UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Título

Ramon de Castro Ramos

Monografia - MBA em Ciência de Dados (CeMEAI)

ICMC USP
SÃO CARLOS

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. 67 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. 67 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

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

AdaBoost	*Adaptive Boosting*
COVID-19	*Coronavirus Disease 2019* - Doença do Coronavírus 2019
CRISP-DM	*Cross-Industry Standard Process for Data Mining* - Processo de Mineração de Dados Padrão entre Indústrias
CSV	*Comma-Separated Values* - 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	*Machine Learning* - Aprendizado de Máquina
ProUni	Programa Universidade Para Todos
RF	*Random Forest* - Floresta Aleatória
SISU	Sistema de Seleção Unificada
UFABC	Universidade Federal do ABC
XGBoost	*Extreme Gradient Boosting*

LISTA DE SÍMBOLOS

α	*Alpha* - Primeiro caractere do alfabeto grego
β	*Beta* - Segundo caractere do alfabeto grego
ϵ	*Epsilon* - Quinto caractere do alfabeto grego
$\leq$	*Menor ou igual a*

SUMÁRIO

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 aos 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.; (ii) 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

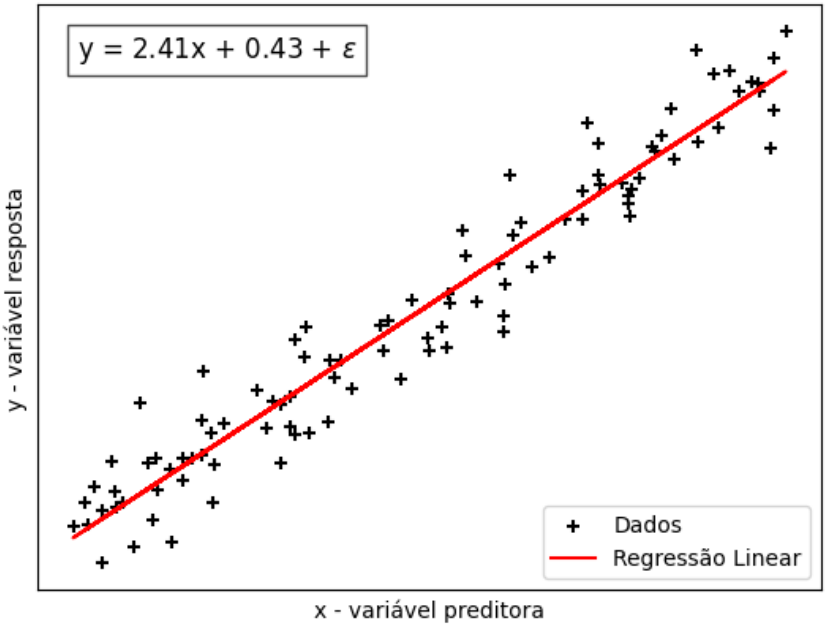
$$Y = \beta_0 + \beta_1 \times X + \epsilon \tag{2.1}$$

onde Y denota a variável resposta, X a variável preditora, β_0 o intercepto (o valor de Y quando $X = 0$), β_1 o coeficiente angular (indicando o impacto de X sobre Y) e ϵ o termo de erro. Em uma regressão múltipla, diversas variáveis independentes são consideradas, cada uma com o seu β_i correspondente.

Por trás da regressão linear, há algumas premissas adotadas, como a linearidade da relação entre X e Y, a independência dos erros, a homocedasticidade e a normalidade

dos resíduos. Essas premissas podem ser interpretadas como desvantagens do modelo de regressão linear, por restringir ou até mesmo 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

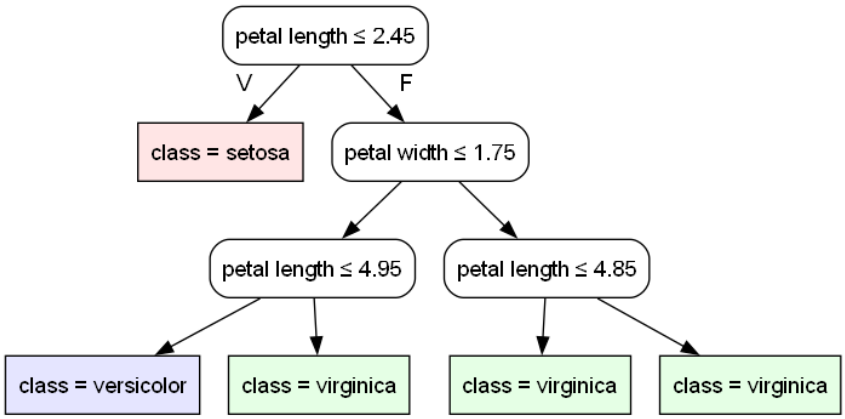
A Árvore de Decisão é um método paramétrico de aprendizado supervisionado que utiliza uma abordagem intuitiva de separação dos dados em grupos semelhantes, através de regras hierárquicas simples e de forma recursiva. Pode ser utilizado para resolver problemas de regressão, com a média da variável resposta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

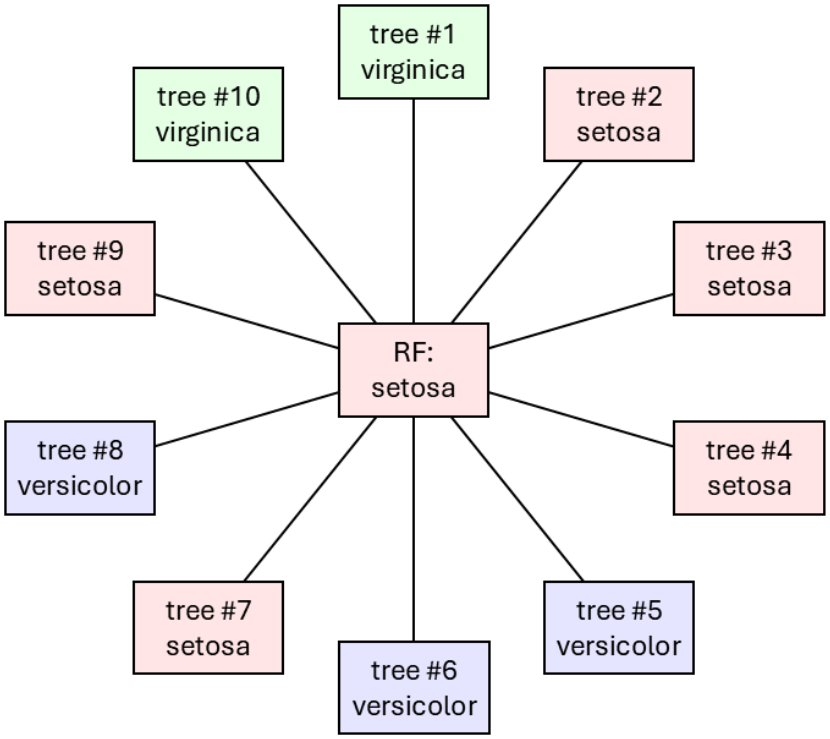
2.7.3 *Random Forest*

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*

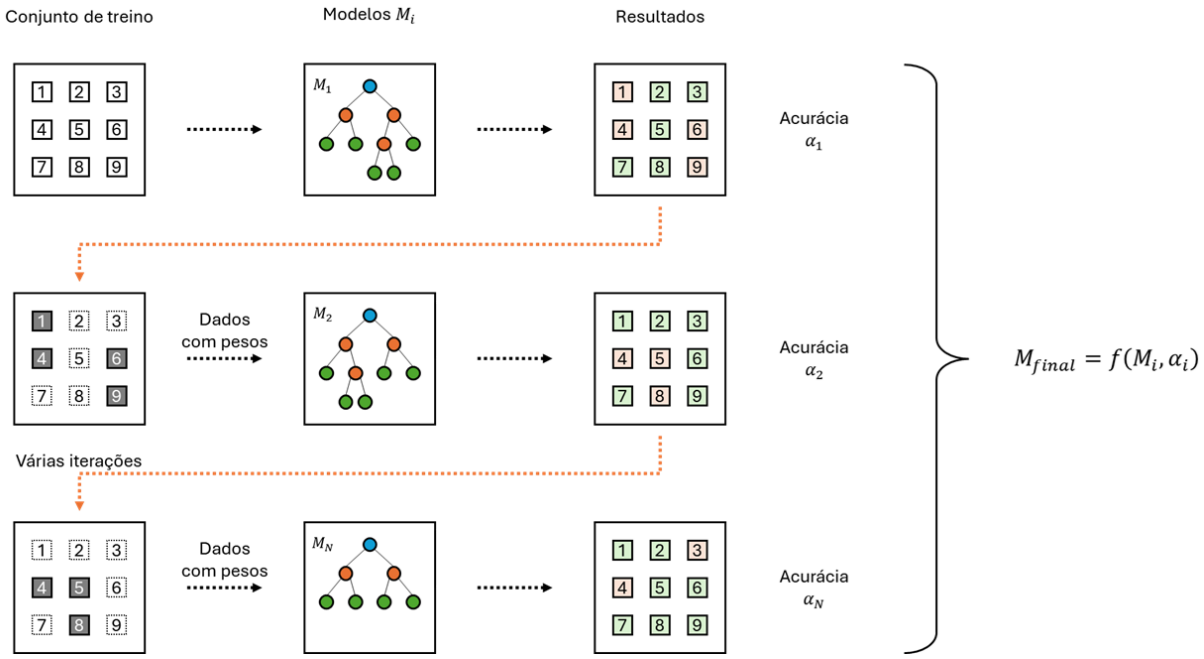
O *Boosting* é uma técnica de *ensemble*, combinando múltiplos modelos fracos para criar um modelo forte. A ideia central é treinar sequencialmente uma série de modelos, onde cada novo modelo foca em corrigir os erros cometidos pelos modelos anteriores. Alguns algoritmos populares de *Boosting* incluem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



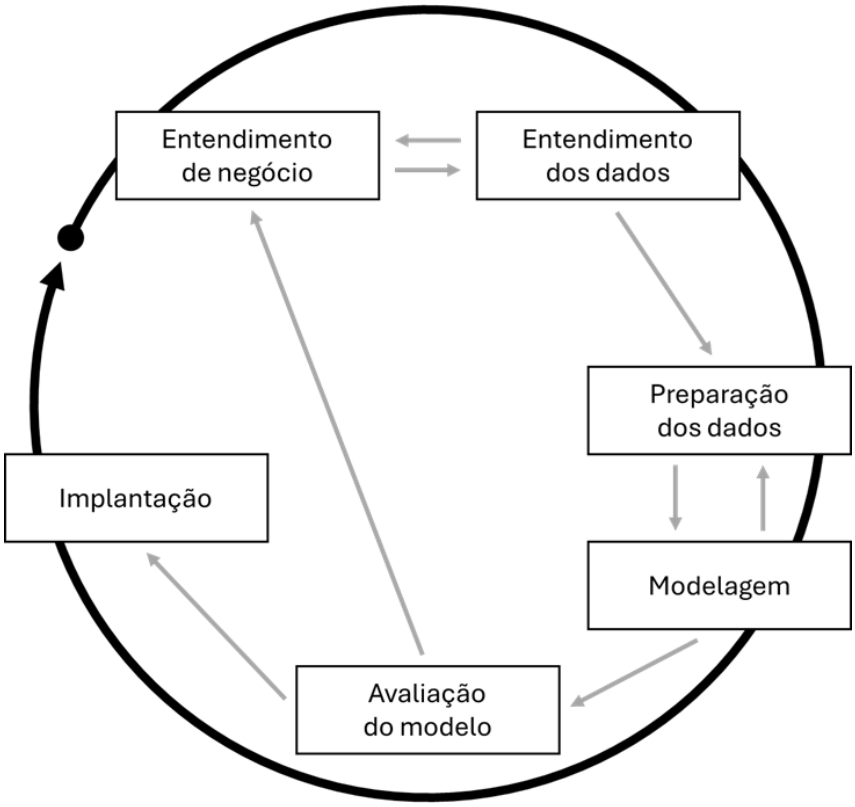
Fonte: elaborado pelo autor.

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

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

O foco principal deste trabalho será a formulação de hipóteses relacionadas aos

fatores que influenciam o desempenho dos estudantes no ENEM, através da análise dos dados de performance dos participantes e suas características socioeconômicas.

Sendo assim, foi necessário formular perguntas de pesquisa específicas que possam ser respondidas através da análise dos dados disponíveis.

3.2 Entendimento dos dados

Com as perguntas de pesquisa definidas, a próxima etapa foi encontrar dados que sejam adequados para responder a essas perguntas e estabelecer, primeiramente, uma forma consistente de coleta e armazenamento para em seguida realizar uma rápida compreensão da estrutura dos dados.

Há depender dos tipos de dados a serem utilizados, é necessário submeter o projeto a um comitê de ética em pesquisa para aprovação, garantindo que todos os aspectos éticos relacionados ao uso dos dados sejam devidamente considerados.

Após essa etapa, foi possível identifcar quais os arquivos são relevantes para a análise e quais variáveis dentro desses arquivos serão utilizadas como variáveis preditoras e como variáveis resposta.

3.3 Preparação dos dados

Com os arquivos relevantes selecionados, passa-se para a etapa de preparação dos dados, que envolve a leitura, limpeza, transformação e integração dos dados para torná-los adequados para a modelagem. Essa etapa é crucial, pois a qualidade dos dados impacta diretamente na eficácia dos modelos preditivos que serão construídos posteriormente.

Para a execução dessa e das etapas posteriores, é necessário preparar um ambiente tecnológico e analítico adequado que permita a manipulação eficiente dos dados e a construção dos modelos preditivos.

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

Em seguida, as variáveis foram renomeadas para nomes mais intuitivos e de fácil compreensão e foi analisado se seria necessário criar variáveis contendo alguma transformação das variáveis originais, por exemplo, criar uma variável que transforme os códigos numéricos de variáveis categóricas em rótulos textuais.

Posteriormente, foi realizada uma análise para identificar e tratar valores nulos, removendo ou imputando valores conforme apropriado. Após o tratamento dos valores nulos, os dados foram separados em diferentes conjuntos de dados, cada um correspondente

a uma variável resposta específica, garantindo que cada conjunto contenha apenas as observações relevantes para a análise daquela variável e sem valores nulos.

3.4 Modelagem

3.5 Avaliação

3.6 Implantação

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 de Negócio

Conforme descrito no Capítulo 3 - Metodologia, foi necessário definir as perguntas de pesquisa que guiaram a análise dos dados disponíveis.

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

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

Tabela 1 – Variáveis socioeconômicas e suas referências

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

Fonte: elaborado pelo autor.

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

- **Pergunta 1:** Quais são os principais fatores socioeconômicos que influenciam o desempenho dos estudantes no ENEM?

- **Pergunta 2:** Qual é a magnitude da influência de cada um desses conjuntos de fatores nas notas dos participantes?

4.2 Entendimento dos dados

4.2.1 Escolha e Coleta dos Dados

Como descrito no Capítulo 3 - Metodologia, foi necessário identificar dados que fossem relevantes para responder as pergunras de pesquisa formuladas. Foi realizada uma busca por bases de dados públicas que contivessem informações detalhadas sobre os participantes do ENEM, incluindo suas características socioeconômicas e desempenho no exame.

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

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

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

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

4.2.2 Compreensão Inicial dos Dados

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

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

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

4.2.2.1 Edicão de 2024 do ENEM e LGPD

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

Isso aconteceu "Devido à vigência da Lei Geral de Proteção de Dados (LGPD), incorporada ao ordenamento jurídico brasileiro por meio da Lei nº 13.709, de 14 de agosto de 2018", conforme descrito no arquivo auxiliar ¨Leia-Me"disponibilizado junto com os microdados do ENEM 2024 (21).

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

4.2.3 Análise dos Dicionários de Dados

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

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

4.2.4 Definição da Variável Resposta

Como esse trabalho pretende avaliar o desempenho dos estudantes no ENEM e os fatores que influenciam esse desempenho, a 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.

Ou seja, 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.

4.3 Preparação dos dados

4.3.1 Preparação do Ambiente Tecnológico e Analítico

Para a execução desse trabalho, foi utilizado um ambiente baseado em `Python` através do gerenciador de ambientes virtuais Miniconda3 (22). Dado o grande volume 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 (23). 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` (24), uma biblioteca para manipulação de `DataFrames` na GPU análoga ao `pandas` (25), e `cuml` (26), que fornece implementações de algoritmos de *Machine Learning* acelerados por GPU, análoga ao `scikit-learn` (27). 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 C.

4.3.2 Leitura dos Dados

Os arquivos CSV dos microdados do ENEM foram lidos utilizando o método `read_csv` da biblioteca `pandas`(25) especificando o separador como ponto e vírgula (`sep = ';'`).

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.

Analisando a Tabela 2 e os dicionários de dados, foi possível observar que todas as edições selecionadas possuem o mesmo esquema, ou seja, as mesmas variáveis e com o mesmo nome estão presentes em todas as edições selecionadas. Portanto, a integração foi

realizada por meio da concatenação vertical dos quatro conjuntos de dados, utilizando o método `concat` da biblioteca `pandas`.

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

4.4 Modelagem

4.5 Avaliação

4.6 Implantação

5 CONCLUSÃO

xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx

REFERÊNCIAS

1 MELO, R. O. *et al.* Impacto das variáveis socioeconômicas no desempenho do ENEM: uma análise espacial e sociológica. **Revista de Administração Pública**, v. 55, n. 6, p. 1271–1294, nov./dez. 2021.

2 ORTEGA, A. *et al.* Análise comparativa: Escola pública x escola privada no ENEM. *In*: **Primeiro Hackthon de Dados pela Universidade Federal do ABC**. São Paulo: [*S.l.: s.n.*], 2025. Relatório.

3 NASCIMENTO, M. M. *et al.* Análise estatística e pluriescalar das desigualdades educacionais: aspirações científicas e desempenho de estudantes no ENEM. **Sociologias**, v. 27, 2025. Disponível em: https://doi.org/10.1590/1807-0337/e130399.

4 INEP. **Microdados ENEM**. Disponível em: https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem.

5 INEP. **Histórico do ENEM**. Disponível em: https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enem/historico.

6 INEP. **ENEM**. Disponível em: https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enem.

7 OLIVEIRA, L. K. S.; CRUZ, R. C. Capital cultural e educação: uma análise da obra de bordieu. *In*: **XIII Encontro Cearense de Historiadores da Educação - ECHE, III Encontro Nacional do Núcleo de História e Memória da Educação - ENHIME, III Simpósio Nacional de Estudos Culturais e Geoeducacionais - SINECGEO**. [*S.l.: s.n.*], 2014. p. 1247–1255. ISBN: 978-85-8126-065-5. Documento de evento, sem data e local de publicação explícitos.

8 VASCONCELLOS, F. **Resultados do ENEM refletem desigualdades comuns no país**. 2013. Disponível em: https://oglobo.globo.com/brasil/educacao/resultados-do-enem-refletem-desigualdades-comuns-no-pais-10445682.

9 JALOTO, A.; PRIMI, R. Fatores socioeconômicos associados ao desempenho no ENEM. **Em Aberto**, v. 34, n. 112, p. 125–141, dec 2021. Disponível em: https://www.researchgate.net/publication/357656960.

10 MORAES, C. P. d. *et al.* Efeito escola a partir de indicadores educacionais: análise entre escolas públicas e privadas no ENEM. **REVISTA META: AVALIAÇÃO**, v. 14, n. 42, p. 67–93, mar 2022.

11 BARTHOLO, T. *et al.* **Oportunidades educacionais de estudantes concluintes do Ensino Médio: Relatório 1-Inscrição e Participação no ENEM entre 2013 e 2021**. Rio de Janeiro, 2023.

12 HIROMI, F. ENEM mais desigual requer atenção dos gestores. **Aprendizagem em Foco**, n. 92, oct 2023. Disponível em: https://www.institutounibanco.org.br/boletim/enem-mais-desigual-requer-atencao-dos-gestores/.

13 ROMERO, M. C. **Aplicando técnicas de Machine Learning para avaliar resultados do ENEM**. 2021. 72 p. Dissertação (Trabalho de Conclusão de Curso (MBA em Ciências de Dados)) — Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2021.

14 FERRAZ, A. P. **Prevendo a aprovação de um participante do ENEM no SISU para o curso de Medicina**. 2020. 70 p. Dissertação (Trabalho de Conclusão de Curso (MBA em Ciências de Dados)) — Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2020.

15 INEP. **Microdados Censo Escolar**. Local: Brasília, DF. [s.d.]. Disponível em: https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/censo-escolar.

16 JAMES, G. *et al.* **An Introduction to Statistical Learning: with Applications in Python**. Boca Raton: CRC Press, 2023.

17 GRUS, J. **Data Science from Scratch: First Principles with Python**. 2. ed. Sebastopol, CA: O'Reilly Media, Inc., 2019.

18 LINDHOLM, A. *et al.* **Machine Learning: A First Course for Engineers and Scientists**. Cambridge, UK; New York, NY: Cambridge University Press, 2022.

19 BREIMAN, L. Random forests. **Machine Learning**, v. 45, n. 1, p. 5–32, oct 2001.

20 CHAPMAN, P. *et al.* **CRISP-DM 1.0: Step-by-step Data Mining Guide**. [S.l.], 2000. Disponível em: https://mineracaodedados.wordpress.com/wp-content/uploads/2012/12/crisp-dm-1-0.pdf.

21 INEP. **Microdados do Enem 2024: Leia-Me**. Brasília, 2025. Disponível em: https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem.

22 Anaconda. **Getting started with Miniconda**. n.d. Disponível em: https://www.anaconda.com/docs/getting-started/miniconda/main.

23 NVIDIA. **CUDA-X Data Science Libraries**. n.d. Disponível em: https://developer.nvidia.com/topics/ai/data-science/cuda-x-data-science-libraries.

24 NVIDIA. **Welcome to the cuDF documentation!** n.d. Disponível em: https://docs.rapids.ai/api/cudf/stable/.

25 The pandas development team. **pandas documentation**. 2025. Disponível em: https://pandas.pydata.org/docs/.

26 NVIDIA. **Welcome to cuML's documentation!** 2023. Disponível em: https://docs.rapids.ai/api/cuml/stable/.

27 PEDREGOSA, F. *et al.* Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

APÊNDICES

APÊNDICE A – DICIONÁRIO DE DADOS DOS MICRODADOS DO ENEM

APÊNDICE B – DICIONÁRIO DE DADOS DO CENSO ESCOLAR

APÊNDICE C – CONFIGURAÇÃO DO AMBIENTE VIRTUAL

```
name: tcc_gpu
channels:
  - conda-forge
  - rapidsai
  - defaults
dependencies:
  - _libgcc_mutex=0.1=conda_forge
  - _openmp_mutex=4.5=2_gnu
  - alsa-lib=1.2.14=hb9d3cd8_0
  - asttokens=3.0.0=pyhd8ed1ab_1
  - attr=2.5.2=h39aace5_0
  - aws-c-auth=0.9.1=h194c533_5
  - aws-c-cal=0.9.8=h346e085_0
  - aws-c-common=0.12.5=hb03c661_1
  - aws-c-compression=0.3.1=h7e655bb_8
  - aws-c-event-stream=0.5.6=h1deb5b9_4
  - aws-c-http=0.10.7=had4b759_1
  - aws-c-io=0.23.2=hbff472d_2
  - aws-c-mqtt=0.13.3=h8ba2272_8
  - aws-c-s3=0.8.6=h493c25d_7
  - aws-c-sdkutils=0.2.4=h7e655bb_3
  - aws-checksums=0.2.7=h7e655bb_4
  - aws-crt-cpp=0.35.0=h719b17a_2
  - aws-sdk-cpp=1.11.606=h522d481_6
  - azure-core-cpp=1.16.1=h3a458e0_0
  - azure-identity-cpp=1.13.2=h3a5f585_1
  - azure-storage-blobs-cpp=12.15.0=h2a74896_1
  - azure-storage-common-cpp=12.11.0=h3d7a050_1
  - azure-storage-files-datalake-cpp=12.13.0=hf38f1be_1
  - bokeh=3.8.1=pyhd8ed1ab_0
  - brotli=1.2.0=h41a2e66_0
  - brotli-bin=1.2.0=hf2c8021_0
  - brotli-python=1.2.0=py311h7c6b74e_0
  - bzip2=1.0.8=hda65f42_8
  - c-ares=1.34.5=hb9d3cd8_0
  - ca-certificates=2025.11.12=hbd8a1cb_0
  - cachetools=6.2.2=pyhd8ed1ab_0
  - cairo=1.18.4=h3394656_0
  - cffi=2.0.0=py311h03d9500_1
  - click=8.3.0=pyh707e725_0
  - cloudpickle=3.1.2=pyhd8ed1ab_0
  - comm=0.2.3=pyhe01879c_0
  - contourpy=1.3.3=py311hdf67eae_3
  - cuda-bindings=12.9.4=py311ha898f3d_1
  - cuda-cccl_linux-64=12.9.27=ha770c72_0
  - cuda-core=0.3.2=py311h2cd87c0_0
  - cuda-crt-dev_linux-64=12.9.86=ha770c72_2
  - cuda-crt-tools=12.9.86=ha770c72_2
  - cuda-cudart=12.9.79=h5888daf_0
  - cuda-cudart-dev=12.9.79=h5888daf_0
  - cuda-cudart-dev_linux-64=12.9.79=h3f2d84a_0
  - cuda-cudart-static=12.9.79=h5888daf_0
  - cuda-cudart-static_linux-64=12.9.79=h3f2d84a_0
  - cuda-cudart_linux-64=12.9.79=h3f2d84a_0
  - cuda-nvcc-dev_linux-64=12.9.86=he91c749_2
```

```
- cuda-nvcc-impl=12.9.86=h85509e4_2
- cuda-nvcc-tools=12.9.86=he02047a_2
- cuda-nvrtc=12.9.86=hecca717_1
- cuda-nvvm-dev_linux-64=12.9.86=ha770c72_2
- cuda-nvvm-impl=12.9.86=h4bc722e_2
- cuda-nvvm-tools=12.9.86=h4bc722e_2
- cuda-pathfinder=1.3.2=pyhcf101f3_0
- cuda-profiler-api=12.9.79=h7938cbb_1
- cuda-python=12.9.4=pyh15a92d1_1
- cuda-version=12.9=h4f385c5_3
- cudf=25.10.00=cuda12_py311_251008_f4e35ca0
- cuml=25.10.00=cuda12_py311_251008_f9fcdbb0
- cupy=13.6.0=py311h72da3fd_2
- cupy-core=13.6.0=py311he30c881_2
- cycler=0.12.1=pyhd8ed1ab_1
- cyrus-sasl=2.1.28=hd9c7081_0
- cytoolz=1.1.0=py311h49ec1c0_1
- dask=2025.9.1=pyhcf101f3_0
- dask-core=2025.9.1=pyhcf101f3_0
- dask-cuda=25.10.00=py313_251008_472ca1ce
- dask-cudf=25.10.00=cuda12_py311_251008_f4e35ca0
- dbus=1.16.2=h3c4dab8_0
- debugpy=1.8.17=py311hc665b79_0
- decorator=5.2.1=pyhd8ed1ab_0
- distributed=2025.9.1=pyhcf101f3_0
- distributed-ucxx=0.46.00=py_251008_64355220_h9c9281c
- dlpack=0.8=h59595ed_3
- double-conversion=3.3.1=h5888daf_0
- executing=2.2.1=pyhd8ed1ab_0
- fastrlock=0.8.3=py311hc665b79_2
- font-ttf-dejavu-sans-mono=2.37=hab24e00_0
- font-ttf-inconsolata=3.000=h77eed37_0
- font-ttf-source-code-pro=2.038=h77eed37_0
- font-ttf-ubuntu=0.83=h77eed37_3
- fontconfig=2.15.0=h7e30c49_1
- fonts-conda-ecosystem=1=0
- fonts-conda-forge=1=hc364b38_1
- fonttools=4.60.1=py311h3778330_0
- freetype=2.14.1=ha770c72_0
- fsspec=2025.10.0=pyhd8ed1ab_0
- gflags=2.2.2=h5888daf_1005
- glog=0.7.1=hbabe93e_0
- graphite2=1.3.14=hecca717_2
- h2=4.3.0=pyhcf101f3_0
- harfbuzz=12.2.0=h15599e2_0
- hpack=4.1.0=pyhd8ed1ab_0
- hyperframe=6.1.0=pyhd8ed1ab_0
- icu=75.1=he02047a_0
- importlib-metadata=8.7.0=pyhe01879c_1
- ipykernel=7.1.0=pyha191276_0
- ipython=9.7.0=pyh53cf698_0
- ipython_pygments_lexers=1.1.1=pyhd8ed1ab_0
- jedi=0.19.2=pyhd8ed1ab_1
- jinja2=3.1.6=pyhd8ed1ab_0
- joblib=1.5.2=pyhd8ed1ab_0
- jupyter_client=8.6.3=pyhd8ed1ab_1
- jupyter_core=5.9.1=pyhc90fa1f_0
- keyutils=1.6.3=hb9d3cd8_0
- kiwisolver=1.4.9=py311h724c32c_2
- krb5=1.21.3=h659f571_0
```

```
- lcms2=2.17=h717163a_0
- ld_impl_linux-64=2.45=h1aa0949_0
- lerc=4.0.0=h0aef613_1
- libabseil=20250512.1=cxx17_hba17884_0
- libarrow=21.0.0=h552f9d5_11_cuda
- libarrow-acero=21.0.0=hb826db4_11_cuda
- libarrow-compute=21.0.0=h58682fd_11_cuda
- libarrow-dataset=21.0.0=hb826db4_11_cuda
- libarrow-substrait=21.0.0=h9d9f3f8_11_cuda
- libblas=3.9.0=39_h4a7cf45_openblas
- libbrotlicommon=1.2.0=h09219d5_0
- libbrotlidec=1.2.0=hd53d788_0
- libbrotlienc=1.2.0=h02bd7ab_0
- libcap=2.77=h3ff7636_0
- libcblas=3.9.0=39_h0358290_openblas
- libclang-cpp21.1=21.1.5=default_h99862b1_1
- libclang13=21.1.5=default_h746c552_1
- libcrc32c=1.1.2=h9c3ff4c_0
- libcublas=12.9.1.4=h676940d_1
- libcublas-dev=12.9.1.4=h676940d_1
- libcudf=25.10.00=cuda12_251008_f4e35ca0
- libcufft=11.4.1.4=hecca717_1
- libcufile=1.14.1.1=hbc026e6_1
- libcufile-dev=1.14.1.1=hecca717_1
- libcuml=25.10.00=cuda12_251008_f9fcdbb0
- libcumlprims=25.10.00=cuda12_py310_251008_7b289cfd
- libcups=2.3.3=hb8b1518_5
- libcurand=10.3.10.19=h676940d_1
- libcurand-dev=10.3.10.19=h676940d_1
- libcurl=8.17.0=h4e3cde8_0
- libcusolver=11.7.5.82=h676940d_2
- libcusolver-dev=11.7.5.82=h676940d_2
- libcusparse=12.5.10.65=hecca717_2
- libcusparse-dev=12.5.10.65=hecca717_2
- libcuvs=25.10.00=cuda12_251008_f245c152
- libdeflate=1.25=h17f619e_0
- libdrm=2.4.125=hb03c661_1
- libedit=3.1.20250104=pl5321h7949ede_0
- libegl=1.7.0=ha4b6fd6_2
- libev=4.33=hd590300_2
- libevent=2.1.12=hf998b51_1
- libexpat=2.7.1=hecca717_0
- libffi=3.5.2=h9ec8514_0
- libfreetype=2.14.1=ha770c72_0
- libfreetype6=2.14.1=h73754d4_0
- libgcc=15.2.0=h767d61c_7
- libgcc-ng=15.2.0=h69a702a_7
- libgfortran=15.2.0=h69a702a_7
- libgfortran5=15.2.0=hcd61629_7
- libgl=1.7.0=ha4b6fd6_2
- libglib=2.86.1=h32235b2_2
- libglvnd=1.7.0=ha4b6fd6_2
- libglx=1.7.0=ha4b6fd6_2
- libgomp=15.2.0=h767d61c_7
- libgoogle-cloud=2.39.0=hdb79228_0
- libgoogle-cloud-storage=2.39.0=hdbdcf42_0
- libgrpc=1.73.1=h3288cfb_1
- libiconv=1.18=h3b78370_2
- libjpeg-turbo=3.1.2=hb03c661_0
- libkvikio=25.10.00=cuda12_9_251008_fb6220c4
```

```
- liblapack=3.9.0=39_h47877c9_openblas
- libllvm21=21.1.5=hf7376ad_0
- liblzma=5.8.1=hb9d3cd8_2
- libnghttp2=1.67.0=had1ee68_0
- libnl=3.11.0=hb9d3cd8_0
- libnsl=2.0.1=hb9d3cd8_1
- libntlm=1.8=hb9d3cd8_0
- libnuma=2.0.18=hb9d3cd8_3
- libnvcomp=5.0.0.6=hb7e823c_3
- libnvcomp-dev=5.0.0.6=hb7e823c_3
- libnvjitlink=12.9.86=hecca717_2
- libnvptxcompiler-dev=12.9.86=ha770c72_2
- libnvptxcompiler-dev_linux-64=12.9.86=ha770c72_2
- libopenblas=0.3.30=pthreads_h94d23a6_4
- libopengl=1.7.0=ha4b6fd6_2
- libopentelemetry-cpp=1.21.0=hb9b0907_1
- libopentelemetry-cpp-headers=1.21.0=ha770c72_1
- libparquet=21.0.0=h31208bf_11_cuda
- libpciaccess=0.18=hb9d3cd8_0
- libpng=1.6.50=h421ea60_1
- libpq=18.0=h3675c94_0
- libprotobuf=6.31.1=h49aed37_2
- libraft=25.10.00=cuda12_251008_521611f8
- libraft-headers=25.10.00=cuda12_251008_521611f8
- libraft-headers-only=25.10.00=cuda12_251008_521611f8
- libre2-11=2025.11.05=h7b12aa8_0
- librmm=25.10.00=cuda12_251008_7aaad1de
- libsodium=1.0.20=h4ab18f5_0
- libsqlite=3.51.0=hee844dc_0
- libssh2=1.11.1=hcf80075_0
- libstdcxx=15.2.0=h8f9b012_7
- libstdcxx-ng=15.2.0=h4852527_7
- libsystemd0=258.2=h6569c3e_1
- libthrift=0.22.0=h454ac66_1
- libtiff=4.7.1=h9d88235_1
- libucxx=0.46.00=cuda12_251008_64355220
- libudev1=258.2=h6569c3e_1
- libutf8proc=2.11.0=hb04c3b8_0
- libuuid=2.41.2=he9a06e4_0
- libvulkan-loader=1.4.328.1=h5279c79_0
- libwebp-base=1.6.0=hd42ef1d_0
- libxcb=1.17.0=h8a09558_0
- libxcrypt=4.4.36=hd590300_1
- libxgboost=3.0.5=rapidsai_h98a37b8_3
- libxkbcommon=1.13.0=hca5e8e5_0
- libxml2=2.15.1=h26afc86_0
- libxml2-16=2.15.1=ha9997c6_0
- libxslt=1.1.43=h711ed8c_1
- libzlib=1.3.1=hb9d3cd8_2
- llvmlite=0.44.0=py311h1741904_2
- locket=1.0.0=pyhd8ed1ab_0
- lz4=4.4.5=py311h1c460e0_0
- lz4-c=1.10.0=h5888daf_1
- markdown-it-py=4.0.0=pyhd8ed1ab_0
- markupsafe=3.0.3=py311h3778330_0
- matplotlib=3.10.8=py311h38be061_0
- matplotlib-base=3.10.8=py311h0f3be63_0
- matplotlib-inline=0.2.1=pyhd8ed1ab_0
- mdurl=0.1.2=pyhd8ed1ab_1
- msgpack-python=1.1.2=py311hdf67eae_1
```

```
- munkres=1.1.4=pyhd8ed1ab_1
- narwhals=2.11.0=pyhcf101f3_0
- nccl=2.28.9.1=h4d09622_0
- ncurses=6.5=h2d0b736_3
- nest-asyncio=1.6.0=pyhd8ed1ab_1
- nlohmann_json=3.12.0=h54a6638_1
- numba=0.61.2=py311h6220fa4_2
- numba-cuda=0.19.1=pyhcf101f3_2
- numpy=2.2.6=py311h5d046bc_0
- nvidia-ml-py=13.580.82=pyhd8ed1ab_0
- nvtx=0.2.13=py311h49ec1c0_1
- openjpeg=2.5.4=h55fea9a_0
- openldap=2.6.10=he970967_0
- openssl=3.6.0=h26f9b46_0
- orc=2.2.1=hd747db4_0
- packaging=25.0=pyh29332c3_1
- pandas=2.3.3=py311hed34c8f_1
- parso=0.8.5=pyhcf101f3_0
- partd=1.4.2=pyhd8ed1ab_0
- patsy=1.0.2=pyhcf101f3_0
- pcre2=10.46=h1321c63_0
- pexpect=4.9.0=pyhd8ed1ab_1
- phik=0.12.5=py311h724c32c_1
- pillow=12.0.0=py311h07c5bb8_0
- pip=25.3=pyh8b19718_0
- pixman=0.46.4=h54a6638_1
- platformdirs=4.5.0=pyhcf101f3_0
- prometheus-cpp=1.3.0=ha5d0236_0
- prompt-toolkit=3.0.52=pyha770c72_0
- psutil=7.1.3=py311haee01d2_0
- pthread-stubs=0.4=hb9d3cd8_1002
- ptyprocess=0.7.0=pyhd8ed1ab_1
- pure_eval=0.2.3=pyhd8ed1ab_1
- py-xgboost=3.0.5=rapidsai_pyh4bd9c9a_3
- pyarrow=21.0.0=py311h38be061_1
- pyarrow-core=21.0.0=py311hbce1db7_1_cuda
- pycparser=2.22=pyh29332c3_1
- pygments=2.19.2=pyhd8ed1ab_0
- pylibcudf=25.10.00=cuda12_py311_251008_f4e35ca0
- pylibraft=25.10.00=cuda12_py311_251008_521611f8
- pyparsing=3.2.5=pyhcf101f3_0
- pyside6=6.9.3=py311he4c1a5a_1
- pysocks=1.7.1=pyha55dd90_7
- python=3.11.14=hd63d673_2_cpython
- python-dateutil=2.9.0.post0=pyhe01879c_2
- python-tzdata=2025.2=pyhd8ed1ab_0
- python_abi=3.11=8_cp311
- pytz=2025.2=pyhd8ed1ab_0
- pyyaml=6.0.3=py311h3778330_0
- pyzmq=27.1.0=py311h2315fbb_0
- qhull=2020.2=h434a139_5
- qt6-main=6.9.3=h5c1c036_1
- raft-dask=25.10.00=cuda12_py311_251008_521611f8
- rapids-dask-dependency=25.10.00=251008_bcfdc48e
- rapids-logger=0.1.1=h98325ef_0
- rdma-core=60.0=hecca717_0
- re2=2025.11.05=h5301d42_0
- readline=8.2=h8c095d6_2
- rich=14.2.0=pyhcf101f3_0
- rmm=25.10.00=cuda12_py311_251008_7aaad1de
```

```
- s2n=1.6.0=h8399546_1
- scikit-learn=1.7.2=py311hc3e1efb_0
- scipy=1.16.3=py311h1e13796_0
- seaborn=0.13.2=hd8ed1ab_3
- seaborn-base=0.13.2=pyhd8ed1ab_3
- setuptools=80.9.0=pyhff2d567_0
- six=1.17.0=pyhe01879c_1
- snappy=1.2.2=h03e3b7b_1
- sortedcontainers=2.4.0=pyhd8ed1ab_1
- stack_data=0.6.3=pyhd8ed1ab_1
- statsmodels=0.14.5=py311h0372a8f_1
- tblib=3.2.2=pyhcf101f3_0
- threadpoolctl=3.6.0=pyhecae5ae_0
- tk=8.6.13=noxft_ha0e22de_103
- toolz=1.1.0=pyhd8ed1ab_1
- tornado=6.5.2=py311h49ec1c0_2
- traitlets=5.14.3=pyhd8ed1ab_1
- treelite=4.4.1=py311hd96400b_1
- typing_extensions=4.15.0=pyhcf101f3_0
- tzdata=2025b=h78e105d_0
- ucx=1.19.0=h63b5c0b_5
- ucxx=0.46.00=cuda12_py311_251008_64355220
- unicodedata2=17.0.0=py311h49ec1c0_1
- urllib3=2.5.0=pyhd8ed1ab_0
- wayland=1.24.0=hd6090a7_1
- wcwidth=0.2.14=pyhd8ed1ab_0
- wheel=0.45.1=pyhd8ed1ab_1
- xcb-util=0.4.1=h4f16b4b_2
- xcb-util-cursor=0.1.5=hb9d3cd8_0
- xcb-util-image=0.4.0=hb711507_2
- xcb-util-keysyms=0.4.1=hb711507_0
- xcb-util-renderutil=0.3.10=hb711507_0
- xcb-util-wm=0.4.2=hb711507_0
- xgboost=3.0.5=rapidsai_pyh13f938f_3
- xkeyboard-config=2.46=hb03c661_0
- xorg-libice=1.1.2=hb9d3cd8_0
- xorg-libsm=1.2.6=he73a12e_0
- xorg-libx11=1.8.12=h4f16b4b_0
- xorg-libxau=1.0.12=hb03c661_1
- xorg-libxcomposite=0.4.6=hb9d3cd8_2
- xorg-libxcursor=1.2.3=hb9d3cd8_0
- xorg-libxdamage=1.1.6=hb9d3cd8_0
- xorg-libxdmcp=1.1.5=hb03c661_1
- xorg-libxext=1.3.6=hb9d3cd8_0
- xorg-libxfixes=6.0.2=hb03c661_0
- xorg-libxi=1.8.2=hb9d3cd8_0
- xorg-libxrandr=1.5.4=hb9d3cd8_0
- xorg-libxrender=0.9.12=hb9d3cd8_0
- xorg-libxtst=1.2.5=hb9d3cd8_3
- xorg-libxxf86vm=1.1.6=hb9d3cd8_0
- xyzservices=2025.10.0=pyhd8ed1ab_0
- yaml=0.2.5=h280c20c_3
- zeromq=4.3.5=h387f397_9
- zict=3.0.0=pyhd8ed1ab_1
- zipp=3.23.0=pyhd8ed1ab_0
- zlib=1.3.1=hb9d3cd8_2
- zlib-ng=2.2.5=hde8ca8f_0
- zstandard=0.25.0=py311haee01d2_1
- zstd=1.5.7=hb8e6e7a_2
- pip:
```

```
      - blessed==1.25.0
      - gpustat==1.1.1
prefix: /home/ramon/miniconda3/envs/tcc_gpu
```