ANÁLISE DAS CORRELAÇÕES NOS RESULTADOS DO ENEM:
UM ESTUDO SOBRE FATORES QUE AFETAM O DESEMPENHO DOS CANDIDATOS DO EXAME[1]

Gabriel Victor Amorim[2]

RESUMO

O acesso ao ensino superior no Brasil tem sido uma das principais ferramentas de emancipação social ao longo dos anos. Desde 2010 o acesso a grande parte das vagas nas instituições públicas brasileiras tem sido pelo SISU (Sistema de Seleção Unificada) através do Exame Nacional do Ensino Médio (ENEM). Este estudo de caso investiga a relação entre variáveis educacionais e socioeconômicas e o desempenho dos estudantes no ENEM no Brasil, analisando dados de 2018 a 2022. O tema central é a compreensão dos fatores que influenciam o desempenho dos candidatos. Diante do papel crucial do ENEM como principal meio de acesso ao ensino superior no país, surge a necessidade de entender como aspectos socioeconômicos e educacionais se entrelaçam nesse contexto. O estudo adota a metodologia CRISP-DM, aplicando técnicas de mineração de dados para identificar padrões nos microdados do exame. Os resultados revelam a influência de variáveis socioeconômicas na média final dos estudantes, ressaltando a importância de considerar esses fatores nas formulações de políticas educacionais.

Palavras-chave: mineração de dados; classificação; ENEM.

ABSTRACT

Access to higher education in Brazil has been one of the main tools for social emancipation over the years. Since 2010, access to a large portion of the slots in Brazilian public institutions has been through SISU (Unified Selection System) based on the National High School Exam (ENEM). This case study investigates the relationship between educational and socioeconomic variables and students' performance in the National High School Exam (ENEM) in Brazil, analyzing data from 2018 to 2022. The central theme is understanding the factors that influence candidates' performance. Given the crucial role of ENEM as the main gateway to higher education in the country, there is a need to understand how socioeconomic and educational aspects intertwine in this context. The study adopts the CRISP-DM methodology, applying data mining techniques to identify patterns in the exam's microdata. The results reveal the influence of socioeconomic variables on students' final scores, emphasizing the importance of considering these factors in the formulation of educational policies.

Keywords: Data mining; classification; ENEM (Brazilian National High School Exam).

1. INTRODUÇÃO

A educação é uma das ferramentas mais importantes para a mudança da realidade do indivíduo e da sociedade. Enquanto instrumento de transformação, assume papel de contínuo movimento de mudanças na estrutura da sociedade no

[1] Trabalho de Conclusão de Curso apresentado como requisito parcial para obtenção do grau de bacharel no Curso de Ciência e Tecnologia, Centro Tecnológico de Joinville (CTJ), Universidade Federal de Santa Catarina (UFSC), sob orientação da Prof. Dr. Benjamin Grando Moreira

[2] Graduando como Bacharel em Ciência e Tecnologia. E-mail: gabriel.amorim777@gmail.com

presente e como resposta às questões que o futuro reserva. Freire (2017) reflete que há contradição na realidade, e que apesar dos desafios das mudanças sociais, a sociedade é produto da ação humana. O homem é consciente, racional e capaz de transformar a si e ao mundo, de modo que a educação emancipatória viabilize as armas necessárias para a mudança da realidade.

A educação emancipatória é uma concepção de educação que tem como objetivo a formação de sujeitos críticos e autônomos, capazes de compreender e transformar a realidade social. Essa concepção é inspirada no pensamento de Paulo Freire, que defende uma educação libertadora, que não se limita à transmissão de conteúdos, mas que também busca despertar a consciência crítica dos alunos. Nesse contexto, conforme observado por Ristoff e Giolo (2006) se insere a educação superior, com a missão de consolidar uma nação soberana, democrática, inclusiva e capaz de gerar emancipação social. Ao avaliar a realidade brasileira, o Instituto do Sindicato das Entidades Mantenedoras de Estabelecimentos de Ensino Superior no Estado de São Paulo - SEMESP (2021) observa expressiva evolução dos índices de empregabilidade e remuneração dos indivíduos que finalizaram o ensino superior.

De acordo com o Brasil (2023), atualmente, o principal caminho para ingresso no ensino superior é através do ENEM (Exame Nacional do Ensino Médio), que oportuniza aos estudantes concorrer para vagas pelo SISU (Sistema de Seleção Unificada), PROUNI (Programa Universidade Para Todos) e FIES (Fundo de Financiamento Estudantil).

Dada a importância do ensino superior para a transformação social e sendo o ENEM o maior exame para ingresso nas instituições públicas brasileiras, o presente estudo busca, a partir dos dados informados pelos candidatos no questionário de inscrição para o ENEM, bem como a partir de suas notas no exame, estimar a contribuição dos aspectos educacionais e socioeconômicos na média final de um candidato.

Diferentemente de grande parte dos estudos sobre o tema, como os realizados por Watanabe (2019), Adeodato e Filho (2020) e Silva (2021), que em geral olham para apenas um único ano de dados disponíveis, no presente trabalho são considerados os últimos cinco anos do exame simultaneamente. Além disso, ajustes de base que reduziam significativamente os dados avaliados com o intuito de utilizar menor capacidade computacional não foram realizados, permitindo assim a disponibilidade e uso de uma quantidade muito maior de dados para análise. De modo objetivo, espera-se responder ao seguinte questionamento: quais aspectos amplos, como condições socioeconômicas e educacionais se relacionam à performance dos estudantes no ENEM?

O capítulo a seguir discorre sobre a contextualização do problema, as bases utilizadas para a referida análise, o tratamento dos dados, premissas e definições bem como a metodologia utilizada para responder à questão levantada.

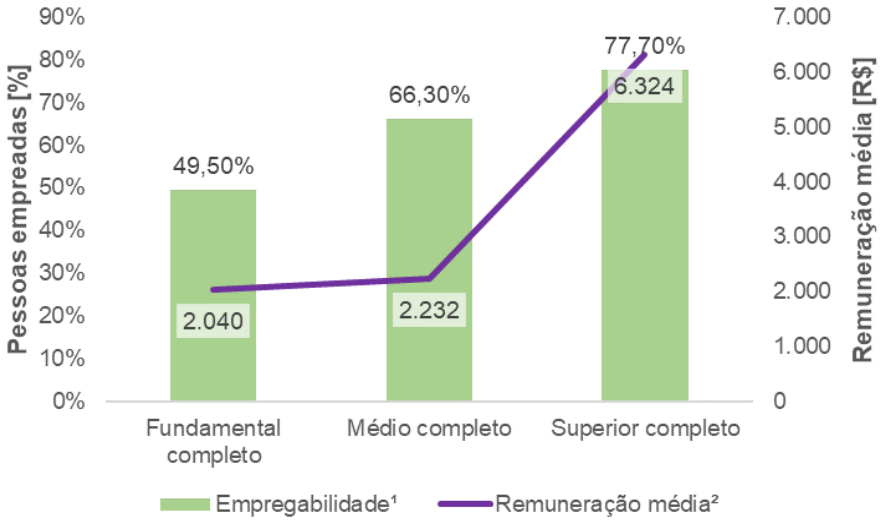
2. FUNDAMENTAÇÃO TEÓRICA

A análise realizada nesse estudo passa, previamente, pela necessidade da contextualização da dinâmica de ingresso no ensino superior no Brasil, bem como as características dos candidatos a uma vaga nas instituições. Por essa razão, o desenvolvimento a seguir será dividido em três partes: contextualização do cenário nacional, mineração de dados educacionais e revisão de trabalhos similares.

2.1. Panorama do acesso ao ensino superior no Brasil

O acesso ao ensino superior no Brasil vem crescendo nas últimas décadas, mas ainda é limitado a uma parcela da população. O retrato da realidade brasileira do ponto de vista de empregabilidade e remuneração média feito pelo Instituto SEMESP (2021) indica, no Mapa do Ensino Superior no Brasil, que pessoas com ensino superior completo apresentam 11,4 mais pontos percentuais de ocupação profissional que pessoas com ensino médio completo; além disso, pessoas com nível superior concluído possuem remuneração 183% superior que pessoas com médio e 210% maior que pessoas com ensino fundamental completo. Os dados são apresentados na Figura 1.

Figura 1 - Empregabilidade e remuneração por nível de escolaridade (2019)



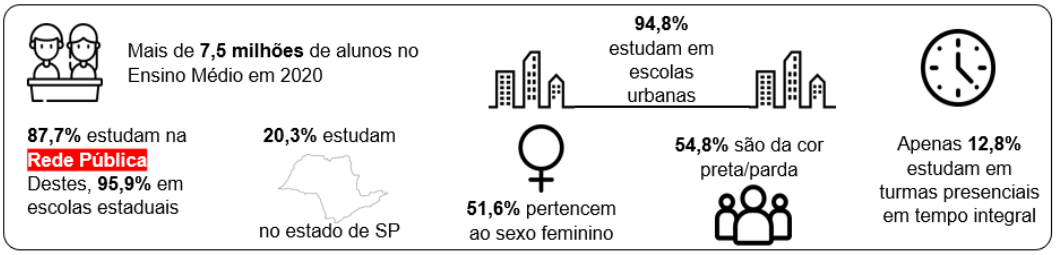
[1] Pessoas com 14 anos ou mais ocupadas na semana de realização do estudo em 2019.
[2] Considera valores de dezembro/19.

Fonte: adaptado do Instituto SEMESP (2021, p.37 - 38).

Para Barth (1990), a inclusão social no contexto da educação, é formada por um conjunto de ações que combatam a exclusão dos benefícios da educação na sociedade; a principal questão reside no fato de o acesso à educação ser estabelecido em padrões igualitários, contradizendo o perfil populacional, claramente consistente de diferenças. Por exemplo, de acordo com o Instituto SEMESP (2021), os alunos matriculados no ensino médio em 2020 eram, em sua maioria, provenientes de famílias de baixa renda (44,2%) e moravam em áreas urbanas (83,4%). A Figura 2 apresenta, a título de exemplo, algumas das características dos alunos matriculados no ensino médio no ano de 2020.

As características exemplificadas nos dados agregados dos alunos matriculados no Ensino Médio em 2020 apresenta, de modo claro, a heterogeneidade dos futuros candidatos a uma vaga no ensino superior.

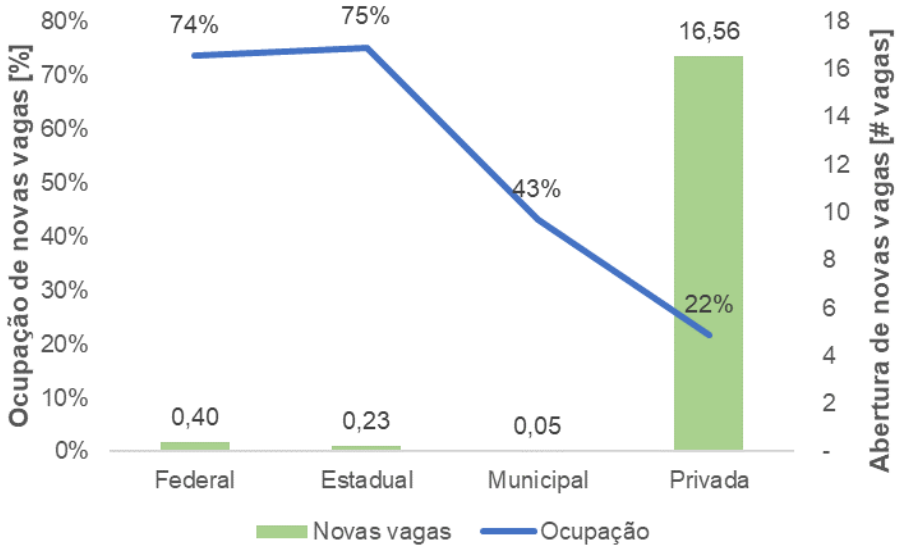
Figura 2 - Perfil dos alunos matriculados no Ensino Médio (2020)



Fonte: adaptado do Instituto SEMESP (2021, p. 52).

Conforme explicitado pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira - INEP (2023) no Censo da Educação Superior de 2022, a disponibilidade de vagas nas instituições privadas de ensino superior supera em mais de 23 vezes a disponibilidade nas instituições públicas e gratuitas; por outro lado, a ocupação das vagas nessas instituições é 50 pontos percentuais superior às instituições privadas. A Figura 3 explicita essa discrepância ao apresentar forte diferença entre a oferta de vagas entre as instituições públicas e as privadas (indicadas nas colunas) e, em contrapartida, índice de ocupação (indicado pela linha) menor nas privadas, revelando assim que, nessas instituições, há grande quantidade de vagas não preenchidas.

Figura 3 - Oferta e preenchimento de novas vagas no ensino superior (2022)



Fonte: adaptado do INEP (2022, p. 17 e 27).

A observação dos dados apresentados sugere lacuna entre a oferta de vagas e sua ocupação e, dado o cenário do perfil heterogêneo dos egressos do ensino médio, oportunidade de entender possíveis características que influenciam o resultado dos alunos no ENEM. As próximas seções abordam aspectos técnicos relacionados com análise de dados e apresenta uma revisão de trabalhos similares.

2.2. Mineração de dados educacionais

A Mineração de Dados Educacionais (EDM - Educational Data Mining) é um campo de pesquisa dedicado à aplicação de técnicas de mineração de dados para extrair conhecimento no contexto educacional. De acordo com Goyal e Vohra (2012), esses dados, oriundos do ambiente educacional, abrem caminho para uma variedade de aplicações, incluindo previsão de desempenho acadêmico, identificação de alunos em risco de evasão, personalização do ensino e avaliação de políticas educacionais.

Um exemplo prático de aplicação da EDM é a previsão de desempenho acadêmico. Através da análise de dados históricos, como notas e frequência, é possível identificar padrões que servem de base para antecipar o desempenho futuro dos alunos. Essas previsões, por sua vez, podem informar ações pedagógicas, tais como a oferta de reforço escolar ou orientação vocacional.

Um estudo conduzido por Barros *et al.* (2020), utilizou técnicas de mineração de dados para prever o rendimento de alunos em disciplinas de lógica de programação, alcançando uma acurácia de até 77% nas previsões.

A EDM tem sido considerada uma área de pesquisa promissora, com potencial para contribuir significativamente para aprimorar a qualidade da educação. Contudo, é imperativo que os usuários dessa abordagem estejam cientes dos potenciais riscos e benefícios associados, utilizando-a de maneira ética e responsável.

Nas subseções seguintes são exploradas técnicas de seleção de características e aplicação de algoritmos de classificação, técnicas que são aplicadas aos conjuntos de dados educacionais analisados neste estudo.

2.2.1. Seleção de características

A seleção criteriosa de variáveis desempenha papel crucial no processo de mineração de dados pois tem como objetivo identificar as variáveis significativas para o estudo em questão. Tais variáveis são vitais, sendo aquelas que têm maior probabilidade de influenciar os resultados do estudo. Segundo Braz e Soares (2016), diversas técnicas de seleção de variáveis estão disponíveis, cada uma com suas próprias vantagens e desvantagens.

Entre as técnicas mais frequentemente empregadas, destaca-se o teste qui-quadrado. Pinto e Rocha (2017) explicam que esse teste é particularmente valioso ao medir a associação entre duas variáveis categóricas. Quando aplicado à seleção de variáveis, o teste qui-quadrado assume a responsabilidade de avaliar a associação entre a variável de resposta e cada variável independente.

Um exemplo notável de algoritmo utilizado nesse contexto é o SelectKBest, uma técnica de seleção de variáveis fundamentada no teste qui-quadrado. Esse algoritmo se destaca ao selecionar as k variáveis que apresentam as pontuações de relevância mais elevadas. A pontuação de relevância de uma variável é determinada pelo valor do teste qui-quadrado específico para essa variável. Braz e Soares (2016) enfatizam que as variáveis com pontuação de relevância mais elevadas são aquelas que possuem uma associação mais forte com a variável de resposta.

Além do teste qui-quadrado, a técnica de seleção de variáveis usando Análise de Variância (ANOVA) também é considerada. ANOVA é particularmente

eficaz quando se lida com variáveis contínuas e categóricas, permitindo a avaliação das diferenças nas médias entre grupos. A combinação de ambas as técnicas, qui-quadrado e ANOVA, proporciona uma abordagem abrangente na identificação de variáveis influentes, enriquecendo a análise de associação no contexto da mineração de dados.

2.2.2. Algoritmos de classificação

Os algoritmos de classificação são modelos de aprendizado de máquina que aprendem a associar uma observação a uma classe. Segundo Braz e Soares (2016), os algoritmos de classificação são usados em uma variedade de aplicações, como reconhecimento de imagem, classificação de texto e filtragem de spam. A seguir, são explicitados os princípios de funcionamento de alguns algoritmos de classificação que estão disponíveis para implementação através da Linguagem Python utilizando a biblioteca scikit-learn conforme discutidos por Chollet (2017). São eles:

a) **Modelo SGD (Gradiente Descendente Estocástico):** O modelo SGD é um dos algoritmos de aprendizado de máquina mais populares e é usado em uma variedade de aplicações, incluindo classificação, regressão e aprendizado não supervisionado. O modelo SGD funciona interativamente, atualizando os parâmetros do modelo de acordo com o gradiente da função de perda. O gradiente é uma medida da direção em que a função de perda está aumentando. Atualizar os parâmetros do modelo de acordo com o gradiente ajuda o modelo a minimizar a função de perda;
b) **Modelo de regressão logística:** O modelo de regressão logística é um algoritmo de aprendizado de máquina supervisionado que usa uma função logística para modelar a relação entre uma variável independente e uma variável dependente categórica. No contexto da classificação, o modelo de regressão logística pode ser usado para prever a probabilidade de uma observação pertencer a uma determinada classe. Por exemplo, o modelo de regressão logística pode ser usado para prever a probabilidade de um aluno ser admitido em uma universidade;
c) **Árvore de decisão:** A árvore de decisão é um algoritmo de aprendizado de máquina supervisionado que usa um conjunto de regras para classificar observações. Uma árvore de decisão é composta de nós internos e nós externos. Os nós internos representam as regras de classificação e os nós externos representam as classes. No contexto da classificação, a árvore de decisão pode ser usada para classificar observações em várias classes. Por exemplo, a árvore de decisão pode ser usada para classificar observações de texto em diferentes classes de tópicos;
d) **Classificador SVM (máquina de vetor de suporte):** O classificador SVM é um algoritmo de aprendizado de máquina supervisionado que usa uma margem para classificar observações. A margem é uma distância entre as fronteiras de decisão e os pontos mais próximos das fronteiras de decisão.No contexto da classificação, o classificador SVM pode ser usado para classificar observações em duas ou mais classes. Por exemplo, o classificador SVM pode ser usado para classificar observações de imagem em diferentes classes de objetos.

2.2.3. Métricas de análise de resultados dos modelos

Um dos principais objetivos dentro da mineração de dados é a seleção de um modelo que defina métricas que possibilitem a comparação entre diferentes modelos. Em geral, a seleção de métricas irá depender do contexto avaliado e de sua aplicação, entretanto, algumas são geralmente usadas. A documentação da biblioteca Scikit-learn (2023) apresenta os conceitos abaixo:

- **Acurácia**: definida pela proporção de previsões corretas feitas pelo modelo. Pode não ser a mais adequada para todos os casos, por exemplo, quando se trata de situações desbalanceadas entre as classes e quando o custo dos erros de diferentes tipos não é uniforme;
- **Precisão**: definida pela proporção de exemplos positivos que foram classificados como positivos corretamente. Importante quando necessário evitar erros de tipo I (erro falso positivo, ocorre quando o modelo classifica erroneamente um exemplo negativo como positivo);
- **Revocação**: definido pela proporção de exemplos positivos que foram classificados como positivos corretamente. Importante quando necessário evitar erros de tipo II (erro falso negativo, ocorre quando o modelo deixa de identificar corretamente exemplos positivos);
- **F1-Score**: O F1-Score é uma combinação da precisão e da revogação. É mais equilibrado do que a precisão ou a revocação isoladamente.

2.3. Revisão de trabalhos similares

O uso da mineração de dados para geração de percepções, explicação de contextos e predição de resultados tem sido utilizada em diversos estudos correlacionando dados educacionais e informações dos estudantes. No Brasil, em especial, a existência de um exame do porte do ENEM que abrange todo o território nacional, ano após ano, desde 1998, tem fomentado a realização de diversos estudos utilizando a base de dados divulgada pelo INEP. Watanabe (2019) realizou testes a partir dos microdados do ENEM de 2016, realizando testes para definição de correlação entre as características e sugerindo o modelo de regressão *Random Forest*. Para esse estudo, o objetivo era fornecer sugestão de ferramenta para análise dos dados em questão, sem a pretensão de fornecer resultado conclusivo.

A análise da influência de características e condições dos alunos sobre os resultados em exames nacionais foram realizadas também no exterior. Brophy e Good (1986), Chudgar, Luschei e Fagioli (2012) e Coleman, Campbell *et al* (1966), realizaram estudos com dados educacionais americanos e observaram alta correlação entre os resultados dos alunos e a infraestrutura das escolas, características socioeconômicas e culturais e o nível de escolaridade dos pais e professores.

Adeodato e Filho (2020) realizaram pesquisa similar para o Brasil com dados do ENEM e utilizaram como base os microdados do ENEM 2017 para treino do modelo e os dados de 2018 para a efetiva análise dos resultados e frisaram o interesse na instituição de ensino da qual o aluno é proveniente, de modo que dados relacionados à instituição disponíveis no Censo do Ensino Médio foram utilizados. Para os autores, tal qual nos estudos norte-americanos (citados no parágrafo

anterior), o foco do estudo residia em três pilares: infraestrutura da escola de origem (foco dos principais investimentos em educação no Brasil), educação humana (mãe, pai e professores) e condições socioeconômicas e culturais.

Adeodato e Filho (2020) evidenciaram a menor importância da infraestrutura para a definição dos resultados no ENEM (excluindo possíveis questões pedagógicas provenientes de infraestrutura diferenciada). Além disso, os autores observaram uma alta correlação entre o nível de escolaridade agrupada de pais, mães e professores e o resultado do exame, entretanto, ao entrar em detalhe neste grupo, foi observado que a escolaridade dos professores tem pouca influência sobre o resultado dos alunos no Brasil.

Silva (2021) realiza estudo similar, entretanto, utiliza ferramentas para seleção de atributos, classificação, clusterização e associação. Utilizando os microdados do ENEM de 2019, o autor apresenta como principais parâmetros a serem analisados o tipo de escola, a renda familiar, escolaridade dos pais e a existência de computador no domicílio. Para a análise em questão, foram filtrados apenas alunos que concluíram o Ensino Médio no ano do exame. Essa filtragem reduz consideravelmente as informações disponíveis para análise, pois exclui os alunos que não concluíram o Ensino Médio no ano do exame, o que representa uma parcela significativa da população.

O diferencial do estudo aqui realizado é utilizar como insumo a base de dados histórica do ENEM entre os anos de 2018 e 2022, a despeito do custo computacional envolvido no processamento dos dados. Além disso, diferentemente do que foi observado em algumas análises, como o objetivo é entender fatores que contribuem para maior chance no ingresso nas instituições de nível superior no Brasil, são considerados como entrada para os modelos, tanto dados de candidatos concluintes do ensino médio, quanto dados de candidatos que já finalizaram essa etapa da vida acadêmica em anos anteriores ao exame para o qual estão prestando.

Assim como observado pelos demais autores, é esperado para esse estudo uma alta correlação dos resultados com os fatores socioeconômicos e educacionais dos candidatos, entretanto, para além de uma pré definição dos parâmetros relevantes, apenas questões relacionadas ao local de realização da prova e o ano de conclusão do ensino médio foram previamente excluídas, de forma a evitar o enviesamento na geração de percepções.

As próximas seções apresentam a aplicação das técnicas no conjunto de dados avaliados nesse estudo, bem como os resultados obtidos e observações coletadas.

3. METODOLOGIA

De acordo com INEP (2023), o ENEM é uma avaliação que abrange todo o território nacional e com números expressivos de participantes, na edição de 2022 foram mais de 2,3 milhões de participantes. Estudar o contexto do exame significa perpassar questões econômicas, geográficas e sociais complexas de serem resumidas e analisadas, portanto, o emprego de mineração de dados aplicado ao estudo dos resultados do exame é um tema relevante e que pode ser tanto extenso quanto complexo.

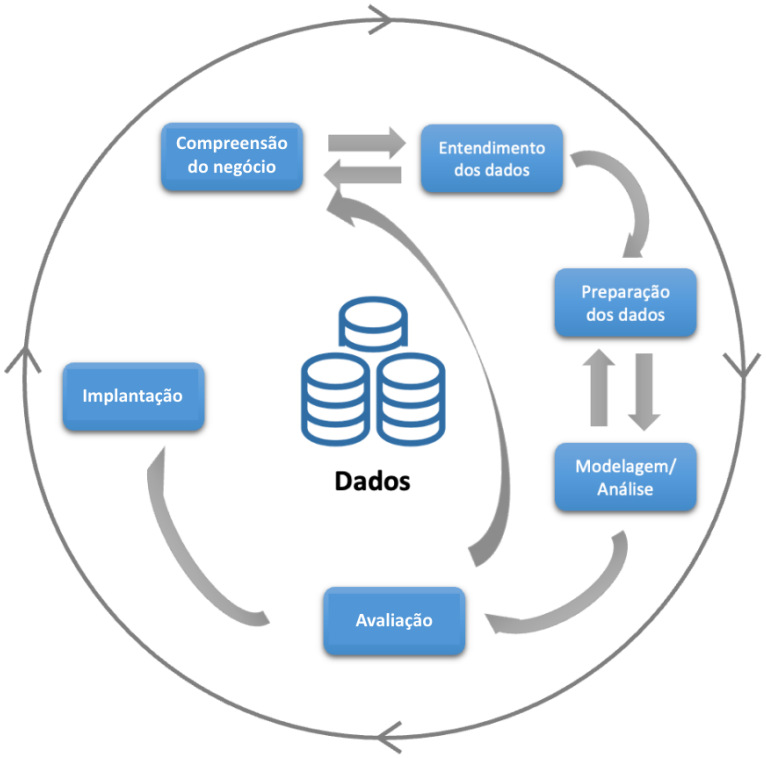
A seguir é apresentada a metodologia aplicada na exploração dos dados em busca de responder ao questionamento central deste estudo: especificar quais

aspectos se relacionam à performance dos estudantes no ENEM visando o acesso ao ensino superior.

3.1. Metodologia CRISP-DM

Diversas ferramentas podem ser utilizadas no processo de mineração de dados. Pádua e Sousa (2018) destacam o uso da Cross-Industry Standard Process of Data Mining (CRISP-DM), criada no século XX com foco na qualidade a partir da padronização de conceitos e técnicas para busca de informações e tomada de decisões. Os autores ressaltam ainda que a estrutura da CRISP-DM auxilia os pesquisadores desde o planejamento até a execução da MD. Para este trabalho, essa metodologia será utilizada seguindo os conceitos apresentados por Chapman (2000), apresentado na Figura 4.

Figura 4 - Etapas do CRISP-DM



Fonte: adaptado de Chapman (2000, p. 13).

Chapman (2000) reforça que a sequência proposta para as fases não é fixa, de modo que grande parte dos projetos avança e retorna para outras fases quando necessário. As seis fases apresentadas pelo autor para a CRISP-DM atuam de maneira cíclica não unidirecional:

a) **Compreensão do negócio**: esta fase é responsável por definir o objetivo do projeto, os stakeholders envolvidos e os requisitos do sistema;
b) **Entendimento dos dados**: esta fase é responsável por explorar os dados e entender suas características;
c) **Preparação dos dados**: esta fase é responsável por preparar os dados para a análise, o que pode incluir etapas como limpeza, transformação e redução de dimensionalidade;

d) **Modelagem**: esta fase é responsável por construir modelos de dados que possam ser usados para gerar conhecimento;
e) **Avaliação**: esta fase é responsável por avaliar os modelos de dados e identificar os melhores modelos para o objetivo do projeto;
f) **Implementação**: esta fase é responsável por implementar os modelos de dados em um sistema produtivo.

3.2. Aplicando a metodologia ao estudo

Na seção de introdução, onde é discutido o cenário heterogêneo dos egressos do Ensino Médio, obtém-se a compreensão do escopo do estudo (fase 1) e a justificativa para sua realização.

As fases de entendimento e preparação de dados (fase 2 e 3) foram conduzidas de maneira integrada e são coletivamente referidas como *tratamento de dados*. Para o desenvolvimento optou-se pelo uso da linguagem de programação Python, juntamente com as bibliotecas pandas e numpy, devido à flexibilidade e facilidade de aplicação em conjuntos de dados extensos.

A modelagem (fase 4) é a fase com mais ajustes de percurso, muito por conta da natureza do estudo, que exige testes dos modelos e avaliação (fase 5) dos resultados. O cerne da modelagem concentra-se no treinamento de modelos de classificação capazes de alcançar métricas de validação substanciais, identificando de maneira eficaz se um determinado perfil de aluno terá um desempenho satisfatório ou abaixo da média.

A variável alvo neste estudo é a classificação da nota média nas quatro provas objetivas do ENEM, que abrangem Ciências Humanas e suas Tecnologias, Ciências da Natureza e suas Tecnologias, Linguagens, Códigos e suas Tecnologias e Matemática e suas Tecnologias. Inicialmente, as notas médias foram agrupadas em quatro categorias (nota baixa, nota média-baixa, nota média-alta, nota alta), mas a avaliação do modelo não atingiu uma acurácia satisfatória. No melhor resultado, a acurácia foi testada em 52%. Buscando aprimorar a acurácia, optou-se por adotar apenas dois agrupamentos (abaixo da nota de corte, a partir da nota de corte) tomando como referência o valor da nota média, dada pela média aritmética das notas de cada prova objetiva, que varia entre 0 e 1.000 pontos. A nota da redação, por sua vez, é composta pela nota de cinco competências, sendo cada competência pontuada de 0 a 200. Os detalhes desse agrupamento serão discutidos na próxima sessão.

3.3. Definindo uma nota de corte

De acordo com o Ministério da Educação (2023), a edição do Sistema de Seleção Unificada (SISU) do 2º Semestre de 2023, cujas inscrições encerraram-se em junho de 2023, teve um total de 51.227 vagas disponibilizadas e distribuídas em 1.666 cursos de graduação em universidades públicas.

Com base nessas informações, adotou-se a nota média da 51.227ª posição, considerando uma ordenação descendente (maior para menor), como a nota de corte; o valor identificado foi 684 pontos, desconsiderando a casa decimal. Assim, define-se a categoria ‘Classificação da Nota Média’ (variável categórica 'NOTA_CLASSIF'), composta por notas abaixo e a partir da nota de corte.

É importante salientar que, no contexto deste estudo, a nota média não incorpora a pontuação da redação. Esta última é empregada como uma variável de treino com o propósito de investigar possíveis correlações entre o desempenho na redação e o desempenho nas provas objetivas. Na seção subsequente, serão discutidos os resultados obtidos por meio da análise e dos métodos empregados na manipulação dos dados.

4. DESENVOLVIMENTO E ANÁLISE DOS DADOS

Nesta seção, adotaremos a metodologia CRISP-DM como guia para explorar os microdados do ENEM. A metodologia oferece uma abordagem sistemática, desde a contextualização até a aplicação dos resultados.

A linguagem utilizada para preparação e modelagem dos dados é o Python e a sua escolha se baseou na extensa documentação disponível e bibliotecas públicas para manipulação, treinamento e avaliação dos dados.

Nas subseções a seguir, é detalhada a exploração destes dados dentre as variáveis disponíveis, dados dos participantes, da escola de origem, das provas objetivas, das redações e do questionário socioeconômico.

4.1. Entendimento e preparação dos dados

Para realizar a análise dos dados do ENEM, foi necessário selecionar as edições que seriam utilizadas. Foram selecionadas as últimas cinco edições, de 2018 a 2022. A escolha de se limitar a essas edições foi feita por dois fatores:

1. O custo computacional e a riqueza dos dados. A análise dos dados de todas as edições do ENEM seria muito custosa computacionalmente, pois envolveria o processamento de um grande volume de informações;
2. Os dados das últimas cinco edições são considerados ricos porque refletem a situação educacional do Brasil nos últimos anos.

A fim de ter a correta interpretação dos dados disponíveis, faz-se necessário a leitura dos dicionários de variáveis. Nas seções a seguir são descritos os tratamentos aplicados para se obter resultados comparáveis entre os Microdados do ENEM analisados neste estudo.

4.1.1. Padronização

Para garantir a confiabilidade e a comparabilidade dos resultados deste estudo, foi necessário padronizar os glossários das edições do ENEM. A padronização dos dicionários foi realizada equiparando as opções de resposta às perguntas que foram alteradas ao longo dos anos.

A padronização privilegiou o formato do ENEM 2022, mais recente, buscando trazer as bases de dados anteriores para o novo formato, ora suprimindo informações, como no caso da variável *tipo de escola do Ensino Médio* que nos anos anteriores dispunha de uma categoria específica para *Educação de Jovens e Adultos*, mas a partir de 2021 passou a dispor apenas das categorias *Ensino Regular* e *Educação Especial - Modalidade Substitutiva*, ora complementando as

opções como no caso da variável *Estado Civil* que não continha a categoria *Não informado* na edição de 2018, como pode ser observado no Quadro 1. Neste caso, os valores nulos na edição de 2018 foram inferidos e transformados na categoria *Não informado* presente nas edições seguintes, conforme mostrado no Quadro 2.

Quadro 1 - Edição de 2018 - Estado Civil

Ano	Nome da variável	Descrição	Categoria	Descrição da categoria
2018	TP_ESTADO_CIVIL	Estado Civil	0	Solteiro(a)
2018	TP_ESTADO_CIVIL	Estado Civil	1	Casado(a)/Mora com companheiro(a)
2018	TP_ESTADO_CIVIL	Estado Civil	2	Divorciado(a)/Desquitado(a)/Separado(a)
2018	TP_ESTADO_CIVIL	Estado Civil	3	Viúvo(a)

Fonte: autoria própria (2023).

Quadro 2 - Edição de 2019 a 2022 - Estado Civil

Ano	Nome da variável	Descrição	Categoria	Descrição da categoria
2019 - 2022	TP_ESTADO_CIVIL	Estado Civil	0	Não informado
2019 - 2022	TP_ESTADO_CIVIL	Estado Civil	1	Solteiro(a)
2019 - 2022	TP_ESTADO_CIVIL	Estado Civil	2	Casado(a)/Mora com companheiro(a)
2019 - 2022	TP_ESTADO_CIVIL	Estado Civil	3	Divorciado(a)/Desquitado(a)/Separado(a)
2019 - 2022	TP_ESTADO_CIVIL	Estado Civil	4	Viúvo(a)

Fonte: autoria própria (2023).

Na padronização, ocorreram alterações nas categorias de quatro variáveis: Estado Civil, Cor/raça, Tipo de escola do Ensino Médio e Tipo de instituição que concluiu ou concluirá o Ensino Médio.

4.1.2. Seleção de dados

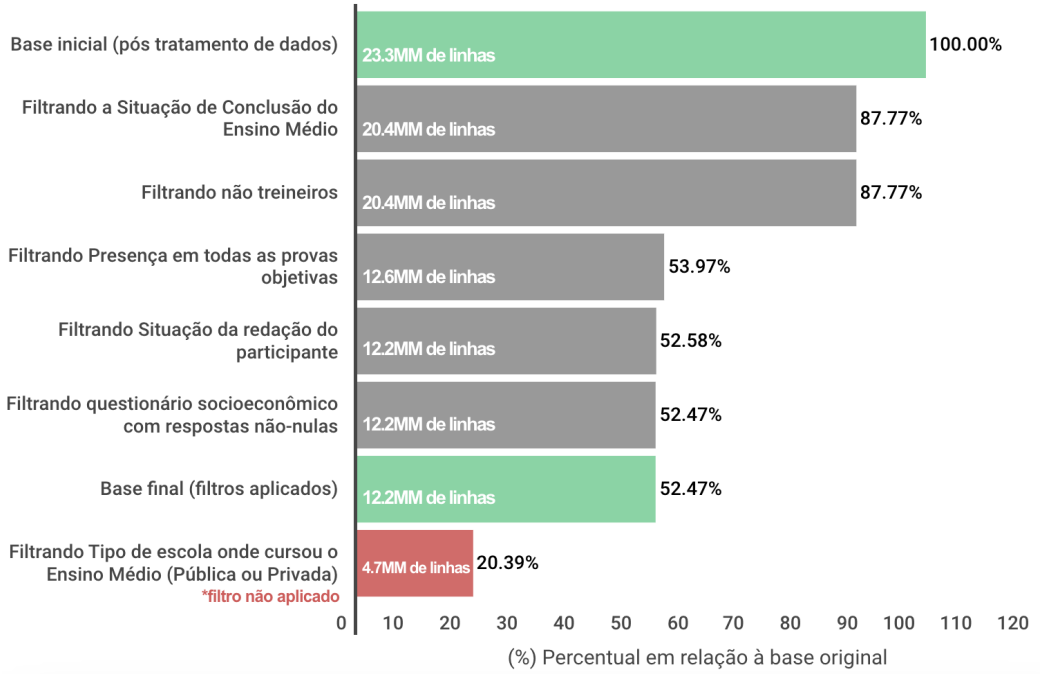
Para realizar esse estudo foram definidos alguns critérios para seleção e limpeza dos dados coletados nas bases disponíveis - partindo da base original com total de 23.257.950 de linhas. Os filtros aplicados foram:

1. Situação de Conclusão do Ensino Médio: foram considerados dados de estudantes apenas com as seguintes opções: *Já concluí o Ensino Médio* e *estou cursando e concluirei o Ensino Médio no ano do exame*;
2. Treineiros: foram desconsiderados estudantes que fizeram a prova com intuito de apenas treinar seus conhecimentos;

3. Realização das diferentes provas por área do conhecimento: foram considerados os estudantes que estiveram presentes em todas as provas do exame;
4. Redação: só foram considerados os estudantes que tiveram a redação classificada como sem problemas, expurgando da base alunos com redação anulada, texto em branco, cópia do texto motivador, fuga ao tema, não atendimento ao tipo textual, texto insuficiente ou com parte desconectada;
5. Questionário socioeconômico: foram considerados apenas os dados dos estudantes que responderam ao questionário socioeconômico.

Ao aplicar essas restrições obtém-se uma base com 52,47% dos dados, conforme mostrado na Figura 5.

Figura 5 - Quantidade de linhas após filtragens



Fonte: autoria própria (2023).

Em contraste com abordagens semelhantes, como as de Silva (2021) e Adeodato e Filho (2020), que aplicaram um filtro à variável tipo de escola onde cursou o Ensino Médio, considerando apenas as categorias (2) - Pública e (3) - Privada, excluindo os estudantes da categoria (4) - Exterior ou (1) - Não Respondeu, esse estudo opta por manter todas as categorias dessa variável. Este filtro é comumente empregado pelos autores ao se concentrarem na comparação entre alunos de escola pública e privada.

No entanto, é importante notar que a pergunta sobre o tipo de escola frequentada só é respondida pelos estudantes que concluíram o ensino médio no ano do exame. Assim, a aplicação desse filtro resultaria em uma redução significativa da base de dados, diminuindo de 52,5% da base original (em relação aos filtros já aplicados anteriormente) para apenas 20,4%. Dado que a amplitude na

análise de dados é um dos principais diferenciais deste estudo, optou-se por não aplicar esse filtro, visando manter a mais ampla base possível para a análise central, que busca entender as variáveis mais fortemente relacionadas ao bom desempenho no exame.

O código de tratamento de dados, que engloba operações como junção, padronização e filtragem dos microdados do ENEM, está publicamente disponível para consulta, podendo ser acessado por meio do seguinte link: https://www.kaggle.com/code/gabrielamorim777/tratamento-de-dados/.

4.1.3. Normalização e codificação ordinal

As variáveis numéricas foram normalizadas enquanto as variáveis categóricas foram transformadas em numéricas através da utilização do *Ordinal Encoder*. Normalizar as variáveis numéricas e aplicar a codificação ordinal nas variáveis categóricas são etapas importantes na preparação dos dados para a modelagem.

4.1.4. Classificação: abaixo da nota de corte e a partir da nota de corte

A aplicação dos algoritmos de classificação visa determinar se um aluno específico, levando em consideração suas características disponíveis, tais como dados pessoais, informações sobre a escola de origem, desempenho nas provas objetivas, redação e respostas ao questionário socioeconômico, obterá uma nota abaixo ou a partir da nota de corte (definida na seção 3.3 como 684 pontos). A Tabela 1 apresenta as classes, quantidade de ocorrência, média, desvio-padrão, pontuação mínima, pontuação máxima e os quartis (25%, 50% e 75%).

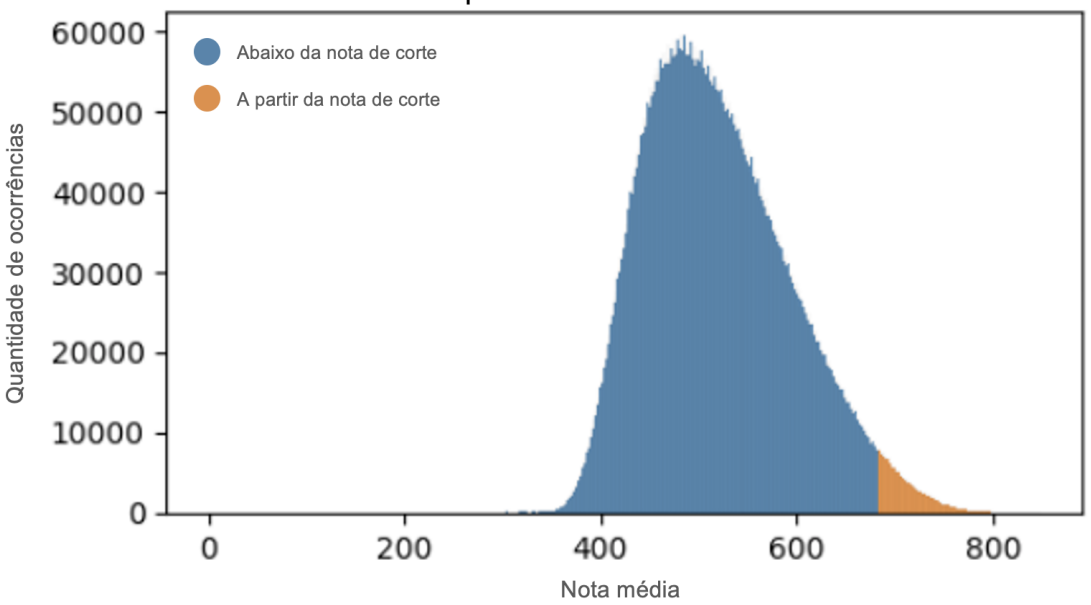
Tabela 1 - Distribuição das notas médias nas classes

Classes	Quantidade	Média	Desvio-padrão	Nota Mín	25%	50%	75%	Nota Máx
A partir da nota de corte	291.951	711	22	684	693	705	724	848
Abaixo da nota de corte	11.911.222	515	68	0	462	508	563	683

Fonte: autoria própria (2023).

Avaliando o desvio-padrão, é perceptível que na classe ‘A partir da nota de corte’ a distribuição das notas é mais homogênea, com menos variação em torno da média, enquanto na classe ‘Abaixo da nota de corte’ a distribuição das notas é mais heterogênea, com maior variação em relação à média. Esse comportamento era esperado dado que a classe ‘a partir da nota de corte’ é a classe minoritária e representa apenas 2,39% do total de candidatos, conforme apresentado na Figura 6.

Figura 6 - Distribuição das notas das classes ‘Abaixo da nota de corte’ e ‘A partir da nota de corte’ pelo número de ocorrência



Fonte: autoria própria (2023).

4.2. Modelagem e avaliação

A fase de modelagem apresenta as técnicas empregadas na base de dados tratada na seção anterior. A junção com a fase de avaliação ocorre porque, neste estudo, a interpretação dos resultados conduz, em geral, à necessidade de retomar a fase de modelagem a fim de buscar resultados melhores. Desse modo, serão apresentadas as duas etapas principais da modelagem:

1. **Aplicação das técnicas de seleção de características**: utilizou-se os métodos ANOVA e Qui-quadrado a fim de obter as variáveis que melhor representassem o modelo;
2. **Algoritmos de Classificação**: executou-se o treinamento de 4 diferentes algoritmos de Classificação: Modelo SGD (Gradiente Descendente Estocástico), Modelo de regressão logística, Árvore de decisão e Classificador SVM (máquina de vetor de suporte).

No que diz respeito aos dados empregados no treinamento e na validação do modelo de classificação, 80% foram destinados ao treinamento, enquanto os 20% restantes foram reservados para a validação.

Já para a fase de avaliação, são consideradas a acurácia, revocação, precisão e *f1-score* obtidos na análise de performance dos modelos construídos na modelagem. Os resultados são detalhados nas subseções a seguir.

4.2.1. Seleção de características usando ANOVA e Qui-2 e Análise de importância das características

Para a seleção de características foi utilizado o *SelectKBest* com os métodos do ANOVA e Qui-2. O *SelectKBest*, pertencente à biblioteca *scikit-learn,* destaca-se por ser útil em uma análise inicial e na redução de dimensionalidade com

base em estatísticas univariadas. Por outro lado, a análise da importância das características, realizada após o treinamento de um modelo, proporciona uma visão holística e iterativa das características, considerando o contexto do modelo escolhido. Ambos os métodos podem ser utilizados de forma complementar, possibilitando uma compreensão abrangente das características do conjunto de dados.

Ao analisar a seleção de características na Tabela 2, observa-se as dez variáveis com as maiores pontuações, obtidas por meio das técnicas ANOVA e Qui-2. Essa abordagem de seleção de características utiliza estatísticas univariadas para identificar as características mais relevantes.

Para avaliar a importância das características, foi utilizado um classificador de árvore de decisão. O treinamento desse modelo permitiu a análise da importância relativa de cada característica para o desempenho geral. O atributo *feature_importances_* da árvore de decisão proporciona uma medida da relevância de cada característica no contexto do modelo treinado.

Tabela 2 - Lista das dez variáveis mais relevantes usando ANOVA e Qui-2

Feature	**ANOVA Rank**	**Qui-2 Rank**	**Average Rank**	**Descrição**
Q006	4	1	**2,5**	Qual é a renda mensal de sua família?
NU_NOTA_COMP5	5	4	**4,5**	Nota da competência 5 da Redação Elaborar proposta de intervenção para o problema abordado, respeitando os direitos humanos.
Q010	7	3	**5**	Na sua residência tem carro?
NU_NOTA_COMP4	1	10	**5,5**	Nota da competência 4 da Redação Demonstrar conhecimento dos mecanismos linguísticos necessários para a construção da argumentação.
NU_NOTA_COMP2	6	5	**5,5**	Nota da competência 2 da Redação Compreender a proposta de redação e aplicar conceitos das várias áreas de conhecimento para desenvolver o tema, dentro dos limites estruturais do texto dissertativo-argumentativo em prosa.
NU_NOTA_COMP3	3	9	**6**	Nota da competência 3 da Redação Selecionar, relacionar, organizar e interpretar informações, fatos, opiniões e argumentos em defesa de um ponto de vista.
Q007	8	6	**7**	Em sua residência trabalha empregado(a) doméstico(a)?
NU_NOTA_COMPI	2	18	**10**	Nota da competência 1 da Redação Demonstrar domínio da modalidade escrita formal da Língua Portuguesa.
Q002	19	2	**10,5**	Até que série sua mãe, ou a mulher responsável por você, estudou?
Q008	14	8	**11**	Na sua residência tem banheiro?

Fonte: autoria própria (2023).

Os resultados apresentados na Tabela 3 foram obtidos utilizando o método de importância de características. O valor de importância representa a pontuação individual atribuída a cada característica, refletindo suas contribuições relativas para as decisões tomadas pelo modelo. Esse enfoque permitiu uma avaliação detalhada da influência de cada característica no desempenho global do modelo, evidenciando sua relevância no contexto da análise apresentada.

Tabela 3 - Lista das dez variáveis mais importantes a partir do método da Importância aplicado a uma árvore de decisão

Feature	Importância	Descrição
NU_NOTA_COMP3	**10,16%**	Nota da competência 3 da Redação Selecionar, relacionar, organizar e interpretar informações, fatos, opiniões e argumentos em defesa de um ponto de vista.
Q006	**9,73%**	Qual é a renda mensal de sua família?
TP_FAIXA_ETARIA	**4,85%**	Faixa etária
Q005	**4,21%**	Incluindo você, quantas pessoas moram atualmente em sua residência?
Q001	**3,89%**	Até que série seu pai, ou o homem responsável por você, estudou?
Q002	**3,69%**	Até que série sua mãe, ou a mulher responsável por você, estudou?
NU_NOTA_COMP5	**3,62%**	Nota da competência 5 da Redação Elaborar proposta de intervenção para o problema abordado, respeitando os direitos humanos.
Q019	**3,54%**	Na sua residência tem máquina de secar roupa?
NU_NOTA_COMP4	**3,20%**	Nota da competência 4 da Redação Demonstrar conhecimento dos mecanismos linguísticos necessários para a construção da argumentação.
Q004	**3,13%**	A partir da apresentação de algumas ocupações divididas em grupos ordenados, indique o grupo que contempla a ocupação mais próxima da ocupação da sua mãe ou da mulher responsável por você.

Fonte: autoria própria (2023).

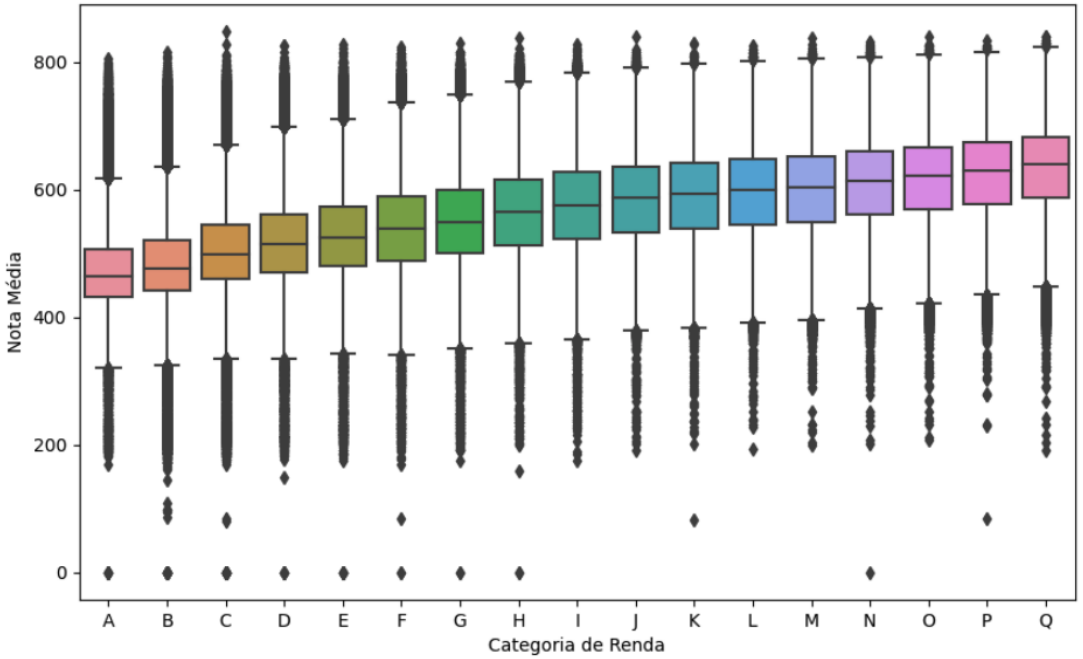
4.2.2. Avaliação dos resultados: seleção de características e importância das características

A variável que se destacou como fundamental nas previsões do modelo foi a renda mensal da família (Q006), a qual desempenhou o papel de segunda variável mais importante na implementação da árvore de decisão, apresentando uma relevância de 9,73%. A Figura 7 ilustra a distribuição das notas em relação às classes de renda familiar, ordenadas de forma crescente. É notável a tendência de aumento na nota média à medida que a renda familiar aumenta.

No Quadro 3, são apresentados os valores associados a cada classe de renda conforme declarado pelos candidatos no ENEM nas edições de 2018 a 2022.

Importante ressaltar que esses valores são baseados no salário mínimo brasileiro, garantindo a comparabilidade dos resultados entre os diferentes anos do exame analisados.

Figura 7 - Distribuição das Notas Médias por Renda Mensal Familiar



Fonte: autoria própria (2023).

Quadro 3 - Categorias da Renda Mensal Familiar

Q006 - Qual é a renda mensal de sua família? (Some a sua renda com a dos seus familiares.)	
A	Nenhuma Renda.
B	Até 1 salário mínimo.
C	Acima de 1 até 1,5 salários mínimos.
D	Acima de 1,5 até 2 salários mínimos.
E	Acima de 2 até 2,5 salários mínimos.
F	Acima de 2,5 até 3 salários mínimos.
G	Acima de 3 até 4 salários mínimos.
H	Acima de 4 até 5 salários mínimos.
I	Acima de 5 até 6 salários mínimos.
J	Acima de 6 até 7 salários mínimos.
K	Acima de 7 até 8 salários mínimos.
L	Acima de 8 até 9 salários mínimos.
M	Acima de 9 até 10 salários mínimos.
N	Acima de 10 até 12 salários mínimos.
O	Acima de 12 até 15 salários mínimos.

P	Acima de 15 até 20 salários mínimos.
Q	Acima de 20 salários mínimos.

Fonte: autoria própria (2023).

A Tabela 4 apresenta as ocorrências de valores atípicos na distribuição das notas médias por renda mensal familiar. Esses valores são visualizados na Figura 7 como pontos fora dos quartis. A Tabela 4 oferece percepções sobre a frequência dessas ocorrências, bem como sua representatividade em relação ao total de itens de cada classe. Destaca-se que o maior valor percentual de ocorrência de valores atípicos superiores equivale a 2,10% do total de itens e ocorre na classe B. De maneira similar, com relação à representatividade de valores atípicos inferiores, o percentual mais elevado é de 1,34% e ocorre na classe Q. Tais observações corroboram com os dados apresentados na Figura 7, onde se observa que os valores atípicos destacados são pouco representativos em comparação com as ocorrências situadas dentro dos quartis.

Tabela 4 - Ocorrência de valores atípicos na Distribuição das Notas Médias por Renda Mensal Familiar

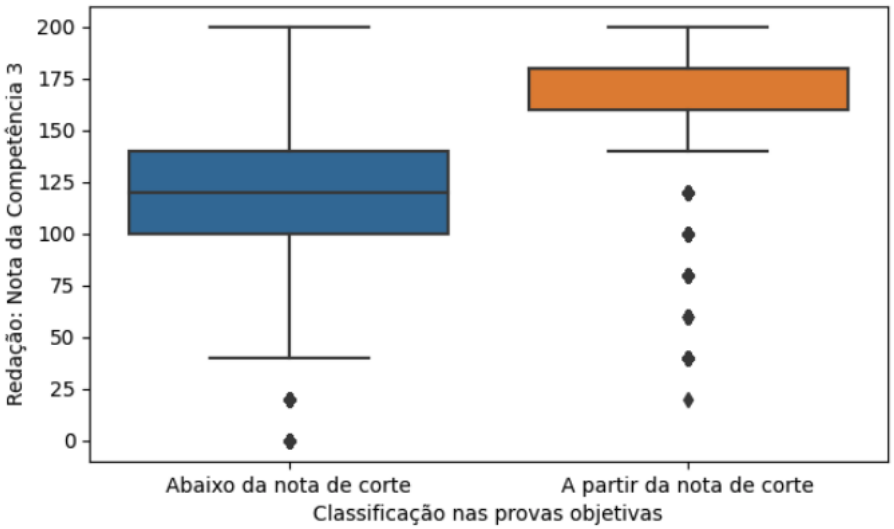
Classe	Limite Inferior	Limite Superior	Quantidade de valores atípicos inferiores	Quantidade de valores atípicos superiores	Quantidade de itens totais da classe	(%) valores atípicos inferiores	(%) valores atípicos superiores
A	320	618	408	12.565	597.982	0,07%	**2,10%**
B	326	637	1.992	42.691	3.175.554	0,06%	1,34%
C	335	670	1.448	22.183	2.649.433	0,05%	0,84%
D	334	699	588	8.102	1.356.264	0,04%	0,60%
E	342	711	480	4.885	1.023.859	0,05%	0,48%
F	341	737	255	1.721	662.583	0,04%	0,26%
G	351	748	286	1.441	768.200	0,04%	0,19%
H	360	770	212	464	528.841	0,04%	0,09%
I	366	784	152	148	347.440	0,04%	0,04%
J	379	792	157	50	206.287	0,08%	0,02%
K	383	798	147	21	147.365	0,10%	0,01%
L	391	803	172	22	125.158	0,14%	0,02%
M	396	806	303	23	133.958	0,23%	0,02%
N	415	807	592	19	133.285	0,44%	0,01%
O	423	812	723	20	116.939	0,62%	0,02%
P	435	817	1.012	8	104.506	0,97%	0,01%
Q	447	824	1.678	12	125.519	**1,34%**	0,01%

Fonte: autoria própria (2023).

Outro aspecto foi a avaliação da nota da competência 3 da redação, que aborda a capacidade de selecionar, relacionar, organizar e interpretar informações, fatos, opiniões e argumentos em defesa de um ponto de vista. Esta variável foi classificada como a mais relevante, com uma importância de 10,16% na aplicação da árvore de classificação. Além disso, vale destacar que ela ocupou a 6ª posição na seleção de características, evidenciando sua influência nas previsões do modelo.

Na Figura 8 pode-se observar os *whiskers*, as linhas que se estendem a partir da caixa até os valores extremos. Nota-se que a faixa de notas, excluindo os *outliers*, para a categoria 'A partir da nota de corte' inicia em torno de 140 pontos, enquanto a caixa da categoria 'Abaixo da nota de corte' finaliza próxima dessa pontuação, evidenciando uma concentração de notas superiores a 140 pontos na competência 3 da redação na categoria 'A partir da nota de corte', em comparação com a categoria 'Abaixo da nota de corte', que apresenta concentração de notas abaixo desse limiar.

Figura 8 - Distribuição das Notas da redação - Competência 3

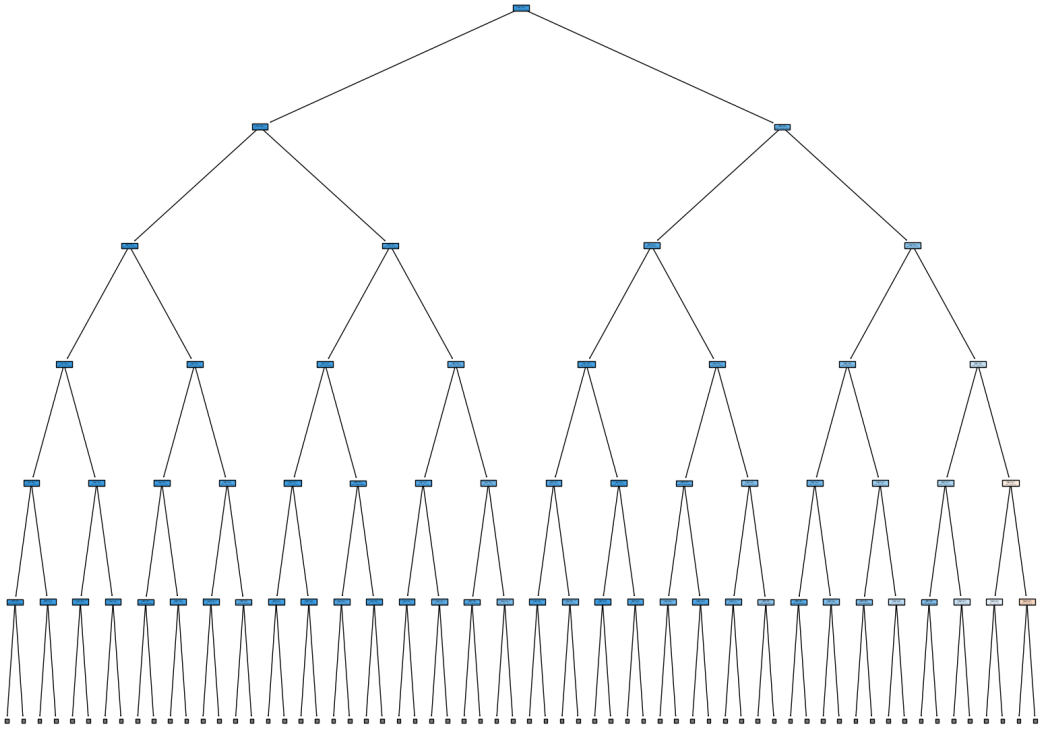


Fonte: autoria própria (2023).

4.2.3. Avaliação dos resultados: Árvore de decisão

A utilização de árvores de decisão na análise de modelos classificatórios oferece uma visão sobre como as decisões são tomadas pelo algoritmo, contudo, é crucial reconhecer que a interpretação dessas árvores pode se tornar desafiadora em cenários de elevada dimensionalidade e complexidade do modelo. Em muitos casos, a profundidade da árvore necessária para representar adequadamente as decisões e diferenciar as classes pode ser extensa, dificultando a compreensão do processo de classificação. É nesse contexto que a Figura 9 é apresentada, ao exemplificar a complexidade alcançada pela árvore de decisão do modelo utilizado neste trabalho com a profundidade 5. Notadamente, a diferenciação entre as duas categorias de desempenho dos estudantes apenas começa a ser percebida a partir da 4ª profundidade (nó de cor alaranjada no canto inferior direito da Figura 9) evidenciando a necessidade de considerar cuidadosamente a interpretação do modelo e sua capacidade de discriminação.

Figura 9 - Árvore de decisão

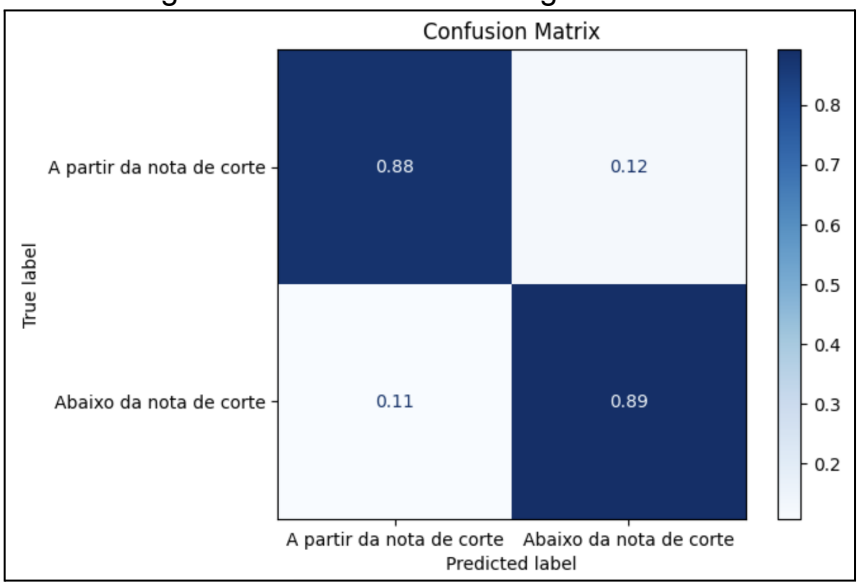


Fonte: autoria própria (2023).

4.2.4. Aplicação do algoritmo: SGD (Stochastic Gradient Descent)

O modelo de classificação SGD demonstrou um bom desempenho geral, com uma acurácia de 89,18%, indicando uma taxa elevada de previsões corretas. O modelo errou 11% das previsões para a classificação *abaixo da nota de corte* e 12% das previsões para a classificação *a partir da nota de corte.* A Figura 10 apresenta a matriz de confusão para o algoritmo em questão.

Figura 10 - Resultados do Algoritmo SGD



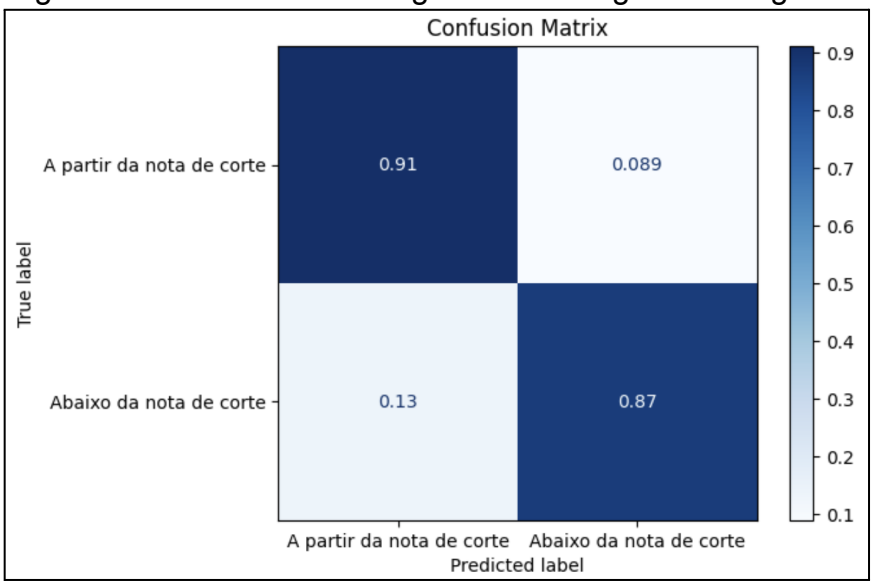
Fonte: autoria própria (2023).

O F1-score, uma métrica que harmoniza precisão e revocação, alcançou 92,57%, refletindo um equilíbrio eficaz entre a capacidade de identificar positivos e evitar falsos positivos. A precisão atingiu 97,68%, destacando a habilidade do modelo em minimizar falsos positivos, enquanto a revocação de 89,18% evidencia sua capacidade de capturar a maioria das instâncias positivas.

4.2.5. Aplicação do algoritmo: Regressão logística

O desempenho utilizando o algoritmo de regressão logística apresentou uma acurácia de 86,72%. O modelo errou 13% das previsões para a classificação *abaixo da nota de corte* e apenas 8,9% das previsões para a classificação *a partir da nota de corte.* O F1-Score foi de 91,09% e a revocação igual a 86,72%. A precisão do modelo alcançou 97,7%, refletindo a capacidade de minimizar falsos positivos. A Figura 11 apresenta a matriz de confusão para o algoritmo em questão.

Figura 11 - Resultados do Algoritmo de Regressão Logística

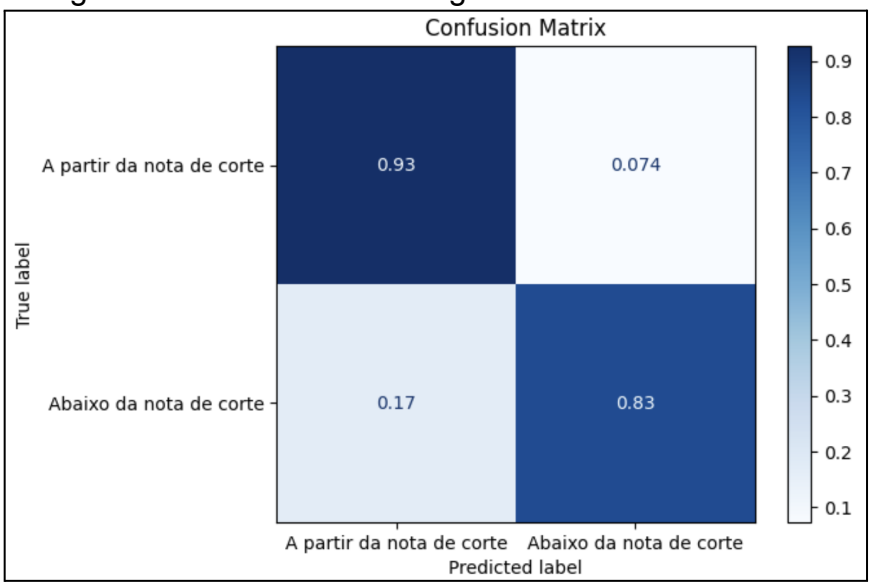


Fonte: autoria própria (2023).

4.2.6. Aplicação do algoritmo: Árvore de decisão

O desempenho utilizando o algoritmo de árvore de decisão com profundidade máxima igual a 7 apresentou uma acurácia de 83,28%. O modelo errou 17% das previsões para a classificação *abaixo da nota de corte* e apenas 7,4% das previsões para a classificação *a partir da nota de corte.* O F1-Score foi de 88,99% e a revocação igual a 83,28%. A precisão do modelo alcançou 97,68%. A Figura 12 apresenta a matriz de confusão para o algoritmo em questão.

Figura 12 - Resultados do Algoritmo Árvore de Decisão

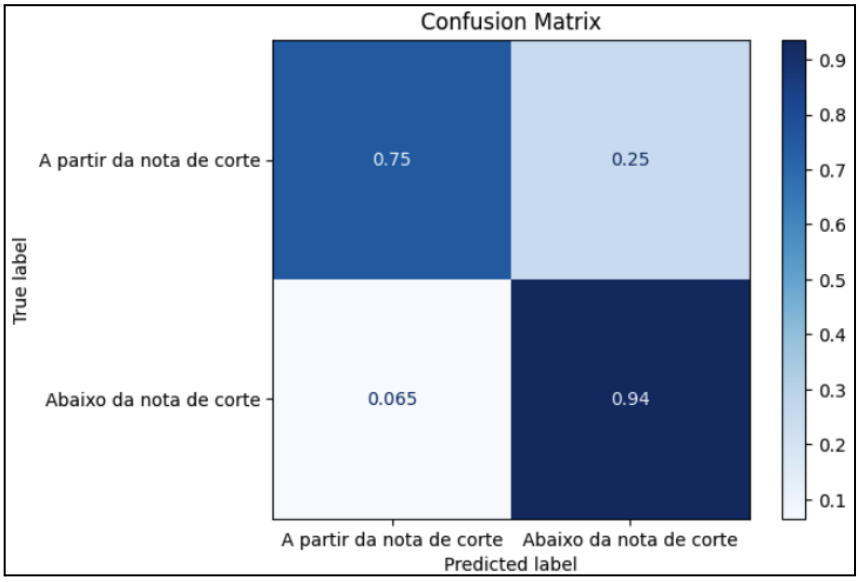


Fonte: autoria própria (2023).

4.2.7. Aplicação do algoritmo: Classificador SVM

O desempenho utilizando o algoritmo do classificador SVM apresentou uma acurácia de 93,08%. Entretanto, o modelo errou 25% das previsões para a classificação *a partir da nota de corte* e apenas 6,5% das previsões para a classificação *abaixo da nota de corte.* O F1-Score foi de 94,86% e a revocação igual a 93,08%. A precisão do modelo alcançou 97,51%. A Figura 13 apresenta a matriz de confusão para o método em questão.

Figura 13 - Resultados do Algoritmo Classificador SVM



Fonte: autoria própria (2023).

4.3. Desafio do processamento massivo de dados

A realização deste estudo envolveu a manipulação de extensas bases de dados, apresentando desafios no tratamento, processamento e treinamento dos modelos. Um desses desafios iniciais foi o consumo de memória de processamento. Por esse motivo, optou-se pela plataforma Kaggle, que possibilita a execução de código em Python e oferece um ambiente virtual com 30GB de memória RAM gratuitamente, no entanto, em determinados cenários, o custo computacional tornou-se mais elevado, levando à adoção de outra ferramenta: o Colab, em sua versão PRO. O Colab, ou Google Colaboratory, é uma plataforma baseada em nuvem que permite a execução de notebooks Jupyter, facilitando o desenvolvimento e a colaboração em projetos de ciência de dados e aprendizado de máquina. Em sua versão paga, o Colab oferece a função *High-RAM*, proporcionando uma memória RAM de 51GB.

Além dos pontos abordados, cabe ainda destacar que o tempo de processamento representou outro desafio significativo. Algumas execuções, como por exemplo o processamento do algoritmo Classificador SVM, demandaram de 6 a 7 horas para serem processadas.

Ao abordar os desafios relacionados ao tratamento de dados, destaca-se o papel fundamental desempenhado pela biblioteca Pandas. Essa biblioteca, especializada em Python, oferece ferramentas para manipulação eficiente de conjuntos de dados, simplificando a tarefa de tratamento de dados.

5. CONSIDERAÇÃO FINAL

Os resultados da análise, com a renda familiar desempenhando um papel significativo, reforçam a importância contínua das políticas de cotas para promover a inclusão e a equidade no acesso à educação superior. A implementação efetiva dessas políticas pode ser crucial para mitigar as disparidades socioeconômicas, proporcionando oportunidades igualitárias para estudantes de diferentes estratos sociais. O acesso equitativo não é apenas uma questão de admissão, mas também de garantir a qualidade e a preparação dos estudantes para os desafios acadêmicos e sociais. Assim, uma abordagem holística é fundamental para enfrentar as raízes da desigualdade e promover um ambiente educacional igualitário.

A elevada importância da competência 3 da redação na previsão do desempenho sugere uma consideração mais profunda das habilidades de expressão escrita no contexto educacional; pode-se explorar a possibilidade de um viés pedagógico que enfatiza a capacidade de selecionar, relacionar, organizar e interpretar informações como critério avaliativo. Essa ênfase pode refletir não apenas as habilidades comunicativas, mas também a capacidade de análise crítica e argumentação, destacando a importância do conhecimento factual e das habilidades de pensamento crítico.

O desafio do processamento massivo de dados destaca a importância de ferramentas e plataformas eficientes para lidar com grandes conjuntos de dados. A escolha do ambiente de execução, como Kaggle e Colab, influenciou diretamente a viabilidade e o custo computacional do estudo.

Como oportunidades futuras, é sugerido aprofundar a análise do desempenho dos candidatos cotistas, cruzando informações do SISU com dados

específicos por curso, tanto na ampla concorrência, quanto nas vagas destinadas às cotas, visando uma compreensão mais detalhada do impacto dessas políticas.

REFERÊNCIAS

ADEODATO, P. J. L.; SILVA FILHO, R. L. C. Where to aim? Factors that influence the performance of Brazilian secondary schools. *In:* **Proceedings** of the 13th INTERNATIONAL CONFERENCE ON EDUCATIONAL DATA MINING. Evento virtual, 2020. Disponível em: https://educationaldatamining.org/files/conferences/EDM2020/papers/paper_55.pdf Acesso em: 1 out. 2023.

BARROS, R.; SANTANA JUNIOR, O. V.; SILVA, I. R. M.; SANTOS, L. F.; CÂMARA NETO, V. R. Predição do rendimento dos alunos em lógica de programação com base no desempenho das disciplinas do primeiro período do curso de ciências e tecnologia utilizando técnicas de mineração de dados. **Brazilian Journal Of Developmen**t. Curitiba, v. 6, n. 1, p. 2523-2534, 2020. Disponível em: https://www.brazilianjournals.com/index.php/BRJD/article/view/6167/5484. Acesso em: 1 out. 2023.

BARTH, R. S. A personal vision of a good school. **Phi Delta Kappan**, n. 71, p. 512-571, 1990..

BRASIL. Ministério da Educação. **ENEM abre as portas da educação superior no Brasil**. Brasília, 2023. Disponível em: https://www.gov.br/mec/pt-br/assuntos/noticias/2023/junho/enem-abre-as-portas-da-educacao-superior-no-brasil. Acesso em: 01 out. 2023.

BRASIL. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. **Divulgado resultado do ENEM 2022**. Brasília, 2023. Disponível em: https://www.gov.br/inep/pt-br/assuntos/noticias/enem/divulgado-resultado-do-enem-2022. Acesso em: 01 out. 2023.

BRASIL. Serviços e Informações do Brasil. **Mais de 51 mil vagas serão ofertadas no SISU do 2º semestre**. Brasília, 2023. Disponível em: https://www.gov.br/pt-br/noticias/educacao-e-pesquisa/2023/02/sistema-de-selecao-unificada-2023-tem-1-073-024-inscritos-e-supera-primeira-edicao-de-2022. Acesso em: 01 out. 2023.

BRAZ, M. A.; SOARES, A. B. **Mineração de dados:** conceitos, métodos e aplicações. São Paulo: Saraiva, 2016.

BROPHY, J.; GOOD, T. Teacher behavior and student achievement in m. witrock. In: BROPHY, J.; GOOD, T. (ed.). **The third handbook of research on teaching**. Nova Iorque: McMillan, 1986.

CHAPMAN, P. et al. CRISP-DM 1.0: Step-by-step data mining guide. **SPSS Inc.**, v. 9, p. 13, 2000.

CHOLLET, F. **Deep learning with Phyton**. Nova Iorque: Manning Publications, 2017.

CHUDGAR, A.; LUSCHEI, T.; FAGILI, L. **Constructing socio-economic status measures using the trends in international mathematics and science study data**. East Lansing: Michigan State University, 2012.

COLEMAN, J. S.; CAMPBELL, E.; HOBSON, C.; MCPARTLAND, J.; MOOD, A.; WEINFELD, F. **Equality of educational opportunity study**. United States: Department of Health, Education, and Welfare, Washington, 1966.

FREIRE. P. **Pedagogia do Oprimido.** 64 ed. Rio de janeiro: paz e terra 2017

GOYAL, M.; VOHRA, R. Applications of data mining in higher education. **International Journal of Computer Applications in Engineering Education**, BahraUniversity, India, v.9, n.1, p. 113 - 120, 2012. Disponível em: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=0fc705ef34776db8f50d64b4b82bbd11ca5ebee5. Acesso em: 12 out. 2023.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA. **Censo da educação superior 2022**. Brasília, 2023. Disponível em: https://download.inep.gov.br/educacao_superior/censo_superior/documentos/2022/apresentacao_censo_da_educacao_superior_2022.pdf. Acesso em: 1 out. 2023.

PÁDUA, A. F. L. O.; SOUSA, F. A. Metodologia CRISP-DM: Potencialidades na descoberta do conhecimento em dados educacionais. *In:* XVI CONGRESSO INTERNACIONAL DE TECNOLOGIA NA EDUCAÇÃO. **Anais...** Recife, 2018. Disponível em: https://www.pe.senac.br/congresso/anais/2018/pdf/poster/METODOLOGIA%20CRISP-DM%20POTENCIALIDADES%20NA%20DESCOBERTA%20DO%20CONHECIMENTO%20EM%20DADOS%20EDUCACIONAIS.pdf.. Acesso em: 1 out. 2023.

PINTO, R. A. M.; ROCHA, A. P. **Mineração de dados:** uma abordagem prática. São Paulo: Pearson, 2017.

RISTOFF, D.; GIOLO, J. O Sinaes como sistema. **Revista Brasileira de Pós Graduação**, Brasília, v. 3, n. 6, p.193-213, 2006. Disponível em:<http://www2.capes.gov.br/rbpg/images/stories/downloads/RBPG/Vol.3_6_dez2006_/Est_Artigo2_n6.pdf>. Acesso em: 10 dez. 2023.

INSTITUTO DO SINDICATO DAS ENTIDADES MANTENEDORAS DE ESTABELECIMENTOS DE ENSINO SUPERIOR NO ESTADO DE SÃO PAULO. **Mapa do Ensino Superior no Brasil**. 11 ed. São Paulo: Convergência 2021.

SCIKIT, L. **Documentação**. Disponível em https://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics. Acesso em: 7 set. 2023.

SILVA, V. A. A. **Uso de aprendizado de máquina para identificar desigualdades sociais na base de dados do ENEM**. Juiz de Fora: Universidade Federal de Juiz de Fora, 2021. Disponível em: http://monografias.ice.ufjf.br/tcc-web/exibePdf?id=549. Acesso em: 12 out. 2023.

SOUZA, V. F.; SOUZA, M.F. Os avanços da mineração de dados educacionais: definições, processo e evolução. **Brazilian Journal of Development**, Curitiba, v.7, n.8, p. 80798-80819, 2021. Disponível em: https://ojs.brazilianjournals.com.br/ojs/index.php/BRJD/article/download/34442/pdf/88019. Acesso em: 12 out. 2023.

WATANABE, W. **Prevendo as notas do ENEM com Machine learning — Data Science**. Medium, 15 de maio de 2019. Disponível em: https://medium.com/@wesleywatanabe/data-science-machine-learning-enem-regressao-linear-5cd140459dc3. Acesso em: 29 out. 2023.