existentes na escola para uso técnico e administrativo - Copiadora	Num	1	0 - Não 1 - Sim	
187	IN_EQUIP_IMPRESSORA	Equipamentos existentes na escola para uso técnico e administrativo - Impressora	Num	1	0 - Não 1 - Sim	
188	IN_EQUIP_IMPRESSORA_MULT	Equipamentos existentes na escola para uso técnico e administrativo - Impressora Multifuncional	Num	1	0 - Não 1 - Sim	
189	IN_EQUIP_SCANNER	Equipamentos existentes na escola para uso técnico e administrativo - Scanner	Num	1	0 - Não 1 - Sim	
190	IN_EQUIP_NENHUM	Nenhum dos equipamentos listados para uso técnico e administrativo - Antena parabólica, Computador, Copiadora, Impressora, Impressora Multifuncional ou Scanner	Num	1	0 - Não 1 - Sim	
191	IN_EQUIP_DVD	Equipamentos existentes na escola para o processo ensino e aprendizagem - DVD/Blu-ray	Num	1	0 - Não 1 - Sim	
192	QT_EQUIP_DVD	Quantidade de Aparelhos de DVD/Blu-ray	Num	5	88888 - registro com marcação de valor extremo (valor superior ao limite máximo de 4 equipamentos para cada 3 salas existentes - foram marcados apenas valores>3)	Em 2019: primeiro ano em que foi aplicado tratamento de valores extremos, que são marcados com o código 88888.
193	IN_EQUIP_SOM	Equipamentos existentes na escola para o processo ensino e aprendizagem - Aparelho de som	Num	1	0 - Não 1 - Sim	
194	QT_EQUIP_SOM	Quantidade de Aparelhos de som	Num	5	88888 - registro com marcação de valor extremo (valor superior ao limite máximo de 4 equipamentos para cada 3 salas existentes - foram marcados apenas valores>3)	Em 2019: primeiro ano em que foi aplicado tratamento de valores extremos, que são marcados com o código 88888.
195	IN_EQUIP_TV	Equipamentos existentes na escola para o processo ensino e aprendizagem - Aparelho de televisão	Num	1	0 - Não 1 - Sim	
196	QT_EQUIP_TV	Quantidade de Aparelhos de televisão	Num	5	88888 - registro com marcação de valor extremo (valor superior ao limite máximo de 4 equipamentos para cada 3 salas existentes - foram marcados apenas valores>3)	Em 2019: primeiro ano em que foi aplicado tratamento de valores extremos, que são marcados com o código 88888.
197	IN_EQUIP_LOUSA_DIGITAL	Equipamentos existentes na escola para o processo ensino e aprendizagem - Lousa digital	Num	1	0 - Não 1 - Sim	
198	QT_EQUIP_LOUSA_DIGITAL	Quantidade de Lousas digitais	Num	5	88888 - registro com marcação de valor extremo (valor superior ao limite máximo de 4 equipamentos para cada 3 salas existentes - foram marcados apenas valores>3)	Em 2019: primeiro ano em que foi aplicado tratamento de valores extremos, que são marcados com o código 88888.
199	IN_EQUIP_MULTIMIDIA	Equipamentos existentes na escola para o processo ensino e aprendizagem - Projetor Multimídia (Datashow)	Num	1	0 - Não 1 - Sim	
200	QT_EQUIP_MULTIMIDIA	Quantidade de Projetores Multimídia (Datashow)	Num	5	88888 - registro com marcação de valor extremo (valor superior ao limite máximo de 4 equipamentos para cada 3 salas existentes - foram marcados apenas valores>3)	Em 2019: primeiro ano em que foi aplicado tratamento de valores extremos, que são marcados com o código 88888.
201	IN_EQUIP_VIDEOCASSETTE	Equipamentos existentes na escola - Videocassete	Num	1	0 - Não 1 - Sim	
202	IN_EQUIP_RETROPROJETOR	Equipamentos existentes na escola - Retroprojetor	Num	1	0 - Não 1 - Sim	
203	IN_EQUIP_FAX	Equipamentos existentes na escola - Fax	Num	1	0 - Não 1 - Sim	
204	IN_EQUIP_FOTO	Equipamentos existentes na escola - Máquina fotográfica/filmadora	Num	1	0 - Não 1 - Sim	
205	QT_EQUIP_VIDEOCASSETTE	Quantidade de Videocassetes	Num	4		
206	QT_EQUIP_PARABOLICA	Quantidade de Antenas parabólicas	Num	4		
207	QT_EQUIP_COPIADORA	Quantidade de Copiadoras	Num	4		
208	QT_EQUIP_RETROPROJETOR	Quantidade de Retroprojetores	Num	4		
209	QT_EQUIP_IMPRESSORA	Quantidade de Impressoras	Num	4		
210	QT_EQUIP_IMPRESSORA_MULT	Quantidade de Impressoras Multifuncionais	Num	4		
211	QT_EQUIP_FAX	Quantidade de Fax	Num	4		
212	QT_EQUIP_FOTO	Quantidade de Máquinas Fotográficas/ Filmadoras	Num	4		
213	QT_COMP_ALUNO	Quantidade de computadores em uso pelos alunos	Num	4	-Não aplicável para escolas que não possuem computador	
214	IN_DESKTOP_ALUNO	Computadores em uso pelos alunos - Computador de mesa (desktop)	Num	1	0 - Não 1 - Sim	
215	QT_DESKTOP_ALUNO	Quantidade de computadores em uso pelos alunos - Computador de mesa (desktop)	Num	5	88888 - registro com marcação de valor extremo (valor superior ao limite máximo de 4 equipamentos para cada 3 matrículas - foram marcados apenas valores>3)	Em 2019: primeiro ano em que foi aplicado tratamento de valores extremos, que são marcados com o código 88888.

216	IN_COMP_PORTATIL_ALUNO	Computadores em uso pelos alunos - Computador portátil	Num	1	0 - Não 1 - Sim	
217	QT_COMP_PORTATIL_ALUNO	Quantidade de computadores em uso pelos alunos - Computador portátil	Num	5	88888 - registro com marcação de valor extremo (valor superior ao limite máximo de 4 equipamentos para cada 3 matrículas - foram marcados apenas valores>3)	Em 2019: primeiro ano em que foi aplicado tratamento de valores extremos, que são marcados com o código 88888.
218	IN_TABLET_ALUNO	Computadores em uso pelos alunos - Tablet	Num	1	0 - Não 1 - Sim	
219	QT_TABLET_ALUNO	Quantidade de computadores em uso pelos alunos - Tablet	Num	5	88888 - registro com marcação de valor extremo (valor superior ao limite máximo de 4 equipamentos para cada 3 matrículas - foram marcados apenas valores>3)	Em 2019: primeiro ano em que foi aplicado tratamento de valores extremos, que são marcados com o código 88888.
220	QT_COMPUTADOR	Quantidade de computadores na escola	Num	4		
221	QT_COMP_ADMINISTRATIVO	Quantidade de computadores de uso administrativo	Num	4		
222	IN_INTERNET	Acesso à Internet	Num	1	0 - Não 1 - Sim	
223	IN_INTERNET_ALUNOS	Acesso à Internet - Para uso dos alunos	Num	1	0 - Não 1 - Sim	
224	IN_INTERNET_ADMINISTRATIVO	Acesso à Internet - Para uso administrativo	Num	1	0 - Não 1 - Sim	
225	IN_INTERNET_APRENDIZAGEM	Acesso à Internet - Para uso nos processos de ensino e aprendizagem	Num	1	0 - Não 1 - Sim	
226	IN_INTERNET_COMMUNIDADE	Acesso à Internet - Para uso da comunidade	Num	1	0 - Não 1 - Sim	
227	IN_ACESSO_INTERNET_COMPUTADOR	Equipamentos que os alunos usam para acessar a internet da escola - Computadores de mesa, portáteis e tablets da escola (no laboratório de informática, biblioteca, sala de aula etc.)	Num	1	0 - Não 1 - Sim 9 - Não informado	
228	IN_ACES_INTERNET_DISP_PESSOAS	Equipamentos que os alunos usam para acessar a internet da escola - Dispositivos pessoais (computadores portáteis, celulares, tablets etc.)	Num	1	0 - Não 1 - Sim 9 - Não informado	
229	TP_REDE_LOCAL	Rede local de interligação de computadores	Num	1	0 - Não há rede local interligando computadores 1 - A Cabo 2 - Wireless 3 - A Cabo e Wireless 9 - Não informado	
230	IN_BANDA_LARGA	Internet Banda Larga	Num	1	0 - Não 1 - Sim - Não aplicável para escolas sem acesso à internet	
231	QT_FUNCIONARIOS	Total de funcionários da escola (inclusive profissionais escolares em sala de aula)	Num	4		
232	IN_PROF_ADMINISTRATIVOS	Profissionais que atuam na escola - Auxiliares de secretaria ou auxiliares administrativos, atendentes	Num	1	0 - Não 1 - Sim	
233	QT_PROF_ADMINISTRATIVOS	Quantidade de profissionais que atuam na escola - Auxiliares de secretaria ou auxiliares administrativos, atendentes	Num	5	88888 - registro com marcação de valor extremo (valor superior ao limite máximo* definido com base na distribuição da razão de profissionais por matrícula)	Em 2019: primeiro ano em que foi aplicado tratamento de valores extremos, que são marcados com o código 88888.
234	IN_PROF_SERVICOS_GERAIS	Profissionais que atuam na escola - Auxiliar de serviços gerais, porteiros(a), zelador(a), faxineiro(a), jardinero(a)	Num	1	0 - Não 1 - Sim	
235	QT_PROF_SERVICOS_GERAIS	Total de profissionais que atuam na escola - Auxiliar de serviços gerais, porteiros(a), zelador(a), faxineiro(a), jardinero(a)	Num	5	88888 - registro com marcação de valor extremo (valor superior ao limite máximo* definido com base na distribuição da razão de profissionais por matrícula)	Em 2019: primeiro ano em que foi aplicado tratamento de valores extremos, que são marcados com o código 88888.
236	IN_PROF_BIBLIOTECARIO	Profissionais que atuam na escola - Bibliotecário(a), auxiliar de biblioteca ou monitor(a) da sala de leitura	Num	1	0 - Não 1 - Sim	
237	QT_PROF_BIBLIOTECARIO	Quantidade de profissionais que atuam na escola - Bibliotecário(a), auxiliar de biblioteca ou monitor(a) da sala de leitura	Num	5	88888 - registro com marcação de valor extremo (valor superior ao limite máximo* definido com base na distribuição da razão de profissionais por matrícula)	Em 2019: primeiro ano em que foi aplicado tratamento de valores extremos, que são marcados com o código 88888.
238	IN_PROF_SAUDE	Profissionais que atuam na escola - Bombeiro(a) brigadista, profissionais de assistência à saúde (urgência e emergência), Enfermeiro(a), Técnico(a) de enfermagem e socorrista	Num	1	0 - Não 1 - Sim	
239	QT_PROF_SAUDE	Quantidade de profissionais que atuam na escola - Bombeiro(a) brigadista, profissionais de assistência à saúde (urgência e emergência), Enfermeiro(a), Técnico(a) de enfermagem e socorrista	Num	5	88888 - registro com marcação de valor extremo (valor superior ao limite máximo* definido com base na distribuição da razão de profissionais por matrícula)	Em 2019: primeiro ano em que foi aplicado tratamento de valores extremos, que são marcados com o código 88888.
240	IN_PROF_COORDENADOR	Profissionais que atuam na escola - Coordenador(a) de turno/disciplina	Num	1	0 - Não 1 - Sim	
241	QT_PROF_COORDENADOR	Quantidade de profissionais que atuam na escola - Coordenador(a) de turno/disciplina	Num	5	88888 - registro com marcação de valor extremo (valor superior ao limite máximo* definido com base na distribuição da razão de profissionais por matrícula)	Em 2019: primeiro ano em que foi aplicado tratamento de valores extremos, que são marcados com o código 88888.
242	IN_PROF_FONAUDIOLOGO	Profissionais que atuam na escola - Fonoaudiólogo(a)	Num	1	0 - Não 1 - Sim	
243	QT_PROF_FONAUDIOLOGO	Quantidade de profissionais que atuam na escola - Fonoaudiólogo(a)	Num	5	88888 - registro com marcação de valor extremo (valor superior ao limite máximo* definido com base na distribuição da razão de profissionais por matrícula)	Em 2019: primeiro ano em que foi aplicado tratamento de valores extremos, que são marcados com o código 88888.
244	IN_PROF_NUTRICIONISTA	Profissionais que atuam na escola - Nutricionista	Num	1	0 - Não 1 - Sim	
245	QT_PROF_NUTRICIONISTA	Quantidade de profissionais que atuam na escola - Nutricionista	Num	5	88888 - registro com marcação de valor extremo (valor superior ao limite máximo* definido com base na distribuição da razão de profissionais por matrícula)	Em 2019: primeiro ano em que foi aplicado tratamento de valores extremos, que são marcados com o código 88888.
246	IN_PROF_PSICOLOGO	Profissionais que atuam na escola - Psicólogo(a) Escolar	Num	1	0 - Não 1 - Sim	
247	QT_PROF_PSICOLOGO	Quantidade de profissionais que atuam na escola - Psicólogo(a) Escolar	Num	5	88888 - registro com marcação de valor extremo (valor superior ao limite máximo* definido com base na distribuição da razão de profissionais por matrícula)	Em 2019: primeiro ano em que foi aplicado tratamento de valores extremos, que são marcados com o código 88888.
248	IN_PROF_ALIMENTACAO	Profissionais que atuam na escola - Profissionais de preparação e segurança alimentar, cozinheiro(a), merendeiro(a) e auxiliar de cozinha	Num	1	0 - Não 1 - Sim	
249	QT_PROF_ALIMENTACAO	Quantidade de profissionais que atuam na escola - Profissionais de preparação e segurança alimentar, cozinheiro(a), merendeiro(a) e auxiliar de cozinha	Num	5	88888 - registro com marcação de valor extremo (valor superior ao limite máximo* definido com base na distribuição da razão de profissionais por matrícula)	Em 2019: primeiro ano em que foi aplicado tratamento de valores extremos, que são marcados com o código 88888.
250	IN_PROF_PEDAGOGIA	Profissionais que atuam na escola - Profissionais de apoio e supervisão pedagógica: pedagogo(s), coordenador(s) pedagógico(s), orientador(s) educacional, supervisor(s) escolar e coordenador(es) de área de ensino	Num	1	0 - Não 1 - Sim	
251	QT_PROF_PEDAGOGIA	Quantidade de profissionais que atuam na escola - Profissionais de apoio e supervisão pedagógica: pedagogo(s), coordenador(s) pedagógico(s), orientador(s) educacional, supervisor(s) escolar e coordenador(es) de área de ensino	Num	5	88888 - registro com marcação de valor extremo (valor superior ao limite máximo* definido com base na distribuição da razão de profissionais por matrícula)	Em 2019: primeiro ano em que foi aplicado tratamento de valores extremos, que são marcados com o código 88888.
252	IN_PROF_SECRETARIO	Profissionais que atuam na escola - Secretário(a) escolar	Num	1	0 - Não 1 - Sim	
253	QT_PROF_SECRETARIO	Quantidade de profissionais que atuam na escola - Secretário(a) escolar	Num	5	88888 - registro com marcação de valor extremo (valor superior ao limite máximo* definido com base na distribuição da razão de profissionais por matrícula)	Em 2019: primeiro ano em que foi aplicado tratamento de valores extremos, que são marcados com o código 88888.
254	IN_PROF_SEGURANCA	Profissionais que atuam na escola - Segurança, guarda ou segurança patrimonial	Num	1	0 - Não 1 - Sim	
255	QT_PROF_SEGURANCA	Quantidade de profissionais que atuam na escola - Segurança, guarda ou segurança patrimonial	Num	5	88888 - registro com marcação de valor extremo (valor superior ao limite máximo* definido com base na distribuição da razão de profissionais por matrícula)	Em 2019: primeiro ano em que foi aplicado tratamento de valores extremos, que são marcados com o código 88888.
256	IN_PROF_MONITORES	Profissionais que atuam na escola - Técnicos(as), monitores(as), supervisores(as) ou auxiliares de laboratório(s), de apoio a tecnologias educacionais ou em multimídia/multimídias eletrônico/digitais	Num	1	0 - Não 1 - Sim	
257	QT_PROF_MONITORES	Quantidade de profissionais que atuam na escola - Técnicos(as), monitores(as), supervisores(as) ou auxiliares de laboratório(s), de apoio a tecnologias educacionais ou em multimídia/multimídias eletrônico/digitais	Num	5	88888 - registro com marcação de valor extremo (valor superior ao limite máximo* definido com base na distribuição da razão de profissionais por matrícula)	Em 2019: primeiro ano em que foi aplicado tratamento de valores extremos, que são marcados com o código 88888.
258	IN_PROF_GESTAO	Profissionais que atuam na escola - Vice-diretor(a) ou diretor(a) adjunto(a), profissionais responsáveis pela gestão administrativa e/ou financeira	Num	1	0 - Não 1 - Sim	
259	QT_PROF_GESTAO	Quantidade de profissionais que atuam na escola - Vice-diretor(a) ou diretor(a) adjunto(a), profissionais responsáveis pela gestão administrativa e/ou financeira	Num	5	88888 - registro com marcação de valor extremo (valor superior ao limite máximo* definido com base na distribuição da razão de profissionais por matrícula)	Em 2019: primeiro ano em que foi aplicado tratamento de valores extremos, que são marcados com o código 88888.
260	IN_PROF_ASSIST_SOCIAL	Profissionais que atuam na escola - Orientador(a) comunitário(a) ou assistente social	Num	1	0 - Não 1 - Sim	
261	QT_PROF_ASSIST_SOCIAL	Quantidade de profissionais que atuam na escola - Orientador(a) comunitário(a) ou assistente social	Num	5	88888 - registro com marcação de valor extremo (valor superior ao limite máximo* definido com base na distribuição da razão de profissionais por matrícula)	Em 2019: primeiro ano em que foi aplicado tratamento de valores extremos, que são marcados com o código 88888.

262	IN_PROF_TRAD_LIBRAS	Profissionais que atuam na escola - Tradutor e Intérprete de Libras para atendimento em outros ambientes da escola que não seja sala de aula	Num	5	0 - Não 1 - Sim	
263	QT_PROF_TRAD_LIBRAS	Quantidade de profissionais que atuam na escola - Tradutor e Intérprete de Libras para atendimento em outros ambientes da escola que não seja sala de aula	Num	5	88888 - registro com marcação de valor extremo (valor superior ao limite máximo* definido com base na distribuição da razão de profissionais por matrícula de aluno com surdez, deficiência auditiva ou surdocegueira).	Em 2023: primeiro ano em que foi aplicado tratamento de valores extremos, que são marcados com o código 88888.
264	IN_PROF_AGRICOLA	Profissionais que atuam na escola - Agrônomo(as), horticultores(as), técnicos ou monitores(as) responsáveis pela gestão da área de horta, plantio e/ou produção agrícola	Num	5	0 - Não 1 - Sim	
265	QT_PROF_AGRICOLA	Quantidade de profissionais que atuam na escola - Agrônomo(as), horticultores(as), técnicos ou monitores(as) responsáveis pela gestão da área de horta, plantio e/ou produção agrícola	Num	5	88888 - registro com marcação de valor extremo (valor superior ao limite máximo* definido com base na distribuição da razão de profissionais por matrícula).	
266	IN_PROF_REVISOR_BRALLE	Profissionais que atuam na escola - Revisor de texto Braille, assistente vidente (assistente de revisão do texto em Braille)	Num	5	0 - Não 1 - Sim	
267	QT_PROF_REVISOR_BRALLE	Quantidade de profissionais que atuam na escola - Revisor de texto Braille, assistente vidente (assistente de revisão do texto em Braille)	Num	5	88888 - registro com marcação de valor extremo (valor superior ao limite máximo* definido com base na distribuição da razão de profissionais por matrícula de aluno com cegueira, surdo-cegueira ou baixa visão).	
268	IN_ALIMENTACAO	Alimentação escolar para os alunos - PNae/FNDE	Num	1	0 - Não oferece 1 - Oferece	
269	IN_SERIE_ANO	Forma de organização do ensino - Série/Ano (séries anuais)	Num	1	0 - Não 1 - Sim 9 - Não informado - Não aplicável para escolas sem matrículas de escolarização	
270	IN_PERIODOS_SEMESTRAIS	Forma de organização do ensino - Períodos semestrais	Num	1	0 - Não 1 - Sim 9 - Não informado - Não aplicável para escolas sem matrículas de escolarização	
271	IN_FUNDAMENTAL_CICLOS	Forma de organização do ensino - Ciclo(s) do Ensino Fundamental	Num	1	0 - Não 1 - Sim 9 - Não informado - Não aplicável para escolas sem matrículas de escolarização	
272	IN_GRUPOS_NAO_SERIADOS	Forma de organização do ensino - Grupos não-seriados com base na idade ou competência (art. 23 LDB)	Num	1	0 - Não 1 - Sim 9 - Não informado - Não aplicável para escolas sem matrículas de escolarização	
273	IN_MODULOS	Forma de organização do ensino - Módulos	Num	1	0 - Não 1 - Sim 9 - Não informado - Não aplicável para escolas sem matrículas de escolarização	
274	IN_FORMACAO_ALTERNANCIA	Forma de organização do ensino - Alternância regular de períodos de estudos (proposta pedagógica de formação por alternância com tempo-escola e tempo-comunidade)	Num	1	0 - Não 1 - Sim 9 - Não informado - Não aplicável para escolas sem matrículas de escolarização	
275	IN_MATERIAL_PED_MULTIMIDIA	Instrumentos e materiais socioculturais e/ou pedagógicos em uso na escola para o desenvolvimento de atividades de ensino e aprendizagem - Acervo multimídia	Num	1	0 - Não 1 - Sim	Em 2019: Existe adicionalmente a categoria "9 - Não informado".
276	IN_MATERIAL_PED_INFANTIL	Instrumentos e materiais socioculturais e/ou pedagógicos em uso na escola para o desenvolvimento de atividades de ensino e aprendizagem - Brinquedos para Educação Infantil	Num	1	0 - Não 1 - Sim	Em 2019: Existe adicionalmente a categoria "9 - Não informado".
277	IN_MATERIAL_PED_CIENTIFICO	Instrumentos e materiais socioculturais e/ou pedagógicos em uso na escola para o desenvolvimento de atividades de ensino e aprendizagem - Conjunto de materiais científicos	Num	1	0 - Não 1 - Sim	Em 2019: Existe adicionalmente a categoria "9 - Não informado".
278	IN_MATERIAL_PED_DIFUSAO	Instrumentos e materiais socioculturais e/ou pedagógicos em uso na escola para o desenvolvimento de atividades de ensino e aprendizagem - Equipamento para amplificação e difusão de som/audio	Num	1	0 - Não 1 - Sim	Em 2019: Existe adicionalmente a categoria "9 - Não informado".
279	IN_MATERIAL_PED_MUSICAL	Instrumentos e materiais socioculturais e/ou pedagógicos em uso na escola para o desenvolvimento de atividades de ensino e aprendizagem - Instrumentos musicais para conjunto, banda/fanfarra e/ou aulas de música	Num	1	0 - Não 1 - Sim	Em 2019: Existe adicionalmente a categoria "9 - Não informado".
280	IN_MATERIAL_PED_JOGOS	Instrumentos e materiais socioculturais e/ou pedagógicos em uso na escola para o desenvolvimento de atividades de ensino e aprendizagem - Jogos Educativos	Num	1	0 - Não 1 - Sim	Em 2019: Existe adicionalmente a categoria "9 - Não informado".
281	IN_MATERIAL_PED_ARTISTICAS	Instrumentos e materiais socioculturais e/ou pedagógicos em uso na escola para o desenvolvimento de atividades de ensino e aprendizagem - Materiais para atividades culturais e artísticas	Num	1	0 - Não 1 - Sim	Em 2019: Existe adicionalmente a categoria "9 - Não informado".
282	IN_MATERIAL_PED_PROFISSIONAL	Instrumentos e materiais socioculturais e/ou pedagógicos em uso na escola para o desenvolvimento de atividades de ensino e aprendizagem - Materiais para Educação Profissional	Num	1	0 - Não 1 - Sim	
283	IN_MATERIAL_PED_DESPORTIVA	Instrumentos e materiais socioculturais e/ou pedagógicos em uso na escola para o desenvolvimento de atividades de ensino e aprendizagem - Materiais para prática desportiva e recreação	Num	1	0 - Não 1 - Sim	Em 2019: Existe adicionalmente a categoria "9 - Não informado".
284	IN_MATERIAL_PED_INDIGENA	Instrumentos e materiais socioculturais e/ou pedagógicos em uso na escola para o desenvolvimento de atividades de ensino e aprendizagem - Indígena	Num	1	0 - Não 1 - Sim	Em 2019: Existe adicionalmente a categoria "9 - Não informado".
285	IN_MATERIAL_PED_ETNICO	Instrumentos e materiais socioculturais e/ou pedagógicos em uso na escola para o desenvolvimento de atividades de ensino e aprendizagem - Materiais pedagógicos para a educação das relações étnico-raciais	Num	1	0 - Não 1 - Sim	Em 2019: Existe adicionalmente a categoria "9 - Não informado".
286	IN_MATERIAL_PED_CAMPO	Instrumentos e materiais socioculturais e/ou pedagógicos em uso na escola para o desenvolvimento de atividades de ensino e aprendizagem - Materiais pedagógicos para a educação do campo	Num	1	0 - Não 1 - Sim	Em 2019: Existe adicionalmente a categoria "9 - Não informado".
287	IN_MATERIAL_PED_BIL_SURDOS	Instrumentos, materiais socioculturais e/ou pedagógicos em uso na escola para o desenvolvimento de atividades de ensino aprendizagem - Materiais pedagógicos para a educação bilíngue de surdos	Num	1	0 - Não 1 - Sim	
288	IN_MATERIAL_PED_AGRICOLA	Instrumentos, materiais socioculturais e/ou pedagógicos em uso na escola para o desenvolvimento de atividades de ensino aprendizagem - Equipamentos e instrumentos para atividades em área de horta, plantio e/ou produção agrícola	Num	1	0 - Não 1 - Sim	
289	IN_MATERIAL_PED_QUILOMBOLA	Instrumentos, materiais socioculturais e/ou pedagógicos em uso na escola para o desenvolvimento de atividades de ensino aprendizagem - Materiais pedagógicos para a educação escolar quilombola	Num	1	0 - Não 1 - Sim	
290	IN_MATERIAL_PED_EDU_ESP	Instrumentos, materiais socioculturais e/ou pedagógicos em uso na escola para o desenvolvimento de atividades de ensino aprendizagem - Materiais pedagógicos para a educação especial	Num	1	0 - Não 1 - Sim	
291	IN_MATERIAL_PED_NENHUM	Instrumentos e materiais socioculturais e/ou pedagógicos em uso na escola para o desenvolvimento de atividades de ensino e aprendizagem - Nenhum	Num	1	0 - Não 1 - Sim	
292	IN_MATERIAL_ESP_QUILOMBOLA	Materiais didáticos específicos para atendimento à diversidade sociocultural - Quilombolas	Num	1	0 - Não 1 - Sim	
293	IN_MATERIAL_ESP_INDIGENA	Materiais didáticos específicos para atendimento à diversidade sociocultural - Indígena	Num	1	0 - Não 1 - Sim	
294	IN_MATERIAL_ESP_NAO_UTILIZA	Materiais didáticos específicos para atendimento à diversidade sociocultural - Não utiliza	Num	1	0 - Não 1 - Sim	
295	IN_EDUCACAO_INDIGENA	Escola Indígena	Num	1	0 - Não 1 - Sim	
296	TP_INDIGENA_LINGUA	Escola Indígena - Língua em que o ensino é ministrado (apenas para escola indígena)	Num	1	1 - Somente em Língua Indígena 2 - Somente em Língua Portuguesa 3 - Em Língua Indígena e em Língua Portuguesa - não aplicável (aplicável apenas para Escola Indígena)	
297	CO_LINGUA_INDIGENA_1	Escola Indígena - Língua em que o ensino é ministrado (apenas para escola indígena) - Código da Língua Indígena 1	Num	4		De 2007 a 2018: Apenas uma Língua Indígena
298	CO_LINGUA_INDIGENA_2	Escola Indígena - Língua em que o ensino é ministrado (apenas para escola indígena) - Código da Língua Indígena 2	Num	4		
299	CO_LINGUA_INDIGENA_3	Escola Indígena - Língua em que o ensino é ministrado (apenas para escola indígena) - Código da Língua Indígena 3	Num	4		
300	IN_BRASIL_ALFABETIZADO	Escola cede espaço para turmas do Programa Brasil Alfabetizado	Num	1	0 - Não 1 - Sim	
301	IN_FINAL_SEMANA	Escola abre aos finais de semana para a comunidade	Num	1	0 - Não 1 - Sim	
302	IN_EXAME_SELECAO	A escola faz exame de seleção para ingresso de seus alunos (avaliação por prova e/ou análise curricular)	Num	1	0 - Não 1 - Sim 9 - Não informado	
303	IN_RESERVA_PPI	Reserva de vagas por sistema de cotas para grupos específicos de alunos - Autodeclarado preto, pardo ou indígena (PPI)	Num	1	0 - Não 1 - Sim - Não aplicável para escolas que não fazem exame de seleção	

304	IN_RESERVA_RENDA	Reserva de vagas por sistema de cotas para grupos específicos de alunos - Condição de Renda	Num	1	0 - Não 1 - Sim - Não aplicável para escolas que não fazem exame de seleção	
305	IN_RESERVA_PUBLICA	Reserva de vagas por sistema de cotas para grupos específicos de alunos - Outro tipo de escola pública	Num	1	0 - Não 1 - Sim - Não aplicável para escolas que não fazem exame de seleção	
306	IN_RESERVA_PCD	Reserva de vagas por sistema de cotas para grupos específicos de alunos - Pessoa com deficiência (PCD)	Num	1	0 - Não 1 - Sim - Não aplicável para escolas que não fazem exame de seleção	
307	IN_RESERVA_OUTROS	Reserva de vagas por sistema de cotas para grupos específicos de alunos - Outros grupos	Num	1	0 - Não 1 - Sim - Não aplicável para escolas que não fazem exame de seleção	
308	IN_RESERVA_NENHUMA	Reserva de vagas por sistema de cotas para grupos específicos de alunos - Sem reservas de vagas para sistema de cotas (ampla concorrência)	Num	1	0 - Não 1 - Sim - Não aplicável para escolas que não fazem exame de seleção	
309	IN_REDES_SOCIAIS	A escola possui site ou blog ou página em redes sociais para comunicação institucional	Num	1	0 - Não 1 - Sim 9 - Não informado	
310	IN_ESPACO_ATIVIDADE	A escola compartilha espaços para atividades de integração escola-comunidade	Num	1	0 - Não 1 - Sim 9 - Não informado	
311	IN_ESPACO_EQUIPAMENTO	A escola usa espaços e equipamentos do entorno escolar para atividades regulares com os alunos	Num	1	0 - Não 1 - Sim 9 - Não informado	
312	IN_ORGAO_ASS_PAIS	Órgãos colegiados em funcionamento na escola - Associação de Pais	Num	1	0 - Não 1 - Sim	Em 2019: Existe adicionalmente a categoria "9 - Não informado".
313	IN_ORGAO_ASS_PAIS_MESTRES	Órgãos colegiados em funcionamento na escola - Associação de Pais e Mestres	Num	1	0 - Não 1 - Sim	Em 2019: Existe adicionalmente a categoria "9 - Não informado".
314	IN_ORGAO_CONSELHO_ESCOLAR	Órgãos colegiados em funcionamento na escola - Conselho Escolar	Num	1	0 - Não 1 - Sim	Em 2019: Existe adicionalmente a categoria "9 - Não informado".
315	IN_ORGAO_GREMIO_ESTUDANTIL	Órgãos colegiados em funcionamento na escola - Grêmio Estudantil	Num	1	0 - Não 1 - Sim	Em 2019: Existe adicionalmente a categoria "9 - Não informado".
316	IN_ORGAO_OUTROS	Órgãos colegiados em funcionamento na escola - Outros	Num	1	0 - Não 1 - Sim	Em 2019: Existe adicionalmente a categoria "9 - Não informado".
317	IN_ORGAO_NENHUM	Órgãos colegiados em funcionamento na escola - Não há órgãos colegiados em funcionamento	Num	1	0 - Não 1 - Sim	Em 2019: Existe adicionalmente a categoria "9 - Não informado".
318	TP_PROPOSTA_PEDAGOGICA	O projeto político pedagógico ou a proposta pedagógica da escola (conforme art. 12 da LDB) foi atualizado nos últimos 12 meses até a data de referência	Num	1	0 - Não 1 - Sim 2 - A escola não possui projeto político pedagógico/proposta pedagógica 9 - Não informado	
319	IN_EDUC_AMBIENTAL	A escola desenvolve ações na área de educação ambiental?	Num	1	0 - Não 1 - Sim	
320	IN_EDUC_AMB_CONTEUDO	Informe de qual(quais) forma(s) a educação ambiental é desenvolvida na escola: Como conteúdo dos componentes/campões de experiências presentes no currículo	Num	1	0 - Não 1 - Sim	
321	IN_EDUC_AMB_CURRICULAR	Informe de qual(quais) forma(s) a educação ambiental é desenvolvida na escola: Como um componente curricular especial, específico, flexível ou eletrônico	Num	1	0 - Não 1 - Sim	
322	IN_EDUC_AMB_EIXO	Informe de qual(quais) forma(s) a educação ambiental é desenvolvida na escola: Como um eixo estruturante do currículo	Num	1	0 - Não 1 - Sim	
323	IN_EDUC_AMB_EVENTOS	Informe de qual(quais) forma(s) a educação ambiental é desenvolvida na escola: Em eventos	Num	1	0 - Não 1 - Sim	
324	IN_EDUC_AMB_PROJETOS	Informe de qual(quais) forma(s) a educação ambiental é desenvolvida na escola: Em projetos transversais ou interdisciplinares	Num	1	0 - Não 1 - Sim	
325	IN_EDUC_AMB_NENHUMA	Informe de qual(quais) forma(s) a educação ambiental é desenvolvida na escola: Nenhuma das opções listadas	Num	1	0 - Não 1 - Sim	
326	TP_AEE	Atendimento Educacional Especializado (AEE)	Num	1	0 - Não oferece 1 - Não exclusivamente 2 - Exclusivamente	
327	TP_ATIVIDADE_COMPLEMENTAR	Atividade Complementar	Num	1	0 - Não oferece 1 - Não exclusivamente 2 - Exclusivamente	
328	IN_MEDIACAO_PRESENCIAL	Mediação didático-pedagógica oferecida pela escola - Presencial	Num	1	0 - Não 1 - Sim	
329	IN_MEDIACAO_SEMIPRESENCIAL	Mediação didático-pedagógica oferecida pela escola - Semipresencial	Num	1	0 - Não 1 - Sim	
330	IN_MEDIACAO_EAD	Mediação didático-pedagógica oferecida pela escola - Educação a Distância - EAD	Num	1	0 - Não 1 - Sim	
331	IN_REGULAR	Modo, maneira ou metodologia de ensino correspondente às turmas com etapas de escolarização consecutivas, Creche ao Ensino Médio	Num	1	0 - Não 1 - Sim	
332	IN_DIURNO	Turno - Diurno - Maior parte das atividades da turma são realizadas no período entre 6h e 17h59	Num	1	0 - Não 1 - Sim	Anteriormente, eram consideradas diurnas, as turmas com horário de início entre 05h e 16h59. Atualmente, nas estatísticas oficiais, passou-se a considerar o período de maior tempo de duração da turma, independentemente do seu horário de início.
333	IN_NOTURNO	Turno - Noturno - Maior parte das atividades da turma são realizadas entre 18h e 5:59h	Num	1	0 - Não 1 - Sim	Anteriormente, eram consideradas noturnas, as turmas com horário de início entre 17h e 04h59. Atualmente, nas estatísticas oficiais, passou-se a considerar o período de maior tempo de duração da turma, independentemente do seu horário de início.
334	IN_EAD	Turno não aplicável para turmas semipresenciais ou de Educação a Distância (EAD)	Num	1	0 - Não 1 - Sim	
335	IN_BAS	Educação Básica (Possui uma ou mais matrículas)	Num	1	0 - Não 1 - Sim	A variável "IN_BAS" foi descontinuada em virtude da inclusão da variável "IN_ESCOLARIZACAO" no Censo Escolar de 2022. Ambas, permitem filtrar as escolas em atividade e que possuem uma ou mais matrículas em turmas de escolarização, obtendo assim, dados estatísticos em consonância com a divulgação oficial.
336	IN_ESCOLARIZACAO	Escola possui uma ou mais matrículas de escolarização em alguma das seguintes etapas de ensino: Creche, Pré-Escola, Ensino Fundamental, Ensino Médio, Educação de Jovens e Adultos (EJA), Curso Técnico Concomitante, Curso Técnico Subsequente, Curso FIC Concomitante	Num	1	0 - Não 1 - Sim	
337	IN_INF	Etapas de Ensino - Educação Infantil (Possui uma ou mais matrículas)	Num	1	0 - Não 1 - Sim	
338	IN_INF_CRE	Etapas de Ensino - Educação Infantil - Creche (Possui uma ou mais matrículas)	Num	1	0 - Não 1 - Sim	
339	IN_INF_PRE	Etapas de Ensino - Educação Infantil - Pré-Escola (Possui uma ou mais matrículas)	Num	1	0 - Não 1 - Sim	
340	IN_FUND	Etapas de Ensino - Ensino Fundamental (Possui uma ou mais matrículas)	Num	1	0 - Não 1 - Sim	
341	IN_FUND_AI	Etapas de Ensino - Ensino Fundamental - Anos Iniciais (Possui uma ou mais matrículas)	Num	1	0 - Não 1 - Sim	
342	IN_FUND_AF	Etapas de Ensino - Ensino Fundamental - Anos Finais (Possui uma ou mais matrículas)	Num	1	0 - Não 1 - Sim	
343	IN_MED	Etapas de Ensino - Ensino Médio (Possui uma ou mais matrículas)	Num	1	0 - Não 1 - Sim	
344	IN_PROF	Educação Profissional - Modo profissionalizante de ensino correspondente às turmas de cursos de formação inicial e continuada ou de qualificação profissional (Cursos FIC) articulados à EJA ou concomitantes; ou de cursos técnicos de nível médio nas formas articulada (integrada ou concomitante) ou subsequente ao ensino médio e de normal/magistério (Possui uma ou mais matrículas)	Num	1	0 - Não 1 - Sim	
345	IN_PROF_TEC	Educação Profissional Técnica - Modo profissionalizante de ensino correspondente às turmas de cursos técnicos de nível médio nas formas articuladas (integrada ou concomitante), ou subsequente ao ensino médio e de normal/magistério (Possui uma ou mais matrículas)	Num	1	0 - Não 1 - Sim	
346	IN_EJA	Educação de Jovens e Adultos (EJA) - Modo, maneira ou metodologia de ensino correspondente às turmas destinadas a pessoas que não cursaram o ensino fundamental e/ou médio em idade própria (Possui uma ou mais matrículas)	Num	1	0 - Não 1 - Sim	
347	IN_EJA_FUND	Educação de Jovens e Adultos (EJA) - Ensino Fundamental (Possui uma ou mais matrículas)	Num	1	0 - Não 1 - Sim	
348	IN_EJA_MED	Educação de Jovens e Adultos (EJA) - Ensino Médio (Possui uma ou mais matrículas)	Num	1	0 - Não 1 - Sim	
349	IN_ESP	Educação Especial - Inclui a Educação Especial Inclusiva (em Classes Comuns) e a Educação Especial Exclusiva (Possui uma ou mais matrículas)	Num	1	0 - Não 1 - Sim	
350	IN_ESP_CC	Educação Especial Inclusiva (em Classes Comuns) - Escola possui um ou mais alunos com deficiência, transtorno global do desenvolvimento ou altas habilidades/superdotação estudando em classes comuns do Ensino Regular e/ou Educação de Jovens e Adultos (Possui uma ou mais matrículas)	Num	1	0 - Não 1 - Sim	
351	IN_ESP_CE	Educação Especial Exclusiva - Escola exclusivamente especializada e/ou que possui classe especial exclusiva para o atendimento de alunos com deficiência, transtorno global do desenvolvimento ou altas habilidades/superdotação (Possui uma ou mais matrículas)	Num	1	0 - Não 1 - Sim	
352	QT_MAT_BAS	Número de Matrículas da Educação Básica	Num	8		Não inclui matrículas em turmas exclusivas de atividade complementar e nem em turmas exclusivos de Atendimento Educacional Especializado. Em síntese, inclui matrículas em turmas de escolarização nas seguintes etapas e modalidades de ensino: Educação Infantil, Ensino Fundamental, Ensino Médio, Educação Profissional Técnica - Curso Técnico Concomitante, na Educação Profissional Técnica - Curso Técnico Subsequente, na Educação Profissional - Curso FIC Concomitante, e na Educação de Jovens e Adultos (EJA).
353	QT_MAT_INF	Número de Matrículas da Educação Infantil	Num	8		
354	QT_MAT_INF_CRE	Número de Matrículas da Educação Infantil - Creche	Num	8		
355	QT_MAT_INF_PRE	Número de Matrículas da Educação Infantil - Pré-Escola	Num	8		

<u>435</u>	<u>QT_DOC_BAS</u>	Número de Docentes da Educação Básica	Num	8	Notas importantes:
<u>436</u>	<u>QT_DOC_INF</u>	Número de Docentes da Educação Infantil	Num	8	1) Os docentes referem-se aos indivíduos que estavam em efetiva regência de classe na data de referência do Censo Escolar da Educação Básica (última quarta-feira do mês de maio).
<u>437</u>	<u>QT_DOC_INF_CRE</u>	Número de Docentes da Educação Infantil - Creche	Num	8	2) Os docentes são contados uma única vez em cada estabelecimento de ensino, contudo, um mesmo docente pode atuar em mais de uma Região Geográfica, Unidade da Federação, Município, estabelecimento de ensino, etapa de ensino, etc. Portanto, a soma dos docentes constantes nos Microdados não representam o total Brasil ou das Unidades da Federação. Para consultar o número de docentes em cada unidade de agregação, de forma distinta, acesse as Sinopses Estatísticas, ou, Inepdata, cujo links estão indicados no cabeçalho deste documento.
<u>438</u>	<u>QT_DOC_INF_PRE</u>	Número de Docentes da Educação Infantil - Pré-Escola	Num	8	Os docentes são contados somente uma vez em cada Etapa de Ensino, independentemente de atuarem em mais de uma delas.■
<u>439</u>	<u>QT_DOC_FUND</u>	Número de Docentes do Ensino Fundamental	Num	8	3) Na Educação Infantil inclui, os docentes que atuam em turmas de Creche, Pré-Escola e unificadas.■
<u>440</u>	<u>QT_DOC_FUND_AI</u>	Número de Docentes do Ensino Fundamental - Anos Iniciais	Num	8	4) Na Pré-Escola da Educação Infantil, inclui os docentes que atuam em turmas unificadas.■
<u>441</u>	<u>QT_DOC_FUND_AF</u>	Número de Docentes do Ensino Fundamental - Anos Finais	Num	8	5) Nos Anos Iniciais do Ensino Fundamental, inclui os docentes que atuam em turmas do 1º ao 5º ano do Ensino Fundamental e em turmas da Educação Infantil e Ensino Fundamental Multietapa.■
<u>442</u>	<u>QT_DOC_MED</u>	Número de Docentes do Ensino Médio	Num	8	6) Nos Anos Finais do Ensino Fundamental, inclui os docentes que atuam em turmas do 6º ao 9º ano do Ensino Fundamental e em turmas do Ensino Fundamental Correção de Fluxo e turmas Multi.■
<u>443</u>	<u>QT_DOC_PROF</u>	Número de Docentes da Educação Profissional	Num	8	7) Na Educação Profissional, inclui os docentes que atuam nas seguintes Etapas de Ensino: Curso Técnico Integrado à Educação Profissional, Ensino Médio Normal/Magistério, Curso Técnico Concomitante, Curso Técnico Subsequente,
<u>444</u>	<u>QT_DOC_PROF_TEC</u>	Número de Docentes da Educação Profissional Técnica	Num	8	
<u>445</u>	<u>QT_DOC_EIA</u>	Número de Docentes da Educação de Jovens e Adultos (EJA)	Num	8	
<u>446</u>	<u>QT_DOC_EJA_FUND</u>	Número de Docentes da Educação de Jovens e Adultos (EJA) - Ensino Fundamental	Num	8	
<u>447</u>	<u>QT_DOC_EJA_MED</u>	Número de Docentes da Educação de Jovens e Adultos (EJA) - Ensino Médio	Num	8	
<u>448</u>	<u>QT_DOC_ESP</u>	Número de Docentes da Educação Especial	Num	8	
<u>449</u>	<u>QT_DOC_ESP_CC</u>	Número de Docentes da Educação Especial Inclusiva	Num	8	
<u>450</u>	<u>QT_DOC_ESP_CE</u>	Número de Docentes da Educação Especial Exclusiva	Num	8	
<u>451</u>	<u>QT_TUR_BAS</u>	Número de Turmas de Educação Básica	Num	8	
<u>452</u>	<u>QT_TUR_INF</u>	Número de Turmas de Educação Infantil	Num	8	
<u>453</u>	<u>QT_TUR_INF_CRE</u>	Número de Turmas de Educação Infantil - Creche	Num	8	
<u>454</u>	<u>QT_TUR_INF_PRE</u>	Número de Turmas de Educação Infantil - Pré-Escola	Num	8	
<u>455</u>	<u>QT_TUR_FUND</u>	Número de Turmas de Ensino Fundamental	Num	8	
<u>456</u>	<u>QT_TUR_FUND_AI</u>	Número de Turmas de Ensino Fundamental - Anos Iniciais	Num	8	
<u>457</u>	<u>QT_TUR_FUND_AF</u>	Número de Turmas de Ensino Fundamental - Anos Finais	Num	8	
<u>458</u>	<u>QT_TUR_MED</u>	Número de Turmas de Ensino Médio	Num	8	
<u>459</u>	<u>QT_TUR_PROF</u>	Número de Turmas de Educação Profissional	Num	8	
<u>460</u>	<u>QT_TUR_PROF_TEC</u>	Número de Turmas de Educação Profissional Técnica	Num	8	
<u>461</u>	<u>QT_TUR_EIA</u>	Número de Turmas de Educação de Jovens e Adultos (EJA)	Num	8	
<u>462</u>	<u>QT_TUR_EJA_FUND</u>	Número de Turmas de Educação de Jovens e Adultos (EJA) - Ensino Fundamental	Num	8	
<u>463</u>	<u>QT_TUR_EJA_MED</u>	Número de Turmas de Educação de Jovens e Adultos (EJA) - Ensino Médio	Num	8	
<u>464</u>	<u>QT_TUR_ESP</u>	Número de Turmas de Educação Especial	Num	8	
<u>465</u>	<u>QT_TUR_ESP_CC</u>	Número de Turmas de Educação Especial Inclusiva	Num	8	
<u>466</u>	<u>QT_TUR_ESP_CE</u>	Número de Turmas de Educação Especial Exclusiva	Num	8	
<u>467</u>	<u>QT_TUR_BAS_D</u>	Número de Turmas da Educação Básica - Turno Diurno	Num	8	Diurno - Quando a maior parte das atividades da turma são realizadas no período entre 6h e 17:59h.
<u>468</u>	<u>QT_TUR_BAS_N</u>	Número de Turmas da Educação Básica - Turno Noturno	Num	8	Noturno - quando a maior parte das atividades da turma são realizadas entre 18h e 5:59h.
<u>469</u>	<u>QT_TUR_BAS_EAD</u>	Número de Turmas da Educação Básica - Turno não aplicável para turmas semipresenciais ou de Educação à Distância (EAD)	Num	8	
<u>470</u>	<u>QT_TUR_INF_INT</u>	Número de Turmas da Educação Infantil - Tempo Integral	Num	8	
<u>471</u>	<u>QT_TUR_INF_CRE_INT</u>	Número de Turmas da Educação Infantil - Creche - Tempo Integral	Num	8	
<u>472</u>	<u>QT_TUR_INF_PRE_INT</u>	Número de Turmas da Educação Infantil - Pré-Escola - Tempo Integral	Num	8	
<u>473</u>	<u>QT_TUR_FUND_INT</u>	Número de Turmas do Ensino Fundamental - Tempo Integral	Num	8	
<u>474</u>	<u>QT_TUR_FUND_AI_INT</u>	Número de Turmas do Ensino Fundamental - Anos Iniciais - Tempo Integral	Num	8	
<u>475</u>	<u>QT_TUR_FUND_AF_INT</u>	Número de Turmas do Ensino Fundamental - Anos Finais - Tempo Integral	Num	8	
<u>476</u>	<u>QT_TUR_MED_INT</u>	Número de Turmas do Ensino Médio - Tempo Integral	Num	8	

Considera-se em tempo integral, as turmas com duração igual ou maior que 7 horas diárias (35 horas semanais).

APÊNDICE C – CONFIGURAÇÃO DO AMBIENTE VIRTUAL

```

name: kernel_4
channels:
- conda-forge
- rapidsai
- defaults
dependencies:
- _libgcc_mutex=0.1=conda_forge
- _openmp_mutex=4.5=2_gnu
- alsa-lib=1.2.15.3=hb03c661_0
- asttokens=3.0.1=pyhd8ed1ab_0
- attr=2.5.2=h39aace5_0
- aws-c-auth=0.9.3=hef928c7_0
- aws-c-cal=0.9.13=h2c9d079_1
- aws-c-common=0.12.6=hb03c661_0
- aws-c-compression=0.3.1=h8b1a151_9
- aws-c-event-stream=0.5.7=h28f887f_1
- aws-c-http=0.10.7=ha8fc4e3_5
- aws-c-io=0.23.3=hdaf4b65_5
- aws-c-mqtt=0.13.3=hc63082f_11
- aws-c-s3=0.11.3=h06ab39a_1
- aws-c-sdkutils=0.2.4=h8b1a151_4
- aws-checksums=0.2.7=h8b1a151_5
- aws-crt-cpp=0.35.4=h8824e59_0
- aws-sdk-cpp=1.11.606=h20b40b1_10
- azure-core-cpp=1.16.2=h206d751_0
- azure-identity-cpp=1.13.3=hed0cdb0_1
- azure-storage-blobs-cpp=12.16.0=hdd73cc9_1
- azure-storage-common-cpp=12.12.0=ha7a2c86_1
- azure-storage-files-datalake-cpp=12.14.0=h52c5a47_1
- brotli=1.2.0=hed03a55_1
- brotli-bin=1.2.0=hb03c661_1
- bzip2=1.0.8=hda65f42_8
- c-ares=1.34.6=hb03c661_0
- ca-certificates=2026.1.4=hbdb8a1cb_0
- cachetools=7.0.1=pyhd8ed1ab_0
- cairo=1.18.4=he90730b_1
- comm=0.2.3=pyhe01879c_0
- contourpy=1.3.3=py311h724c32c_4
- cuda-bindings=12.9.5=py311hff0572f_0
- cuda-cccl_linux-64=12.5.39=ha770c72_0
- cuda-core=0.5.1=cuda12_py311hc26999e_1
- cuda-crt-dev_linux-64=12.5.82=ha770c72_0
- cuda-crt-tools=12.5.82=ha770c72_0
- cuda-cudart=12.5.82=he02047a_0
- cuda-cudart-dev=12.5.82=he02047a_0
- cuda-cudart-dev_linux-64=12.5.82=h85509e4_0
- cuda-cudart-static=12.5.82=he02047a_0
- cuda-cudart-static_linux-64=12.5.82=h85509e4_0
- cuda-cudart_linux-64=12.5.82=h85509e4_0
- cuda-nvcc-dev_linux-64=12.5.82=ha770c72_0
- cuda-nvcc-impl=12.5.82=hd3aeb46_0
- cuda-nvcc-tools=12.5.82=hd3aeb46_0
- cuda-nvrtc=12.5.82=he02047a_0
- cuda-nvvm-dev_linux-64=12.5.82=ha770c72_0
- cuda-nvvm-impl=12.5.82=h59595ed_0

```

```
- cuda-nvvm-tools=12.5.82=h59595ed_0
- cuda-pathfinder=1.3.4=pyhcf101f3_0
- cuda-profiler-api=12.5.39=ha770c72_0
- cuda-python=12.9.5=pyh698daf1_0
- cuda-version=12.5=hd4f0392_3
- cudf=25.12.00=cuda12_py311_251210_580975be
- cuml=25.12.00=cuda12_py311_251211_5c22c200
- cupy=13.6.0=py311h72da3fd_2
- cupy-core=13.6.0=py311he30c881_2
- cycler=0.12.1=pyhcf101f3_2
- cyrus-sasl=2.1.28=hd9c7081_0
- dbus=1.16.2=h24cb091_1
- debugpy=1.8.20=py311hc665b79_0
- decorator=5.2.1=pyhd8ed1ab_0
- dlpack=0.8=h59595ed_3
- double-conversion=3.4.0=hecca717_0
- executing=2.2.1=pyhd8ed1ab_0
- fastrlock=0.8.3=py311hc665b79_2
- font-ttf-dejavu-sans-mono=2.37=hab24e00_0
- font-ttf-inconsolata=3.000=h77eed37_0
- font-ttf-source-code-pro=2.038=h77eed37_0
- font-ttf-ubuntu=0.83=h77eed37_3
- fontconfig=2.15.0=h7e30c49_1
- fonts-conda-ecosystem=1=0
- fonts-conda-forge=1=hc364b38_1
- fonttools=4.61.1=py311h3778330_0
- freetype=2.14.1=ha770c72_0
- fsspec=2026.2.0=pyhd8ed1ab_0
- gflags=2.2.2=h5888daf_1005
- glog=0.7.1=hbabe93e_0
- graphite2=1.3.14=hecca717_2
- harfbuzz=12.3.2=h6083320_0
- icu=78.2=h33c6efd_0
- ipykernel=7.2.0=pyha191276_1
- ipython=9.10.0=pyh53cf698_0
- ipython_pygments_lexers=1.1.1=pyhd8ed1ab_0
- jedi=0.19.2=pyhd8ed1ab_1
- joblib=1.5.3=pyhd8ed1ab_0
- jupyter_client=8.8.0=pyhcf101f3_0
- jupyter_core=5.9.1=pyhc90fa1f_0
- keyutils=1.6.3=hb9d3cd8_0
- kiwisolver=1.4.9=py311h724c32c_2
- krb5=1.21.3=h659f571_0
- lcms2=2.18=h0c24ade_0
- ld_impl_linux-64=2.45.1=default_hbd61a6d_101
- lerc=4.0.0=h0aef613_1
- libabseil=20260107.1=cxx17_h7b12aa8_0
- libarrow=21.0.0=h2603568_18_cpu
- libarrow-acero=21.0.0=h635bf11_18_cpu
- libarrow-compute=21.0.0=h53684a4_18_cpu
- libarrow-dataset=21.0.0=h635bf11_18_cpu
- libarrow-substrait=21.0.0=hb4dd7c2_18_cpu
- libblas=3.11.0=5_h4a7cf45_openblas
- libbrotlicommon=1.2.0=hb03c661_1
- libbrotlidec=1.2.0=hb03c661_1
- libbrotlienc=1.2.0=hb03c661_1
- libcap=2.77=h3ff7636_0
- libcblas=3.11.0=5_h0358290_openblas
- libclang-cpp21.1=21.1.8=default_h99862b1_3
- libclang13=21.1.8=default_h746c552_3
```

```

- libcrc32c=1.1.2=h9c3ff4c_0
- libcublas=12.5.3.2=he02047a_0
- libcublas-dev=12.5.3.2=he02047a_0
- libcurlf=25.12.00=cuda12_251210_580975be
- libcufft=11.2.3.61=he02047a_0
- libcufile=1.10.1.7=he02047a_0
- libcufile-dev=1.10.1.7=he02047a_0
- libcuml=25.12.00=cuda12_251211_5c22c200
- libcumlprims=25.12.00=cuda12_py310_251211_7e1ef293
- libcurls=2.3.3=hb8b1518_5
- libcurl=10.3.6.82=he02047a_0
- libcurl-dev=10.3.6.82=he02047a_0
- libcurl=8.18.0=h4e3cde8_0
- libcusolver=11.6.3.83=he02047a_0
- libcusolver-dev=11.6.3.83=he02047a_0
- libcusparse=12.5.1.3=he02047a_0
- libcusparse-dev=12.5.1.3=he02047a_0
- libcuvs=25.12.00=cuda12_251211_fc27938b
- libcuvs-headers=25.12.00=cuda12_251211_fc27938b
- libdeflate=1.25=h17f619e_0
- libdrm=2.4.125=hb03c661_1
- libedit=3.1.20250104=pl5321h7949ede_0
- libegl=1.7.0=ha4b6fd6_2
- libev=4.33=hd590300_2
- libevent=2.1.12=hf998b51_1
- libexpat=2.7.3=hecca717_0
- libffi=3.5.2=h3435931_0
- libfreetype=2.14.1=ha770c72_0
- libfreetype6=2.14.1=h73754d4_0
- libgcc=15.2.0=he0feb66_17
- libgcc-ng=15.2.0=hb9a702a_17
- libgfortran=15.2.0=hb9a702a_17
- libgfortran5=15.2.0=h68bc16d_17
- libgl=1.7.0=ha4b6fd6_2
- libglib=2.86.3=h6548e54_1
- libglvnd=1.7.0=ha4b6fd6_2
- libglx=1.7.0=ha4b6fd6_2
- libgomp=15.2.0=he0feb66_17
- libgoogle-cloud=2.39.0=h9d11ab5_1
- libgoogle-cloud-storage=2.39.0=hdbdcf42_1
- libgrpc=1.78.0=h1d1128b_1
- libiconv=1.18=h3b78370_2
- libjpeg-turbo=3.1.2=hb03c661_0
- libkvikio=25.12.00=cuda12_251210_61297197
- liblapack=3.11.0=5_h47877c9_openblas
- libllvm21=21.1.8=hf7376ad_0
- liblzma=5.8.2=hb03c661_0
- libnghttp2=1.67.0=had1ee68_0
- libnl=3.11.0=hb9d3cd8_0
- libnsl=2.0.1=hb9d3cd8_1
- libntlm=1.8=hb9d3cd8_0
- libnuma=2.0.18=hb9d3cd8_3
- libnvcomp=5.0.0.6=hb7e823c_3
- libnvcomp-dev=5.0.0.6=hb7e823c_3
- libnvjitlink=12.9.86=hecca717_2
- libopenblas=0.3.30=pthreads_h94d23a6_4
- libopengl=1.7.0=ha4b6fd6_2
- libopentelemetry-cpp=1.21.0=h9692893_2
- libopentelemetry-cpp-headers=1.21.0=ha770c72_2
- libparquet=21.0.0=h7376487_18_cpu

```

```
- libpciaccess=0.18=hb9d3cd8_0
- libpng=1.6.55=h421ea60_0
- libpq=18.2=hb80d175_0
- libprotobuf=6.33.5=h2b00c02_0
- libraft=25.12.00=cuda12_251211_d226bc96
- libraft-headers=25.12.00=cuda12_251211_d226bc96
- libraft-headers-only=25.12.00=cuda12_251211_d226bc96
- libre2-11=2025.11.05=h0dc7533_1
- librmm=25.12.00=cuda12_251210_86731e05
- libsodium=1.0.20=h4ab18f5_0
- libsqlite=3.51.2=hf4e2dac_0
- libssh2=1.11.1=hcf80075_0
- libstdcxx=15.2.0=h934c35e_17
- libstdcxx-ng=15.2.0=hdf11a46_17
- libsystemd0=259.1=h6569c3e_0
- libthrift=0.22.0=h454ac66_1
- libtiff=4.7.1=h9d88235_1
- libucxx=0.47.00=cuda12_251210_5d6f0af3
- libudev1=259.1=h6569c3e_0
- libutf8proc=2.11.3=hfe17d71_0
- libuuid=2.41.3=h5347b49_0
- libvulkan-loader=1.4.341.0=h5279c79_0
- libwebp-base=1.6.0=hd42ef1d_0
- libxcb=1.17.0=h8a09558_0
- libxcrypt=4.4.36=hd590300_1
- libxgboost=3.1.2=rapidsai_hc3bde56_1
- libxkbcommon=1.13.1=hca5e8e5_0
- libxml2=2.15.1=he237659_1
- libxml2-16=2.15.1=hca6bf5a_1
- libxslt=1.1.43=h711ed8c_1
- libzlib=1.3.1=hb9d3cd8_2
- llvmlite=0.44.0=py311h1741904_2
- lz4-c=1.10.0=h5888daf_1
- markdown-it-py=4.0.0=pyhd8ed1ab_0
- matplotlib=3.10.8=py311h38be061_0
- matplotlib-base=3.10.8=py311h0f3be63_0
- matplotlib-inline=0.2.1=pyhd8ed1ab_0
- mdurl=0.1.2=pyhd8ed1ab_1
- munkres=1.1.4=pyhd8ed1ab_1
- nccl=2.29.3.1=h4d09622_0
- ncurses=6.5=h2d0b736_3
- nest-asyncio=1.6.0=pyhd8ed1ab_1
- nlohmann_json=3.12.0=h54a6638_1
- numba=0.61.2=py311h6220fa4_2
- numba-cuda=0.19.2=pyhcf101f3_0
- numpy=2.2.6=py311h5d046bc_0
- nvidia-ml-py=13.590.48=pyhd8ed1ab_0
- nvtx=0.2.14=py311h49ec1c0_1
- openjpeg=2.5.4=h55fea9a_0
- opendldap=2.6.10=he970967_0
- openssl=3.6.1=h35e630c_1
- orc=2.2.2=hb90d81_1
- packaging=26.0=pyhcf101f3_0
- pandas=2.3.3=py311hed34c8f_2
- parso=0.8.6=pyhcf101f3_0
- patsy=1.0.2=pyhcf101f3_0
- pcre2=10.47=haa7fec5_0
- pexpect=4.9.0=pyhd8ed1ab_1
- pillow=12.1.1=py311hf88fc01_0
- pip=26.0.1=pyh8b19718_0
```

```
- pixman=0.46.4=h54a6638_1
- platformdirs=4.5.1=pyhcf101f3_0
- prometheus-cpp=1.3.0=ha5d0236_0
- prompt-toolkit=3.0.52=pyha770c72_0
- psutil=7.2.2=py311haee01d2_0
- pthread-stubs=0.4=hb9d3cd8_1002
- ptyprocess=0.7.0=pyhd8ed1ab_1
- pure_eval=0.2.3=pyhd8ed1ab_1
- py-xgboost=3.1.2=rapidsai_pyh395bae7_1
- pyarrow=21.0.0=py311h38be061_3
- pyarrow-core=21.0.0=py311h342b5a4_3_cpu
- pygments=2.19.2=pyhd8ed1ab_0
- pylibcudf=25.12.00=cuda12_py311_251210_580975be
- pylibraft=25.12.00=cuda12_py311_251211_d226bc96
- pyparsing=3.3.2=pyhcf101f3_0
- pyside6=6.10.2=py311he4c1a5a_0
- python=3.11.14=hd63d673_3_cpython
- python-dateutil=2.9.0.post0=pyhe01879c_2
- python-tzdata=2025.3=pyhd8ed1ab_0
- python_abi=3.11=8_cp311
- pytz=2025.2=pyhd8ed1ab_0
- pyzmq=27.1.0=py311h2315fb0_0
- qhull=2020.2=h434a139_5
- qt6-main=6.10.2=hb82b983_4
- rapids-logger=0.2.3=h98325ef_0
- rdma-core=61.0=h192683f_0
- re2=2025.11.05=h5301d42_1
- readline=8.3=h853b02a_0
- rich=14.3.2=pyhcf101f3_0
- rmm=25.12.00=cuda12_py311_251210_86731e05
- s2n=1.6.2=he8a4886_1
- scikit-learn=1.8.0=np2py311ha15b03d_1
- scipy=1.16.3=py311hbe70eeb_2
- seaborn=0.13.2=hd8ed1ab_3
- seaborn-base=0.13.2=pyhd8ed1ab_3
- setuptools=82.0.0=pyh332efcf_0
- six=1.17.0=pyhe01879c_1
- snappy=1.2.2=h03e3b7b_1
- stack_data=0.6.3=pyhd8ed1ab_1
- statsmodels=0.14.6=py311h0372a8f_0
- threadpoolctl=3.6.0=pyhecae5ae_0
- tk=8.6.13=noxft_h366c992_103
- tornado=6.5.4=py311h41d9c34_0
- traitlets=5.14.3=pyhd8ed1ab_1
- treeelite=4.6.1=py311h72b0140_0
- typing_extensions=4.15.0=pyhcf101f3_0
- tzdata=2025c=hc9c84f9_1
- ucx=1.19.1=h63b5c0b_0
- ucxx=0.47.00=cuda12_py311_251210_5d6f0af3
- unicodedata2=17.0.1=py311h49ec1c0_0
- wayland=1.24.0=hd6090a7_1
- wcwidth=0.6.0=pyhd8ed1ab_0
- wheel=0.46.3=pyhd8ed1ab_0
- xcb-util=0.4.1=h4f16b4b_2
- xcb-util-cursor=0.1.6=hb03c661_0
- xcb-util-image=0.4.0=hb711507_2
- xcb-util-keysyms=0.4.1=hb711507_0
- xcb-util-renderutil=0.3.10=hb711507_0
- xcb-util-wm=0.4.2=hb711507_0
- xkeyboard-config=2.46=hb03c661_0
```

```
- xorg-libice=1.1.2=hb9d3cd8_0
- xorg-libsm=1.2.6=he73a12e_0
- xorg-libx11=1.8.13=he1eb515_0
- xorg-libxau=1.0.12=hb03c661_1
- xorg-libxcomposite=0.4.7=hb03c661_0
- xorg-libxcursor=1.2.3=hb9d3cd8_0
- xorg-libxdamage=1.1.6=hb9d3cd8_0
- xorg-libxdmcp=1.1.5=hb03c661_1
- xorg-libxext=1.3.7=hb03c661_0
- xorg-libxfixed=6.0.2=hb03c661_0
- xorg-libxi=1.8.2=hb9d3cd8_0
- xorg-libxrandr=1.5.5=hb03c661_0
- xorg-libxrender=0.9.12=hb9d3cd8_0
- xorg-libxtst=1.2.5=hb9d3cd8_3
- xorg-libxxf86vm=1.1.7=hb03c661_0
- zeromq=4.3.5=h387f397_9
- zlib=1.3.1=hb9d3cd8_2
- zlib-ng=2.3.3=hceb46e0_1
- zstd=1.5.7=hb78ec9c_6
- pip:
  - cloudpickle==3.1.2
  - duckdb==1.4.4
  - lightgbm==4.6.0
  - mpmath==1.3.0
  - shap==0.50.0
  - slicer==0.0.8
  - sympy==1.14.0
  - tqdm==4.67.3
```

prefix: /home/ramon/miniconda3/envs/kernel_4

APÊNDICE D – LINK PARA O GITHUB

O código-fonte do projeto está disponível no seguinte repositório do Github: https://github.com/sudruder/TCC_MBA_Ciencia_Dados_ICMC