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. 51 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. 51 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
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, características da escola e particularidades regionais no desempenho dos estudantes no ENEM. A pergunta central que guia esta pesquisa é: “Quais são os principais fatores socioeconômicos, características da escola e particularidades regionais 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) e do Censo Escolar (5), 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).

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 (6).

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 (6).

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 (7). 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) (8).

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 etc.; (ii) o cpital 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 conheicmento (8).

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” (9) 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 (9).

- **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 (10).

- **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 (11).

- **Sexo:** Diferenças de desempenho por sexo são notadas, especialmente na prova de Matemática, com vantagem para os homens (até 36 pontos a mais) (11).

- **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 (10).

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 (11).

Achados relevantes incluem:

- **Dependência Administrativa (Pública vs. Privada):** Alunos de escolas privadas consistentemente superam os de escolas públicas (11). Em Matemática, a diferença pode ser de aproximadamente 83,9 pontos entre alunos de escolas privadas e estaduais (10). O 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 (11). 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 (11).

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 (11). 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 (12).

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 (13). 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 (12).

2.6 Aplicações de Ciência de Dados na Análise do ENEM

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, 5, 10, 11, 14, 15)

A análise exploratória de dados e a seleção de *features* são etapas cruciais para o sucesso dessas análises (16). A aplicação de *Machine Learning* permite ir além da correlação estatística, possibilitando a construção de modelos preditivos com alta acurácia (15) e a quantificação da importância relativa de cada fator. Isso é crucial para o objetivo de determinar “quanto que esses fatores influenciam”, oferecendo uma abordagem mais robusta para a tomada de decisão baseada em dados, pois quantifica a contribuição de cada variável para o poder preditivo do modelo, mesmo em relações não lineares.

2.7 Métodos de *Machine Learning*

Essa seção pretende apresentar, de forma não exaustiva, alguns métodos de *Machine Learning* utilizados em trabalhos similares. Foram usadas as referências (17–20) para a fundamentação teórica dos métodos apresentados.

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.

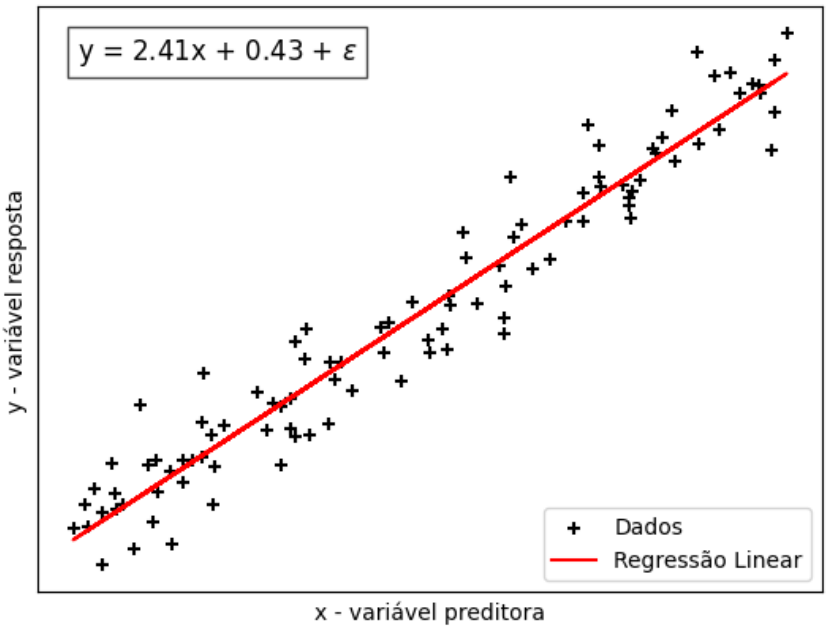
A implementação mais básica é expressa pela equação

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

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 a sua aplicação 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: do autor.

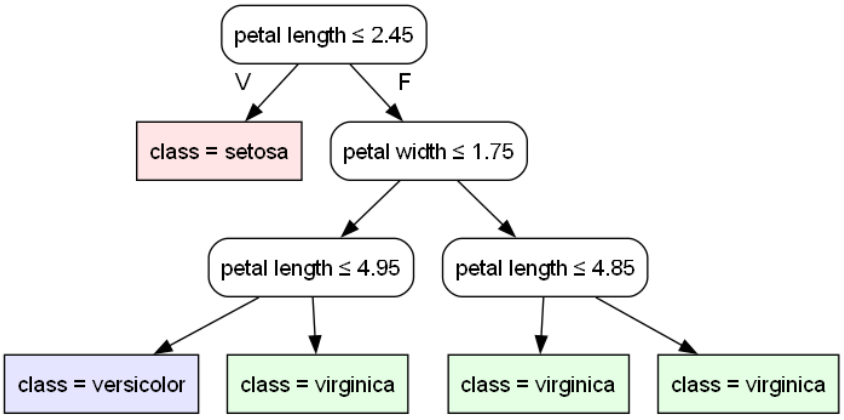
2.7.2 Árvore de Decisão

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 maior 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 feita baseada em alguma medida de impureza, como a Entropia ou o Índice de Gini.

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 pode ser propensa *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: do autor.

2.7.3 *Random Forest*

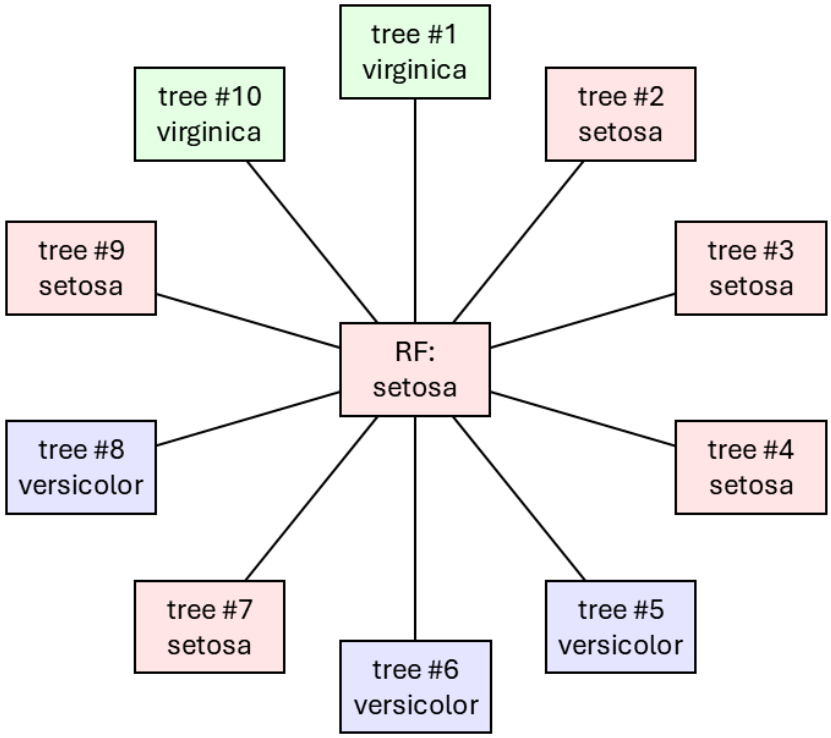
O *Random Forest* é um método derivado da Árvore de Decisão, sendo um dos algoritmos mais populares e eficazes em *Machine Learning*. Ele adota uma abordagem de *ensemble*, ou seja, combina múltiplos modelos para melhorar a precisão e a robustez das previsões. A ideia central é criar uma "floresta" de árvores de decisão, onde a decisão final é feita pela média 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: do 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*.

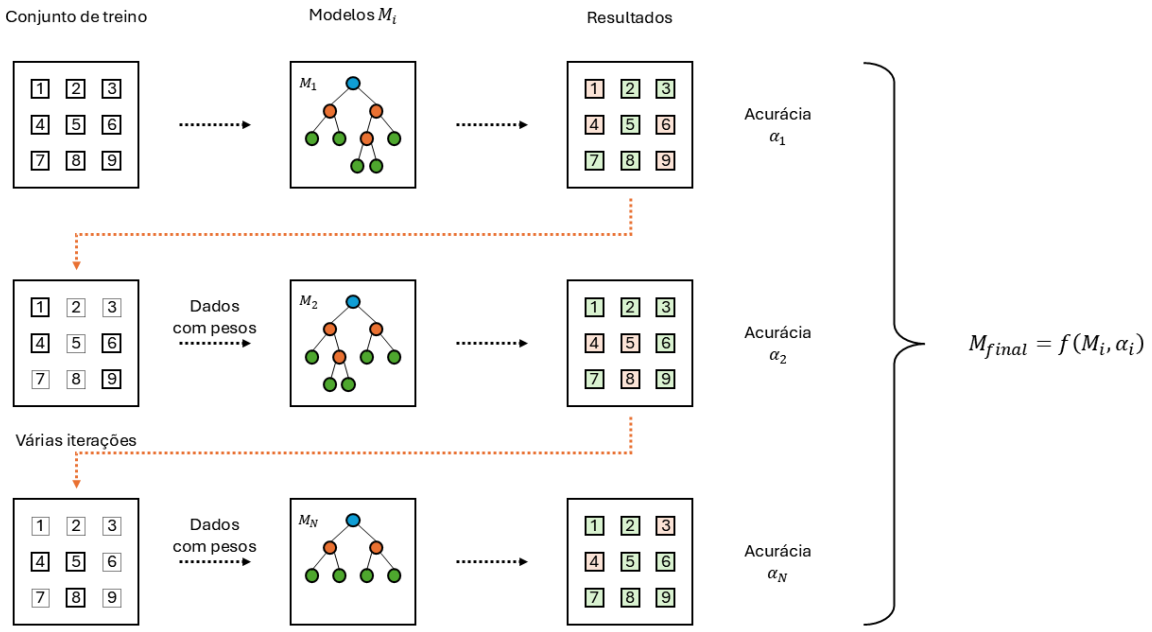
O *AdaBoost* (*Adaptive Boosting*) é um dos primeiros algoritmos de *Boosting* e funciona aumentando o peso dos dados de treinamento que foram classificados incorreta-

mente 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: do autor.

3 METODOLOGIA

Este capítulo detalha o delineamento da pesquisa, a coleta, o processamento e as técnicas analíticas empregadas para responder às perguntas de pesquisa.

3.1 Tipo de Pesquisa e Abordagem

A pesquisa será de natureza quantitativa, com caráter exploratório e descritivo, incorporando elementos de modelagem preditiva. A escolha dessa abordagem justifica-se pela necessidade de identificar e, crucialmente, quantificar a influência de diversos fatores no desempenho dos estudantes no ENEM. A abordagem quantitativa permite a análise de grandes volumes de dados, a identificação de padrões estatisticamente significativos e a construção de modelos que podem prever resultados e mensurar o impacto das variáveis.

3.2 Fonte e Coleta de Dados

A fonte primária de dados para este estudo serão os microdados do ENEM, disponibilizados publicamente pelo INEP. (4) Serão utilizados os dados dos últimos 3 exames: 2022 a 2024. Usarei esse histórico para afastar quaisquer efeitos vindo da pandemia de COVID-19 e poder construir um comparativo entre dos efeitos influenciadores em cada edição.

Para enriquecer as variáveis de características escolares e regionais, faremos a integração dos microdados do ENEM com outras bases de dados públicas, como o Censo Escolar. (5) A integração de dados de diferentes fontes é um passo metodológico avançado que permite uma análise mais holística.

Ao combinarmos informações sobre o aluno (do ENEM) e a escola (do Censo Escolar), é possível desvendar interações complexas e o "Efeito Escola" de forma mais granular, superando as limitações de analisar apenas um conjunto de dados. Isso proporciona uma compreensão mais completa do ambiente educacional, incluindo infraestrutura, qualificação docente e práticas de gestão, que complementam os dados de desempenho individual.

4 CONCLUSÃO

xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx

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. **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.

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

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

8 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.

9 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.

10 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.

11 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.

12 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.

13 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/.

14 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.

15 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.

16 DUTRA, J. F.; FERNANDES, D. Y. d. S.; JUNIOR, J. B. F. Fatores que podem interferir no desempenho de estudantes no enem: uma revisão sistemática da literatura. **Revista Brasileira de Informática na Educação**, jul 2023. Disponível em: http://br-ie.org/pub/index.php/rbie.

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

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

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

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

APÊNDICES

APÊNDICE A – EXEMPLO 1

APÊNDICE B – EXEMPLO 2

APÊNDICE C – EXEMPLO 3