Instituto Federal do Norte de Minas Gerais
campus Montes Claros
Curso de Bacharelado em Ciência da Computação

PREVISÃO DO DESEMPENHO ESCOLAR DO 9º ANO EM ESCOLAS PÚBLICAS COM BASE EM FATORES SOCIOECONÔMICOS E CULTURAIS: UM ESTUDO DE CASO COM DADOS DO SAEB

WELINGTON JUNIO ALVES DE SOUZA
Orientador: Luciana Balieiro Cosme

Montes Claros
Março de 2025

WELINGTON JUNIO ALVES DE SOUZA

PREVISÃO DO DESEMPENHO ESCOLAR DO 9º ANO EM ESCOLAS PÚBLICAS COM BASE EM FATORES SOCIOECONÔMICOS E CULTURAIS: UM ESTUDO DE CASO COM DADOS DO SAEB

Projeto de Monografia apresentado ao Curso de Graduação em Ciência da Computação do Instituto Federal do Norte de Minas Gerais – Campus Montes Claros, como requisito parcial para a obtenção do grau de Bacharel em Ciência da Computação.

Orientador: Luciana Balieiro Cosme

Montes Claros

Março de 2025

Dedico este trabalho à minha família, pelo apoio incondicional em cada passo desta jornada; à minha orientadora, pela paciência e orientação, fundamentais para o aprimoramento deste estudo; e aos meus amigos, pelas dicas e companheirismo ao longo desse caminho.

Agradecimentos

Agradeço a Deus pelo suporte e força ao longo de toda essa caminhada. À minha família, que sempre esteve ao meu lado, acreditando em mim e tornando cada etapa mais leve. Aos amigos, por tornarem os momentos difíceis mais suportáveis com suas companhias e conselhos. À professora orientadora Luciana, pela orientação dedicada, paciência e honestidade, essenciais para a realização deste trabalho.

"A verdadeira motivação vem de realização, desenvolvimento pessoal, satisfação no trabalho e reconhecimento."
(Frederick Herzberg)

Resumo

Este trabalho exploram os impactos dos fatores socioeconômico-culturais no desempenho em Matemática de alunos do 9º ano de escolas públicas, utilizando dados do Sistema de Avaliação da Educação Básica (SAEB) de 2021. Para facilitar a interpretação, as notas contínuas foram convertidas em categorias binárias ("Proficiente" e "Insuficiente"). A pesquisa integra uma análise exploratória dos dados, seleção de características (por meio do algoritmo CART) e a aplicação de modelos preditivos de aprendizado de máquina, como Regressão Logística e *Random Forest*, combinados com técnicas de balanceamento (SMOTE) para corrigir o desbalanceamento das classes. Os resultados revelam que variáveis relacionadas a atividades extracurriculares e estudo extraclasse apresentam associação positiva com a proficiência, enquanto fatores como idade elevada e baixa condição socioeconômica estão negativamente relacionados ao desempenho. Os achados podem oferecer subsídios para a implementação de políticas educacionais que visem promover a equidade e melhorar a qualidade do ensino.

Palavras-chave: SAEB, Regressão, Educação, Análise Estatística, Modelagem Preditiva.

Abstract

This study analyzes the impact of socioeconomic and cultural factors on the mathematics performance of 9th-grade students in public schools, using data from the Sistema de Avaliação da Educação Básica (SAEB) 2021. To enhance interpretability, continuous scores were transformed into binary categories ("Proficient" and "Insufficient"). The research integrates an exploratory data analysis, feature selection via the CART algorithm, and the application of machine learning predictive models, including Logistic Regression and Random Forest, combined with SMOTE techniques to address class imbalance. The results indicate that variables related to extracurricular activities and out-of-school study are positively associated with proficiency, while higher age and lower socioeconomic status are negatively associated with performance. These findings provide valuable insights for developing educational policies aimed at promoting equity and improving the quality of education.

Keywords: SAEB, Regression, Education, Statistical Analysis, Predictive Modeling.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

A educação brasileira tem experimentado avanços nos indicadores educacionais, o que é evidente ao considerarmos o progresso no Sistema de Avaliação da Educação Básica (SAEB) no período de 2011 a 2020, conforme ilustrado nas Figuras 1.1 e 1.2, que mostram a evolução das notas de matemática e língua portuguesa ao longo desse intervalo. Observa-se um aumento nas médias do 9º ano, com destaque para os anos de 2013 a 2019. No entanto, o cenário revela uma reversão em 2021, com uma redução nos níveis de desempenho comparáveis aos patamares de 2015 para matemática e 2017 para língua portuguesa. Este retrocesso pode estar relacionado a eventos recentes, como a pandemia, conforme mencionado pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira [2022].

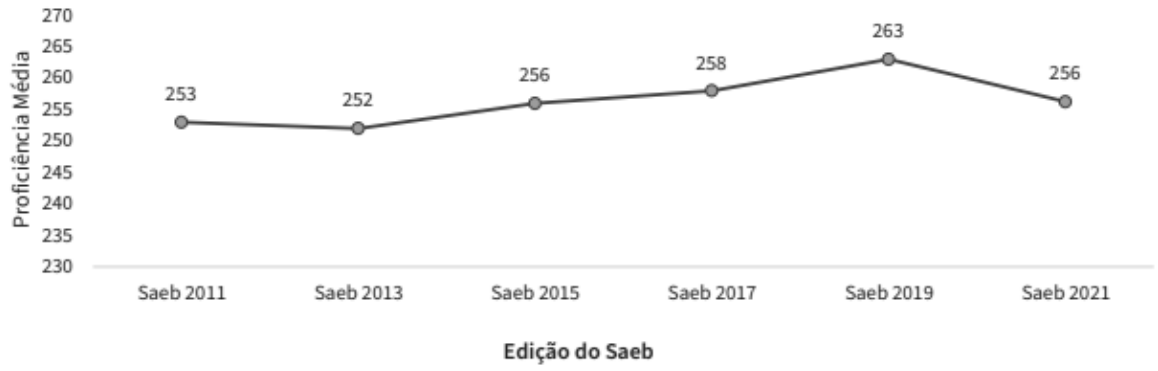

Figura 1.1. Evolução das Proficiências Médias no SAEB, em Matemática, no 9º Ano do Ensino Fundamental – Brasil – 2011 A 2021. Fonte: Direitoria de Avaliação da Educação Básica [2023]

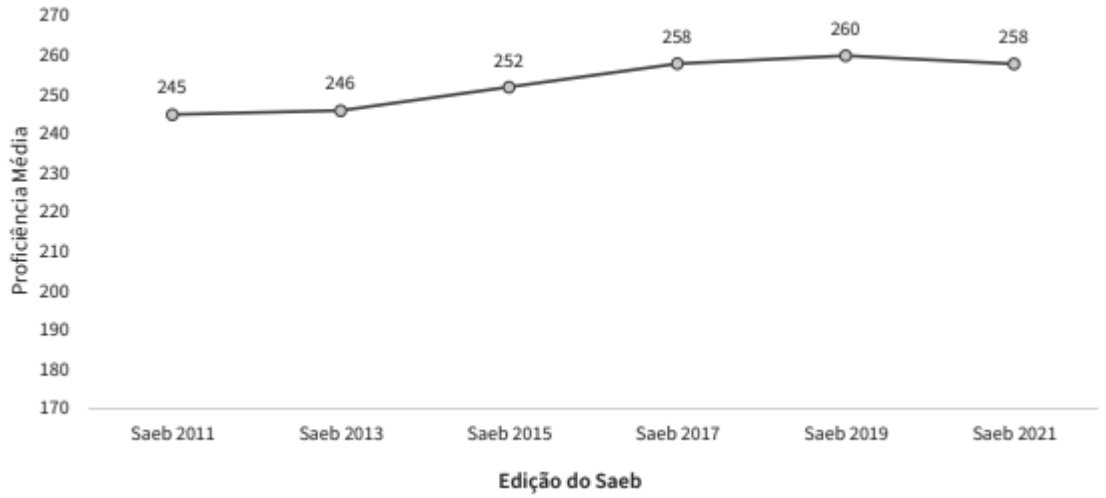

Figura 1.2. Evolução das Proficiências Médias no SAEB, em Língua Portuguesa, no 9º Ano do Ensino Fundamental – Brasil – 2011 A 2021. Fonte: Direitoria de Avaliação da Educação Básica [2023]

Durante o período analisado pelas Figuras 1.1 e 1.2, de 2011 a 2021, o Brasil aprovou projetos nacionais e leis com o intuito de aprimorar os indicadores educacionais no país. Iniciativas como o Plano Nacional de Educação (PNE), estabelecido pela Lei nº 13.005/2014, que definiu metas e estratégias para o avanço da educação [Brasil [2014]] e a implementação da Base Nacional Comum Curricular (BNCC), regulamentada pela Lei nº 13.415/2017 que trouxe diretrizes para o currículo escolar, demonstrando o alinhamento com as necessidades do sistema educacional [Brasil [2017]] que podem ter contribuído para a padronização e qualidade do ensino em todo o país.

No contexto educacional brasileiro, o Artigo 205 da Constituição Federal do Brasil cita "Art. 205. A educação, direito de todos e dever do Estado e da família, será promovida e incentivada com a colaboração da sociedade, visando ao pleno desenvolvimento da pessoa, seu preparo para o exercício da cidadania e sua qualificação para o trabalho."[Brasil [1988]]. No entanto, a legislação precisa lidar com desafios decorrentes das complexidades socioeconômico-culturais, que podem representar obstáculos à efetiva implementação desse direito fundamental.

Essas complexidades referem às condições socioeconômicas das famílias dos alunos e aos aspectos culturais presentes em seu ambiente, como considerado por Soares [2004]. A interseção desses fatores podem influenciar as experiências educacionais dos estudantes, contribuindo para a formação de disparidades no desempenho e no acesso a recursos educacionais. No âmbito socioeconômico, pode haver impacto ao acesso a oportunidades educacionais. Estudantes de contextos mais favorecidos podem ter acesso a recursos adicionais, como material didático de qualidade e apoio suplementar.

Por outro lado, aqueles de origens socioeconômicas mais desafiadoras podem enfrentar obstáculos relacionados à falta de recursos. Quanto aos elementos culturais, tradições e práticas familiares também podem exercer influência.

Diante do contexto atual, marcado pela relevância da educação básica, pelas mudanças legislativas e pelos desafios impostos pela pandemia, além de diversos fatores socioeconômicos e culturais, o objetivo central deste estudo é identificar fatores mais relevantes que podem influenciar o desempenho dos alunos em matemática e prever a proficiência desses alunos com base nesses fatores, com foco específico nos estudantes do 9º ano das escolas públicas, utilizando os dados do SAEB de 2021. A escolha de focar no 9º ano do Ensino Fundamental para a análise da previsão da nota de Matemática no SAEB é importante, pois, segundo dados recentes, quase metade dos alunos brasileiros não conclui o Ensino Fundamental na idade adequada [Palhares [2024]]. Além disso, a matemática é uma disciplina avaliada em exames educacionais internacionais como o PISA (*Programa Internacional de Avaliação de Estudantes*). O desempenho em matemática pode refletir a qualidade geral da educação em um país. De acordo com a OCDE (Organização para a Cooperação e Desenvolvimento Econômico), que realiza o PISA, o Brasil está abaixo de sua média, como mostrado por Letícia Mori [2023].

Para alcançar o objetivo deste trabalho, são aplicadas técnicas de seleção de características, análise exploratória de dados e modelagem preditiva, com ênfase na aplicação da regressão logística, visando prever o desempenho dos alunos com base nas variáveis analisadas. Essa abordagem permitiu não apenas identificar a relação entre as variáveis socioeconômicas e culturais e o desempenho dos estudantes, mas também pode contribuir para a criação de modelos preditivos que possam subsidiar políticas públicas e estratégias educacionais voltadas à melhoria da proficiência em matemática.

1.1 Organização do documento

Após esta introdução, o Capítulo 2 apresenta os conceitos fundamentais sobre o SAEB, análise exploratória e modelos de aprendizado utilizados. O Capítulo 3 discute trabalhos relacionados ao desempenho escolar e fatores socioeconômico-culturais. No Capítulo 4, detalha-se a metodologia adotada. O Capítulo 5 expõe os resultados da seleção de características, da análise exploratória fornecendo visualizações sobre a relação entre diferentes fatores e o desempenho em matemática, além da aplicação dos modelos de aprendizado para previsão das notas dos alunos. Por fim, o Capítulo 6 apresenta as considerações finais do estudo.

Capítulo 2

Conceitos Básicos

2.1 Sistema de Avaliação da Educação Básica (SAEB)

O Sistema de Avaliação da Educação Básica (SAEB) é um instrumento para avaliar o desempenho acadêmico. Realizada a cada dois anos, esta avaliação nacional abrange alunos do 2º, 5º e 9º anos do Ensino Fundamental, além dos alunos do 3º ano do Ensino Médio [Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira [2021]Diretoria de Gestão e Planejamento [2021]].

Ao longo de sua história, o SAEB passou por diversas reestruturações. Criado em 1990 como uma única avaliação, o sistema foi reformulado em 2005. Em 2013, ocorreu uma significativa transformação com a inclusão da Avaliação Nacional da Alfabetização (ANA), ampliando o SAEB para englobar três avaliações externas em larga escala: a Avaliação Nacional da Educação Básica (ANEB), a Avaliação Nacional do Rendimento Escolar (ANRESC - conhecida como Prova Brasil) e a própria ANA. Em 2019, o SAEB passou por uma nova reestruturação para alinhar-se à Base Nacional Comum Curricular (BNCC), que agora tem um papel central na formulação dos itens avaliativos, especialmente nos 2º e 9º anos do ensino fundamental, cobrindo áreas como Língua Portuguesa, Matemática, Ciências da Natureza e Ciências Humanas. As siglas ANA, ANEB e ANRESC foram unificadas sob o nome SAEB, identificando as diferentes etapas, áreas de conhecimento e tipos de instrumentos aplicados [Diretoria de Avaliação da Educação Básica [2023]]. Em 2021, foi introduzida a avaliação da educação infantil, utilizando questionários eletrônicos exclusivamente para professores, diretores e gestores educacionais em nível municipal e estadual [Diretoria de Gestão e Planejamento [2021]]. A avaliação é realizada a cada dois anos e abrange escolas urbanas e rurais, tanto da rede pública quanto privada, seguindo critérios específicos estabelecidos. Ela

é composta tanto por uma forma censitária, que inclui escolas públicas com pelo menos 20 estudantes do 5º e 9º anos do Ensino Fundamental, quanto por amostragem, que engloba escolas privadas do 5º e 9º anos do Ensino Fundamental com no mínimo 10 estudantes em turmas regulares, além de escolas públicas e privadas da 3ª série do Ensino Médio com 10 ou mais estudantes [Diretoria de Avaliação da Educação Básica [2023]]. Critérios como série, unidade federativa, dependência administrativa, localização e tamanho da escola são considerados [Diretoria de Gestão e Planejamento [2021]].

O SAEB utiliza uma variedade de instrumentos e metodologias. A Teoria de Resposta ao Item (TRI) é um alicerce metodológico que permite comparar o desempenho dos alunos em diferentes períodos por meio da coerência de suas respostas, enquanto a análise hierárquica, realizada por meio de modelos de regressão, aprofunda a compreensão, controlando variáveis relevantes. As Matrizes de Referência estabelecem o referencial curricular mínimo, descrevendo competências e habilidades por disciplina e série. Os Testes Padronizados, compostos por itens de múltipla escolha, passam por revisões e validações estatísticas. Os Questionários de Contexto coletam dados socioeconômicos, culturais e escolares para a compreensão do ambiente educacional, visando capturar a diversidade e representatividade do sistema educacional brasileiro [Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira [2021]] [Diretoria de Gestão e Planejamento [2021]].

Os resultados do SAEB são apresentados por uma escala de desempenho, conforme ilustrado na Tabela 2.1. Esta tabela resume as competências e habilidades esperadas para cada nível, refletindo o desenvolvimento dos estudantes. Cada série e disciplina possui sua própria escala, que atua como uma régua para verificar o domínio dos alunos em determinados assuntos. As competências são classificadas em níveis de proficiência, onde cada nível representa um avanço em relação ao anterior [Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira [2021]] [Diretoria de Gestão e Planejamento [2021]]. É importante observar que o SAEB não avaliou habilidades do Nível 0 para o 9º ano. Estudantes com desempenho abaixo de 200 pontos requerem atenção especial, pois ainda não demonstram habilidades elementares esperadas para essa etapa escolar, conforme cita Direitoria de Avaliação da Educação Básica [2020].

Nível	Desempenho	Competências Resumidas
1	Entre 200 e 225	Reconhecer números racionais maiores/menores e interpretar dados em tabelas e gráficos.
2	Entre 225 e 250	Reconhecer frações equivalentes e associar dados em gráficos e tabelas.
3	Entre 250 e 275	Identificar ângulos, localizar objetos em gráficos e resolver problemas proporcionais.
4	Entre 275 e 300	Localizar pontos em um plano cartesiano e resolver problemas com números racionais.
5	Entre 300 e 325	Resolver problemas com ângulos e converter unidades de medida.
6	Entre 325 e 350	Trabalhar com coordenadas e converter unidades de massa e volume.
7	Entre 350 e 375	Resolver problemas complexos com ângulos e áreas de figuras.
8	Entre 375 e 400	Trabalhar com propriedades de triângulos e conversões de capacidade.
9	Maior que 400	Resolver problemas sobre ângulos de polígonos e expressões algébricas em sequências.

Tabela 2.1. Escala Resumida de Proficiência de Matemática do 9º Ano do Ensino Fundamental. Fonte: Adaptada de Direitoria de Avaliação da Educação Básica [2020].

2.2 Análise Exploratória de Dados

A Análise Exploratória de Dados (AED) é uma etapa fundamental no processo de análise de dados, cuja importância foi destacada pelas ideias de John W. Tukey na década de 1960. Em sua obra seminal "*Exploratory Data Analysis*", Tukey propôs uma abordagem, enfatizando a importância de explorar os dados antes de realizar inferências estatísticas formais [Andrew Bruce & Peter Bruce [2019]].

Ao longo do tempo, a AED evoluiu, especialmente com o advento da capacidade computacional e o desenvolvimento de softwares especializados. Essa evolução permitiu lidar com conjuntos de dados cada vez maiores e mais complexos, possibilitando uma compreensão mais profunda da natureza dos dados, conforme cita Andrew Bruce & Peter Bruce [2019]. Os autores ainda agregam que na análise exploratória, são examinadas diversas características dos dados, incluindo seus tipos e estruturas. Dados contínuos, discretos, categóricos, binários e ordinais são alguns dos tipos comuns discutidos. Além disso, a AED se concentra em entender as estruturas dos dados, com destaque para os dados retangulares, sendo uma matriz bidimensional com linhas indicando registros e colunas indicando características, amplamente utilizados na ciência de dados.

Na visão de Andrew Bruce & Peter Bruce [2019], métricas como média, mediana, desvio padrão e intervalo interquartil são ferramentas essenciais na estimação da localização e variabilidade dos dados. Essas medidas fornecem informações sobre a distribuição e dispersão dos dados, ajudando a identificar padrões, *outliers* e possíveis relações entre as variáveis. A AED permite compreender a estrutura dos dados, identificar padrões e detectar anomalias, destacando sua relevância na seleção de técnicas adequadas, na identificação de problemas nos dados, na geração de hipóteses e na visualização dos dados para facilitar a interpretação.

2.3 Modelos de Aprendizado de Máquina Supervisionados

Os modelos de aprendizado de máquina supervisionados são construídos a partir de um conjunto de dados rotulado, no qual cada entrada está associada a uma saída conhecida. O objetivo é identificar uma função preditiva que relacione essas variáveis, permitindo a generalização para novos dados. Esse processo envolve a busca por uma hipótese que melhor represente a relação subjacente nos dados, minimizando erros e maximizando a capacidade preditiva do modelo. Entre os métodos mais utilizados nesse contexto estão os classificadores lineares, como a regressão logística, além de técnicas mais avançadas, como árvores de decisão e redes neurais. A escolha do modelo mais adequado depende das características dos dados e do problema a ser resolvido [Russell & Norvig [2013]].

2.3.1 Desbalanceamento de Classes

O desbalanceamento de classes é um problema em aprendizado de máquina e ocorre quando a distribuição das classes dentro de um conjunto de dados é desigual, com uma classe majoritária possuindo um número maior de instâncias do que a classe minoritária. Essa discrepância pode impactar negativamente o desempenho dos modelos de classificação, pois muitos algoritmos supervisionados baseiam suas predições em padrões estatísticos gerais dos dados e, quando treinados em conjuntos desbalanceados, tendem a priorizar a classe mais frequente, ignorando ou classificando erroneamente a classe minoritária. Como resultado, os modelos podem apresentar alta acurácia geral, mas desempenho insatisfatório na identificação dos exemplos da classe menos representada [Hasib et al. [2020]].

Para contornar esse problema, diferentes abordagens têm sido propostas, incluindo técnicas de balanceamento no nível dos dados, como o *Random Under-Sampling* (RUS) e o *Synthetic Minority Over-sampling Technique* (SMOTE). O RUS atua removendo

aleatoriamente amostras da classe majoritária, reduzindo a discrepância entre as classes. Embora essa técnica possa melhorar a imparcialidade do modelo, também pode resultar na perda de informações valiosas, caso dados relevantes sejam descartados [Hasib et al. [2020]].

Por outro lado, o SMOTE busca aumentar a representatividade da classe minoritária através da geração de novas amostras sintéticas, criadas a partir da interpolação dos exemplos existentes. Em vez de replicar simplesmente as instâncias minoritárias, o algoritmo seleciona, para cada exemplo, seus k-vizinhos mais próximos dentro da mesma classe e gera novas amostras interpolando linearmente entre a instância original e esses vizinhos, utilizando um fator de ponderação aleatório entre 0 e 1. Esse procedimento promove a expansão contínua do espaço amostral da classe minoritária, contribuindo para a formação de uma fronteira de decisão mais suave e robusta, além de reduzir o risco de *overfitting*. A escolha adequada dos parâmetros – como o número de vizinhos (k) e a taxa de oversampling – é fundamental para evitar a introdução de ruídos ou a criação de regiões sintéticas pouco representativas [Hasib et al. [2020]][Chawla et al. [2002]].

2.3.2 Regressão Logística

A regressão logística é uma técnica estatística utilizada para modelar a relação entre uma variável dependente qualitativa e um conjunto de variáveis explicativas. Diferente da regressão linear, que pressupõe uma variável dependente quantitativa, a regressão logística é aplicada quando a variável de interesse é categórica, como no caso de eventos dicotômicos (sim/não) ou multinomiais (três ou mais categorias). A estimação dos parâmetros do modelo é realizada pelo método de máxima verossimilhança, garantindo que as probabilidades previstas estejam sempre no intervalo entre 0 e 1. Além disso, a interpretação dos coeficientes é feita por meio da razão de chances (*odds ratio*), que indica o impacto das variáveis explicativas na probabilidade de ocorrência do evento de interesse [Fávero & Belfiore [2017]].

No modelo de regressão logística, o objetivo é estimar a probabilidade p de ocorrência do evento de interesse. Para isso, utiliza-se a transformação logarítmica das chances (ou *odds*), denominada *logito*, definida como:

$$logit(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k, \tag{2.1}$$

onde $x_1, x_2, \ldots, x_k$ são as variáveis explicativas e $\beta_0, \beta_1, \ldots, \beta_k$ são os coeficientes do modelo.

A partir da Equação 2.3.2 do *logito*, pode-se obter a função que relaciona os preditores à probabilidade do evento:

$$p = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}}{1 + e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}}.$$

(2.2)

Os parâmetros do modelo são estimados pelo método de máxima verossimilhança, que escolhe os valores de $\beta_0, \beta_1, \ldots, \beta_k$ de forma a maximizar a probabilidade de observar os dados amostrais, garantindo que as probabilidades previstas estejam sempre entre 0 e 1.

A interpretação dos coeficientes é feita por meio da razão de chances (*odds ratio*). Para um coeficiente β_j, o valor $\exp(\beta_j)$ indica o efeito multiplicativo sobre as chances do evento ocorrer para um aumento unitário na variável x_j, mantendo as demais variáveis constantes [Fávero & Belfiore [2017]][James et al. [2013]].

2.3.3 Árvore de Decisão

As árvores de decisão são algoritmos amplamente utilizados em aprendizado de máquina para tarefas de classificação e regressão, pois oferecem interpretabilidade e simplicidade na modelagem de dados. Elas funcionam segmentando iterativamente o espaço de atributos, criando regras de decisão hierárquicas que permitem a classificação ou previsão de novas amostras com base em padrões extraídos dos dados de treinamento. A construção de uma árvore de decisão envolve a escolha do melhor atributo para divisão em cada nó, como mostra a Figura 2.1, levando em consideração métricas como ganho de informação ou redução de impureza [Ragsdale [2017]].

2.3.3.1 Algoritmo CART

Dentre os algoritmos de indução de árvores de decisão, o CART (*Classification and Regression Trees*) destaca-se por sua abordagem baseada em divisões binárias, onde cada nó interno possui exatamente duas ramificações [Wu & Kumar [2009]]. Ele busca minimizar a heterogeneidade dentro dos subconjuntos gerados a cada divisão. Essa característica torna o CART adequado para lidar com bases de dados complexas e de alta dimensionalidade, garantindo uma separação eficaz das classes. Além disso, o algoritmo permite a poda da árvore, removendo divisões que não contribuem significativamente para a precisão do modelo, o que reduz o risco de *overfitting* e melhora sua capacidade de generalização [Breiman et al. [1984]].

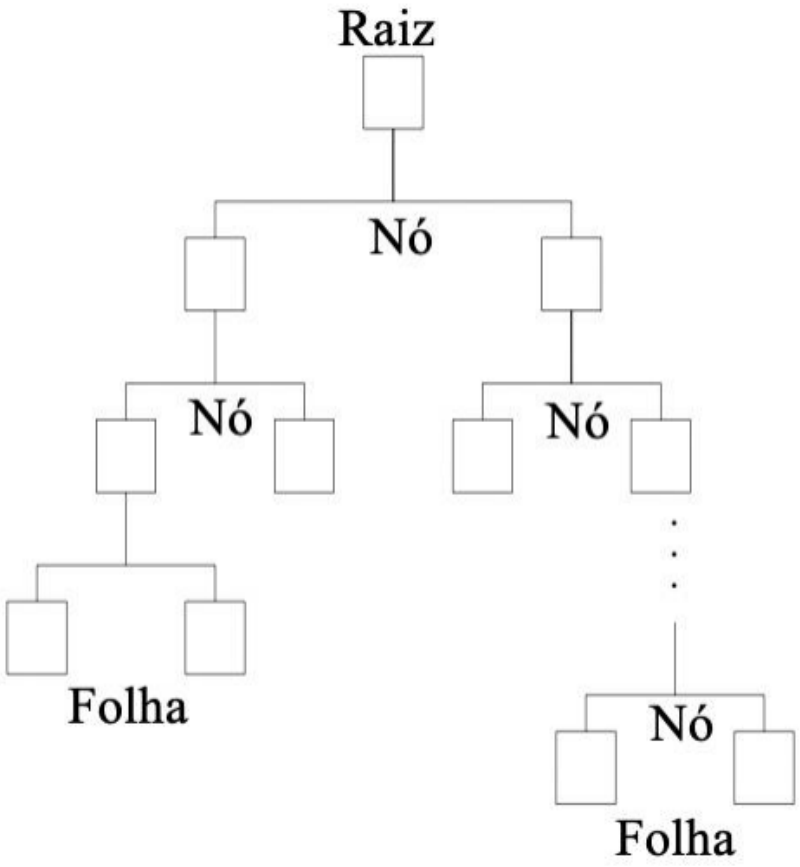

Figura 2.1. Exemplo de subdivisão da raiz em nós até chegar a folha em uma árvore de decisão. Fonte: Autoria Própria

2.3.3.2 Random Forest

O *random forest* é um algoritmo de aprendizado de máquina baseado em *ensemble* que combina o resultado de múltiplas árvores de decisão para melhorar a acurácia e reduzir o risco de *overfitting*. Cada árvore é construída a partir de uma amostra *bootstrap* dos dados e, em cada nó, é considerado apenas um subconjunto aleatório de variáveis para a divisão. Essa estratégia aumenta a diversidade entre as árvores, tornando o modelo final possivelmente mais robusto e estável [Breiman [2001]].

2.3.4 Otimização de Hiperparâmetros: Grid Search

No treinamento de modelos de aprendizado de máquina, a escolha adequada dos hiperparâmetros é fundamental para alcançar um desempenho ótimo. O *grid search* é uma técnica exaustiva e sistemática que consiste em definir uma grade de valores candidatos para cada hiperparâmetro e, em seguida, testar todas as combinações possíveis. Antes de ajustar os hiperparâmetros, o *grid search* permite explorar de forma organizada o

espaço de configurações, avaliando o desempenho do modelo para cada combinação por meio de validação cruzada. Dessa forma, mesmo que seja computacionalmente custoso para espaços de parâmetros muito grandes, essa abordagem possibilita a identificação da configuração que melhor otimiza a métrica de interesse [Pedregosa et al. [2011]].

2.3.5 Métricas de Avaliação

Na avaliação de modelos de classificação, é fundamental utilizar métricas que permitam mensurar o desempenho do modelo de forma abrangente. Entre as principais métricas, destacam-se a acurácia, o *recall*, o *F1-score* e a matriz de confusão.

A **matriz de confusão** é uma ferramenta que resume os resultados de um classificador, comparando as classes reais com as predições do modelo. Em um problema binário, a matriz é composta por quatro elementos:

Verdadeiros Positivos (VP): número de instâncias positivas corretamente classificadas.

Falsos Positivos (FP): número de instâncias negativas incorretamente classificadas como positivas.

Verdadeiros Negativos (VN): número de instâncias negativas corretamente classificadas.

Falsos Negativos (FN): número de instâncias positivas incorretamente classificadas como negativas.

A **acurácia** é definida como a razão entre o número de predições corretas e o total de instâncias:

$$Acurcia = \frac{VP + VN}{VP + VN + FP + FN}. \tag{2.3}$$

Embora seja uma medida intuitiva, a acurácia pode ser enganosa em situações de desbalanceamento de classes.

O ***recall*** (ou sensibilidade) mede a capacidade do modelo em identificar corretamente as instâncias positivas, sendo definido por:

$$Recall = \frac{VP}{VP + FN}. \tag{2.4}$$

A **precisão** quantifica a proporção de instâncias positivas corretamente identificadas dentre todas as instâncias classificadas como positivas:

$$Preciso = \frac{VP}{VP + FP}. \tag{2.5}$$

O ***F1-score*** é a média harmônica entre precisão e recall, oferecendo uma medida equilibrada em situações de desbalanceamento entre as classes:

$$F1 = 2 \times \frac{Preciso \times Recall}{Preciso + Recall}. \tag{2.6}$$

Essas métricas fornecem uma visão abrangente do desempenho do modelo, permitindo identificar pontos fortes e limitações na capacidade de generalização do classificador [Pedregosa et al. [2011]][Fawcett [2006]].

Capítulo 3

Trabalhos Relacionados

A abordagem do desempenho dos alunos no contexto educacional tem sido objeto de diversos estudos, os quais adotam métodos variados e exploram diversos conjuntos de dados. Neste contexto, investigações como as conduzidas por Karakolidis et al. [2016], Santos, A [2018], Gomes et al. [2020], Melo et al. [2021] e Capelato & Périco [2023], têm contribuído para a exploração das variáveis envolvidas no desempenho dos alunos.

No texto sobre o estudo de Karakolidis et al. [2016], uma regressão linear múltipla foi empregada para examinar os possíveis determinantes do desempenho em matemática dos alunos, considerando fatores individuais e escolares. Os resultados, baseados nos dados do Programa Internacional de Avaliação de Estudantes(PISA), destacaram a importância de variáveis como autoconfiança em matemática, status socioeconômico da escola, gênero e participação em educação pré-primária na predição do desempenho em matemática dos alunos na Grécia.

A pesquisa conduzida por Santos, A [2018] concentrou-se na evolução do desempenho em matemática no Brasil, utilizando dados de proficiência do SAEB ao longo do período de 2007 a 2017. O estudo oferece uma análise do desempenho dos estudantes brasileiros do 5º ano do Ensino Fundamental em Matemática. O estudo revela tendências e padrões no desenvolvimento das habilidades matemáticas dos alunos, adotando uma abordagem descritiva numérica para avaliar as notas, analisando dados agrupados por diferentes categorias, como escolas públicas, privadas, estaduais e o total. Revelou-se que houve um pequeno aumento no desempenho dos estudantes em Matemática a cada edição do SAEB. Em média, as notas dos estudantes brasileiros matriculados no 5º ano do Ensino Fundamental em Matemática foram de 179,17 em 2007 para 211,51 em 2017 na região Norte, de 179,19 em 2007 para 207,12 em 2017 na região Nordeste e de 202,31 em 2007 para 233,92 em 2017 na região Sudeste. Além disso, o estudo indicou que há uma defasagem de aprendizagem nas diferentes regiões brasileiras, especificamente, as médias de desempenho em Matemática dos estudantes nas regiões

norte e nordeste foram observadas como menores que a média nacional, enquanto nas regiões sul, sudeste e centro-oeste as médias ficaram acima da média nacional.

A abordagem preditiva adotada por Gomes et al. [2020] utilizou a Regressão em árvore com o algoritmo CART com dados do ENEM para analisar o desempenho em matemática dos estudantes do Ensino Médio. O modelo desenvolvido explicou 29,97% da variância do desempenho nessa disciplina. O estudo considerou 53 preditores, dos quais apenas 7 foram identificados como tendo importância preditiva. Esses preditores incluíam aspectos socioeconômicos, características escolares e motivacionais dos estudantes.

Um estudo conduzido por Melo et al. [2021] investigou o impacto das variáveis socioeconômicas no desempenho dos alunos no Exame Nacional do Ensino Médio (ENEM). Os pesquisadores adotaram uma abordagem que envolveu a organização dos dados a nível municipal e a aplicação de modelos de regressão linear. O estudo examinou uma série de variáveis e identificou aquelas com poder explicativo significativo para as pontuações obtidas na prova objetiva do ENEM.

Os resultados revelaram que o nível de escolaridade e profissionalização da mãe, a raça do estudante e a renda média da família estão entre as variáveis mais relevantes para o desempenho dos alunos. Além disso, o percentual de estudantes matriculados em escolas particulares com bolsa também demonstrou ter impacto significativo, embora em menor escala. Observou-se ainda que a presença de infraestrutura escolar, como refeitório e laboratório de ciências, e o ensino de espanhol no terceiro ano do ensino médio, embora tenham um impacto menor, ainda influenciam o desempenho.

Quanto à redação, as variáveis associadas ao desempenho são semelhantes às da prova objetiva, porém com um impacto menos pronunciado na média. A introdução de um componente de localização no modelo para as notas de redação revelou que fatores regionais, distintos dos socioeconômicos, também desempenham um papel importante no desempenho das pontuações entre os municípios.

Capelato & Périco [2023] conduziram uma análise do desempenho dos estudantes do terceiro ano do Ensino Médio em escolas públicas do estado de São Paulo, utilizando dados do Sistema de Avaliação do Rendimento Escolar do Estado de São Paulo (SARESP), uma prova aplicada anualmente aos alunos do 2º e 5º anos do ensino fundamental, dos 6º ao 9º ano do ensino fundamental e das 1ª a 3ª séries do ensino médio, em turmas regulares em escolas públicas do Estado de São Paulo. Através de uma análise técnica, o estudo revelou que as variáveis que mais influenciaram o desempenho em matemática foram, em ordem decrescente de importância, a qualificação dos professores (39,3%), o esforço dos professores (27%), a quantidade de alunos por turma (22,4%) e o nível socioeconômico da escola (11,3%). Ademais, a diminuição da quantidade de

alunos por turma nas regiões de São Paulo, Campinas, Barretos e Franca demonstrou potencial para aumentar tanto a eficiência quanto a pontuação final dos alunos.

A Tabela 3.1 resume os trabalhos analisados nesta seção, destacando características como a avaliação do SAEB em escolas públicas, foco na nota de matemática (Mat.) e o uso de métodos de regressão. A última linha da tabela apresenta a proposta deste trabalho, que se diferencia dos demais ao considerar todos esses critérios, mesmo que a literatura tenha contribuído para seu desenvolvimento.

Autor	**SAEB**	**Escola**	**Mat.**	**Método de Regressão**
1. Karakolidis et al. [2016]	Não (PISA)	Não informa	Sim	Sim
Santos, A [2018]	Sim (5° Ano)	Pública e Privada	Sim	Não
Gomes et al. [2020]	Não (Enem)	Pública e Privada	Sim	Sim
Melo et al. [2021]	Não (Enem)	Pública e Privada	Não	Sim
Capelato & Périco [2023]	Não (SARESP 3° Ano Ensino Médio)	Pública	Sim	Não
Este Estudo	Sim (9° Ano)	Pública	Sim	Sim

Tabela 3.1. Resumo dos trabalhos relacionados, incluindo as principais características analisadas: avaliação do SAEB, tipo de escola, nota de matemática (Mat.), e uso de métodos de regressão. A última linha compara a proposta deste trabalho com os critérios dos demais estudos, destacando as diferenças e contribuições específicas. Fonte: Autoria Própria.

Analisando os estudos, replicar o estudo de Karakolidis et al. [2016] no Brasil pode ser desafiador devido às dificuldades em obter dados precisos, como em relação à frequência na educação pré-primária e até mesmo à medição da autoconfiança em matemática. Essas variáveis podem não estar disponíveis de forma abrangente ou padronizada, o que tornaria complexa a comparação com os resultados obtidos na pesquisa grega mas as outras caracteristicas estão presentes no nosso estudo e seria interessante compará-las. Já nos estudos de Gomes et al. [2020] e Melo et al. [2021], embora o ENEM e a aréa de estudo desse trabalho sejam distintas, certas variáveis socioeconômicas podem influenciar de maneira semelhante. De forma similar, o estudo de Capelato & Périco [2023] concentra na eficiência escolar utilizando outra fonte

de dados, apesar de utilizar outra técnica e não envolver diretamente a regressão, ele aborda fatores importantes para o desempenho escolar, a seleção e análise de variáveis relevantes e a interpretação de resultados. Embora o estudo de Santos, A [2018] não realizar cálculos específicos ou uma seleção de características, ele contextualiza o SAEB de forma descritiva no contexto educacional brasileiro. Essa abordagem descritiva e exploratória oferece uma compreensão abrangente das tendências e padrões no desempenho dos alunos ao longo do tempo, o que pode servir como uma referência para complementar os métodos de regressão utilizados nessa pesquisa.

Em suma, os estudos relacionados oferecem uma visão das variáveis que podem influenciar o desempenho dos alunos em diferentes contextos educacionais. Enquanto alguns se concentram em métodos de regressão, outros exploram orientações descritivas e contextuais no desempenho dos alunos ao longo do tempo. Essa diversidade de abordagens pode ser visualizada de forma sumarizada na Tabela 3.1 de trabalhos relacionados, juntamente com a Tabela 3.3, que apresenta as características selecionadas por cada estudo e se são utilizadas nesse estudo. Este trabalho busca combinar abordagens preditivas com uma análise exploratória das características. Dessa forma, contribuirá para a compreensão dos possíveis fatores que podem influenciar o desempenho dos alunos no contexto específico dos 9º anos das escolas públicas do Ensino Fundamental.

Estudo	**Trabalho**
1	Karakolidis et al. [2016]
2	Santos, A [2018]
3	Gomes et al. [2020]
4	Melo et al. [2021]
5	Capelato & Périco [2023]

Tabela 3.2. Tabela de Estudos para referência da Tabela 3.3. Fonte: Autoria Própria.

Característica	**1**	**2**	**3**	**4**	**5**	**Este Estudo**
Gênero	Sim	Não	Sim	Não	Não	**Sim**
Status imigração	Sim	Não	Não	Não	Não	Não
Nível econômico	Sim	Não	Sim	Não	Não	**Sim**
Autoconceito em matemática	Sim	Não	Não	Não	Não	Não
Autoeficácia em matemática	Sim	Não	Não	Não	Não	Não
Ansiedade em matemática	Sim	Não	Não	Não	Não	Não
Frequência do ensino pré-escolar	Sim	Não	Não	Não	Não	Não
Frequência do ensino pré-primário	Sim	Não	Não	Não	Não	Não
Nível econômico da escola	Sim	Não	Não	Não	Sim	Não
Região do Brasil	Não	Sim	Sim	Não	Não	**Sim**
Rede(Pública, Privada, Estadual)	Não	Sim	Não	Não	Não	Não
Características escolares	Não	Não	Sim	Não	Não	Não
Características motivacionais dos estudantes	Não	Não	Sim	Sim	Não	Não
Renda familiar	Não	Não	Sim	Sim	Não	Não
Escola frequentada no ensino fundamental e médio	Não	Não	Sim	Não	Não	**Sim**
Estudaram com bolsa	Não	Não	Não	Sim	Não	Não
Variáveis econômicas	Não	Não	Não	Sim	Não	Não
Variáveis raciais	Não	Não	Não	Sim	Não	**Sim**
Variáveis de perfil instrucional da mãe	Não	Não	Não	Sim	Não	**Sim**
Variáveis de infraestrutura e ensino escolares	Não	Não	Não	Sim	Não	Não
Adequação da formação docente	Não	Não	Não	Não	Sim	Não
Quantidade de alunos por turma	Não	Não	Não	Não	Sim	Não
Esforço docente	Não	Não	Não	Não	Sim	Não
Região(Estado de São Paulo)	Não	Não	Não	Não	Sim	Não

Tabela 3.3. Tabela de seleção das variáveis estudadas pelos Trabalhos Relacionados e se são utilizadas neste estudo. Tabela 3.2 de referências dos estudos. Fonte: Autoria Própria.

Capítulo 4

Metodologia

4.1 Metodologia

Este capítulo descreve a metodologia utilizada para a análise exploratória e aplicação dos métodos preditivos na previsão da proficiência em matemática, com base nos dados do SAEB. O estudo foca nos alunos do 9º ano de escolas públicas buscando identificar e analisar os fatores mais relevantes que podem influenciar o desempenho dos alunos em matemática.

A metodologia foi estruturada em quatro etapas, resumidas na Figura 4.1:

1. **Preparação dos Dados**: Realizou-se a análise, extração e limpeza dos dados para garantir sua qualidade, incluindo a classificação da proficiência com base nas notas, remoção de valores ausentes e padronização das variáveis, conforme descrito na Seção 4.2.

2. **Seleção de Características**: Utilizou-se o algoritmo CART para identificar as questões mais relevantes entre as 31 disponíveis, avaliando a importância relativa de cada uma e o impacto da perda de dados, conforme detalhado na Seção 5.1.

3. **Análise Exploratória de Dados**: Foram calculadas estatísticas descritivas e realizadas visualizações para compreender a distribuição e padrões das variáveis, conforme descrito na Seção 5.2.

4. **Modelos Preditivos**: Aplicaram-se quatro modelos de classificação: (i) Regressão Logística, (ii) *Random Forest* com Hiperparâmetros Padrão, (iii) Regressão Logística otimizada com *Grid Search* e (iv) *Random Forest* otimizado com *Grid Search*. Diferentes técnicas de balanceamento de classes foram testadas, e os modelos comparados com base em métricas como acurácia, precisão, *recall* e

F1-Score. Os resultados são apresentados em tabelas e matrizes de confusão, conforme descrito na Seção 5.3.

As etapas foram implementadas com ferramentas computacionais como *Minitab* (Minitab, LLC [2023]), *Jupyter Notebook* (Jupyter [2015]), *Google Colab* (Google [2023]) e a linguagem *Python*(Python [2023]).

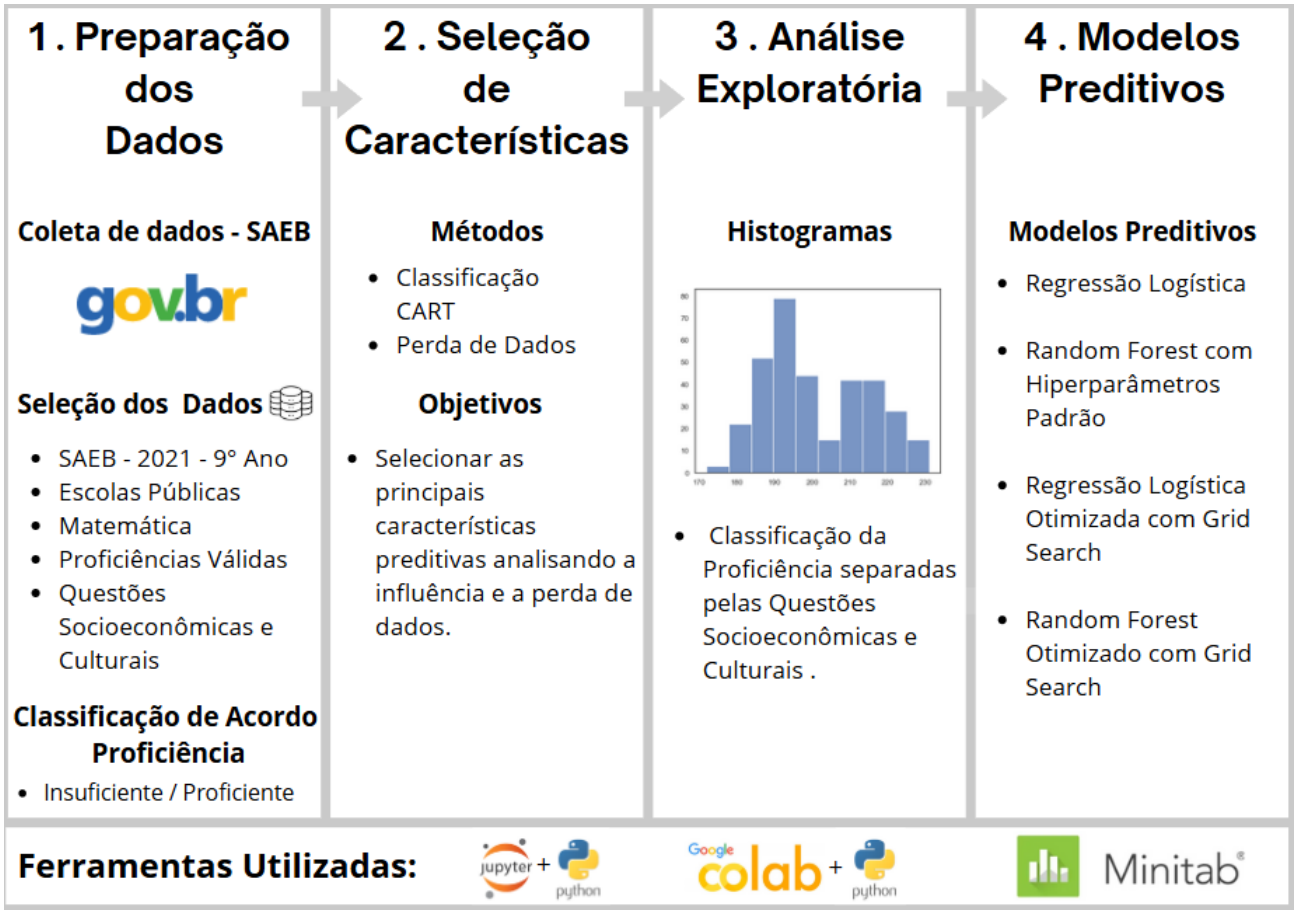

Figura 4.1. Resumo das etapas metodológicas do estudo, desde a preparação dos dados até a aplicação dos métodos preditivos. Fonte: Autória Própria

4.2 Preparação dos Dados

4.2.1 Preparação Inicial

Após o *download* da base de dados do SAEB [Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira [2023]], foram selecionados os registros dos alunos do 9º ano das escolas públicas do Ensino Fundamental que participaram do exame. A etapa de preparação incluiu o tratamento de registros ausentes ou inconsistentes, resultando em um conjunto de dados limpo e adequado para análise.

A divisão entre alunos com proficiências válidas e inválidas está ilustrada na Figura 4.2. As proficiências inválidas correspondem a registros com ausência de respostas ou respostas insuficientes, impossibilitando o cálculo de uma nota. Foram considerados apenas os registros com proficiências válidas, totalizando 1.886.179 observações.

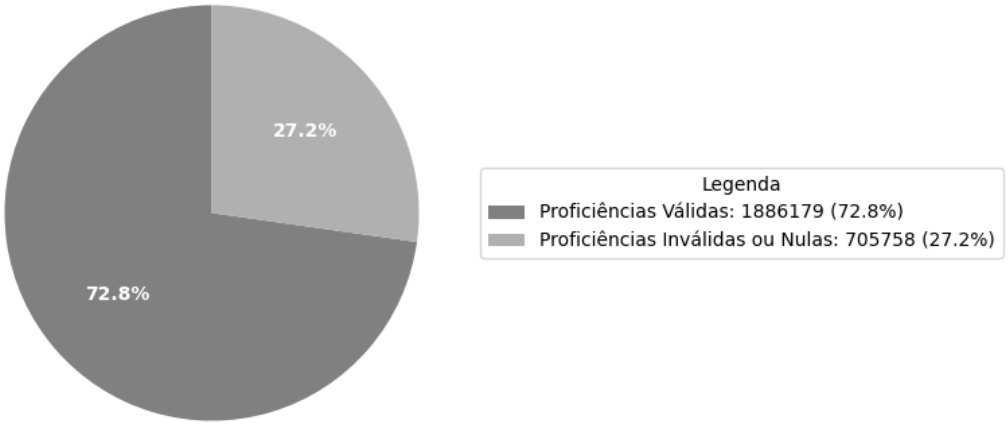

Figura 4.2. Quantidade de alunos participantes do SAEB, de Escolas Públicas, em Matemática, no 9º Ano do Ensino Fundamental, Brasil. Fonte: Autoria Própria com base nos dados do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira [2023].

Na Tabela 4.1, são apresentadas as variáveis do estudo no formato de saída *head*. Exceto a variável de proficiência, todas incluem alternativas codificadas numericamente ou alfabeticamente, representando os fatores socioeconômicos e culturais analisados.

Outro aspecto relevante é a análise das respostas das questões, conforme ilustrado na Tabela 4.1. Algumas alternativas contêm os símbolos '.' e '*', indicando valores nulos ou em branco. Esses valores serão tratados de forma específica para cada situação, evitando a remoção generalizada, que poderia resultar em grande perda de dados. Essa abordagem visa preservar o maior número possível de registros úteis, garantindo a qualidade do conjunto de dados para as etapas seguintes.

Variável	Aluno 1	Aluno 2	Aluno 3 . . .
ID_{REGIAO}	1	1	1
ID_{AREA}	2	2	2
$\text{ID}_{LOCALIZACAO}$	1	1	1
$\text{PROFICIENCIA}_{MTSAEB}$	234.70	165.20	208.60
$\text{NU}_{TIPONIVELINSE}$	4	5	4
$\text{TX}_{RESPQ}01$	A	A	B
$\text{TX}_{RESPQ}02$	B	E	B
$\text{TX}_{RESPQ}04$	A	C	A
$\text{TX}_{RESPQ}05$	B	B	B
$\text{TX}_{RESPQ}07$	D	F	B
$\text{TX}_{RESPQ}08$	F	F	B
$\text{TX}_{RESPQ}09a$	B	B	A
$\text{TX}_{RESPQ}09b$	B	C	B
$\text{TX}_{RESPQ}09c$	C	C	B
$\text{TX}_{RESPQ}09d$	A	C	C
$\text{TX}_{RESPQ}09e$	B	C	A
$\text{TX}_{RESPQ}09f$	C	C	B
$\text{TX}_{RESPQ}10a$	A	A	B
$\text{TX}_{RESPQ}10b$	A	B	A
$\text{TX}_{RESPQ}10c$	A	A	A
$\text{TX}_{RESPQ}13$	C	B	C
$\text{TX}_{RESPQ}14$	D	D	A
$\text{TX}_{RESPQ}15$	B	B	A
$\text{TX}_{RESPQ}16$	A	B	C
$\text{TX}_{RESPQ}17$	A	C	A
$\text{TX}_{RESPQ}18$	A	C	.
$\text{TX}_{RESPQ}19$	A	A	A
$\text{TX}_{RESPQ}20a$	C	B	B
$\text{TX}_{RESPQ}20b$	B	D	*
$\text{TX}_{RESPQ}20c$	D	D	A
$\text{TX}_{RESPQ}20d$	D	D	A
$\text{TX}_{RESPQ}20e$	B	C	D

Tabela 4.1. Tabela de saída da função *head* do *DataFrame* após preparação anterior dos dados. O Dicionário para a referência das variáveis pode ser encontrado no Anexo A e para referência das alternativas, no Anexo B. Fonte: Autoria Própria.

O conjunto de dados utilizado neste estudo abrange 31 questões relacionadas a aspectos socioeconômicos e culturais, além da proficiência dos alunos em matemática. A Tabela 4.2 apresenta as variáveis independentes com suas definições formais, classificadas pela natureza das perguntas. A previsão será realizada exclusivamente para a proficiência em matemática, que é a única variável dependente deste estudo.

Visões	**Variáveis Independentes**
Localização	Área
	Local
	Região
Socioeconômico	INSE
Social	Sexo
	Idade
	Cor ou raça
	Necessidade especial
Escolaridade dos pais	Escolaridade da mãe
	Escolaridade do pai
Envolvimento dos pais na educação	Pais leem em casa
	Pais conversam sobre o que acontece na escola
	Pais incentivam a estudar
	Pais incentivam a fazer tarefa de casa
	Pais incentivam a ir às aulas
	Pais vão a reunião na escola
Infraestrutura na rua de residência	Asfalto ou calçamento
	Água tratada
	Iluminação
Deslocamento para a escola	Tempo pra chegar à escola
	Vai para a escola de (Maior distância)
	Utiliza transporte ou passe escolar
Trajetória escolar	Idade que entrou na escola
	A partir do 1°, tipo de escola que estudou
	Já foi reprovado
	Já abandonou a escola até o final do ano
Tempo em atividades diárias	Estudar, quando fora da escola
	Fazer cursos, quando fora da escola
	Trabalhar em casa, quando fora da escola
	Trabalhar fora de casa, quando fora da escola
	Lazer, quando fora da escola

Tabela 4.2. Variáveis independentes a serem estudadas. Fonte: Autoria Própria com base nos dados do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira [2023].

4.2.2 Classificação da Proficiência

Após a preparação dos dados, a proficiência dos alunos em matemática — inicialmente representada por valores contínuos fornecidos pelo SAEB, conforme destacado na Tabela 4.1 — foi convertida em categorias interpretáveis. Essa etapa tem o objetivo de transformar as notas brutas em faixas de desempenho, facilitando a aplicação de métodos de classificação no modelo preditivo e permitindo análises mais direcionadas sobre os fatores que influenciam o aprendizado.

A classificação agrupa os alunos em faixas que refletem seus níveis de aprendizado, possibilitando identificar os fatores socioeconômicos e culturais que impactam cada grupo de desempenho. Neste estudo, adota-se a classificação adaptada da proposta pelo site QEdu [QEdu [2024]], baseada nos cálculos do INEP, que organiza os níveis de proficiência em categorias pedagógicas. A Figura 4.3 ilustra essas faixas, utilizadas como referência para as análises subsequentes, enquanto a Tabela 4.3 apresenta a classificação adaptada binária adotada, que simplifica a análise ao concentrar os alunos em dois grandes grupos: "Proficiente" e "Insuficiente".

Figura 4.3. Classificação da proficiência dos alunos participantes do SAEB em Matemática, no 9º Ano do Ensino Fundamental, Brasil. Fonte: QEdu [2024] com base nos dados do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira [2023].

Essa redução de classes oferece vantagens, pois simplifica a interpretação dos resultados e contribui para um desempenho e estabilidade maiores dos modelos preditivos. Com menos categorias, há uma redução na granularidade e no impacto de ruído associado a classes de menor representatividade, o que facilita a identificação de padrões gerais e tendências amplas, essenciais para a formulação de estratégias e intervenções

Classificação que será utilizada	Classificação Base do QEdu [2024]
Insuficiente	Insuficiente
	Básico
Proficiente	Proficiente
	Avançado

Tabela 4.3. Classificação adaptada que será utilizada neste estudo. Fonte: Autória Própria.

educacionais.

Capítulo 5

Resultados

Neste trabalho, buscamos identificar os fatores que podem impactar o desempenho escolar com base em dados do SAEB e avaliar modelos preditivos em diferentes cenários de balanceamento e ajustes de hiperparâmetros. Adotamos a proposta de classificação apresentada na Tabela 4.3 (Seção 4.2.2), utilizando uma abordagem binária. As variáveis socioeconômicas e culturais foram consideradas, e a variável resposta foi definida como a proficiência classificada.

Nas Seções 5.1 e 5.3, para padronizar o processamento e garantir a consistência do tratamento das categorias, todas as questões foram convertidas em variáveis binárias (*dummificação*). Respostas como "Não Sei" e "Outros" foram preservadas para minimizar perdas de dados e manter a representatividade do conjunto, mesmo com a introdução de algum ruído.

5.1 Seleção de Características

Foram analisadas as 31 questões disponíveis (conforme descritas na Tabela 4.2), buscando um equilíbrio entre a redução da dimensionalidade e a preservação da qualidade da informação. Utilizamos o método CART, executado no *Minitab*, para classificar as variáveis por ordem de importância (Figura 5.1).

As questões incluíram alternativas numéricas e alfabéticas, conforme descrito no Anexo B, e apresentaram valores como "." e "*", indicando dados ausentes ou nulos, visíveis no *head* da Tabela 4.1. A seleção levou em conta a perda de dados associada à exclusão desses registros. Quando todas as 31 questões foram incluídas, a perda de dados foi de 31%, enquanto ao considerar menos de 20 questões, essa perda foi reduzida para menos de 20%.

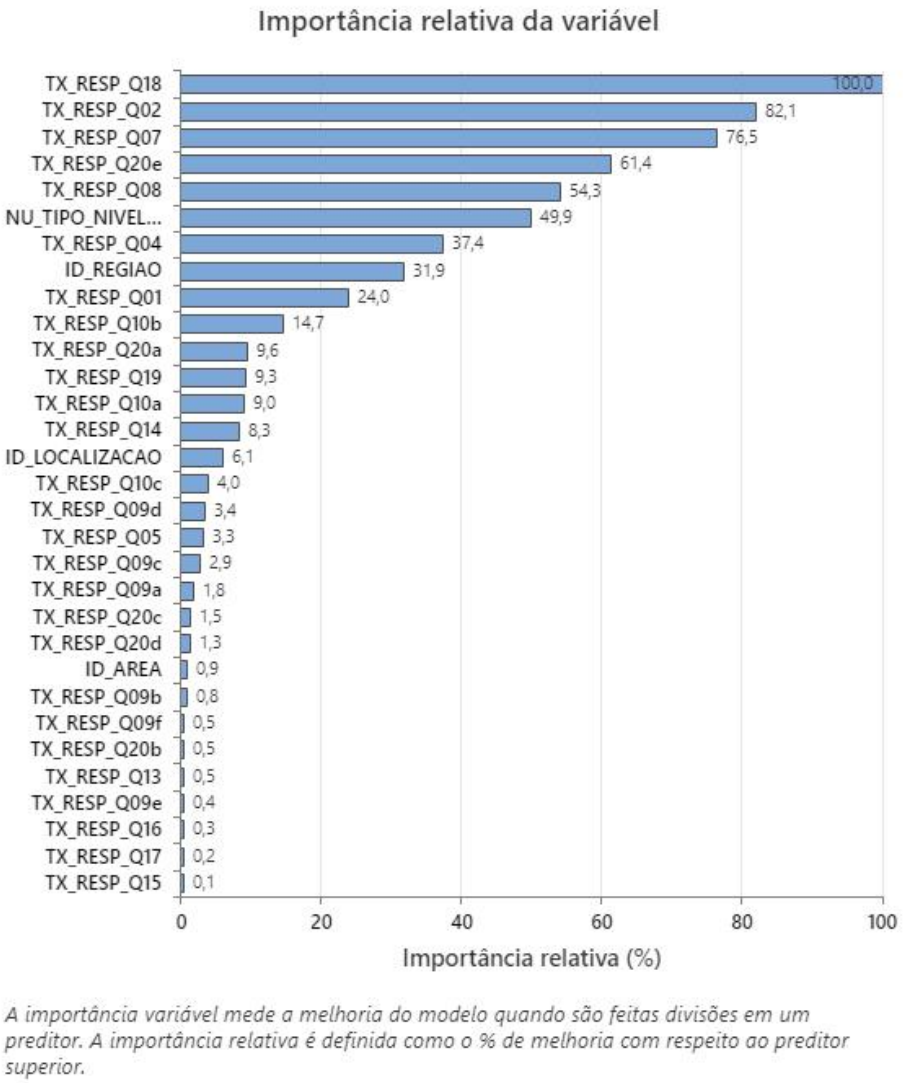

Figura 5.1. Gráfico das questões socioeconômicas e culturais ordenadas pela importância das variáveis: A relevância relativa das questões foi determinada utilizando o método de classificação CART no *software Minitab*. O dicionário das questões pode ser encontrado no Anexo A. Fonte: Autoria Própria.

A Tabela 5.1 mostra a porcentagem de perda de dados associada à exclusão de registros nulos, com a quantidade de questões selecionadas variando conforme a importância identificada pelo método CART. A perda aumentou com o número de questões, alcançando 31,23% ao considerar todas as 31 questões. Com 15 questões, a perda foi reduzida para 19,19%, tornando esse conjunto ideal para o estudo, mantendo o equilíbrio entre a perda de dados e dados mais "limpos".

Abordagem	Porcentagem de Perda %
5 Questões	10,01
7 Questões	11,73
9 Questões	12,03
11 Questões	16,03
13 Questões	17,17
15 Questões	19,19
17 Questões	20,84
31 Questões	31,23

Tabela 5.1. Tabela de resultados da perda de dados com diferentes quantidades de questões. Fonte: Autoria Própria.

A escolha das 15 questões mais relevantes representa um equilíbrio entre a preservação dos dados e a inclusão das variáveis significativas, minimizando vieses e simplificando o modelo preditivo. A Tabela 5.2 apresenta as 15 questões selecionadas, que serão utilizadas nas análises subsequentes e na modelagem preditiva.

Questão	Variável	Importância (%)
TX_RESP_Q18	Já foi reprovado	100
TX_RESP_Q02	Idade	82,1
TX_RESP_Q07	Escolaridade da Mãe	76,5
TX_RESP_Q20e	Lazer, quando fora da escola	61,4
TX_RESP_Q08	Escolaridade do Pai	54,3
NU_TIPO_NIVEL_INSE	Nível Econômico	49,9
TX_RESP_Q04	Cor ou Raça	37,4
ID_REGIAO	Região	31,9
TX_RESP_Q01	Sexo	24
TX_RESP_Q10b	Água Tratada	14,7
TX_RESP_Q20a	Estudar, quando fora da escola	9,6
TX_RESP_Q19	Já abandonou a escola até o final do ano	9,3
TX_RESP_Q10a	Asfalto ou Calçamento	9
TX_RESP_Q14	Vai para a Escola de (Maior Distância)	8,3
ID_LOCALIZACAO	Local	6,1
Perda após remoção de registros com alternativas com (".e "*"): 19,19%		

Tabela 5.2. Tabela das questões socioeconômicas e culturais selecionadas, ordenadas pela importância das variáveis. Fonte: Adaptada da Figura 5.1.

5.2 Análise Exploratória

O objetivo dessa fase foi explorar as relações entre fatores socioeconômicos e culturais e o desempenho escolar dos alunos, comparando as respostas para as variáveis categóricas

nos grupos de desempenho (“Proficiente” e “Insuficiente”).

Nas Figuras 5.2 e 5.3, são apresentados os histogramas que ilustram a distribuição dessas variáveis. Cada gráfico mostra a quantidade de alunos que escolheu cada alternativa, segmentada pela classificação do desempenho. As frequências absolutas estão no eixo y, enquanto as porcentagens em relação ao total são indicadas acima das barras, sendo exibidas apenas para valores superiores a 1%.

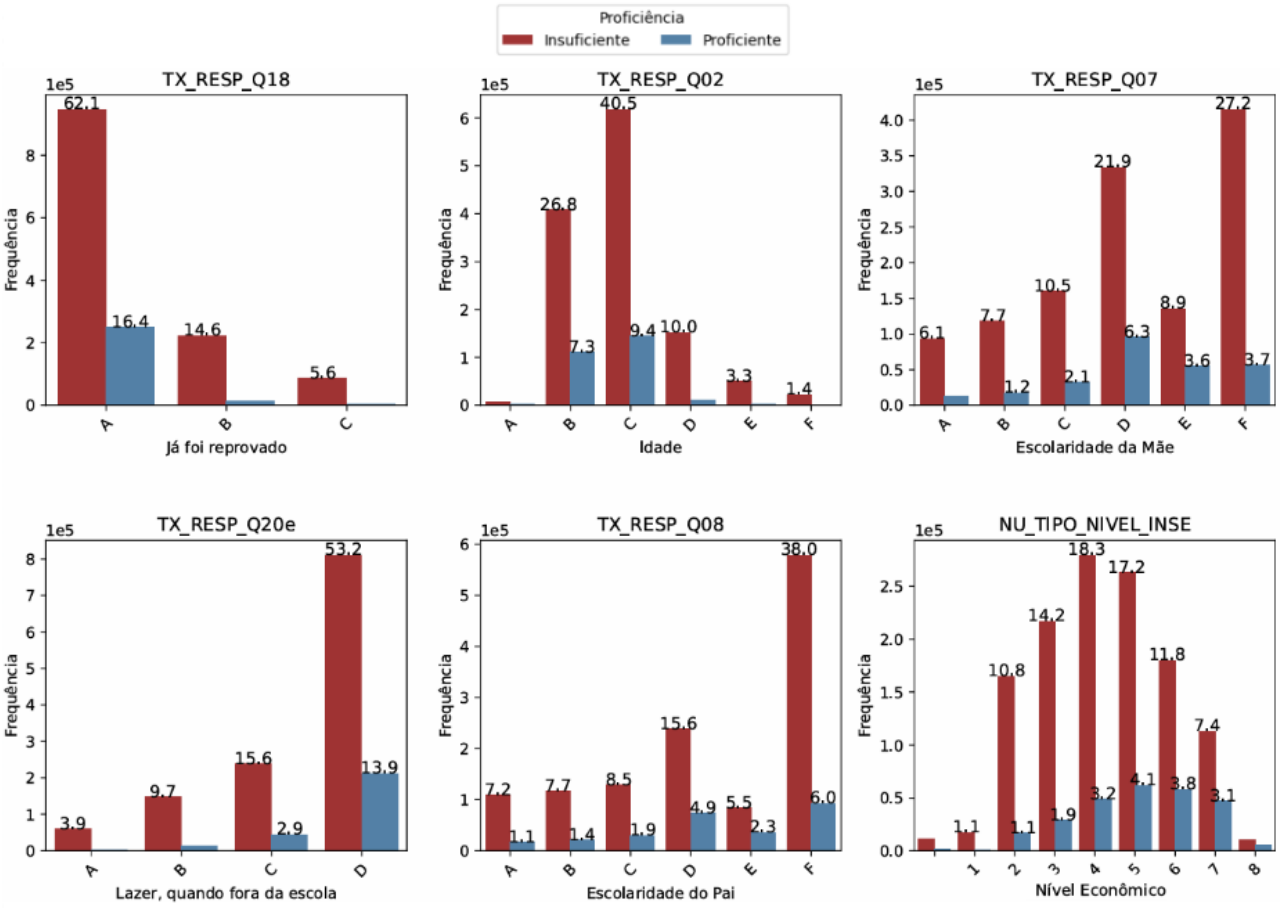

Figura 5.2. Histograma da classificação dos alunos de acordo sua proficiência para algumas questões socioecônomicas e culturais. Os números acima das barras, representam a porcentagem do total de alunos. O dicionário das questões para referência pode ser encontrado no Anexo B. Fonte: Autoria Própria.

Mais de 80% dos alunos de escolas públicas que realizaram o SAEB, em matemática, estão classificados como “Insuficiente”, o que resulta em uma predominância dessa classe nos gráficos. No entanto, padrões relevantes podem ser identificados. Observa-se que a maior parte dos alunos reside em áreas urbanas, mas a proporção de “Proficiente” é ligeiramente maior nessas áreas. A idade também mostra um padrão claro: alunos mais jovens (até 15 anos) apresentam maior proporção de “Proficiente”, enquanto os mais velhos (16 anos ou mais) predominam na classe “Insuficiente”.

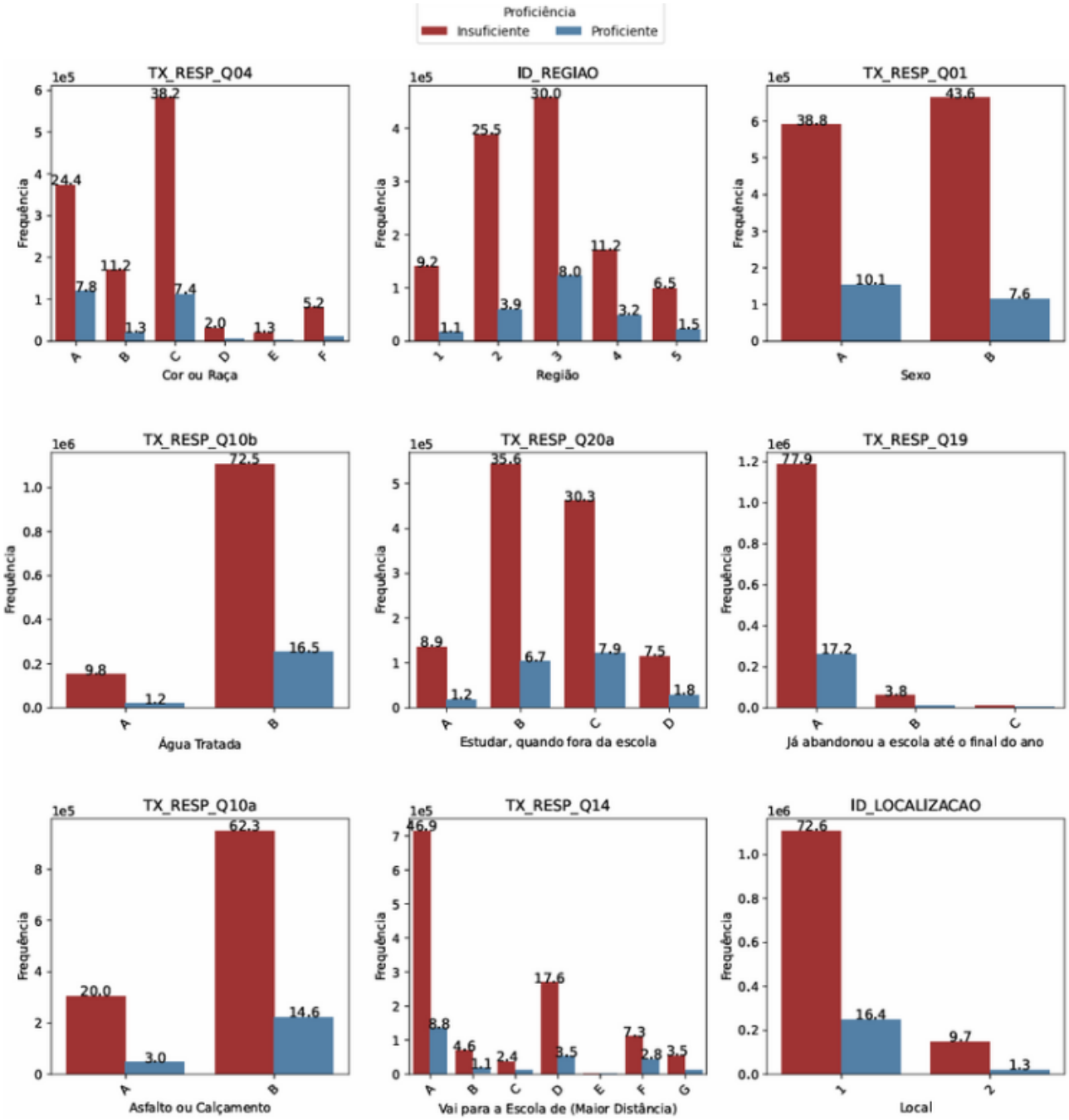

Figura 5.3. Histograma da classificação dos alunos de acordo sua proficiência para algumas questões socioecônomicas e culturais. Os números acima das barras, representam a porcentagem do total de alunos. O dicionário das questões para referência pode ser encontrado no Anexo B. Fonte: Autoria Própria.

Quanto à escolaridade dos pais, alunos cujos pais possuem ensino superior têm maior proporção de "Proficiente". Embora a classe "Insuficiente" seja majoritária em todos os gráficos, isso apenas indica uma maior chance de ser "Proficiente" para aqueles com determinados fatores, como escolaridade mais alta.

Fatores como o histórico escolar também revelam tendências importantes: alunos nunca reprovados apresentam uma maior proporção de proficiência, enquanto múltiplas reprovações estão associadas à insuficiência. Ao analisarmos a Figura 5.2, observamos que, na ausência de reprovação, a cada 5% de alunos classificados como "Insuficiente", há em média 1% de alunos classificados como "Proficiente". Em contrapartida, quando há múltiplas reprovações, a cada 62% de alunos "Insuficiente", há 16% de alunos "Proficiente". O tempo dedicado ao lazer também segue um padrão: alunos "Insuficiente" tendem a passar menos de 2 horas diárias nessa atividade, enquanto, aqueles com maior chance de ser "Proficiente" dedicam mais tempo. Níveis econômicos mais altos estão associados a melhores desempenhos, embora com uma menor quantidade de alunos nos níveis mais altos.

5.3 Modelos Preditivos

Devido à predominância da classe "Insuficiente" no conjunto de dados, foram testadas três abordagens de balanceamento aplicadas à Regressão Logística: sem balanceamento, *undersampling* com o *RandomUnderSample* e *oversampling* com o SMOTE. Na Tabela 5.3 são apresentadas as métricas (acurácia, precisão, *recall* e *F1-Score*) de cada abordagem.

Abordagem		**Precisão**	**Recall**	**F1-Score**	**Acurácia**
Sem Balanceamento	**Proficiente**	51	3	6	82,32
	Insuficiente	83	99	90	
Undersampling	**Proficiente**	29	72	41	63,99
	Insuficiente	91	62	74	
Oversampling	**Proficiente**	28	64	39	65
	Insuficiente	89	65	75	

Tabela 5.3. Resultados da RL para distintas abordagens de balanceamento. Os resultados são apresentados em porcentagem. Fonte: Autória Própria.

Embora o modelo sem balanceamento tenha atingido uma acurácia de 82,32%, as métricas para a classe minoritária "Proficiente" foram muito baixas – com *recall* de apenas 3% e *F1-Score* de 6% –, evidenciando a dificuldade do modelo em identificar corretamente essa classe. Em contraste, as abordagens com *undersampling* e *oversampling* melhoraram o equilíbrio entre as classes, sendo que a técnica de *oversampling* foi selecionada para as etapas subsequentes devido à sua leve vantagem de desempenho e à preservação de um número maior de instâncias.

Na Tabela 5.4 são apresentados os resultados dos experimentos realizados com a biblioteca *scikit-learn*, que incluiu os seguintes cenários:

- (i) Regressão Logística (RL) com parâmetros padrão (*baseline*);
- (ii) *Random Forest* (RF) com hiperparâmetros padrão;
- (iii) RL otimizada por *Grid Search*; e
- (iv) RF otimizada por *Grid Search*.

Abordagem		Precisão	Recall	F1-Score	Acurácia
i	**Proficiente**	28	64	39	65
	Insuficiente	89	65	75	
ii	**Proficiente**	28	53	37	67,37
	Insuficiente	87	70	78	
iii	**Proficiente**	28	64	39	65,01
	Insuficiente	89	65	75	
iv	**Proficiente**	27	74	39	59,56
	Insuficiente	91	56	70	

Tabela 5.4. Resultados da regressão logística para distintas abordagens. Os resultados são apresentados em porcentagem. Fonte: Autória Própria.

Os resultados mostraram que o modelo de RF com hiperparâmetros padrão teve a maior acurácia geral (67,37%), seguido pela RL e pela versão otimizada com *Grid Search* (cerca de 65%). O modelo de RF otimizado teve a menor acurácia (59,56%). Em termos de *F1-Score*, a RL apresentou desempenho mais equilibrado entre as classes: 39% para a classe "Proficiente" e 75% para a classe Insuficiente. O modelo de RF otimizado teve o maior recall para a classe "Proficiente"(74%), mas uma queda no *F1-Score* da classe Insuficiente (70%).

A Tabela 5.5 detalha os hiperparâmetros testados durante o *Grid Search*, incluindo valores de regularização (C) e penalidade para a RL e parâmetros para o RF.

A Tabela 5.6 apresenta as matrizes de confusão, revelando que o RF com hiperparâmetros padrão teve a melhor taxa de acerto para a classe "Insuficiente" (70,47%), enquanto o RF otimizado teve a melhor taxa de acerto para a classe "Proficiente" (74,44%). No entanto, o modelo otimizado apresentou queda no desempenho da classe "Insuficiente", sugerindo possível desequilíbrio após o ajuste dos hiperparâmetros.

Para a interpretação dos modelos preditivos, apresentamos na Tabela 5.7 os *odds ratios* e os *p-values*, com intervalo de confiança de 95%, de cada variável incluída na

	Hiperparâmetros Testados	Melhores Hiperparâmetros
Grid Search	C: [0.01, 0.1, 1, 10] penalty: ['l1','l2'] solver: ['liblinear','lbfgs']	C: [0.1] penalty: ['l2'] solver: ['liblinear']
Random Forest	n_estimators: [10, 50, 100, 200] max_depth: [None, 5, 10, 20] min_samples_slpit: [2, 5, 10] min_samples_leaf: [1, 2, 4] bootstrap: [True, False] max_features: ['sqrt', 'log2']	n_estimators: [50] max_depth: [5] min_samples_slpit: [2] min_samples_leaf: [1] bootstrap: [True] max_features: ['sqrt', 'log2']

Tabela 5.5. Hiperparâmetros utilizados nos modelos preditivos. Fonte: Autória Própria.

		Insuficiente	Proficiente
i	**Insuficiente**	65,14	34,86
	Proficiente	35,62	64,38
ii	**Insuficiente**	70,47	29,53
	Proficiente	47	53
iii	**Insuficiente**	65,15	34,85
	Proficiente	35,63	64,37
iv	**Insuficiente**	56,37	43,63
	Proficiente	25,56	74,44

Tabela 5.6. Matriz de confusão para as diferentes abordagens deste estudo.(i) RL com parâmetros padrão; (ii) RF com hiperparâmetros padrão; (iii) RL otimizada por Grid Search; e (iv) RF otimizada por Grid Search. Fonte: Autória Própria.

abordagem (i). Essa tabela, portanto, permite identificar quais variáveis têm impacto relevante no desempenho escolar.

Observa-se que a constante apresenta um *odds ratio* elevado (168,50), indicando um forte efeito base no modelo. Entre os preditores, variáveis com *odds ratios* superiores a 1, como lazer fora da escola por mais de 2 horas (*TX_RESP_Q20e_D* – Odds = 1,3458) e estudo extraclasse entre 1 e 2 horas (*TX_RESP_Q20a_C* – Odds = 1,3325; *TX_RESP_Q20a_D* – Odds = 1,2207), estão positivamente associadas à proficiência. Isso sugere que alunos que dedicam mais tempo ao estudo ou ao lazer fora da escola têm maior probabilidade de alcançar um desempenho satisfatório.

Por outro lado, variáveis com *odds ratios* inferiores a 1 indicam uma relação negativa com a proficiência. Destaca-se que ter 18 anos ou mais (*TX_RESP_Q02_F* – Odds = 0,0781) e pertencer ao nível socioeconômico mais baixo (*NU_TIPO_NIVEL_INSE_1* – Odds = 0,0992) reduzem significativamente a chance de um aluno ser classificado como "Proficiente". Esses achados reforçam a influência da idade e das condições socioeconômicas no desempenho escolar.

A maioria dos preditores apresenta *p-values* extremamente baixos ($p < 0{,}005$), indicando forte significância estatística das associações encontradas. No entanto, algumas variáveis, como lazer entre 1 e 2 horas diárias ($TX\_RESP\_Q20e\_C$ – $p = 0{,}1234$), estudo extraclasse por menos de 1 hora ($TX\_RESP\_Q20a\_B$ – $p = 0{,}0109$) e a região geográfica do aluno (ID_REGIAO – $p = 0{,}00199$), não atingiram o nível usual de significância, sugerindo que seu impacto sobre a proficiência pode ser menos relevante ou sujeito a variabilidade amostral.

Em resumo, os modelos preditivos desenvolvidos demonstraram que a combinação de técnicas de balanceamento e otimização por *Grid Search* aprimora a identificação da classe minoritária "Proficiente", evidenciando a influência de variáveis socioeconômicas e comportamentais sobre o desempenho dos alunos. Esses resultados corroboram os achados de Gomes et al. [2020] e Melo et al. [2021], que também apontam para a relevância de fatores como tempo dedicado ao estudo e condições socioeconômicas na predição do desempenho escolar. Ademais, a análise dos *odds ratios*, com destaque para o efeito positivo de atividades extracurriculares e o efeito negativo de variáveis como idade elevada e baixo nível socioeconômico, alinha-se com os resultados de Capelato & Périco [2023], que enfatizam a importância de aspectos contextuais e estruturais na eficiência escolar. Por outro lado, a abordagem descritiva de Santos, A [2018] e as dificuldades apontadas por Karakolidis et al. [2016] para replicar certos determinantes em contextos distintos reforçam a necessidade de uma análise abrangente e adaptada às especificidades dos dados locais.

Variável	Odds Ratio	p-valor
const	168.496593	0.000000e+00
TX_RESP_Q20e_D	1.345827	7.135990e-203
TX_RESP_Q20a_C	1.332494	0.000000e+00
TX_RESP_Q20a_D	1.220672	1.095888e-153
TX_RESP_Q10b_B	1.210153	1.981996e-221
ID_REGIAO	0.994665	1.996863e-03
TX_RESP_Q20a_B	0.985125	1.096252e-02
TX_RESP_Q20e_C	0.984269	1.234054e-01
TX_RESP_Q14_F	0.918757	1.296736e-54
TX_RESP_Q10a_B	0.907463	3.320647e-103
TX_RESP_Q07_E	0.873213	1.681576e-64
TX_RESP_Q14_D	0.808156	0.000000e+00
TX_RESP_Q07_D	0.798336	1.481745e-221
TX_RESP_Q08_E	0.761524	5.027819e-239
TX_RESP_Q08_D	0.722964	0.000000e+00
TX_RESP_Q14_C	0.650782	0.000000e+00
TX_RESP_Q07_C	0.642334	0.000000e+00
TX_RESP_Q08_F	0.624840	0.000000e+00
TX_RESP_Q08_C	0.597787	0.000000e+00
TX_RESP_Q04_C	0.587186	0.000000e+00
TX_RESP_Q14_G	0.575823	0.000000e+00
ID_LOCALIZACAO	0.569130	0.000000e+00
TX_RESP_Q14_B	0.549792	0.000000e+00
TX_RESP_Q08_B	0.544747	0.000000e+00
TX_RESP_Q07_B	0.534211	0.000000e+00
TX_RESP_Q07_F	0.493694	0.000000e+00
TX_RESP_Q01_B	0.467720	0.000000e+00
TX_RESP_Q19_B	0.465126	0.000000e+00
TX_RESP_Q20e_B	0.462698	0.000000e+00
TX_RESP_Q19_C	0.344629	4.863923e-185
TX_RESP_Q04_F	0.306601	0.000000e+00
TX_RESP_Q14_E	0.304255	2.936245e-75
TX_RESP_Q04_B	0.289046	0.000000e+00
NU_TIPO_NIVEL_INSE_7	0.266798	0.000000e+00
NU_TIPO_NIVEL_INSE_6	0.244867	0.000000e+00
TX_RESP_Q02_C	0.241940	0.000000e+00
NU_TIPO_NIVEL_INSE_8	0.239635	0.000000e+00
TX_RESP_Q04_D	0.235084	0.000000e+00
TX_RESP_Q02_B	0.233572	0.000000e+00
TX_RESP_Q18_C	0.218830	0.000000e+00
TX_RESP_Q02_D	0.212044	0.000000e+00
NU_TIPO_NIVEL_INSE_5	0.211624	0.000000e+00
TX_RESP_Q18_B	0.202386	0.000000e+00
NU_TIPO_NIVEL_INSE_4	0.185087	0.000000e+00
NU_TIPO_NIVEL_INSE_3	0.165475	0.000000e+00
NU_TIPO_NIVEL_INSE_2	0.152666	0.000000e+00
TX_RESP_Q04_E	0.135631	0.000000e+00
TX_RESP_Q02_E	0.124084	0.000000e+00
NU_TIPO_NIVEL_INSE_1	0.099168	0.000000e+00
TX_RESP_Q02_F	0.078119	0.000000e+00

Tabela 5.7. *Odds Ratio* e p-valor do modelo de regressão de logística ordenados por *Odds Ratio* em ordem decrescente. O Dicionário para a referência das variáveis e alternativas podem ser encontrados no Anexos A e B. Fonte: Autoria Própria.

Capítulo 6

Considerações Finais

Este estudo visou identificar os fatores socioeconômicos e culturais que podem influenciam o desempenho dos alunos do 9º ano em Matemática, utilizando dados do SAEB 2021 e modelos preditivos ajustados com técnicas de aprendizado de máquina. Foram realizadas etapas de preparação dos dados, seleção de características e aplicação de modelos preditivos (Regressão Logística e *Random Forest*), com balanceamento dos dados via SMOTE para mitigar o desbalanceamento das classes. Os principais achados indicam que variáveis como lazer fora da escola por mais de 2 horas e estudo extraclasse estão positivamente associadas à proficiência, enquanto fatores como idade elevada e baixa condição socioeconômica reduzem a chance de alcançar um desempenho satisfatório. A análise dos *odds ratios* e *p-values* reforça a significância estatística dos preditores, permitindo identificar aqueles que mais impactam a classificação dos alunos. Esses resultados sugerem que intervenções direcionadas — como o incentivo a atividades extraclasse e políticas de apoio às famílias em situação de vulnerabilidade — podem contribuir para a melhoria dos índices de proficiência. Entretanto, a transformação das notas em categorias binárias, embora facilite a análise, pode simplificar demais a realidade, e o uso de SMOTE, embora necessário, pode introduzir alguns vieses. Em síntese, o estudo pode fornecer importantes subsídios para a compreensão dos determinantes do desempenho escolar em Matemática, bem como, pode apontar caminhos para a implementação de políticas educacionais mais eficazes, promovendo a equidade e a qualidade do ensino no Brasil. Espera-se que os achados sirvam de base para comparações ou um norte para futuras pesquisas sobre o desempenho dos alunos em matemáticae e intervenções no setor educacional.

Referências Bibliográficas

Andrew Bruce & Peter Bruce (2019). *Estatística Prática para Cientistas de Dados: 50 Conceitos Essenciais*. Alta Books, 1ª edição.

Brasil (1988). Art. 205. https://portal.stf.jus.br/constituicao-supremo/artigo.asp?abrirBase=CF&abrirArtigo=205#:~:text=Da%20EducaÃğÃčo-,Art.,sua%20qualificaÃğÃčo%20para%20o%20trabalho.&text=Lei%20nÂž%2014.172%2C%20de%2010%20de%20junho%20de%202021. Accessed: 14-03-2024.

Brasil (2014). Lei nº 13.005, de 25 de junho de 2014. https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2014/lei/l13005.htm. Accessed: 14-03-2024.

Brasil (2017). Lei nº 13.415 de 16 de fevereiro de 2017. https://legislacao.presidencia.gov.br/atos/tipo=LEI&numero=13415&ano=2017&ato=115MzZE5EeZpWT9be. Accessed: 14-03-2024.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Breiman, L.; Friedman, J.; Stone, C. & Olshen, R. (1984). *Classification and Regression Trees*. Taylor & Francis.

Capelato, & Périco, A. E. (2023). Eficiência escolar no estado de São Paulo: fatores interescolares e desempenho em matemática. *Amazônia: Revista de Educação em Ciências e Matemáticas*, 19.

Chawla, N. V.; Bowyer, K. W.; Hall, L. O. & Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

Direitoria de Avaliação da Educação Básica (2020). Escalas de Proficiência do SAEB. https://download.inep.gov.br/publicacoes/institucionais/avaliacoes_e_exames_da_educacao_basica/escalas_de_proficiencia_do_saeb.pdf. Accessed: 14-03-2024.

Direitoria de Avaliação da Educação Básica (2023). Relatório de Resultados do SAEB 2021 | Volume 1. https://download.inep.gov.br/educacao_basica/saeb/2021/resultados/relatorio_de_resultados_do_saeb_2021_volume_1.pdf. Accessed: 14-03-2024.

Diretoria de Avaliação da Educação Básica (2023). Relatório de Amostragem do SAEB 2021. https://download.inep.gov.br/publicacoes/institucionais/avaliacoes_e_exames_da_educacao_basica/relatorio_de_amostragem_saeb_2021.pdf. Accessed: 14-03-2024.

Diretoria de Gestão e Planejamento (2021). Cartilha SAEB 2021. https://download.inep.gov.br/publicacoes/institucionais/avaliacoes_e_exames_da_educacao_basica/cartilha_saeb_2021.pdf. Accessed: 14-03-2024.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874.

Fávero, L. P. & Belfiore, P. (2017). *Manual de Análise de Dados: Estatística e Modelagem Multivariada com Excel®, SPSS® e Stata®*. GEN LTC, Rio de Janeiro, 1ª edição.

Gomes, C. M. A.; Fleith, D. d. S.; Marinho-Araujo, C. M. & Rabelo, M. L. (2020). Predictors of Students' Mathematics Achievement in Secondary Education. *Psicologia: Teoria e Pesquisa*, v.36.

Google (2023). Google colaboratory documentation. https://colab.research.google.com/notebooks/welcome.ipynb?hl=pt-BR. Accessed: 24-08-2024.

Hasib, K. M.; Iqbal, M. S.; Shah, F. M.; Mahmud, J. A.; Popel, M. H.; Showrov, M. I. H.; Ahmed, S. & Rahman, O. (2020). A survey of methods for managing the classification and solution of data imbalance problem. *Journal of Computer Science*, 16(11):1546–1557. Accessed: 13-02-2025.

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (2021). Microdados do SAEB - Leia-me. https://download.inep.gov.br/publicacoes/institucionais/avaliacoes_e_exames_da_educacao_basica/cartilha_saeb_2021.pdf. Accessed: 14-03-2024.

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (2022). Nota Informativa dos Resultados do SAEB 2021 - Versão Retificada. https://download.inep.gov.br/saeb/outros_documentos/nota_explicativa_saeb_2021.pdf. Accessed: 14-03-2024.

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (2023). Microdados SAEB. https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/saeb. Accessed: 10-12-2023.

James, G.; Witten, D.; Hastie, T. & Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R*. Springer, New York.

Jupyter, P. (2015). Jupyter notebook: An open source platform for interactive computing. https://jupyter.org. Acessado em: 26/12/2024.

Karakolidis, A.; Pitsia, V. & Emvalotis, A. (2016). Examining students' achievement in mathematics: A multilevel analysis of the Programme for International Student Assessment (PISA) 2012 data for Greece. *International Journal of Educational Research*, 79:106–115.

Letícia Mori (2023). Pisa: até alunos mais ricos no Brasil estão abaixo da média global em Matemática. https://www.bbc.com/portuguese/articles/cv2zx819rg4o. BBC News Brasil. Accessed: 13-04-2024.

Melo, R. O.; Freitas, A. C. D.; Francisco, E. D. R. & Motokane, M. T. (2021). Impacto das variáveis socioeconômicas no desempenho do Enem: uma análise espacial e sociológica. *Revista de Administração Pública*, 55(6):1271–1294.

Minitab, LLC (2023). Minitab statistical software. https://www.minitab.com. Versão 21.3.

Palhares, I. (2024). Quase metade dos alunos brasileiros não termina ensino fundamental na idade certa. https://www1.folha.uol.com.br/educacao/2024/03/quase-metade-dos-alunos-brasileiros-nao-termina-ensino-fundamental-na-idade-certa.shtml. Folha de S.Paulo. Accessed: 25-03-2024.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M. & Duchesnay, E. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830.

Python (2023). Python documentation. https://www.python.org/doc/. Accessed: 24-08-2024.

QEdu (2024). Qedu - ideb brasil. Acesso em: 27 dez. 2024.

Ragsdale, C. (2017). *Spreadsheet Modeling & Decision Analysis: A Practical Introduction to Business Analytics*. Cengage Learning, Boston, MA, 8th edição.

Russell, S. & Norvig, P. (2013). *Inteligência Artificial*. Elsevier, Rio de Janeiro, 3ª edição.

Santos, A (2018). O Panorama das Notas em Matemática no SAEB dos Estudantes do 5º Ano do Ensino Fundamental no Período de 2007 a 2017. *Revista Educação Matemática em Foco*.

Soares, J. F. (2004). O efeito da escola no desempenho cognitivo de seus alunos. *REICE. Revista Iberoamericana sobre Calidad, Eficacia y Cambio en Educación*, v.2. Accessed: 13-04-2024.

Wu, X. & Kumar, V., editores (2009). *The Top Ten Algorithms in Data Mining*. CRC Press, Boca Raton, FL.

Anexo A

Dicionário de Associação entre Questões e Variáveis

Este dicionário foi elaborado como um recurso complementar para auxiliar na compreensão das nomenclaturas utilizadas nas questões socioeconômico-culturais do estudo, baseado nos dados do SAEB [Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira [2023]]. Ele fornece uma descrição de cada questão e sua descrição.

Cada entrada neste dicionário inclui a sigla da questão a ser estudada, seguida de uma descrição do seu conteúdo. As questões são separadas por temas que agregam tais conteúdos. As opções das alternativas podem ser encontradas no Anexo B.

LOCALIZAÇÃO	
Questão	**Variável(Descrição)**
ID_AREA	Área
ID_LOCALIZACAO	Local
ID_REGIAO	Região
SOCIOECONÔMICO	
Questão	**Variável(Descrição)**
NU_TIPO_NIVEL_INSE	Nível Econômico
SOCIAL	
Questão	**Variável(Descrição)**
TX_RESP_Q01	Sexo
TX_RESP_Q02	Idade
TX_RESP_Q04	Cor ou Raça
TX_RESP_Q05	Necessidade Especial
ESCOLARIDADE DOS PAIS	
Questão	**Variável(Descrição)**
TX_RESP_Q07	Escolaridade da Mãe
TX_RESP_Q08	Escolaridade do Pai
ENVOLVIMENTO DOS PAIS NA EDUCAÇÃO	
Questão	**Variável(Descrição)**
TX_RESP_Q09a	Pais leem em casa
TX_RESP_Q09b	Pais conversam sobre o que acontece na escola
TX_RESP_Q09c	Pais incentivam a estudar
TX_RESP_Q09d	Pais incentivam a fazer tarefa de casa
TX_RESP_Q09e	Pais incentivam a ir às aulas
TX_RESP_Q09f	Pais vão a reunião na escola
INFRAESTRUTURA NA RUA DE RESIDÊNCIA	
Questão	**Variável(Descrição)**
TX_RESP_Q10a	Asfalto ou Calçamento
TX_RESP_Q10b	Água Tratada
TX_RESP_Q10c	Iluminação
DESLOCAMENTO PARA A ESCOLA	
Questão	**Variável(Descrição)**
TX_RESP_Q13	Tempo para chegar à Escola
TX_RESP_Q14	Vai para a Escola de (Maior Distância)
TX_RESP_Q15	Utiliza Transporte ou Passe Escolar

TRAJETÓRIA ESCOLAR	
Questão	**Variável(Descrição)**
TX_RESP_Q16	Idade que entrou na escola
TX_RESP_Q17	A partir do 1°, tipo de escola que estudou
TX_RESP_Q18	Já foi reprovado
TX_RESP_Q19	Já abandonou a escola até o final do ano
TEMPO EM ATIVIDADES DIÁRIAS	
Questão	**Variável(Descrição)**
TX_RESP_Q20a	Estudar, quando fora da escola.
TX_RESP_Q20b	Fazer cursos, quando fora da escola.
TX_RESP_Q20c	Trabalhar em casa, quando fora da escola.
TX_RESP_Q20d	Trabalhar fora de casa, quando fora da escola
TX_RESP_Q20e	Lazer, quando fora da escola

Tabela A.1. Dicionário dos dados utilizados no estudo sobre as questões do SAEB. Fonte: Autoria Própria com base nos dados do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira [2023].

Anexo B

Dicionário de Alternativas de Respostas das Variáveis

Este dicionário foi elaborado como um recurso complementar para auxiliar na compreensão das alternativas de resposta utilizadas nas questões socioeconômico-culturais do estudo, baseado nos dados do SAEB [Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira [2023]]. Ele fornece uma descrição detalhada de cada alternativa, seja ela numérica ou alfabética, permitindo uma interpretação dos dados apresentados.

Cada entrada neste dicionário inclui a variável a ser estudada, seguida da sigla da alternativa e, posteriormente, uma explicação do seu significado ou conteúdo. É importante observar que as alternativas podem variar de acordo com o contexto da pergunta e o tipo de informação solicitada, e as questões são separadas por temas que agregam tais conteúdos.

LOCALIZAÇÃO	
Variável	**Código de Preenchimento**
Aréa	1 - Capital
	2 - Interior
Local	1 - Urbana
	2 - Rural
Região	1 - Norte
	2 - Nordeste
	3 - Sudeste
	4 - Sul
	5 - Centro-Oeste
SOCIOECONÔMICO	
Variável	**Código de Preenchimento**
Nível Econômico	1 - Nível I
	2 - Nível II
	3 - Nível III
	4 - Nível IV
	5 - Nível V
	6 - Nível VI
	7 - Nível VII
	8 - Nível VIII
SOCIAL	
Variável	**Código de Preenchimento**
Sexo	A - Masculino
	B - Feminino
Idade	A - 13 anos ou menos
	B - 14 anos
	C - 15 anos
	D - 16 anos
	E - 17 anos
	F - 18 anos ou mais
Cor ou Raça	A - Branca
	B - Preta
	C - Parda
	D - Amarela
	E - Indígena
	F - Não quero declarar
Necessidade Especial	A - Sim
	B - Não

ESCOLARIDADE DOS PAIS	
Variável	**Código de Preenchimento**
Escolaridade da Mãe	A - Não completou o 5º ano do Ensino Fundamental.
Escolaridade do Pai	B - Ensino Fundamental, até o 5º ano.
	C - Ensino Fundamental completo.
	D - Ensino Médio completo.
	E - Ensino Superior completo (faculdade ou graduação).
	F - Não sei.
ENVOLVIMENTO DOS PAIS NA EDUCAÇÃO	
Variável	**Código de Preenchimento**
Pais leem em casa	A - Nunca ou quase nunca.
Pais conversam sobre o que acontece na escola	B - De vez em quando.
Pais incentivam a estudar	C - Sempre ou quase sempre.
Pais incentivam a fazer tarefa de casa	
Pais incentivam a ir às aulas	
Pais vão a reunião na escola	
INFRAESTRUTURA NA RUA DE RESIDÊNCIA	
Variável	**Código de Preenchimento**
Asfalto ou Calçamento	A - Não
Água Tratada	B - Sim
Iluminação	
DESLOCAMENTO PARA A ESCOLA	
Variável	**Código de Preenchimento**
Tempo para chegar à Escola	A - Menos de 30 minutos.
	B - Entre 30 minutos e uma hora.
	C - Mais de uma hora.
Vai para a Escola de (Maior Distância)	A - À pé.
	B - De bicicleta.
	C - De Van (ou Kombi).
	D - De ônibus.
	E - De metrô (ou trem urbano).
	F - De carro.
	G - Outros meios de transporte (barco, motocicleta, etc.)
Utiliza Transporte ou Passe Escolar	A - Não.
	B - Sim.

TRAJETÓRIA ESCOLAR	
Variável	**Código de Preenchimento**
Idade que entrou na escola	A - 3 anos ou menos.
	B - 4 ou 5 anos.
	C - 6 ou 7 anos.
	D - 8 anos ou mais.
A partir do 1°, tipo de escola que estudou	A - Somente em escola pública.
	B - Somente em escola particular.
	C - Em escola pública e em escola particular.
Já foi reprovado	A - Não.
	B - Sim, uma vez.
	C - Sim, duas vezes ou mais.
Já abandonou a escola até o final do ano	A - Nunca.
	B - Sim, uma vez.
	C - Sim, duas vezes ou mais.
TEMPO EM ATIVIDADES DIÁRIAS	
Variável	**Código de Preenchimento**
Estudar, quando fora da escola.	A - Não uso meu tempo para isso.
Fazer cursos, quando fora da escola.	B - Menos de 1 hora.
Trabalhar em casa, quando fora da escola.	C - Entre 1 e 2 horas.
Trabalhar fora de casa, quando fora da escola	D - Mais de 2 horas.
Lazer, quando fora da escola	

Tabela B.1. Dicionário dos dados utilizados no estudo sobre variáveis utilizadas. Fonte: Autoria Própria com base nos dados do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira [2023].