Revista Brasileira de Informática na Educação – RBIE
Brazilian Journal of Computers in Education
(ISSN online: 2317-6121; print: 1414-5685)
http://br-ie.org/pub/index.php/rbie

Submission: 30/Dec/2022; 1st round notif.: 15/Mar/2023; New version: 14/Apr/2023; 2nd round notif.: 15/May/2023
Camera ready: 26/May/2023; Edition review: 25/Jun/2023; Available online: 03/Jul/2023; Published: 03/Jul/2023

Fatores que podem interferir no desempenho de estudantes no ENEM: uma revisão sistemática da literatura

Title: Factors that may affect students performance in the ENEM Exam: a systematic review of the literature

Janderson Ferreira Dutra
Instituto Federal da Paraíba (IFPB)
ORCID: 0000-0003-4601-1461
janderson.dutra@ifpb.edu.br

João Batista Firmino Junior
Instituto Federal da Paraíba (IFPB)
ORCID: 0000-0001-8038-8516
batista.firmino@academico.ifpb.edu.br

Damires Yluska de Souza Fernandes
Instituto Federal da Paraíba (IFPB)
ORCID: 0000-0001-9305-5323
damires@ifpb.edu.br

Resumo

Avaliar o desempenho de estudantes em diversos contextos é algo muito complexo. Isso não é diferente quando se discute fatores associados ao desempenho de estudantes em provas como a do Exame Nacional do Ensino Médio (ENEM). Diversos fatores como, por exemplo, o conhecimento adquirido do estudante ao longo de sua trajetória acadêmica, assim como outros oriundos de suas experiências, situação social ou econômica podem impactar em resultados diferenciados na prova. Dados históricos do ENEM disponibilizados compõem informações diversas sobre os resultados individuais dos estudantes assim como incluem respostas a questionários formulados no momento da inscrição. Diante da dimensionalidade dos dados e da complexidade de análises que podem ser realizadas a partir desses conjuntos de dados, uma questão essencial é identificar quais fatores são realmente mais relevantes para tais análises. Técnicas de mineração de dados, a exemplo de modelos preditivos e seleção de features, têm sido usadas como meio para ajudar na obtenção das análises. Neste cenário, este trabalho apresenta uma revisão sistemática da literatura com o intuito de identificar os principais fatores que podem influenciar no desempenho dos estudantes na prova do ENEM, considerando estudos publicados nos últimos dez anos. Os resultados obtidos mostraram que os fatores mais relevantes estão relacionados às questões socioeconômicas, sendo os atributos em maior evidência os seguintes: renda familiar, idade, sexo e raça. O nível de escolaridade dos pais também ganha destaque. Atributos relacionados às notas nas provas e caracterização das escolas de origem dos estudantes relativos à estrutura física e pedagógica são igualmente destacados. O presente estudo evidencia alguns caminhos que podem ser conduzidos em pesquisas complementares.

Palavras-Chave: *ENEM; Desempenho de estudantes; Fatores; Análise de dados; Mineração de dados.*

Abstract

Evaluating student performance in diverse contexts is a very hard task. This is not different when discussing factors associated with student performance regarding tests such as the National High School Examination (ENEM). Several factors such as the students own knowledge acquired throughout their academic career, as well as other ones arising from their experiences, or even social or economic situation, may impact on diverse results in the test. ENEM historical data made available comprise diverse information about the individual results of students as well as answers obtained by means of questionnaires formulated at registration time. Due to the high dimensionality of the data and the complexity of analyzes that might be carried out from these datasets, an essential question is how to identify which factors are really more relevant for such analyses. Data mining techniques, such as predictive models and feature selection, have been used as a means to help obtaining such analyses. In this light, this work presents a systematic literature review in order to identify the main factors that may influence the performance of students in the ENEM test, considering studies published along the last ten years. The results obtained showed that the most relevant factors are related to socioeconomic issues, with the following attributes being most evident: family income,

Cite as: Dutra, J. F., Firmino Júnior, J. B. & Fernandes, D. Y. S. Fatores que podem interferir no desempenho de estudantes no ENEM: uma revisão sistemática da literatura. Revista Brasileira de Informática na Educação, 31, 323-351. DOI: 10.5753/rbie.2023.3087

age, sex and race. Parents' level of education is also highlighted. Attributes related to test scores and characterization of the students' schools of origin with respect to the physical and pedagogical structure are additionally emphasized. This study points out some paths that may be accomplished on complementary research.
Keywords: *ENEM; Student performance; Factors; Data analysis; Data mining.*

1 Introdução

Ingressar na Educação Superior é o sonho de muitos jovens brasileiros. Construir uma formação sólida em curso superior e contribuir com conhecimento são questões extremamente importantes para a evolução social (Coutinho et al., 2018). No Brasil, o Ministério da Educação (MEC) possui diversos programas e ações voltados a todos os níveis de educação. Um dos principais órgãos de controle de dados educacionais é o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), que é responsável por desenvolver e publicar informações sobre as avaliações relacionadas à Educação mais importantes do país.

Sabe-se que a educação busca promover inclusão social e ser fonte de inovações no desenvolvimento de soluções tecnológicas nos campos técnico, científico, social e cultural. Os princípios das políticas públicas de educação desenvolvidas para a sociedade têm um papel estratégico no contexto socioeconômico das suas respectivas regiões, contribuindo com o desenvolvimento de um determinado território em acordo com as políticas nacionais (Medeiros, 2021). Para avaliar aspectos associados ao ensino, pesquisa, extensão, desempenho de estudantes, entre outros relativos às Instituições de Ensino Superior (IES) foi criado o Sistema Nacional de Avaliação da Educação Superior (SINAES) - Lei n° 10.861, de 14 de abril de 2004. O SINAES possui alguns instrumentos de avaliação de cursos, entre eles, o Exame Nacional do Ensino Médio (ENEM) que avalia o conhecimento universal e habilidades adquiridas ao longo da Educação Básica. O exame também verifica o nível de interpretação sobre acontecimentos em relação à realidade brasileira e mundial. Além disso, é a principal porta de acesso de estudantes ao nível superior. Além do ENEM, outros exames são realizados pelo INEP como, por exemplo, o Exame Nacional de Desempenho dos Estudantes (ENADE), a Prova Brasil, o Exame Nacional para Certificação de Competências de Jovens e Adultos (ENCCEJA) e a Avaliação da Educação Básica (SAEB). Neste trabalho, o foco de estudo é concentrado no ENEM.

Avaliar o desempenho de estudantes em diversos contextos é algo muito complexo. Isso não é diferente quando se discute aspectos associados ao desempenho do estudante em provas do ENEM. Diversos fatores como, por exemplo, o próprio conhecimento adquirido do estudante seja por meio da sua vida acadêmica, seja por meio de experiências extras, além de fatores relacionados ao local onde vive, à situação social e econômica dele e da família podem implicar em resultados individuais da prova. Li & Patel (2021), por exemplo, investigaram se fatores relativos ao clima têm impacto significativo no desempenho de estudantes na prova do ENEM. O resultado mostrou que o efeito negativo das condições climáticas, particularmente da temperatura, não é substancial nem conclusivo para justificar o desempenho na prova. Melo et al. (2022), por outro lado, buscaram identificar variáveis com maior impacto no desempenho dos municípios no ENEM com base em técnicas de estatística espacial e na perspectiva da sociologia da educação. Os resultados apresentados sugerem que questões socioculturais e econômicas podem ser fatores importantes quanto às desigualdades de desempenho dos estudantes em avaliações.

Particularmente, quando se discute aspectos que podem interferir positiva ou negativamente no desempenho de estudantes no ENEM, diversas análises têm sido realizadas como no tocante a fatores socioeconômicos (Jaloto & Primi, 2021; Lucena & Santos, 2020). Entretanto, são vários os fatores que cercam estudantes quando se busca entender impactos e interferências de modo a se buscar mitigar essas questões. Os dados disponibilizados pelo INEP, ao longo dos anos, oferecem diversas informações e parâmetros que podem ser analisados, como, por exemplo, renda, escolaridade dos pais, tipo de escola e notas de provas. Para lidar de forma mais eficiente com a multidimensionalidade envolvida na disponibilização destas informações e sua avaliação quanto ao desempenho de estudantes no ENEM, análises baseadas em extração de conhecimentos podem ser usadas. Nesse panorama, técnicas da Mineração de Dados podem ajudar a analisar os dados e obter indicadores aptos a assistir tomadas de decisão de modo mais ágil e assertivo (Soares

et al., 2021; Ferreira et al., 2021). Particularmente, a identificação de fatores considerados mais relevantes pode apoiar profissionais da educação, como professores e gestores, em ações que visem minimizar problemas associados ao desempenho escolar dos estudantes ou mesmo nas questões pedagógicas e de infraestrutura das escolas.

Neste cenário, este trabalho apresenta uma Revisão Sistemática da Literatura (RSL) que buscou identificar o estado da arte em termos de estudos que investigaram os principais fatores que podem influenciar no desempenho de estudantes no ENEM. A presente RSL considerou o período dos últimos 10 anos para a seleção de trabalhos primários. A questão principal da RSL objetiva responder quais fatores têm sido considerados como mais relevantes pelos estudos no entendimento sobre desempenhos de estudantes no ENEM. Para isso, foram selecionados trabalhos que explorassem fatores ou atributos considerados importantes e que utilizassem técnicas de mineração de dados como apoio computacional a estas análises e extrações de conhecimento. No contexto de mineração de dados, uma importante área de pesquisa associada à determinação de fatores relevantes é a seleção de *features* (Chandrashekar & Sahin, 2014; Liu & Motoda, 2007) associada a técnicas supervisionadas ou não supervisionadas de aprendizado de máquina.

Além deste contexto introdutório, este trabalho está organizado da seguinte forma: na Seção 2 encontra-se a fundamentação teórica, onde são apresentados alguns conceitos necessários à compreensão do trabalho; a Seção 3 descreve alguns trabalhos relacionados a esta RSL; na Seção 4, é mostrada a metodologia usada para condução da presente pesquisa; na Seção 5, encontram-se a apresentação dos resultados obtidos e discussões empregadas; por fim, na Seção 6, são tecidas considerações finais e indicações de trabalhos futuros.

2 Fundamentação Teórica

Nesta seção, são introduzidos conceitos acerca dos temas deste trabalho, com destaque ao ENEM e algumas técnicas associadas à Mineração de Dados (MD) e ao Aprendizado de Máquina (AM).

2.1 ENEM

O ENEM é um exame realizado anualmente pelo MEC sob organização do INEP. Ele foi criado em 1998 com o objetivo central de avaliar o desempenho escolar dos estudantes ao término da educação básica. Porém, foi apenas em 2009 que o INEP passou a usar o exame como instrumento de avaliação para ingresso dos estudantes na educação superior (Inep, 2022).

A nota do ENEM passou então a ser usada por instituições públicas através do Sistema de Seleção Unificada (SISU), programa responsável por selecionar candidatos para ingresso em cursos superiores. Posteriormente outros programas surgiram, a exemplo do Programa Universidade para Todos (PROUNI) e do Financiamento ao Estudante do Ensino Superior (FIES), que usam a nota do ENEM, respectivamente, para oferta de bolsas de estudo e para o financiamento em instituições de ensino superior privadas (Miranda & Azevedo, 2020).

Atualmente, a prova do ENEM avalia cinco competências, o que inclui provas de redação e provas objetivas nas áreas de conhecimento de Ciências Humanas e suas Tecnologias, Ciências da Natureza e suas Tecnologias, Matemática e suas Tecnologias, e Linguagens, Códigos e suas Tecnologias. Sendo um instrumento de avaliação, o ENEM possibilita que professores e especialistas em educação acompanhem mais de perto o desempenho dos estudantes sobre pontos importantes considerados nas provas como, por exemplo, interdisciplinaridade, contextualização e resolução de problemas (Castro & Tiezzi, 2004).

Os resultados individuais e as respostas ao questionário socioeconômico pelos inscritos foram sendo disponibilizados a cada edição do ENEM no formato de microdados no Portal do INEP1. Os microdados, de acordo com esse portal, "são o menor nível de desagregação de dados recolhidos por meio do exame. Eles atendem à demanda por informações específicas ao disponibilizar as provas, os gabaritos, as informações sobre os itens, as notas e o questionário respondido pelos inscritos no Enem." São, dessa forma, disponibilizados, no formato "zip", por ano, desde 1998. Em atendimento às diretrizes da Lei Geral de Proteção de Dados Pessoais2 (LGPD), os microdados de todas as edições foram revisados e, conforme as necessidades apresentadas, alguns foram reestruturados com o objetivo de evitar a identificação de pessoas.

Tendo em vista o volume e a complexidade desse conjunto de dados educacionais, não é trivial identificar padrões nos referidos dados de modo explícito, o que torna interessante o uso de técnicas de análises mais implícitas. Para isso, técnicas pertinentes ao escopo da MD e do AM podem ser utilizadas (Baradwaj & Pal, 2012; Han et al., 2012; Fayyad et al., 1996).

2.2 Mineração de Dados

Fayyad et al. (1996) definem *Knowledge Discovery in Databases* (KDD) como o processo de descoberta de conhecimento a partir de dados, incluindo neste processo como os dados são selecionados, transformados, como algoritmos de aprendizado de máquina podem ser utilizados e executados com eficiência para identificar padrões, e como resultados podem ser interpretados. Nesta definição, os autores consideram a MD como uma única etapa deste processo. No entanto, a MD tem sido normalmente usada como sinônimo de KDD (Martinez-Plumed et al., 2019; Cao, 2017). Sendo assim, neste trabalho não há distinção entre os dois termos, ou seja, considera-se a MD como um processo iterativo e incremental de descoberta do conhecimento que inclui etapas e resultados para cada uma delas.

Trabalhos e projetos de MD têm sido realizados usando outros modelos de processo que estendem o KDD, a exemplo do *Cross Industry Standard Process for Data Mining* (CRISP-DM) (Martinez-Plumed et al., 2019). O CRISP-DM estende as etapas da proposta original de KDD em seis etapas (Chapman et al., 2000): entendimento do negócio, entendimento dos dados, preparação dos dados, modelagem do aprendizado de máquina, avaliação e implantação. Na literatura, há vários trabalhos que consideram este modelo de processo um dos mais utilizados em contextos de MD e até mesmo na área "guarda-chuva" recente denominada Ciência de Dados (Martinez-Plumed et al., 2019; Cao, 2017).

No contexto de MD, o AM fornece a base técnica para a extração de padrões que irão gerar conhecimento. Isso é realizado por meio de algoritmos que extraem padrões de comportamento a partir de dados (exemplos) sem necessidade de programação explícita (Mitchell, 1997). Além do AM, na MD, outras atividades e técnicas associadas à preparação dos dados (e.g., limpeza, normalização, imputação de valores) assim como à engenharia de atributos são importantes (Goldschmidt et al., 2015). A seleção de *features*, por sua vez, é uma tarefa essencial na eliminação de atributos irrelevantes/redundantes em um conjunto de dados (Liu & Motoda, 2007).

Quando esses dados são provenientes da área educacional, a exemplo dos dados do ENEM, autores têm considerado a subárea intitulada Mineração de Dados Educacionais (MDE) (Ramos et al., 2020; Romero et al., 2014; Baker, 2010). Para fins de generalização, o termo MD será

[1] Microdados do ENEM disponíveis em: gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem. Acesso em 20 mar. 2023.

[2] Informações complementares em: gov.br/inep/pt-br/assuntos/noticias/institucional/inep-republica-conjunto-de-microdados-do-enem-e-saeb. Acesso em 20 mar. 2023.

utilizado no decorrer do texto deste trabalho quando referente à Mineração de Dados. Algumas técnicas consideradas da MD neste trabalho são resumidamente abordadas na seção seguinte.

2.3 Algumas técnicas de MD

Predição é a arte de construir e usar modelos que fazem previsões baseadas em padrões identificados a partir de dados históricos (Kelleher et al., 2020). Neste sentido, a classificação e a regressão constituem tarefas de predição. A classificação é usada para predizer valores discretos. A regressão, por outro lado, é usada para estimar valores contínuos (numéricos), em vez de rótulos de classe (discretos) (Han et al., 2012).

Classificação e regressão são exemplos de tarefas de aprendizado supervisionado, onde, a partir de um conjunto de exemplos pré-rotulados, é possível desenvolver um modelo que permite classificar um conjunto de novas instâncias, rotulando a partir de uma dada classe ou estimando valores contínuos. Para isso, são usados métodos de AM supervisionados como, por exemplo, *k-nearest neighbors* (KNN), regressão logística, *naive bayes*, árvore de decisão, *ensembles* e redes neurais artificiais (Han et al., 2012).

No aprendizado não supervisionado, não há rótulo ou valor de destino fornecido para os dados, estando incluídas neste grupo as tarefas de agrupamento e regras de associação (Harrington, 2012). O Agrupamento, ou *Clustering*, de acordo com Harrington (2012), é um tipo de aprendizado não-supervisionado em que se formam grupos de instâncias similares. O *k-means* é, por exemplo, um algoritmo aplicado em tarefas de agrupamento. Regras de associação têm por objetivo encontrar relações entre elementos de um conjunto de dados, de modo a descobrir regras que estabeleçam combinações de valores com uma certa frequência (Harrington, 2012). O *Apriori* e o *Frequent Pattern-growth* (FP-growth) são exemplos de algoritmos usados para mineração de regras de associação (Rahman et al., 2022).

Em muitos casos de MD, os conjuntos de dados a analisar são de dimensões elevadas (muitos atributos) e as variáveis ou *features* podem ter dependências entre si. Nestes casos, é comum a aplicação ou estudo de técnicas para redução da dimensionalidade, capazes de, com a redução do número de *features* ou variáveis, minimizar os ruídos e melhorar a performance dos algoritmos de AM (Harrington, 2012).

Alpaydin (2010) explica que existem dois métodos principais para reduzir a dimensionalidade: seleção de *features* e extração de *features*. Esses métodos podem ser supervisionados ou não supervisionados, dependendo de usarem ou não as informações de saída. Na seleção de *features*, busca-se encontrar K das *D* dimensões que impactam mais positivamente o modelo, descartando-se as demais dimensões (D - K). Os métodos de seleção de *features* são normalmente categorizados em (Liu & Motoda, 2007; Chandrashekar & Sahin, 2014): *Filter*, *Wrapper* e *Embedded*. No método *Filter* é realizada a seleção sem a necessidade de treinamento de algoritmos de aprendizagem, onde os atributos com menor variação (*score*) podem ser eliminados. O *Wrapper* faz uso de algoritmo de classificação para escolher os melhores atributos, otimizando assim a complexidade computacional. Já o *Embedded* corresponde a uma junção dos métodos anteriores, onde o processo de seleção dos atributos resultantes já está contido no algoritmo de aprendizado. Na extração de *features*, busca-se um novo conjunto de K dimensões que são combinações das *D* dimensões originais. A extração dessas *features* corresponde a um dos métodos de combinação envolvendo, por exemplo, *Principal Component Anlysis* (PCA) (Albuquerque et al., 2022).

3 Trabalhos Relacionados

Os trabalhos relacionados a esta RSL incluem estudos secundários formulados como revisões sistemáticas da literatura ou mapeamentos sistemáticos (Kitchenham & Charters, 2007) que apresentam uma relação com o objetivo de identificar fatores de desempenho de estudantes no ENEM.

O grupo de Noguera et al. (2019), motivado pela discrepância entre o ingresso de homens e mulheres em universidades na área de ciências exatas, desenvolveram um Mapeamento Sistemático da Literatura (MSL) para analisar os fatores possíveis que pudessem justificar o desempenho dos participantes do ENEM entre os anos de 2013 e 2017. Os estudos selecionados foram classificados e agrupados em cinco categorias: pedagogia; inteligência artificial; psicologia; políticas públicas educacionais; e desempenho dos participantes. As autoras não deixaram explícito como chegaram a essa classificação, embora ela tenha sido aplicada ao final do processo de seleção dos estudos primários. A última categoria, desempenho dos participantes, inclui 8 trabalhos, sendo ela a mais relacionada ao tema desta revisão. Os resultados das análises demonstraram que o desempenho dos participantes masculinos foi discretamente superior ao das participantes femininas em matemática e ciências da natureza. Já para linguagens e códigos, o desempenho de ambos os sexos foi quase equivalente. Segundo os estudos desta categoria, os principais fatores socioeconômicos dos participantes do ENEM que foram usados na avaliação do impacto no desempenho dos estudantes foram: sexo; tipo de escola; região dos participantes; raça; classe social; renda. As autoras avaliaram, no estudo, outros fatores, a saber: relação entre sexo e às ciências exatas; ano; região; cor-raça; tipo de ensino médio; e renda mensal. Em especial, sexo e ciências exatas são usados para mostrar suas variações no desempenho dos participantes no ENEM, de modo a embasar discussões, decisões e proposições de ações que minimizem o desequilíbrio de desempenho de gêneros em cursos de exatas.

Lima et al., (2019) desenvolveram uma RSL com objetivo de identificar como os dados do ENEM e do ENADE, disponibilizados pelo INEP, vêm sendo utilizados em estudos. Para isso, identificaram quais os tipos de análises foram realizadas com base nesses dados, considerando o período de 2005 a 2016. Particularmente, em relação ao ENEM, os principais problemas abordados nos estudos encontrados estão relacionados à dificuldade na visualização e análise de dados escolares, à qualidade do ensino, ao desempenho dos estudantes na prova e dificuldades administrativas das instituições. O levantamento feito pelos autores mostrou que, entre os dados do ENEM, os mais usados foram: notas, questões da prova, número de inscritos no ENEM por ano e dados do questionário socioeconômico. Quanto ao desempenho dos estudantes no ENEM, o principal foco dos trabalhos foi avaliar os resultados dos estudantes e das instituições na prova. Em geral, os trabalhos que abordaram esses tipos de problemas, mostraram a relação entre o desempenho dos alunos e a sua situação socioeconômica. Os problemas associados ao desempenho envolvem um melhor entendimento de como os estudantes se desenvolveram na academia. Para os estudos que envolveram os dados do ENEM, a estatística descritiva e a MD foram os principais meios utilizados para prover as análises. Os autores destacam que análises sobre os dados do ENEM e do ENADE em sua maioria são feitas com o objetivo de melhorar o desempenho dos estudantes nos exames na busca por melhor qualidade na educação. A quantidade de estudos feitos sobre o ENADE é bem maior do que em relação ao ENEM. Devido ao número maior de estudos que usaram estatística descritiva para as análises, os autores apontam também a necessidade de se investir em mais pesquisas associadas à MD. Essas pesquisas podem ser realizadas no campo de predição que permitam auxiliar, por exemplo, na detecção de problemas associados ao baixo desempenho dos estudantes.

Ferreira et al. (2021) realizaram um MSL a partir de dados abertos educacionais brasileiros. Mesmo constatando o crescimento de estudos nesta área, a pesquisa revelou que dados da

Educação Básica ainda são pouco explorados quando comparados ao uso de dados da Educação Superior. Com o crescente aumento de dados, entre eles os dados abertos educacionais brasileiros, os autores buscaram mostrar, em uma visão ampla, como a comunidade científica está usando esses dados. Para isso, apresentaram as fontes de dados, *softwares*, algoritmos, métodos, ferramentas e produtos que foram mais usados nos estudos primários. Entre as fontes de dados públicas brasileiras mais utilizadas estão as bases do ENEM e do ENADE, seguidas pelas bases da Prova Brasil e do Censo de Educação Básica. Os autores identificaram que poucos trabalhos desenvolveram ferramentas ou produtos para visualização de dados com bases abertas no Brasil. Com isso, evidenciam a necessidade de investimento em soluções que possam auxiliar gestores e profissionais da educação por meio da análise de dados educacionais.

Já Soares et al. (2021) detectaram, por meio de uma RSL, que existe uma tendência em grande parte dos artigos selecionados de tratar problemas relacionados ao baixo desempenho escolar dos alunos. As principais soluções propostas são baseadas em criação de modelos preditivos que possam, por exemplo, prever notas e tentar melhorar as taxas de desempenho nas provas. Os autores buscaram trabalhos que usaram MD em fontes disponibilizadas pelo INEP, sendo elas: Índice de Desenvolvimento da Educação Básica (IDEB), SAEB, Censo Escolar e Indicadores Educacionais. Entre os principais problemas encontrados, estão fatores relacionados à fragilidade do sistema educacional, evasão e reprovação de estudantes, problemas com a formação dos docentes e baixo investimento na educação. O baixo desempenho escolar foi um problema frequente tratado nos artigos revisados. Dentre as técnicas de MD utilizadas pelos estudos selecionados, houve uma predominância do modelo supervisionado de regressão e análise de correlação. Ambas as técnicas foram muito utilizadas, por exemplo, na análise de desempenho estudantil e de infraestrutura escolar. Apesar de uma maior exploração do desenvolvimento de modelos de regressão, existe uma grande variedade de técnicas que foram aplicadas como, por exemplo, classificação, agrupamento e análises estatística. Os autores limitaram o estudo a quatro bases de dados da Educação Básica, porém sugerem que novos estudos sejam realizados com outras bases mantidas pelo INEP, a exemplo do ENEM, ENADE, Censo dos Profissionais do Magistério e ENCCEJA.

Comparando os trabalhos descritos com a presente RSL, pode-se destacar alguns diferenciais, a saber: (i) A presente RSL considera um período de 10 anos para seleção dos estudos primários que analisaram fatores que podem interferir no desempenho de estudantes no ENEM; (ii) São mapeados, contabilizados e identificados os atributos considerados mais relevantes à análise de desempenho dos estudantes, considerando, para isso, categorias de dados socioeconômicos, de localização, de notas em todas as áreas de conhecimento e do perfil das escolas; (iii) Do ponto de vista do uso de MD, são identificadas as técnicas mais utilizadas pelos estudos cujo objetivo seja a análise e/ou descoberta de fatores associados ao desempenho de estudantes no ENEM; (iv) Por fim, alguns caminhos para novas pesquisas com dados do ENEM são discutidos.

4 Metodologia

Esta RSL seguiu o protocolo descrito por Kitchenham & Charters (2007), sendo executado em três fases principais: planejamento, condução e relatório. Cada fase, com as atividades desenvolvidas, está representada na Figura 1.

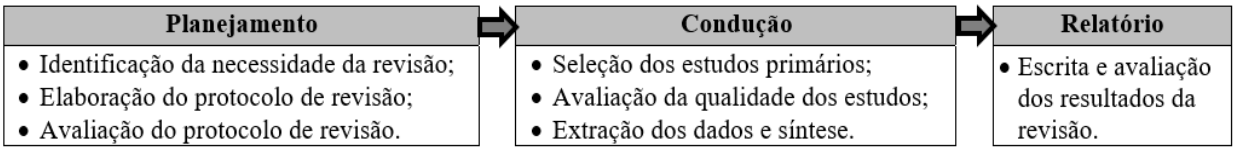

Figura 1: Etapas seguidas na revisão [adaptado de Kitchenham & Charters (2007)].

Durante a fase de planejamento, nenhum dos estudos secundários encontrados estavam totalmente relacionados à identificação de fatores relevantes ao desempenho de estudantes que realizaram o ENEM. Diante desta lacuna percebida, foi definido o protocolo da revisão, bem como as questões de pesquisas. A avaliação do protocolo permitiu realizar algumas adaptações como, por exemplo, nos descritores e *strings* de busca, devido à necessidade de calibrar e obter os estudos relacionados à temática desta RSL.

Na fase de condução, iniciou-se a etapa de seleção mediante os critérios de inclusão/exclusão e de qualidade para se chegar à seleção final de estudos para extração e síntese dos resultados. Por fim, na última fase do processo, os resultados da revisão foram descritos e avaliados.

4.1 Questões de pesquisa

A Questão de Pesquisa (QP) base definida para este trabalho foi: *Quais fatores interferem no desempenho de estudantes em provas do ENEM considerando trabalhos da literatura neste tema nos últimos 10 anos?*

Quatro questões mais específicas foram definidas a partir desta pergunta maior. São elas:

QP1*: Quais atributos foram normalmente considerados na seleção dos dados do ENEM?*

QP2*: Quais atributos, dentre os comumente usados, foram identificados como relevantes à análise de desempenho de estudantes no ENEM?*

QP3*: Quais técnicas e métodos de MD foram utilizados como meio para análises de fatores de desempenho?*

QP4*: Quais desafios de pesquisa foram apontados?*

A QP1 busca reunir os atributos oriundos de *datasets* do ENEM mais frequentemente utilizados nas pesquisas para se ter uma visão geral sobre eles. De modo complementar à QP1, a QP2 objetiva identificar, entre os atributos normalmente considerados, aqueles que, de fato, foram elencados como determinantes à análise de desempenho do estudante no ENEM, sendo este um dos pontos de destaque desta pesquisa. A QP3 buscou identificar as técnicas de MD usadas nos artigos como meio para analisar e identificar os fatores supracitados. Por fim, a QP4 procurou evidenciar e discutir os desafios ou lacunas apontados pelos autores, bem como indicar potenciais trabalhos de pesquisa a serem explorados nesta temática.

4.2 Bases de busca

As bases de busca foram definidas mediante os principais indexadores de publicações disponíveis na literatura na área de Computação e com foco voltado à área de Informática na Educação. Foram consideradas para esta RSL as seguintes bases: *Association for Computing Machinery* (ACM) *Digital Library*, *Institute of Electrical and Electronics* (IEEE) *Xplore*, Portal de periódicos da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Plataforma SOL (biblioteca pertencente à Sociedade Brasileira de Computação - SBC), Revista Brasileira de Informática na Educação (RBIE), Revista Novas Tecnologias na Educação (RENOTE), SCOPUS e Google SCHOLAR.

No entanto, há duas observações que devem ser consideradas: (i) CAPES e SCHOLAR são indexadores e foram usados de modo complementar às bases de busca, o que permitiu a localização de artigos pertencentes a revistas de outras áreas diferentes da computação; (ii) A revista RBIE foi integrada ao sistema de submissão de artigos da Plataforma SOL a partir de junho de 2021. Logo, as buscas realizadas nessas bases retornaram, em algumas situações, estudos duplicados. Nesta revisão esses estudos foram integrados e considerados como oriundos da RBIE.

4.3 *Strings* de busca

Buscou-se tornar a *string* de busca mais simples possível para que fosse encontrado o máximo de resultados que estivessem associados aos descritores: *ENEM, mineração de dados (data mining) e desempenho (performance)*. Esses descritores foram usados nas *strings* de busca, porém foi preciso calibrar a pesquisa em cada uma das bases e indexadores complementares até que fossem encontrados os resultados mais relevantes ao tema desta RSL.

Para as bases ACM e SCOPUS foi usada a *string*: *(ENEM) AND (performance OR data mining)*.

No IEEE Xplore foi usada a *string*: *("All Metadata":ENEM) AND ("All Metadata":DATA MINING) AND ("All Metadata":BRAZIL)*. Nessa base foi necessário acrescentar o descritor "BRAZIL", visto que sem esse termo, o resultado foi de quase seis mil trabalhos.

No portal da SBC e nas revistas RENOTE e RBIE usou-se apenas um único descritor: *ENEM*. Ressalta-se ainda que foi possível identificar algumas publicações das revistas RENOTE e RBIE que apareceram nos resultados dos indexadores e no portal da SBC.

Quanto aos indexadores Google SCHOLAR e CAPES usou-se a *string*: *ENEM AND ((performance AND data mining) OR (desempenho AND mineração de dados))*. O uso dos indexadores permitiu expandir a busca de estudos relacionados ao tema e que estivessem disponíveis em outras bases, além das que foram consideradas.

4.4 Critérios de inclusão e exclusão

Os Critérios de Inclusão (CI) considerados para o processo de filtragem foram: (i) estudos com resultados completos e publicados na íntegra nos idiomas inglês ou português; (ii) estudos publicados nos últimos 10 anos; (iii) estudos que avaliaram o desempenho de estudantes associados diretamente ao ENEM. Os Critérios de Exclusão (CE), por sua vez, foram: (i) estudos cuja disponibilidade seja somente paga; estudos que não usaram dados do ENEM; (ii) estudos secundários, terciários e trabalhos de conclusão de curso; (iii) estudos que não evidenciam dados obtidos por meio do ENEM.

As *strings* de busca foram aplicadas na primeira fase da seleção, onde foi realizada a leitura dos títulos, resumos e palavras-chaves para identificar se o contexto do estudo estaria relacionado às QP definidas. Na segunda fase foi lida a introdução e conclusão dos estudos. Na terceira e última fase, foi realizada a leitura completa de todos os estudos, com o objetivo de apresentar os pontos essenciais, assim como dados relevantes, fatos, ideias e técnicas relacionadas às QP. Após a terceira fase de seleção, os estudos foram avaliados por meio dos Critérios de Qualidade (CQ) descritos na Tabela 1.

Tabela 1: Critérios de qualidade adotados na revisão.

CQ	Descrição
CQ1	A problemática e o objetivo do trabalho são apresentados de forma clara?
CQ2	Há discussão detalhada sobre os resultados obtidos?
CQ3	Objetiva claramente a análise de desempenho do estudante?
CQ4	Foi realizada a avaliação do modelo ou da proposta da solução apresentada?

Kitchenham & Charters (2007) indicam investigar o nível de associação entre os resultados do estudo primário e a qualidade por meio de um escore. Os CQ foram usados como recurso auxiliar para se chegar à seleção final dos estudos primários, bem como para identificar informações durante a etapa de extração e síntese dos dados.

Assim, para cada estudo selecionado na terceira fase, foi aplicado uma pontuação definida por um cálculo simples. A partir de cada CQ, os estudos foram pontuados com os valores: 2 (atende ao CQ); 1 (atende parcialmente ao CQ); 0 (não atende ao CQ). Considerou-se 5 o valor mínimo a ser atingido. Aqueles estudos que atingiram um valor abaixo do mínimo foram descartados. Na última fase de seleção, apenas dois estudos obtidos da base de dados SCOPUS estão entre os que não atingiram a pontuação limite. A decisão pela eliminação foi motivada pelo fato de os estudos terem usado os dados do ENEM para avaliação de redações, no entanto os estudos não objetivaram, ainda que parcialmente, à análise de desempenho do estudante (CQ3), bem como não foram avaliados os modelos apresentados (CQ4). Os dados extraídos foram tabulados e são apresentados neste artigo na seção seguinte.

5 Resultados e discussões

Ao final, foram selecionados 19 estudos primários que atenderam aos critérios de seleção e de qualidade e que, consequentemente, respondem às QP desta RSL. A Tabela 2 mostra a quantidade de estudos selecionados em cada base de dados e indexadores nas três fases da revisão.

Tabela 2: Quantidade de estudos selecionados em cada fase.

Base/Indexador	Fase 1	Fase 2	Fase 3	Seleção final
ACM	26	4	0	0
SCOPUS	94	8	2	0
IEEE	133	18	2	2
SBC	56	11	7	7
RENOTE	7	5	2	2
RBIE	5	1	0	0
SCHOLAR	302	37	5	5
CAPES	22	4	3	3
Total	**645**	**88**	**21**	**19**

A base da SBC reuniu a maior parte dos estudos selecionados (7). As bases internacionais tiveram poucas pesquisas publicadas nesta temática (4), provavelmente pelo fato do ENEM ser um exame aplicado em nível nacional. A base ACM, considerada uma das principais fontes de pesquisa da área da Computação, não teve nenhum estudo selecionado nesta temática. Revistas nacionais também recuperaram um número reduzido de publicações (2) nos últimos dez anos, período de busca considerado nesta RSL. O uso dos indexadores permitiu selecionar 8 estudos relevantes ao tema. Ressalta-se ainda que, destes, 2 foram publicados em revistas internacionais e 3 nas últimas edições do Congresso Brasileiro de Informática na Educação (CBIE). Observa-se que não houve estudos selecionados na revista nacional RBIE e nas bases internacionais ACM e SCOPUS. Além disso, em outras bases, é notório, pelo quantitativo de resultados, o número reduzido de pesquisas, segundo a temática apresentada por esta RSL.

Os dados coletados por meio das leituras dos estudos foram extraídos e organizados de maneira a mostrar e discutir aspectos sobre: a origem dos dados; os principais atributos usados e quais destes foram considerados relevantes à análise de desempenho dos estudantes no ENEM; e as técnicas e métodos de AM usados. Também foram destacados desafios de pesquisas nesta área. Estes dados são apresentados e discutidos no decorrer desta seção.

A Tabela 3 apresenta a lista dos 19 estudos selecionados nesta RSL. Foi atribuído um identificador (ID) para cada estudo, que será usado para referenciá-lo no decorrer deste artigo. As demais colunas indicam os autores, seguidos do ano de publicação do trabalho, o título do artigo e a base onde foi publicado.

Tabela 3: Estudos primários selecionados.

ID	Autor(es)	Título	Base
A1	Silva Filho & Adeodato, 2019	Data Mining Solution for Assessing the Secondary School Students of Brazilian Federal Institutes	IEEE
A2	Santos et al., 2019	Analysis of Candidates Profile for the National Entrance Exams for Admission to Brazilian Universities	IEEE
A3	Gomes et al., 2020	Applying the Regression Tree Method to Predict Students Science Achievement	CAPES (Revista Trends in Psychology)
A4	Pimentel et al., 2021	Learning Time Acceleration in Support Vector Regression: A Case Study in Educational Data Mining	CAPES (Revista Stats)
A5	De Oliveira et al., 2020	Análise do desempenho de pessoas com deficiência que prestaram o exame nacional do ensino médio - ENEM	CAPES (Revista de Educação, Ciência e Tecnologia)
A6	Banni et al., 2021	Uma Análise Experimental Usando Mineração de Dados Educacionais sobre os Dados do ENEM para Identificação de Causas do Desempenho dos Estudantes	SBC (Plataforma SOL)
A7	Franco et al., 2020	Usando Mineração de Dados para Identificar Fatores mais Importantes do Enem dos Últimos 22 Anos	SBC (Plataforma SOL)
A8	Ferreira et al., 2021	A interdisciplinaridade no desempenho da nota de matemática: um olhar para evolução do processo de ensino por meio de modelos regressivos	SBC (Plataforma SOL)
A9	Silva et al., 2020	Identificação de Desigualdades Sociais a partir do desempenho dos alunos do Ensino Médio no ENEM 2019 utilizando Mineração de Dados	SBC (Plataforma SOL)
A10	Lima et al., 2020	Analysis of ENEM's attendants between 2012 and 2017 using a clustering approach	SBC (Plataforma SOL)
A11	Stearns et al., 2017	Prevendo Desempenho dos Candidatos do ENEM Através de Dados Socioeconômicos	SBC (Plataforma SOL)
A12	Gomes et al., 2017	Dados Educacionais Abertos: Associações em dados dos inscritos do Exame Nacional do Ensino Médio	SBC (Plataforma SOL)
A13	Alexandrino Garcia et al., 2021	School performance, infrastructure and teaching practice effects on secondary education in Brazil	SCHOLAR (Revista Brasileira de Estudos de População)
A14	Júnior et al., 2017	Identificando correlações e outliers entre bases de dados educacionais	SCHOLAR (CBIE)
A15	Markoski et al., 2019	Descoberta de Indicadores e Padrões nos Participantes do ENEM	SCHOLAR (Revista Eletrônica de Sistemas de Informação e Gestão Tecnológica)
A16	Silva et al., 2014	Prática de mineração de dados no Exame Nacional do Ensino Médio	SCHOLAR (CBIE)
A17	Alves et al., 2018	Predição do desempenho de Matemática e Suas Tecnologias do ENEM utilizando técnicas de Mineração De Dados	SCHOLAR (CBIE)
A18	Filho et al., 2021	Utilização de notas escolares para predição da nota ENEM em ciências humanas	RENOTE
A19	Vinicios do Carmo et al., 2021	Uma Análise do Desempenho dos Estudantes do Rio Grande do Sul no ENEM 2019	RENOTE

A seguir, são discutidos os resultados obtidos para cada uma das Questões de Pesquisa deste trabalho e, em seguida, considerações sobre os resultados obtidos no tocante à questão de pesquisa principal (QP).

5.1 Atributos normalmente considerados na seleção dos dados do ENEM

A QP1 mostrou todos os atributos extraídos das bases de dados do ENEM que foram utilizados nos 19 estudos selecionados, para demonstrar quais fatores foram considerados na análise de desempenho de estudantes nesta prova. A Figura 2 apresenta uma nuvem de palavras como resultado da frequência de ocorrências de termos correspondentes aos atributos.

Figura 2: Nuvem de palavras gerada por meio dos atributos extraídos dos estudos selecionados.

Para gerar esse conjunto de termos (atributos), foi realizado manualmente um processo básico de integração de dados, visto que os termos podiam corresponder à mesma fonte de origem, mas possuírem nomenclaturas diferentes entre as edições do ENEM. Para isso, foram realizadas três etapas: (i) todos os atributos foram extraídos de cada artigo conforme haviam sido descritos; (ii) os atributos que possuíam a mesma nomenclatura (nome do atributo) e domínio foram agrupados; já os atributos que tiveram a nomenclatura diferente foram comparados e inseridos nos grupos correspondentes; (iii) os atributos agrupados foram fundidos em um único termo (atributo) e contabilizados em relação à quantidade de ocorrências de cada um para geração da nuvem de palavras. Buscou-se manter o termo mais recente entre os atributos agrupados.

Alguns atributos foram nomeados de modo diferente entre os trabalhos. Isso se deu por dois motivos: (i) As bases de dados do ENEM mudaram a nomenclatura desses atributos durante suas reformulações, porém representando as mesmas informações e semântica. Alguns exemplos de como foram considerados os termos (atributos) são vistos em: ‘notas por prova’ que equivale a ‘notas por área’; para os atributos ‘idioma’ e ‘tipo de língua estrangeira’, conveio chamá-las de ‘língua estrangeira’; para os atributos ‘cor da pele’ e ‘raça’, representou-se como ‘raça’. (ii) Em alguns trabalhos houve a derivação de novos atributos para análises específicas sobre eles. Por exemplo, o atributo ‘nota média’ não faz parte dos microdados do ENEM, porém foi derivado por meio das ‘notas por área’ e ‘nota redação’.

5.2 Atributos identificados como relevantes

Partindo dos atributos identificados na QP1, a segunda questão de pesquisa buscou mostrar o subconjunto mais relevante de atributos que determinam o resultado da análise de desempenho do estudante no ENEM. São considerados relevantes todos os atributos das bases de dados do ENEM, que foram indicados ou citados como fatores mais importantes para avaliar o desempenho dos estudantes na prova, embora não tenha sido este exatamente o objetivo principal de todos os estudos selecionados. A RSL permitiu agrupar os atributos em quatro categorias: notas, socioeconômicos, localização e perfil das escolas. A Tabela 4 contém a distribuição dos estudos primários por cada categoria.

Tabela 4: Categorias de atributos usados para análises de dados nas pesquisas.

Categoria	Estudos	Total
Notas	[A1], [A2], [A3], [A4], [A5], [A6], [A8], [A9], [A10], [A11], [A12], [A13], [A14], [A15], [A16], [A17], [A18], [A19]	18
Socioeconômicos	[A1], [A2], [A3], [A4], [A5], [A6], [A7], [A9], [A10], [A11], [A12], [A13], [A15], [A16], [A17], [A18], [A19]	17
Localização	[A2], [A3], [A6], [A10], [A11], [A13], [A14], [A15], [A16], [A19]	10
Perfil das escolas	[A1], [A2], [A3], [A5], [A7], [A12], [A15]	7

A categoria notas contém atributos associados ao desempenho dos estudantes nas provas. Os atributos desta categoria foram os mais considerados pelos estudos (94,7%), sendo eles: notas por área (13), nota da redação (8), nota média (8), nota em língua estrangeira (5). A nota média não pertence aos microdados do ENEM, porém foi derivada das notas por área e redação e usada nas análises de alguns trabalhos.

A categoria socioeconômica compreende atributos sobre a vida escolar e condições sociais, econômicas e culturais do estudante. Citados em 89,5% dos estudos, os atributos socioeconômicos identificados e evidenciados como mais relevantes foram: renda familiar mensal (13), idade (10), sexo (10), raça (9), escolaridade dos pais (9), número de pessoas que moram no domicílio (4) e estado civil (4).

Algumas características socioeconômicas junto aos resultados apresentados pelos estudantes por meio das notas mostram a ligação existente entre a desigualdade social e desigualdade educacional no Brasil, e reforça que há influência das origens sociais e heranças familiares que afetam o desempenho educacional de estudantes, principalmente das classes de renda mais baixa (Castro, 2014). Logo, entende-se que os atributos identificados, nesta revisão, como mais relevantes à análise de desempenho dos estudantes fazem parte destas duas categorias apresentadas.

A categoria localização compreende os atributos que armazenam dados sobre localização geográfica do estudante e da escola. Essa categoria foi considerada como relevante em mais da metade dos estudos (52,6%). Região/uf/município da escola (8), região/uf/município de domicílio do estudante (6) e região de nascimento (2) foram os atributos mais considerados pelos estudos nas análises de desempenho dos estudantes.

Sobre a categoria perfil das escolas, são mencionadas características físicas e pedagógicas, cujos atributos em parte foram retirados de bases de dados, não somente do ENEM, mas das próprias escolas. Os atributos sobre as escolas correspondem à categoria menos evidenciada pelos estudos (36,8%). Foram identificados como mais relevantes neste grupo: tipo de escola (9) e tipo administrativo da escola (3). No caso das escolas, é bem provável que quanto mais investimentos e recursos ela receber, melhor será o resultado apresentado por estudantes e mais bem preparados eles estarão para o mundo do trabalho. No entanto, faz-se necessário o uso de dados e de outros estudos que possam confirmar essa hipótese.

A Tabela 5 mostra a lista de atributos considerados relevantes por cada estudo.

Tabela 5: Lista de atributos relevantes por estudo.

Estudos	Atributos
A1	notas por área; nota da redação; nota média; nota em língua estrangeira; renda familiar mensal; sexo; escolaridade dos pais; raça; número de pessoas que moram no domicílio; tipo de escola.
A2	notas por área; renda familiar mensal; sexo; idade; escolaridade dos pais; região de nascimento.
A3	notas por área; nota em língua estrangeira; renda familiar mensal; sexo; escolaridade dos pais; região/uf/município de domicílio do estudante; tipo de escola; tipo administrativo da escola.
A4	nota média; nota em língua estrangeira; renda familiar mensal; sexo; idade; escolaridade dos pais; raça número de pessoas que moram no domicílio; estado civil; tipo de escola.
A5	nota média; renda familiar mensal; idade; raça; tipo de escola.
A6	notas por área; nota média; nota em língua estrangeira; renda familiar mensal; sexo; idade; escolaridade dos pais; raça; tipo de escola.
A7	nota em língua estrangeira; renda familiar mensal; sexo; raça; tipo de escola.
A8	notas por área; nota da redação.
A9	notas por área; nota da redação; nota média; renda familiar mensal; idade; escolaridade dos pais; raça; estado civil; região/uf/município da escola; região/uf/município de domicílio do estudante.
A10	notas por área; nota da redação; sexo; idade; raça; estado civil; região/uf/município da escola; região/uf/município de domicílio do estudante.
A11	notas por área; nota da redação; renda familiar mensal; sexo; escolaridade dos pais; idade; raça; número de pessoas que moram no domicílio; estado civil; região/uf/município da escola; região/uf/município de domicílio do estudante; região de nascimento.
A12	notas por área; nota da redação; renda familiar mensal; idade; região/uf/município da escola; região/uf/município de domicílio do estudante; tipo de escola.
A13	região/uf/município da escola; tipo administrativo da escola.
A14	notas por área; nota da redação; nota média.
A15	renda familiar mensal; sexo; idade; região/uf/município da escola; tipo administrativo da escola.
A16	nota média; renda familiar mensal; escolaridade dos pais; número de pessoas que moram no domicílio; região/uf/município de domicílio do estudante; tipo de escola.
A17	região/uf/município da escola; notas por área; tipo de escola.
A18	notas por área.
A19	notas por área; nota da redação; nota média; renda familiar mensal; sexo; idade; escolaridade dos pais; raça; região/uf/município da escola.

Em [A1] os principais atributos usados envolveram dados socioeconômicos, notas obtidas e dados sobre a escola. As principais variáveis e regras usadas foram: número de funcionários da escola; renda familiar mensal, ingresso em Universidade Pública; número de computadores nos laboratórios da escola; raça indígena; número de pessoas que moram no domicílio. Ao comparar com resultados obtidos na literatura, os autores concluíram que as características da escola e o nível socioeconômico do estudante estão associados ao desempenho acadêmico dele.

Os aspectos motivacionais, como discutido por [A8], e o sexo fazem parte de características socioculturais na sociedade contemporânea e influenciam diretamente nos resultados alcançados pelos estudantes. Sobre a influência do nível socioeconômico, os resultados mostraram que o valor de corte da renda mensal familiar de 5 salários-mínimos é o valor máximo para discriminar o aproveitamento em ciências pelos alunos. Rendas familiares mensais mais altas não fazem diferença na discriminação do desempenho dos alunos em ciências, segundo os resultados apresentados. Logo, esse fator é visto como de alta relevância para populações com grande desigualdade de renda ou que possuem uma quantidade substancial de pessoas com baixa renda familiar.

O tipo de escola frequentada pelos alunos do Ensino Médio apresentou valor preditivo relevante na discriminação dos estudantes em termos de aproveitamento na prova de Ciências no trabalho de [A3]. Além disso, as regiões brasileiras, o gênero, a motivação, a finalização da

educação secundária até 2011, incluindo uma variação do salário mínimo de acordo com a escola cursada, por exemplo, abaixo de um e meio salário mínimo para uma escola e abaixo de 5 salários mínimos em outra escola.

O trabalho de [A2] apresentou uma visão sobre como o perfil do estudante mudou ao longo de 1998 a 2017 em termos de escolaridade, inclusão social e totalidade de participantes por região. Nos três primeiros anos de aplicação do ENEM, as regiões Sul e Sudeste detinham mais de 80% dos candidatos. Nos demais anos houve o aumento gradual de candidatos nas demais regiões do brasileiras, principalmente no Nordeste. Em 2007, com o programa de Reestruturação e Expansão das Universidades Federais (REUNI) houve o aumento na participação de candidatos do Norte e Nordeste e a ampliação do número de universidades, chegando a mais de 7 milhões de alunos a partir de 2013. Na distribuição de candidatos por gênero, houve a predominância de participantes do sexo feminino. Quanto à distribuição de candidatos por idade, até 2004, 70% tinham no máximo 20 anos. A partir de 2005, houve um aumento no percentual de candidatos na faixa de 21 a 30 anos. Quanto à distribuição de candidatos por cor/raça, até 2002 havia a predominância de candidatos de cor branca, com mais de 60%. A partir de 2003, candidatos declarados pardos e negros têm aumentado consistentemente. Os resultados por renda familiar indicam um aumento constante na participação de pessoas oriundas de famílias mais carentes, no entanto os dados apontam que 65% dos candidatos possuem renda familiar mais alta.

O perfil dos estudantes que participaram do ENEM 2019 foi avaliado em [A4]. As variáveis consideradas no estudo foram: idade; gênero; grupo étnico; estado civil; renda familiar; status de conclusão do ensino médio; ano de conclusão; tipo de escola do ensino médio; tipo de língua estrangeira; nível educacional do pai; nível educacional da mãe; número de residentes na família. Os resultados mostraram que as variáveis etnia, renda familiar e nível de escolaridade dos pais são as mais relevantes quanto ao desempenho obtidos pelos estudantes no exame. A etnia parda é predominante, seguido das etnias branca, preta, amarela e indígena. Há maior concentração de participantes cuja renda familiar está na faixa de R$ 998,00 a R$ 2.994,00. As mães têm maior nível de escolaridade em relação aos pais.

O objetivo do trabalho [A9] foi identificar correlações entre variáveis relacionadas a aspectos socioeconômicos, e de desempenho no exame, ao grupo ao qual o estudante está inserido. As variáveis socioeconômicas, o tipo de administração de escola e a renda familiar foram consideradas as mais preponderantes nas correlações com as notas obtidas. No geral, as variáveis consideradas mais relevantes foram as que evidenciam as desigualdades, tais como: nota média por administração da escola; nota por escolaridade da mãe; e autodeclaração de raça por nota média.

O trabalho de [A5] buscou entender características que interferem no desempenho final do candidato do ENEM realizado em 2018. Esse estudo restringiu-se aos estudantes que possuíam deficiências, ou seja, aqueles participantes que declararam possuir alguma deficiência e que necessitavam de atendimento especial para fazer o exame. Os dados considerados à MD foram: média final (variável alvo para predição), idade, cor da pele, tipo de escola e classe econômica. Para os autores, o problema investigado pode contribuir em novas buscas por soluções de inclusão desses estudantes em cursos superiores, bem como propiciar uma visão geral das suas caraterísticas que podem influenciar no rendimento escolar.

O estudo de [A10] analisou os participantes do ENEM (entre 2012 e 2017), baseado em áreas do conhecimento, tipo de escola e acessibilidade. Os atributos considerados foram: residência por estado, idade, sexo, estado civil, cor/raça, tipo da escola, Unidade Federativa (UF) da escola, dependência administrativa, localização da escola, tipos de deficiência, presença do estudante no exame, o score de 0 a 1000 em cada disciplina de prova (competência), atributos relacionados ao peso dos scores, de 0 a 200 e status da redação. Os resultados mostraram que a

performance geral aumentou durante os anos, e as regiões Sul e Sudeste obtiveram as melhores performances. No Norte e Nordeste houve crescimento gradual do desempenho durante o intervalo de anos considerado no estudo. Os grupos de baixa e média performance de todas as regiões têm, por maioria, estudantes das escolas públicas. Somente 10% são de escolas privadas. Mas no grupo de alta performance a taxa de participação entre estudantes oriundos de escolas públicas e privadas não varia muito. Em 2017 as escolas privadas do Nordeste apresentaram a menor participação, enquanto, por todos os anos anteriores, eram as escolas privadas do Norte com a menor participação. Houve uma mudança no tipo de escola dos estudantes do grupo de alta performance, entre os deficientes. As escolas privadas, em 2017, eram predominantes para os alunos deficientes.

Em [A6] foram identificados atributos mais relevantes para mensurar o desempenho dos estudantes na prova de 2018. Semelhante a [A10], os resultados apontam novamente para os atributos socioeconômicos como fatores significativos ao desempenho apresentado pelos estudantes. Os autores constataram que estudantes de maior poder aquisitivo, oriundos de escolas privadas e federais, das raças branca, parda ou amarela, apresentaram melhor desempenho nessa edição de 2018. Os resultados da pesquisa mostraram que o nível de escolaridade dos pais e localização geográfica influenciam no desempenho dos estudantes em todos os estados brasileiros. A escolha da língua estrangeira inglesa é determinante para um melhor desempenho. De modo geral, o estudo mostrou uma alta correlação entre os atributos escolaridade dos pais, renda familiar, se o estudante é recém formado, raça e faixa etária, com o desempenho obtido.

O estudo de [A7] identificou as vinte principais características de estudantes que concorrem ao ENEM que contribuem para o desempenho alto ou baixo, considerando os dados de vinte e dois anos (1998-2019) de aplicação dos exames pelo INEP. Os atributos mais relevantes à tarefa de classificação identificadas foram: (1) Língua Estrangeira; (2) Grau de importância quanto aos motivos que levaram a participar do ENEM como, por exemplo, para conseguir uma bolsa de estudos (ProUni, outras); (3) O quanto se interessa e acompanha a política internacional; (4) Se indicou ser indígena, qual(is) língua(s) domina; (5) Motivos que levaram a participar do ENEM: conseguir uma bolsa de estudos (ProUni, outras). Os demais quinze atributos foram classificados por maior relevância em ordem decrescente. Os autores destacam os resultados da edição 2012, considerando que se o "estudante cursou o ensino fundamental" foi o fator de classificação mais importante para 2012. Os microdados referentes às primeiras edições do ENEM consideravam essa informação sobre a conclusão do nível fundamental. Neste mesmo ano, o desempenho do aluno foi considerado alto se obteve média maior ou igual a 570,3 pontos. Os autores inferem que provavelmente esses resultados se aplicam nos anos seguintes, baseando-se na hipótese de que o Brasil não avançou muito em termos de qualidade do Ensino Fundamental.

Considerando o exame de 2019, o trabalho de [A8] demonstrou que a nota obtida pelo estudante em Matemática está diretamente relacionada ao desempenho obtido nas demais áreas. Os autores destacam que a desmotivação e questões pessoais são os principais fatores que influenciam no baixo desempenho dos estudantes. Não obstante, a motivação também está relacionada ao bom desempenho escolar. Mostraram também que o desempenho dos estudantes brasileiros nas áreas de leitura, matemática e ciência está muito abaixo nos dados apresentados pelo Programa Internacional de Avaliação de Estudantes (PISA). Entre as áreas avaliadas, a matemática possui a média mais baixa. O Brasil está entre os dez piores países no ranking mundial do PISA 2018 em matemática. O estudo não considerou os dados socioeconômicos. Esse detalhe difere dos outros estudos, visto que os dados socioeconômicos são predominantemente usados. Os únicos atributos considerados para predição foram as notas em: (i) Ciências Naturais (CN): média 477,9; desvio padrão 76,3; (ii) Ciências Humanas (CH): médias 507,3 pontos; desvio padrão não informado; (iii) Linguagens e Códigos (LC): média 520; desvio padrão 64,5; (iv)

Matemática (MT) [variável alvo do modelo]: média 523,5; desvio padrão 109,4; (v) Redação: não foi informada.

Já em [A17], os autores usaram os dados referentes somente à prova de Matemática e dados das escolas dos estudantes para encontrar padrões e gerar um modelo preditivo do indicador de desempenho considerando dados do ENEM 2015. As categorias criadas foram: baixo (instâncias com nota de até no máximo 451); média (instâncias com nota maior que 451 e menor ou igual a 502); alto (nota maior que 502). As variáveis mais significativas quanto ao desempenho na prova de matemática foram: "dependencia_administrativa", "indicador_ de_nivel_socioeconomico" e "categorizacao_taxa_de_participacao".

Em [A11] os autores buscaram predizer a média de estudantes considerando o ENEM 2014 e utilizando somente os dados do questionário socioeconômico. O estudo se restringiu apenas à prova de matemática, cujas notas apresentaram maior variação. O pré-processamento realizado considerou: (i) Filtro de Instâncias: desconsideravam os estudantes que faltaram algum dia; (ii) Seleção de *Features*: não deixaram explícitas as variáveis usadas, mas mostraram o ranqueamento das 10 *features* mais importantes encontradas pelo modelo de predição: Longitude; Latitude; Idade; Motivo para realizar o ENEM: ingressar no ensino privado; Ano de conclusão do Ensino Médio; Renda mensal familiar; Motivo para realizar o Enem: financiamento do FIES; Quantidade de pessoas que moram na mesma residência; Motivo para realizar o ENEM: aumentar possibilidade de emprego; Idade que começou a exercer atividade remunerada; (iii) Discretização: os dados foram discretizados de categóricos para numéricos; (iv) Data Normalization: usaram normalização *Min-Max* para padronização de cada *feature*. Os autores indicam que as informações presentes nos dados socioeconômicos permitem uma predição da nota do estudante.

O estudo de [A12] buscou encontrar possíveis relações entre o desempenho do estudante na prova de Matemática e o seu local de residência (interior/capital), a renda familiar, a escola onde o estudante cursou os ensinos médio e o fundamental. Foram usados os dados do questionário socioeconômico do ENEM 2013 e 2014. Após o pré-processamento dos dados, as variáveis consideradas foram um conjunto de atributos de: dados pessoais; socioeconômicos; notas; e o ano de realização de cada exame. Os resultados sugerem uma relação forte entre o desempenho dos candidatos e a renda familiar, principalmente entre os provenientes de estudantes de escolas públicas. Também foi observado que o desempenho do sexo feminino se concentrou entre a nota mínima e média em matemática.

No trabalho de [A16] foram usados dados do questionário socioeconômico e do desempenho (nota média) do ENEM 2010, a saber: a inscrição do candidato, com sua identificação, unidade da federação, atributos do questionário socioeconômico e nota. Os resultados da pesquisa indicaram que a renda familiar baixa, a escolaridade dos pais de nível primário e a quantidade alta de pessoas no mesmo domicílio, que moram com o estudante, são fatores que implicaram no seu baixo desempenho.

Em [A19] os autores analisaram e compararam o perfil educacional e socioeconômico dos estudantes do estado do Rio Grande do Sul, considerando os dados do ENEM 2019. Foram 136 variáveis consideradas, que podem ser resumidas em: nome do município da escola, a UF da escola, as notas das provas objetivas, e as respostas para o questionário socioeconômico. Os resultados mostraram que o perfil socioeconômico influencia diretamente no desempenho dos estudantes. O desempenho tende a aumentar conforme o aumento da renda familiar. Menos de 1% dos estudantes sem acesso à internet aparecem entre os melhores desempenhos. O grau de estudo maior dos pais também reflete para estudantes que apresentaram melhores desempenhos. Sobre o tipo de escola frequentada, estudantes de escolas privadas e públicas federais se destacam entre os melhores desempenhos, enquanto aqueles vindos de escolas públicas municipais e estaduais aparecem em maior proporção no grupo de piores desempenhos.

Em muitos trabalhos o desempenho é associado às características das escolas. Em [A15] buscou-se a obtenção de indicadores que possam estar relacionados ao desempenho dos participantes do ENEM 2014, evidenciando dados sobre as escolas. Foram usados os atributos: sexo; idade; certificado emitido (atributo discretizado para identificar se o estudante obteve reprovação ou aprovação, segundo as regras que permitem emitir certificado do ensino médio por meio da prova do ENEM); dependência administrativa da escola; tipo de escola; localização (se urbana ou rural); tipo de ensino; classe social pela renda mensal e as notas. Chegaram a resultados que, sintetizando, comprovaram uma porcentagem de aprovados maior na Rede Federal de ensino.

O trabalho de [A18] considerou a relação entre notas do ENEM 2019 e uma escola particular do estado de São Paulo para avaliar a procedência de um modelo preditivo de nota, a partir de dados pedagógicos em relação à área de ciências humanas. Para isso, atributos sobre notas e dados dos estudantes provenientes da escola foram usados, destacando-se: dados de estudantes que iniciaram e terminaram o ensino médio na instituição; sem reprovação; que só prestaram o ENEM no terceiro ano, e não em anos subsequentes; notas de Avaliação Mensal e Avaliação Trimestral. Quanto aos dados do ENEM foi considerada apenas a nota média em Ciências Humanas, obtidas por cada estudante.

O estudo realizado em [A13] analisou os diferenciais de desempenho no ENEM de escolas que oferecem ensino médio, segundo os tipos de ensino e de administração. Para isso, os autores identificaram fatores de qualidade do ensino médio no Brasil, como rendimento escolar, infraestrutura e prática docente. Os atributos considerados no estudo foram: código da escola (usada para junção das bases de dados do Censo Escolar e do ENEM); UF da escola; dependência administrativa; total de alunos concluintes participantes; tipo de ensino médio; indicador de desempenho escolar. Sobre a identificação de rendimento escolar dos alunos da rede pública de ensino, perceberam que não houve grandes disparidades estaduais, mas que houve diferenças regionais. Quanto à infraestrutura, as escolas estaduais de ensino médio sofrem efeitos negativos em relação às demais escolas municipais, federais e privadas. Já quanto à prática docente, a qualificação se apresenta como o fator mais importante.

Em [A14] os autores buscaram identificar discrepâncias entre as notas e as características de infraestrutura das escolas. No entanto, observou-se que as escolas e notas inferidas apresentaram características semelhantes entre si. Presença de laboratórios de computação e ciências, salas de leitura e biblioteca nas escolas estão entre os atributos normalmente considerados para o tipo de estudo em questão, além do próprio desempenho de cada estudante no ENEM. Já os atributos identificados como relevantes foram os seguintes: desempenho no ENEM (nota) por área; presença de água filtrada; se há água da rede pública; se há energia elétrica; se há sistema de esgoto; se há coleta de lixo; se há laboratório de informática; se há laboratório de ciências; se há bibliotecas e se há sala de leitura. Os resultados mostraram que a presença de atributos como laboratórios de computação e ciências, salas de leitura e biblioteca nas escolas indicam um melhor desempenho dos alunos nas provas de Ciências da Natureza e Matemática, Linguagens e Tecnologias e Redação, respectivamente. Outra correlação encontrada foi que as escolas que possuíam água filtrada recebiam água da rede pública, bem como energia elétrica, rede de esgoto e coleta de lixo.

Os estudos selecionados buscaram analisar o desempenho e características dos estudantes no ENEM, cujos resultados foram obtidos, em ampla maioria, por meio de técnicas de MD. A QP3 discute técnicas que foram utilizadas.

5.3 Técnicas utilizadas

Na busca de soluções para as problemáticas identificadas, cada estudo primário usou uma ou várias técnicas de MD e AM a partir de dados disponibilizados do ENEM. De modo mais geral,

as tarefas comumente utilizadas para algum tipo de análise de desempenho de estudantes utilizou: classificação; regressão; seleção de *features*; extração de *features*; associação e/ou agrupamento.

A classificação foi a técnica de AM supervisionado utilizada para predizer o desempenho e identificar características que possam influenciar neste resultado, sendo aplicada em sete estudos: [A1], [A5], [A6], [A7], [A13], [A15], [A17]. Como método, a regressão logística foi usada em [A1], [A6] e [A13] e os demais trabalhos utilizaram os métodos de árvore de decisão como técnica de classificação.

Em [A1] foi calculado o nível de desempenho de um aluno a partir dos atributos com maior influência e a indução de regras de classificação foi usada para identificar nichos de alunos com alto e baixo desempenho. A técnica de árvores de decisão foi usada para explicar o processo de decisão que um especialista humano usaria em uma tomada de decisão baseada em condições. Em [A6], os autores identificaram que, dentre os métodos de AM utilizados na criação dos modelos preditivos, a regressão logística reportou os melhores resultados nos experimentos, quando comparados aos trabalhos que usaram o mesmo método para busca de padrões e geração de modelos preditivos com dados do ENEM.

Em [A13] foi utilizado um modelo de Regressão Logística Múltipla (RLM) para decomposição do efeito escola sobre os alunos nos três componentes (infraestrutura, rendimento e prática docente). As variáveis dependentes foram as classes de nível da qualidade do ensino médio (baixo, intermediário e alto). Os autores apontaram o nível intermediário como a classe de referência do nível da qualidade de ensino médio das escolas públicas, para efeitos analíticos do modelo.

O estudo de [A7] considerou identificar os 20 fatores mais importantes em todas as edições do ENEM, até 2019, avaliando o modelo por meio de algoritmos de classificação para rotular estudantes com desempenho alto ou baixo. O trabalho [A15] apresenta uma proposta de uso da técnica de classificação para realizar a comparação e validação da árvore de decisão gerada pelo algoritmo de mineração com consultas em *Structured Query Language* (SQL). As consultas tinham o objetivo de verificar os padrões encontrados na árvore de decisão, comparando com toda a base de dados. Em [A17], os autores adotaram a técnica de classificação com árvores de decisão para gerar um modelo preditivo do indicador de desempenho das notas da prova de Matemática. Em [A5], os autores usaram a árvore de decisão para predizer qual foi o desempenho final do candidato (ruim, regular, bom, ótimo ou excelente), de acordo com as suas características e se tais atributos influenciam no rendimento final.

A tarefa de regressão foi usada por cinco estudos para estimativa de desempenho. A regressão linear foi usada nos estudos de [A8] e [A18]. O uso de regressão com árvores de decisão foi aplicado em [A3] e [A11]. Em [A4] foi usada a Regressão por Vetores de Suporte (SVR). Em [A18] a técnica de regressão linear demonstrou uma correspondência moderada entre as notas escolares e o desempenho do ENEM em ciências humanas. Em [A8], os autores aplicaram a técnica de regressão linear nas notas de matemática a partir das notas oriundas das demais áreas para saber se o desempenho nas notas de matemática pode estar associado ao melhor desempenho nas demais áreas. A técnica de árvore de regressão usada em [A3] teve o objetivo de realizar uma análise preditiva para estimar a nota (desempenho) dos estudantes em ciências no ENEM 2011, onde foram consideradas variáveis relativas ao questionário socioeconômico. No estudo desenvolvido por [A11] foram usadas árvores de decisão, na tarefa de regressão para predição da nota. Os autores escolheram modelos de regressão baseados em árvore de decisão combinados através de técnicas de *boosting*, aumentando o poder preditivo do modelo. O estudo de [A4] buscou melhorar o desempenho computacional do processo de aprendizagem com um modelo de SVR para prever as notas médias dos candidatos ao ENEM no Brasil.

No contexto de AM supervisionado, a seleção de *features* também foi aplicada em cinco estudos: [A5], [A6], [A7], [A11] e [A19]. Em [A5] foi usada para visualizar quais atributos fornecem um ganho de informação maior entre as classes geradas pela variável média. Por outro lado, em [A7], a técnica foi empregada para identificar quais fatores mais contribuem para o desempenho obtido pelo estudante. Em [A19], os autores mostraram os resultados das análises de distribuições dos resultados dos estudantes com base na influência de atributos socioeconômicos. Em [A6], os autores usaram algoritmos de seleção de *features* para avaliar a taxa de ganho e correlação de informações em relação à classe alvo; com isso, foi possível identificar quantitativamente o nível de influência de um atributo no desempenho dos estudantes. Em [A11] foram identificadas as *features* mais relevantes para o modelo, baseando-se na quantidade de vezes que cada uma foi utilizada em um nó de decisão da árvore.

As regras de associação foram aplicadas em três estudos: [A9], [A12], [A16]. Os resultados obtidos em [A12] indicaram relações entre o desempenho dos candidatos e a renda familiar, sobretudo candidatos provenientes de escolas públicas. Em [A16], a associação de dados permitiu analisar a causa e efeito sobre o desempenho na prova com os fatores socioeconômicos. Por exemplo, um dos resultados obtidos indicou que a renda familiar baixa, a escolaridade dos pais de nível primário e a quantidade alta de pessoas no mesmo domicílio do estudante são fatores que favorecem para um desempenho baixo.

Por fim, técnicas de agrupamento foram aplicadas em apenas dois estudos: [A9] e [A10]. Em [A9] os autores usaram agrupamento juntamente com regras de associação. Para o agrupamento foram consideradas as notas das provas de cada inscrito, enquanto o uso das regras de associação teve o objetivo de identificar os itens frequentes na base de dados e o nível de afinidade entre esses elementos. A partir dos grupos identificados e do conjunto de regras de associação extraídas da base de dados, os autores identificaram correlações entre variáveis relacionadas a aspectos socioeconômicos com o desempenho do estudante no exame e com o grupo ao qual o estudante está inserido. Em [A10] os estudantes foram agrupados baseando-se no desempenho em cada área do conhecimento, a fim de se descobrir grupos com performance similar em cada região do Brasil.

Apenas dois estudos usaram técnicas diferentes das descritas anteriormente: [A2] e [A14]. Em [A2] os autores realizaram uma análise quantitativa dos participantes do ENEM, ao passo que descreve o quanto o perfil do estudante mudou desde a sua implantação do exame. Os resultados sobre a renda familiar, por exemplo, indicaram que houve um aumento constante na participação de estudantes pertencentes a famílias mais carentes. No trabalho de [A14] o objetivo foi adaptar um *framework* de análise de dados a partir de dados do ENEM 2014 e do Censo Escolar em Pernambuco, por meio de análises estatísticas, cálculo de coeficientes de correlação e *outliers* entre informações obtidas das bases do Censo Escolar brasileiro e do ENEM de 2014, porém apenas com dados do estado de Pernambuco.

De modo geral, a Tabela 6 mostra as técnicas e métodos utilizados nos estudos selecionados. Foram identificados 25 métodos que foram agrupados e quantificados a partir das técnicas consideradas.

Para as técnicas supervisionadas de predição (classificação e regressão), percebe-se que há uma predominância no uso de métodos que dão suporte às técnicas de árvores de decisão, bem como para os métodos de *ensemble* na temática desta pesquisa. As árvores de decisão facilitam o entendimento e interpretação dos resultados. Os métodos de árvores de decisão, seguido do número de vezes em que foram usados nos estudos, foram: *J48* (4), *DecisionTree* (1) e *CART* (1). Já os *ensembles* foram: *Random Forest* (2), *XGBoost* (1), *LightGBM* (1) *Gradient Boostin*g (2) e *AdaBoost* (1). Outros métodos utilizados foram: Regressão Linear (3), por meio da Regressão Linear Multivariada, Regressão Linear *Lasso* e *SimpleLinearRegression,* Regressão Logística (3),

Regressão por Vetores Suporte (1), Redes Neurais Artificiais (Perceptron Multi-Camadas) (1), *Naive Bayes* (2) e KNN (1).

Tabela 6: Utilização de técnicas e métodos de MD pelos estudos selecionados.

Método/Técnica	Classificação	Regressão	Seleção de Features	Extração de Features	Associação	Agrupamento
Ensembles	3	4	-	-	-	-
Árvore de decisão	5	1	-	-	-	-
Regressão Linear	-	3	-	-	-	-
Regressão Logística	3	-	-	-	-	-
Naive Bayes	2	-	-	-	-	-
KNN	1	-	-	-	-	-
Regressão por Vetores de Suporte	-	1	-	-	-	-
Redes Neurais Artificiais	-	1	-	-	-	-
Filter	-	-	4	-	-	-
Wrapper	-	-	1	-	-	-
Embedded	-	-	2	-	-	-
PCA	-	-	-	1	-	-
Apriori	-	-	-	-	3	-
K-means	-	-	-	-	-	2
Total	**14**	**10**	**7**	**1**	**3**	**2**

Quanto à técnica de seleção de *features* foram usados métodos para medir o ganho de informação dos atributos em relação às classes alvo. Os métodos *filter* foram: *InfoGainAttributeEval* (1), *GainRatioAttributeEval* (1), *CorrelationAttributeEval* (1) e *Ranker* (1). Quanto ao *wrapper* foi usado o *SequentialFeatureSelector* (1). Já em relação ao *embedded* foram usados os modelos de árvores*: XGBoost* (1) e *ExtraTreesClassifier* (1). Por fim, quanto a técnica de extração de *features*, foi usado o método PCA (1).

Entre os modelos não-supervisionados, destacam-se os métodos: *Apriori* (3), para as regras de associação, e o *K-means* (2), para a técnica de agrupamento.

5.4 Desafios de pesquisa apontados

A maioria dos trabalhos selecionados criou modelos ou aplicou técnicas de MD para descoberta de conhecimento, onde foi possível identificar diversas abordagens sobre como os dados do ENEM têm sido utilizados na busca por analisar fatores associados ao desempenho de estudantes. Diante dos resultados dessas descobertas, a QP4 evidencia alguns desafios apontados a partir desses estudos. Entre esses desafios, destacam-se: (i) necessidade de integração entre os dados do ENEM e de outras fontes; (ii) gestão administrativa e de ensino nas escolas; (iii) análise de *features*; (iv) avaliação de desempenho por área de conhecimento; (v) avaliação de desempenho de estudantes por localidade.

O primeiro desafio destaca a necessidade de integração de dados para o processo de MD do ENEM. Conforme apresentado por [A1], é possível considerar a integração de dados do ENEM com outras bases de dados como, por exemplo, o Censo Escolar, PROUNI e SISU. Com isso, em concordância com [A6], por exemplo, pode-se realizar a predição do resultado do estudante no exame e verificar se é possível indicar os possíveis cursos e instituições de ensino superior em que ele teria condições de disputar uma vaga.

Nessa perspectiva de uso de fontes externas integradas ao ENEM, informações dos estudantes sobre motivação, atitudes e questões sobre transtornos ou deficiências podem auxiliar na investigação do nível de interferência desses fatores quanto ao desempenho do estudante no

ENEM, como destacado nos trabalhos [A3], [A5] e [A10]. Em [A5], por exemplo, foi observado que o tipo de escola, no caso de escolas privadas, contribui consideravelmente para um melhor desempenho do público de estudantes com deficiência e que a renda familiar também é um fator relevante. Os autores mencionam a necessidade de um maior investimento em ferramentas que auxiliem o aprendizado e em políticas de inclusão do estudante deficiente. Em [A3], um desafio foi a correlação entre fatores que influenciam no desempenho em Ciências. Mais especificamente, como interpretar a relação entre escola e renda familiar mensal? Estudantes em determinada escola cuja família receba menos que 5 salários mínimos aparecem com prejuízo tanto quanto estudantes em outra escola que receba menos que um e meio salário mínimo. São dois fatores em que um influencia o outro, apesar de um não ser a causa do outro. Além disso: o que está implícito entre renda familiar mensal e escola específica? Ou seja, não se trata apenas da parte metodológica, mas também do significado socioeconômico das variáveis.

Observou-se alguns desafios sobre gestão administrativa e de ensino nas escolas, sendo esta fundamental para mensurar o desempenho do estudante. Em [A9], os autores apontam desafios relacionados à gestão administrativa e de ensino, bem como sobre a necessidade de se aprofundar nas questões socioeconômicas dos alunos. Um exemplo sobre desafios no ensino é indicado em [A8], onde os autores reforçam a importância de atividades que promovam a interdisciplinaridade entre as áreas de conhecimento, visto que isso poderá ajudar no aumento das notas das disciplinas que compõem os testes avaliativos nacionais e internacionais, como o próprio ENEM e o PISA. Sobre a gestão administrativa, no trabalho de [A18], foram comparados dados socioeconômicos e dados sobre práticas pedagógicas, possibilitando novas investigações acerca de estudos que relacionem esses dados com as notas obtidas pelos estudantes na escola e no ENEM. Os resultados evidenciados por [A12] propõem que haja novas ferramentas de gestão administrativa no intuito de diminuição da evasão dos estudantes e no auxílio para tomadas de decisão pelos gestores.

Ainda sobre os fatores escolares, o trabalho de [A13] mostra que a qualificação docente se mostrou o fator mais impactante no desempenho escolar dos alunos. Os autores afirmam que, devido à grande oferta do ensino médio realizada por escolas públicas, cabe muito mais ao poder público do que à escola a promoção da capacitação, alocação e criação de condições para que docentes qualificados permaneçam nas escolas públicas. Logo, desenvolver estudos que investiguem a adoção de práticas pedagógicas, semelhantes às de escolas de destaque, podem contribuir positivamente em relação ao desempenho de estudantes nas escolas ou em exames classificatórios.

Alguns desafios relacionados à análise de *features* foram identificados. O grupo de [A11] apresentou um ranqueamento das dez *features* mais importantes que podem ser utilizadas como base para entender o viés na nota em matemática do ENEM. Este estudo serve como base para estudos científicos na área de educação e sociedade, com possibilidade de aplicações para melhorar o sistema de ensino brasileiro. Outra possibilidade, segundo o trabalho [A7], seria verificar, de modo complementar, quais interações os fatores podem ter entre si, permitindo compreender melhor cada atributo e sua influência no desempenho dos estudantes no exame.

O estudo de [A10] propõe que novas análises possam ser realizadas no contexto do ENEM, considerando principalmente os dados após 2018. Foi nessa edição que o exame passou por uma das reformulações. Um dos exemplos de mudanças foi a inclusão do critério de avaliação individual por área de conhecimento. Outro exemplo foi a aplicação das provas em dois finais de semanas. Essas mudanças no exame podem ser consideradas para novas investigações a respeito do desempenho de estudantes em edições posteriores.

Observa-se que os estudos selecionados voltados à geração de modelos preditivos do desempenho dos estudantes buscaram trabalhar principalmente com um tipo de área do conhecimento, neste caso, principalmente com a prova de matemática, como visto em [A8],

[A12], [A14] e [A17]. Logo, é importante que se investigue melhor o desempenho acadêmico estudantil, não somente na matemática, mas também nas demais áreas de conhecimento que foram pouco referenciadas, como linguagens, códigos e suas tecnologias e a redação.

Pouco estudos foram aplicados em contextos locais, como visto nos trabalhos de [A14] e [A19]. Esses autores sugerem que os estudos realizados sejam ampliados em relação à abrangência do país. No entanto, também há possibilidade de considerar estudos em nível de estado, ou região, para elaborar outros tipos de análises, incluindo algumas análises visuais dos dados, para avaliar o desempenho, ou identificar perfis de estudantes de uma localidade mais específica. Com isso é possível comparar os resultados entre regiões e estados brasileiros e fornecer evidências que possam apoiar no entendimento da situação educacional desses estudantes no ENEM no Brasil.

5.5 Considerações sobre a QP principal da RSL

De modo mais geral, no tocante à QP principal deste estudo, observa-se que os principais fatores identificados como mais relevantes remetem às questões socioeconômicas, sendo os atributos em maior evidência particularmente os seguintes, em ordem de citação: (i) renda familiar mensal, (ii) idade, (iii) sexo e (iv) raça.

As características dos estudantes, evidenciadas por esses quatro principais atributos, reforçam a indicação de que os diferenciais de desempenho escolar no ENEM estão principalmente relacionados às questões socioeconômicas, principalmente no que diz respeito à renda familiar. Estudantes oriundos de famílias com maior renda têm maiores condições de acesso às escolas privadas, cujos índices de desempenho são maiores, quando comparados, por exemplo, aos estudantes que cursaram o ensino médio em escolas públicas estaduais e municipais. O fator de desempenho pela faixa de idade pode ser explicado pelo fato de que a maior parte dos estudantes que apresentam melhores resultados são aqueles que estão em uma faixa de idade adequada para o período escolar de nível médio. Quanto ao fator sexo, percebe-se haver uma preocupação em investigar o que leva cada grupo de estudantes a apresentarem índices variados de desempenho em algumas áreas como, por exemplo, em Ciências Exatas. O fator relacionado à raça reforça a necessidade de discussões e ações públicas para equilibrar as diferenças de desempenho entre os estudantes de etnias diferentes no ENEM.

Os fatores relacionados à caracterização das escolas de origem dos estudantes, como ‘tipo de escola’ e ‘localização da escola’ também foram evidenciados pelos estudos, que mostram que os melhores desempenhos são de estudantes que estudaram o ensino médio em escola privada ou pública federal. Considerando o nível de renda maior e nível de escolaridade dos pais, neste caso, essa relação implica em pais que oferecem melhores recursos aos filhos, como escolas de boas estruturas física e pedagógica, bons materiais escolares, cursos preparatórios para o ENEM, acesso a computadores e a redes banda larga. Atributos relacionados às notas como, ‘notas por área’, ‘nota da redação’ e ‘nota média’, também estão entre os mais considerados pelos estudos, porém ressalta-se que esses atributos são usados com maior frequência por serem os próprios valores de representação do desempenho dos estudantes na prova do ENEM.

6 Considerações e Trabalhos Futuros

Esta RSL buscou identificar os principais fatores que influenciam no desempenho de estudantes na prova do ENEM. Para isso considerou estudos primários publicados nos últimos 10 anos, nas principais bases de buscas nacionais e internacionais relevantes na área de Informática na Educação. A abordagem da pesquisa discutiu análises sobre atributos considerados relevantes ao

desempenho do estudante, bem como sobre técnicas e métodos de MD comumente utilizadas nos trabalhos como meio para as análises. Ao final, o estudo evidenciou alguns caminhos que podem ser conduzidos em pesquisas complementares.

De acordo com o levantamento, os estudos selecionados mostraram que o desempenho dos estudantes no ENEM é influenciado principalmente por questões socioeconômicas (renda familiar mensal, nível de escolaridade dos pais, sexo, raça), idade e rendimento obtidos nas notas por área de conhecimento e redação. A caracterização das escolas também foi bastante discutida pelos autores, mostrando que há questões pedagógicas e administrativas que impactam diretamente na formação e consequentemente no desempenho dos estudantes no ENEM.

Fatores socioeconômicos como, por exemplo, renda familiar, idade, sexo, raça e escolaridade dos pais, foram alguns dos atributos mais citados nos estudos. Uma investigação pode ser realizada no sentido de rastrear a influência destes atributos em todas as edições do ENEM, ou seja, se realmente há a prevalência destes atributos como sendo os mais relevantes quanto ao estudante apresentar um bom ou mau desempenho no ENEM.

Um desafio interessante seria também investigar o desempenho apresentado por estudantes no ENEM por meio das notas em cada área do conhecimento. Neste caso, seriam analisados os fatores que influenciaram o resultado individualmente por área. Seria possível, por exemplo, constatar se haveria similaridade ou variação destes fatores mais influentes em relação a cada área de conhecimento.

Ao estudar os dados provenientes do ENEM, em várias edições no contexto de desempenho apresentado, surgem algumas questões a se investigar que possam mostrar uma extensão da relação dos fatores de desempenho em específico para estudantes de uma determinada escola ou instituição: É possível aferir outros fatores que influenciaram no desempenho dos estudantes, que não estejam ligados aos que foram identificados por essa RSL? A presença de especialistas ligados à educação poderia auxiliar na busca de soluções para problemas específicos de uma escola?

Poucos estudos selecionados trabalharam com públicos-alvo específicos de estudantes como deficientes, adultos, idosos, presidiários, ou portadores de transtornos como, por exemplo, o autismo. Abordagens que tratem de desempenhos dessas categorias (quantitativo menor) de participantes nas provas podem mostrar resultados importantes. Por exemplo, em novas edições do ENEM foram disponibilizados novos recursos de acessibilidade, logo é possível investigar, historicamente, a influência dos recursos de acessibilidade sobre o desempenho apresentado por estudantes com deficiência e, assim, comparar se a disponibilidade do recurso favoreceu para um melhor desempenho no exame. Além disso, pesquisas como essa podem contribuir no fomento de políticas públicas que acolham esses públicos.

Sugere-se ainda que possa ser realizada a integração dos microdados do ENEM com outras fontes educacionais como, por exemplo, sistemas de gestão acadêmica e sistemas de educação a distância. Uma vez integrados e em conformidade com a LGPD, esses dados podem ser usados para se obter uma melhor compreensão do perfil de estudantes e, consequentemente, avaliar o seu desempenho.

Dessarte, os resultados apresentados por essa revisão contribuem com o estado da arte sobre pesquisas no desenvolvimento de soluções educacionais baseadas em MD que possam melhorar o desempenho do estudante no ENEM, assim como podem servir de referência a outras revisões que usem fontes de dados digitais ou educacionais.

Agradecimentos

Os autores agradecem ao Programa de Pós-graduação em Tecnologia da Informação (PPGTI) e ao Instituto Federal da Paraíba (IFPB) pelo apoio durante a realização desta pesquisa.

Referências

Albuquerque, D., Tarrataca, L., Brandão, D., & Coutinho, R. (2022). A Genetic Algorithm with Flexible Fitness Function for Feature Selection in Educational Data: Comparative Evaluation. *Journal of Information and Data Management, 13*(3). doi: 10.5753/jidm.2022.2480 [GS Search]

Alexandrino Garcia, R., Luiz Gonçalves Rios-Neto, E., & Miranda-Ribeiro, A. de. (2021). School performance, infrastructure and teaching practice effects on secondary education in Brazil. *Brazilian Journal of Population Studies, 38*, 1–32. doi: 10.20947/S0102-3098a0152 [GS Search]

Alpaydin, E. (2010). Introduction to machine learning (2nd ed). MIT Press. [GS Search]

Alves, R. D., Cechinel, C., & Queiroga, E. (2018). Predição do desempenho de Matemática e Suas Tecnologias do ENEM utilizando técnicas de Mineração De Dados. *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*, *7*(1), 469. [GS Search]

Baker, R. S. J. D. (2010). Data Mining for Education. *International Encyclopedia of Education,* 7(3), 112-118. [GS Search]

Banni, M. R., Oliveira, M. V. dos P., & Bernardini, F. C. (2021). Uma análise experimental usando mineração de dados educacionais sobre os dados do enem para identificação de causas do desempenho dos estudantes. *Anais do Workshop sobre as Implicações da Computação na Sociedade (WICS)*, 57–66. doi: 10.5753/wics.2021.15964 [GS Search]

Baradwaj, B. K., & Pal, S. (2012). Mining educational data to analyze students' performance. doi: 10.48550/ARXIV.1201.3417 [GS Search]

Castro, V. G. de. (2014). Determinantes do sucesso educacional: Reflexões teóricas sobre as possibilidades de sucesso escolar em contextos de desvantagem social. *Sociologias Plurais, 2*(1). doi: 10.5380/sclplr.v2i1.64758 [GS Search]

Castro, M. H. G., & Tiezzi, S. (2004). A reforma do ensino médio e a implantação do Enem no Brasil. *Desafios, 65*(11), 46-115. [GS Search]

Cao, L. (2017). Data science: a comprehensive overview. *ACM Computing Surveys (CSUR), 50*(3), 1-42. doi: 10.1145/3076253 [GS Search]

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering, 40*(1), 16-28. doi: 10.1016/j.compeleceng.2013.11.024 [GS Search]

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. *SPSS inc, 9*(13), 1-73. Recuperado de http://www.statoo.com/CRISP-DM.pdf

Coutinho, F. L., Leite, R. S., & Souza Filho, S. A. (2018). Intenção em ingressar no ensino superior: Uma análise sob a perspectiva dos valores e dos fatores motivacionais. *Revista Gestão Universitária na América Latina - GUAL*, 122–145. doi: 10.5007/1983-4535.2018v11n3p122 [GS Search]

De Oliveira, C. G., Barwaldt, R., & Lucca, G. (2020). Análise do desempenho de pessoas com deficiência que prestaram o exame nacional do ensino médio - ENEM. *#Tear: Revista de Educação, Ciência e Tecnologia*, *9*(1). doi: 10.35819/tear.v9.n1.a4038 [GS Search]

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine, 17*(3), 37. doi: 10.1609/aimag.v17i3.1230 [GS Search]

Ferreira, L. A., Rodrigues, R. L., & Souza, R. N. P. M. de. (2021). Dados abertos educacionais brasileiros: Um mapeamento sistemático da literatura. *Anais do XXXII Simpósio Brasileiro de Informática na Educação (SBIE 2021)*, 1186–1195. doi: 10.5753/sbie.2021.218158 [GS Search]

Ferreira, M. B., Amorim, M., Ogasawara, E., & Barbastefano, R. (2021). A interdisciplinaridade no desempenho da nota de matemática: Um olhar para evolução do processo de ensino por meio de modelos regressivos. *Anais da Escola Regional de Informática do Rio de Janeiro (ERI-RJ)*, 41–48. doi: 10.5753/eri-rj.2021.18773 [GS Search]

Filho, J. C., Penteado, B. E., Bittencourt, I. I., & Isotani, S. (2021). Utilização de notas escolares para predição da nota ENEM em ciências humanas. *RENOTE*, *19*(2), 223–233. doi: 10.22456/1679-1916.121211 [GS Search]

Franco, J. J., Miranda, F. L. de A., Stiegler, D., Dantas, F. R., Brancher, J. D., & Nogueira, T. do C. (2020). Usando Mineração de Dados para Identificar Fatores mais Importantes do Enem dos Últimos 22 Anos. *Anais do Simpósio Brasileiro de Informática na Educação*, 1112–1121. doi: 10.5753/cbie.sbie.2020.1112 [GS Search]

Goldschmidt, R., Passos, E., & Bezerra, E. (2015). *Data mining*. Elsevier Brasil. [GS Search]

Gomes, C. M. A., Amantes, A., & Jelihovschi, E. G. (2020). Applying the regression tree method to predict students' science achievement. *Trends in Psychology*, *28*(1), 99–117. doi: 10.9788/s43076-019-00002-5 [GS Search]

Gomes, T., Gouveia, R., & Batista, M. C. (2017). Dados Educacionais Abertos: Associações em dados dos inscritos do Exame Nacional do Ensino Médio. *Anais do Workshop de Informática na Escola*, 895–904. doi: 10.5753/cbie.wie.2017.895 [GS Search]

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques*. (3 ed) Elsevier Inc. (USA): Morgan Kaufmann. doi: 10.1016/C2009-0-61819-5

Harrington, P. (2012). *Machine learning in action*. (Vol. 5). Greenwich, CT: Manning. [GS Search]

Inep | Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Exame Nacional do Ensino Médio (Enem): Apresentação. (2022). Recuperado de: https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enem/enem. Acesso em: 20 mar. 2023.

Inep | Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Exame Nacional do Ensino Médio (Enem): Microdados. 2020. Recuperado de: https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem. Acesso em: 20 mar. 2023.

Jaloto, A., & Primi, R. (2021). Fatores socioeconômicos associados ao desempenho no Enem. *Em Aberto, 34*(112). doi: 10.24109/2176-6673.emaberto.34i112.5002 [GS Search]

Júnior, G. C., Nascimento, R., Alves, G., & Gouveia, R. (2017). Identificando correlações e outliers entre bases de dados educacionais. *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*, *6*(1), 694. [GS Search]

Kelleher, J. D., Mac Namee, B., & D'Arcy, A. (2020). Fundamentals of machine learning for predictive data analytics: Algorithms, worked examples, and case studies. The MIT Press. [GS Search]

Kitchenham, B. A. & Charters, S. (2007). Guidelines for performing Systematic Literature Reviews in Software Engineering (EBSE 2007-001). *Keele University and Durham University Joint Report*. Recuperado de: https://www.elsevier.com/__data/promis_misc/525444systematicreviewsguide.pdf. Acesso em: 12 ago. 2022. [GS Search]

Li, X., & Patel, P. C. (2021). Weather and high-stakes exam performance: Evidence from student-level administrative data in Brazil. *Economics Letters*, *199*, 109698. doi: 10.1016/j.econlet.2020.109698 [GS Search]

Lima, A. M. S., Florez, A. Y. C., Lescano, A. I. A., Novaes, J. V. de O., Martins, N. de F., Junior, C. T., Sousa, E. P. M. de, Junior, J. F. R., & Cordeiro, R. L. F. (2020). Analysis of ENEM's attendants between 2012 and 2017 using a clustering approach. *Journal of Information and Data Management*, *11*(2). doi: 10.5753/jidm.2020.2023 [GS Search]

Lima, P. D. S. N., Ambrósio, A. P. L., Ferreira, D. J., & Brancher, J. D. (2019). Análise de dados do Enade e Enem: Uma revisão sistemática da literatura. *Avaliação: Revista da Avaliação da Educação Superior (Campinas), 24*(1), 89–107. doi: 10.1590/s1414-40772019000100006 [GS Search]

Liu, H., & Motoda, H. (Eds.). (2007). *Computational methods of feature selection*. CRC Press. [GS Search]

Lucena, J. P. O., & Dos Santos, H. N. L. (2020). A relação entre desempenho no Exame Nacional do Ensino Médio e o perfil socioeconômico: Um estudo com os microdados de 2016. *Revista de Gestão e Secretariado, 11*(2), 1–23. doi: 10.7769/gesec.v11i2.994 [GS Search]

Markoski, A., Zancanaro, L., Guerra, P. A. C., Bertolini, C., & Silveira, S. R. (2019). Descoberta de Indicadores e Padrões nos Participantes do ENEM. *Revista Eletrônica de Sistemas de Informação e Gestão Tecnológica, 10*(1). Recuperado de: https://periodicos.unifacef.com.br/index.php/resiget/article/view/1668 [GS Search]

Martinez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernandez-Orallo, J., Kull, M., Lachiche, N., Ramirez-Quintana, M. J., & Flach, P. (2019). Crisp-dm twenty years later: From data mining processes to data science trajectories. *IEEE Transactions on Knowledge and Data Engineering, 33*(8), 3048–3061. doi: 10.1109/TKDE.2019.2962680 [GS Search]

Medeiros, I. C. (2021). O ciclo da inclusão digital: Social-digital-social / Digital inclusion cycle: social-digital-social. *Brazilian Journal of Development*, *7*(8), 75705–75714. doi: 10.34117/bjdv7n8-002 [GS Search]

Melo, R. O., Freitas, A. C. de, Francisco, E. de R., & Motokane, M. T. (2022). Impacto das variáveis socioeconômicas no desempenho do Enem: Uma análise espacial e sociológica. *Revista de Administração Pública*, *55*, 1271–1294. doi: 10.1590/0034-761220200843 [GS Search]

Miranda, P. R., & Azevedo, M. L. N. D. (2020). Fies e Prouni na expansão da educação superior brasileira: Políticas de democratização do acesso e/ou de promoção do setor privado-mercantil? *Educação Formação*, *5*(3), e1421. doi: 10.25053/redufor.v5i15set/dez.1421 [GS Search]

Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill.

Noguera, V., Branco, K., & Ciferri, C. (2019). Gêneros e suas nuances no ENEM. *Anais do Women in Information Technology (WIT)*, 41–50. doi: 10.5753/wit.2019.6711 [GS Search]

Penteado, B. E. (2016). Correlational analysis between school performance and municipal indicators in brazil supported by linked open data. *Proceedings of the 25th International Conference Companion on World Wide Web - WWW '16 Companion*, 507–512. doi: 10.1145/2872518.2890459 [GS Search]

Pimentel, J. S., Ospina, R., & Ara, A. (2021). Learning time acceleration in support vector regression: A case study in educational data mining. *Stats*, *4*(3), 682–700. doi: 10.3390/stats4030041 [GS Search]

Rahman, Md. M., Watanobe, Y., Matsumoto, T., Kiran, R. U., & Nakamura, K. (2022). Educational data mining to support programming learning using problem-solving data. *IEEE Access, 10*, 26186–26202. doi: 10.1109/ACCESS.2022.3157288 [GS Search]

Ramos, J., Rodrigues, R., Silva, J., & Oliveira, P. (2020). CRISP-EDM: uma proposta de adaptação do Modelo CRISP-DM para mineração de dados educacionais. In *Anais do XXXI Simpósio Brasileiro de Informática na Educação*, (pp. 1092-1101). Porto Alegre: SBC. doi: 10.5753/cbie.sbie.2020.1092 [GS Search]

Romero, C., Romero, J. R., & Ventura, S. (2014). A survey on pre-processing educational data. Em A. Peña-Ayala (Org.), Educational Data Mining (Vol. 524, p. 29–64). Springer International Publishing. doi: 10.1007/978-3-319-02738-8_2 [GS Search]

Santos, B., Oliveira, C. G., Topin, L. O. H., Mendizabal, O. M., & Barwaldt, R. (2019). Analysis of candidates profile for the national entrance exams for admission to brazilian universities. *2019 IEEE Frontiers in Education Conference (FIE)*, 1-8. doi: 10.1109/FIE43999.2019.9028381 [GS Search]

Silva Filho, R. L. C., & Adeodato, P. J. L. (2019). Data mining solution for assessing the secondary school students of brazilian federal institutes. *2019 8th Brazilian Conference on Intelligent Systems (BRACIS)*, 574–579. doi: 10.1109/BRACIS.2019.00106 [GS Search]

Silva, L. A., Morino, A. H., & Sato, T. M. C. (2014). Prática de mineração de dados no exame nacional do ensino médio. *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*, *3*(1), 651. [GS Search]

Silva, V. A. A. da, Moreno, L. L. O., Gonçalves, L. B., Soares, S. S. R. F., & Júnior, R. R. S. (2020). Identificação de Desigualdades Sociais a partir do desempenho dos alunos do Ensino Médio no ENEM 2019 utilizando Mineração de Dados. *Anais do Simpósio Brasileiro de Informática na Educação*, 72–81. doi: 10.5753/cbie.sbie.2020.72 [GS Search]

Soares, R. D. C., Neto, N. W., Coutinho, L. R., e Silva, F. J. D. S., dos Santos, D. V., & Teles, A. S. (2021). Mineração de dados da educação básica brasileira usando as bases do INEP: Uma revisão sistemática da literatura. *RENOTE*, *19*(1), 361–370. doi: 10.22456/1679-1916.118526 [GS Search]

Stearns, B., Rangel, F., Firmino, F., Rangel, F., & Oliveira, J. (2017). Prevendo Desempenho dos Candidatos do ENEM Através de Dados Socioeconômicos. In *Anais do XXXVI Concurso de Trabalhos de Iniciação Científica da SBC*. Porto Alegre: SBC. [GS Search]

Vinicios do Carmo, R., Felipe Heckler, W., & Varella de Carvalho, J. (2021). Uma Análise do Desempenho dos Estudantes do Rio Grande do Sul no ENEM 2019. *RENOTE*, *18*(2), 378–387. doi: 10.22456/1679-1916.110257 [GS Search]