SEVEN
publicações acadêmicas

APLICAÇÃO DE TÉCNICAS DE APRENDIZADO DE MÁQUINA PARA CLASSIFICAÇÃO DO DESEMPENHO ACADÊMICO NO ENEM: UM ESTUDO COM DADOS DO MARANHÃO

APPLICATION OF MACHINE LEARNING TECHNIQUES TO CLASSIFY ACADEMIC PERFORMANCE IN ENEM: A STUDY WITH DATA FROM MARANHÃO

APLICACIÓN DE TÉCNICAS DE APRENDIZAJE AUTOMÁTICO PARA CLASIFICAR EL RENDIMIENTO ACADÉMICO EN LA ENEM: UN ESTUDIO CON DATOS DE MARANHÃO

https://doi.org/10.56238/sevened2025.030-001

Ernandes Guedes Moura
Doutor em Estatística e Experimentação Agropecuária
Instituição: Instituto Federal de Educação, Ciência e Tecnologia do Piauí
Endereço: Uruçuí – Piauí, Brasil
E-mail: ernandes.guedes@ifipi.edu.br
Orcid: https://orcid.org/0000-0003-2057-5736

Hedley Lima Cunha
Graduando em Análise de Dados e Inteligência Artificial
Instituição: Universidade Federal do Maranhão
Endereço: Fortaleza dos Nogueiras – Maranhão, Brasil
E-mail: hedley@hlc.dev.br

Bruno Roberto Silva de Moraes
Doutorando em Ciência da Computação
Instituição: Universidade Federal do Maranhão Endereço: São Luís – Maranhão, Brasil
E-mail: brs.moraes@ufma.br
Orcid: https://orcid.org/0000-0002-1040-7790

RESUMO
Este estudo tem como objetivo aplicar e comparar técnicas de aprendizado de máquina para classificar o desempenho dos estudantes maranhenses no Exame Nacional do Ensino Médio (ENEM) 2023, com foco na prova de Matemática. A partir de uma base com mais de 112 mil participantes, foram utilizados três modelos preditivos: Regressão Logística, Random Forest e XGBoost. Após o pré-processamento dos dados e a conversão das variáveis categóricas, os modelos foram treinados para classificar os estudantes em duas categorias: bom e ruim, com base em uma nota de corte de 500 pontos. A análise revelou padrões regionais de desigualdade educacional e destacou variáveis socioeconômicas como renda e idade como os principais preditores do desempenho. A Regressão Logística obteve maior acurácia, enquanto o XGBoost apresentou melhor equilíbrio entre precisão e sensibilidade. Os resultados reforçam a utilidade do aprendizado de máquina para análises educacionais e fornecem subsídios importantes para políticas públicas voltadas à melhoria da equidade e da qualidade da educação no estado do Maranhão.

Palavras-chave: Aprendizado de máquina. Análise de dados educacionais. ENEM. Educação no Maranhão. Desempenho acadêmico.

ABSTRACT
This study aims to apply and compare machine learning techniques to classify the performance of students from Maranhão in the 2023 National High School Exam (ENEM), with a focus on the Mathematics test. Using a dataset with over 112,000 participants, three predictive models were employed: Logistic Regression, Random Forest, and XGBoost. After data preprocessing and conversion of categorical variables, the models were trained to classify students into two categories: good and poor, based on a cutoff score of 500 points. The analysis revealed regional patterns of educational inequality and highlighted socioeconomic variables such as income and age as the main predictors of performance. Logistic Regression achieved the highest accuracy, while XGBoost demonstrated a better balance between precision and recall. The results reinforce the usefulness of machine learning for educational analysis and provide important support for public policies aimed at improving equity and the quality of education in the state of Maranhão.

Keyword: Machine learning. Educational data analysis. ENEM. Education in Maranhão. Academic performance.

RESUMEN
Este estudio tiene como objetivo aplicar y comparar técnicas de aprendizaje automático para clasificar el rendimiento de los estudiantes de Maranhão en el Examen Nacional de Enseñanza Media (ENEM) 2023, centrándose en la prueba de Matemáticas. Se utilizaron tres modelos predictivos a partir de una base de datos de más de 112.000 participantes: Regresión Logística, Random Forest y XGBoost. Tras preprocesar los datos y convertir las variables categóricas, los modelos se entrenaron para clasificar a los estudiantes en dos categorías: buenos y malos, basándose en una puntuación de corte de 500 puntos. El análisis reveló patrones regionales de desigualdad educativa y destacó variables socioeconómicas como la renta y la edad como principales predictores del rendimiento. La regresión logística logró una mayor precisión, mientras que XGBoost mostró un mejor equilibrio entre precisión y sensibilidad. Los resultados refuerzan la utilidad del aprendizaje automático para los análisis educativos y proporcionan información importante para las políticas públicas destinadas a mejorar la equidad y la calidad de la educación en el estado de Maranhão.

Palabras clave: Aprendizaje automático. Análisis de datos educativos. ENEM. Educación en Maranhão. Rendimiento académico.

1 INTRODUÇÃO

O Exame Nacional do Ensino Médio (ENEM), criado em 1998 pela Portaria nº 438 do Ministério da Educação (MEC), consolidou-se como o principal instrumento de avaliação da educação básica no Brasil. Inicialmente, seu objetivo era oferecer aos estudantes uma ferramenta de autoavaliação e apoio à inserção no mundo do trabalho. Com o passar do tempo, o exame ampliou sua função, tornando-se uma peça central nos processos de acesso ao ensino superior, sendo utilizado em programas como o Sistema de Seleção Unificada (SISU), o Programa Universidade para Todos (ProUni) e o Fundo de Financiamento Estudantil (FIES) (Brasil, 1998; De Sousa Fernandes et al., 2023).

A reformulação promovida pela Portaria nº 462/2009 consolidou o modelo de avaliação com base em competências e habilidades, alinhando o ENEM a uma proposta de formação mais crítica e cidadã (Brasil, 2009). Desde 2014, seus resultados passaram a ser utilizados como critério de ingresso em instituições de ensino superior de Portugal, por meio de acordos firmados com o INEP, ampliando sua relevância internacional (INEP, 2024).

Nesse contexto, o ENEM passou a exercer não apenas o papel de avaliação acadêmica, mas também tornou-se um instrumento relevante de inclusão social e de compreensão das desigualdades educacionais no Brasil. Diversos estudos têm apontado que fatores como renda familiar, escolaridade dos pais, tipo de escola frequentada e até mesmo a região onde o estudante vive exercem forte influência sobre seu desempenho nas provas (Silva, 2013; De Sousa Fernandes et al., 2023; Lima et al., 2019). Trata-se de uma relação complexa entre condições socioeconômicas e oportunidades educacionais, amplamente discutida na literatura, e que reforça a urgência de políticas públicas mais justas e baseadas em evidências.

Nos últimos anos, novas abordagens vêm sendo exploradas por pesquisadores da área da educação. Entre elas, o uso de técnicas de aprendizado de máquina (Machine Learning) tem ganhado destaque, especialmente por sua capacidade de analisar grandes volumes de dados e revelar padrões que muitas vezes passam despercebidos. Estudos recentes demonstram, por exemplo, que modelos como Regressão Logística, Random Forest e XGBoost (Ziegel, 2003; Breiman, 2001; Chen e Guestrin, 2016) têm sido aplicados com sucesso na previsão do desempenho estudantil, contribuindo para o desenvolvimento de estratégias pedagógicas mais personalizadas e eficazes (Teixeira e Cavique, 2023; Lopez-Arevalo et al., 2020).

Diante disso, este trabalho tem como objetivo comparar diferentes algoritmos de aprendizado de máquina na tarefa de classificação do desempenho dos estudantes do estado do Maranhão na prova de Matemática do ENEM 2023, categorizando-os entre “bom” e “ruim”, com base em uma nota de corte de 500 pontos. Para tanto, foram aplicados três modelos: a Regressão Logística, como abordagem tradicional em tarefas classificatórias; o Random Forest, representando os métodos de bagging (uma

classe de modelo de aprendizado de máquinas que funciona criando várias versões de um modelo utilizando amostras aleatórias (bootstrap) dos dados originais e, em seguida, combina as previsões desses modelos por média para regressão e por votação majoritária para classificação para produzir uma previsão final).

Adicionalmente, foi empregado o XGBoost, caracterizando os modelos baseados em boosting, uma técnica que visa aprimorar o desempenho preditivo ao combinar vários modelos fracos de forma sequencial, de modo que cada novo modelo corrige os erros cometidos pelos anteriores, gerando assim um conjunto de alta precisão. A escolha desses algoritmos visa avaliar qual abordagem oferece melhor desempenho preditivo diante das características da base analisada, além de identificar as variáveis mais relevantes para a explicação das desigualdades educacionais observadas e mapear espacialmente os municípios maranhenses com os melhores e piores desempenhos, fornecendo subsídios para políticas públicas educacionais mais direcionadas.

A escolha da nota 500 como nota de corte para classificação fundamenta-se no fato de que a pontuação máxima da prova é 1000, e a média geral dos estudantes foi de 482, o que permite aproximar para 500 como um valor representativo da metade da escala total. Trabalho anterior, como o de De Sousa Fernandes et al. (2023), utilizaram critérios similares para classificação de alunos com base em dados do ENEM.

Este artigo está estruturado da seguinte maneira: Introdução (Seção 1), Seção 2, são apresentados os dados analisados e os procedimentos metodológicos adotados, incluindo etapas como o pré-processamento, a seleção das variáveis e a aplicação dos modelos de aprendizado de máquina. A Seção 3 traz os principais resultados obtidos, acompanhados de análises descritivas, comparativas e geográficas, bem como a discussão sobre o desempenho dos modelos e a relevância das variáveis preditoras. Por fim, a Seção 4 reúne as considerações finais, nas quais são destacadas as contribuições do estudo, suas limitações e possíveis caminhos para investigações futuras.

2 MATERIAIS E MÉTODOS

Nesta seção são apresentados os materiais utilizados e a metodologia aplicada para o desenvolvimento deste estudo. Descrevem-se os procedimentos de coleta, tratamento e análise de dados, bem como os métodos de aprendizado de máquina empregados.

2.1 CONJUNTO DE DADOS E PRÉ-PROCESSAMENTO

Os dados analisados neste estudo foram obtidos da base de dados disponibilizada pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep) (https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enem) e complementados por informações geográficas processadas com o uso da biblioteca Geopandas e da ferramenta geobr (Jordahl, 2021;

Pereira et al., 2019). O estudo concentra-se nos candidatos que participaram do exame em todo o estado do Maranhão no ano de 2023, com as coordenadas geográficas dos municípios calculadas e projetadas para análises espaciais (construção de mapas).

O conjunto de dados inicial continha 165.756 linhas (candidatos) e 76 colunas (características). Inicialmente, 33 colunas (como 'NU_INSCRICAO', 'NU_ANO', 'CO_MUNICIPIO_PROVA', etc.) foram removidas por serem consideradas irrelevantes para o processo de classificação. Em seguida, os dados passaram por um pré-processamento para eliminar redundâncias e inconsistências causadas pela ausência de informações. Adicionalmente, quatro colunas foram excluídas devido a apresentarem aproximadamente 50% de valores ausentes. Após a remoção de linhas com dados faltantes, resultou-se em 112.482 indivíduos e 39 variáveis de interesse.

Para preparar os dados para as análises de classificação, foi necessário transformar diversas variáveis. Variáveis ordinais, como níveis de escolaridade (Q001) e categorias de renda (Q006), foram convertidas em valores numéricos utilizando a técnica de codificação ordinal (Ordinal Encoding). Conforme Teixeira e Cavique (2023), essa abordagem é adequada, pois preserva a ordem natural das categorias, representando de forma precisa a hierarquia das variáveis. Além dessas, foram transformadas variáveis como 'TP_FAIXA_ETARIA', 'TP_ANO_CONCLUIU', 'Q002', 'Q005', 'Q008', 'Q009', 'Q010', 'Q011', 'Q012', 'Q013', 'Q014', 'Q015', 'Q016', 'Q017', 'Q019', 'Q022' e 'Q024', garantindo que todas fossem tratadas de maneira apropriada para o modelo de classificação.

Além disso, variáveis nominais, como "TP_SEXO" e "TP_ESTADO_CIVIL", foram transformadas em variáveis binárias utilizando a técnica de codificação One-Hot Encoding. Conforme destacado por Lopez-Arevalo et al. (2020), esse método converte dados categóricos em formato numérico, aumentando a dimensionalidade do conjunto de dados e evitando problemas de multicolinearidade entre as colunas, garantindo sua adequação para modelos de regressão ou classificação. Outras variáveis nominais transformadas incluem "TP_COR_RACA", "TP_NACIONALIDADE", "TP_ST_CONCLUSAO", "TP_ESCOLA", "Q003", "Q004", "Q007", "Q018", "Q020", "Q021", "Q023" e "Q025". Essas transformações foram fundamentais para adequar os dados às análises de classificação. A Tabela 1 abaixo, apresenta as variáveis que foram transformadas em variáveis Dummy com seus respectivos significados.

Tabela 1. Variáveis e suas descrições no questionário ENEM.

Variável	Descrição
TP_FAIXA_ETARIA	Faixa etária do participante
TP_ANO_CONCLUIU	Ano de conclusão do ensino médio
TP_COR_RACA	Cor/Raça declarada
TP_NACIONALIDADE	Nacionalidade do participante
TP_ST_CONCLUSAO	Situação de conclusão do ensino médio
TP_ESCOLA	Tipo de escola em que concluiu ou concluirá o ensino médio
Q001	Escolaridade do pai
Q002	Escolaridade da mãe

Q003	Ocupação da pai ou homem responsável
Q004	Ocupação da mãe ou mulher responsável
Q005	Número de pessoas que moram na residência
Q006	Renda familiar mensal
Q007	Presença de empregado(a) doméstico(a)
Q008	Possui banheiro em casa
Q009	Possui quartos para dormir
Q010	Possui automóvel
Q011	Possui motocicleta
Q012	Possui geladeira
Q013	Possui freezer (independente ou segunda porta da geladeira)
Q014	Possui máquina de lavar roupas
Q015	Possui máquina de secar roupas
Q016	Possui forno micro-ondas
Q017	Possui lava-louças
Q018	Possui aspirador de pó
Q019	Possui televisão em cores
Q020	Possui aparelho de DVD
Q021	Possui TV por assinatura
Q022	Possui telefone celular
Q023	Possui telefone fixo
Q024	Possui computador
Q025	Possui acesso à internet

Fonte: Questionário Socioeconômico do ENEM, adaptado pelo autor, 2025.

Para integrar as informações geográficas dos municípios ao conjunto de dados principal, foi realizada uma operação de mesclagem, utilizando a coluna de municípios como chave de correspondência. Esse processo garantiu que os dados geográficos fossem adicionados ao conjunto original, mantendo todas as observações do conjunto principal. Essas etapas de pré- processamento e transformação foram fundamentais para assegurar que os dados estivessem bem estruturados e interpretáveis, garantindo a robustez e a qualidade das análises subsequentes.

Além disso, foi criada uma nova coluna para categorizar a variável de interesse, a nota de matemática. Os candidatos com notas superiores a 500 pontos (metade do valor total da prova) foram classificados como 'bom', enquanto aqueles com pontuações abaixo desse valor foram categorizados como 'ruim'. Assim, o foco da análise está na avaliação da proporção de alunos com desempenho acima da nota de corte e na comparação entre modelos de aprendizado de máquina para prever corretamente se um candidato pertence ou não ao evento de interesse.

2.2 MODELOS DE APRENDIZADO DE MÁQUINAS

Os modelos de aprendizado de máquina utilizados neste estudo foram selecionados com base em sua eficiência e capacidade de classificação. Três técnicas principais foram empregadas: Regressão Logística, Random Forest e XGBoost.

2.2.1 Regressão logística

A regressão logística é uma técnica estatística amplamente empregada na análise de dados quando a variável de desfecho é binária, ou seja, assume apenas dois valores possíveis, como "sim" ou "não", "presente" ou "ausente" (Hosmer; Lemeshow e Sturdivant, 2013). Nesse modelo, o objetivo é estimar a probabilidade condicional de ocorrência de um evento em função de uma ou mais variáveis independentes. Para isso, utiliza-se a função logística, que transforma uma combinação linear das variáveis preditoras em um valor entre 0 e 1 — intervalo adequado para representar probabilidades.

A transformação central do modelo logístico é o logit, definido como o logaritmo da razão entre a probabilidade de ocorrência e a de não ocorrência do evento. Essa transformação resulta em uma equação linear nos parâmetros, mantendo muitas das propriedades desejáveis dos modelos de regressão linear tradicional, com a vantagem de ser apropriada para desfechos dicotômicos.

Segundo os autores, a regressão logística busca o modelo mais ajustado, simples e interpretável clinicamente possível, preservando os princípios da análise de regressão mesmo sob as especificidades de uma variável resposta binária (Hosmer; Lemeshow e Sturdivant, 2013).

Como destacado por De Sousa Fernandes et al., (2023), é comum definir um limiar (threshold) para a probabilidade estimada com o intuito de realizar a classificação das observações. No presente estudo, adotou-se o valor de 0,5 como ponto de corte: se a probabilidade prevista (p) for maior que 0,5, o estudante é classificado como "bom", representando o evento de interesse.

2.2.2 Random forest

O método Bagging (bootstrap aggregating) consiste na criação de diversos subconjuntos de dados por meio de reamostragem com reposição (bootstrapping), nos quais são ajustadas várias árvores de decisão cuja média (ou votação) é utilizada para melhorar a estabilidade e a precisão das previsões (Breiman, 1996; Prasad et al., 2006).

O Random Forest é uma extensão do Bagging. Ele também utiliza amostras bootstrap para construir múltiplas árvores de decisão, mas introduz um elemento adicional de aleatoriedade: em cada divisão da árvore, é considerado apenas um subconjunto aleatório de variáveis preditoras (Breiman, 2001). Essa técnica reduz a correlação entre as árvores e melhora o desempenho do modelo.

A previsão final do Random Forest é obtida por meio de votação majoritária no caso de classificação, ou pela média das previsões, no caso de regressão. Comparado ao Bagging, o Random Forest é mais robusto, menos sensível a variações nos dados e notoriamente resistente ao overfitting.

2.2.3 XGBoost (Extreme Gradient Boosting)

O XGBoost (*Extreme Gradient Boosting*) é um algoritmo de aprendizado de máquina baseado em árvores de decisão que implementa o método de boosting por gradiente de forma otimizada. Ele se

destaca por sua eficiência computacional, capacidade de paralelização, controle de *overfitting* e desempenho preditivo elevado em tarefas tanto de classificação quanto de regressão (Chen e Guestrin, 2016).

O princípio do boosting consiste na construção sequencial de modelos fracos (geralmente árvores rasas), em que cada novo modelo tenta corrigir os erros cometidos pelos anteriores. O XGBoost aprimora esse processo ao introduzir regularização explícita na função de custo, suporte a missing values, e otimizações de memória e processamento, o que o torna amplamente utilizado em competições de ciência de dados e aplicações em larga escala.

2.3 IMPLEMENTAÇÃO DA ANÁLISE

As análises foram realizadas utilizando a linguagem de programação Python 3, com o auxílio de diversas bibliotecas amplamente usadas em ciência de dados e aprendizado de máquina. A biblioteca scikit-learn (versão 0.24.2) foi empregada, junto com outras ferramentas como Pandas, NumPy, Matplotlib e XGBoost. Além disso, foram utilizadas bibliotecas de georreferenciamento, como GeoPandas para identificar as coordenadas geográficas das cidades do Maranhão (MA) e os códigos dos municípios. As análises foram realizadas em um desktop com as seguintes especificações: processador Intel Core i7, memória DDR4 com capacidade de 8 gb.

2.4 PADRONIZAÇÃO

Para minimizar possíveis diferenças de escala entre as variáveis explicativas, os dados foram padronizados utilizando a fórmula:

$$Z = \frac{X-\mu}{\sigma} \qquad (1)$$

em que:

- Z: valor padronizado,
- X: valor original da variável,
- μ: média da variável,
- σ: desvio padrão da variável.

Esse processo transforma os dados em uma escala com média 0 e desvio padrão 1, permitindo que os modelos processem as variáveis com diferentes magnitudes de forma eficiente.

2.5 ACURÁCIA PREDITIVA

Para avaliar a eficiência das técnicas utilizadas, o treinamento e a validação foram realizados por meio da divisão dos dados em treino e teste. A divisão seguiu a proporção de 70% dos dados destinados para treinamento do modelo e 30% para teste. Essa abordagem permite medir o desempenho do modelo em dados não vistos, garantindo uma validação mais robusta dos resultados. Para uma melhor compreensão das métricas, segue a tabela com a matriz de confusão genérica:

Tabela 2. Matriz de Confusão

	Previsto: Classe A	**Previsto: Classe B**
Real: Classe A	**VP**	FN
Real: Classe B	FP	**VN**

Fonte: elaborada pelo autor, 2025.

A tabela acima permite uma análise detalhada dos erros cometidos pelo modelo. Ela possui quatro elementos principais:

1. VP significa verdadeiros positivos;
2. FN significa falsos negativos;
3. FP significa falsos positivos;
4. VN significa verdadeiros negativos.

A partir da tabela podem-se obter várias métricas de qualidade do modelo. As mais utilizadas são:

Acurácia: Mede a proporção de previsões corretas em relação ao total de observações. É dada por (VP + VN) / (VP + VN + FP + FN);

Precisão: Indica o quão preciso o modelo é na identificação da classe positiva. É calculada como VP / (VP + FP);

Recall (Sensibilidade): Mede a capacidade do modelo em identificar corretamente a classe positiva. Sua fórmula é VP / (VP + FN);

F1 Score: É a média harmônica entre precisão e recall, sendo especialmente útil em situações onde existe um equilíbrio entre FP e FN. É calculada como 2* (Precisão * Recall) / (Precisão + Recall)

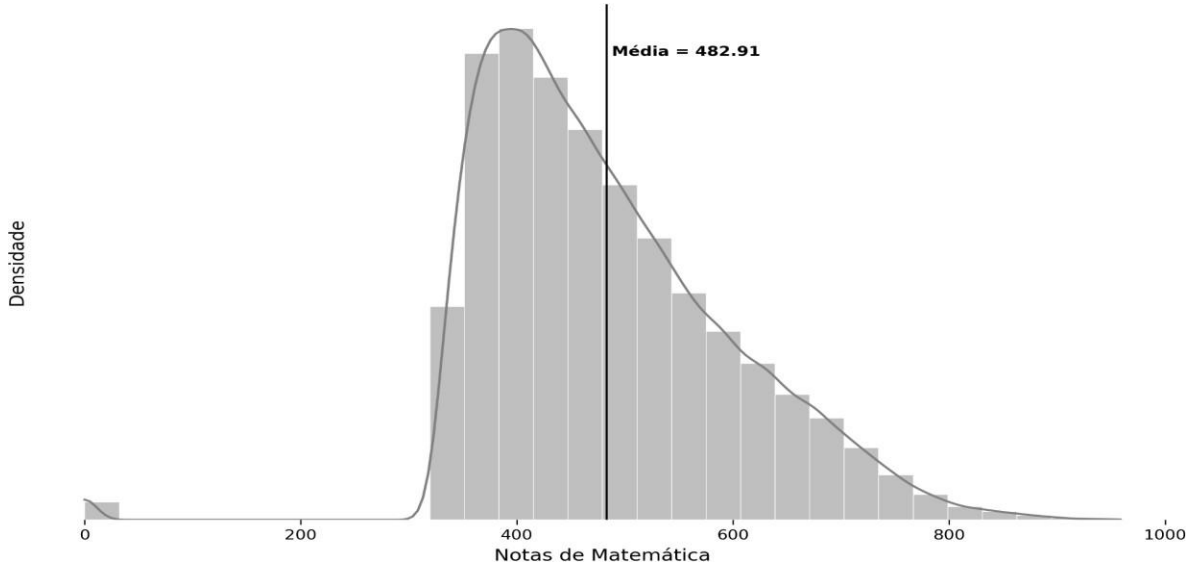
Outra maneira de avaliar modelos de classificação, é por meio de uma curva ROC. Uma curva ROC é uma representação gráfica do desempenho do modelo. A AUC da curva ROC, representa a área sob a curva ROC, que mede o desempenho geral do modelo em distinguir entre classes positivas e negativas. Um valor mais próximo de 1 indica um modelo melhor.

3 RESULTADOS E DISCUSSÃO

3.1 ESTATÍSTICA DESCRITIVA

Para compreender a distribuição das notas de Matemática dos candidatos maranhenses no ENEM 2023, foi construído um histograma acompanhado por uma curva de densidade. A análise buscou identificar padrões no desempenho geral dos participantes. Conforme apresentado na Figura 01, a maior concentração de notas encontra-se entre aproximadamente 398 e 553 pontos, refletindo o desempenho típico da maioria dos candidatos. Um total de 554 participantes obteve nota zero. Excluindo esses casos, as notas variaram de 322,7 a 958,6 pontos, com média geral de 485,30.

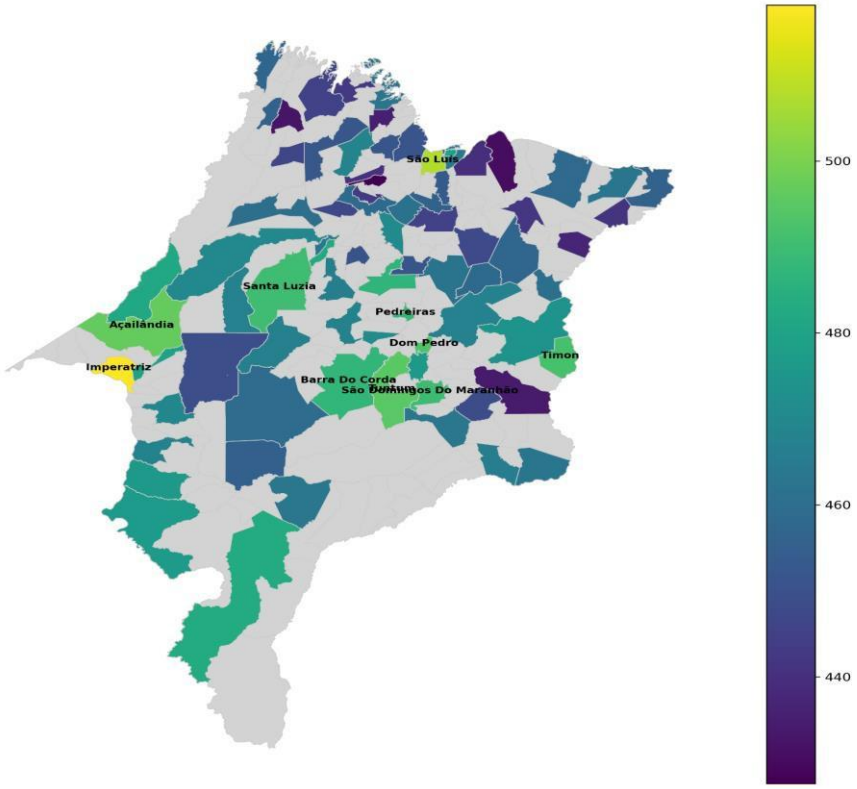
Figura 1. Distribuição das Notas de Matemática do ENEM 2023 no Maranhão.



Fonte: elaborada pelo autor, 2025.

A Figura 2 apresenta um mapa de calor que ilustra o desempenho dos estudantes no ENEM 2023, considerando a nota média de matemática nos municípios do Maranhão (MA). Os resultados ilustram uma variação expressiva entre as cidades, destacando desigualdades educacionais dentro do estado.

Figura 2. Mapa de Calor do Desempenho em Matemática no ENEM 2023 (Maranhão). As 10 cidades com maiores médias estão destacadas. Tons mais claros (amarelos) indicam notas mais altas, enquanto tons mais escuros refletem menor desempenho.



Fonte: elaborada pelo autor, 2025.

Imperatriz lidera com uma média de 518,13, seguida por São Luís, com 507,70. Açailândia ocupa o terceiro lugar com 496,82, enquanto Tuntum e Dom Pedro aparecem em posições intermediárias, com médias de 494,90 e 492,70, respectivamente. Municípios como Timon, Santa Luzia e São Domingos do Maranhão mantêm médias próximas, variando entre 490 e 491. Pedreiras e Barra do Corda também estão entre as dez cidades com maiores notas médias do estado (Top 10), com valores ligeiramente inferiores.

3.2 MODELOS DE APRENDIZADO DE MÁQUINAS

Foi realizada uma avaliação de diferentes modelos de aprendizado de máquina com o objetivo de prever os resultados com maior precisão. A Tabela 1 apresenta a comparação entre os modelos testados — XGBoost, Regressão Logística e Random Forest — com base em métricas como acurácia, precisão, recall e F1 Score. Esses indicadores permitem identificar qual modelo oferece o melhor equilíbrio entre os acertos e os erros na classificação dos dados, sendo fundamentais para apoiar decisões em contextos preditivos.

Tabela 3. Acurácia, Precisão, Recall e F1-Score dos Modelos Avaliados

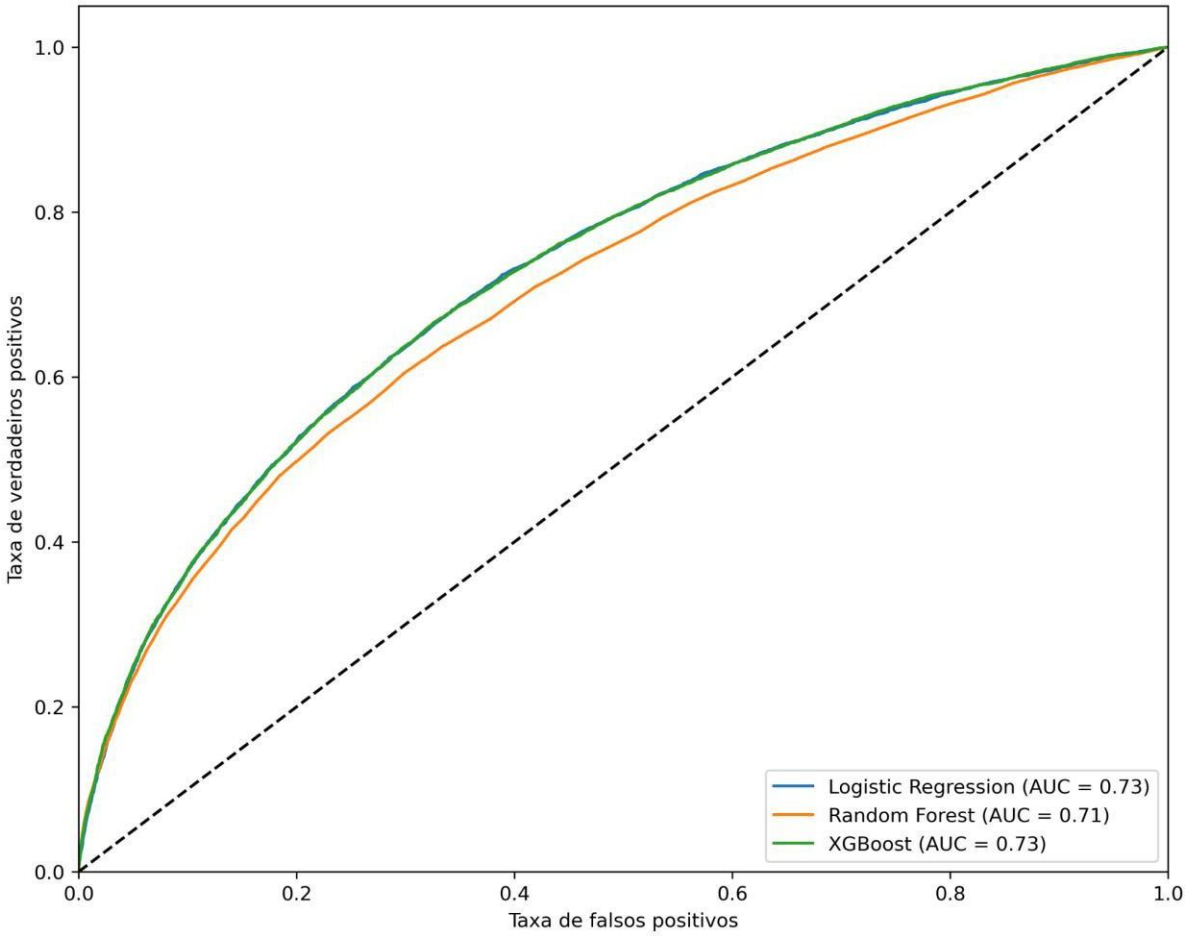
Modelo	Acurácia	Precisão	Recall	F1 Score
XGBoost	0.729	0.635	**0.3465**	**0.448**
Logistic Regression	**0.731**	**0.654**	0.3213	0.431
Random Forest	0.724	0.619	0.3384	0.438

Fonte: elaborada pelo autor, 2025.

A Regressão Logística destacou-se como o modelo com maior acurácia (0,7312) e boa precisão (0,6544), demonstrando eficácia na identificação correta dos casos negativos. Contudo, seu recall baixo (0,3213) aponta limitações significativas na detecção de verdadeiros positivos, o que pode ser problemático em cenários onde é crucial identificar todos os casos positivos. Por outro lado, o XGBoost apresentou um equilíbrio mais satisfatório entre as métricas avaliadas, alcançando um F1 Score de 0,4484 e mantendo uma alta acurácia (0,7299), o que o torna uma escolha mais adequada para situações que demandam equilíbrio entre precisão e recall. Já o Random Forest, embora tenha exibido a menor acurácia entre os modelos testados (0,7246), alcançou um F1 Score (0,4378), evidenciando sua robustez em cenários com maior complexidade e variabilidade dos dados.

Para melhor compreensão, a Figura 3 ilustra as curvas ROC (Receiver Operating Characteristic) dos três modelos analisados, permitindo uma comparação visual de suas capacidades de discriminação entre as classes. A curva ROC fornece uma representação gráfica da relação entre a taxa de verdadeiros positivos (Recall) e a taxa de falsos positivos, avaliando como os modelos equilibram essas taxas em diferentes pontos de corte.

Figura 3. apresenta as curvas ROC dos três modelos de classificação: Random Forest (amarela), Regressão Logística (azul) e XGBoost (verde). A linha cinza tracejada ilustra o desempenho de um classificador aleatório, com AUC igual a 0.50. Cada curva reflete diferentes características de discriminação entre as classes.



Fonte: elaborada pelo autor, 225.

A Figura 3 ilustra as curvas ROC dos modelos XGBoost, Regressão Logística e Random Forest, destacando suas capacidades na discriminação entre classes. Tanto o XGBoost quanto a Regressão Logística alcançaram áreas sob a curva (AUC) de 0.73, demonstrando superioridade na separação das classes. Visualmente, as curvas desses modelos permanecem consistentemente acima da linha diagonal tracejada, reforçando sua eficiência.

Por outro lado, o Random Forest apresentou a menor AUC, de 0.71, com uma curva mais próxima da linha diagonal, indicando desempenho inferior na distinção entre as classes. Essa análise evidencia que o XGBoost e a Regressão Logística são melhores na discriminação de classes, enquanto o Random Forest tem eficácia mais limitada nesse aspecto.

3.3 IMPORTÂNCIA DAS VARIÁVEIS

A regressão logística foi escolhida como modelo final por apresentar o melhor equilíbrio entre desempenho e interpretabilidade. Embora o XGBoost tenha se destacado no recall e F1 Score, a regressão logística obteve a maior acurácia (0,731) e precisão (0,654), além de permitir maior clareza sobre o impacto das variáveis.

As importâncias das variáveis na regressão logística foram avaliadas pelos coeficientes do modelo, que indicam sua influência na predição do resultado. A variável Q006 (renda familiar mensal) destacou-se como o principal preditor, seguida por TP_FAIXA_ETARIA e TP_ANO_CONCLUIU, refletindo a relevância de fatores socioeconômicos e educacionais no desempenho dos participantes. Por outro lado, diversas variáveis apresentaram coeficientes próximos de zero, como Q007 (número de banheiros) e Q011 (posse de motocicleta), sugerindo baixa relevância. Dummies categóricas como TP_COR_RACA e TP_ST_CONCLUSAO também mostraram impacto reduzido, embora revelem nuances entre subgrupos.

Vale destacar que algumas variáveis apresentaram coeficientes negativos, como TP_ESTADO_CIVIL e Q004_B, mas com magnitudes muito baixas, indicando efeito marginal. Esses resultados reforçam a importância de selecionar variáveis realmente relevantes, e sugerem que, em estudos futuros, variáveis com baixa contribuição podem ser descartadas para tornar o modelo mais enxuto e interpretável.

4 CONSIDERAÇÕES FINAIS

As análises realizadas neste estudo mostraram que os algoritmos de aprendizado de máquina podem ser ferramentas para compreender o desempenho dos estudantes no ENEM. Entre os modelos testados, o XGBoost foi o mais equilibrado em termos de sensibilidade e precisão, mas a Regressão Logística apresentou maior acurácia e clareza na interpretação dos resultados. Desta forma, evidenciou-se que aspectos socioeconômicos, como a renda familiar e a idade dos estudantes, influenciam no desempenho em Matemática mais do que fatores relacionados à infraestrutura das residências. Além disso, a análise espacial evidenciou discrepância entre os municípios maranhenses, indicando que algumas regiões, como São Luís e Imperatriz, têm desempenho superior à média estadual. Esses resultados podem contribuir para o planejamento de políticas públicas educacionais mais eficazes.

Embora o estudo utilize uma abordagem de classificação binária e esteja baseado em dados de um único ano, essas escolhas foram feitas para garantir clareza e foco na análise. No entanto, essas características também oferecem margem para futuras investigações que explorem modelos mais sofisticados, integrem dados de múltiplos anos e ampliem o escopo para outras regiões do país.

REFERÊNCIAS

BRASIL. Ministério da Educação. Portaria nº 438, de 28 de maio de 1998. Institui o Exame Nacional do Ensino Médio – ENEM. Disponível em: <http://www.crmariocovas.sp.gov.br/pdf/diretrizes_p0178-0181_c.pdf>. Acesso em: 19 abr. 2025.

BRASIL. Ministério da Educação. Portaria nº 462, de 27 de maio de 2009. Altera a Portaria nº 438/1998 e define a matriz de competências do ENEM. Disponível em: <http://portal.mec.gov.br/dmdocuments/port462.pdf>. Acesso em: 19 abr. 2025.

BREIMAN, L. Bagging predictors. Machine learning, v. 24, p. 123-140, 1996. BREIMAN, L. Random forests. Machine learning, v. 45, p. 5-32, 2001.

CHEN, T; GUESTRIN, C. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2016. p. 785–794. DOI: 10.1145/2939672.2939785.

DE SOUSA FERNANDES, L C; MENDES, F B; SILVA, J A da; SILVA, R F da; DAMACENO, G P; MOURA, E G. Análise do desempenho em matemática e suas tecnologias dos participantes do ENEM 2021 em Barra do Corda, Maranhão: uma comparação entre alunos de escolas públicas e privadas por meio de regressão logística. Contribuciones a Las Ciencias Sociales, São José dos Pinhais, v. 16, n. 12, p. 33822-33835, 2023. DOI: 10.55905/revconv.16n.12-282.

DE SOUSA FERNANDES, L C; PEREIRA, L B; MOURA, E G. VARIÁVEIS QUE INFLUENCIARAM O DESEMPENHO DOS CANDIDATOS NO ENEM EM MATEMÁTICA E SUAS TECNOLOGIAS: UM ESTUDO COM CANDIDATOS QUE REALIZARAM A PROVA EM BARRA DO CORDA MARANHÃO EM 2021. Revista Ibero-Americana de Humanidades, Ciências e Educação, v. 9, n. 9, p. 3587-3599, 2023.

HOSMER, D W.; LEMESHOW, S; STURDIVANT, R X. Applied logistic regression. 3. ed. Hoboken: John Wiley & Sons, 2013.

JORDAHL, K et al. geopandas/geopandas: v0. 5.0. Zenodo, 2021.

LIMA, P da S N et al. Analise de dados do Enade e Enem: uma revisão sistemática da literatura. Avaliacão: Revista da Avaliacão da Educação Superior (Campinas), v. 24, p. 89- 107, 2019.

LOPEZ-AREVALO, I. et al. Um método de codificação com eficiência de memória para processamento de dados de tipo misto em aprendizado de máquina. Entropy, v. 22, 2020. https://doi.org/10.3390/e22121391.

PEREIRA, R.H.M.; GONÇALVES, C.N.; et al. Geobr: Loads Shapefiles of Official Spatial Data Sets of Brazil. GitHub repository, 2019. Disponível em: <https://github.com/ipeaGIT/geobr>. Acesso em: 19 abr. 2025.

PRASAD, A M.; IVERSON, L R.; LIAW, A. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. Ecosystems, v. 9, p. 181- 199, 2006.

SILVA, R F da. Fatores que influenciam o desempenho acadêmico. 2013. Dissertação (Mestrado) – Insper Instituto de Ensino e Pesquisa, São Paulo. Disponível em: <https://repositorio-api.insper.edu.br/server/api/core/bitstreams/7dee910b-b1cf-4ba4-8f27-7ef3f318ad6a/content>. Acesso em: 7 abr. 2025.

TEIXEIRA, M; CAVIQUE, L. Feature engineering: techniques and applications. Revista de Ciências da Computação, p. 43-54, 2023.

ZIEGEL, E R. The elements of statistical learning. 2003.