



INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO PIAUÍ
CAMPUS URUÇUI
CURSO LICENCIATURA EM MATEMÁTICA

FRANCISCO BEZERRA MENDES

**INVESTIGAÇÃO DOS PRINCIPAIS FATORES QUE INFLUENCIAM O
DESEMPENHO DOS ESTUDANTES NO ENEM NO ESTADO DO PIAUÍ, COM
ÊNFASE NA UTILIZAÇÃO DE MARCHINE LEARNING PARA PREVISÃO DAS
NOTAS.**

URUÇUI

2024

FRANCISCO BEZERRA MENDES

INVESTIGAÇÃO DOS PRINCIPAIS FATORES QUE INFLUENCIAM O DESEMPENHO
DOS ESTUDANTES NO ENEM NO ESTADO DO PIAUÍ, COM ÊNFASE NA
UTILIZAÇÃO DE MARCHINE LEARNING PARA PREVISÃO DAS NOTAS.

Trabalho de Conclusão de Curso (artigo científico)
apresentado como exigência parcial para obtenção do
diploma do Curso de Licenciatura em Matemática do
Instituto Federal de Educação, Ciência e Tecnologia do
Piauí, Campus Uruçuí.

Orientadora: Prof. Dr. Ernandes Guedes Moura.

Coorientador: Prof. Dr. Bruno Ribeiro de Mesquita.

URUÇUÍ

2024

FRANCISCO BEZERRA MENDES

INVESTIGAÇÃO DOS PRINCIPAIS FATORES QUE INFLUENCIAM O DESEMPENHO
DOS ESTUDANTES NO ENEM NO ESTADO DO PIAUÍ, COM ÊNFASE NA
UTILIZAÇÃO DE MARCHINE LEARNING PARA PREVISÃO DAS NOTAS.

Trabalho de conclusão de curso
apresentado como exigência para
aprovação na disciplina TCC II do Curso
de Licenciatura em Matemática do
Instituto Federal de Educação, Ciência e
Tecnologia do Piauí, Campus Uruçuí.

Aprovada em: 30 / 08 / 2024 .

BANCA EXAMINADORA

Prof. Dr. Ernandes Guedes Moura (Orientador)
Instituto Federal do Piauí (IFPI)

Prof. Dr. Bruno Ribeiro de Mesquita (Coorientador)
Instituto Federal do Piauí (IFPI)

Prof. Ms. Gabriel dos Santos Pinto
Instituto Federal do Piauí (IFPI)

Prof. Erivan Silva Costa
Instituto Federal do Piauí (IFPI)

Francisco Bezerra Mendes 1
Prof. Dr. Ernandes Guedes Moura 2
Prof. Dr. Bruno Ribeiro de Mesquita 3

RESUMO

Este projeto tem como objetivo analisar os fatores que influenciam o desempenho dos alunos no Exame Nacional do Ensino Médio (ENEM), no Estado do Piauí, e desenvolver um modelo de aprendizado de máquina (*Machine Learning* -ML) para prever as notas dos participantes. O projeto envolve a coleta e análise de dados, pré-processamento, seleção de características, treinamento e avaliação do modelo, com o intuito de compreender as principais variáveis que afetam o desempenho dos alunos e fornecer uma ferramenta preditiva útil para educadores e gestores.

Palavras-chave: ENEM; Desempenho de estudantes; Análise de dados.

ABSTRACT

This project aims to analyze the factors that influence student performance in the National High School Exam (ENEM) in the state of Piauí, and to develop a machine learning (ML) model to predict student grades. The project involves data collection and analysis, preprocessing, feature selection, training, and model evaluation, with the aim of understanding the main variables that affect student performance and providing a useful predictive tool for educators and managers.

Keywords: ENEM; Student performance; Data analysis.

Data de aprovação: 30 / 08 / 2024

1 INTRODUÇÃO

O Exame Nacional do Ensino Médio (ENEM) desempenha um papel fundamental na avaliação do conhecimento e habilidades dos estudantes brasileiros. O ENEM é utilizado como processo seletivo de grande parte das universidades públicas e privadas do país.

Compreender os fatores que afetam o desempenho dos alunos no ENEM é importante para melhorar a qualidade do ensino. Dados históricos do ENEM, disponibilizados pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira – INEP, compõem informações diversas sobre os resultados individuais dos estudantes e várias variáveis caracterizam cada participante. Diante da dimensionalidade dos dados, uma questão essencial é identificar quais fatores são mais relevantes para o desempenho dos estudantes.

Muitos estudos já foram realizados com o objetivo de investigar os principais fatores que contribuem no desempenho educacional dos alunos e determinar as possíveis diferenças no desempenho educacional, em relação às diferentes características dos indivíduos (SILVA, 2017). Todavia, fazer previsões das notas dos candidatos ao ENEM via aprendizado de máquinas (*Machine Learning*-ML) tem sido ainda muito pouco explorado e pode

proporcionar elementos adicionais que ajudem entender melhor os padrões subjacentes e fornecer informações úteis para os candidatos, instituições de ensino e sugerir políticas públicas voltadas para a educação.

A *Machine Learning* é um sub-campo da inteligência artificial, dedicado ao desenvolvimento de algoritmos e técnicas que permitem ao computador “aprender” com dados históricos e aperfeiçoar o desempenho de algumas tarefas de forma automática (AMORIM *et.al*, 2008). Outra aplicação importante da *Machine Learning* é fazer previsões e, no contexto da educação, o uso desta técnica pode ser empregado para prever notas dos alunos, risco de evasão escolar, recomendar estratégias educacionais, entre outros. As técnicas de *Machine Learning* podem encontrar padrões complexos nos dados e aprender a relação entre variáveis para realizar previsões com base em novos dados.

Diante desse contexto, este trabalho teve como objetivo identificar e compreender os fatores que influenciam o desempenho dos estudantes que realizaram o último ENEM no estado do Piauí. Para isso, foi realizada uma análise exploratória dos dados e uma comparação de modelos de *Machine Learning* para prever as notas dos participantes, oferecendo informações relevantes que podem contribuir para tomada de decisões e melhorias na educação.

2 ESTRUTURA DO TEXTO

2.1 APRENDIZADO DE MÁQUINA

O aprendizado de máquinas é um subcampo da ciência da computação, que busca uma similaridade das técnicas matemáticas e estatísticas com os algoritmos computacionais (DEO, 2015; OBERMEYER, EMANUEL, 2016). O ML abrange o estudo e explora a construção de algoritmos computacionais a partir do aprendizado através de dados ao invés de instruções pré-programadas (MITCHELL, 2006; HASTIE, TIBSHIRANI, FRIEDMAN, 2009; DEO, 2015), em que o objetivo principal consiste em construir um sistema computacional que aprenda com um banco de dados pré-definidos e a partir deles ter potencial para produzir modelos de predição, classificação ou detecção (PAIXÃO *et al.*, 2022).

Na educação, técnicas de mineração de dados, a exemplo de modelos preditivos e seleção de features (características), têm sido usadas como meio para ajudar na obtenção da compreensão das associações de variáveis relacionadas (DUTRA *et.al*, 2023). O ML pode ser implementado em várias áreas, como algoritmos para previsão de rendimentos, algoritmo de reconhecimento de imagem, entre outros. Stearns *et.al* (2017) realizaram um estudo sobre a capacidade de prever a nota de um aluno utilizando somente as informações do questionário socioeconômico como features. Para isso, usaram modelo de regressão capaz de trabalhar com um volume grande de dados. Dessa forma, para trabalhar com modelos de ML, é importante saber que há três tipos de aprendizado de máquinas: aprendizado supervisionado, não supervisionado e aprendizado por reforço.

2.2 APRENDIZADO SUPERVISIONADO, APRENDIZADO NÃO SUPERVISIONADO E APRENDIZADO POR REFORÇO

O ML pode ser classificado em três tipos: aprendizado supervisionado, aprendizado não supervisionado e aprendizado reforçado (CRAVERO *et al.*, 2022). A linha da Tabela 01, chamada “Tarefa de processamento de dados”, faz alusão a problemas que precisam ser resolvidos, e a linha, “Algoritmos de aprendizado”, descreve os métodos que podem ser usados para resolver os problemas.

Tabela 01 – Tipos de classificação do aprendizado de máquinas

Tipo de classificação	Aprendizado Supervisionado	Aprendizado não supervisionado	Aprendizado por Reforço
Tarefas de processamento de dados	Estimativa Classificação Regressão	Agrupamento Predição	Tomando uma decisão
Algoritmos de aprendizagem	Máquina de vetores de suporte Redes Bayesianas Redes neurais Baías ingênuas Modelo oculto de Markov	Modelo de mistura de processo Dirichlet X-means K-means Modelo de mistura gaussiana	Aprendizagem TD Aprendizagem de sarsa Q – learning R – learning

Fonte: Cravero et al. (2022)

Tanto o aprendizado supervisionado quanto o aprendizado não supervisionado fundamentam-se, principalmente, na análise de dados. Já o aprendizado por reforço é preferível em problemas relacionados à tomada de decisões (BENOS *et al.*, 2021). Para que os dados brutos se tornem dados valiosos são necessárias algumas etapas a serem feitas, como por exemplo: (i) limpeza de dados para remoção de elementos e ruídos inconsistentes ou faltantes, (ii) integração de dados para reunir dados de várias fontes e (iii) transformação de dados para normalização de dados (BENOS *et al.*, 2021).

3 METODOLOGIA

3.1 DADOS

Os dados utilizados para o presente estudo foram extraídos da base de dados aberta no site do Inep, conhecida como microdados do ENEM. Esses microdados contêm informações detalhadas sobre os participantes e seus desempenhos no Exame Nacional do Ensino Médio, e são frequentemente usados para análises estatísticas, pesquisas acadêmicas e desenvolvimento de políticas educacionais. Após a coleta dos dados, foi realizado um pré-processamento que

incluiu a limpeza, a remoção de redundâncias e inconsistências e, se necessário, a aplicação de adequadas transformações nos dados. Os dados pré-processados foram então submetidos às técnicas de análise apropriadas.

Inicialmente, os dados do ENEM referentes a todo o Brasil em 2023 tinham uma dimensão de 3.933.955 participantes e 76 características, que incluíam dados socioeconômicos e informações sobre a escola onde o participante fez a prova. Após a delimitação para o estado do Piauí, o número de participantes foi reduzido para 99.639, mantendo as 76 características. Com o tratamento dos dados, incluindo a remoção de inconsistências e o tratamento de valores ausentes (NAs), o conjunto de dados foi reduzido para 71.048 participantes e 76 características.

Dessas 76 características, várias foram excluídas por não contribuírem explicitamente para o desempenho dos candidatos na nota de matemática (como número de inscrição, UF da escola, gabaritos e notas de outras disciplinas, entre outras). Assim, foram mantidas 13 variáveis consideradas mais relevantes para explicar a nota de matemática. Após um tratamento mais refinado, essas variáveis foram utilizadas para análises posteriores.

3.2 MODELO

A metodologia utilizada no presente estudo será focada em selecionar as variáveis mais relevantes para o desempenho dos alunos que realizaram a prova no estado do Piauí e ajustar modelos de Machine Learning com base nessas variáveis. Inicialmente, será realizada uma análise de correlação para verificar quais variáveis explicativas possuem uma associação linear com a variável resposta (nota dos estudantes). As correlações serão determinadas pela fórmula:

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} \quad (1)$$

Onde:

- x_i e y_i são os valores individuais das variáveis x e y .
- $\bar{x}$ e $\bar{y}$ são as médias das variáveis x e y .
- O numerador representa a covariância entre x e y .
- O denominador normaliza a variância de cada variável.

Em que $-1 < r < 1$ e, se $r \rightarrow 1$ ou $r \rightarrow -1$, significa que a variável explicativa tem alta associação com a variável resposta positiva ou negativa, respectivamente. Porém, se $r \rightarrow 0$, significa que há baixa associação, ou mesmo que não há associação entre variável.

Após a análise de correlação, para a construção dos modelos de Machine Learning, os dados serão divididos em conjunto de treinamento (70%) e conjunto de teste (30%). Serão ajustados e comparados os modelos de Regressão Linear, Random Forest e XGBoost para avaliar o desempenho na predição das notas de matemática dos alunos. As variáveis pré-selecionadas serão analisadas para determinar quais delas têm maior impacto na previsão das notas, e os modelos serão ajustados e comparados com base nessa análise.

3.3 MÉTRICAS

Para avaliar o desempenho dos modelos ajustados, serão utilizadas as métricas mais comuns para regressão.

1. Erro Quadrático Médio (Mean Squared Error, MSE): Média das diferenças ao quadrado entre as previsões e os valores reais. O modelo que obtiver o menor MSE será considerado melhor ajustado aos dados.

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (2)$$

Onde:

- n é o número total de observações.
- y_i são os valores reais (verdadeiros).
- $\hat{y}_i$ são os valores estimados pelo modelo.
- A soma calcula o erro ao quadrado para cada ponto e, depois, faz a média.

2. Raiz do Erro Quadrático Médio (Root Mean Squared Error, RMSE): Raiz quadrada do MSE fornece uma medida do erro médio em unidades originais. O modelo que obtiver o menor RMSE será considerado melhor ajustado aos dados.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (3)$$

Onde:

- n é o número total de observações.
- y_i são os valores reais.
- $\hat{y}_i$ são os valores estimados pelo modelo.
- O MSE é calculado primeiro e depois extraímos a raiz quadrada para obter o RMSE.

3. Coeficiente de Determinação (R^2): Mede a proporção da variabilidade dos valores da variável dependente explicada pelas previsões do modelo. O modelo que obtiver o maior valor de R^2 será considerado melhor ajustado aos dados. Vale ressaltar que o R^2 varia entre 0 e 1.

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (4)$$

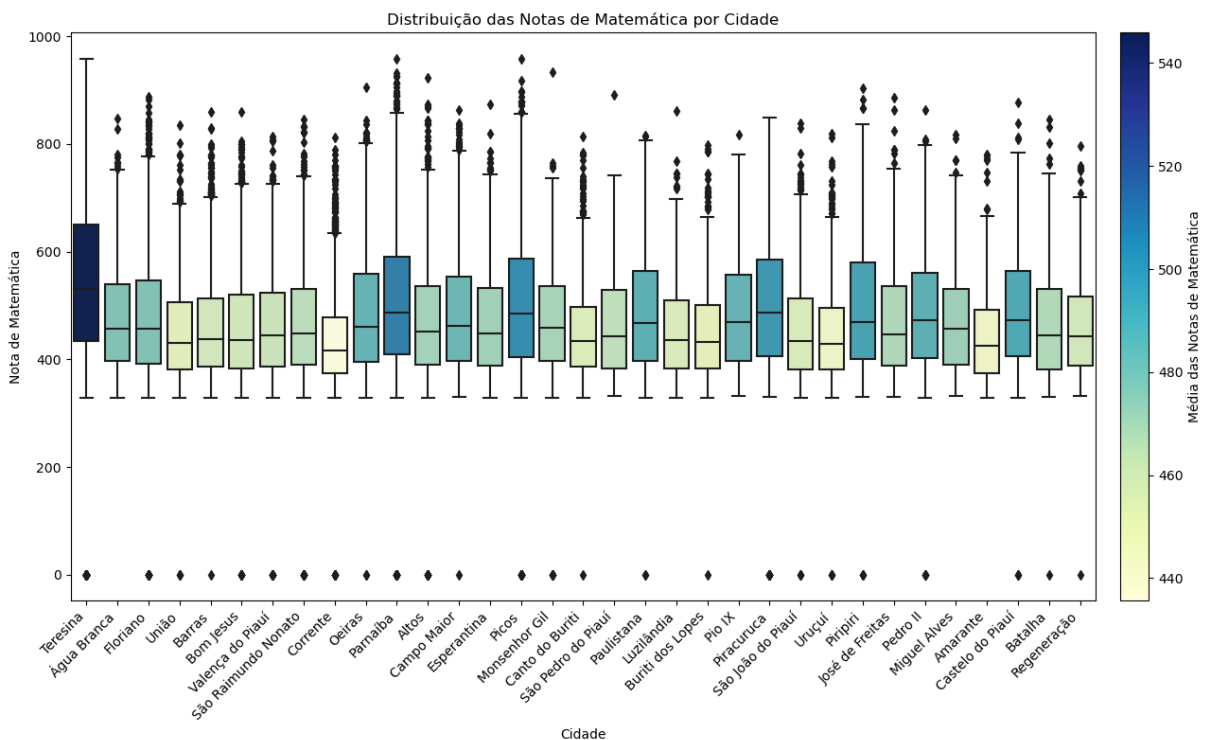
Onde:

- O numerador é a soma dos quadrados dos resíduos (Erro entre valores reais e preditos).
- O denominador é a soma dos quadrados totais (Variabilidade total dos valores reais em relação á média).

4 RESULTADOS E DISCURSÃO

A Figura 1 abaixo mostra um boxplot combinado com um gráfico de calor. No gráfico, a cidade com a maior nota de matemática no ENEM é destacada em azul escuro, enquanto a cidade com a nota mais baixa é representada em amarelo claro. O boxplot fornece uma visão clara das diferenças no desempenho das cidades e pode servir como uma base para discutir a qualidade e a equidade do sistema educacional.

Figura 01 – Boxplot das notas de matemática, agrupadas por cidades onde o exame foi realizado.



Fonte: Autor (2024)

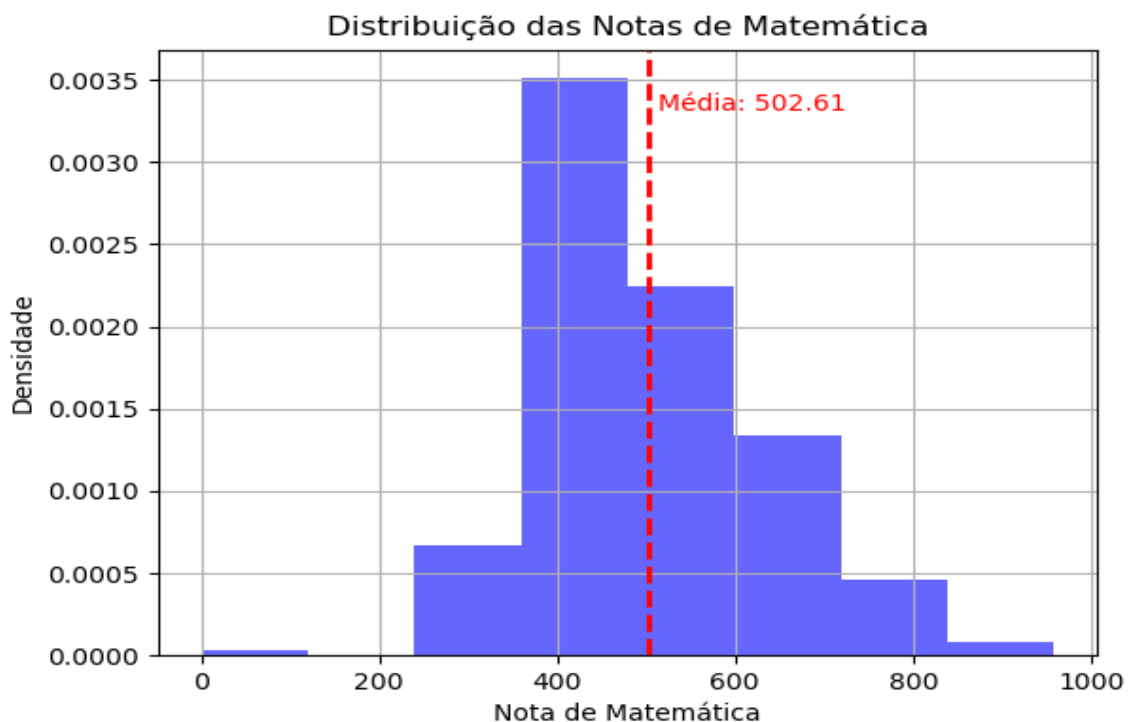
O boxplot revela que Teresina apresenta as maiores médias de notas de matemática, como era de se esperar, dada sua condição de capital. Em seguida, está Oeiras, Esperantina, Pio IX, Piracuruca, Piripiri e Castelo do Piauí. Destaca-se que Teresina também apresenta a maior

variabilidade nas notas, o que pode ser atribuído à diversidade de instituições de ensino, tanto públicas quanto privadas, e à variedade de recursos educacionais disponíveis na cidade. Essa variabilidade também pode refletir diferenças socioeconômicas entre os alunos, impactando o desempenho acadêmico.

Além disso, a mediana das notas de matemática por cidade mostra que Teresina tem a maior mediana, 531.30, indicando um desempenho superior em comparação com outras cidades. Outras cidades com medianas elevadas incluem Picos e Piracuruca, com valores de 484.60 e 487.35, respectivamente. Por outro lado, cidades como Corrente e Buriti dos Lopes apresentam medianas mais baixas, 415.50 e 432.25, sugerindo possíveis desafios educacionais nessas localidades.

A seguir, abordaremos a distribuição das notas de matemática do ENEM no Piauí. A Figura 2 apresenta um histograma que ilustra essa distribuição.

Figura 2 - Histograma de notas de matemática dos participantes do ENEM 2023.



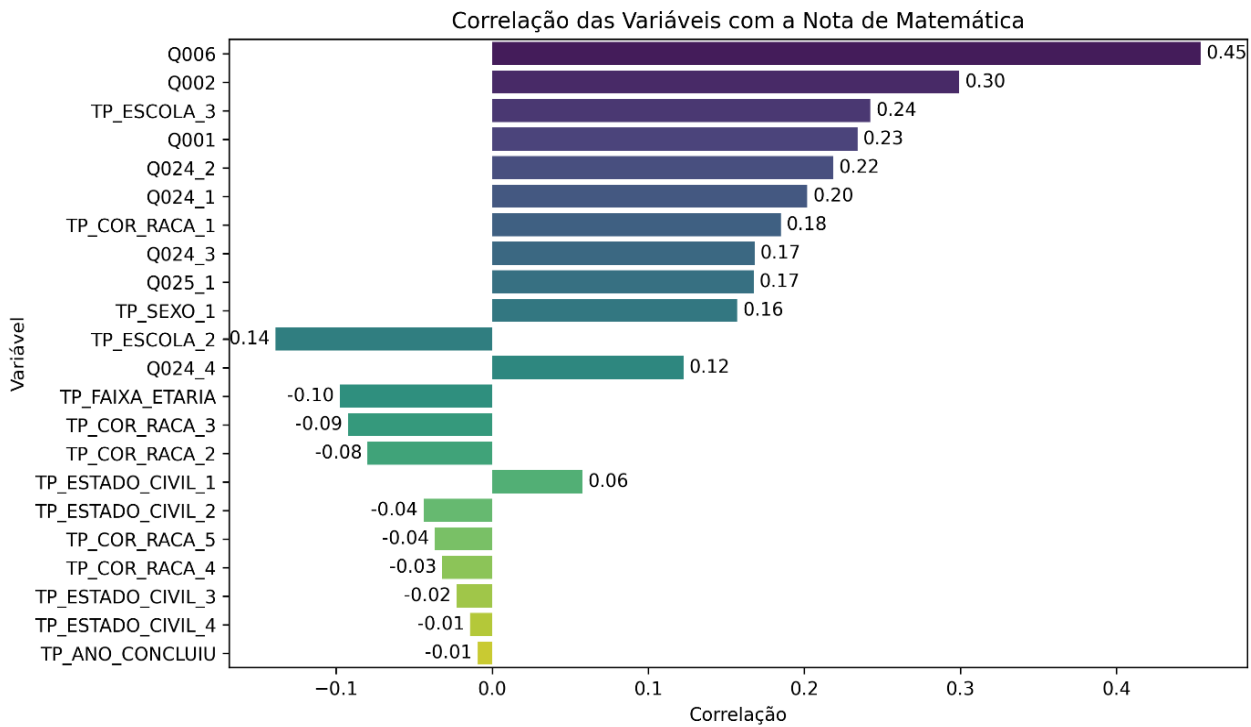
Fonte: Autor (2024)

O histograma das notas dos estudantes apresenta uma distribuição centralizada em torno da média de 502,61. A maioria das notas está concentrada entre os quartis 25% (403,3) e 75% (585,33), sugerindo que a maior parte dos alunos obteve notas intermediárias. A mediana, 477,4, indica que metade dos estudantes teve notas abaixo desse valor, refletindo uma leve assimetria à direita.

A dispersão dos dados, evidenciada pelo desvio padrão de 127,87, mostra que há variação significativa nas notas. Notas mais extremas são menos frequentes, como indicam os valores mínimos e máximos de 0 e 958,6, respectivamente. Esses dados sugerem que, embora a maioria dos estudantes tenha obtido notas próximas da média, uma pequena parcela apresentou desempenho muito acima ou muito abaixo, resultando em uma cauda longa à direita no histograma.

Aqui apresentamos a distribuição geral das notas de matemática. No entanto, é importante entender quais variáveis têm maior influência na previsão dessas notas. O gráfico abaixo destaca as correlações entre as variáveis estudadas.

Figura 3 - Correlação entre Variáveis na Previsão das Notas de Matemática do ENEM 2023 no Piauí.



Fonte: Autor (2024)

Tabela 1- Tabela explicativa das siglas da figura 3.

SIGLA	SIGNIFICADO
Q006	Qual a renda mensal da sua família.
Q002	Escolaridade da mãe ou responsável do participante.
TP_ESCOLA_3	Tipo de escola onde concluiu o ensino médio (categoria 3: Privada).
Q001	Escolaridade da mãe ou responsável do participante.
Q024_2	Frequência com que estuda sozinho

	(categoria 2: Poucas vezes).
Q024_1	Frequência com que estuda sozinho (categoria 1: Nenhuma vez).
TP_COR_RACA_1	Cor/ raça do participante (categoria 1: Branca).
Q024_3	Sua residência tem computador (categoria 3: sim, dois).
Q025_1	Na sua residência tem acesso a Internet (categoria 1: Não).
TP_SEXO_1	Sexo do participante (categoria 1: Masculino).
TP_ESCOLA_2	Tipo de escola onde concluiu o ensino médio(categoria 2: Pública).
Q024_4	Sua residência tem computador (categoria 4: sim, três).
TP_FAIXA_ETARIA	Faixa etária do participante.
TP_COR_RACA_3	Cor/raça autodeclarada pelo participante (categoria 3: Parda).
TP_COR_RACA_2	Cor/raça autodeclarada pelo participante (categoria 2: Preta).
TP_ESTADO_CIVIL_1	Estado civil do participante (categoria 1: Solteira(o)).
TP_ESTADO_CIVIL_2	Estado civil do participante (categoria 2: Casada(o)).
TP_COR_RACA_5	Cor/raça autodeclarada pelo participante (categoria 5: Indígena).
TP_COR_RACA_4	Cor/raça autodeclarada pelo participante (categoria 4: Amarela).
TP_ESTADO_CIVIL_3	Estado civil do participante (categoria 3: Divorciada(o)/ Desquitada(o)/Separada(o)).
TP_ESTADO_CIVIL_4	Estado civil do participante (categoria 4: Viúva(o)).
TP_ANO_CONCLUIU	Ano de conclusão do ensino médio pelo participante do ENEM.

Fonte: Autor (2024)

A análise das correlações entre as variáveis e a nota de matemática (NU_NOTA_MT) revela diversas relações que fornecem insights sobre os fatores associados ao desempenho dos alunos nesta disciplina. A correlação é uma medida que indica a força e a direção da relação linear entre duas variáveis. A renda familiar (Q006) apresenta a maior correlação positiva com a nota de matemática, com um coeficiente de 0.454. Esta relação sugere que existe uma associação substancial entre a renda familiar e o desempenho em matemática. Alunos provenientes de famílias com maior renda tendem a obter melhores notas, possivelmente devido ao acesso a recursos educacionais adicionais e suporte acadêmico.

A escolaridade dos pais (Q002 e Q001) mostra correlações positivas de 0.299 e 0.234, respectivamente. Esses resultados indicam que um maior nível educacional dos pais está

positivamente relacionado ao desempenho dos alunos em matemática. Essa correlação pode refletir o impacto do ambiente educacional e do suporte familiar no desempenho acadêmico dos filhos. A variável que representa a frequência a escolas particulares (TP_ESCOLA_3) tem uma correlação positiva de 0.242 com a nota de matemática. Alunos matriculados em escolas particulares tendem a apresentar um desempenho superior em matemática, o que pode ser atribuído a diferenças nos recursos e na qualidade da educação oferecida por essas instituições.

A correlação positiva de 0.168 associada ao estado civil do próprio candidato (TP_ESTADO_CIVIL) sugere que, em média, alunos casados têm um desempenho ligeiramente superior em matemática. Contudo, a correlação é relativamente fraca, indicando que outros fatores podem estar mais significativamente envolvidos. A leitura semanal de livros pelos pais (Q024_2) mostra uma correlação positiva de 0.219 com a nota de matemática. Esse achado sugere que pais que incentivam a leitura regular podem impactar positivamente o desempenho de seus filhos em matemática, possivelmente através do estímulo ao desenvolvimento cognitivo e ao hábito de estudo.

A posse de computador em casa (Q0024) apresenta uma correlação positiva de 0.167, indicando que a presença de um computador em casa está associada a um desempenho ligeiramente melhor em matemática. Esse recurso pode facilitar o acesso a ferramentas educacionais e a realização de atividades relacionadas ao aprendizado. Outras variáveis apresentam correlações mais fracas ou negativas com a nota de matemática. Por exemplo, a faixa etária (TP_FAIXA_ETARIA) tem uma correlação negativa de -0.098, sugerindo que a idade não é um fator determinante significativo para o desempenho em matemática dentro do conjunto de dados analisado. Da mesma forma, o sexo (TP_SEXO_1) apresenta uma correlação positiva modesta de 0.157, indicando uma leve diferença de desempenho entre os sexos.

Esses resultados fornecem uma visão geral sobre como diferentes fatores podem estar associados ao desempenho em matemática. No entanto, é crucial notar que correlação não implica causalidade. Fatores adicionais, interações complexas e variáveis não observadas podem influenciar essas relações. Portanto, uma análise mais aprofundada é necessária para compreender essas associações de maneira mais abrangente. Para isso, foram aplicados modelos de aprendizado de máquina com o objetivo de prever a nota de matemática e comparar a capacidade preditiva desses modelos.

A Tabela 1 apresenta os resultados de três modelos de aprendizado de máquina (Regressão Linear, Random Forest e XGBoost) usados para prever a nota de matemática no ENEM dos participantes do Piauí. Os resultados são avaliados com base em três métricas: Erro Quadrático Médio (MSE), Raiz do Erro Quadrático Médio (RMSE) e o Coeficiente de Determinação (R^2).

Tabela 2 - Coeficiente de determinação (R^2) e Erro Quadrático Médio (MSE) entre os valores verdadeiros (nota real) e preditos (nota prevista), usando os métodos Regressão linear, Random Forest e XGBoost.

MODELO	MSE	RMSE	R^2 (Percentual)
Regressão Linear	11958.66	109.36	27.45%
Random Forest	13444.97	115.95	18.43%

MODELO	MSE	RMSE	R ² (Percentual)
XGBoost	11355.85	106.56	31.11%

De acordo com a Tabela acima, o modelo de Regressão Linear apresentou um MSE de 11.958,66 e um RMSE de 109,36, com um R² de 27,45%. Isso indica que o modelo conseguiu explicar 27,45% da variação nas notas, mas ainda deixou uma margem significativa de erro. O modelo de Random Forest teve um desempenho inferior, com um MSE de 13.444,97, um RMSE de 115,95 e um R² de apenas 18,43%. Esses resultados sugerem que o Random Forest não foi eficaz em capturar as variações nas notas, resultando em maior erro nas previsões.

Por outro lado, o XGBoost demonstrou o melhor desempenho, com o menor MSE de 11.355,85 e o menor RMSE de 106,56, além do maior R² de 31,11%. Isso significa que o XGBoost explicou mais de 31% da variabilidade das notas de matemática, indicando uma maior precisão nas previsões.

Em resumo, o XGBoost se destacou como o modelo mais eficiente, fornecendo previsões mais precisas e uma melhor explicação da variação nas notas em comparação aos outros modelos, especialmente ao Random Forest. No entanto, apesar da utilidade desses modelos para a previsão, nenhum deles alcançou uma capacidade preditiva excepcional. Isso sugere que deve-se considerar a análise de outras variáveis e explorar diferentes modelos de aprendizado de máquina para um estudo mais abrangente e aprofundado.

5 CONSIDERAÇÕES FINAIS

Os resultados indicam que o desempenho em matemática no ENEM está fortemente relacionado a fatores socioeconômicos e educacionais. Em Teresina, a capital, observam-se as maiores médias e maior variabilidade nas notas, refletindo a diversidade de instituições e recursos disponíveis. A renda familiar e a escolaridade dos pais apresentam correlações positivas significativas com as notas, sugerindo que alunos de famílias com maior renda e maior nível de escolaridade tendem a alcançar melhores resultados.

A distribuição das notas de matemática é concentrada em torno da média, com uma leve assimetria à direita, indicando que a maioria dos alunos obteve notas intermediárias. A variabilidade nas notas revela diferenças significativas no desempenho, possivelmente atribuídas ao acesso desigual a recursos educacionais e suporte acadêmico. Para estudos futuros, é recomendável explorar técnicas avançadas e analisar dados de outras regiões do Brasil, além de investigar a evolução do desempenho ao longo do tempo. Este estudo contribui para a aplicação do aprendizado de máquina na educação, auxiliando na compreensão dos fatores que influenciam o desempenho dos estudantes e orientando melhorias no sistema educacional.

REFERÊNCIAS

- AMORIM, Maurício JV; BARONE, Dante; MANSUR, André Uebe. **Técnicas de aprendizado de máquina aplicadas na previsão de evasão acadêmica. In: Brazilian Symposium on Computers in Education** (Simpósio Brasileiro de Informática na Educação/SBIE). sn, 2008. p. 666-674. Disponível em: <http://milanesa.ime.usp.br/rbie/index.php/sbie/article/view/755>. Acesso em: 08 Jan. 2023.
- BENOS, L.; TAGARAKIS, A.C.; DOLIAS, G.; BERRUTO, R.; KATERIS, D.; BOCHTIS, D. **Machine learning in agriculture: A comprehensive updated review**. Sensors, 2021. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/34071553/>. Acesso em 04 Dez. 2023.
- CRAVERO, A.; PARDO, S.; SEPÚLVEDA, S.; MUÑOZ, L. **Challenges to use Machine Learning in agricultural Big Data: A systematic literature review**. Agronomy, 2022. Disponível em: https://www.researchgate.net/publication/359386300_Challenges_to_Use_Machine_Learning_in_Agricultural_Big_Data_A_Systematic_Literature_Review. Acesso em: 06 de Abril de 2024.
- DEO, R. C. JANDERSON, F. JOÃO, B. F. JÚNIOR, S. F. DAMIRES, Y. **Fatores que podem interferir no desempenho de estudantes no ENEM: uma revisão sistemática da literatura**. Machine Learning in medicine. Circulation, 2015 DUTRA, Revista Brasileira de Informática na Educação, v. 31, p. 323-351, 2023. Disponível em: <https://www.ahajournals.org/doi/10.1161/CIRCULATIONAHA.115.001593>. Acesso em: 13 de Junho de 2023.
- HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning: Data Mining, inference, and prediction**. Nova York: Springer, 2009. Disponível em: <https://link.springer.com/book/10.1007/978-0-387-84858-7>. Acesso em: 10 de Abril de 2023.
- MITCHELL, T. M. **The discipline of Machine Learning**. Pittsburgh: Carnegie Mellon University, School of Computer Science, Machine Learning Department, 2006. Disponível em: <https://www.cs.cmu.edu/~tom/pubs/MachineLearning.pdf>. Acesso em: 28 de Outubro de 2023.
- OBERMEYER, Z.; EMANUEL, E. J. **Predicting the future - Big Data, Machine Learning, and clinical medicine**. The New England Journal of Medicine, 2016. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/27682033/>. Acesso em: 27 de Outubro de 2023.
- PAIXÃO, G. M. M.; SANTOS, B. C.; ARAUJO, R. M.; RIBEIRO, M. H.; MORAES, J. L.; RIBEIRO, A. L. **Machine Learning na Medicina: Revisão e Aplicabilidade**. Arquivos Brasileiros de Cardiologia, 2022. Disponível em: https://www.researchgate.net/publication/357782657_Machine_Learning_na_Medicina_Revisao_e_Aplicabilidade. Acesso em: 08 de Dezembro de 2023.
- SCHWARTZMAN, S. **Educação e desenvolvimento: desafios para o Brasil**. Revista Brasileira de Educação, v. 52, p. 643-662; 2013. Disponível em: <http://schwartzman.org.br/simon/desafios/1desafios.pdf>. Acesso em: 04 de Dezembro de 2023.
- SILVA, I. C. S. **Fatores que influenciam o desempenho dos participantes da prova ENEM: um estudo para os residentes do Ceará no ano de 2015**. 2017. Orientador: Prof. Dr. João Mário Santos de França. 2017.29. Monografia - Bacharel em Finanças, Faculdade de Economia, Administração, Atuária e Contabilidade, Universidade Federal do Ceará, Fortaleza – CE, 2017.
- STEARNS, Bernardo et al. Prevendo desempenho dos candidatos do ENEM através de dados socioeconômicos. In: Anais do XXXVI Concurso de Trabalhos de Iniciação Científica da SBC. SBC, 2017. Disponível em: <https://repositorio.ufc.br/handle/riufc/33971>. Acesso em: 20 de Setembro de 2023.
- SOARES, J. F. **O impacto do ENEM na educação brasileira**. Revista Brasileira de Avaliação Educacional, v. 39, p. 65-87; 2020. Disponível em:

https://www.researchgate.net/publication/237074231_Qualidade_e_equidade_na_educacao_basica_brasileira_fatos_e_possibilidades. Acesso em: 12 de Agosto de 2023.