



DOI: 10.5281/zenodo.18929527

Mineração de dados educacionais do ENEM: Análise de caso do estado da Bahia

Jeilly de Almeida Costa¹

Cedma Ranielly Santos Firmino²

Rogério Luiz Cardoso Silva Filho³

Karla Patrícia Santos Oliveira Rodriguez Esquerre⁴

Resumo

O Exame Nacional do Ensino Médio (ENEM) e os microdados da educação são recursos fundamentais para a análise abrangente do desempenho educacional em nível nacional. Esses dados não apenas oferecem uma visão ampla da situação da educação básica em todo o país, mas também possibilitam estudos detalhados em níveis mais regionais e locais. Neste artigo, é apresentado o resultado da mineração desses dados para a obtenção de informações sobre o desempenho dos estudantes das escolas públicas estaduais da Bahia, um dos estados mais populosos do Brasil. Ao utilizar tanto o modelo de regressão logística quanto o modelo de floresta aleatória, foi possível identificar padrões e correlações significativas entre variáveis socioeconômicas e características da unidade de ensino com o desempenho dos alunos no ENEM. Fatores como a renda familiar, raça, nível educacional dos pais, além de aspectos específicos da escola, como qualificação dos professores, emergiram como influências diretas no rendimento dos estudantes. A análise revelou que estudantes oriundos de famílias com maior poder aquisitivo e com pais com maior grau de instrução formal tendem a obter melhores resultados. Da mesma forma, escolas com um corpo docente mais qualificado tendem a produzir melhores resultados acadêmicos entre os alunos. Esses achados têm implicações significativas para políticas educacionais, destacando a necessidade de medidas que visem a redução das disparidades socioeconômicas e melhoria da qualidade do ensino público. Os resultados obtidos estão alinhados com a literatura, e sugerem que investimentos em programas de capacitação docente e políticas de inclusão social podem contribuir para um ambiente educacional mais equitativo e de melhor qualidade. **Palavras-chave:** ENEM, Aprendizado de Máquinas, Mineração de Dados Educacionais, Regressão Logística, Floresta Aleatória.

¹Universidade Federal da Bahia (UFBA), jeilly.costa@ufba.br

²Universidade Federal da Bahia (UFBA), cedma.santos@gmail.com

³Instituto Federal do Norte de Minas Gerais (IFNMG), rogerio.l Luiz@ifnmg.edu.br

⁴Universidade Federal da Bahia (UFBA), karlaesquerre@ufba.br

Abstract

The Exame Nacional do Ensino Médio (ENEM) and education microdata are essential tools for analyzing educational performance at the national level. These data not only provide a broad overview of basic education across Brazil but also support detailed investigations at regional and local levels. This article presents an analysis of these data focusing on students from Bahia's state public schools, one of Brazil's most populous states. Using logistic regression and random forest models, significant patterns and correlations were identified between student performance on the ENEM and both socioeconomic variables and school characteristics. Factors such as family income, race, parental educational level, and teacher qualifications were found to have a direct impact on student outcomes. The analysis showed that students from wealthier families and with more educated parents tend to achieve better results. Similarly, schools with better-qualified teaching staff tend to achieve stronger academic results. These findings have significant implications for educational policy, emphasizing the need for initiatives that address socioeconomic disparities and enhance the quality of public education. The results are consistent with the existing literature and suggest that investments in teacher training programs and social inclusion policies may foster a more equitable and higher-quality educational environment. **Keywords:** ENEM, Machine Learning, Educational Data Mining, Logistic Regression, Random Forest.

Introdução

O Exame Nacional do Ensino Médio (ENEM), criado em 1998, surgiu como uma ferramenta crucial para avaliar não apenas a aprendizagem dos estudantes ao término do ciclo básico de ensino, mas também para fornecer indicadores abrangentes da qualidade do ensino em todo o território nacional. Ao longo dos anos, consolidou-se como uma referência fundamental no cenário educacional brasileiro, tanto para alunos quanto para instituições de ensino.

Além de servir como um instrumento de avaliação da aprendizagem, a nota do ENEM adquiriu uma nova dimensão ao ser aceita como critério de seleção para ingresso em instituições de ensino superior. No processo de inscrição para o exame, os estudantes são solicitados a preencher um formulário com informações pessoais e familiares. Questões sobre renda familiar, raça e gênero estão entre os itens presentes nesse formulário, fornecendo dados valiosos para análises socioeconômicas e demográficas.

A alta participação de alunos concluintes do Ensino Médio no ENEM possibilita a obtenção de um panorama abrangente da qualidade da educação básica no país. Além disso, o formulário preenchido pelos próprios alunos proporciona uma fonte rica de informações para estudos e pesquisas sobre diversos aspectos relacionados à educação.

Neste sentido, este trabalho faz uso analítico desse material para obter informações relevantes sobre os aspectos envolvidos no desempenho escolar dos alunos que se declara-

ram concluintes do ciclo básico de ensino no estado da Bahia durante o período de 2009 a 2019. Essa análise, conhecida como mineração de dados, permite identificar tendências, padrões e possíveis áreas de melhoria no sistema educacional, contribuindo assim para o desenvolvimento de políticas e estratégias mais eficazes no campo da educação e o uso mais eficiente de esforços governamentais.

A Bahia, de acordo com o censo do IBGE de 2022, é o quarto estado mais populoso do país e tem composição étnica majoritariamente parda (57,3%) e preta (22,4%). Brancos (19,6%) e indígenas (0,6%) compõem em menor proporção o panorama étnico do estado. O Brasil possui composição étnica majoritária de pessoas pardas (45,3%) e brancas (43,5%), seguida de pessoas pretas (10,2%), indígenas (0,6%) e amarelas (0,4%) em menor quantidade. Se torna pertinente, assim, olhar com especial atenção para o estado visto que sua composição racial difere da composição étnica brasileira, em especial com relação às populações preta e branca.

Estudos utilizando mineração de dados educacionais já foram feitos buscando analisar quais fatores socioeconômicos e espaciais mais impactam no desempenho escolar de discentes do ensino secundário público brasileiro. Silva Filho *et al.* (2023a) investigaram extensivamente essas questões para escolas estaduais brasileiras. Cunha *et al.* (2016) e Silva Filho *et al.* (2019) empregaram mineração de dados para detectar, respectivamente, causas de abandono escolar e de boa performance escolar em institutos federais de educação. Rodrigues *et al.* (2021) avaliaram a performance em todo o território nacional de estudantes no ENEM em 2017 e 2018. O presente trabalho contribui no desenvolvimento da compreensão da conjuntura educacional do país ao examinar dados de um dos estados mais populosos do Brasil, avaliados em um longo período de tempo - 11 anos contínuos - ao mesmo tempo em que obtém um panorama mais amplo da situação da educação pública no estado nos anos finais do ciclo básico de ensino.

Objetivo

O presente trabalho procurou obter informações relevantes sobre o aprendizado dos estudantes concluintes do ensino médio das escolas públicas do estado da Bahia através da investigação das suas características sociais e econômicas, bem como dos aspectos estruturais das escolas em que estudaram no último ano do ciclo básico. Para tanto, por meio da mineração de dados educacionais, buscou compreender quais desses fatores socioeconômicos e estruturais mais influenciaram no desempenho desses estudantes com base nas notas gerais obtidas no ENEM durante o período analisado.

Material e Método

Dados

Para a realização dessa pesquisa, foram utilizados dois bancos de dados com dados pré-processados e harmonizados de bases públicas (Silva Filho, 2022). Neles estão concatenados dados dos microdados do ENEM e dos microdados da educação básica de todos os estados brasileiros entre os anos de 2009, quando o exame sofreu reformulação para permitir sua comparação ao longo do tempo (Silva Filho *et al.*, 2023b), e 2019, último ano presente em ambas as bases de dados usadas (Silva Filho *et al.*, 2023a). Os bancos de dados contém, dentre outros dados, informações sobre a nota geral e das demais áreas de conhecimento de avaliação do exame, além da nota da redação. Há também informações socioeconômicas dos estudantes, da infraestrutura das escolas e do corpo docente das mesmas.

O primeiro banco de dados explorado contém informações individuais de cada estudante da educação básica que prestou o ENEM no período. O segundo contém dados com informações consolidadas das escolas.

Filtragem e Pré-Processamento

Na análise exploratória, foram utilizados os dados individuais de cada aluno. Para obter apenas os resultados dos alunos das escolas estaduais baianas prestando o ENEM no último ano do ciclo básico, foi necessário filtrar os dados para retirar da amostra as informações de alunos de outros estados da federação e de colégios particulares, municipais e de institutos federais, bem como dos alunos não-concluintes, ou “treineiros” - estudantes que ainda não estão no último ano do Ensino Médio e fazem a prova apenas como forma de preparação.

Na etapa da modelagem e mineração dos dados utilizando os modelos de classificação, foi utilizado apenas o banco de dados com informações das escolas. De modo a obter dados contendo apenas escolas com quantidade significativa de alunos para mineração, com pelo menos uma mínima infraestrutura básica de funcionamento e para excluir estudantes não-concluintes, os seguintes filtros foram aplicados:

- Escolas com menos de 10 alunos presentes no banco de dados foram removidos;
- Escolas sem rede de esgoto, água e energia elétrica foram removidas;
- Apenas escolas com alunos concluintes em cada ano foram mantidas.

Além disso, foram removidas da amostra as variáveis com alta frequência de moda (maior do que 90%). Essas variáveis apresentam baixa variabilidade, pois um único valor ocorre em mais de 90% dos registros, o que reduz sua relevância para a análise.

Antes da aplicação dos filtros, a base de dados contava com 220.319 registros de escolas de todo o Brasil. Ao todo, foram encontrados 12.473 registros de escolas baianas, número que foi reduzido para 8.967 ao final.

Para que pudessem ser empregados os métodos de classificação, foi feita uma etapa de pré-processamento antes da realização da modelagem dos dados. Neste estudo, as notas gerais do ENEM foram utilizadas como base para avaliar o desempenho dos estudantes. Para fins de análise, esses valores foram dicotomizados, classificando os resultados como de bom ou insatisfatório desempenho. Para tanto, os dados foram divididos em quartis. A estratégia de separar os dados em quartis foi escolhida porque estes são resistentes aos valores extremos (Johnson *et al.*, 2002) que estão presentes durante todos os anos analisados. Foi atribuído o valor 1 a todas as notas acima do terceiro quartil em cada ano. Para as demais notas, foi atribuído o valor 0. A proporção final das notas após o pré-processamento foi de 75%/25%.

Análise Exploratória de Dados

Visando a uma melhor compreensão da composição étnica e da segmentação por sexo do corpo discente do estado, foi empregada a análise exploratória de dados com os dados individuais de cada aluno presente no banco. Os estudantes foram agrupados com base no sexo - feminino ou masculino - e na raça - não declarada, branca, preta, parda, amarela ou indígena - ao qual indicaram pertencer no ato de inscrição do exame. Em seguida, foi avaliado o desempenho desses grupos com relação à nota geral.

Modelagem e Importância das Variáveis

Foi necessário selecionar modelos já testados e validados na literatura para a mineração dos dados. A regressão logística é um método amplamente utilizado na classificação de dados educacionais (Gorostiaga e Rojo-Álvarez, 2016). Rodrigues *et al.* (2021) destacaram a clareza e a concisão da obtenção de resultados através do uso deste modelo para análise de características relacionadas ao desempenho estudantil. O mecanismo conhecido como floresta aleatória se caracteriza como um método de classificação que se destaca pela baixa incerteza e alta performance (Guidotti *et al.*, 2018). Silva Filho *et al.* (2023b) utilizaram da floresta aleatória para minerar dados educacionais de múltiplos anos do ENEM. Assim, foram escolhidos para mineração os métodos de regressão logística, aplicado sem seleção de variáveis, e floresta aleatória pelo emprego já consolidado desses modelos na descoberta de informações educacionais.

A abordagem de winsorização dos dados foi aplicada para valores acima de 97,5% e abaixo de 2,5% da distribuição nas variáveis numéricas de modo que os valores da cauda, isto é, os valores extremos de cada variável, não influenciassem na robustez qualitativa, como Wilcox (2017) pontuou. Por fim, essas variáveis foram normalizadas entre 0 e 1.

A etapa de modelagem foi realizada com 37 variáveis. A relação dessas variáveis pode ser verificada no **Apêndice**. Foram utilizadas as implementações da biblioteca Scikit-learn da linguagem Python com os hiperparâmetros padrão da aplicação para a floresta aleatória e para a regressão logística.

O método escolhido para avaliação da performance dos modelos foi o escore da Área Sob A Curva Característica de Operação do Receptor (AUC ROC). Esta métrica de avaliação consegue mensurar o desempenho de modelos de classificação binários, como os usados neste trabalho, e retornar um único valor (Silva Filho e Adeodato, 2019). O valor final do escore AUC, avaliado e ajustado para cada ano, foi obtido com base na média desse escore em cada iteração considerando as 5 dobras de uma validação cruzada *K-fold* estratificada. Esta estratégia foi utilizada tanto para a regressão logística quanto para a floresta aleatória. A validação estratificada se mostrou útil uma vez que os dados da nota geral ficaram propositalmente desbalanceados após a categorização.

Com a etapa de modelagem concluída, foi necessário escolher uma forma de desvendar quais as variáveis mais importantes para os modelos. O procedimento utilizado para tal fim foi a importância das variáveis. Este procedimento, de acordo com Guidotti *et al.*, 2018, é um modo eficaz de descobrir a relevância de uma determinada característica para o modelo, pois quantifica o peso das variáveis utilizadas no processo de modelagem. A importância das variáveis para a floresta aleatória foi calculada por meio da Importância de Gini, baseado na redução da impureza nos nós, enquanto que para a regressão logística foi utilizado a magnitude dos coeficientes das variáveis para esse fim. A média dos coeficientes, suas razões de chance e os respectivos intervalos de confiança de 95% podem ser consultados no **Apêndice**.

Linguagem de Programação

Todas as etapas desta pesquisa foram realizadas utilizando a linguagem de programação Python 3.11.11 e suas bibliotecas. Para a etapa de filtragem e análise exploratória, foram empregadas as bibliotecas Pandas, Matplotlib, Seaborn e NumPy. Já na etapa de modelagem e avaliação da importância das variáveis, utilizou-se a biblioteca Scikit-learn juntamente com as bibliotecas já citadas.

As bibliotecas Pandas e Numpy foram escolhidas, respectivamente, por sua eficiência na manipulação e análise de dados e pela sua capacidade de realizar operações matemáticas de forma eficiente. A Scikit-learn foi escolhida devido à sua especialização em machine learning e modelagem estatística. Por fim, Seaborn e Matplotlib foram escolhidas pelas suas funcionalidades na criação de gráficos estatísticos.

Resultados e Discussão

A análise exploratória de dados revelou que, durante o período de tempo avaliado, a maior parte dos estudantes inscritos no ENEM foram do sexo feminino e autodeclarados pardos e pretos (ver **Tabela 1**). Em alguns anos, o número das estudantes inscritas chegou a ser o dobro do número dos estudantes. A nota geral das estudantes foi, em média, 13,59 pontos abaixo da nota geral dos estudantes do outro sexo.

Tabela 1: Distribuição percentual anual dos inscritos no ENEM na Bahia por sexo.

Sexo (%)	Ano										
	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Feminino	68,20	66,98	66,22	64,90	64,84	64,05	63,36	63,68	63,12	62,07	62,62
Masculino	31,80	33,02	33,78	35,10	35,16	35,95	36,64	36,32	36,88	37,93	37,38

Fonte: De autoria própria, 2025

A **Figura 1** apresenta a distribuição das notas gerais de estudantes do sexo masculino e feminino. Ao longo de todo o período, os estudantes do sexo masculino obtiveram medianas superiores às das estudantes do sexo feminino e essa diferença se estende ao longo dos anos. Esses resultados sugerem a persistência de desigualdades de desempenho entre os sexos, podendo suas causas serem multifatoriais.

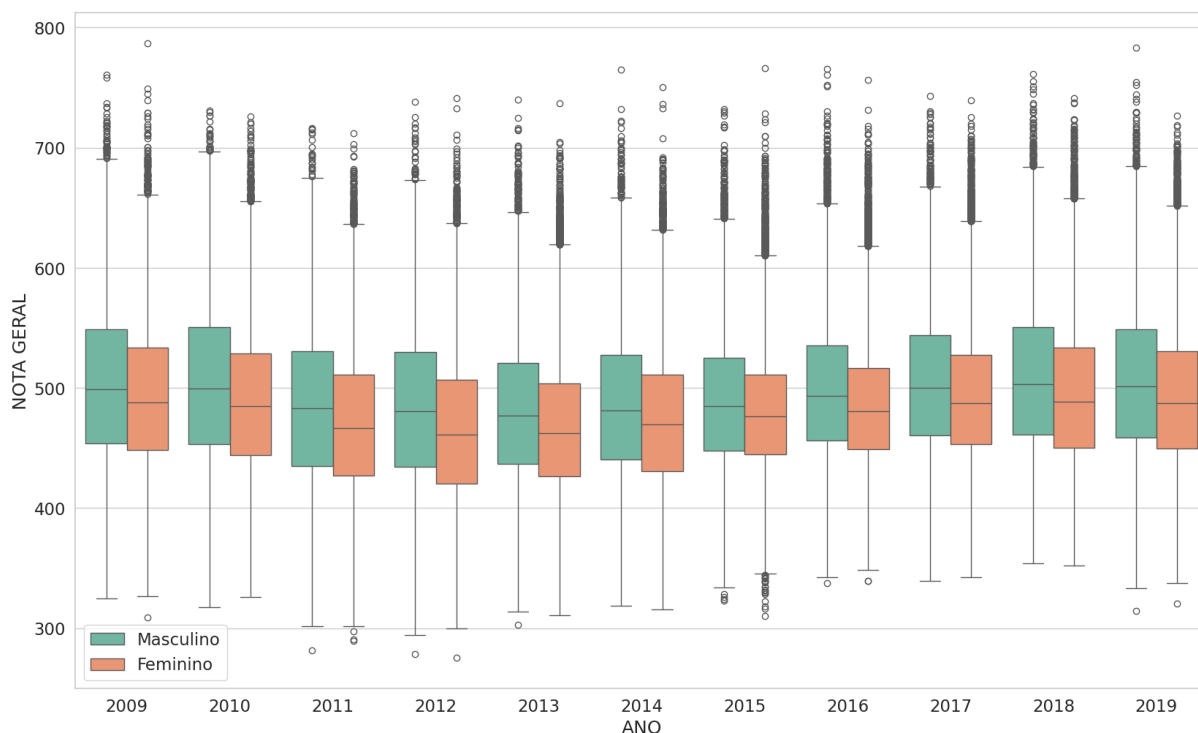


Figura 1: Comparação da nota geral por sexo.

Fonte: De autoria própria, 2025

Com relação à raça dos participantes, a proporção de estudantes pardos e pretos aumentou ao longo do tempo, atingindo, juntos, mais de 78% dos inscritos em 2019 (ver Tabela 2). Já a participação de estudantes brancos permaneceu relativamente estável, enquanto a categoria “não declarada” apresentou queda acentuada — de 15,2% em 2009 para apenas 2,8% em 2019 — o que pode indicar avanços no processo de autodeclaração racial, possivelmente influenciados por políticas afirmativas.

Tabela 2: Distribuição percentual anual dos inscritos no ENEM na Bahia por raça/cor.

Raça (%)	Ano										
	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Não Declarada	15,16	4,92	3,71	2,83	2,38	2,09	1,87	2,09	2,89	2,95	2,82
Branca	13,61	16,07	16,32	17,03	15,17	14,65	13,42	13,20	14,48	14,12	14,64
Preta	21,45	24,01	25,15	24,05	25,63	25,85	27,84	28,08	25,40	25,98	26,86
Parda	46,25	51,98	51,40	52,48	52,82	53,73	53,22	53,25	53,67	53,24	51,83
Amarela	2,70	2,08	2,53	2,80	3,04	2,88	2,92	2,72	2,88	2,89	3,07
Indígena	0,82	0,95	0,89	0,81	0,95	0,80	0,73	0,66	0,68	0,81	0,79

No que se refere a raça e o desempenho, apresentado na **Figura 2**, os estudantes declarados brancos mantiveram, ao longo do período, medianas de nota geral superiores às dos demais grupos raciais. Os estudantes declarados amarelos e pardos obtiveram desempenho intermediário, seguidos pelos estudantes pretos. Os estudantes autodeclarados indígenas obtiveram as piores notas gerais com exceção de 2011 e 2014 no qual estudantes sem raça declarada obtiveram as piores notas.

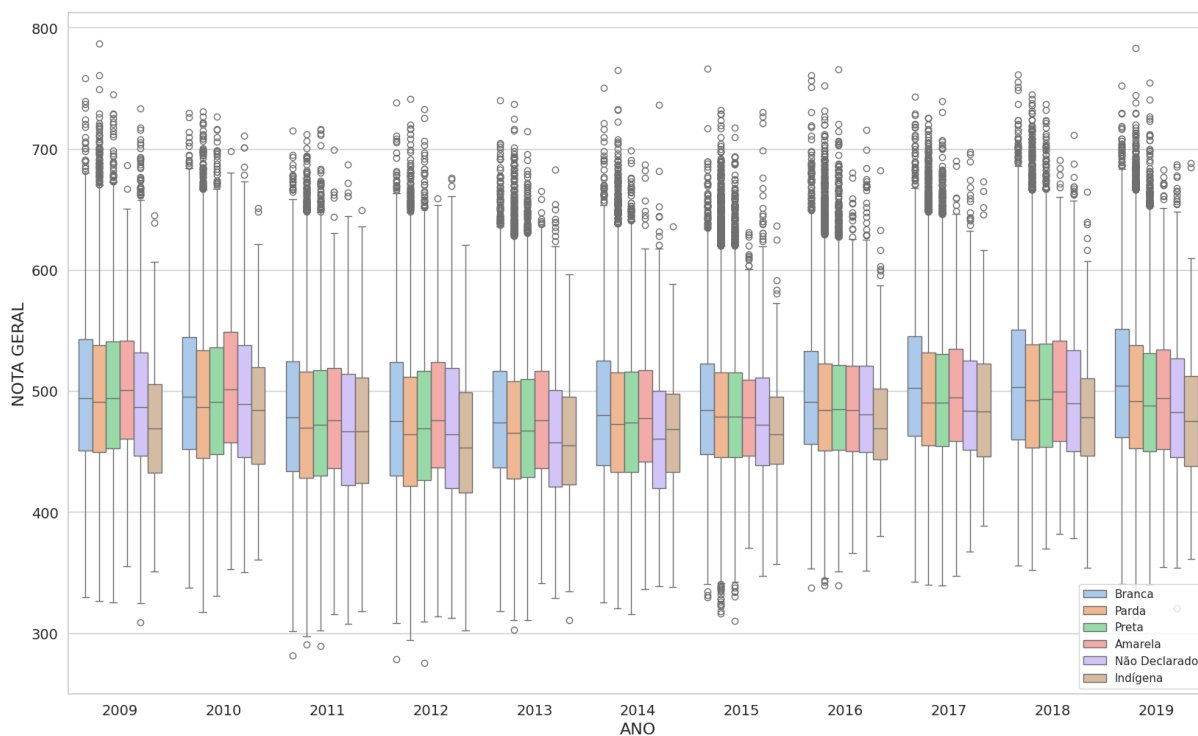


Figura 2: Comparação da nota geral por raça.

Fonte: De autoria própria, 2025

Com relação à modelagem, a **Tabela 3** separa a quantidade de registros presentes por ano, com diferenciação da quantidade de escolas nacionais e estaduais após as filtrações.

Tabela 3: Quantidade de escolas nacionais e estaduais por ano.

Ano	Escolas no Banco de Dados	Escolas Baianas pós Filtros
2009	15.413	604
2010	17.957	723
2011	19.584	760
2012	20.120	757
2013	21.085	845
2014	21.383	860
2015	21.843	897
2016	22.476	920
2017	21.692	900
2018	20.617	860
2019	18.149	841

Fonte: De autoria própria, 2025

Para cada ano analisado, foram treinados ambos os modelos, utilizando apenas os dados correspondentes àquele ano. A avaliação do desempenho foi realizada por meio de

validação cruzada *K-fold*, aplicada aos dados de cada ano separadamente. Os valores de AUC apresentados na **Figura 3** correspondem às médias dos escores obtidos nas iterações da validação cruzada, sendo calculados com base nas previsões feitas sobre os subconjuntos de teste de cada dobra. Dessa forma, os AUCs refletem o desempenho médio dos modelos em dados de teste não vistos durante o treino dentro do processo de validação cruzada.



Figura 3: Comparação AUC score do modelo de regressão logística e do modelo de floresta aleatória. Comparação AUC score do modelo de regressão logística e do modelo de floresta aleatória.

Fonte: De autoria própria, 2025

O escore AUC ROC é um valor que mede quão bem um modelo de classificação conseguiu desempenhar na predição dos dados. Ele avalia a sensibilidade, ou seja, de acordo com James *et al.* (2023), quão bem o modelo conseguiu detectar corretamente os resultados avaliados como positivos (1 para o caso da análise), e a especificidade, métrica que avalia quão bem o modelo conseguiu detectar corretamente os resultados negativos - 0 para o caso da análise. Na análise empregada, o modelo de regressão logística conseguiu melhor desempenho, ou seja, obteve maior AUC score, do que o modelo de floresta aleatória.

Após a obtenção do desempenho dos modelos, foram identificadas as variáveis com

maior impacto considerando todos os anos analisados. A comparação gráfica das variáveis mais relevantes identificadas por cada modelo está apresentada na **Figura 4**. No caso da regressão logística, a importância das variáveis foi estimada com base no valor absoluto dos coeficientes associados a cada uma delas. O **Apêndice** apresenta o valor original dos coeficientes da regressão, incluindo seus sinais.

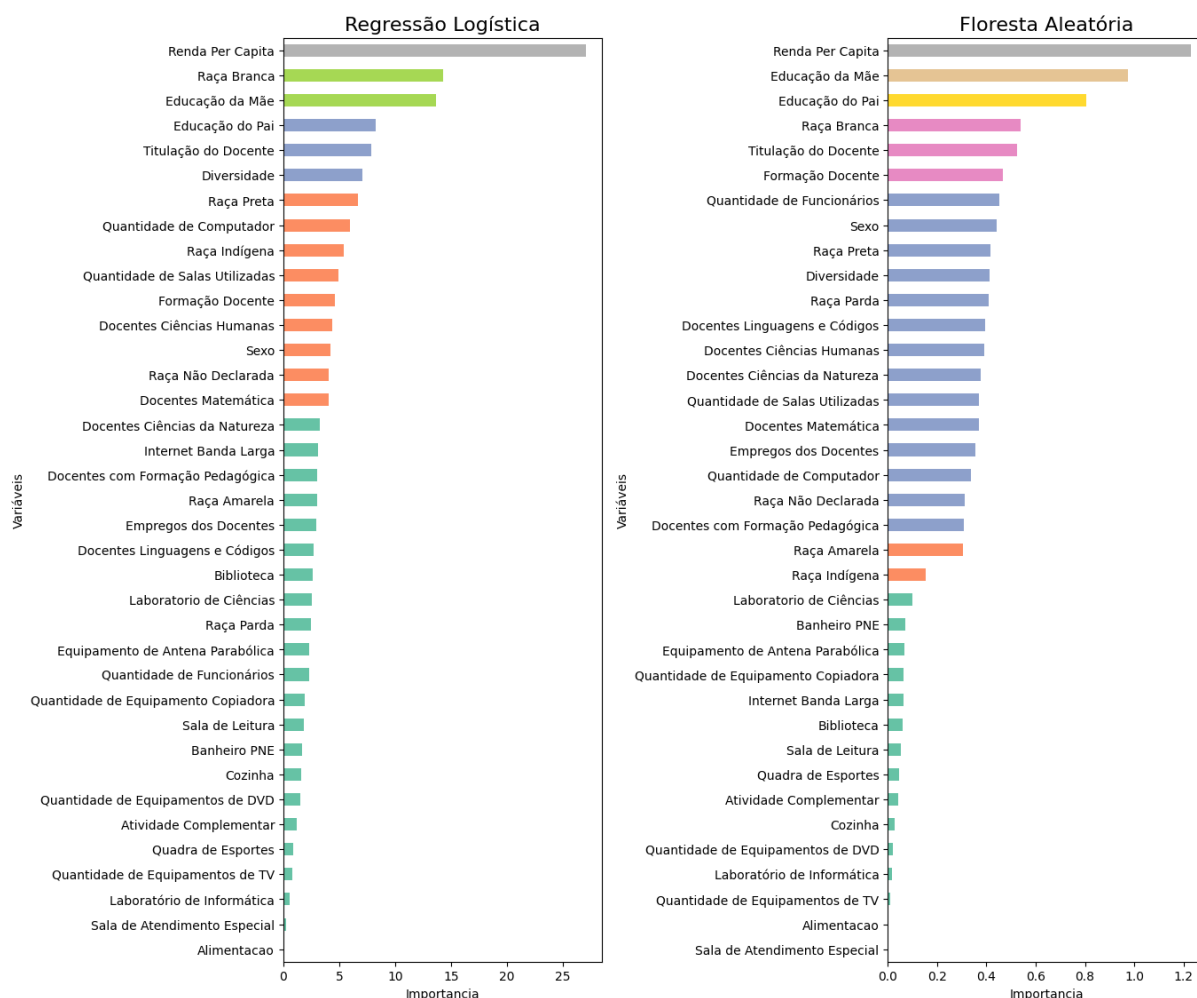


Figura 4: Avaliação da importância de cada variável para todo o período de tempo analisado pelo modelo de regressão logística e pelo modelo de floresta aleatória.

Fonte: De autoria própria, 2025

A análise detalhada dos resultados revela pontos interessantes sobre a relevância das características socioeconômicas e educacionais no desempenho dos estudantes ao longo dos anos. Embora a renda per capita seja identificada como o fator de maior relevância em ambos os modelos, é possível notar que cada modelo atribui uma importância diferente a essa variável.

Além disso, ao comparar a ordem de importância das variáveis entre os modelos, observa-se divergências. Enquanto o modelo de regressão logística destaca a **renda per capita** como o fator mais determinante, seguido pela **raça branca**, **educação da mãe**, **educação do pai** e **titulação docente**, o modelo de floresta aleatória prioriza a **renda**

per capita, seguida pela **educação da mãe**, **educação do pai**, **raça branca** e **titulação do professor**, ou seja, mesmas características mas em ordens ligeiramente diferentes. Essa diferença enfatiza a complexidade dos dados analisados e a necessidade de considerar diferentes modelos para avaliação.

No entanto, é interessante notar que, num todo, ambos os modelos concordam sobre as cinco variáveis mais importantes, indicando uma certa consistência nas características que mais influenciaram o desempenho dos alunos ao longo do tempo.

A análise ressaltou resultados já observados em estudos anteriores que buscaram avaliar o mesmo tema. Em Silva Filho *et al.* (2023a) Silva Filho *et al.* (2023b), estudos realizados com escolas públicas nacionais, a forte influência de atributos como renda, raça e educação dos pais foi destacada no desempenho escolar. Fernandes *et al.* (2019) evidenciaram que o meio social dos estudantes influencia diretamente no sucesso escolar. Cunha *et al.* (2016) e Rodrigues *et al.* (2021) apontaram fatores como renda familiar e raça como aspectos importantes na boa performance escolar. O estudo de Silva Filho e Adeodato (2019), realizado com dados do ENEM, encontrou relação entre a educação do pai e a renda familiar com o êxito no exame. Gomes e Jelihovschi (2020) também sinalizaram a renda da família do aluno como relacionada à obtenção de uma maior pontuação no ENEM.

Por outro lado, a questão da titulação dos professores foi pouco mencionada ou está ausente nos estudos citados. Silva Filho *et al.* (2023b) refletem que esta característica pode não estar sendo computada em estudos na área devido à pouca proporção de professores com alta titulação na educação pública. Machado (2014) aponta o problema da falta de titulação dos professores que atuam na educação básica no estado da Bahia, com apenas 54% dos profissionais apresentando graduação acadêmica. Além disso, ela também aborda os problemas com a formação especializada de professores no estado. Na época analisada pela pesquisadora, o estado apresentou índices de formação especializada menor proporcionalmente em comparação aos números registrados no Brasil e na região Nordeste. Neste sentido, considerando os resultados evidenciados pelos dois modelos da importância da formação docente no sucesso profissional dos estudantes, entende-se que a melhora no desempenho dos discentes de escolas públicas estaduais passa pela melhoria da preparação docente para o ensino.

Conclusão

O presente estudo utilizou da mineração de dados educacionais para investigar a relação entre as características estruturais das escolas públicas do estado da Bahia, as características socioeconômicas dos alunos dessas escolas e o desempenho desses estudantes no Exame Nacional do Ensino Médio (ENEM). Os dois modelos de classificação empregados nesta análise enfatizaram a **renda per capita**, a **raça branca**, o **grau de educação**

dos pais do aluno e a **titulação do professor** como os elementos mais importantes para o bom desempenho estudantil. Os resultados apresentados possibilitam um melhor entendimento das estatísticas de desempenho dos estudantes com recortes de sexo e raça, bem como o impacto que o quadro docente e a infraestrutura dos ambientes de ensino exercem no aprendizado estudantil, podendo apoiar a proposição de estratégias e de políticas públicas educacionais.

Agradecimentos

Esta pesquisa foi desenvolvida no âmbito de dois projetos: Educação Científica para a Justiça Social: Ciência de Dados e Inteligência Artificial no Estudo de Questões Raciais, Sociais e de Gênero no Ensino Público, vinculado ao Edital CNPq/MCTI/FNDCT N° 18/2021, e Meninas na Ciência (de Dados) e Inteligência Artificial: STEM, Ética e Inovação Social, financiado pela Chamada CNPq N° 31/2023 – Modalidade Projeto Individual. Agradecemos às agências de fomento, colaboradores e instituições parceiras pelo apoio essencial à realização deste trabalho.

Referências

- CUNHA, J. A. da; MOURA, E.; ANALIDE, C. **Data Mining in Academic Databases to Detect Behaviors of Students Related to School Dropout and Disapproval**. *Advances in Intelligent Systems and Computing*, p. 189–198, 2016.
- FERNANDES, E. et al. **Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil**. *Journal of Business Research*, v. 94, p. 335–343, jan. 2019.
- GOMES, C. M. A.; JELIHOVSCHI, E. **Presenting the Regression Tree Method and its application in a large-scale educational dataset**. *International Journal of Research & Method in Education*, v. 43, n. 2, p. 201–221, 13 ago. 2019.
- GOROSTIAGA, Arantza; ROJO-ÁLVAREZ, José Luis. **On the use of conventional and statistical-learning techniques for the analysis of PISA results in Spain**. *Neurocomputing*, v. 171, p. 625–637, 1 jan. 2016.
- GUIDOTTI, R. et al. **A Survey of Methods for Explaining Black Box Models**. *ACM Computing Surveys*, v. 51, n. 5, p. 1–42, 22 ago. 2018.
- Instituto Brasileiro de Geografia e Estatística (IBGE). **Censo Demográfico 2022**. Rio de Janeiro: IBGE, 2022.

JAMES, Gareth; WITTEN, Daniela; HASTIE, Trevor; TIBSHIRANI, Robert; TAYLOR, Jonathan. **An Introduction to Statistical Learning: with Applications in Python.** [s.l.] Springer, 2023.

JOHNSON, R. A. et al. **Applied multivariate statistical analysis.** 5. ed. Pearson, 2002.

MACHADO, Cristiane Brito. **A Política Nacional de Formação de Professores da Educação Básica: a implementação do PARFOR-Presencial no Estado da Bahia** 208 pp. ill. 2014. Tese (Doutorado) – Faculdade de Educação, Universidade Federal da Bahia, Salvador, 2014.

RODRIGUES, D. et al. **Data Mining on the Prediction of Student's Performance at the High School National Examination.** PUCRS Repository (Pontifical Catholic University of Rio Grande do Sul), 1 jan. 2021.

SILVA FILHO, R.L. **Enem Census 2009 - 2019** 2022. Harvard Dataverse, V1, UNF:6:lvdu04mKdagEgl70si/7bA==[fileUNF]. Disponível em: <https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/WEWDHL>.

SILVA FILHO, R. L. C.; ADEODATO, P. J. L. **Data Mining Solution for Assessing the Secondary School Students of Brazilian Federal Institutes.** 2019 8th Brazilian Conference on Intelligent Systems (BRACIS), p. 574–579, out. 2019.

SILVA FILHO, Rogério Luiz Cardoso; BRITO, Kellyton; ADEODATO, Paulo Jorge Leitão. **A data mining framework for reporting trends in the predictive contribution of factors related to educational achievement.** Expert Systems with Applications, 2023

SILVA FILHO, R. L. C.; BRITO, K.; ADEODATO, P. J. L. **Leveraging Causal Reasoning in Educational Data Mining: An Analysis of Brazilian Secondary Education.** Applied Sciences, v. 13, n. 8, p. 5198, 21 abr. 2023.

WILCOX, R. R. **Introduction to Robust Estimation and Hypothesis Testing.** 4. ed. Cambridge: Elsevier, 2017.

Anexo

O código referente a este artigo encontra-se disponível no link <https://github.com/JdeACosta/Data-mining-Bahia>.

Apêndice

Variáveis utilizadas na modelagem

Variáveis	Coeficientes	Intervalo de Confiança [0,025 - 0,975]		Razão de Chance
		Inferior	Superior	
Renda Per Capita	2,8678	1,7963	3,9393	17,5983
Raça Branca	1,1619	-0,3834	2,7073	3,1961
Educação da Mãe	1,1333	-0,0523	2,5456	3,106
Titulação do Docente	0,6439	-0,2032	1,7771	1,9038
Educação do Pai	0,5045	-0,7252	1,9585	1,6562
Quantidade de Computador	0,4152	-0,4209	1,2513	1,5147
Diversidade	0,3887	-0,4819	1,5507	1,475
Quantidade de Salas Utilizadas	0,3308	-0,5692	1,3779	1,3921
Quantidade de Equipamentos de TV	0,2281	-0,4358	0,8920	1,2563
Sexo	0,2050	-0,5731	1,1368	1,2275
Formação Docente	0,2028	-0,7705	1,5139	1,2248
Biblioteca	0,1798	-0,2998	0,6954	1,1970
Laboratório de Ciências	0,1656	-0,2493	0,6137	1,1801
Sala de Atendimento Especial	0,1527	-0,4306	0,7359	1,1650
Equipamento de Antena Parabólica	0,1150	-0,2678	0,5207	1,1219
Empregos dos Docentes	0,0992	-0,6782	0,9509	1,1043
Quadra de Esportes	0,0710	-0,3728	0,5716	1,0736
Atividade Complementar	0,0637	-0,3444	0,5227	1,0657
Quantidade de Funcionários	0,0502	-0,8495	1,0336	1,0515
Docentes com Formação Pedagógica	0,0121	-0,9118	0,9451	1,0122
Laboratório de Informática	-0,0033	-0,5462	0,5351	0,9967
Quantidade de Equipamento Copiadora	-0,0088	-0,4491	0,4276	0,9912
Docentes Linguagens e Códigos	-0,0105	-1,0017	0,9556	0,9896
Banheiro PNE	-0,0273	-0,4186	0,3586	0,9731
Docentes Matemática	-0,0411	-1,0642	0,9136	0,9598
Cozinha	-0,0464	-0,6653	0,5262	0,9547
Sala de Leitura	-0,0549	-0,5428	0,4087	0,9466
Alimentação	-0,0626	-0,5769	0,4516	0,9393
Raça Amarela	-0,0963	-0,9508	0,6860	0,9082
Docentes Ciências da Natureza	-0,2133	-1,3614	0,5791	0,8079
Raça Parda	-0,2175	-2,9116	2,1140	0,8045
Raça Não Declarada	-0,2364	-1,4364	0,7862	0,7894
Quantidade de Equipamentos de DVD	-0,2545	-0,8584	0,2222	0,7753
Internet Banda Larga	-0,2660	-0,7896	0,2043	0,7664

Variáveis utilizadas na modelagem

Variáveis	Coeficientes	Intervalo de Confiança [0,025 - 0,975]		Razão de Chance
		Inferior	Superior	
Raça Indígena	-0,4618	-1,5174	0,2474	0,6301
Docentes Ciências Humanas	-0,4785	-1,5962	0,4265	0,6197
Raça Preta	-0,6546	-2,7539	1,3137	0,5196

Fonte: De autoria própria, 2025