# Relatório Analítico : ENEM por Escolas

Guilherme Fogaça[1], Milena Churata[2]
[1]Departamento Acadêmico de Computação (DACOM)
Universidade Tecnológica Federal do Paraná (UTFPR)

**Abstract**

This article details the data analysis process related to ENEM by schools. It expands on the variables that permeate the exam process in order to find correlations between reality and the available dataset.

**Resumo**

O presente artigo detalha o processo de análise de dados referente ao ENEM por escolas. Expandindo as variáveis que permeiam o processo da prova, a fim de encontrar correlações entre a realidade e o conjunto de dados disponível.

## 1 Introdução e Exaplanação do Problema

O ENEM ( Exame Nacional do Ensino Médio), principal prova do sistema de educação brasileiro, é um possível termômetro para análise de métricas referente ao nível de ensino. Possibilitando adentrar a programas governamentais, que disponibilizam vagas diretas para as principais faculdades do país. Contudo, as escolas possuem diferentes subordinações, sendo elas Federais, Estaduais e Municipais, além do setor privado. Essa pluralidade, permite diferentes abordagens de ensino, administração e alocação orçamentária.

### 1.1 Perguntas e Hipótese de Pesquisa

Dado o contexto, as perguntas inferidas a esse trabalho são as seguintes:

- **P1 (Gestão e Localização):** Como o tipo de administração (Pública/Privada) e a localização (Urbana/Rural) influenciam a média do ENEM?
- **P2 (Porte):** O porte da escola (tamanho) está associado ao desempenho médio, independentemente da dependência administrativa?
- **H1:** Escolas privadas e federais apresentam médias mais altas nas provas do ENEM do que escolas estaduais e municipais.
- **H2:** Escolas urbanas obtêm desempenhos superiores às rurais em todas as áreas do ENEM.
- **H3:** Existe uma correlação positiva entre o porte da escola (PORTE_ESCOLA) e as notas médias do ENEM, indicando que escolas maiores tendem a ter melhor desempenho.

## 2 Esqueleto e Limpeza de Dados

O Dataset comprime mais de 172.305 registros de todas as escolas presentes nos vinte e sete estados brasileiros. Os dados são inicialmente datados em 2005 e se estende até 2015. Ao todo são 27 variáveis, a respeito de médias em sessões da prova, porte, estado, taxa de aprovação entre outros campos que caracterizam a escola.

**Corresponding author(s):** Guilherme Fogaça & Milena Churata, Universidade Tecnológica Federal do Paraná (UTFPR)

## 2.1 Contexto de Limpeza

No processo de limpeza, foi constado um alto volume de escolas com dados faltantes. Em avaliação geral, foi verificado alta concentração em valores anteriores ao ano de 2009. Para identificar possíveis causas, em pesquisa, no ano de 2009, o ENEM passou por uma reestruturação, onde a prova que inicialmente era concebida por 2 grandes blocos, foi substituída pela qual como é hoje. Portanto, dados como médias nos âmbitos de ciências humanas, matemática entre outros estão completamente vazios, desta forma, os dados anteriores a 2009, não participaram da análise. Com o script para verificar valores, constatou as seguintes reproduções conforme a Figura 1.

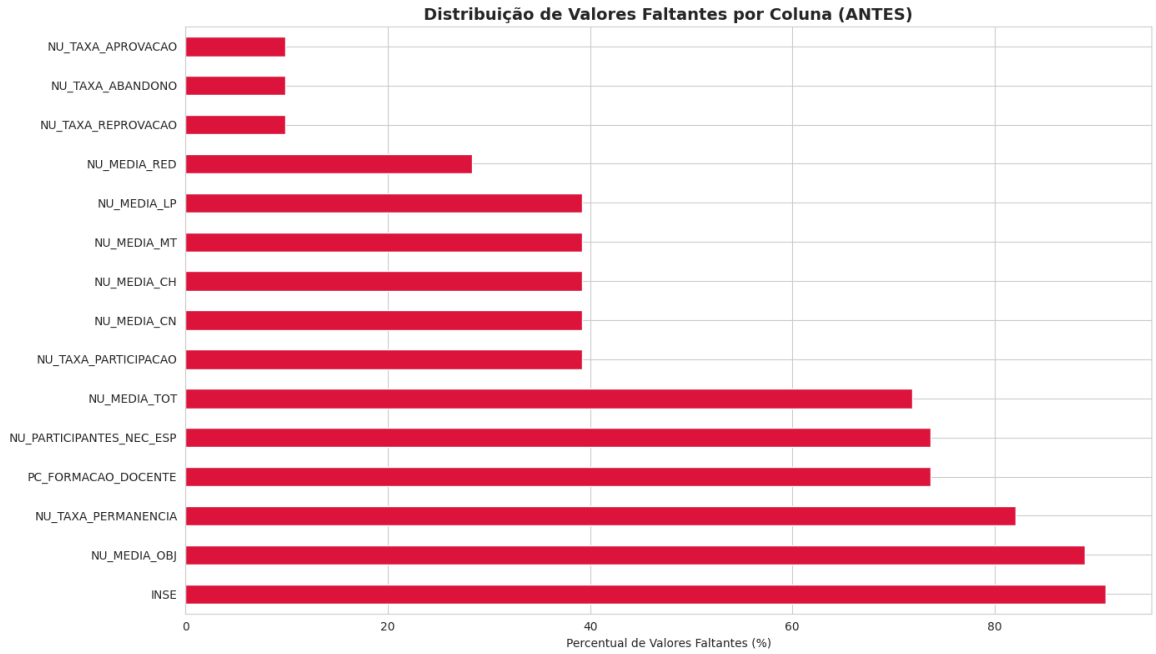


Figura 1: Quantidade de Dados Faltantes

Campos com altos missing values foram prontamente descartados, uma vez ao analisar o impacto para o problema desse artigo, foi categorizado como prescindível. Resultando em 67.618 dados removidos, 16 variáveis viáveis e apenas 0.4% de valores faltantes. Segue a diferença na representação gráfica na Figura 2.

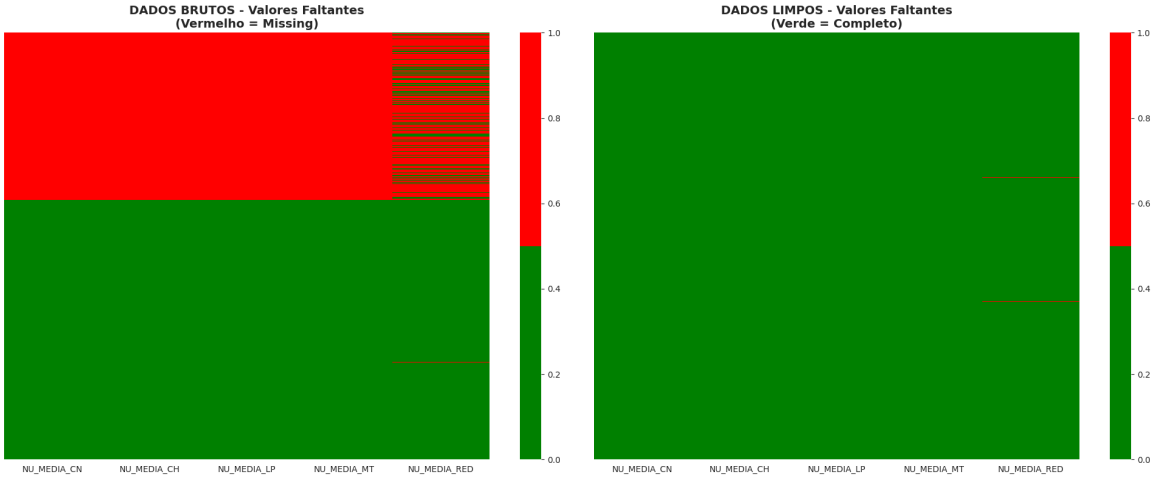


Figura 2: Antes x Depois Missing Values

### 2.2 Outliers

Após o procedimento de limpeza e análise inicial, para verificar integridade e confiabilidade dos dados, que deu resultado no agora Dataset limpo, foi calculado a métrica Outlier, em que no seu caso mais critico evidenciou 1.4% de variação. Segue a representação na Figura 3.

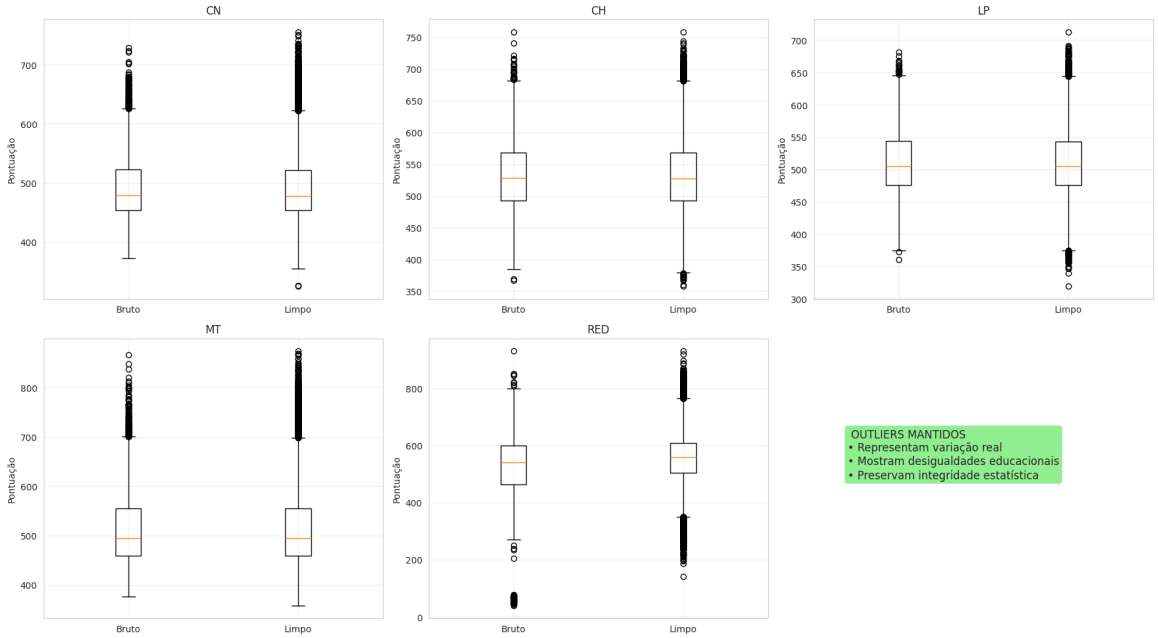


Figura 3: Métrica Outlier

O conjunto de dados resultante, foi categorizado como confiável, uma vez que sua variância demonstrou um comportamento natural, representando fielmente em dados a possível realidade.

## 3 Análise Exploratórias

Com os processos previamente listados, o Dataset foi transformado em Tidy Data, além da exportação para formato Parquet. Com dados prontos para análise, executou-se a integração junto a consultas SQL com o auxílio do DuckDB para possíveis insights.

### 3.1 Dependência Administrativa

A princípio foi executado uma exploração em sentido a dependência administrativa, visando buscar diferença em possíveis abordagens de ensino e gestões. Com os dados, é possível ver clara discrepância na variação pública, onde enquanto a Federal navega próximo a 600 e os demais setores públicos como Municipal e Estadual lutam para alcançar 500. O setor privado é o mais próximo da Federal. Evidenciando assim, uma clara diferença de modelo, ligado ao críterio administrativo.

Tabela 1: Média das Notas por Dependência Administrativa

| Tipo | CN | CH | LP | MT | Redação |
|---|---|---|---|---|---|
| Federal | 543.14 | 588.18 | 555.00 | 586.35 | 620.24 |
| Privada | 537.12 | 577.19 | 552.48 | 576.55 | 613.81 |
| Municipal | 480.13 | 518.66 | 503.28 | 500.94 | 555.90 |
| Estadual | 461.68 | 502.90 | 483.08 | 472.67 | 520.87 |

## 3.2 Análise Temporal

Outra análise executada, é em sentido a possível evolução nas áreas de conhecimento ao decorrer dos anos. A priori, nenhuma discrepância que evidencie algum comportamento além do natural.

Tabela 2: Evolução Anual das Médias

| Ano | CN | CH | LP | MT | Redação |
|---|---|---|---|---|---|
| 2009 | 496.60 | 496.66 | 495.14 | 495.71 | 579.91 |
| 2010 | 481.82 | 534.53 | 505.92 | 504.59 | 591.03 |
| 2011 | 489.48 | 494.46 | 537.22 | 555.24 | 559.45 |
| 2012 | 491.43 | 536.72 | 505.01 | 540.99 | 533.04 |
| 2013 | 488.30 | 529.77 | 502.03 | 535.44 | 537.90 |
| 2014 | 498.74 | 555.98 | 520.02 | 495.66 | 515.40 |
| 2015 | 490.87 | 566.76 | 515.31 | 492.57 | 563.77 |

## 3.3 Média Geral por Estado

A fim de reconhecer alguma dependência geográfica, uma análise em virtude da média de cada estado foi elencado. Como resultado, médias maiores reconhecidas na região sul/suldeste e distrito federal, enquanto, menores localizadas mais na região norte/nordeste.

Tabela 3: Rank dos Estados - Parte 1

| Rank | Estado | Média |
|---|---|---|
| 1 | DF | 544.15 |
| 2 | RJ | 543.11 |
| 3 | SP | 540.82 |
| 4 | MG | 532.38 |
| 5 | RS | 525.56 |
| 6 | SC | 523.94 |
| 7 | PR | 518.32 |
| 8 | BA | 513.92 |
| 9 | GO | 511.17 |
| 10 | PE | 507.38 |
| 11 | MS | 507.00 |
| 12 | ES | 504.28 |
| 13 | AL | 503.01 |
| 14 | PA | 502.87 |

Tabela 4: Rank dos Estados - Parte 2

| Rank | Estado | Média |
|---|---|---|
| 15 | RN | 501.10 |
| 16 | SE | 498.18 |
| 17 | MT | 496.99 |
| 18 | PB | 496.91 |
| 19 | CE | 493.88 |
| 20 | RO | 492.64 |
| 21 | PI | 490.49 |
| 22 | AM | 485.76 |
| 23 | RR | 485.64 |
| 24 | MA | 483.52 |
| 25 | AP | 481.93 |
| 26 | AC | 479.06 |
| 27 | TO | 475.79 |

## 3.4 Análise Univariada, Bivariada e Multivariada

- **Univariada:** Distribuições, medidas de tendência central e dispersão, identificação de padrões e valores atípicos.
- **Bivariada:** Correlações entre variáveis, visualizações de relacionamentos, comparações entre grupos.
- **Multivaraida:** Matriz de correlações, análises de agrupamento, técnicas de redução dimensional .

### 3.4.1 Distribuição Geral das Notas - Univariada

Com as distribuições, diferenças claras entre as áreas do ENEM é identificada, notas de Linguagens e Redação tendem a ser mais altas e menos dispersas, enquanto Matemática apresenta maior variabilidade e medianas mais baixas.

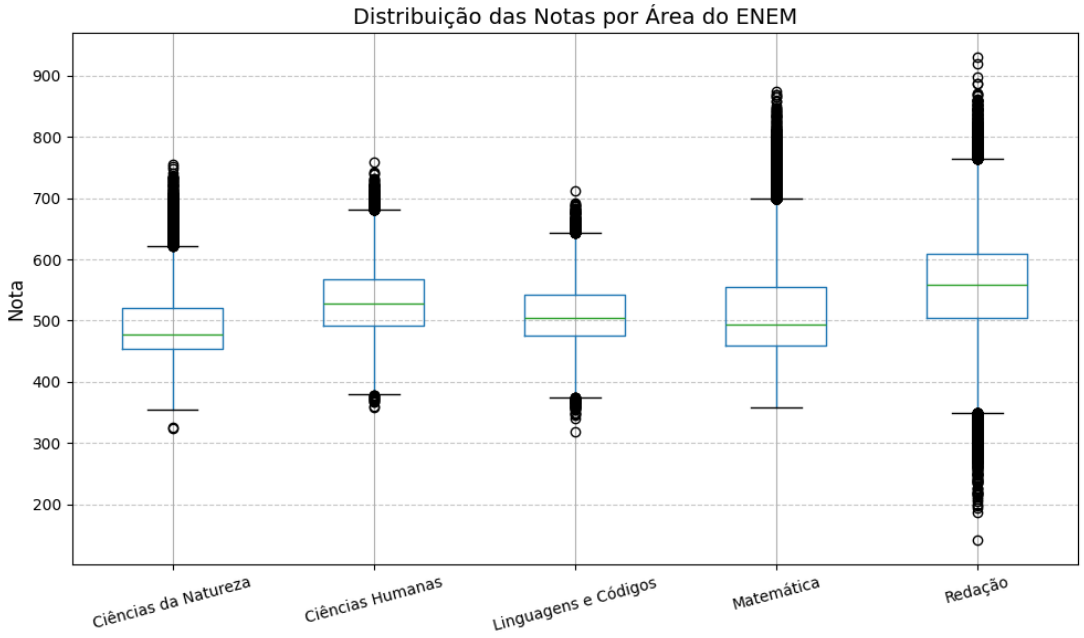


Figura 4: Distribuição por Área

### 3.4.2 Média por Tipo de Escola e Localização - Bivariada

Com os dados, escolas privadas e federais apresentam, em média, notas mais altas que as estaduais e municipais. E escolas localizadas no perímetro urbano superam consistentemente as rurais, independentemente de seu tipo administrativo. Reforçando desigualdades estruturais entre redes e contextos geográficos de ensino.

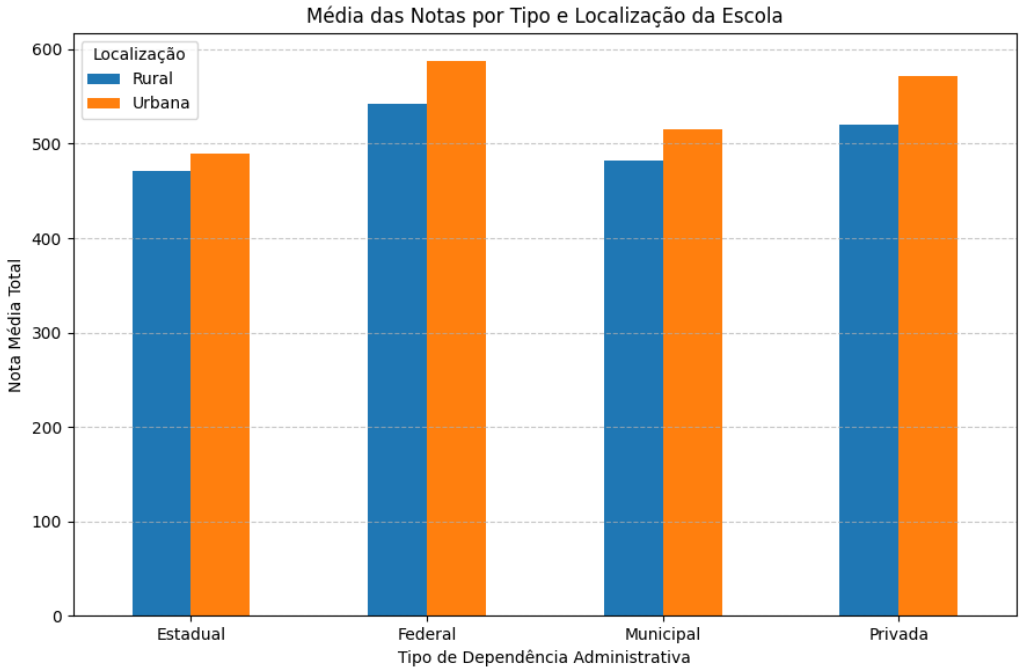


Figura 5: Tipo e Localização

### 3.4.3 Porte da Escola e Correlação com Desempenho Médio - Bivariada

Os dados explicitam, uma tendência de aumento da nota média conforme o porte da escola cresce, principalmente nas redes Federal e Privada. Sugerindo que diferentes tipos gestão, onde que com escolas maiores podem prover melhor infraestrutura, corpo docente e recursos, impactando diretamente no desempenho médio dos alunos.

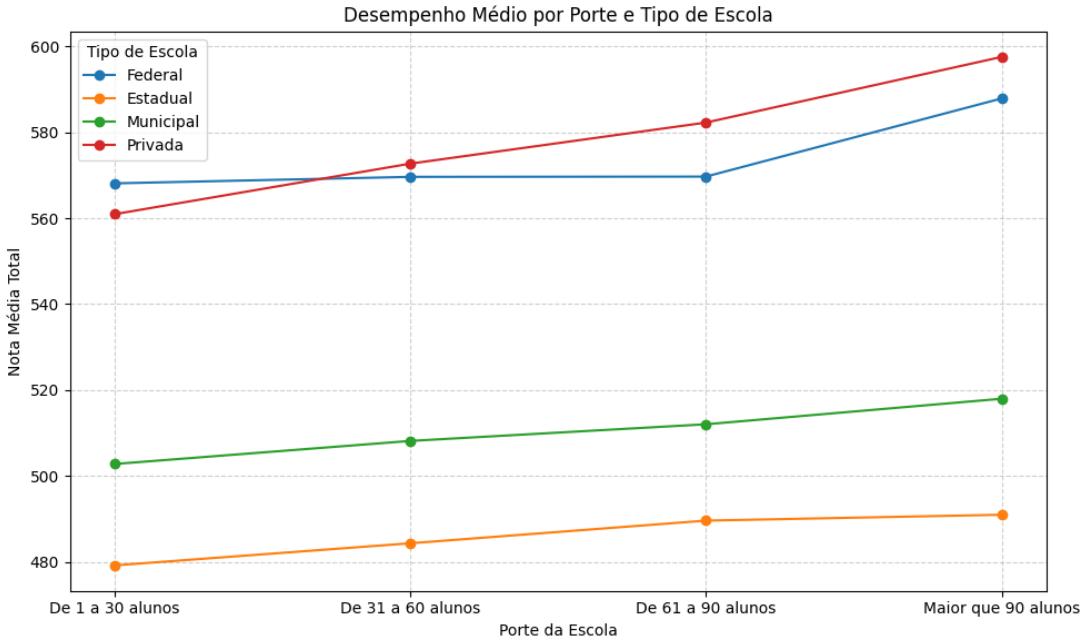


Figura 6: Porte x Tipo

### 3.4.4 Correlação de Áreas - Multivariada

Existem correlações fortes e positivas entre todas as áreas, principalmente entre Linguagens e Redação. Indicando que alunos com bom desempenho tendem a se destacar em múltiplas áreas, inferindo em habilidades gerais de leitura, interpretação e raciocínio.

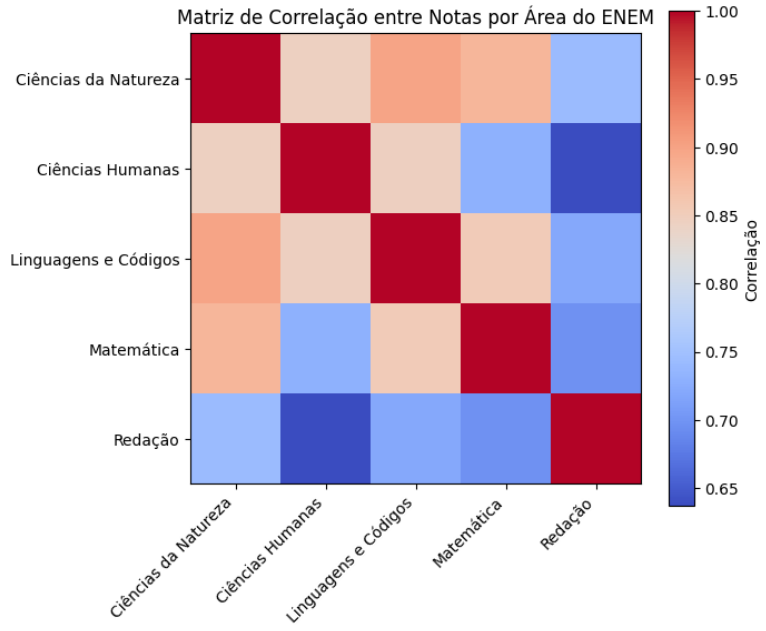


Figura 7: Matriz de Correlação

# 4 Teste de Hipótese

Tabela 5: Resultados dos Testes de Hipóteses Estatísticos

| Teste | Hipóteses | Estatística | p-valor | Conclusão |
|---|---|---|---|---|
| **Tipo de Escola** | $H_0$: Médias iguais<br>$H_1$: Privada/Federal > | t = 315,63 | < 0,001 | **Rejeita Ho** |
| **Localização** | $H_0$: Desempenho igual<br>$H_1$: Urbana > Rural | t = 55,35 | < 0,001 | **Rejeita Ho** |
| **Porte vs Notas** | $H_0$: Sem correlação<br>$H_1$: Correlação positiva | $\rho$= -0,257 | < 0,001 | **Não rejeita Ho** |

# 5 Modelos Preditivos e Análise de Performance

Com base nas análises exploratórias e testes de hipótese realizados, foi dado continuidade com a modelagem preditiva para estimar o desempenho médio das escolas no ENEM. Foram explorados três modelos com diferentes níveis de complexidade, permitindo uma avaliação comparativa do ganho preditivo.

## 5.1 Regressão Linear: Modelo Baseline

Com início na Regressão Linear como nosso Modelo Baseline. O objetivo de um baseline é estabelecer um ponto de referência simples e rápido, para podermos medir o ganho de complexidade ao usar modelos mais avançados.

- **$R^2$ = 0.660 (66.0%):** O modelo de Regressão Linear conseguiu explicar 66% da variação na nota média das escolas.
- **MSE = 1082.84:** O Erro Quadrático Médio é de 1082.84.

**Conclusão:** Já é promissor, um $R^2$ de 0.660 sugere que existe uma relação linear significativa entre as características da escola e o desempenho, confirmando que os dados que escolhemos são altamente preditivos.

## 5.2 Random Forest: Modelo Complexo

Dando sequência, testamos um modelo de ensemble mais sofisticado, o Random Forest. Tivemos que otimizar os hiperparâmetros como *n_estimators*, *max_depth*, entre outros. O Random Forest é computacionalmente mais custoso e, devido ao tamanho do nosso dataset, reduzimos a complexidade para que o treinamento fosse viável no ambiente que utilizamos, que para esse artigo foi o Google Colab.

- **$R^2$ = 0.720 (72.0%):** O Random Forest melhorou a performance, explicando 72% da variação total.
- **MSE = 891.99:** O erro diminuiu significativamente (de 1082.84 para 891.99).

**Ganhos:** Com a melhoria de 6 pontos percentuais no $R^2$ em relação ao baseline prova que as relações no nosso dataset são não-lineares e que modelos mais complexos capturam melhor esses padrões.

### 5.3 HistGradientBoostingRegressor: Modelo Otimizado

E por fim, implementamos o HistGradientBoostingRegressor (HGB), um modelo de boosting otimizado. Modelos de boosting constroem árvores de decisão sequencialmente, corrigindo os erros das árvores anteriores.

- **$R^2$ = 0.733 (73.3%):** O HGB alcançou a melhor performance de todos, explicando 73.3% da variação na nota média.
- **MSE = 851.15:** O menor erro de todos.

**Vantagem:** O HGB não apenas superou o Random Forest em precisão ($R^2$ de 0.733 vs. 0.720), mas se demonstrou muito eficiente em termos de tempo de treinamento em grandes conjuntos de dados, o que é um ponto crucial para a nossa aplicação.

### 5.4 Conclusão dos Modelos Preditivos

O HistGradientBoostingRegressor é nosso modelo de produção. Ele demonstrou o maior poder preditivo, provando que é o mais adequado para estimar o desempenho de uma escola com base em suas características administrativas, geográficas e de porte. A evolução do $R^2$ de 66.0% para 73.3% entre os modelos confirma que relações não-lineares estão presentes nos dados e que algoritmos mais sofisticados conseguem obter essas complexidades de forma eficiente.

Tabela 6: Comparação de Performance dos Modelos Preditivos

| Modelo | $R^2$ Score | MSE | Ganho em $R^2$ |
|---|---|---|---|
| Regressão Linear | 0.660 | 1082.84 | — |
| Random Forest | 0.720 | 891.99 | +6.0% |
| HistGradientBoosting | 0.733 | 851.15 | +7.3% |

## 6 Trabalhos Futuros

- Expandir a análise incluindo variáveis socioeconômicas dos estudantes
- Desenvolver um sistema de predição com novos dados, buscando features como nível de formação de professores.

## 7 Conclusão

Este estudo demonstra que fatores administrativos, geográficos e de porte escolar são determinantes no desempenho do ENEM. As análises realizadas fornecem uma visão valiosos para tomadores de decisão no setor educacional, enquanto os modelos desenvolvidos oferecem ferramentas práticas para predição. Os resultados reforçam a necessidade de políticas educacionais direcionadas para o sentido correto, com base na análise robusta de dados, obtendo maior assertividade de medidas e tomada de decisões. Constatando achismos em dados.

## Referências

learn Developers, Scikit. 2023a. Histogram-based gradient boosting. Documentação oficial do HistGradientBoostingRegressor. https://scikit-learn.org/stable/modules/ensemble.html#histogram-based-gradient-boosting.

learn Developers, Scikit. 2023b. Scikit-learn: Machine learning in python. Documentação oficial da biblioteca scikit-learn. https://scikit-learn.org/stable/.

GOV, INEP. 2020. Enem por escola. https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem-por-escola.

Science, Towards Data. 2023a. Evaluation metrics for regression models. Métricas de avaliação para modelos de regressão. https://towardsdatascience.com/evaluation-metrics-for-regression-models-77e85b4b5e5e.

Science, Towards Data. 2023b. Gradient boosting explained. Explicação sobre Gradient Boosting. https://towardsdatascience.com/gradient-boosting-explained-9d8a4e8c9c45.

Science, Towards Data. 2023c. Linear regression for machine learning. Explicação detalhada sobre Regressão Linear. https://towardsdatascience.com/linear-regression-for-machine-learning-1cdd64a0b7b3.

Science, Towards Data. 2023d. Understanding random forests. Guia completo sobre Random Forest. https://towardsdatascience.com/understanding-random-forest-58381e0602d2.