

# **UNIVERSIDADE DE SÃO PAULO**

Instituto de Ciências Matemáticas e de Computação

**Título**

**Ramon de Castro Ramos**

Monografia - MBA em Ciência de Dados (CeMEAI)



**Ramon de Castro Ramos**

## **Título**

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**  
**2025**

É possível elaborar a ficha catalográfica em LaTeX ou incluir a fornecida pela Biblioteca. Para tanto observe a programação contida nos arquivos USPSC-modelo.tex e fichacatalografica.tex e/ou gere o arquivo fichacatalografica.pdf.

A biblioteca da sua Unidade lhe fornecerá um arquivo PDF com a ficha catalográfica definitiva, que deverá ser salvo como fichacatalografica.pdf no diretório do seu projeto.

## **Ramon de Castro Ramos**

### **Title**

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  
2025**

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**

Primeira frase do agradecimento ....

Segunda frase ....

Outras frases ....

Última frase ....



*“Be yourself, everyone else is already taken.”*

*Oscar Wilde*



## **RESUMO**

RAMOS, R. C. **Título.** 2025. **63** 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, 2025.

xxx

**Palavras-chave:** x. x. x. x. x. x.



## **ABSTRACT**

RAMOS, R. C. **Title.** 2025. 63 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, 2025.

xxx

**Keywords:** x. x. x. x. x. x.



## **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 – Esquema ilustrativo do funcionamento do <i>AdaBoost</i> . . . . .	39
Figura 5 – Modelo CRISP-DM . . . . .	41



## **LISTA DE TABELAS**

Tabela 1 – Variáveis socioeconômicas e suas referências . . . . .	42
Tabela 2 – Quantidade de observações e variáveis por edição do ENEM . . . . .	43
Tabela 3 – Percentual inicial de valores nulos por variável . . . . .	50
Tabela 4 – Quantidade de observações por conjunto de dados . . . . .	51



## **LISTA DE QUADROS**



## LISTA DE ABREVIATURAS E SIGLAS

AdaBoost	<i>Adaptive Boosting</i>
COVID-19	<i>Coronavirus Disease 2019</i> - Doença do Coronavírus 2019
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
ENEM	Exame Nacional do Ensino Médio
Fies	Fundo de Financiamento Estudantil
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
ML	<i>Machine Learning</i> - Aprendizado de Máquina
ProUni	Programa Universidade Para Todos
RF	<i>Random Forest</i> - Floresta Aleatória
SISU	Sistema de Seleção Unificada
UFABC	Universidade Federal do ABC
XGBoost	<i>Extreme Gradient Boosting</i>



## **LISTA DE SÍMBOLOS**

$\alpha$	<i>Alpha</i> - Primeiro caractere do alfabeto grego
$\beta$	<i>Beta</i> - Segundo caractere do alfabeto grego
$\epsilon$	<i>Epsilon</i> - Quinto caractere do alfabeto grego
$\leq$	<i>Menor ou igual a</i>



## SUMÁRIO

<b>1</b>	<b>INTRODUÇÃO</b>	<b>29</b>
<b>2</b>	<b>FUNDAMENTAÇÃO TEÓRICA</b>	<b>31</b>
<b>2.1</b>	<b>O ENEM no Cenário Educacional Brasileiro</b>	<b>31</b>
<b>2.2</b>	<b>Teorias sobre Desigualdades Educacionais: O Capital Cultural de Bourdieu</b>	<b>32</b>
<b>2.3</b>	<b>Fatores Socioeconômicos e Desempenho no ENEM</b>	<b>32</b>
<b>2.4</b>	<b>Características escolares e o “Efeito Escola”</b>	<b>33</b>
<b>2.5</b>	<b>Disparidades Regionais e a Participação no ENEM</b>	<b>34</b>
<b>2.6</b>	<b>Aplicações de Ciência de Dados na Análise do ENEM e resultados obtidos</b>	<b>34</b>
<b>2.7</b>	<b>Métodos de <i>Machine Learning</i></b>	<b>35</b>
<b>2.7.1</b>	<b>Régressão Linear</b>	<b>35</b>
<b>2.7.2</b>	<b>Árvore de Decisão</b>	<b>36</b>
<b>2.7.3</b>	<b>Random Forest</b>	<b>37</b>
<b>2.7.4</b>	<b>Boosting</b>	<b>38</b>
<b>3</b>	<b>METODOLOGIA</b>	<b>41</b>
<b>3.1</b>	<b>Entendimento de Negócio</b>	<b>41</b>
<b>3.2</b>	<b>Entendimento dos Dados</b>	<b>42</b>
<b>3.2.1</b>	<b>Estrutura dos Dados</b>	<b>43</b>
<b>3.2.2</b>	<b>Definição da Variável Resposta</b>	<b>44</b>
<b>3.3</b>	<b>Preparação do Ambiente Python</b>	<b>44</b>
<b>3.4</b>	<b>Preparação dos Dados</b>	<b>44</b>
<b>3.4.1</b>	<b>Coleta, Leitura e Integração dos Dados</b>	<b>45</b>
<b>3.4.2</b>	<b>Transformação das Variáveis Categóricas</b>	<b>45</b>
<b>3.4.3</b>	<b>Tratamentos dos Valores Nulos</b>	<b>45</b>
<b>3.4.4</b>	<b>Separação dos Conjuntos de Dados por Variável Resposta</b>	<b>46</b>
<b>3.4.5</b>	<b>Exploração Inicial</b>	<b>46</b>
<b>3.5</b>	<b>Análise de correlação das Variáveis Preditoras</b>	<b>46</b>
<b>3.5.1</b>	<b>Imputação de Valores Ausentes</b>	<b>46</b>
<b>3.6</b>	<b>Modelagem</b>	<b>47</b>
<b>3.6.1</b>	<b>Seleção de Modelos</b>	<b>47</b>
<b>3.6.2</b>	<b>Otimização dos Hiperparâmetros</b>	<b>47</b>
<b>3.6.3</b>	<b>Treino dos Modelos</b>	<b>47</b>
<b>3.6.4</b>	<b>Avaliação dos Modelos</b>	<b>47</b>

3.7	<b>Magnitude da Influência das Variáveis Preditoras</b>	48
3.8	<b>Limitações e Considerações Éticas</b>	48
4	<b>RESULTADOS</b>	49
4.1	<b>Entendimento dos Dados</b>	49
4.1.1	Tratamento dos Valores nulos	49
4.1.2	Separação dos Conjuntos de Dados por Variável Resposta	49
4.1.3	Exploração Inicial	49
5	<b>CONCLUSÃO</b>	53
	<b>REFERÊNCIAS</b>	55
	<b>APÊNDICES</b>	59
	<b>APÊNDICE A – DICIONÁRIO DE DADOS DOS MICRODADOS DO ENEM</b>	61
	<b>APÊNDICE B – CONFIGURAÇÃO DO AMBIENTE VIRTUAL</b>	63

## 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, buscam-se 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 extraindo 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 irão apresentar a metodologia adotada neste trabalho, os resultados obtidos e a discussão desses resultados, culminando nas conclusões e recomendações para futuros trabalhos.

## 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 salta para 93 instituições no ano seguinte. A importância do ENEM cresce 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 assume o formato que tem hoje: 180 questões objetivas divididas em 4 áreas do conhecimento 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 adotam o ENEM como critério de seleção. Duas universidade portuguesas, a Universidade de Coimbra e Universidade de Algrave, passar a usar o ENEM como critério de seleção em 2014, número que chega a 35 instituições portuguesas em 2018 (5).

É evidente que o ENEM deixa de ser apenas uma ferramenta de avaliação e transforma-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.), 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 instucionalizado, caracterizado por diplomas e títulos de conhecimento (7).

A acumulação 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 estão 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 passa 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 intra-escolares quanto os extra-escolares.

## 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 similares 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 socioeconômico 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*

Essa seção pretende apresentar, 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

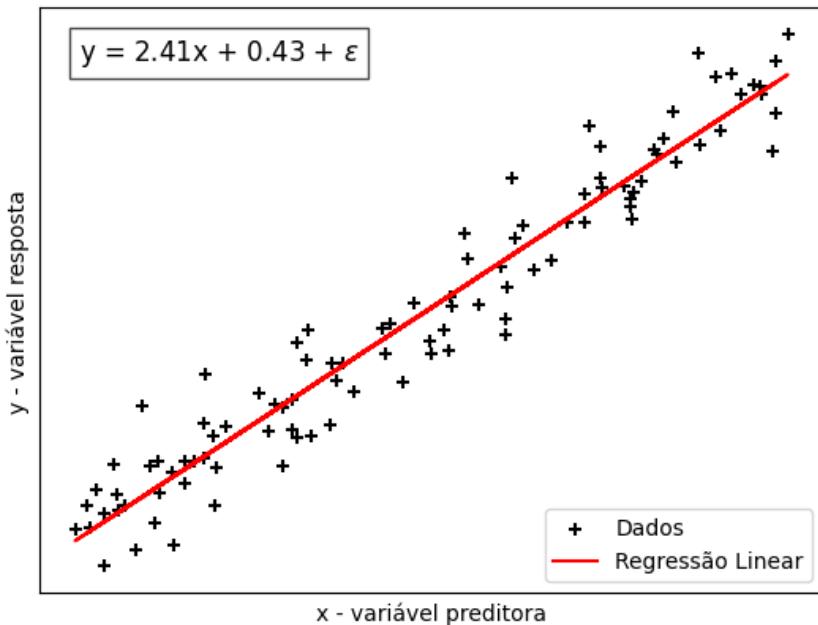
$$Y = \beta_0 + \beta_1 \times X + \epsilon \quad (2.1)$$

onde  $Y$  denota a variável resposta,  $X$  a variável preditora,  $\beta_0$  o intercepto (o valor de  $Y$  quando  $X = 0$ ),  $\beta_1$  o coeficiente angular (indicando o impacto de  $X$  sobre  $Y$ ) e  $\epsilon$  o termo de erro. Em uma regressão múltipla, diversas variáveis independentes são consideradas, cada uma com o seu  $\beta_i$  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 a 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

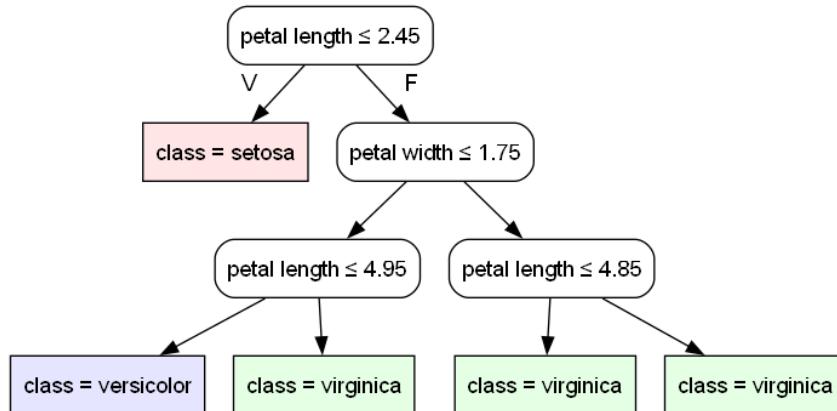
A Árvore de Decisão é um método paramétrico de aprendizado supervisionado que utiliza uma abordagem intuitiva de separação dos dados em grupos semelhantes, através de regras hierárquicas simples e de forma recursiva. Pode ser utilizado para resolver problemas de regressão, com a média da variável resposta 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 a esquerda; se não, então este segue pelo caminho a 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.

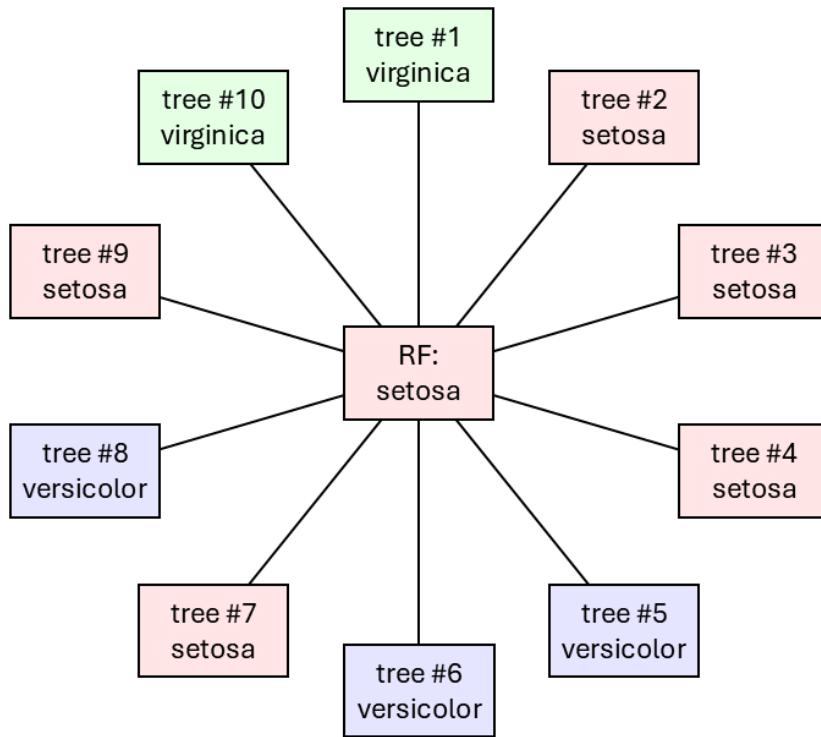
### 2.7.3 Random Forest

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

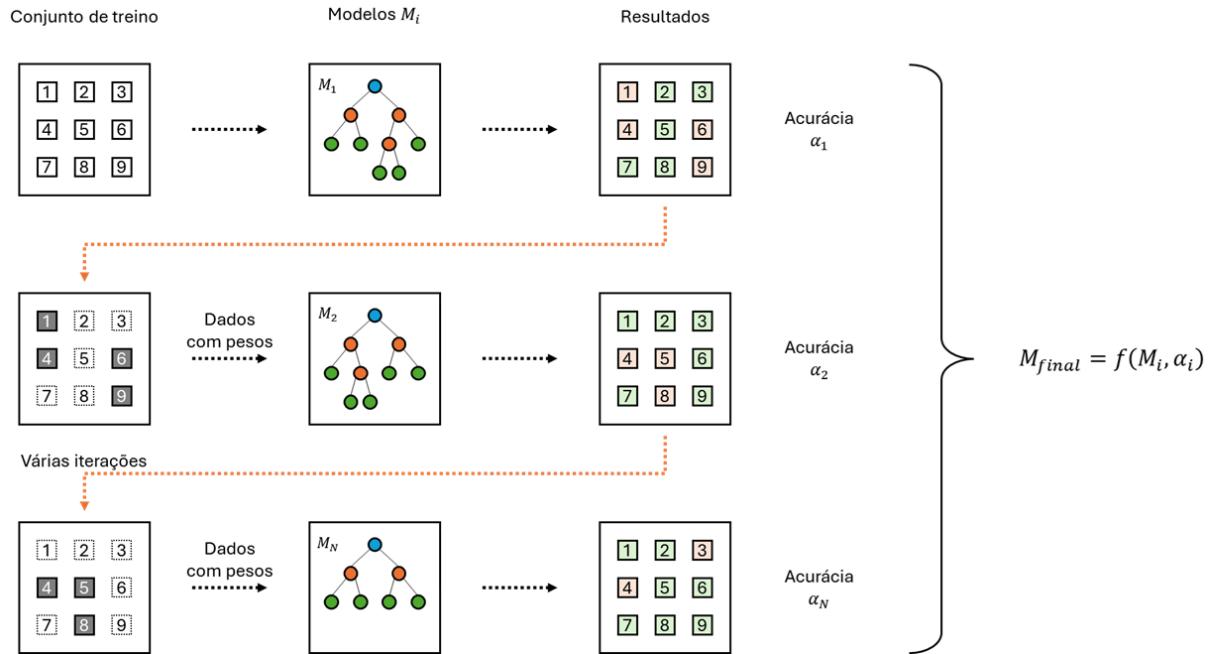
O *Boosting* é uma técnica de *ensemble*, combinando múltiplos modelos fracos para criar um modelo forte. A ideia central é treinar sequencialmente uma série de modelos, onde cada novo modelo foca em corrigir os erros cometidos pelos modelos anteriores. Alguns algoritmos populares de *Boosting* incluem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

Figura 4 – Esquema ilustrativo do funcionamento do *AdaBoost*



Fonte: elaborado pelo autor.

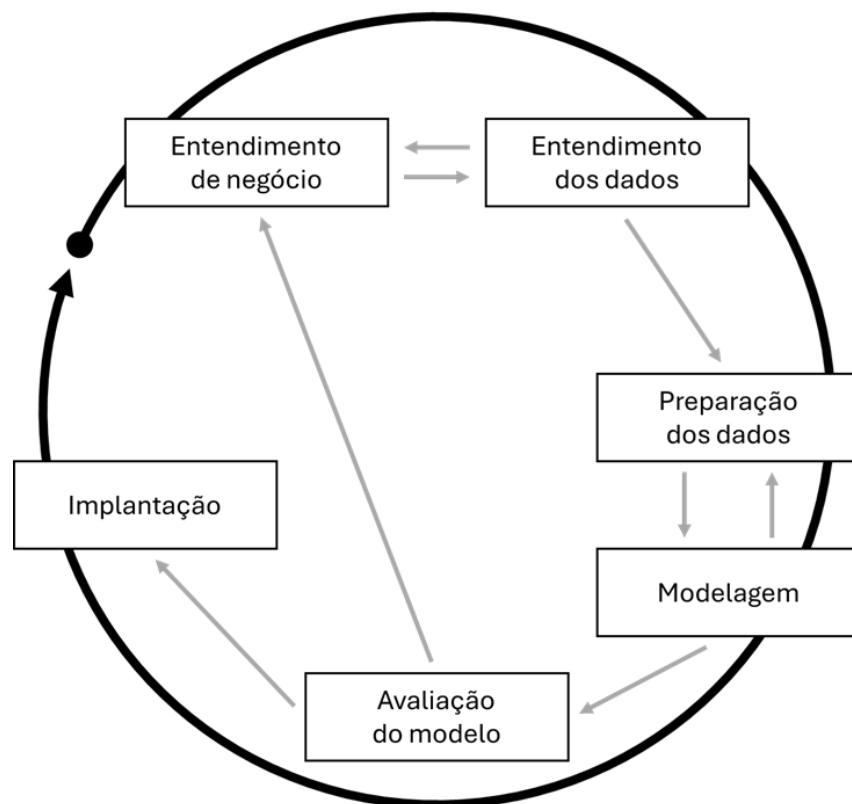


### 3 METODOLOGIA

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 será baseada 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 5 – Modelo CRISP-DM



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

#### 3.1 Entendimento de Negócio

O entendimento de negócio envolve a definição clara dos objetivos do projeto, a compreensão do contexto em que a pesquisa está inserida e a identificação das partes interessadas.

O foco principal deste trabalho acadêmico será a formulação de hipóteses relacionadas aos fatores que influenciam o desempenho dos estudantes no ENEM.

Conforme já 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 algumas dessas variáveis socioeconômicas 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)
Administração: Pública vs. Privada	Moraes <i>et al.</i> (10) Jaloto e Primi (9) Ortega <i>et al.</i> (2)
Atributos Escolares	Moraes <i>et al.</i> (10)

Fonte: elaborado pelo autor.

Assim, para este trabalho, são 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?

Com as perguntas de pesquisa definidas, o próximo passo é compreender os dados disponíveis para análise, conforme descrito na Seção 3.2.

### 3.2 Entendimento dos Dados

Nesta etapa, o foco será a coleta e compreensão dos dados disponíveis para análise. Foram utilizados os microdados do ENEM como fonte principal, que são disponibilizados anualmente pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP).

### 3.2.1 Estrutura dos Dados

Os dados se encontram em arquivos CSV (*Comma-Separated Values*) (4), o que facilita a sua manipulação e análise. Serão utilizados os dados das edições de 2020 a 2023 e possuem, respectivamente, as seguintes quantidade de observações e variáveis:

Tabela 2 – Quantidade de observações e variáveis por edição do ENEM

Edição	Observações	Variáveis
2020	5.783.109	76
2021	3.389.832	76
2022	3.476.105	76
2023	3.933.955	76

Fonte: microdados do INEP; elaborado pelo autor.

O dicionário de dados também é disponibilizado e descreve detalhadamente cada variável presente no conjunto de dados. A seguir, é apresentado um resumo do dicionário de dados. O dicionário completo está disponível no Apêndice A.

- **Dados do participante:** Número de inscrição mascarado, ano do exame, faixa etária, sexo, estado civil, cor/raça, nacionalidade, situação de conclusão do Ensino Médio, ano de conclusão do Ensino Médio, tipo de escola do Ensino Médio, tipo de instituição que concluiu ou concluirá o Ensino Médio e se o inscrito fez a prova como treineiro.
- **Dados da escola:** Código e nome da escola, código e sigla da Unidade da Federação, código do município, nome do município, dependência administrativa, localização e situação de funcionamento.
- **Dados do local de aplicação da prova:** Código e nome da escola onde a prova foi aplicada, código e sigla da Unidade da Federação, código do município e nome do município.
- **Dados das provas objetivas:** Presença e código do tipo de prova objetiva, nota das provas objetivas, vetor com as respostas, língua estrangeira escolhida e vetor com o gabarito.
- **Dados da redação:** Situação da redação (se o candidato apresentou fuga do tema, deixou em branco, copiou o texto motivador, entre outras), nota das cinco competências e nota final da prova de redação.
- **Dados do questionário socioeconômico:** Respostas do questionário socioeconômico aplicado aos participantes do ENEM.

A partir da interpretação do dicionário de dados, já é possível identificar diversas variáveis que podem ser utilizadas para responder às perguntas de pesquisa formuladas na Seção 3.1, assim como as variáveis resposta.

### 3.2.2 Definição da Variável Resposta

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

Assim, teremos cinco variáveis resposta distintas para análise: (i) Nota da prova de Ciências da Natureza; (ii) Nota da prova de Ciências Humanas; (iii) Nota da prova de Linguagens e Códigos; (iv) Nota da prova de Matemática; e (v) Nota da Redação.

## 3.3 Preparação do Ambiente Python

Para a execução desse trabalho, foi utilizado um ambiente virtual baseado em Miniconda3 (21). Dado o volume grande de dados (mais de 16 milhões de observações, 6,4 GB de tamanho), foi necessário utilizar uma GPU para acelerar o processamento dos dados e a modelagem. A GPU utilizada foi uma NVIDIA GeForce RTX 4070 Ti Super, com 16 GB de memória dedicada.

Para possibilitar essa execução, o ambiente foi especificamente configurado com o ecossistema NVIDIA CUDAX (22). 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` (23), uma biblioteca para manipulação de `DataFrames` na GPU análoga ao `pandas` (24), 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 13.0, com as bibliotecas e dependências gerenciadas diretamente pelo Conda.

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

## 3.4 Preparação dos Dados

A preparação dos dados é uma etapa crucial no processo de análise, pois envolve a limpeza, transformação e integração dos dados para torná-los adequados para a modelagem. Nesta seção, detalharei as etapas realizadas para preparar os microdados do ENEM para análise.

### 3.4.1 Coleta, Leitura e Integração dos Dados

A coleta dos dados foi realizada por meio do download dos arquivos CSV disponibilizados pelo INEP para as edições de 2020 a 2023 do ENEM. Os arquivos foram armazenados localmente para facilitar o acesso durante o processo de análise.

A leitura dos dados foi feita utilizando o método `read_csv`, da biblioteca `pandas`, especificando o separador como ponto e vírgula (`sep = ', ;'`).

Os conjuntos de dados do ENEM de 2020 a 2023 possuem o mesmo esquema, ou seja, as mesmas variáveis estão presentes em ambas as edições e com o mesmo nome. Portanto, a integração foi realizada por meio da concatenação vertical dos dois 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.

### 3.4.2 Transformação das Variáveis Categóricas

Analizando o dicionário de dados, foi possível identificar diversas variáveis categóricas representadas por códigos numéricos. Essas variáveis precisam ser transformadas para garantir que os modelos de *Machine Learning* as interpretem corretamente como categóricas e não como variáveis numéricas.

Os códigos numéricos foram então substituídos por rótulos textuais mais descritivos utilizando o método `map` com um dicionário de mapeamento específico para cada variável categórica. Alguns códigos representam ausência de informação e foram substituídos por `None` para um futuro tratamento de valores ausentes adequado.

### 3.4.3 Tratamentos dos Valores Nulos

Após a integração dos dados, foi realizada uma análise inicial para identificar a presença de valores nulos em cada variável. Dependendo do percentual de valores nulos e da natureza da variável, diferentes estratégias de tratamento foram adotadas.

As variáveis com mais de 50% de valores nulos foram removidas do conjunto de dados, pois a alta proporção de valores ausentes poderia comprometer a análise e a modelagem subsequente.

As variáveis de notas das provas objetivas e da redação passaram por uma análise mais detalhada de valores nulos em conjunto com as variáveis de presença nas respectivas provas para entender o significado dos valores nulos e decidir o tratamento adequado. No fim, as observações com notas nulas foram removidas, enquanto as observações com notas zero foram mantidas, pois representam casos distintos.

As demais variáveis com valores nulos foram consideradas de maneira consolidada, ou seja, caso uma observação possua valor ausente em qualquer uma dessas variáveis, ela será considerada como tendo valor ausente. Essas observações foram então removidas do conjunto de dados.

#### 3.4.4 Separação dos Conjuntos de Dados por Variável Resposta

Como possuímos cinco variáveis resposta distintas, é necessário preparar os dados para cada uma delas. Assim, foram criados cinco conjuntos de dados distintos, um para cada variável resposta. O filtro aplicado é o mesmo para todos os conjuntos de dados: foram mantidas apenas as observações que possuem presença na prova correspondente. Para a nota da redação, foi utilizada a variável de presença na prova de Linguagens e Códigos, que é quando a redação é aplicada.

Além disso, foram retiradas as variáveis de nota das demais provas, assim como as variáveis de presença nas demais provas, para evitar qualquer tipo de *data leakage* durante a modelagem.

Foi utilizado a estrutura de dicionário para armazenar os cinco conjuntos de dados distintos, facilitando o acesso e a manipulação dos dados durante a análise.

#### 3.4.5 Exploração Inicial

Com os dados no formato desejado, foi realizada uma exploração inicial para compreender a estrutura dos dados e detectar possíveis inconsistências.

Primeiramente, foi analisada a distribuição das notas para cada variável resposta, utilizando histogramas e boxplots. Foi constatado que a nota da redação é aplicada em intervalos de 20 pontos, enquanto que as demais notas são aplicadas de maneira contínua. Isso se dá pela quantidade muito diferente de notas distintas existentes em cada prova: a redação possui 50 valores distintos, enquanto que as demais provas possuem mais de 5.200 valores distintos cada.

### 3.5 Análise de correlação das Variáveis Preditoras

Antes de iniciar a modelagem, foi realizada uma análise de correlação entre as variáveis preditoras para identificar possíveis multicolinearidades que poderiam afetar o desempenho dos modelos. A correlação foi avaliada utilizando a biblioteca `phik` (27) (28), que permite calcular a correlação entre variáveis categóricas e numéricas.

#### 3.5.1 Imputação de Valores Ausentes

Para as variáveis com valores ausentes que não foram removidas, foi adotado o método de *KNN Imputer*, que utiliza a média/moda dos  $k$  vizinhos mais próximos para

preencher os valores ausentes. Esse método foi escolhido por sua capacidade de preservar a distribuição dos dados e considerar a correlação entre as variáveis.

O *KNN Imputer* foi implementado utilizando a classe `KNNImputer` da biblioteca `sklearn.impute`. Para se determinar o hiperparâmetro  $k$ , foi feita uma rápida análise de sensibilidade, testando diferentes valores de  $k$  e avaliando o impacto na distribuição das variáveis imputadas e na performance dos modelos preditivos. Para cada variável a ser imputada, foi selecionado o valor de  $k$  que apresentou o melhor equilíbrio entre acurácia e estabilidade.

### 3.6 Modelagem

A modelagem é a etapa onde os dados preparados são utilizados para construir modelos preditivos que possam responder às perguntas de pesquisa formuladas na Seção ???. Nesta seção, serão detalhadas as técnicas de modelagem empregadas e os critérios utilizados para a seleção dos modelos.

#### 3.6.1 Seleção de Modelos

Como as nossas variáveis respostas são numéricas e contínuas, estamos em um problema de regressão de aprendizado supervisionado. Assim, é preciso selecionar modelos de regressão, que são os adequados para prever variáveis contínuas. Os modelos escolhidos foram a Regressão Linear com Regularizações (como *benchmark*) e o XGBoost Regressor (29), ambos amplamente utilizados devido à sua eficácia e interpretabilidade.

#### 3.6.2 Otimização dos Hiperparâmetros

Para otimizar o desempenho dos modelos selecionados, foi realizada uma busca em grade (*Grid Search*) para identificar os melhores hiperparâmetros. A busca foi realizada utilizando validação cruzada para garantir que os resultados fossem robustos e generalizáveis.

#### 3.6.3 Treino dos Modelos

Definidos os hiperparâmetros ótimos, os modelos foram treinados utilizando o conjunto de dados preparado na Seção ???. O treinamento foi realizado em validação cruzada de cinco *folds*, utilizando a biblioteca `cuml` para se beneficiar da aceleração por GPU. Foi selecionado como modelo final aquele modelo que apresentou o melhor desempenho médio nos *folds* de validação para cada variável resposta.

#### 3.6.4 Avaliação dos Modelos

A avaliação dos modelos foi realizada utilizando métricas de desempenho apropriadas para problemas de regressão, como a Raiz do Erro Quadrático Médio (RMSE) e o

Coeficiente de Determinação ( $R^2$ ). Essas métricas fornecem uma visão clara da precisão das previsões dos modelos em relação aos valores reais das notas do ENEM.

Também foram realizadas análises de resíduos para verificar a adequação dos modelos e identificar possíveis padrões não capturados pelas previsões, bem como a análise gráfica das previsões versus os valores reais para avaliar visualmente o desempenho dos modelos.

### **3.7 Magnitude da Influência das Variáveis Preditoras**

Para medir o efeito de cada variável preditora nas notas do ENEM, foram utilizadas técnicas de interpretação de modelos, que permitem identificar quais variáveis têm maior impacto nas previsões dos modelos e como elas influenciam as notas dos estudantes. Entre as técnicas empregadas estão a importância das características (*feature importance*), a análise dos valores SHAP (*SHapley Additive exPlanations*) e os gráficos de dependência parcial (*partial dependence plots*).

Além disso, foi realizada uma análise de sensibilidade para avaliar como mudanças nas variáveis preditoras afetam as previsões dos modelos. Para isso, foi construída uma matriz de cenários hipotéticos, onde cada variável preditora é alterada sistematicamente enquanto as outras são mantidas constantes. As previsões resultantes foram então analisadas para quantificar o impacto de cada variável nas notas do ENEM.

### **3.8 Limitações e Considerações Éticas**

Algumas limitações devem ser destacadas. Primeiro, apesar do grande volume de dados, a presença de vieses de seleção (por exemplo, diferença entre participantes regulares e treineiros, ou entre ausentes e presentes nas provas) pode influenciar as inferências; as análises procuram mitigar esses efeitos, mas não os eliminam completamente. Segundo, variáveis censuradas, inconsistências de registro e códigos especiais para ausência exigiram tratamentos que podem introduzir perdas de informação.

Quanto às considerações éticas, todos os dados utilizados são microdados públicos disponibilizados pelo INEP, já anonimizados para preservar a privacidade dos participantes. Recomenda-se cautela na interpretação dos resultados para evitar conclusões simplistas sobre individualidades dos estudantes ou estigmatização de grupos e instituições.

No Capítulo 4 serão apresentados os resultados empíricos da metodologia empregada: (i) a descrição dos dados após o pré-processamento, (ii) a análise de correlação entre as variáveis preditoras, (iii) o desempenho dos modelos preditivos na previsão das notas do ENEM, e (iv) a análise da magnitude da influência das variáveis preditoras nas notas dos estudantes.

## 4 RESULTADOS

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

### 4.1 Entendimento dos Dados

#### 4.1.1 Tratamento dos Valores nulos

Conforme descrito na Seção 3.4.3, os dados nulos foram tratados de acordo com a natureza de cada variável. A Tabela 3 apresenta o percentual de valores nulos por variável na conjunto de dados integrados incial, antes dos tratamentos.

A quantidade de nulos nas variáveis varia significativamente, com algumas 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 realizado alguma imputação de valores.

Nas variáveis de notas, foi observado que há valores preenchidos com zero e valores nulos. Após análise conjunta com as variáveis de presença nas respecitivas provas, constatou-se que os valores nulos indicam ausência ou eliminação do participante na prova, enquanto os valores zero indicam que o participante esteve presente na prova, mas obteve nota zero. Dessa forma, as observações com notas nulas foram removidas, enquanto as observações com notas zero foram mantidas.

Após a aplicação dos tratamentos descritos na Seção 3.4.3, restaram 12.241.442 observações e 45 variáveis no conjunto de dados integrado.

#### 4.1.2 Separação dos Conjuntos de Dados por Variável Resposta

Após a separação dos conjuntos de dados por variável resposta, conforme descrito na Seção 3.4.4, foram criados cinco conjuntos de dados distintos. A Tabela 4 apresenta a quantidade de observações e variáveis em cada conjunto de dados.

#### 4.1.3 Exploração Inicial

Tabela 3 – Percentual inicial de valores nulos por variável

<b>Variável</b>	<b>Percentual de nulos</b>
sigla_uf_escola	78,1%
cod_municipio_escola	78,1%
tp_adm_escola	78,1%
nome_municipio_escola	78,1%
cod_uf_escola	78,1%
tp_local_escola	78,1%
funcionamento_escola	78,1%
tp_ensino	69,8%
tp_escola	68,2%
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%
06_renda_familiar	0,6%
07_qtde_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.

Tabela 4 – Quantidade de observações por conjunto de dados

<b>Conjunto de Dados</b>	<b>Observações</b>	<b>Variáveis</b>
Ciências Humanas	7.895.093	36
Ciências Natureza	7.500.050	36
Linguagem e Código	7.895.093	36
Matemática	7.500.050	36
Redação	7.895.093	36

Fonte: elaborado pelo autor.



## 5 CONCLUSÃO

XX



## 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**. Local: Brasília, DF. [s.d.]. 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 Anaconda. **Getting started with Miniconda**. n.d. Disponível em: <https://www.anaconda.com/docs/getting-started/miniconda/main>.
- 22 NVIDIA. **CUDA-X Data Science Libraries**. n.d. Disponível em: <https://developer.nvidia.com/topics/ai/data-science/cuda-x-data-science-libraries>.
- 23 NVIDIA. **Welcome to the cuDF documentation!** n.d. Disponível em: <https://docs.rapids.ai/api/cudf/stable/>.
- 24 The pandas development team. **pandas documentation**. 2025. Disponível em: <https://pandas.pydata.org/docs/>.
- 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 KPMG Advisory N.V. **Phi\_K Correlation Constant**. 2024. Disponível em: <https://phik.readthedocs.io/en/latest/>.
- 28 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.

29 xgboost developers. **XGBoost Tutorials**. 2025. Disponível em: <https://xgboost.readthedocs.io/en/stable/tutorials/index.html>.



## **APÊNDICES**



**APÊNDICE A – DICIONÁRIO DE DADOS DOS MICRODADOS DO ENEM**



**APÊNDICE B – CONFIGURAÇÃO DO AMBIENTE VIRTUAL**