Instituto Federal de Educação, Ciência e Tecnologia da Paraíba

Campus João Pessoa

Programa de Pós-Graduação em Tecnologia da Informação

Nível Mestrado Profissional

JANDERSON FERREIRA DUTRA

ANÁLISE DE PERFIS DE ESTUDANTES NO ENEM CONSIDERANDO HÁBITOS DE ESTUDO

DISSERTAÇÃO DE MESTRADO

JOÃO PESSOA – PB

2023

Janderson Ferreira Dutra

Análise de perfis de estudantes no ENEM considerando hábitos de estudo

Dissertação de Mestrado apresentada como requisito final para obtenção do título de Mestre em Tecnologia da Informação pelo Programa de Pós-Graduação em Tecnologia da Informação do Instituto Federal de Educação, Ciência e Tecnologia da Paraíba - IFPB.

Orientadora: Profª Dra. Damires Yluska de Souza Fernandes

João Pessoa – PB

2023

Dados Internacionais de Catalogação na Publicação – CIP
Biblioteca Nilo Peçanha – IFPB, *campus* João Pessoa

D978a Dutra, Janderson Ferreira.
Análise de perfis de estudantes no ENEM considerando hábitos de estudo / Janderson Ferreira Dutra. – 2023.
124 f. : il.

Dissertação (Mestrado em Tecnologia da Informação) – Instituto Federal da Paraíba – IFPB / Programa de Pós-Graduação em Tecnologia da Informação Nível - PPGTI.

Orientadora: Profa. Dra. Damires Yluska de Souza Fernandes.

1. Desempenho acadêmico. 2. Perfis de estudantes. 3. Hábitos de estudo. 4. ENEM. 5. Pandemia Covd-19. I. Título.

CDU 37.046

Bibliotecária responsável Josinete Nóbrega de Araújo – CRB15/116

MINISTÉRIO DA EDUCAÇÃO
SECRETARIA DE EDUCAÇÃO PROFISSIONAL E TECNOLÓGICA
INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DA PARAÍBA

PROGRAMA DE PÓS-GRADUAÇÃO *STRICTO SENSU*

MESTRADO PROFISSIONAL EM TECNOLOGIA DA INFORMAÇÃO

JANDERSON FERREIRA DUTRA

ANÁLISE DE PERFIS DE DESEMPENHO DE ESTUDANTES NO ENEM CONSIDERANDO HÁBITOS DE ESTUDO

Dissertação apresentada como requisito para obtenção do título de Mestre em Tecnologia da Informação, pelo Programa de Pós- Graduação em Tecnologia da Informação do Instituto Federal de Educação, Ciência e Tecnologia da Paraíba – IFPB - Campus João Pessoa.

Aprovado em 28 de Dezembro de 2023

Membros da Banca Examinadora:

Dra. Damires Yluska de Souza Fernandes

IFPB - PPGTI

Dra. Patrícia Cabral de Azevedo Restelli Tedesco

UFPE

Dr. Francisco Petrônio Alencar de Medeiros

IFPB - PPGTI

João Pessoa/2023

Dedico este trabalho: aos meus pais, João e Rosangela; ao meu irmão, Jamerson; e, de modo muito especial, à minha esposa, Karla, e aos meus amados filhos Nícolas, ☆ e Benício. Vocês são luzes que me guiam na vida.

Amo vocês!

AGRADECIMENTOS

Agradeço primeiramente a Deus, pois sem Sua permissão divina não haveria realizações em nossas vidas.

À minha família, pela confiança e apoio. À minha companheira Karla Keliane por sua força e disposição durante as minhas limitações diárias. Aos meus filhos que muito sentiram minha ausência nos momentos de árduo trabalho. Nossa família é o que mais importa!

Aos meus amigos e colegas do campus Cajazeiras, pela amizade e confiança que sempre depositaram no meu trabalho.

A todos os meus colegas de curso, os quais agora posso chamá-los de amigos, e em especial ao colega João Firmino. Compartilhamos muito conhecimento em pouco tempo, juntos aprendemos e dividimos experiências diante das atividades do mestrado.

A todos os professores do PPGTI pelo valioso conhecimento mediado nas maravilhosas aulas, bem como aos demais colaboradores que fazem o curso. Agradeço de modo especial aos professores da banca de defesa pelas valiosas sugestões.

À professora Damires Souza, particularmente, meu muito obrigado por se disponibilizar na orientação deste trabalho e conduzi-lo com muita sapiência e tranquilidade. Compartilharei os seus valiosos ensinamentos, os quais tive o privilégio de adquirir durante nossas conversas e orientações de pesquisa.

Foram todos vocês, professores, que durante esta trajetória no curso nos serviram como guias. Vocês sempre serão nossos grandes mestres que tomaremos como exemplo em nossa vida profissional.

Finalmente, ao Instituto Federal da Paraíba - campus João Pessoa - pelo curso ofertado e pela oportunidade em ampliar o conhecimento discente nesta Instituição à qual tenho grande apreço. A finalização deste curso reafirma o alto grau de comprometimento que temos para conosco e para com a nossa família e, em especial, à comunidade do IFPB.

Muito obrigado!

Janderson Ferreira Dutra.

RESUMO

O Exame Nacional do Ensino Médio é um processo abrangente de avaliação do desempenho de estudantes da Educação Básica. Durante a pandemia, os estudantes tiveram o seu desempenho impactado por diversos fatores, onde são incluídos neste aspecto questões socioeconômicas e de hábitos de estudos. Neste contexto, este trabalho propõe uma abordagem baseada em aprendizado de máquina não supervisionado para identificação de perfis de desempenho de estudantes e associações de características de hábitos de estudo durante um período da pandemia da COVID-19. A abordagem se baseia na identificação de atributos relevantes ao entendimento do desempenho do estudante a partir de microdados comuns obtidos no ENEM e, com base neles, na agregação de dados relacionados aos hábitos de estudo durante a pandemia. Métodos de agrupamento e de regras de associação são utilizados para a identificação dos perfis e das associações entre as características mais relevantes ao entendimento do desempenho do estudante no período referido. Os resultados mostraram que os estudantes que estabelecem melhores gestão e práticas de atividades em sua rotina de estudos, mesmo diante de dificuldades de infraestrutura e pouco acesso a meios tecnológicos, puderam almejar bons resultados no exame. Em relação aos fatores que mais influenciam na caracterização dos perfis estão os hábitos de estudo dos grupos temáticos referente à gestão do tempo, planejamento e práticas de estudo e pesquisa. Considerando o cenário mais geral com todos os dados obtidos, prevalecem as regras de associação que envolvem os fatores considerados essenciais às atividades de estudo durante a pandemia como, por exemplo, presença de celulares, computadores e acesso à internet na residência. Em relação aos dados de hábitos de estudo, em geral, são destacados aqueles relacionados à gestão de tempo, organização e aproveitamento dos estudos na preparação para a prova. Já em relação aos dados do ENEM, os resultados endossam que fatores socioeconômicos, principalmente a renda familiar, impactam fortemente no desempenho. O panorama de impactos de hábitos de estudo apontado neste trabalho apresenta informações que podem auxiliar na definição de diretrizes e políticas públicas de apoio aos estudantes da Educação Básica rumo ao ENEM.

Palavras-chaves: Perfis de estudantes; Desempenho acadêmico; ENEM; Pandemia; Hábitos de Estudo.

ABSTRACT

The National High School Examination (ENEM) is a comprehensive process for evaluating the performance of Basic Education students. During the pandemic, students' performance was impacted by several factors, including socioeconomic issues and study habits. In this context, this work proposes an approach based on unsupervised machine learning to identify student performance profiles and associations of study habit characteristics during a period of the COVID-19 pandemic. The approach is based on the identification of attributes relevant to understanding student performance based on common microdata obtained from ENEM and, based on them, the aggregation of data related to study habits during the pandemic. Clustering methods and association rules are used to identify profiles and associations among the most relevant characteristics for understanding the student's performance in the referred period. The results showed that students who establish better management and activity practices in their study routine, despite infrastructure difficulties and little access to technological means, were able to achieve good results in the exam. With respect to the factors that most influence the characterization of profiles are the study habits of thematic groups regarding time management, planning and study and research practices. Considering the more general scenario with all the data obtained, the association rules involve factors considered essential to study activities during the pandemic, such as, the presence of cell phones, computers and internet access at home. Regarding data on study habits, in general, those related to time management, organization and use of studies in preparation for the test are highlighted. In relation to ENEM data, the results endorse that socioeconomic factors, mainly family income, have a strong impact on student’s performance. The impacts of study habits highlighted in this work present information that may help in defining guidelines and public policies to support Basic Education students towards ENEM.

Keywords: Student profiles; Academic performance; ENEM; Pandemic; Study Habits.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

SUMÁRIO

1. INTRODUÇÃO

Este capítulo apresenta a contextualização na qual esta pesquisa está inserida, incluindo a motivação, as questões de pesquisa e proposta de solução, as contribuições esperadas, os objetivos, a metodologia e, ao final, a estrutura em que a dissertação está organizada.

1.1. CONTEXTUALIZAÇÃO E MOTIVAÇÃO

O Exame Nacional do Ensino Médio (ENEM) caracteriza-se como um dos principais e mais complexos sistemas de avaliação de desempenho acadêmico, visto que alcança uma maior diversidade de estudantes participantes nas mais diversas classes sociais distribuídas em praticamente todos os municípios brasileiros. Além disso, o exame vem sendo aplicado e melhorado há mais de 25 anos, atingindo um nível de confiabilidade também no que se refere a seus microdados publicados (INEP, 2022).

Os microdados relativos às edições do ENEM estão disponíveis publicamente e podem ser usados para extrair diversas informações sobre os estudantes possibilitando que, por exemplo, gestores e profissionais da educação possam ter subsídios para buscar entender problemas educacionais e, assim, possam desenvolver estratégias que visem à melhoria de infraestrutura e da qualidade do processo de ensino e aprendizagem. Por meio dos microdados, busca-se, por exemplo, investigar quais aspectos inerentes aos estudantes, como dados escolares e socioeconômicos, podem influenciar em seu desempenho acadêmico medido por meio das provas do ENEM.

O ENEM é um exemplo de processo de avaliação de ensino, tendo em vista que, desde sua criação, tem como objetivo fundamental “avaliar o desempenho do aluno ao término da escolaridade básica, para aferir o desenvolvimento de competências fundamentais ao exercício pleno da cidadania (BRASIL, 2002, p.5)”. De modo geral, desempenho acadêmico refere-se ao rendimento do estudante em ciclos de educação e expressa a aprendizagem obtida em um processo de ensino que envolve alguns fatores como, por exemplo, a estrutura da instituição, o grau de qualificação do corpo docente, questões sociodemográficas da família e da escola, o nível de escolaridade dos pais e a renda familiar (ARGÔLO, 2017).

Nos últimos dez anos, alguns estudos buscaram identificar fatores que podem influenciar no desempenho de estudantes no exame do ENEM. A Revisão Sistemática da Literatura (RSL) realizada por este autor e outros (Dutra et al., 2023) identificou que, de modo mais geral, os atributos pertencentes aos microdados do ENEM nos últimos anos considerados mais relevantes à análise de desempenho dos estudantes são: renda familiar mensal, idade, sexo, raça, tipo de escola, tipo administrativo da escola, localização da escola, nível de escolaridade dos pais, e notas por área de conhecimento. Os autores categorizam tais atributos em dados socioeconômicos, de localização, de notas em todas as áreas de conhecimento e de perfil das escolas.

Em março de 2020 ocorreu a confirmação do primeiro caso do novo coronavírus (Sars-Cov-2) no Brasil, período no qual a Organização Mundial da Saúde (OMS) declarou oficialmente a pandemia da COVID-19 (HINAZZI et al., 2020; JARDIM; BUCKERIDGE, 2020; WHO, 2020). Devido à pandemia de COVID-19 e suas consequências, a rotina de estudos de estudantes mudou, implicando na sua preparação acadêmica e, em alguns casos, supostamente afetando o seu desempenho.

Algumas medidas consideradas importantes foram adotadas para deter o rápido avanço das infecções entre a população. Uma das principais medidas foi o distanciamento social (KISSLER et al., 2020). Assim como nos mais diversos setores sociais, as instituições de ensino buscaram adaptar estratégias metodológicas para que houvesse a continuidade das atividades acadêmicas, amenizando o impacto causado pelo fechamento das escolas. Vale ressaltar que muitas escolas, em atendimento às recomendações de distanciamento social, precisaram suspender por tempo indeterminado as atividades presenciais como medida de evitar o contágio entre estudantes, professores e demais funcionários.

Com o objetivo de promover a retomada das atividades escolares, o Ministério da Educação e Cultura (MEC) publicou a Portaria[1] nº 343, de 17 de março de 2020, como forma de normativa que dispôs sobre a substituição das aulas presenciais por aulas em meios digitais enquanto durasse a pandemia do novo coronavírus. Com isso, as instituições passaram a utilizar o ensino remoto emergencial como alternativa durante o estado de calamidade proveniente da pandemia. No entanto, nem todas as instituições de ensino e professores estavam preparados para oferecer essa modalidade de ensino aos estudantes, que por sua vez também não dispunham, em algumas situações, de infraestrutura adequada para participar das atividades acadêmicas. Exemplos de dificuldades foram, entre outras, o acesso à internet e a equipamentos computacionais (SOARES; SILVA, 2020).

De modo geral, o ensino remoto não pode garantir o acompanhamento adequado do aprendizado do estudante, pois, durante o processo de ensino, algumas lacunas podem surgir, não sendo possível, por exemplo, prever as dificuldades enfrentadas pelos estudantes diante da sua realidade familiar. Pode-se destacar pontos positivos e negativos em relação ao método de ensino remoto. Entre os pontos positivos está o uso de meios tecnológicos como forma segura de comunicação e interação nas aulas sem contato físico evitando-se a proliferação de doenças, enquanto que a falta de acesso à tecnologia a todos os estudantes está entre os pontos negativos.

Ainda que algumas escolas tenham conseguido se adaptar com menor dificuldade à modalidade de ensino remota, essa não foi a realidade de algumas instituições de ensino cujas atividades escolares eram apoiadas por toda uma infraestrutura voltada para o ensino presencial.

Na edição do ENEM do ano de 2022, além dos microdados tradicionais, um novo conjunto de dados sobre hábitos de estudo foi publicado pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). Tais dados foram coletados por meio de um questionário

[1] Disponível em: https://www.planalto.gov.br/ccivil_03/portaria/prt/portaria%20n%C2%BA%20343-20-mec.htm. Acesso em: 02 jul. 2023.

que buscou compreender aspectos da rotina de estudos e preparação dos estudantes para o ENEM durante a pandemia (INEP, 2022).

O desempenho acadêmico no ENEM pode ser influenciado por diversos fatores, sendo os hábitos de estudo citados aspectos adicionais que podem indicar como foi possível o envolvimento acadêmico dos estudantes nos estudos durante a pandemia.

Partindo-se do pressuposto de que os atributos considerados mais relevantes identificados por este autor e outros (Dutra et al., 2023) em uma RSL podem ser usados como uma lista consolidada de fatores importantes à análise de desempenho de estudantes no ENEM, este trabalho busca agregar mais informação e perspectivas considerando os hábitos de estudo durante um período da pandemia. Com a integração desses dois conjuntos de dados (Microdados do ENEM + hábitos de estudo) busca-se identificar e entender melhor perfis de estudantes durante a pandemia com respeito a seu desempenho no ENEM e como esses fatores podem estar ou não associados.

1.2. QUESTÕES DE PESQUISA E PROPOSTA DE SOLUÇÃO

Após a realização do levantamento do estado da arte da pesquisa em relação aos fatores que impactam o desempenho do estudante no ENEM por meio da RSL supracitada por meio da análise de alguns outros trabalhos relacionados, discutidos no Capítulo 3 deste documento, não foram encontrados trabalhos que avaliassem o desempenho de estudantes no ENEM, considerando perfis de estudantes com respeito ao hábitos de estudo durante a pandemia.

Neste panorama, duas questões de pesquisa foram definidas para nortear este trabalho. A primeira refere-se a:

QP1: ***Considerando dados do ENEM e de hábitos de estudo durante a pandemia da COVID-19, como identificar perfis associados ao desempenho de estudantes para o referido exame?***

Para responder a esta questão, pressupõe-se que os dados de hábitos de estudo podem agregar na identificação de perfis de estudantes, tanto para mostrar características e dificuldades que possam levar ao insucesso na prova durante o período da pandemia, quanto também para casos de sucesso em relação ao desempenho.

De modo complementar à QP1, a segunda questão foi definida como segue:

QP2: ***Quais regras de associações de características podem ajudar a ratificar ou entender melhor os perfis de desempenho de estudantes no exame do ENEM durante a pandemia?***

Para ajudar a responder as questões postas, este trabalho propõe uma abordagem baseada em aprendizado não supervisionado, usando métodos de agrupamento e regras de associação (HARRINGTON, 2012). Este trabalho tem como principal propósito compreender melhor os impactos causados pela pandemia em estudantes que realizaram o ENEM a partir da identificação de perfis associados ao desempenho no ENEM. Para isso, constrói a abordagem proposta considerando apenas a edição do ENEM referente ao ano de 2022, quando os microdados do exame foram disponibilizados juntamente aos dados de hábitos de estudo. Os resultados

encontrados podem auxiliar educadores e estudantes quanto ao melhor entendimento acerca dos impactos acadêmicos no ENEM durante a pandemia.

1.3. OBJETIVOS

1.3.1. Objetivo geral

Propor e desenvolver uma abordagem baseada em aprendizado de máquina não supervisionado para identificação de perfis de desempenho de estudantes no ENEM e associações de características de hábitos de estudo durante um período da pandemia da COVID-19.

1.3.2. Objetivos específicos

Para atingir o objetivo geral, busca-se contemplar os seguintes objetivos específicos:

- Identificar os atributos relevantes ao entendimento do desempenho do estudante a partir de microdados do ENEM e de hábitos de estudo durante a pandemia;
- Coletar e preparar dados do ENEM e de hábitos de estudo para identificar padrões e associações de dados;
- Elaborar uma abordagem de identificação de perfis de desempenho de estudantes quando atuam no ENEM considerando hábitos de estudo;
- Identificar as principais associações de características mais relevantes que implicam no desempenho dos perfis de estudantes considerando hábitos de estudo;
- Avaliar os resultados obtidos a partir da abordagem e discorrer sobre o perfil e regras de associação mais importantes ao desempenho acadêmico dos grupos de estudantes no ENEM durante um período da pandemia considerando hábitos de estudo.

1.4. METODOLOGIA DA PESQUISA

Este trabalho de pesquisa é de natureza aplicada e do tipo exploratória, tendo em vista que trata-se de uma investigação direcionada à construção prática da solução para o objeto de investigação. Quanto à finalidade, ela é experimental (KÖCHE, 2016). Sobre a abordagem do problema, trata-se de uma pesquisa quantitativa, cuja análise e interpretação dos resultados são apoiadas por meio de tabelas e visualizações gráficas (PRODANOV e DE FREITAS, 2013).

1.5. ORGANIZAÇÃO DA DISSERTAÇÃO

Neste capítulo inicial da dissertação, foram destacadas a contextualização e motivações que levaram à pesquisa, bem como os seus objetivos. Os capítulos subsequentes estão organizados conforme a seguinte estrutura: no Capítulo 2 está disposta a fundamentação teórica, contendo os conceitos essenciais à compreensão deste trabalho; no Capítulo 3 encontram-se os trabalhos relacionados, onde foram realizadas comparações com o intuito de identificar características principais acerca das implicações da pandemia no ENEM; no Capítulo 4 é apresentada a abordagem proposta, bem como é mostrado todo o processo de desenvolvimento do estudo e os

principais diferenciais deste trabalho; no Capítulo 5 são discutidos e descritos os resultados obtidos com a avaliação experimental. Por fim, o Capítulo 6 contém as considerações finais, onde se discorre sobre as principais contribuições obtidas e algumas propostas de continuidade da pesquisa.

2. FUNDAMENTAÇÃO TEÓRICA

Este capítulo introduz conceitos relacionados aos temas associados ao presente trabalho. Inicialmente são mostrados, de modo geral, aspectos que envolvem o ENEM e o recente questionário associado aos hábitos de estudo. Em seguida, são discutidos fundamentos voltados à área de descoberta e extração de conhecimento em bases de dados, particularmente, em conjuntos de dados educacionais. Logo após, o capítulo apresenta conceitos voltados ao Aprendizado de Máquina (AM). Alguns dos algoritmos de AM são descritos e exemplificados, evidenciando principalmente a abordagem não supervisionada, foco deste trabalho, bem como algumas medidas de avaliação que foram usadas nos experimentos da pesquisa.

2.1. ENEM

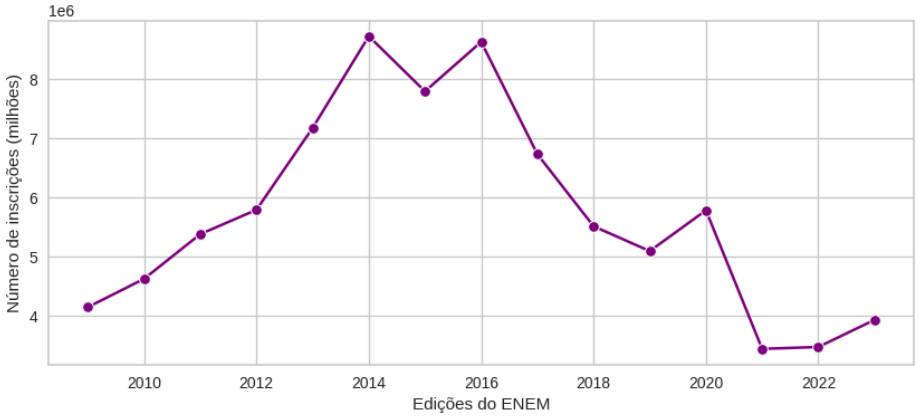
Para avaliar aspectos de ensino, pesquisa, extensão, responsabilidade social, desempenho de estudantes, gestão, professores e infraestrutura das Instituições de Ensino Superior (IES), foi criado o Sistema Nacional de Avaliação da Educação Superior (SINAES) - Lei n° 10.861, de 14 de abril de 2004. O SINAES possui alguns instrumentos de avaliação de cursos, entre eles, o Exame Nacional do Ensino Médio (ENEM), criado pelo MEC por meio da Portaria nº 438, de 1998. O ENEM é um meio de avaliação do conhecimento universal e das habilidades adquiridas ao longo da Educação Básica (BRASIL, 2002). A proposta inicial foi usar o exame como um instrumento técnico para avaliar as competências dos estudantes concluintes do ensino médio. Desde então, anualmente, o exame é usado como instrumento de coleta de dados sobre estudantes egressos da Educação Básica e para verificar o nível de interpretação sobre acontecimentos em relação à realidade brasileira e mundial. Além disso, ele é a principal porta de acesso à Educação Superior.

O ENEM passou por diversas mudanças desde a sua criação. Um marco importante, por exemplo, ocorreu em 2009, quando: (i) o número de questões objetivas da prova passou de 63 para 180; a aplicação passou a ser realizada em dois dias sequenciais (sábado e domingo); (iii) foi adotada a Teoria da Resposta ao Item (TRI) na elaboração das provas (INEP, 2021).

No ano de 2020 houve a inclusão, em fase de testes, do ENEM Digital. Este formato de aplicação foi limitado apenas para estudantes que concluíram ou estavam para concluir o ensino médio até o final de 2020. A prova foi realizada por meio de computadores sem acesso à internet, inicialmente em algumas capitais brasileiras e em datas diferentes da aplicação impressa. No entanto, o INEP cancelou o ENEM Digital para 2023, devido ao alto custo e baixo interesse dos estudantes (RODRIGUES, 2023).

A Figura 2.1 mostra o número de estudantes inscritos no ENEM a partir de 2009. Historicamente, o ENEM vinha evoluindo quanto ao quantitativo de participantes porém, a partir de 2017, a participação diminuiu devido a problemas relacionados, por exemplo, à crise econômica de 2015, à Pandemia iniciada em 2020 e a fatores políticos e socioeconômicos que vieram nos anos que se sucederam (LEAL, 2022).

Figura 2.1 - Número de estudantes inscritos no ENEM.



Fonte: Elaborado pelo autor, com dados de INEP (2023a).

Para a edição mais atual, neste ano de 2023, cerca de 3,9 milhões de estudantes confirmaram a inscrição no exame. Observa-se que, a partir de 2021, houve uma retomada no que diz respeito ao aumento de estudantes inscritos, se comparado aos dados de anos anteriores (INEP, 2023a). No entanto, a quantidade de inscrições em 2023 ainda é inferior ao ano de 2009. Esse cenário mostra que ainda há muitos desafios para gestores e estudantes diante dos problemas relacionados, entre eles, por exemplo, o período de pós-pandemia.

As IES brasileiras têm usado este exame no decorrer dos últimos anos como instrumento seletivo para ingresso em cursos superiores. Da mesma forma, os resultados obtidos pelos estudantes no ENEM são usados para seleção de bolsas integrais ou parciais no Programa Universidade para Todos (ProUni), bem como para financiamento dos estudos durante a graduação através do Fundo de Financiamento Estudantil (FIES). As notas do exame também são aceitas em mais de 50 instituições de educação superior portuguesas (INEP, 2021).

Criado em 2010, o Sistema de Seleção Unificado (SISU) passou a ser o principal meio de seleção dos candidatos às vagas do ensino superior. Com base no desempenho obtido no ENEM, o SISU permite o ingresso do estudante por cotas ou por ampla concorrência (BRASIL, 2023). Um dos aspectos positivos da relação entre ENEM e SISU é o favorecimento da mobilidade de estudantes para IES nos mais variados locais do país, possibilitando o deslocamento de estudantes para regiões mais desenvolvidas, promovendo um ambiente multicultural nas universidades (SILVEIRA; BARBOSA; SILVA, 2015).

Enquanto 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 (DE CASTRO; TIEZZI, 2004).

O ENEM pode ser realizado por quaisquer pessoas que concluíram o ensino médio, ou que estejam concluindo, para ter acesso ao ensino superior. Os participantes que ainda não concluíram o ensino médio podem participar como “treineiros”, porém o resultado serve somente como instrumento de preparação e de autoavaliação de conhecimentos. Além dos treineiros, há grupos

de participantes em menor quantidade que recebem apoio legal para realizarem a prova, como por exemplo, pessoas com deficiência, que são devidamente assistidas pela Política de Acessibilidade e Inclusão do INEP e pessoas privadas de liberdade (BRASIL, 2023).

Entre os motivos favoráveis à adoção do ENEM, ANDRIOLA (2011) destaca que os itens (questões) componentes da prova do ENEM buscam avaliar habilidades e competências, a partir de problemas cujas soluções transcendam o conhecimento formal sobre os conteúdos escolares. As questões permitem que o candidato possa interpretar, inferir, deduzir, comparar, julgar, aplicar e resolver situações-problema. Sob essa nova ótica, o estudante terá que demonstrar suas competências para, a partir de informações que lhe foram apresentadas, empregá-las a fim de propor soluções factíveis para problemas que envolvem conteúdos curriculares.

Os candidatos são avaliados a partir de cinco provas: Ciências Humanas e suas Tecnologias (CH); Ciências da Natureza e suas Tecnologias (CN); Linguagens, Códigos e suas Tecnologias (LC); Matemática e suas Tecnologias (MT); e Redação. Ao todo são 180 questões objetivas (45 para cada uma das quatro áreas de conhecimento) e uma redação, às quais são atribuídas notas. As notas médias das áreas de conhecimento são calculadas por meio da Teoria de Resposta ao Item (TRI) (ANDRADE; TAVARES; VALLE, 2000). Já a redação corresponde a um texto dissertativo-argumentativo a partir de uma situação-problema (BRASIL, 2022).

De acordo com o Guia do Participante, o modelo matemático TRI calcula o item de resposta de acordo com três informações: (i) parâmetro de discriminação, que é a capacidade de um item distinguir os estudantes que têm a habilidade requisitada numa questão daqueles que não a têm; (ii) parâmetro de dificuldade, que refere-se à dificuldade da habilidade avaliada na questão, onde, quanto maior seu valor, mais difícil é a questão; (iii) parâmetro de acerto casual (chute), que ocorre quando um participante, que não domina a habilidade avaliada em uma determinada questão objetiva, pode responder corretamente um item por acerto casual. Esse parâmetro representa a probabilidade de um estudante acertar a questão não dominando a habilidade exigida (INEP, 2021).

Para que o estudante compreenda as diretrizes de habilidades e competências exigidas nas provas, o INEP disponibiliza um documento intitulado Matriz de Referência do ENEM[2]. Este documento pode servir de apoio para o estudante que busque obter um melhor desempenho na prova, tendo em vista que é nele onde estão localizadas as competências e habilidades exigidas nas questões da prova, contribuindo para que estudantes e professores consigam entender a estrutura da prova à qual está sendo avaliado.

Segundo Brasil (2020), os dados do ENEM estão disponíveis sob a regulação do Plano de Dados Abertos (PDA), que é um instrumento que operacionaliza a Política de Dados Abertos do Poder Executivo Federal, pois planeja as ações que visam a abertura e sustentação de dados abertos nas organizações públicas. A disponibilização desses dados para a sociedade em formato aberto incentiva a participação social através de estudos e pesquisas. Para isso, o INEP disponibiliza esses dados de modo estruturado e sem possibilidade de identificação de pessoas que

[2] Disponível em https://download.inep.gov.br/download/enem/matriz_referencia.pdf. Acesso em: 08 jul. 2023.

participam do exame, atendendo às normas da Lei Geral de Proteção de Dados Pessoais (LGPD)[3]. Informações sensíveis, como o código de identificação da escola e dados pessoais dos estudantes, dentre outros, foram retiradas de todas as edições dos microdados, após alterações realizadas pelo INEP no ano de 2020.

Durante a inscrição, os estudantes preenchem um questionário que contempla questões sobre a caracterização socioeconômica familiar. Em 2022 o questionário era composto por 25 questões. Os resultados individuais no exame, bem como as respostas ao questionário são disponibilizados a cada edição no formato de microdados no portal do INEP[4].

2.2. QUESTIONÁRIO DE HÁBITOS DE ESTUDO

Na edição de 2022 do ENEM, foi disponibilizado aos estudantes um questionário intitulado "Hábitos de estudo dos participantes do Enem em contexto de pandemia". O questionário de Hábitos de Estudo (HE)[5] contém 34 perguntas de marcação única ou de múltipla escolha e complementa o conjunto de dados do ENEM de 2022. Durante a etapa de confirmação do local de prova, o estudante que optasse pela participação na pesquisa poderia preencher o questionário.

O questionário tinha como objetivo principal fazer um levantamento de informações não-cognitivas sobre a rotina de estudos e estratégias de preparação para o ENEM adotadas pelos estudantes durante o segundo ano de pandemia, considerando que as instituições de ensino estavam realizando as atividades educacionais, por meio das modalidades de ensino presencial, híbrida ou totalmente remota. Os resultados objetivam oferecer subsídios para políticas públicas de acesso e permanência na Educação Básica, diante dos problemas ocorridos nos processos educacionais devido à pandemia (INEP, 2022).

Durante o processo de elaboração desta pesquisa, em agosto de 2023, o INEP (2023b) publicou o "Painel da Pesquisa Enem 2022 – Hábitos de Estudo na Pandemia"[6]. Trata-se de um painel para visualização dos dados de 2022, que foi organizado a partir dos Grupos Temáticos (GT), sendo esses: monitoramento da pesquisa; matrícula escolar; gestão do tempo e planejamento de estudos; práticas de estudo e pesquisa; tecnologias e tipo de acesso/problemas na rotina; dificuldades de infraestrutura/ajuda de terceiros. Para cada GT há uma aba que contém os elementos dos dados da pesquisa para visualização que são representados em forma de gráficos e tabelas (INEP, 2023b).

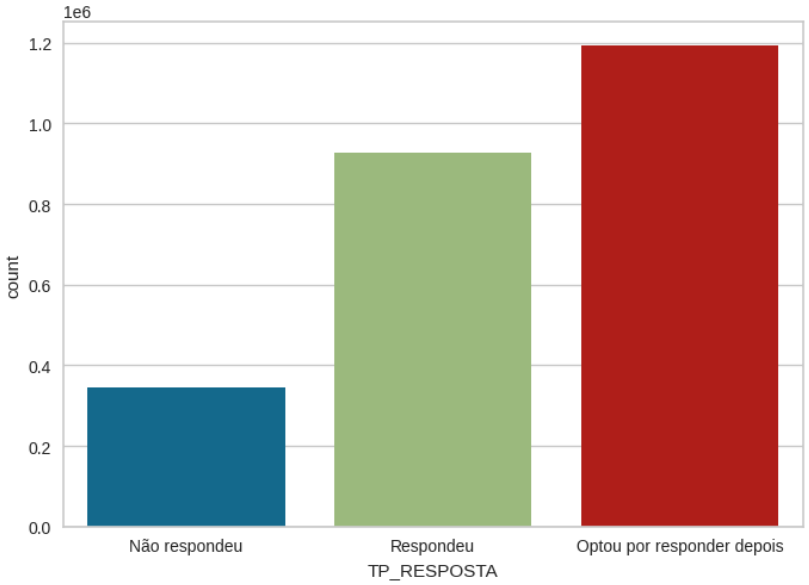
O Painel permite que as informações quantitativas possam ser analisadas de forma combinada a partir do cruzamento entre as variáveis correspondentes aos dados de HE e algumas variáveis dos microdados do ENEM, sendo essas: sexo, cor/raça, faixa etária e região. As

[3] Disponível em: planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Acesso em: 08 jul. 2023.
[4] Todas as edições dos microdados do ENEM estão disponíveis em: gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem. Acesso em: 18 jun. 2023.
[5] Os microdados referentes ao questionário de HE de 2022 está no arquivo compactado, disponível em: https://download.inep.gov.br/microdados/microdados_enem_2022.zip. Acesso em: 27 jun. 2023.
[6] Disponível em: gov.br/inep/pt-br/assuntos/noticias/enem/painel-apresenta-pesquisa-sobre-estudo-na-pandemia. Acesso em: 17 out. 2023.

visualizações geradas contribuem para reflexões sobre a situação educacional durante o segundo ano da pandemia (INEP, 2023b).

A Figura 2.2 mostra que, dentre o total de 2.467.086 estudantes que consultaram o local de prova, 928.564 (37,6%) responderam ao menos uma entre as perguntas do questionário de HE.

Figura 2.2 - Número de estudantes por tipo de resposta ao questionário de HE.



Fonte: Elaborado pelo autor, adaptado de (INEP, 2023b).

Nos microdados de HE há estudantes que responderam entre uma e até todas as questões. Logo, esses estudantes são considerados respondentes da pesquisa, ainda que tenham respondido parcialmente o questionário de HE.

2.3. MINERAÇÃO DE DADOS

Fayyad, Piatetsky-Shapiro e Smyth (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 AM podem ser utilizados e executados com eficiência para identificar padrões, e como resultados podem ser interpretados.

O processo de KDD representa uma visão abrangente para um conjunto de etapas não trivial de identificação de padrões, a partir de dados, de modo que esses sejam válidos, novos, úteis e compreensíveis (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). Os padrões devem ser novos, compreensíveis e úteis para usuários que desejam realizar tomadas de decisões em prol de alguma meta.

Fayyad, Piatetsky-Shapiro e Smyth (1996) definem dois objetivos pelo que se deseja obter de retorno (conhecimento) a partir da aplicação do processo de KDD: (i) verificação e (ii) descoberta. Com a verificação busca-se verificar hipóteses definidas por usuários. Por exemplo, um especialista do domínio pode confirmar ou refutar hipóteses por meio da averiguação de

padrões pré-estabelecidos nos dados. Com a descoberta, busca-se encontrar novos padrões nos dados de forma automática, de modo que tais padrões possam induzir subsídios para tomada de decisão.

O objetivo da descoberta de conhecimento pode ser subdividido ainda nas tarefas de aprendizado preditiva ou descritiva (ALFRED, 2005). Na tarefa preditiva há um atributo específico do conjunto de dados, denominado atributo ou objeto alvo. Dado um número de instâncias em um conjunto de dados, para as quais o valor do atributo alvo é conhecido, a tarefa é gerar um modelo preditivo que seja capaz de predizer valores (nominais ou numéricos) para novas instâncias. Na tarefa descritiva, não existe um atributo alvo conhecido, o objetivo é encontrar padrões que descrevem os dados para apresentação a um usuário de forma compreensível. Tal modelo pode ser expresso por meio de associações, grupos, ou dependências probabilísticas entre os dados (ALFRED, 2005).

As tarefas de AM são essenciais no processo de MD para construção de aplicações (ALFRED, 2005). Além do AM, outras atividades e técnicas associadas à preparação dos dados são importantes como, por exemplo, limpeza, normalização e imputação de valores. Grande parte do esforço na construção de um pipeline de AM é gasto na preparação e limpeza de dados (ZHENG; CASARI, 2018).

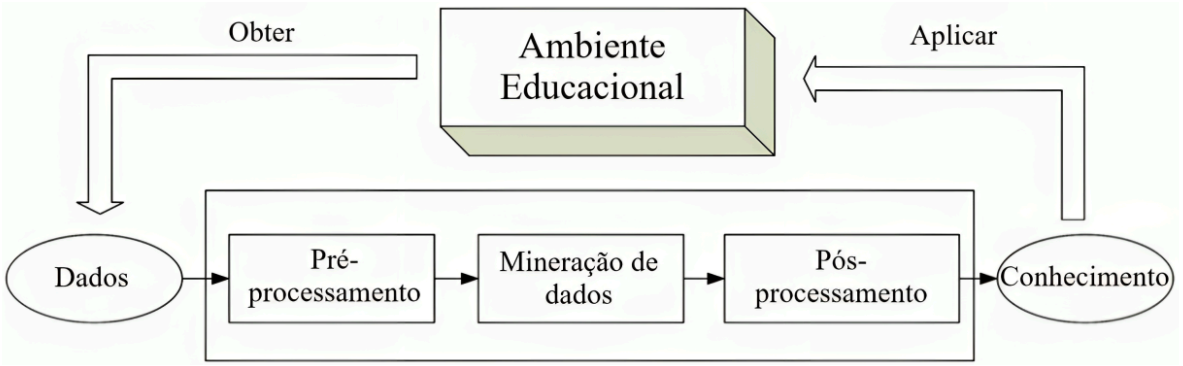
2.3.1. Mineração de dados educacionais

Ambientes e tecnologias para o ensino utilizadas em salas de aula presenciais tradicionais e em plataformas de aprendizagem online são conhecidas como *e-learning,* ou educação baseada na *web* (RODRIGUES; ISOTANI; ZÁRATE, 2018). O aumento crescente do *e-learning* favorece a geração de grandes repositórios ou conjuntos de dados, podendo ser importantes fontes de informação para apoio à decisão e melhoria do processo de ensino-aprendizagem. Como ilustração, as melhorias podem ser obtidas por meio da análise dos dados dos estudantes sobre fatores como comportamento, satisfação e desempenho (KOEDINGER et al., 2008).

Com o crescimento do *e-learning* e da educação flexível (aulas remotas e/ou híbridas) durante a pandemia, informações sobre os estudantes têm se tornado cada vez mais importantes. Novos dados têm sido gerados a respeito da situação dos estudantes, como, por exemplo, na geração de dados sobre hábitos de estudo pelo INEP. Quando os dados são provenientes da área educacional, a exemplo dos dados do ENEM e de hábitos de estudo, os autores têm considerado a subárea intitulada Mineração de Dados Educacionais (MDE) (RAMOS et al., 2020; ROMERO et al., 2014; BAKER, 2010).

De acordo com García et al. (2011), o processo de MDE é constituído por três etapas principais (Figura 2.3):

Figura 2.3 - Processo geral de Mineração de Dados Educacionais.



Fonte: Adaptado de GARCÍA et al. (2011).

No pré-processamento os dados são obtidos do Ambiente Educacional e devem primeiramente ser pré-processados para serem transformados em um formato adequado para a tarefa de MD. Na etapa central são aplicadas as tarefas de mineração de dados com os dados previamente pré-processados. São exemplos dessas tarefas: regressão, classificação, agrupamento, mineração de regras de associação, mineração de padrões sequenciais, mineração de texto, entre outras. No pós-processamento os resultados ou modelo obtidos são interpretados e usados para tomar decisões e aplicá-las sobre o Ambiente Educacional.

2.3.2. MODELO CRISP-EDM

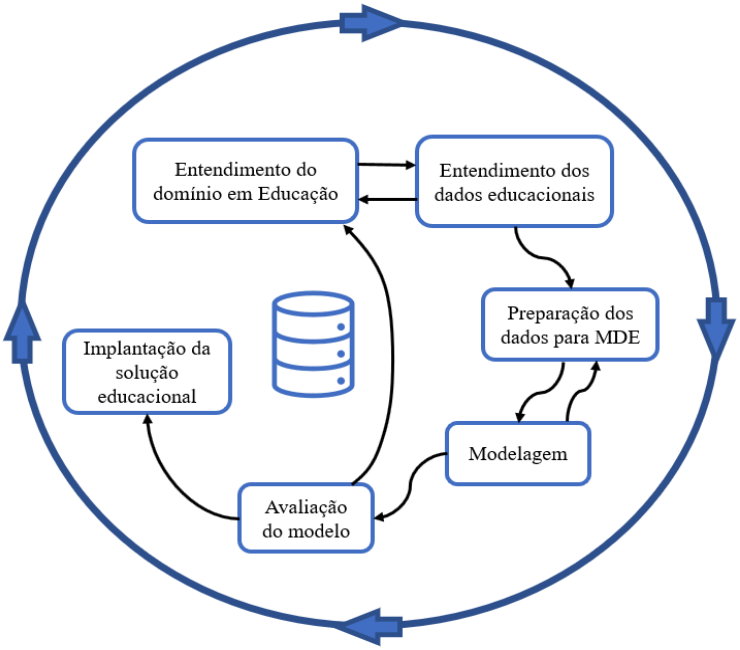
Criado em 1996, o modelo de processo de desenvolvimento intitulado *Cross Industry Standard Process for Data Mining* (CRISP-DM)[7] consiste em uma metodologia (conjunto de fases) para se executar um projeto de descoberta de conhecimento (CHAPMAN et al., 2000). Segundo a IBM *Documentation* (2021), o processo CRISP-DM permite a organização e orientação durante a condução de um processo de MD, auxiliando na descrição e relações entre as tarefas definidas para cada etapa. 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.

Quando esses processos de MD estão associados a contextos educacionais, o processo CRISP-DM pode ser adaptado à MDE, sendo essa área muito usada para fornecer *feedback* a gestores e professores no que diz respeito a tomadas de decisões estratégicas mais assertivas. Considerando que a MDE é uma extensão da MD, os autores RAMOS et al. (2020) propuseram uma extensão do CRISP-DM, chamado de CRISP-EDM.

A Figura 2.4 representa as etapas definidas para o modelo CRISP-EDM.

[7] Informações adicionais disponíveis em: https://ibm.com/docs/en/spss-modeler/saas?topic=dm-crisp-help-overview. Acesso em: 10 jul. 2013.

Figura 2.4 - Modelo do CRISP-EDM.



Fonte: Adaptado de RAMOS et al. (2020).

Ramos et al. (2020) definem cada etapa da seguinte forma:

- **Entendimento do domínio em Educação**: nessa primeira fase busca-se o entendimento completo do problema de MD a ser investigado e aplicado no campo educacional, a fim de descobrir fatores importantes que possam interferir no resultado que se pretende alcançar com a MD. Um maior esforço empregado nessa etapa de conhecimento do domínio da aplicação implica na possibilidade de alcançar os objetivos do projeto com maior precisão.
- **Entendimento dos dados educacionais**: a segunda etapa compreende a análise dos dados educacionais que podem ter origem em fontes como, por exemplo, um AVA ou dados gerados por um censo escolar. A análise nesta etapa visa a compreensão dos dados dentro do domínio onde eles se encontram.
- **Preparação dos dados para MDE**: esta etapa envolve toda a preparação do(s) conjunto(s) de dados a ser(em) minerado(s). Envolve todo o pré-processamento dos dados, onde são realizadas atividades de limpeza, transformações, discretizações, integração e outros ajustes necessários ao conjunto de dados.
- **Modelagem**: refere-se à etapa onde são definidas e aplicadas técnicas de modelagem de AM. Comumente, a execução dos algoritmos de AM é repetida várias vezes até que se obtenha o melhor ajuste dos parâmetros dos algoritmos, a fim de refinar os resultados finais ou mesmo comparar os algoritmos usados para a técnica de MD em estudo.
- **Avaliação dos modelos**: após a elaboração dos modelos de AM é preciso avaliar seu comportamento. Para isso, a avaliação da qualidade do modelo de AM faz-se necessário antes da implantação em um ambiente de produção educacional. Métricas de avaliação podem ser usadas para avaliar a qualidade do modelo de AM.

- **Implantação da solução educacional**: a implantação do modelo ocorre, por exemplo, a partir do uso de *plug-ins* ou módulos de sistemas que realizam as tarefas de mineração na plataforma de dados educacionais. Os resultados da aplicação dos modelos nos dados podem ser usados pelos atores envolvidos no contexto educacional em tomadas de decisão que resultem em intervenções para melhoria do processo educacional. Os resultados do modelo de AM podem ser apresentados de maneira simplificada, geralmente por meio de relatórios ou visualizações (*dashboards*). A avaliação da solução educacional por diversos usuários integrantes do processo educacional implica em uma validação da implantação nessa última etapa.

A metodologia de desenvolvimento da abordagem proposta tem como base as etapas definidas pelo processo CRISP-EDM. O Capítulo 4 contém o detalhamento da adaptação desse processo na abordagem definida.

2.4. APRENDIZADO DE MÁQUINA

Devido à complexidade de problemas a serem tratados computacionalmente e também ao aumento do volume de dados oriundos de diferentes setores, tornou-se clara a necessidade de melhoria das técnicas e das ferramentas computacionais, de maneira que essas possam auxiliar em tomadas de decisões por humanos, a partir de experiência passada, hipótese ou função, capaz de resolver o problema que se deseja tratar (FACELI et al., 2011).

Um exemplo de descoberta de uma hipótese está na identificação de atributos relevantes para predição de evasão através de dados educacionais utilizando, para isso, dados de estudantes registrados na base de dados de uma instituição acadêmica. Logo, com o AM é possível extrair informações de modo automático por meio de dados de entrada.

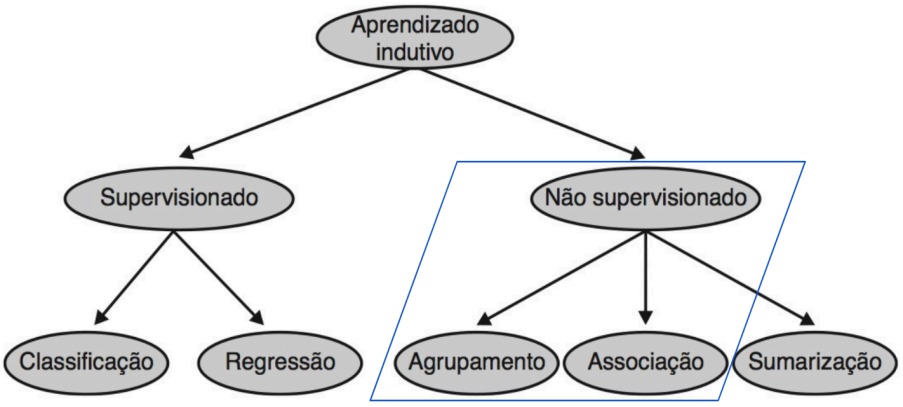
Segundo Monard e Baranauskas (2003) o método indutivo é a forma de inferência lógica que permite generalizar conclusões e aprender a partir de um conjunto particular de exemplos. Na indução de hipóteses geradas pode-se haver a preservação ou não da verdade.

A esse processo de indução de uma hipótese a partir da experiência passada, dá-se o nome Aprendizado de Máquina (FACELI et al., 2011). Mitchell (1997) define o AM como a capacidade de melhorar o desempenho na realização de alguma tarefa por meio da experiência.

Segundo Alpaydin (2010), o AM pode ser categorizado, de modo geral, em: supervisionado e não-supervisionado. Para qualquer um desses tipos de AM, o objetivo é ensinar a máquina (algoritmo de AM) a processar os dados e produzir algum tipo de decisão ou resultado baseado nos próprios dados (SRIVASTAVA, JOSHI e GAUR, 2014) .

A Figura 2.5 apresenta uma taxonomia de tipos de AM, de acordo com Faceli et al. (2011). No topo aparece o aprendizado indutivo, processo pelo qual são realizadas as generalizações a partir dos dados. Em seguida, o aprendizado é dividido em dois tipos: o supervisionado, voltado para tarefas preditivas, e o não supervisionado, para tarefas descritivas.

Figura 2.5 - Taxonomia dos tipos de AM.



Fonte: FACELI et al. (2011, p. 6).

As tarefas supervisionadas (preditivas) se distinguem pelo tipo dos rótulos dos dados: discreto, no caso de classificação; e contínuo, no caso de regressão. As tarefas não supervisionadas (descritivas) são, de modo geral, divididas em: agrupamento, em que os dados são agrupados sob o critério de similaridade; regras de associação, que consiste em encontrar padrões frequentes de associações entre os itens de um conjunto de dados; e sumarização, que objetiva encontrar uma descrição simples e compacta de conjunto de dados (FACELI et al., 2011). O destaque na Figura 2.5 assinala as tarefas não supervisionadas consideradas neste trabalho.

2.4.1. AM supervisionado

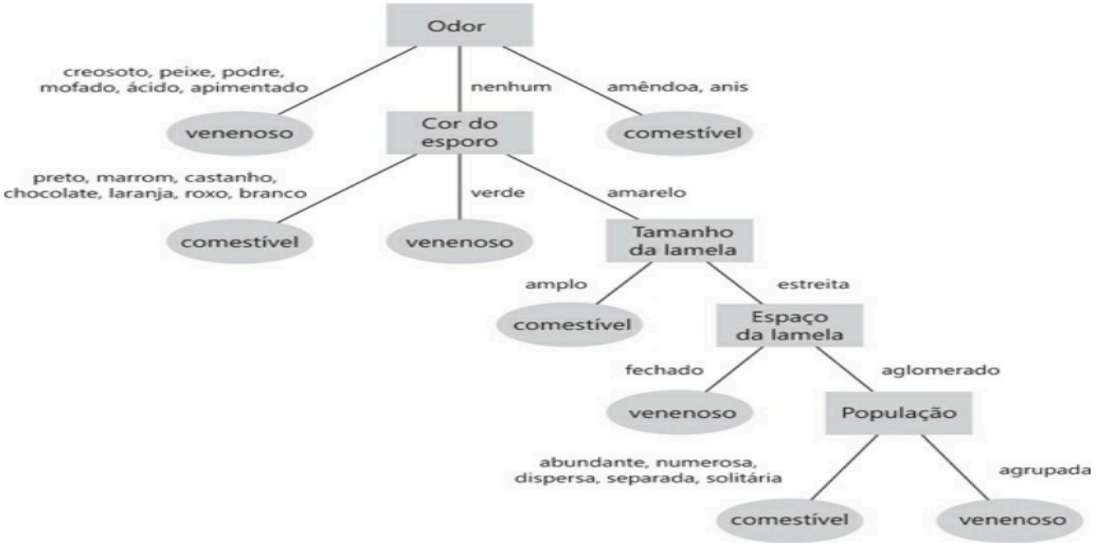
O AM supervisionado tem como base a geração de um modelo que aprende a partir de um conjunto de instâncias rotuladas, por meio de uma variável alvo, que permite classificar um conjunto de novas instâncias. Neste sentido, a classificação e a regressão são tarefas de predição no contexto de AM supervisionado (KELLEHER; MAC NAMEE; D'ARCY, 2020). Para aplicação dessas tarefas são usados algoritmos de AM supervisionados como, por exemplo, *k-nearest neighbors* (KNN), regressão logística, *naive bayes*, árvore de decisão, *ensembles* ou redes neurais artificiais (HAN; KAMBER; PEI, 2012).

Tomando como exemplo, a estratégia por *decision tree induction* (indução de árvore de decisão) possibilita que algoritmos de árvore de decisão construam uma estrutura hierárquica de nós. Cada nó interno (não-folha) denota uma condição em um atributo e cada ramificação (folha) refere-se a um resultado da condição, ao passo que cada nó folha indica uma predição de classe. Em cada nó, o algoritmo escolhe o "melhor" atributo para particionar os dados em classes individuais (HAN; KAMBER; PEI, 2012).

A Figura 2.6 exemplifica uma árvore de decisão para classificar cogumelos em comestíveis ou venenosos (CASTRO; FERRARI, 2016). Considerando a classificação de uma primeira instância de cogumelo como exemplo, de acordo com a árvore, o atributo classificado como mais relevante foi 'Odor', que possui o valor 'podre', portanto, essa instância de cogumelo é classificada como 'venenoso'. Considerando uma segunda instância como exemplo, o atributo 'Odor' possui valor 'nenhum', portanto, é necessário caminhar na árvore para analisar os valores

do próximo atributo chave. Para o atributo ‘Cor do esporo’, a instância tem valor marrom, o que leva à instância de cogumelo a ser classificada como comestível, que é a classe correta.

Figura 2.6 - Exemplo de uma árvore de decisão



Fonte: CASTRO e FERRARI (2016).

Na regressão ocorre a estimação de valores contínuos a partir de um conjunto de dados históricos como entrada (ELMASRI e NAVATHE, 2019). Um exemplo de aplicação dessa tarefa supervisionada está em problemas de indicadores econômicos ou de mercado financeiro, onde tenta-se prever valores (numéricos) estimados por meio de dados históricos de um conjunto de dados. Outro exemplo de aplicação de regressão está na estimação da quantidade de produtos que deve ser estocado para determinado período, ou na predição do número de clientes que visitará uma loja em uma ou mais datas específicas (SILVA, PERES e BOSCARIOLI, 2020).

2.4.2. AM não supervisionado

No aprendizado não supervisionado, não há rótulo ou valor de destino definido para os dados, estando incluídas neste grupo as tarefas de agrupamento e regras de associação (HARRINGTON, 2012).

O Agrupamento, ou *Clustering* é um tipo de aprendizado não-supervisionado em que se formam grupos de instâncias similares (HARRINGTON, 2012). As estratégias de resolução da tarefa de agrupamento podem ser encontradas de forma variada na literatura de AM sendo que, de modo geral, são divididos em (SILVA, PERES e BOSCARIOLI, 2020): hierárquicos, partitivos ou baseados em densidade.

A estratégia hierárquica considera a criação de uma decomposição hierárquica dos dados, podendo ainda ser subdivididos nas abordagens divisiva ou aglomerativa. A abordagem divisiva é do tipo *top-down*, onde o processo de divisão se inicia colocando todas as instâncias de um conjunto de dados em um único grupo e, iterativamente, são gerados grupos menores, até que cada instância constitua um único grupo. Por outro lado, a abordagem aglomerativa é do tipo

bottom-up, onde, no início do processo, cada instância faz parte de um grupo separado e, iterativamente, aglomera grupos similares até formar um único grupo com todos as instâncias. Os algoritmos clássicos AGNES (*AGlomerative NESting*) e DIANA (*DIvisive ANAlysis*), por exemplo, usam essa estratégia.

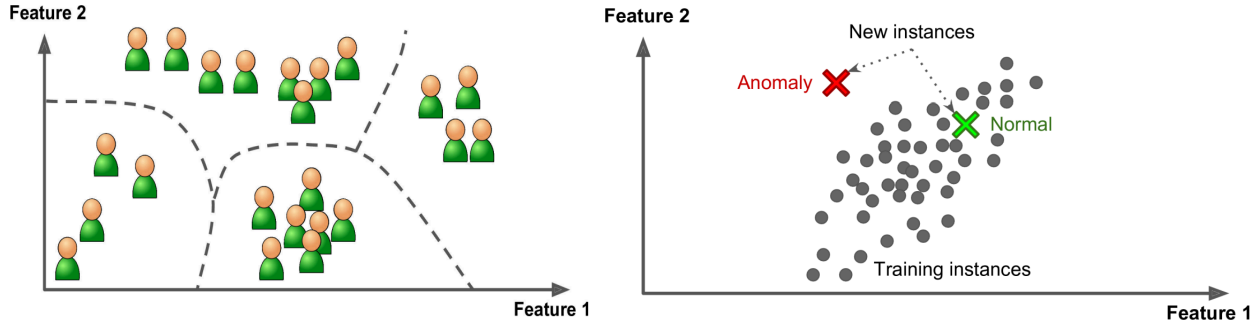
Na estratégia partitiva são gerados grupos (partições) de instâncias similares de acordo com um critério de particionamento. Iterativamente, as instâncias são realocadas nos grupos, de maneira que se ajustem melhor ao objetivo de maximização de similaridade intragrupo. Alguns algoritmos se baseiam em centróides, instâncias centradas em pontos específicos de cada grupo, como ocorre com o K-means (KAUFMAN, 1990) e o K-medoids (MACQUEEN et al.,1967).

Por fim, na estratégia baseada em densidade, grupos com uma ou poucas instâncias são inicialmente formados e, iterativamente, recebem mais instâncias (localizadas na vizinhança do grupo) e crescem até que um limiar seja atingido. Algoritmos com esse tipo de estratégia assumem que um grupo é uma região densa formada por instâncias similares, sendo que uma das principais vantagens é a possibilidade de identificação de outliers. O algoritmo DBSCAN (*Density Based Spatial Clustering of Applications with Noise*) (ESTER et al., 1996) é um exemplo desse tipo de estratégia.

A Rede Neural Artificial Mapas Auto Organizáveis, também conhecida como SOM (*Self Organizing Maps*) é um algoritmo conexionista, muito aplicado na visualização de conjuntos de dados de alta dimensão, já que é capaz de executar projeções dos dados em espaços de baixa dimensão (SILVA, PERES e BOSCARIOLI, 2020).

A Figura 2.7 (GÉRON, 2019) representa um exemplo de um administrador de um blog que busca detectar grupos de visitantes com perfis semelhantes. Neste caso, um algoritmo de agrupamento poderia indicar que existem quatro perfis de visitantes, sendo que 40% dos seus visitantes do sexo masculino preferem ler histórias em quadrinhos durante a noite, e que entre esses, 20% são jovens amantes de ficção científica que visitam o blog em finais de semana com maior frequência (Figura 2.7(a)). Um visitante, que está frequentemente conectado ao seu blog realizando muitas solicitações, poderia ser identificado como um potencial ataque (anomalia) ao blog (Figura 2.7(b)). Os grupos podem ser avaliados de diferentes formas por meio das características (*features*) do conjunto de dados.

Figura 2.7 - Representação simplificada de agrupamentos.



(a) (b)

Fonte: Adaptado de GÉRON (2019).

As organizações, em geral, buscam obter lucro e promover satisfação de clientes por meio de seus produtos e serviços, para isso é preciso que se tenha uma ideia do perfil e comportamento de compra dos clientes. Esse conhecimento pode ser extraído a partir dos dados e aplicado, por exemplo, em divulgação de promoções, ajuste de preços e gerenciamento de estoque. Essa descoberta de conhecimento em grandes conjuntos de dados é conhecida como análise de associação ou mineração de regras de associação (HARRINGTON, 2012). O objetivo é extrair as regras de conjuntos de dados, onde cada exemplo consiste em um conjunto de itens, por meio do poder computacional em um período razoável de tempo (AGRAWAL; IMIELIŃSKI; SWAMI, 1993).

Os algoritmos Apriori (AGRAWAL et al., 1996), Frequent Pattern-growth (FP-growth) (HAN; PEI; YIN, 2000) e Equivalence CLAss Transformation (ECLAT) (ZAKI, 2000) são exemplos de algoritmos usados para mineração de regras de associação.

O Apriori é um dos algoritmos mais aplicados em tarefas de regras de associação, sendo o tempo de execução uma desvantagem em relação ao FP-growth, que faz uso de uma árvore de busca para realizar uma otimização. Por fim, o ECLAT realiza uma combinação de busca em profundidade com “intersecções rápidas” entre os itens do conjunto de dados (GRAHNE; ZHU, 2003).

As subseções seguintes detalham e exemplificam algoritmos de AM não supervisionado, bem como as medidas de avaliação que foram usadas neste trabalho.

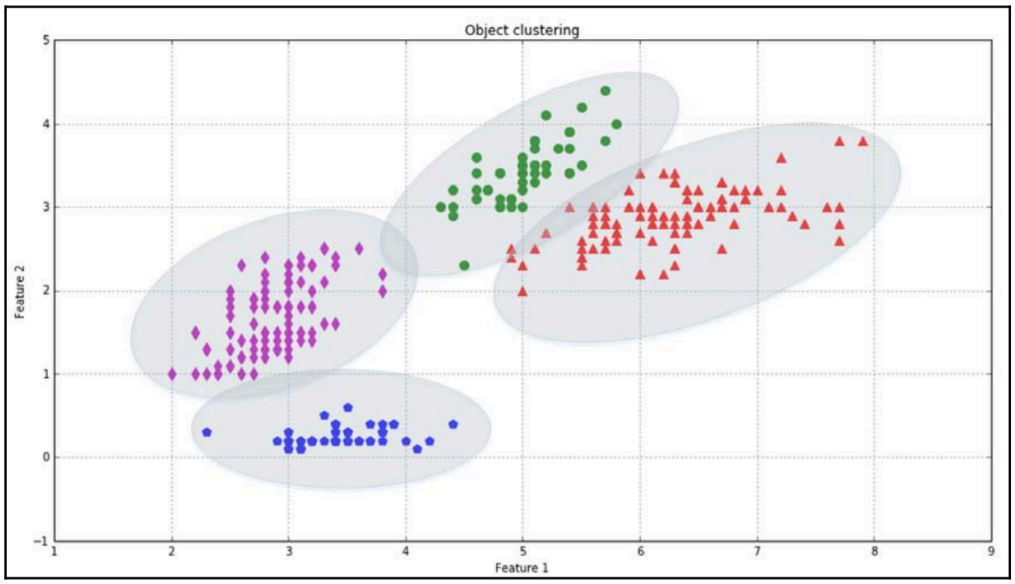
2.4.3. Agrupamento

O objetivo do agrupamento é identificar grupos (*clusters*) de instâncias que sejam semelhantes. Para isso, tipicamente é usada uma medida de distância para determinar o quão próximo (similares) são as instâncias que formam os grupos. Uma vez definidos os clusters, novas instâncias podem ser incluídas no grupo de maior similaridade (ROMERO; VENTURA, 2013).

Os algoritmos de agrupamento atribuem (ou prevêem) um valor numérico a cada uma das instâncias, que define o grupo a qual pertence. A tarefa de agrupamento pode ser usada, por exemplo, na identificação de grupos de estudantes com base no desempenho, estilos de aprendizagem e comportamento (DUTT; ISMAIL; HERAWAN, 2017).

A Figura 2.8 ilustra a disposição de objetos (instâncias) em um espaço de dados hipotético (BONACCORSO, 2018).

Figura 2.8 - Exemplo de instâncias agrupadas em quatro clusters.



Fonte: BONACCORSO (2018).

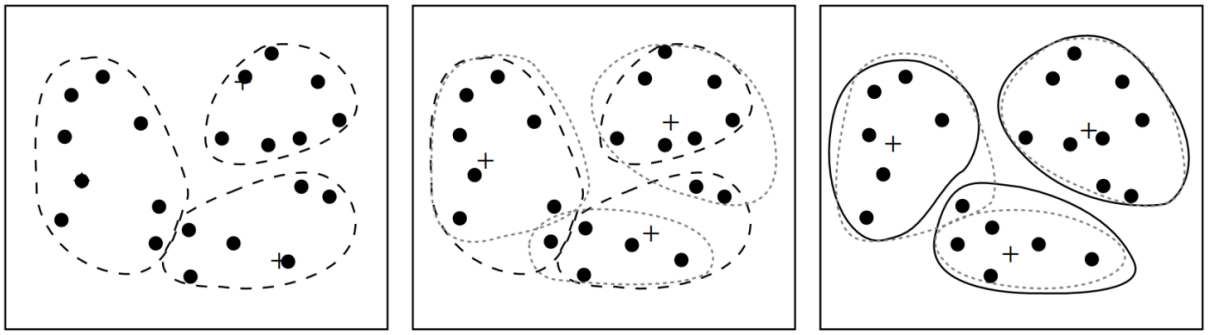
As elipses representam os clusters e delimitam as áreas compostas pelos objetos (instâncias), estando cada um em um único grupo. Para os pontos de fronteira (triângulos sobrepostos em mais de um grupo) deve ser definido um critério específico como, por exemplo, a medida de distância, para determinar o cluster adequado. Outra característica que deve ser considerada na tarefa de agrupamento é a presença de *outliers* (pontos isolados, mais distantes dos demais). Na Figura 2.8, todos os triângulos limites estão próximos uns dos outros, então o vizinho mais próximo é outro triângulo. No entanto, a depender do domínio de aplicação, alguns pontos têm um alto grau de incerteza (ruído) devido à natureza de seus valores (BONACCORSO, 2018).

2.4.3.1. K-means

O K-means (MACQUEEN et al.,1967) é um algoritmo que determina um número k de clusters para um determinado conjunto de dados, sendo o valor de k definido obrigatoriamente pelo usuário. Cada cluster é definido por meio de pontos conhecidos como centróides. Cada centróide está no centro de todos os pontos de um dado cluster (HARRINGTON, 2012).

A Figura 2.9 ilustra uma representação resumida da formação de clusters pelo K-means (HAN; KAMBER; PEI, 2012).

Figura 2.9 - Exemplo de formação de clusters com o K-means.



(a) Agrupamento inicial (b) Iteração (c) Agrupamento final

Fonte: HAN; KAMBER; PEI (2012).

O K-means define aleatoriamente os k centróides iniciais (+) dos clusters (Figura 2.9(a)). Cada um dos objetos representa inicialmente uma média ou centro do cluster. Para cada um dos objetos restantes, um objeto é atribuído ao cluster ao qual é mais semelhante, com base na distância, normalmente euclidiana, entre o objeto e a média do cluster. O algoritmo K-means melhora iterativamente a variação dentro do cluster. Para cada cluster, é calculada a nova média usando os objetos atribuídos ao cluster na iteração anterior. Todos os objetos são (re)atribuídos usando as médias atualizadas como os novos centros de cluster (Figura 2.9(b)). As iterações continuam até que a atribuição fique estável, ou seja, os clusters formados na rodada atual sejam os mesmos formados na rodada anterior (Figura 2.9(c)).

A Figura 2.10 mostra o funcionamento do K-means (CASTRO; FERRARI, 2016).

Figura 2.10 - Algoritmo K-means.

procedure k-Means(I,k,AG)
Input: I = {I_1, I_2, ..., I_N} *%conjunto com N instâncias de dados a serem agrupadas*
k *% número de grupos a serem criados*
Output: AG = {G_1,G_2,...G_k} *%agrupamento formado por k grupos induzidos a partir de I*
begin
% Inicialização
% no passo (1) cada grupo é definido apenas pelo centroide
(1) escolha arbitraria de k instâncias do conjunto I, como centroides dos grupos G_1,G_2,...G_k

% Indução do agrupamento AG
(2) **repeat**
(3) (re)atribuir cada instância $I_i \in I$ (i=1, ..., N) ao grupo cujo centroide que lhe seja mais próximo;
(4) atualizar os centroides de cada grupo, como a média dos valores das suas instâncias
(5) **until** nenhuma alteração aconteça.
end.
return AG = {G_1,G_2,...G_k}
end_procedure

Fonte: MATTE; NICOLETTI (2019).

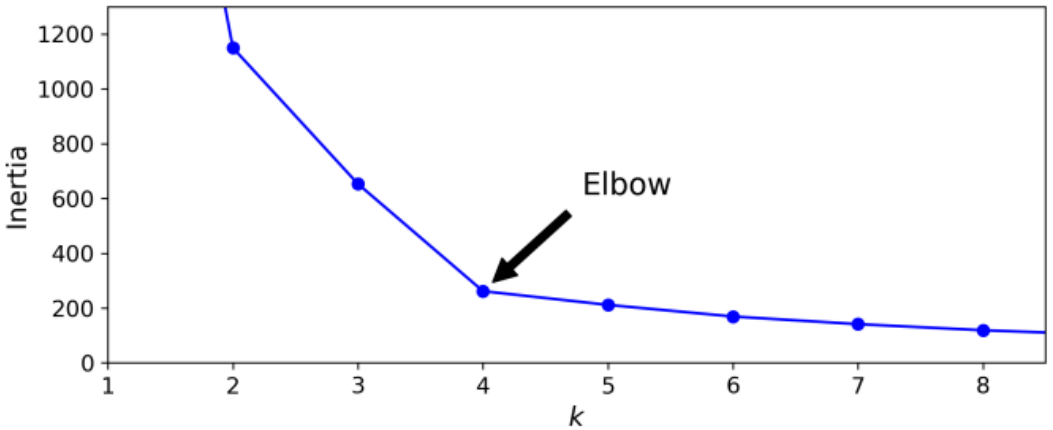
Entre as vantagens do agrupamento com o K-means estão a capacidade de ser relativamente escalável e rápido porque não realiza muitos cálculos. Entre as desvantagens ele pode convergir para mínimos locais, tornar-se muito lento para conjuntos de dados de alta

dimensionalidade, apresentar problemas quando os dados contêm ruídos; e prover baixa eficiência para representar dados em formas visuais não-esféricas (HAN; KAMBER; PEI, 2012; HARRINGTON, 2012). Vale ressaltar que extensões e implementações mais recentes baseadas no K-Means buscam avançar na melhoria dessas desvantagens (EZUGWU, 2022). Por exemplo, tem-se o algoritmo G-means (HAMERLY; ELKAN, 2003) que realiza um teste de razão de verossimilhança para decidir se um grupo deve ser dividido ou mesclado com outro grupo. Essa abordagem é mais robusta quanto ao tratamento de grupos não-esféricos e ruídos. O algoritmo X-means (PELLEG, 2000) realiza decisões locais sobre quais grupos os centróides atuais devem se dividir para obter um melhor ajuste, após cada iteração do K-means, por meio do cálculo de Critério de Informação Bayesiano (BIC).

Conforme mostrado anteriormente, o K-means requer que seja especificado o número k de clusters. Muitas vezes esse valor pode ser identificado conforme as regras de negócio por um especialista do domínio. O especialista pode definir que deseja avaliar, por exemplo, 3, 4, ou 5 grupos de clientes em uma concessionária de automóveis com base na lógica da análise do negócio a ser realizada. Porém, na ausência de um especialista ou de definições do próprio contexto do negócio ou do problema em questão, uma abordagem estatística comumente usada na identificação do k ideal é chamada de método *elbow* (cotovelo). O objetivo desse método é identificar qual o conjunto de clusters explica "a maior parte" da variação nos dados. O cotovelo é o ponto onde a variância cumulativa se estabiliza após um aumento acentuado de grupos, daí o nome do método (BRUCE; BRUCE; GEDECK, 2020).

O método baseia-se na suposição de que um número apropriado de clusters deve produzir uma pequena inércia. Porém, este valor atinge seu mínimo (0,0) quando o número de clusters é igual ao número de amostras; portanto, não se pode procurar o mínimo, mas sim um valor que é uma compensação entre a inércia e o número de clusters (BONACCORSO, 2018). A Figura 2.11 ilustra um exemplo para um conjunto de dados, com recorte de intervalo de k clusters especificado de 2 a 8, onde são calculadas e coletadas as inércias.

Figura 2.11 - Exemplo do método elbow.



Fonte: GÉRON (2019).

A maior redução ocorre entre os números de clusters 2 e 4, em seguida, a inclinação começa a se achatar, sem que haja muita variação. O objetivo então é avaliar qual o valor que, se reduzido, leve a um grande aumento inercial (distorção) e, se aumentado, produza uma redução inercial muito pequena. Logo, a escolha mais indicada seria 4, enquanto valores maiores provavelmente produzirão divisões intracluster indesejadas, até a situação extrema em que cada ponto se torna um único cluster (BONACCORSO, 2018).

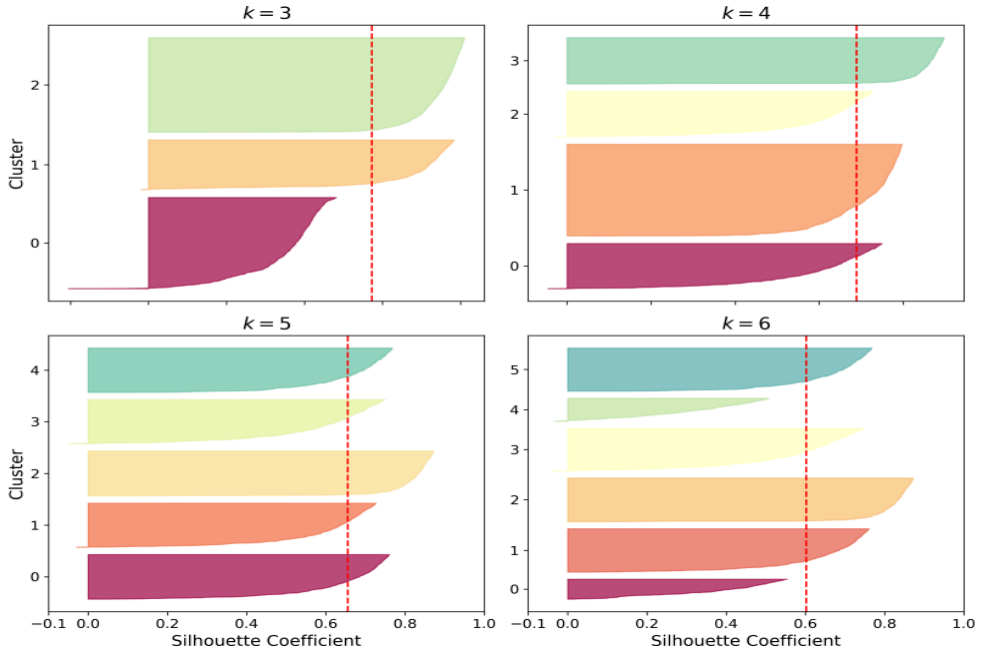
O método elbow é muito simples e pode ser empregado como primeira abordagem para determinar uma faixa de k ideal. De maneira complementar, as próximas estratégias mostradas são mais complexas e podem ser utilizadas para confirmar ou indicar melhor o número ideal de clusters.

O coeficiente da silhueta (ROUSSEEUW, 1987) é uma medida de avaliação de clusters que tem o objetivo de mostrar o fator de pertinência das instâncias aos grupos, informando quão bem agrupados elas estão (CASTRO; FERRARI, 2016).

O cálculo é feito por meio da distância média intracluster e da distância média do cluster mais próximo para cada amostra. O coeficiente de silhueta pode variar entre -1 e +1: um coeficiente próximo de +1 significa que a instância está bem definida dentro de seu próprio cluster e longe de outros clusters, enquanto um coeficiente próximo de 0 significa que está perto de um limite de cluster, e finalmente um coeficiente próximo de -1 significa que a instância pode ter sido atribuída ao cluster errado ou que estejam sobrepostos (GÉRON, 2019). Ao fornecer todas as instâncias do conjunto de dados e os rótulos que foram atribuídos, a função retorna o coeficiente de silhueta médio de todas as amostras.

A Figura 2.12 mostra uma comparação de diagramas de silhueta para quatro valores possíveis de k clusters.

Figura 2.12 - Exemplo de comparação de coeficientes de silhueta.



Fonte: GÉRON (2019).

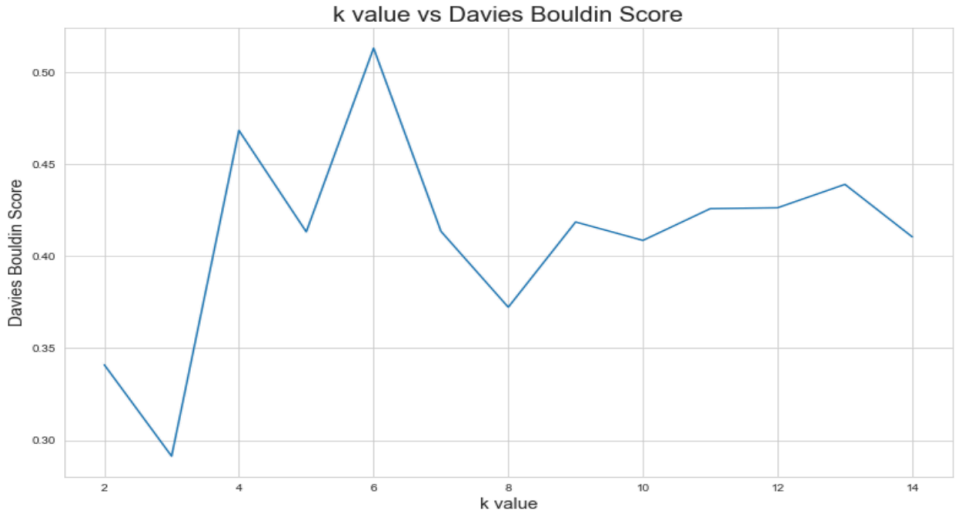
Observa-se que este tipo de medida de avaliação é muito rica em detalhes. Percebe-se que k=4 seria a melhor escolha, tendo em vista que apresenta o valor médio mais próximo a 1. Avaliando de maneira comparativa, k=5 também é um bom número e é melhor que k=6 e 7. Tal comparação não é visivelmente simples ao comparar as inércias obtidas com o método elbow.

As linhas tracejadas em vermelho na vertical representam a pontuação média da silhueta para cada número de clusters. Quando a maioria das instâncias em um cluster tem um coeficiente inferior a esta pontuação (ou seja, se muitas das instâncias param antes da linha tracejada, terminando à esquerda dela), então é dito que o cluster é mal definido, visto que as instâncias estão muito próximas de outros clusters, com é o caso da formação de clusters com k=3 e k=6. Quando k=4 ou k=5, os clusters parecem melhor definidos, tendo em vista que a maioria das instâncias se estende além da linha tracejada, para a direita e mais perto de 1,0. Caso houvesse a necessidade de se obter clusters de tamanhos semelhantes, a melhor escolha seria usar k=5 (GÉRON, 2019).

Outras medidas de avaliação de clusters complementares podem ser aplicadas para reduzir o grau de incerteza na definição de k ideal e para avaliar a qualidade dos clusters. O Índice Davies-Bouldin (DBI) é uma medida de avaliação muito aplicada em tarefas de agrupamento (DAVIES; BOULDIN, 1979). Para definir o índice é necessário antes definir a medida de dispersão e de similaridade do cluster (SHEIKH; GHANBARPOUR; GHOLAMIANGONABADI, 2019). O valor do índice deve ser minimizado podendo assumir valores no intervalo de 0 ao infinito. A maior proximidade do valor 0 (zero) indica agrupamentos mais bem particionados (CASTRO; FERRARI, 2016).

A Figura 2.13 ilustra um exemplo da distribuição de k clusters segundo o DBI. Como pode ser observado o ponto k=3 tem a pontuação mais baixa do DBI, o que reflete que existem três clusters ideais que podem ser identificados a partir deste exemplo de conjunto de dados.

Figura 2.13 - Exemplo da distribuição de k clusters pelo DBI.

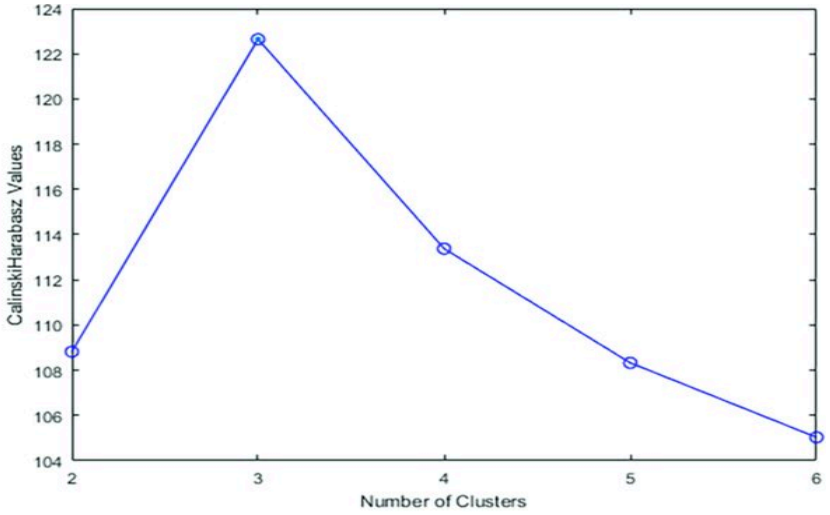


Fonte: AOUEDI (2021)

O Índice Calinski and Harabasz (ICH) - também conhecido como Critério de Razão de Variância - pode ser usado para avaliar o modelo de agrupamento, onde uma pontuação maior indica um modelo com clusters melhor definidos (CALIŃSKI; HARABASZ, 1974). O índice é calculado como a razão entre a soma da dispersão entre os clusters e a dispersão dentro de cada cluster para todos os clusters. Nesta medida, quanto maior for o valor do índice, melhor é o agrupamento e mais bem separados estão os clusters, e quanto mais próximo de zero, menos bem definidos estarão os clusters.

A Figura 2.14 ilustra um exemplo da distribuição de k clusters segundo o ICH. Como pode ser observado o ponto k=3 tem o ICH maior, sendo este valor o ideal para determinar o número ideal de clusters no conjunto de dados de exemplo (YAN et al., 2019).

Figura 2.14 - Exemplo da distribuição de k clusters pelo ICH.



Fonte: YAN et al. (2019).

2.4.3.2. CLARA

Os algoritmos K-means e K-medoids se baseiam em centróides e são muito úteis para produzir resultados de agrupamentos em muitas situações. No entanto, quando aplicado a conjuntos de dados de alta dimensionalidade, o tempo de execução é maior e é exigida uma quantidade maior de memória. Por esta razão, outros algoritmos baseados em centróides foram desenvolvidos e adaptados, especialmente para grandes conjuntos de dados (KAUFMAN; ROUSSEEUW, 1990).

O CLARA (KAUFMAN; ROUSSEEUW, 1990) (do inglês, *Clustering LARge Applications*) é um exemplo de algoritmo que estende o K-medoids usando uma abordagem de amostragem de instâncias e pode ser mais escalável em grandes conjuntos de dados (RENJITH; SREEKUMAR; JATHAVEDAN, 2022).

A Figura 2.15 representa o funcionamento do CLARA.

Figura 2.15 - Algoritmo CLARA.

(1) Para i=1 até 5, repita os seguintes passos:
(2) Obtenha amostra de 40+2*k* de instâncias da base, e execute o K-medoids para encontrar os *k* centróides da amostra.
(3) Para cada objeto *Oj* na base, determine qual dos *k* centróides é o mais similar a ele.
(4) Calcule a dissimilaridade média do agrupamento obtido no passo anterior. Se este valor for menor que o atual mínimo, utilize este valor como o atual mínimo e guarde os *k* centróides encontrados no passo 2 como os melhores centróides.
(5) Retornar ao passo 1 seguindo com a próxima iteração.

Fonte: Adaptado de (NG; HAN, 2002).

O CLARA considera uma pequena amostra de dados com tamanho pré-determinado (*n_sampling*), ao invés de encontrar centróides para todo o conjunto de dados, e aplica o algoritmo K-medoids para gerar um conjunto ideal de centróides para a amostra. A média (equivalente à soma) das diferenças das instâncias com seu centróide mais próximo é usada como uma medida de qualidade do cluster. Cada amostra do conjunto de dados é forçada a conter os centróides obtidos da melhor amostra de dados a cada iteração. Instâncias sorteadas aleatoriamente são adicionadas a este conjunto até que o tamanho da amostra seja alcançado. Os autores do CLARA sugerem que seja usada uma amostra de tamanho 40+2*k* de instâncias (KAUFMAN; ROUSSEEUW, 1990).

A maior vantagem do algoritmo CLARA, como destacado anteriormente, é a capacidade de trabalhar com conjuntos de dados de maior dimensionalidade, preservando os benefícios do K-medoids, e é mais robusto quanto a presença de valores discrepantes. A principal desvantagem está na dependência do algoritmo quanto ao tamanho da amostra escolhida, porém é possível calibrar o tamanho da amostra para um valor maior. Além disso, qualquer provável viés na seleção da amostra pode influenciar a qualidade geral do processo de agrupamento (RENJITH; SREEKUMAR; JATHAVEDAN, 2022).

2.4.3.3. DBSCAN

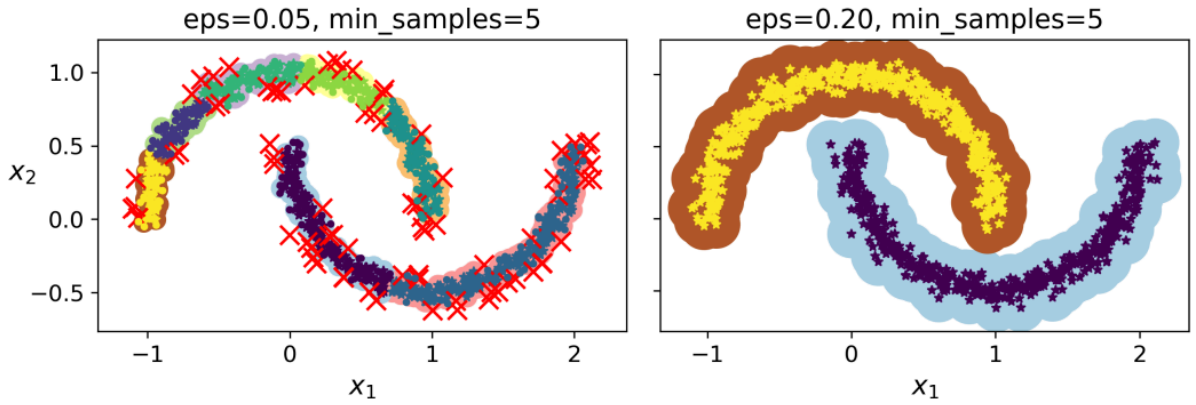
O DBSCAN (ESTER et al., 1996) (do inglês, *Density-Based Spatial Clustering of Applications with Noise*), que significa “Clusterização Espacial Baseada em Densidade de Aplicações com Ruído”, é mais recente do que o K-Means. O DBSCAN funciona identificando pontos que estão em regiões muito próximas (densas) em um espaço de dados. O princípio do DBSCAN está na formação de regiões densas de dados (clusters), separados por regiões relativamente vazias (MÜLLER e GUIDO, 2016).

Diferentemente dos algoritmos divisivos (K-means e CLARA) o DBSCAN não requer que o usuário defina o número de clusters. Esta abordagem permite que o algoritmo faça a identificação do número de clusters de forma arbitrária (GÉRON, 2019).

Para determinar a vizinhança de um objeto, o usuário precisará definir dois parâmetros principais no DBSCAN: *min_samples* e *eps*. O parâmetro *min_samples* determina o número mínimo de instâncias que devem estar ao redor de um ponto central (*core*) em uma distância *eps*, que determina o raio de vizinhança de um determinado ponto. A distância euclidiana é usada como cálculo padrão (SAMMUT; WEBB, 2017).

A Figura 2.16 ilustra uma representação resumida da formação de clusters pelo DBSCAN. O agrupamento representado no gráfico da esquerda, identificou muitas anomalias, além de 7 clusters diferentes. Ao ampliar a vizinhança de cada instância aumentando o eps para 0.2, no gráfico à direita, observa-se a formação de 2 clusters.

Figura 2.16 - Exemplos de formação de clusters com o DBSCAN.



Fonte: GÉRON (2019).

O DBSCAN funciona escolhendo um ponto arbitrário para começar. Em seguida (GÉRON, 2019):

- Para cada instância, o algoritmo conta quantas instâncias estão localizadas dentro de um pequena distância *eps* (*épsilon*) dela. Esta região é chamada de vizinhança da instância;
- Identifica os vizinhos de cada ponto no conjunto de dados verificando quais pontos estão a uma distância menor do valor de *eps* e, com isso, detecta os pontos centrais. Os pontos centrais, em um raio menor ou igual ao valor de *eps* apresentam uma quantidade de vizinhos menor ou igual a *min_samples* (incluindo o próprio ponto central). Em outras palavras, as instâncias centrais são aquelas que estão localizadas em regiões densas;
- Todas as instâncias na vizinhança de uma instância principal pertencem ao mesmo cluster. Isto pode incluir outras instâncias centrais, portanto, uma longa sequência de instâncias vizinhas formam um único cluster;
- Qualquer instância que não seja uma instância central e não tenha uma em sua vizinhança é considerada uma anomalia (*outlier*, ruído).

A Figura 2.17 representa o funcionamento do DBSCAN.

Figura 2.17 - Algoritmo DBSCAN.

```
Entrada
   data : base de dados com n objetos e m atributos (n x m)
   minPts : quantidade mínima de objetos na vizinhança
   raio : raio de vizinhança
Saída
   G : vetor com o rótulo dos objetos (n x 1)
Passos
   // Calcular a distância entre os objetos da base D(n x n)
   D = dist(data,data);

   // Controle dos objetos já visitados (vetor booleano com tamanho n)
   visitado[1:n] = false;

   // variáveis de controle
   rotulo = 1; // controla o rotulo do grupo atual

   // Análise dos objetos
   Para i=1:n Faça
   {
      Se (visitado[i]) Então continuar;
      // Marcar que o objeto já foi analisado
      visitado[i] = true;
      // Buscar os objetos dentro do raio de vizinhança
      vizinhos = ø;
      Para j=1:n Faça
         Se (D[i][j] <= raio) Então
            vizinhos.Add( j );

      // Verificar se o objeto é núcleo de grupo
      Se (vizinhos.size() < minPts) Então continuar;

      // Marcar o objeto com o rótulo do grupo atual
      G[i] = rotulo;

      v = 1;
      // Processo de expansão do grupo procurando por novos vizinhos
      Enquanto (v < vizinhos.size()) Faça
      {
         Se (visitado[vizinhos[v]] == false) ou (G[vizinhos[v]]==0) Então
         {
            // Buscar por novos vizinhos para formar o grupo
            aux = ø;
            Para j=1:n Faça
               Se (D[vizinhos[v]][j] <= raio) Então
                  aux.Add( j );

            // Adicionar os novos vizinhos na busca
            vizinhos.Add( aux );

            visitado[vizinhos[v]] = true;
         }
         G[vizinhos[v]] = rotulo;
         v = v +1;
      }

      rotulo = rotulo + 1;
   }
```

Fonte: CASTRO; FERRARI (2016).

O DBSCAN é mais lento que o K-means, mas pode ser dimensionado para conjuntos de dados relativamente grandes. Além disso, consegue identificar clusters de formas arbitrárias,

sendo ideal para identificação de outliers, sendo muitas vezes usado com essa finalidade. No entanto, ele pode não ser tão preciso ao definir os grupos tendo em vista que alguns pontos podem estar no limite entre um cluster e outro, havendo uma troca a depender da ordem de processamento dos dados (PATLOLLA, 2023; PYTHON, 2023).

2.4.4. Regras de associação

Resumidamente, a tarefa de regras de associação tem duas etapas (FACELI et al., (2011): (i) encontrar o conjunto de itens frequentes, mediante um limite mínimo definido pelo usuário; (ii) encontrar o conjunto de regras de associação, com um limite mínimo de uma medida definida pelo usuário.

Formalmente, as regras de associação possuem a forma: X → Y (lê-se: "Se X, então Y", onde X é denominado o antecedente da regra e Y o consequente da regra). Os itens frequentes que compõem X e Y devem ser disjuntos e com ao menos um elemento (AGRAWAL; IMIELIŃSKI; SWAMI, 1993).

Um exemplo de regra obtida a partir do comportamento do consumidor no contexto de um supermercado seria: {Fraldas} → {Cerveja}. Esta regra sugere que pessoas que compram fraldas também compram cerveja. Outro exemplo de regra é: {Pão, Manteiga} → {Leite}, ou seja, quem compra pão e manteiga, provavelmente compra leite. Um exemplo aplicado ao domínio educacional para identificar o desempenho de estudantes em disciplinas pode ser formulado assim: {Matemática, Física} → {Programação}. Essa regra poderia indicar que estudantes com boas notas em disciplinas associadas à área das exatas têm bom desempenho em Programação.

Para definir os itens frequentes e avaliar as regras de associação, algumas medidas foram definidas, sendo suporte e confiança as mais utilizadas (HARRINGTON, 2012). Essas e outras medidas são explicadas a seguir.

Para facilitar a explicação sobre medidas de avaliação, a Figura 2.18 será usada como meio de representar um conjunto de dados.

Figura 2.18 - Conjunto de dados com 10 transações e itens frequentes com suporte de 30%.

TID	Itens
1	{a,d,e}
2	{b,c,d}
3	{a,c,e}
4	{a,c,d,e}
5	{a,e}
6	{a,c,d}
7	{b,c}
8	{a,c,d,e}
9	{b,c,e}
10	{a,d,e}

0 itens	1 item	2 itens	3 itens
Ø: 10	{a}: 7	{a,c}: 4	{a,c,d}: 3
	{b}: 3	{a,d}: 5	{a,c,e}: 3
	{c}: 7	{a,e}: 6	{a,d,e}: 4
	{d}: 6	{b,c}: 3	
	{e}: 7	{c,d}: 4	
		{c,e}: 4	
		{d,e}: 4	

Fonte: FACELI et al. (2011).

Cada linha da tabela à esquerda da Figura 2.18 corresponde a uma transação (10 no total) onde, na primeira coluna, cada transação é identificada por um ID e, na segunda coluna, está o conjunto de itens de cada transação. Na tabela à direita estão enumerados todos os conjuntos de itens frequentes usando o suporte mínimo de 30%.

O suporte é definido como a proporção (frequência) em que um conjunto de itens X ocorre (suporta) no conjunto de dados com N transações (AGRAWAL; IMIELIŃSKI; SWAMI, 1993). O cálculo do suporte é mostrado na Fórmula (1).

$$Suporte(X) = \frac{X}{N} \tag{1}$$

O valor de X corresponde ao total de transações em que o(s) item(ns) X aparece(m) no conjunto de dados; e N corresponde ao número total de transações.

Observando a Figura 2.18 tem-se, por exemplo:

- Suporte({a,e}) => 6/10 = 0.6. Ou seja os itens {a,b} ocorrem em 6 de um total de 10 transações, tendo o suporte de 60%;
- Suporte({a,d,e}) => 4/10 = 0.4. Dessa forma os itens {a,d,e} ocorrem em 4 de um total de 10 transações, tendo o suporte de 40%;
- Suporte({a,c,d,e}) => 2/10 = 0.2. Assim os itens {a,c,d,e} ocorrem em apenas 20% das transações. Para esse exemplo foi definido um valor mínimo de suporte em 30%, por isso os itens {a,c,d,e} não aparecem na tabela à direita;

Normalmente, o suporte mínimo (*minimum support threshold*) é usado para se obter o conjunto de itens frequentes, onde são recuperados os itens cujo suporte seja maior ou igual ao valor de suporte mínimo especificado pelo usuário (HIPP; GÜNTZER; NAKHAEIZADEH, 2000). Quando aplicadas às regras, itens antecedentes e consequentes com baixos valores de suporte também são de pouco interesse sob a perspectiva do negócio, por isso o suporte também pode ser usado para eliminar regras pouco interessantes (CASTRO; FERRARI, 2016).

A confiança serve para medir a força de uma regra, ou seja, é um número que expressa a possibilidade de ocorrer o consequente em uma transação, dado que ela também contém o antecedente (HIPP; GÜNTZER; NAKHAEIZADEH, 2000), conforme a Fórmula (2).

$$Confiança(X \rightarrow Y) = \frac{Suporte(X\ U\ Y)}{Suporte(X)} \tag{2}$$

A medida de confiança de uma regra X→Y é a probabilidade de ver o consequente (Y) ocorrer em uma transação dado que ela também contém o antecedente (X).

Para ilustrar, a Tabela 2.1 foi elaborada com base nos itens frequentes (com suporte mínimo de 30%) definido anteriormente na Figura 2.18. Para cada transação, o valor 1 indica a presença do item na transação e o valor 0 implica na ausência desse item na transação[8].

[8] As variáveis dummies (fictícias) são usadas para transformar cada valor exclusivo de uma variável categórica em sua própria coluna que é verdadeira (valor 1) ou falsa (valor 0) (OZDEMIR, 2016).

Tabela 2.1 - Itens frequentes por registros nas transações do exemplo.

TID	**a**	**b**	**c**	**d**	**e**
1	1	0	0	1	1
2	0	1	1	1	0
3	1	0	1	0	0
4	1	0	1	1	1
5	1	0	0	0	1
6	1	0	1	1	1
7	0	1	1	0	0
8	1	0	1	1	1
9	0	1	1	0	1
10	1	0	0	1	1

Fonte: Elaborado pelo autor, baseado em FACELI et al. (2011).

Ao gerar regras de associação derivadas desse conjunto de itens frequentes considerando o valor mínimo de confiança (*minimum confidence threshold*) maior ou igual a 70%, tem-se os resultados indicados na Tabela 2.2.

Tabela 2.2 - Exemplos de medidas de avaliação para regras de associação.

X (antecedente)	Y (consequente)	Confiança	Lift
{b}	{c}	1.0	1.43
{e}	{a}	0.86	1.22
{d}	{a}	0.83	1.19
{e,c}	{d,a}	0.75	1.50
{a}	{d,e}	0.71	1.43

Fonte: Elaborado pelo autor.

Os resultados ilustrados indicam, por exemplo, que a confiança foi:

- Confiança({b}→{c}) = (3)/(3) = 1.0. Ou seja, em 100% das vezes que ocorre o item {b} implica na ocorrência do item {c};
- Confiança({a}→{d,e}) = (5)/(7) = 0.71. Ou seja, em 71% das vezes que ocorre o item {a} implica na ocorrência do item {d,e};

- Confiança({e,c}→{d,a}) = (3)/(4) = 0.75. Ou seja, em 75% das vezes que ocorre o item {e,c} implica na ocorrência do item {d,e}.

Outras medidas podem ser usadas na avaliação das regras de associação. O Lift, também conhecido como interesse da regra de associação, é representado pela Fórmula (3).

$$Lift(X \rightarrow Y) = \frac{Suporte(X\ U\ Y)}{Suporte(X) * Suporte(Y)} \quad (3)$$

A medida Lift indica qual a chance de Y (consequente) ocorrer, se X (antecedente) ocorrer, considerando toda ocorrência de Y. Essa medida avalia a chance dos itens ocorrerem juntos (ALPAYDIN, 2010). Se X e Y forem independentes, então espera-se que o Lift seja próximo de 1; se o valor do Lift for maior que 1, pode-se afirmar que X torna Y mais provável, e se o aumento for menor que 1, ter X torna Y menos provável (ALPAYDIN, 2010).

Retomando aos resultados ilustrados na Tabela 2.2, tem-se que:

- Lift({d}→{a}) = (0.5) / (0.6 * 0.7) = 1.19. Ou seja, há 1.19 chances de ocorrer {a}, quando ocorre {d};
- Lift({e,c}→{d,a}) = (0.3) / (0.4 * 0.5) = 1.5. Ou seja, há 1.5 chances de ocorrer {d,a}, quando ocorre {e,c}.

O principal desafio quando se trata de regras de associações é o grande número de regras que teoricamente devem ser consideradas. O número de regras cresce exponencialmente a depender do número de itens. Como não é desejável explorar um grande número de regras, os limites mínimos (*minimum threshold*) para as medidas de avaliação (suporte, confiança, lift e outras), como mencionado anteriormente, podem ser aplicadas para as medidas (HIPP; GÜNTZER; NAKHAEIZADEH, 2000).

Muitos algoritmos têm sido propostos para aplicação das regras de associação. Neste trabalho foram usados dois dos principais algoritmos descritos na literatura (HARRINGTON, 2012): o Apriori e o FP-growth. A escolha por esses algoritmos se deu mediante as informações obtidas com a RSL realizada, com foco na análise de desempenho de estudantes no ENEM quando implementada por meio da tarefa de regras de associação. Os trabalhos relacionados apresentaram evidências de bons resultados com o uso desses algoritmos, especialmente no uso do Apriori.

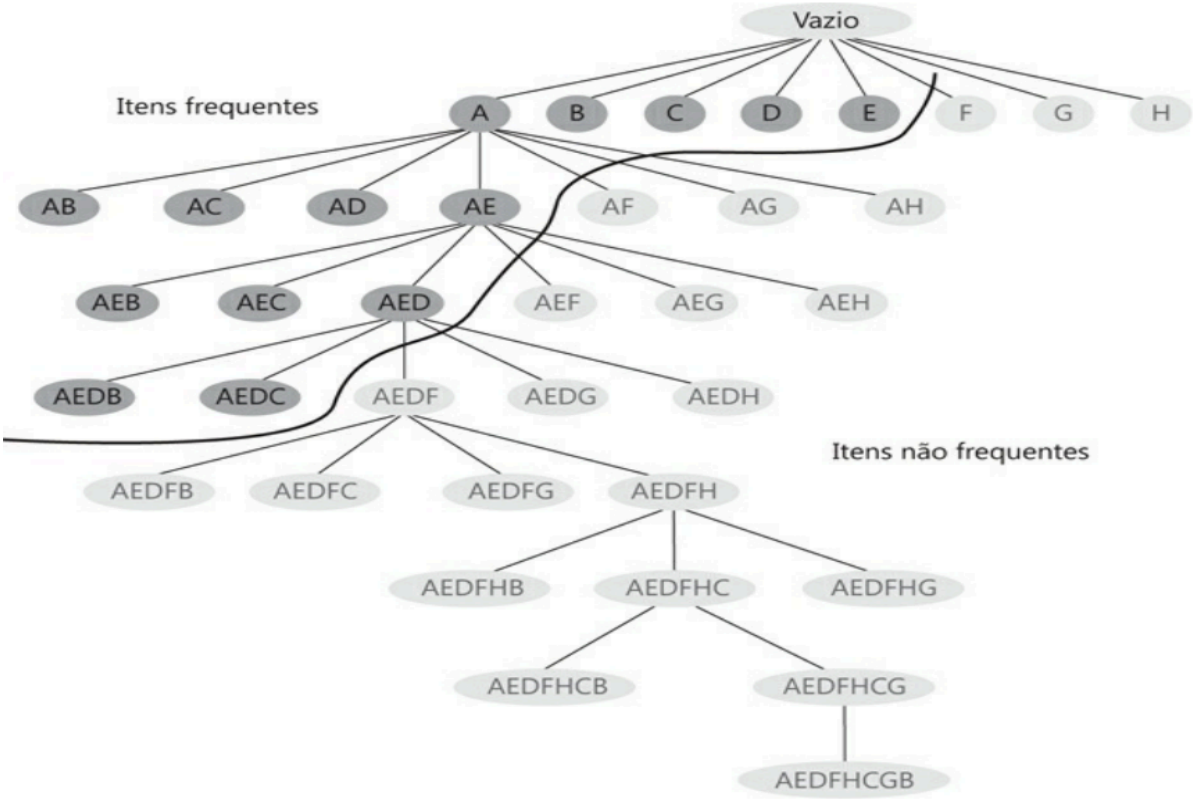
O algoritmo FP-Growth foi analisado complementarmente, pois, devido à sua estratégia de dividir para conquistar, consegue reduzir substancialmente o tamanho dos conjuntos de dados a serem pesquisados, bem como das regras de associação a serem avaliadas (HAN; KAMBER; PEI, 2012). Isso favorece um menor tempo de execução em relação ao Apriori. Esses algoritmos são explicados nas seções seguintes.

2.4.4.1. Apriori

A mineração de regras de associação pode ser aplicada por meio de diferentes algoritmos, assim como ocorre nas demais tarefas de AM supervisionadas e não supervisionadas. O algoritmo mais clássico voltado à tarefa de regras de associação é o Apriori, porém há outros algoritmos que são variações e melhorias dele (AGRAWAL et al., 1996; ALPAYDIN, 2010).

O Apriori realiza múltiplas interações no conjunto de dados, sendo essa varredura um tipo de busca em profundidade. A cada iteração são gerados conjunto de itens candidatos de k elementos a partir de conjuntos de itens com k-1 elementos. Na Figura 2.19, após a primeira iteração, são encontrados 8 itens {A, B, C, D, E, F, G, H}. Na segunda iteração, para o item A, são geradas as combinações possíveis, resultando nos itens {AB, AC, AD, AE, AF, AG, AH}. Nas demais iterações, em ordem crescente de tamanho o Apriori gera novos itens, baseado nos itens candidatos da iteração anterior, até que todos os elementos sejam conhecidos.

Figura 2.19 - Representação de busca de itens frequentes com Apriori.



Fonte: CASTRO; FERRARI (2016).

Uma propriedade muito significativa para redução do espaço de busca, chamada de antimonoticidade, indica que se um item é frequente, logo os itens gerados por meio dele também são frequentes e, se um item não é frequente, logo os itens gerados por meio dele também serão identificados como não frequentes. A medida de suporte pode ser usada para eliminar os itens menos frequentes (RAO; GUPTA, 2012). A Figura 2.19 ilustra uma poda, eliminando os itens não frequentes.

A Figura 2.20 mostra a função principal do Apriori para geração de itens frequentes.

Figura 2.20 - Função do Apriori para geração de itens frequentes.

```
1.  Entrada: Uma base de dados D e o valor de suporte mínimo min_sup.
2.  Saída: O conjunto L com todos os itemsets frequentes.
3.  Função apriori-main(D, min_sup)
4.  L_1 = {conjunto dos itemsets frequentes de tamanho 1 contidos em D};
5.  para (k = 2; L_{k-1} ≠ ∅; k++)
6.      C_k = apriori-gen(L_{k-1});
7.      para todas transações t ∈ D fazer
8.          C_t = subset(C_k, t);
9.          para todos candidatos c ∈ C_t fazer
10.             c.count++;
11.         fim para
12.     fim para
13.     L_k = {c ∈ C_k | c.count ≥ min_sup};
14. fim para
15. retorne L = ∪_k L_k;
```

Fonte: MARIANO (2011).

Basicamente, a função tem o objetivo de identificar os itemsets frequentes, filtrando pela medida de suporte e, em seguida, construir regras de associações com esses itemsets selecionados. Em seguida é realizada uma varredura no conjunto de dados para cada itemset de tamanho k. Há duas funções adicionais: apriori-gen (linha 6) que retorna todos os k-itemsets candidatos e subset (linha 8) que gera as combinações possíveis para os itemsets candidatos e elimina aqueles itensets não frequentes (MARIANO, 2011).

A segunda parte do Apriori, representada na Figura 2.21, corresponde à função *ap-genrules*. que recebe uma medida mínima de confiança e filtra os itemsets possíveis gerando as regras de associação, que são mostradas ao final da execução do algoritmo.

Figura 2.21 - Função do Apriori para geração das regras de associação.

```
1.  Entrada: Um conjunto de itemsets L e a confiança mínima da regra min_conf.
2.  Saída: O conjunto de regras R.
3.  Função ap-genrules(L, min_conf)
4.  para todos k-itemsets ∈ L fazer
5.      para (i = k-1; i ≥ 1; i--)
6.          para todos i-itemsets ⊂ k-itemset fazer
7.              conf = suporte(k-itemset) / suporte(i-itemset);
8.              se (conf ≥ min_conf) então
9.                  adicione i-itemset → (k-itemset − i-itemset) em R;
10.             fim se
11.         fim para
12.     fim para
13. fim para
14. retorne R;
```

Fonte: MARIANO (2011).

Para uma melhor compreensão do funcionamento do Apriori, considera-se um exemplo a partir da Figura 2.18 mostrada anteriormente. A Tabela 2.3 mostra como seria a disposição dos itens de cada transação com a lista completa de itemsets candidatos e, após a aplicação da medida de suporte mínimo de 30%, a lista de itens frequentes selecionados (destacados) pela primeira função do Apriori.

Tabela 2.3 - Exemplo de itens frequentes obtidos com o Apriori.

0 itens	1 item	2 itens	3 itens	4 itens
-	**{a} (70%)**	**{a,e} (60%)**	**{d,a,e} (50%)**	**{d,a,c,e} (30%)**
-	**{c} (70%)**	**{d,a} (50%)**	**{d,a,c} (30%)**	-
-	**{e} (70%)**	**{d,e} (50%)**	**{d,c,e} (30%)**	-
-	**{d} (60%)**	**{a,c} (40%)**	**{a,c,e} (30%)**	-
-	**{b} (30%)**	**{d,c} (40%)**	{d,b,c} (10%)	-
-	-	**{c,e} (40%)**	{b,c,e} (10%)	-
-	-	**{b,c} (30%)**	-	-
-	-	{d,b} (10%)	-	-
-	-	{d,e} (10%)	-	-

Fonte: Elaborado pelo autor.

Seguindo o exemplo, para a segunda parte do Apriori, que se refere à geração de regras de associação, caso aplicada a medida mínima de confiança de 70%, seriam obtidas as cinco regras destacadas na Tabela 2.4. A primeira coluna contém os itemsets antecedentes das regras, a segunda, por sua vez, os consequentes. As demais colunas apresentam as medidas de avaliação para as regras obtidas. As regras estão ordenadas pela medida de confiança.

Tabela 2.4 - Exemplos de regras de associação obtidas com Apriori.

Antecedente	Consequente	Suporte	Confiança	Lift
{b}	{c}	30%	100%	1.43
{a}	{e}	60%	86%	1.22
{d}	{a}	50%	83%	1.19
{a,c}	{d,e}	30%	75%	1.50
{e}	{d,a}	50%	71%	1.42

Fonte: Elaborado pelo autor.

Um conjunto de dados que contém N itens possíveis pode gerar até 2^N-1 itens candidatos (HARRINGTON, 2012). No exemplo ilustrado anteriormente o conjunto de dados contém apenas 5 itens, o que resultaria em até 2^5-1 = 31 possibilidades de itens. Se isso fosse aplicado, por exemplo, a um conjunto contendo 30 itens, geraria até 1.073.741.823 de itens possíveis. Isso levaria muito tempo para ser computado. Para reduzir o tempo necessário para calcular este valor, outros algoritmos buscam reduzir o tempo de execução especialmente quando há um grande número de itens candidatos e de regras, o que acaba sendo muito dispendioso percorrer várias vezes o conjunto de dados (HAN; KAMBER; PEI, 2012).

2.4.4.2. FP-growth

O algoritmo de FP-growth armazena os itens em uma estrutura de dados compacta chamada *FP-tree* (árvore "padrão frequente"). Uma *FP-tree* é semelhante à estrutura de dados de árvores, porém contém links que conectam itens frequentes. Os itens frequentes podem ser considerados uma lista de itens vinculada (HARRINGTON, 2012).

A Figura 2.22 corresponde à função *FP-Tree*. que recebe um conjunto de dados e uma medida de mínima de suporte, e posteriormente, na segunda fase, a função *FP-Growth* é então usada para encontrar as regras de associação.

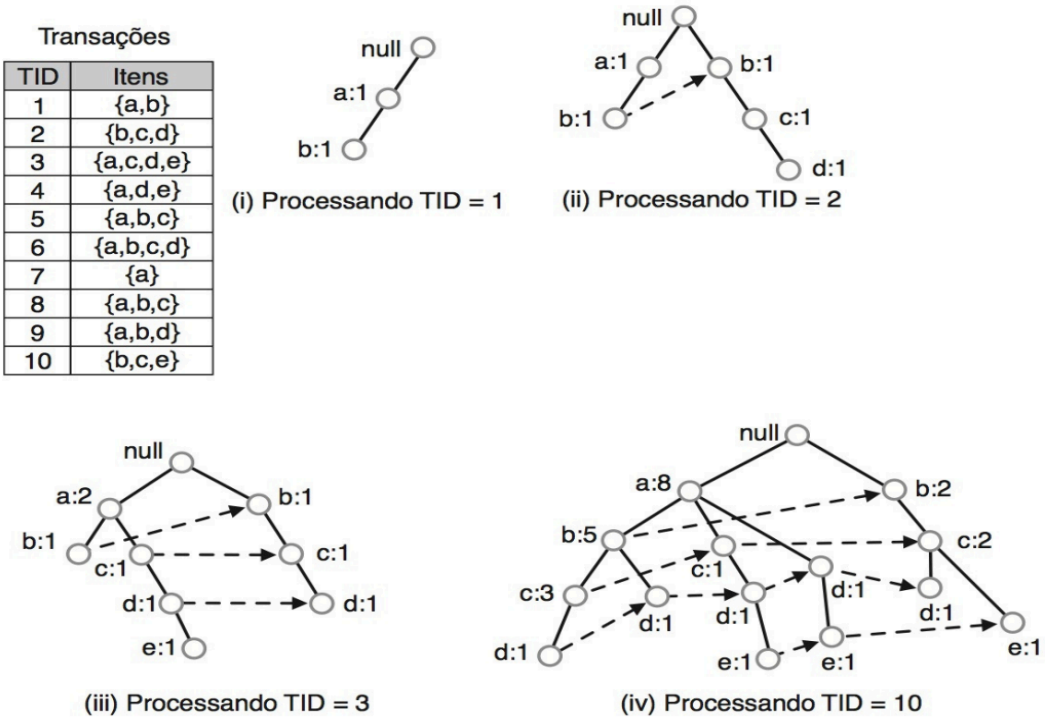
Figura 2.22 - Algoritmo FP-growth.

1. **Entrada**: Uma base de dados D e o valor de suporte mínimo *min_sup*.
2. **Saída**: O conjunto L com todos os *itemsets* frequentes.
// Fase 1 - Construção da *FP-Tree*
3. **Função** *FP-Tree*(D, *min_sup*)
4. **percorra** a base de dados D uma vez;
5. **determine** o conjunto de itens frequentes F e seus suportes;
6. **ordene** F em ordem decrescente em função do suporte e chamá-la de L;
7. **crie** a raiz da *FP-Tree* T e coloque como "*null*";
8. **para cada** transação t em D **faça**
9. **selecione e ordene** os itens frequentes em t de acordo com a ordem de L, tornando a lista de itens frequentes em t igual a $[p|P]$, onde p é o primeiro elemento e P é o resto da lista;
10. **execute** *insere_tree*($[p|P]$, T).
11. **se** T tiver um filho N em que $N.nome\_item = p.nome\_item$ **então**
12. **incremente** o contador de N por em 1;
13. **senão**
14. **crie** um novo nó N, e inicie seu contador com 1;
15. **ligue** o seu *parent-link* a T, e seu *node-link* aos nós de mesmo (*nome_item*) através da estrutura dos *node-links*;
16. **fim se**
17. **se** P não for vazio **então**
18. **chame** *insere_tree*(P, N) recursivamente;
19. **fim se**
20. **fim para**
// Fase 2 - Mineração da *FP-Tree*
21. **Função** *FP-Growth*(*Tree*, α)
22. **se** *Tree* contém apenas um caminho P **então**
23. **para** cada combinação β de nós no caminho P **faça**
24. **gere** o padrão $\beta \cup \alpha$ com suporte = *min_sup* dos nós em β;
25. **fim para**
26. **senão**
27. **para** cada a_i na tabela de *node-links* de *Tree* **faça**
28. **gere** o padrão $\beta = a_i \cup \alpha$ com suporte = $a_i.suporte$;
29. **construa** a base de padrões condicionada de β e crie a *FP-Tree* condicionada de β chamada de $Tree_\beta$;
30. **se** ($Tree_\beta \neq \emptyset$) **então**
31. *FP-growth*($Tree_\beta$, β);
32. **fim se**
33. **fim para**
34. **fim se**

Fonte: MARIANO (2011).

Um exemplo de construção de uma *FP-tree* é ilustrado na Figura 2.23, conforme Faceli et al. (2011).

Figura 2.23 - Exemplo de construção de uma FP-tree.



Fonte: FACELI et al. (2011).

Faceli et al. (2011) explicam que a propriedade da monotonicidade do suporte sugere uma representação sumarizada do conjunto de itemsets frequentes em dois tipos:

- Maximais: um itemset é denominado maximal se ele é frequente, mas nenhum dos seus superconjuntos próprios é frequente;
- Fechados: um conjunto de itens frequente é chamado fechado se, e somente se, ele não tem superconjuntos frequentes com a mesma frequência.

No exemplo comentado, há 13 itens frequentes fechados (suporte mínimo de 30%): {b,c}, {d,c,a}, {e,c,a}, {c,a}, {d,c}, {e,c}, {d,e,a}, {d,a}, {d}, {e,a}, {a}, {c}, {e}, e 4 conjuntos de itens maximais, que são: {b, e}{a, e, d}{a, e, e}{a, d, e}. Todos os itemsets frequentes podem ser vistos como um subconjunto de pelo menos um dos conjuntos maximais. Assim, os itemsets maximais são um subconjunto dos itemsets fechados. A partir dos itemsets maximais, é possível derivar todos os itemsets frequentes (mas não o seu suporte) calculando todas as interseções não vazias, mantendo o mesmo suporte (FACELI et al., 2011).

Seguindo o exemplo, para a geração de regras com FP-Growth, aplicando uma medida mínima de confiança de 70%, seriam obtidas as quatro regras destacadas na Tabela 2.5.

Tabela 2.5 - Exemplos de regras de associação obtidas com FP-Growth.

Antecedente	Consequente	Suporte	Confiança	Lift
{c}	{b}	50%	83%	1.19
{d}	{a}	40%	80%	1.00
{c,a}	{b}	30%	75%	1.07
{b}	{a}	50%	71%	0.89

Fonte: Elaborado pelo autor.

Este estudo não envolveu uma abordagem comparativa entre os algoritmos de regras de associação. Entretanto, em se tratando de desempenho computacional (memória e tempo de execução) o FP-Growth requer menos recursos quando comparado ao algoritmo Apriori (ALDINO et al., 2021).

3. TRABALHOS RELACIONADOS

Este capítulo apresenta trabalhos relacionados à temática desta pesquisa. Esses estudos têm sido realizados na busca por uma melhor compreensão a respeito de diferentes implicações da pandemia na Educação e, em alguns casos, no desempenho dos estudantes no ENEM. Ao final, um quadro comparativo entre os trabalhos relacionados é mostrado e discutido, ao passo que é feita uma síntese com respeito aos diferenciais desta dissertação.

Os estudos foram obtidos por meio de buscas nas bases: ACM Digital Library, IEEE Explore, Biblioteca SOL-RBIE e Biblioteca Digital Brasileira de Teses e Dissertações (BDTD). Os indexadores SCHOLAR e CAPES também foram usados para ampliar as buscas por estudos em outras fontes.

3.1. AED para identificar o impacto da Pandemia no ENEM em três estados

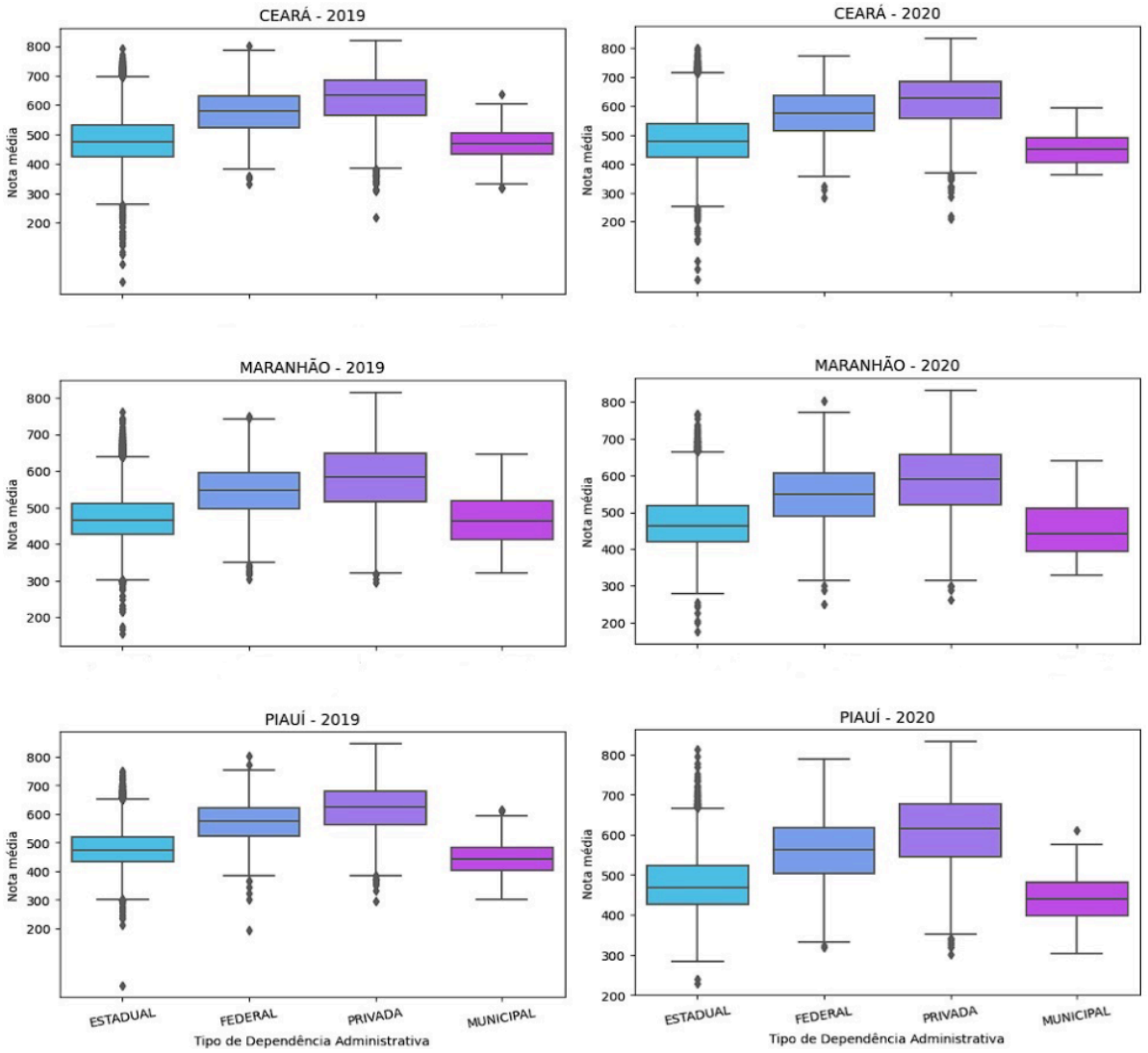
Os autores Weber Neto et al. (2022a) buscaram identificar como a pandemia interferiu no desempenho dos estudantes que realizaram o ENEM, comparando o desempenho entre os anos de 2019 e 2020 (primeiro ano da Pandemia). Para isso, conduziram um processo de Análise Exploratória de Dados (AED) com dados referentes aos estados do Maranhão, Ceará e Piauí.

Sem que ainda houvesse dados oficiais de hábitos de estudo referentes a 2020, os autores usaram apenas dados socioeconômicos, dados sobre escolas do ensino médio e notas por área de conhecimento.

Os resultados mostraram que o número de ausências na prova mais que dobrou nos três estados, sendo esse um impacto evidente causado pelo primeiro ano de aplicação do ENEM durante a pandemia. Esses estados não sofreram alteração significativa quanto ao desempenho dos estudantes pelo tipo de dependência administrativa das escolas. As escolas privadas obtiveram as maiores notas, seguidas das escolas públicas federais, estaduais e municipais. Os gráficos de *boxplot* ilustrados na Figura 3.1 mostram as notas médias nos três estados nos anos de 2019 e 2020.

Nos três estados houve aumento de participantes de escolas privadas. Quanto às escolas públicas, o Ceará teve menos participantes, já Maranhão e Piauí tiveram um pequeno aumento de participantes apenas em escolas públicas municipais.

Figura 3.1 - Relação entre notas médias por estado e dependência administrativa das escolas.



Fonte: WEBER NETO et al. (2022a, p.7).

Os autores sugerem que, em trabalhos futuros, possam ser incluídas análises a partir de um processo que contemple as demais etapas do CRISP-DM, tendo em vista esse ser ainda um trabalho inicial no contexto da pandemia e limitado à uma análise mais exploratória. Sugerem ainda a inclusão de microdados do ENEM do ano de 2021, referente ao segundo ano de pandemia.

3.2. Análise comparativa sobre como a Pandemia impactou o ENEM

A segunda publicação de Weber Neto et al. (2022b) teve como objetivo realizar uma análise comparativa de dados dos anos de 2019 (anterior ao início da pandemia) e de 2020 (primeiro ano da pandemia) 2019 em nível nacional. O estudo considerou o cenário de pausa das aulas presenciais e, posteriormente, a execução do ensino de forma remota, de maneira a identificar mudanças no panorama geral do ENEM em consequência à pandemia e como foi afetado o processo de ensino e aprendizagem dos estudantes.

A presença de inscritos no ENEM foi um dos principais destaques sobre o impacto causado pela pandemia. Para o ano de 2019, 77% dos candidatos inscritos estiveram presentes na prova, já no ano de 2020 apenas 48% dos estudantes inscritos realizaram o exame. Os autores sugerem que a ausência se deve a questões como, por exemplo, mudança na data da prova, falta de preparação dos candidatos, medo de contrair o coronavírus e a situação econômica desfavorável de parte dos estudantes. Houve também um aumento considerável de solicitação de reaplicação da prova para casos de incidentes, falhas e surtos de COVID em algumas regiões. Em 2019 foram realizadas 150 solicitações, enquanto que em 2020 foram 235.204 solicitações.

Sobre as mudanças de participantes em relação a raça e sexo, não houveram mudanças significativas quando comparados esses atributos individualmente. Quando relacionados esses atributos, a mudança mais significativa foi entre o número de autodeclarados como brancos, que aumentou para o sexo feminino e reduziu para o masculino em 2020.

Os atributos que tiveram maior influência na nota média do ENEM nos dois anos citados foram, respectivamente, a renda familiar, a quantidade de computadores e celulares na residência, acesso à internet e escolaridade dos pais. Foi usada a correlação de Pearson (BENESTY et al., 2009) para esse levantamento. A Tabela 3.1 mostra as correlações obtidas para cada uma dessas variáveis principais.

Tabela 3.1 - Maiores correlações entre as variáveis socioeconômicas e a nota média.

Questão Socioeconômica	**ENEM 2019**	**ENEM 2020**
Renda Familiar	0,48	0,51
Possui Computadores	0,41	0,46
Possui Celulares	0,28	0,32
Acesso à Internet	0,26	0,40
Escolaridade Pai	0,25	0,29
Escolaridade Mãe	0,31	0,34

Fonte: Adaptado de WEBER NETO et al. (2022b, p.8).

De modo geral, os resultados apontam para o aumento da diferença de notas entre os estudantes de maior e menor renda familiar entre os dois anos, sendo essa a variável de maior correlação. A variável ‘Acesso à Internet’ foi a que sofreu maior alteração entre os anos. Considerando o segundo ano de pandemia e continuidade da modalidade de ensino remoto, essa variável passou a influenciar ainda mais no desempenho dos estudantes, bem como o uso de computadores e celulares, instrumentos essenciais à condução das atividades acadêmicas.

3.3. Data Analysis to Identify the Impact of the Pandemic in 3 States

Em outro estudo mais recente, Weber Neto et al. (2023) ampliaram e atualizaram a investigação em Weber Neto et al. (2022a) ao realizarem novamente um estudo comparativo entre os três estados (Maranhão, Ceará e Piauí.) considerando as 3 últimas edições do exame antes da Pandemia (2017, 2018 e 2019) e 2 anos pós-pandemia (2020 e 2021).

Os resultados da análise exploratória foram apresentados individualmente para os três estados, considerando atributos socioeconômicos, informações sobre escolas do ensino médio dos participantes e o desempenho (média entre as áreas de conhecimento). Os autores identificaram que a presença dos participantes no ENEM em 2020 foi bastante afetada pela pandemia nos três estados, caindo ainda mais a participação em 2021. Quanto ao desempenho, os resultados mostraram, de modo geral, pouco impacto causado pela pandemia, visto que a nota média se manteve com valores semelhantes nos três estados no período anterior e durante a pandemia.

Quanto ao desempenho por tipo administrativo, as escolas particulares tiveram as notas mais altas, seguidas pelas públicas federais, estaduais e municipais nos dois anos de pandemia avaliados. Os autores não chegaram a desenvolver modelos de aprendizado de máquina ou algum tipo de análise mais analítica, conforme indicaram para trabalhos futuros nos estudos ora descritos.

3.4. Análise dos Perfis de Alunos do Ensino Superior na Modalidade Remota

Em Pereira Junior et al. (2021), os autores buscaram identificar características relevantes sobre o perfil de estudantes universitários em relação ao ensino remoto.

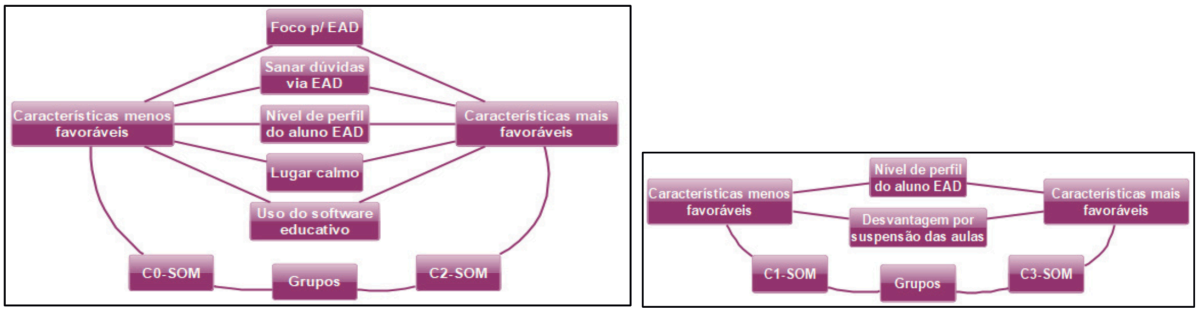
Os dados foram obtidos por meio de um formulário de 30 questões elaborado pelos autores. Foi gerada uma base de dados com 483 instâncias. As questões envolviam: informações pessoais, acadêmicas e socioeconômicas de estudantes; condições de acesso a recursos tecnológicos; condições psicológicas; informações sobre adaptação, conhecimento, engajamento e interação em relação à plataforma de ensino remoto; interação com colegas e professores; informações sobre a situação do estudante em relação às aulas e opinião sobre a modalidade de ensino remoto.

Os autores usaram os algoritmos Self-organizing map (SOM) e K-means para realizar uma análise de perfis dos estudantes. Para cada algoritmo foram criados grupos diferentes e identificadas as quantidades de instâncias pertencentes aos grupos e distribuição dos valores de cada atributo. O algoritmo SOM foi usado para definição automática de um número ideal de grupos, sendo definido 4 grupos. Baseado neste número, os autores consideraram usar, para o K-means, valores de k mais próximos, sendo definidos os valores de k com 3 e 5 grupos. A partir da quantidade de instâncias de cada grupo foi realizada a análise de cada um dos grupos/perfis, destacando-se os atributos que representavam as características mais favoráveis e menos favoráveis à modalidade de ensino remoto.

Para os 4 grupos formados com o SOM, dois se mostraram contrários à aplicação de ensino remoto durante a pandemia e fora dela (C0-SOM e C2-SOM). Os outros dois grupos se mostraram mais favoráveis (C1-SOM e C3-SOM). Considerando as quantidades de instâncias pertencentes aos grupos, a Figura 3.2(a) representa os principais atributos divisores dos grupos contrários ao ensino remoto. que foram ‘foco_para_ead’, ‘sanar_duvidas_ead’, ‘nivel_perfil_aluno_ead’, ‘lugar_calmo’ e ‘uso_software_educativo’. Para os grupos mais

favoráveis ao ensino remoto, ilustrados na Figura 3.2(b) destacaram-se como divisores os atributos ‘nivel_perfil_aluno_ead’ e ‘desvantagem_suspensao_aulas’.

Figura 3.2 - Atributos destacados em grupos formados com SOM.

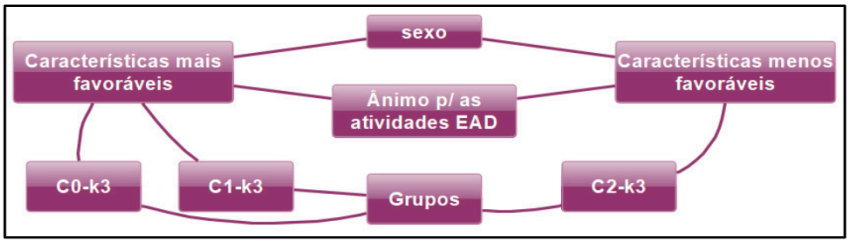


(a) Grupos contrários (b) Grupos favoráveis

Fonte: PEREIRA JUNIOR et al. (2021, p.6).

Para o agrupamento feito com K-means que resultou em 3 grupos, um grupo não optaria pelo ensino remoto na pandemia (C2-k3), enquanto os outros dois optariam por esta modalidade na pandemia, mas não fora dela (C0-k3 e C1-k3). A Figura 3.3 mostra que os principais atributos separadores que dividiram as instâncias em três grupos foram: ‘sexo’ e ‘animo_para_atividades_ead’.

Figura 3.3 - Atributos destacados em grupos formados com K-means (k=3).



Fonte: PEREIRA JUNIOR et al. (2021, p.7).

No cenário com 5 grupos, dois grupos optariam apenas pelo EAD durante a pandemia, outros dois grupos não optariam pelo EAD e um grupo optaria pelo EAD durante a pandemia e fora dela. Os grupos foram definidos principalmente pelos atributos ‘nivel_academico’, ‘situacao_emprego’ e ‘nivel_estimulo_negativo_estresse’.

Ao final, os autores identificaram três grupos ideais, sendo estes formados por estudantes que optariam pelo ensino remoto durante a pandemia e fora de pandemia, outro composto por aqueles optariam pelo ensino remoto apenas durante a pandemia e o último composto por estudantes que não optariam pelo ensino remoto em ambas as situações

Os resultados mostraram que estudantes com menor ânimo para atividades no ensino remoto pertenciam aos grupos de estudantes com condições de adaptação menos favoráveis do ponto de vista econômico e psicológico. Já o grupo de estudantes com características favoráveis e mais engajadas com o ensino remoto tinham melhores condições de acesso às tecnologias e pertenciam às escolas de melhor infraestrutura.

Apesar da resposta positiva de diversos estudantes quanto ao nível de adaptação, muitos preferem não optar pela modalidade de ensino remoto, sendo este um ponto em aberto para trabalhos futuros.

3.5. Analysis of ENEM's attendants using a clustering approach

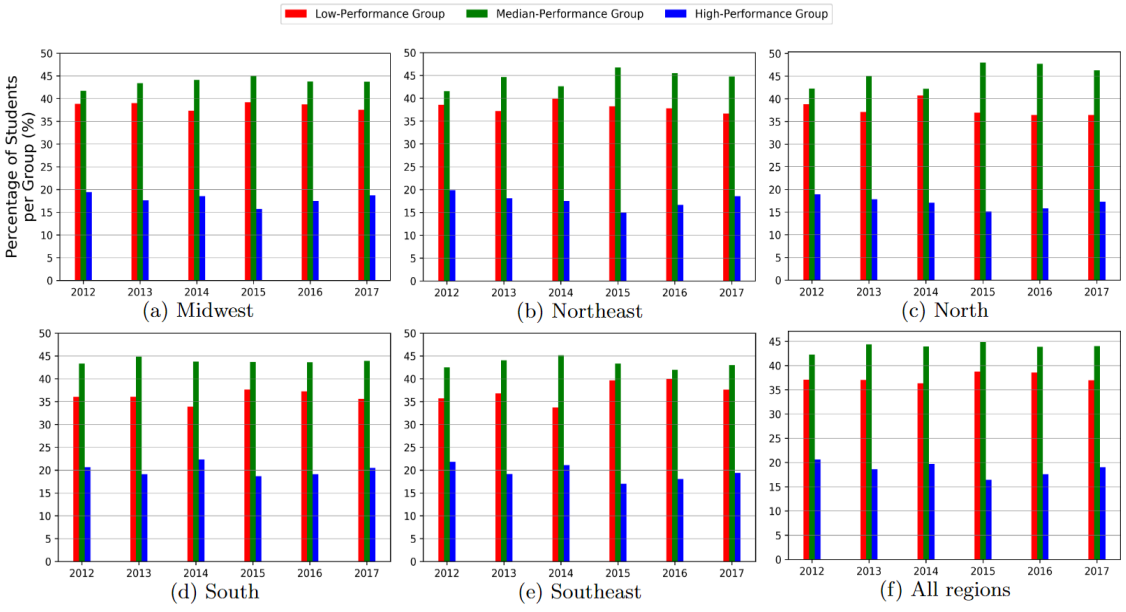
O grupo de Lima et al. (2020) analisou os participantes do ENEM (entre 2012 e 2017), baseado em áreas do conhecimento, tipo de escola e acessibilidade. Os estudantes foram agrupados baseando-se no desempenho em cada área do conhecimento, possibilitando uma compreensão mais aprofundada deste item em cada região do Brasil no ENEM.

O K-means foi usado para geração de grupos de estudantes. Ficou definido um total de 3 clusters, conforme indicado pelo método do cotovelo, cujos grupos foram identificados com relação a um baixo, médio ou alto desempenho. Os atributos usados para o agrupamento foram apenas as quatro notas das áreas de conhecimento e a nota da redação, com a finalidade de definir os grupos com baixo, médio ou alto desempenho e analisar qual área tem maior impacto no resultado final obtido pelos estudantes no exame.

Os atributos considerados foram: residência por estado, idade, sexo, estado civil, cor/raça, tipo da escola, Unidade Federativa 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.

A Figura 3.4 mostra em porcentagem a composição dos grupos de desempenho formados por estudantes de cada região entre anos avaliados. Os resultados mostraram que o desempenho geral aumentou durante os anos, e as regiões Sul e Sudeste obtiveram os melhores desempenhos. No Norte e Nordeste houve crescimento gradual do desempenho durante o intervalo de anos considerado no estudo.

Figura 3.4 - Evolução do percentual de estudantes em cada grupo de desempenho.



Fonte: LIMA et al. (2020, p.10).

Os grupos de baixo e médio desempenho de todas as regiões têm, por maioria, estudantes das escolas públicas. Somente 10% são de escolas privadas. Mas, no grupo de alto desempenho, 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 que, 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 alto desempenho, entre os deficientes. As escolas privadas, em 2017, eram predominantes para os estudantes deficientes.

Quanto ao desempenho por áreas de conhecimento, os resultados mostram que a redação apresenta valores mais discrepantes entre os grupos de desempenho. Com isso, os autores inferem que ter uma nota alta na redação pode ser um fator determinante para o desempenho do ENEM.

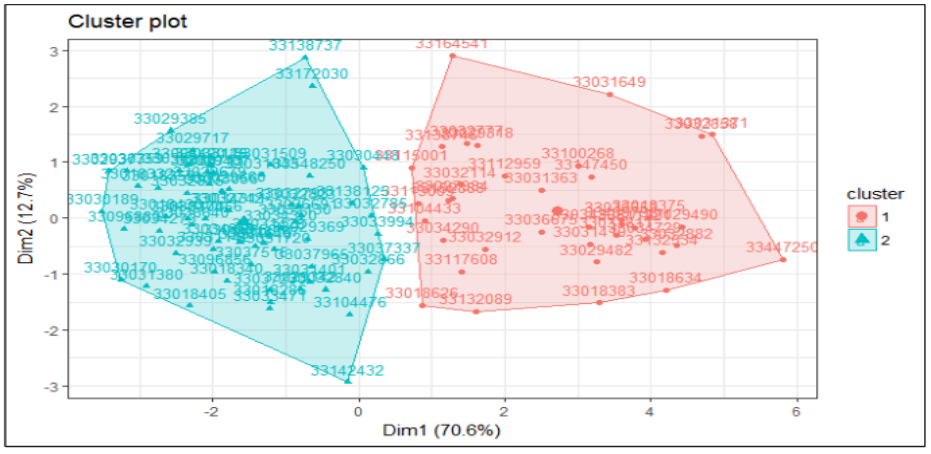
3.6. Desempenho das escolas públicas e privadas: agrupamentos com K-means

Ainda que não tenham sido desenvolvidos no contexto atual da pandemia, alguns trabalhos podem ser destacados devido às análises com dados do ENEM, por meio de métodos de AM não supervisionado e por considerarem questões associadas ao desempenho de estudantes.

Os autores Leoni e Sampaio (2017) buscaram identificar grupos de estudantes por tipos de escolas (públicas e privadas), com desempenhos similares por meio de variáveis indicadoras do Enem. Essas variáveis são descritas brevemente como: (i) indicador de proficiência média por escolas: corresponde às notas médias obtidas por cada escola em cada área de conhecimento do ENEM; (ii) indicadores contextuais: se referem à formação docente e às taxas de rendimento escolar de aprovação; (iii) indicador para taxas de participação: envolve a taxa de participação de estudantes da escola que fizeram o ENEM e não zeraram em nenhuma das provas. Consideraram como recorte as escolas da região sul fluminense tendo em vista que estas apresentaram o melhores resultados quanto aos indicadores PIB per capita de 2014 e IDH de 2010.

Para definir o número ideal de clusters foram usados o método elbow, coeficiente de silhueta e estatística *Gap* (lacuna) (TIBSHIRANI; WALTHER; HASTIE, 2001). Todos indicaram o valor ideal de 2 clusters. A Figura 3.5 ilustra os grupos formados a partir da alocação das escolas.

Figura 3.5 - Grupos formados após a aplicação do K-means.



Fonte: LEONI e SAMPAIO (2017, p.7).

O Cluster 1 resultou em 38 escolas (maior desempenho) e o Cluster 2 em 65 escolas (menor desempenho). Em relação aos indicadores de taxas de participação, reprovação e proficiência média por escolas, o Cluster 1 apresentou níveis superiores em relação ao Cluster 2, já o índice de formação docente foi equivalente em ambos. Quanto ao desempenho apresentado pelas escolas por dependência administrativa, o Cluster 1 é basicamente caracterizado por escolas em sua maioria privadas, com indicador socioeconômico alto ou muito alto e porte não maior que 60 alunos.

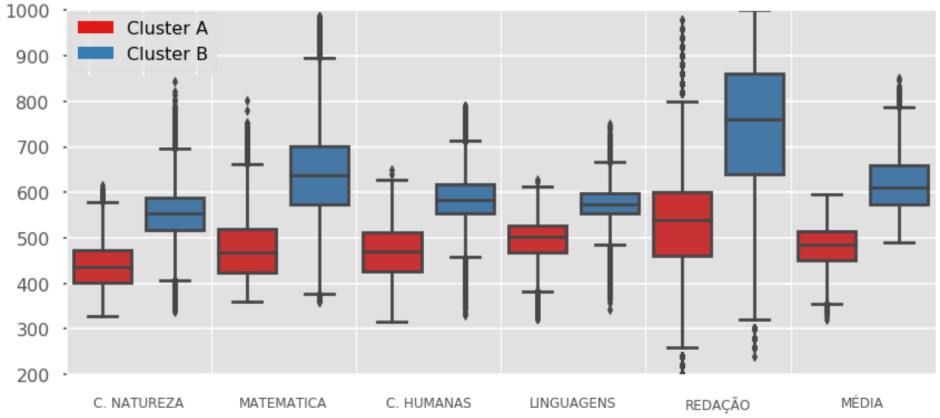
Os autores sugerem realizar agrupamentos considerando as escolas particulares com desempenho similar às algumas escolas públicas, que em geral, não apresentam bom desempenho no ENEM e, com isso, identificar fatores críticos de sucesso que possam ser determinantes para o bom desempenho das escolas no ENEM.

3.7. Identificação de Desigualdades Sociais a partir do desempenho no ENEM

O trabalho de Silva et al. (2020) buscou identificar grupos de estudantes a partir das notas obtidas em cada área de conhecimento e, a partir desses grupos formados, caracterizá-los e identificar correlações entre variáveis referentes a aspectos socioeconômicos e de notas de estudantes de Minas Gerais no ENEM de 2019. Para isso, dois métodos de aprendizado não supervisionados foram usados: agrupamento e regras de associação.

Para o agrupamento foi usado o método K-means, onde chegaram à definição de dois clusters, avaliados pelo método de silhueta, cujo melhor valor do coeficiente foi próximo a 0,42 para o cenário com dois clusters. Os rótulos "Cluster A" (piores notas) e "Cluster B" (melhores notas) foram adicionados à base de dados para identificar as instâncias. Os estudantes do Cluster B apresentaram desempenho superior em todas as provas por área de conhecimento. A Figura 3.6 apresenta a distribuição de notas em cada área de conhecimento e a nota média geral nos dois clusters.

Figura 3.6 - Distribuição de notas por área de conhecimento.



Fonte: SILVA et al. (2020, p.5).

Já para a mineração de regras de associação foi usado o algoritmo Apriori, com o objetivo de mostrar associações de características presentes nos clusters a partir dos atributos

socioeconômicos. As variáveis socioeconômicas, o tipo de administração de escola e a renda familiar foram consideradas as mais preponderantes nas regras obtidas. No geral, os atributos considerados mais relevantes foram os que evidenciam as desigualdades, tais como: nota média por administração da escola; nota por escolaridade da mãe; e autodeclaração de raça por nota média. As principais regras obtidas foram discutidas em três cenários. A Figura 3.7 contém uma amostra de 8 entre as 80 regras de associação obtidas.

Figura 3.7 - Regras de associação obtidas com o Apriori.

#	Antecedente → Consequente	Suporte	Conf.	*Lift*
	(a) Regras considerando toda a base			
1	mediaNota=585,6 - 718,21 → Cluster=B	28,5%	99,9%	2,152
2	mediaNota=452.99 - 585.6 → Cluster=A	39.4%	72,0%	1,344
3	Cluster=A → ADM_ESC=Estadual	50,6%	94,5%	1,227
4	RACA=NãoBranco, ADM_ESC=Estadual → Cluster=A	34,2%	70,1%	1,309
5	EstudoMae=Medio_Inc, Classe=E → Cluster=A	22,4%	73,6%	1,373
6	TemPC=Sim. Cluster=B → EstudoMae=Medio_Comp.	28,0%	75,7%	1,334
7	Cluster=B → TemPC=Sim	37,0%	79,6%	1,222
8	TemPC=Não → Cluster=A	25,4%	72,9%	1,361
	(b) Regras considerando apenas o Cluster A			
1	RACA=NãoBranco → Classe=E	50%	75%	1,051
2	Classe=E → RACA=Não-Branco	50%	75%	1,051
3	EstudoMae=Medio_Comp. → mediaNota=585,6 - 718,21	34%	78%	1,062
4	mediaNota=585,6 - 718,21 → Classe=E	20%	78%	1,099
	(c) Regras considerando apenas o Cluster B			
1	NotaMedia=452,99 - 585,6 → ADM_ESC=Estadual	27%	84%	1,470
2	ADM_ESC=Particular → NotaMedia=585,6 - 781,21	25%	75%	1,226

Fonte: SILVA et al. (2020, p.7).

No primeiro cenário, considerando todo o conjunto de dados, obteve-se que: estudantes com médias entre 585,6 a 718,21 estão no Cluster B; estudantes com média entre 452,99 e 585,6, encontram-se no Cluster A; com 94,5% de confiança, estudantes do Cluster A vem de escolas estaduais, sendo 70,1% destes declarantes negros, pardos, amarelos ou indígenas (não-brancos). Além disso, o Cluster A, com confiança de 73,6%, contém estudantes cujas mães não concluíram o ensino médio; estudantes do Cluster B que possuem PC em casa têm mãe que, ao menos, completou o ensino médio e 79,9% possuem computador.

No segundo cenário, com apenas dados de estudantes do Cluster A, as regras mostram que: há uma relação forte da autodeclaração da raça e a classe econômica, indicando que 75% dos autodeclarados não-brancos são da classe E e vice-versa; 78% dos estudantes do Cluster A cujas mães concluíram pelo menos o ensino médio têm nota média entre 585,6 e 718,21(as maiores notas do cluster), sendo que isso ocorre em 34% dos registros do Cluster A.

No terceiro cenário, considerando o Cluster B: há variação maior das faixas de notas médias, onde 27% dos estudantes apresentaram nota média na faixa baixa (452,99 a 585,6) e estudam em escolas estaduais; 84% dos estudantes com baixo desempenho são de escolas estaduais, enquanto 75% dos estudantes de escolas particulares atingiram desempenho na faixa entre 585,6 e 781,21 de nota média.

3.8. Associações em dados dos inscritos do ENEM

O estudo de Gomes et al. (2017) objetivou identificar relações entre o desempenho do estudante na prova de Matemática e o seu local de residência (interior/capital), renda familiar, escola onde cursou os ensinos médio e o fundamental. Foram usados dados do questionário socioeconômico do ENEM 2013 e 2014, considerando a região Nordeste e o estado de Pernambuco como recorte. Considerando os dados pessoais, socioeconômicos, de notas e o ano de realização de cada exame, os resultados obtidos pelos autores 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 a nota média em matemática.

Foi usado o algoritmo Apriori para identificação das regras de associação. Como exemplo dessas regras, destacam-se: (i) estudantes que cursaram o ensino médio apenas em escolas públicas apresentam forte tendência a serem oriundos de famílias com até dois salários mínimos; (ii) estudantes que realizaram seus estudos na modalidade regular tem tendência a residirem na zona urbana, já aqueles residentes no interior tem tendência a pertencerem a famílias com renda de até dois salários mínimos; (iii) estudantes com nota entre 400 e 500 pontos em Ciências da Natureza e em Matemática tem forte tendência a terem renda familiar de até dois salários mínimos.

Em geral, as regras obtidas para o estado de Pernambuco são muito similares àquelas identificadas em nível Nordeste.

3.9. Eficácia escolar e características familiares em tempos de pandemia

O estudo de De Moraes et al. (2021) traz uma visão complementar para trabalhos relacionados, sob o conceito de eficácia escolar, porém sem uso de algum método de AM. Os autores fizeram uma análise de dados sobre fatores familiares, escolares e de desempenho de estudantes na área de conhecimento de Matemática e Suas Tecnologias em meio à pandemia. Avaliaram as condições do Brasil, um dos países que mais sofreram com as consequências da pandemia, frente às desigualdades sociais e aos sérios problemas educacionais já presentes antes da pandemia. Os atributos usados foram obtidos do ENEM e do Censo Escolar de 2017. A intenção dos autores com o uso dos dados de 2017 é para que houvesse uma projeção dos resultados no período da pandemia.

Os dados usados na análise foram agrupados considerando as cinco regiões brasileiras, com apenas estudantes pertencentes a municípios com população entre 50 e 500 mil habitantes. Entre as variáveis usadas no estudo, os autores consideram aquelas que são usualmente citadas pela literatura em avaliação educacional, a saber: escolaridade da mãe e renda familiar, como fatores familiares; a escola ter biblioteca e sala de leitura, como fatores de infraestrutura escolar; as notas em matemática, como fator de desempenho escolar.

Os resultados mostraram que estudantes oriundos de famílias de renda superior alta possuem notas melhores em relação aos estudantes de renda inferior. A perspectiva dos autores,

diante do primeiro ano da pandemia, foi de um cenário desanimador para jovens de baixa renda, tendo em vista, por exemplo, a falta de acesso à internet e a outros meios tecnológicos, dificultando mais ainda a probabilidade de um bom desempenho no exame.

3.10. SÍNTESE SOBRE OS TRABALHOS RELACIONADOS

Algumas informações importantes sobre os Trabalhos Relacionados (TR) foram extraídas e organizadas no Quadro 3.1, a saber: (i) identificação do TR e a edição dos dados do ENEM considerada; (ii) o tipo de análise realizada e quais técnicas/algoritmos de AM não supervisionados foram usados; (iii) a lista de atributos considerados relevantes ao desempenho de estudantes; (iv) sendo o caso, a identificação de perfis/grupos de estudantes; (v) e o tipo de associações de características entre os atributos, quando for o caso.

Quadro 3.1 - Síntese sobre os trabalhos relacionados à temática da pesquisa.

TR/ edição dos dados	**Tipo de análise/ Algoritmo(s) de AM**	**Atributos relevantes identificados**	**Perfis de estudantes**	**Associações de características**
- Weber Neto et al. (2022a) - 2019 e 2020	- Exploratória	SG_UF_ESC, TP_PRESENCA (nas provas), TP_DEPENDENCIA_ADM_ESC, TP_ESCOLA, TP_COR_RACA, RENDA_FAMILIAR, NOTA_MEDIA.	- - -	- - -
- Weber Neto et al. (2022b) - 2019 e 2020	- Exploratória	TP_SEXO, TP_DEPENDENCIA_ADM_ESC, TP_ESCOLA, TP_COR_RACA, RENDA_FAMILIAR, Possui Computadores, Possui Celulares, Acesso à Internet, Escolaridade Pai, Escolaridade Mãe, NOTA_MEDIA.	- - -	- - -
- Weber Neto et al. (2023) - 2017, 2018, 2019 - 2020 e 2021	- Exploratória	SG_UF_ESC, TP_PRESENCA (nas provas), TP_DEPENDENCIA_ADM_ESC, TP_ESCOLA, RENDA_FAMILIAR, NOTA_MEDIA.	- - -	- - -
- Pereira Junior et al. (2021) - 2020 (gerou dataset próprio)	- Análise de clusters/ SOM e K-means	foco_para_ead, sanar_duvidas_ead, nivel_perfil_aluno_ead, lugar_calmo, uso_software_educativo, desvantagem_suspensao_aulas, sexo, animo_para_atividades_ead, nivel_academico, situacao_emprego, nivel_estimulo_negativo_estresse, nivel_perfil_aluno.	Favoráveis / Não favoráveis ao ensino remoto.	- - -
- Lima et al. (2020) - 2012 a 2017	- Análise de clusters/ K-means	TP_PRESENCA (nas provas), NU_IDADE, TP_SEXO, TP_ESTADO_CIVIL, TP_COR_RACA, TP_ESCOLA, SG_UF_ESC, TP_DEPENDENCIA_ADM_ESC, TP_LOCALIZACAO_ESC, NOTAS (de cada área), Atributos relacionados a tipos de deficiências, NOTA_MEDIA.	Baixo / Médio / Alto desempenho.	- - -

- Leoni e Sampaio (2017) - 2015	- Análise de clusters/ K-means	TP_ESCOLA, TAXA_PARTICIPACAO, IND_FORM_DOCENTE, TAXA_APROVACAO, MEDIA_CH, MEDIA_CN, MEDIA_LC, MEDIA_MT, MEDIA_RED.	Pior / Melhor desempenho por tipo de escola.	- - -
- Silva et al. (2020) - 2019	- Análise de clusters/ K-means. - Análise de regras de associação / Apriori	TP_COR_RACA, TemPC, RENDA_FAMILIAR, Escolaridade Mãe, NOTAS (de cada área), NOTA_MEDIA.	Melhor / Pior desempenho.	Atributos socioeconômicos.
- Gomes et al. (2017) - 2013 e 2014	- Análise de regras de associação/ Apriori	NU_IDADE, TP_ESCOLA, TP_LOCALIZACAO_ESC, TP_ENSINO, RENDA_FAMILIAR, NOTAS (de cada área).	- - -	Atributos socioeconômicos, de escola e de local de residência.
- De Moraes et al. (2021) - 2017	- Exploratória	Escolaridade Mãe, Renda familiar, Escola possui biblioteca, Escola possui sala de leitura, Faixa etária, Nota em Matemática.	- - -	- - -

Fonte: Elaborado pelo autor.

Observa-se que, em relação ao uso dos dados mais recentes do ENEM (atualmente disponíveis até 2022), pouco tem sido explorado no contexto da pandemia. Apenas os trabalhos de Weber Neto et al. (2022a, 2022b, 2023) consideram esse contexto pandêmico com os dados do ENEM e o trabalho de Pereira Junior et al. (2021), porém considerando a produção de dados próprios, não fazendo uso de dados do ENEM. Os trabalhos descritos, em geral, consideraram dados regionais, alguns chegaram a criar modelos de AM não supervisionados, outros ficaram restritos a análises mais exploratórias.

O cenário gerado pela pandemia durante esses últimos anos de aplicação da prova do ENEM pode gerar novas questões relacionadas tanto ao desempenho quanto a outros fatores que devem ser investigados com mais detalhes.

Os trabalhos relacionados envolveram a investigação do desempenho dos estudantes utilizando-se de tarefas e estratégias variadas. Em relação aos algoritmos de AM não supervisionado, o K-means (análise de clusters) foi o mais utilizado, com apenas um trabalho que também aplicou o algoritmo SOM, muito usado para agrupar dados e realizar redução de dimensionalidade. O algoritmo Apriori foi aplicado por dois trabalhos para análise de regras de associação. Os demais trabalhos realizaram análises exploratórias e comparativas dos dados, que fazem parte das atividades ligadas à análise de dados.

Os trabalhos analisados geraram novos atributos, além daqueles que foram usados e nomeados neste estudo como atributos relevantes. Um exemplo de atributo gerado é a ‘nota média’, obtido a partir da média aritmética das notas das provas objetivas e da redação. Buscou-se na literatura algo que pudesse embasar o que seria, por exemplo, um bom ou mau desempenho do estudante no ENEM. Os trabalhos, em geral, não justificam, com base na literatura, como chegam a essa divisão para definir o desempenho dos estudantes, sendo uma abordagem mais empírica dos autores.

4. ABORDAGEM PROPOSTA

A RSL proporcionou uma visão mais abrangente sobre o que os estudos associados à análise de desempenho de estudantes no ENEM contribuíram em relação aos seguintes aspectos (DUTRA; FIRMINO JUNIOR; SOUZA, 2023): (i) elaboração de modelos preditivos de AM de modo a auxiliar educadores a, por exemplo, antever o desempenho, mitigar o problema de evasão estudantil e acompanhar o estudante durante o seu percurso escolar; (ii) análise de fatores principais que justificam o desempenho; (iii) identificação de perfis de estudantes que possam apoiar o entendimento da situação educacional no Brasil; e (iv) integração de dados do ENEM com outras bases educacionais para enriquecimento de informações.

Particularmente, em relação a análise de fatores principais que justificam o desempenho, a RSL apontou que os principais atributos identificados nos estudos remetem às questões socioeconômicas, sendo mais destacados, nesta ordem, os seguintes: renda familiar mensal, idade, sexo e raça. Com destaque posterior a esses atributos principais aparecem: o nível de educação dos pais, tipo de escola, localização da escola e outros atributos relativos à estrutura física e pedagógica das escolas.

Para chegar a essa lista de atributos mais relevantes, este autor e os demais realizaram um processo de integração de dados dos resultados da RSL. Os atributos foram extraídos de cada artigo conforme haviam sido descritos, sendo posteriormente agrupados conforme a nomenclatura (nome do atributo) mais recente dos microdados. Após a fusão dos atributos, esses foram contabilizados em relação à quantidade de ocorrências de cada um, estabelecendo-se, assim, a lista de atributos mais relevantes.

Após a RSL, em novas pesquisas na literatura, não foi identificado nenhum trabalho com foco na análise dos dados associados aos hábitos de estudos durante a pandemia. A partir dessa constatação, este trabalho trouxe à tona esta lacuna e questionou como tais hábitos poderiam impactar no desempenho dos estudantes durante a pandemia além dos já normalmente percebidos pela comunidade de pesquisa?

Isso se transcreveu nas duas questões de pesquisa especificadas no Capítulo 1 e relembradas a seguir:

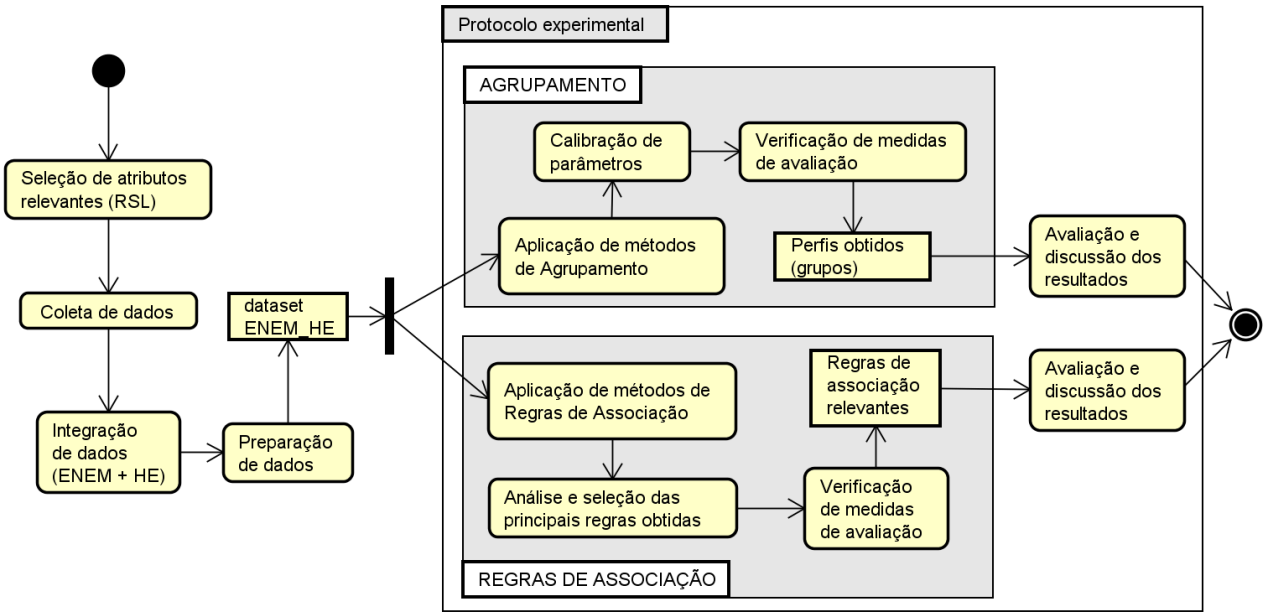
QP1: Considerando dados do ENEM e de hábitos de estudo durante a pandemia da COVID-19, como identificar perfis associados ao desempenho de estudantes para o referido exame?

e

QP2: Quais regras de associações de características podem ajudar a ratificar ou entender melhor os perfis de desempenho de estudantes no exame do ENEM durante a pandemia?

A abordagem proposta neste trabalho se baseia nas etapas definidas pelo processo CRISP-EDM, com adaptações ao contexto para resposta às questões de pesquisa formuladas. A Figura 4.1 mostra a visão geral da abordagem proposta.

Figura 4.1 - Etapas da abordagem proposta.



Fonte: Elaborado pelo autor.

O objetivo geral da abordagem proposta é identificar perfis de estudantes a partir de atributos relevantes ao entendimento do desempenho do estudante com base nos microdados comuns obtidos no ENEM e na agregação de dados relacionados aos hábitos de estudo durante a pandemia. Para isso, após a seleção dos atributos relevantes da RSL, são coletados e integrados os dois datasets (ENEM e hábitos de estudo), considerando apenas estudantes que responderam ao questionário de hábitos de estudo. Em seguida, na preparação dos dados são realizadas atividades de: eliminação de instâncias duplicadas; criação de variável; padronização de nomes, normalização e transformação de variáveis. O dataset gerado (ENEM_HE) é usado em cada uma das tarefas de AM não supervisionado. Métodos de agrupamento e de regras de associação são utilizados em um protocolo experimental para a identificação dos perfis e das associações entre as características mais relevantes ao entendimento do desempenho do estudante no período referido. Por fim, os resultados são avaliados e discutidos. As questões de pesquisa estão diretamente relacionadas às duas tarefas de modelagem do AM não supervisionados e aos experimentos definidos para a avaliação da pesquisa.

A seguir, cada etapa da abordagem é descrita e as atividades realizadas são explicadas.

4.1. IDENTIFICAÇÃO DE ATRIBUTOS MAIS RELEVANTES

Os microdados[9] do ENEM da edição 2022 são compostos por 76 atributos, organizados e distribuídos por dados sobre o(a): participante, escola, local de aplicação da prova, prova objetiva e questionário socioeconômico. O arquivo completo referente aos microdados contém 1,45 *gigabytes* e possui 2.467.086 instâncias, que se referem aos estudantes e ao resultado obtido na prova por cada um em nível nacional.

[9] Os microdados e o dicionário de dados estão disponíveis em: https://download.inep.gov.br/microdados/microdados_enem_2022.zip . Acesso em: 14 dez. 2023.

A lista de atributos relevantes identificados na RSL foram categorizadas em (DUTRA, et al., 2023): dados socioeconômicos, de localização, de notas em todas as áreas de conhecimento e de perfil das escolas. Os autores da RSL atribuíram uma contagem para cada atributo, a partir do número de trabalhos que os identificaram como relevantes. No total, 17 atributos foram identificados como relevantes. O Quadro 4.1 mostra como ficou a distribuição dos atributos, com a quantidade de ocorrências de trabalhos que os identificaram como relevantes ao desempenho, por cada categoria de dados.

Quadro 4.1 - Lista de atributos relevantes por categoria de dados.

Categoria	**Atributos relevantes**
Socioeconômicos	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).
Localização	Região/uf/município da escola (8), região/uf/município de domicílio do estudante (6) e região de nascimento (2).
Notas	notas por área (13), nota da redação (8), nota média (8), nota em língua estrangeira (5).
Perfil das escolas	tipo de escola (9) e tipo administrativo da escola (3).

Fonte: Adaptado de DUTRA et al. (2023).

A seleção de atributos relevantes para compor o dataset ENEM_HE foi avaliada mediante a importância (maior quantidade de ocorrências de atributos) em relação ao desempenho e à relação de hábitos de estudo durante as aulas remotas no período da pandemia. O atributos selecionados para este fim, então, foram: tp_faixa_etaria (idade), tp_sexo (sexo), tp_cor_raca (raça); Q001 (escolaridade do pai), Q002 (escolaridade da mãe), Q006 (renda familiar), Q022 (celular na residência), Q024 (computador na residência), Q025 (internet na residência); notas por área; nota da redação; nota média.

As questões Q022, Q024 e Q025 estão incluídas no questionário socioeconômico e, por terem uma forte relação com o uso de meios de acesso às aulas remotas, foram selecionadas para integrar também o dataset ENEM_HE.

As notas de cada área de conhecimento e da redação são incluídas, pois, a partir delas gerou-se a nota média, que determina o desempenho do estudante na prova do ENEM.

Avaliou-se a inclusão de outros atributos relevantes como, por exemplo, tp_dependencia_adm_esc (tipo administrativo da escola) e cod_uf_esc (Região/uf/município da escola), no entanto, eles foram retirados por conterem muitos valores nulos. Para o atributo tp_dependencia_adm_esc citado, 72,27% dos valores são nulos. Quanto ao atributo co_uf_esc (UF de localização da escola), não seria necessária a sua inclusão, por ser considerado o escopo de dados em nível nacional neste trabalho.

Para seleção dos atributos de hábitos de estudo mais representativos ao contexto desta pesquisa, foi realizada uma análise de correlação em relação ao desempenho dos estudantes, tomando como referência a nota média. Observou-se correlação forte positiva da nota média com

os atributos de hábitos de estudos dos grupos temáticos de práticas e gestão de tempo para as atividades durante a pandemia. A correlação forte negativa com a nota média envolveu atributos dos grupos temáticos de situação de matrícula escolar, ajuda de terceiros e dificuldades de infraestrutura durante a pandemia.

Foi observado que haveria pouca relevância em usar atributos colocados como subquestões dos hábitos de estudo, que detalham especificamente alguns atributos com respostas mais gerais, sendo então desconsiderados o seu uso. Por exemplo, foi considerado apenas se o estudante recebeu, ou não, ajuda de terceiros, sendo descartados atributos que detalham de quem partiu a ajuda, se foi de pais, irmãos, primos, tios, avós, etc. Do mesmo modo, para as subquestões com respeito aos detalhes de dificuldades de infraestrutura, não sendo considerados atributos que detalham sobre, por exemplo, se a dificuldade foi a falta de um equipamento compartilhado ou com configuração ruim, conexão lenta com internet, material pedagógico insuficiente, local com condições inadequadas, etc. Detalhes de tipos de tecnologias e meios de acesso mais usados pelo estudante para estudar também foram descartados como, por exemplo, rádio, programas de televisão, telefone, tablet, computador, etc.

Esses atributos, que se referem a subquestões, também continham muitos valores nulos, tendo em vista que alguns deles também tratam de questões de múltipla escolha.

4.2. COLETA E INTEGRAÇÃO DE DADOS

Para este trabalho, estão sendo utilizados dois conjuntos de dados do INEP (2022): (i) microdados do ENEM de 2022, que fornecem as notas de cada área do conhecimento e redação por estudante, além de dados socioeconômicos e sobre escolas; (ii) microdados de hábitos de estudo, que contém respostas de parte dos estudantes de como foi sua rotina de estudos e preparação para o ENEM durante a pandemia.

Após suas coletas, a integração dos conjuntos de dados foi realizada por meio de uma variável em comum, a NU_INSCRICAO. Os valores contidos nesta variável se referem a uma máscara e não ao número de inscrição real do estudante no ENEM, no entanto, devido à correspondência com os microdados de hábitos de estudo, verifica-se que, de fato, esta variável se refere às mesmas instâncias de estudantes que realizaram a prova somente na edição de 2022. Cabe ressaltar que os valores presentes no atributo NU_INSCRICAO na edição de 2022 não tem relação com números de inscrições de quaisquer edições anteriores do exame (INEP, 2022).

O conjunto de dados gerado apresenta somente estudantes que optaram por responder ao questionário de hábitos de estudo. O filtro foi realizado por meio da variável TP_RESPOSTA com valor igual a 1 (responderam o questionário). Obteve-se o total de 928.564 (37,6%) estudantes.

4.3. PREPARAÇÃO DE DADOS

Conforme explicado anteriormente, nem todos os estudantes responderam por completo o questionário de hábitos de estudo, com isso o conjunto de dados continha muitos campos nulos. Desse modo, com base em um um filtro para se obter todos os estudantes que responderam o

questionário de hábitos de estudo por completo, foram excluídas as instâncias que não tinham todas as colunas preenchidas. Havia duas instâncias duplicadas (NU_INSCRICAO repetido) que também foram eliminadas.

Estudantes reprovados, que zeraram em pelo menos uma das provas ou redação, foram removidos. Neste caso, a nota média (aritmética) foi calculada para todos aqueles estudantes com notas por área de conhecimento e redação maiores que 0 (zero). Uma nova variável, NOTA_MEDIA, contendo os valores da média de cada estudante foi acrescentada ao dataset.

Conforme explicado na Seção 4.1, para os dados referentes ao ENEM, os atributos relevantes da RSL mantidos foram: 'TP_FAIXA_ETARIA', 'TP_SEXO', 'TP_COR_RACA', 'Q001', 'Q002' (escolaridade pai e mãe), 'Q006' (renda), 'Q022' (celular), 'Q024' (quantidade de computadores), 'Q025' (acesso à internet).

Após o processo de integração dos microdados, alguns atributos referentes ao questionário socioeconômico do ENEM (descritos com a terminação "_x", por exemplo, Q001_x) foram renomeados para que se tornassem mais representativos e para diferenciá-los dos atributos de hábitos de estudo (descritos com a terminação "_y", por exemplo, Q001_y). Para os atributos referente aos microdados padrão do ENEM definiu-se as assinaturas:

- Q001 → NIVEL_ESC_PAI (nível de escolaridade do pai/homem responsável);
- Q002 → NIVEL_ESC_MAE (nível de escolaridade do mãe/mulher responsável);
- Q006 → RENDA_FAMILIAR (renda mensal da família);
- Q022 → NU_CELULAR (total de telefones celular na residência);
- Q024 → NU_COMPUTADOR (total de computadores na residência);
- Q025 → ACESSO_INTERNET (tem acesso à internet na residência? [sim=A/não=B]).

Para os dados de hábitos de estudo, as novas descrições das variáveis seguem uma padronização para que fosse identificado o grupo temático ao qual pertence cada variável. Alguns exemplos são informados a seguir:

- 'Q004_y' → 'APR_MATR_APREND_PANDEMIA'
    - GT: situação de MATRricula escolar e perceção da própria APRENDizagem
    - Descrição: Como o estudante percebe o seu processo de aprendizagem durante a PANDEMIA.
- 'Q021_y' → 'GEST_TEMP_PONTUAL_AULA_ONLINE'
    - GT: GESTão do TEMPo e planejamento de estudos
    - Descrição: PONTualidade do estudante em entrar nas AULAs ONLINE por videoconferência sem atraso.
- 'Q029' → 'DIF_INFRA_PANDEMIA'
    - GT: DIFiculdades de INFRAestrutura
    - Descrição: O estudante teve dificuldades de infraestrutura para estudar ou manter-se informado em 2021 durante a PANDEMIA

As variáveis categóricas tiveram seus valores transformadas em numéricos 0 ou 1 (*one hot encoding*) e apenas três variáveis numéricas (TP_FAIXA_ETARIA; TP_COR_RACA;

NOTA_MEDIA) foram normalizadas por meio do método *MinMaxScaler*[10], cujos valores ficaram dimensionados no intervalo entre zero e um.

Algumas variáveis, principalmente as numéricas, tiveram os seus valores discretizados em níveis de categorias mais reduzidas, de maneira a viabilizar que os algoritmos tivessem resultados melhores e mais compreensíveis. Para isso, foi reduzida a granularidade (amplitude) dos valores dos atributos, tendo em vista que, por exemplo, para as regras de associação, estes compunham os itens do dataset. Do ponto de vista da interpretação dos resultados, para o agrupamento, a redução da granularidade facilita a elucidação da caracterização dos perfis de estudantes, e para as regras de associação há a redução da complexidade de itens candidatos.

A Figura 4.2 mostra um fragmento do dataset ENEM_HE resultante da preparação dos dados, que ficou com 332.793 (46,69%) de instâncias e 136 atributos, com valores booleanos (0 - falso; 1 - verdadeiro) para presença do item (questão respondida pelo estudante).

Figura 4.2 - Fragmento do dataset ENEM_HE.

	TP_FAIXA_ETARIA=faixaAdultosIdosos	TP_FAIXA_ETARIA=faixaMédio	TP_FAIXA_ETARIA=faixaProfissional	TP_SEXO=femino	...	AUTOAV_PREPARACAO_APRENDIZ=poucoPreparado	AUTOAV_PREPARACAO_APRENDIZ=totalmentePreparado	NOTA_MEDIA=baixo
0	0	0	1	1	...	1	0	1
1	0	0	1	1	...	1	0	1
2	0	0	1	0	...	0	0	0
3	0	0	1	1	...	0	0	0
4	0	0	1	1	...	1	0	1
...	...	...	...	...	...	...	...	...
332788	0	1	0	0	...	0	0	1
332789	0	1	0	1	...	1	0	0
332790	0	1	0	0	...	0	0	0
332791	0	0	1	0	...	1	0	0
332792	0	1	0	0	...	1	0	0

332793 rows × 136 columns

Fonte: Elaborado pelo autor.

O Anexo A mostra o dicionário de dados (Quadro 4.3) do conjunto ENEM_HE usado como entrada para a aplicação dos algoritmos de AM não supervisionados. O Apêndice A apresenta os atributos e os respectivos valores categorizados (Quadro 4.4), que favorece um melhor entendimento sobre os dados e a forma como são apresentados neste trabalho. A lista final dos atributos está organizada por dados relevantes obtidos por meio da RSL e de hábitos de estudo, distribuídos nos grupos temáticos (monitoramento da pesquisa; matrícula escolar; gestão do tempo e planejamento de estudos; práticas de estudo e pesquisa; tecnologias e tipo de acesso/problemas na rotina; dificuldades de infraestrutura/ajuda de terceiros).

4.4. PROTOCOLO EXPERIMENTAL

O protocolo experimental usado no desenvolvimento da abordagem possui duas etapas que objetivam apoiar as respostas à QP1, para identificar os perfis de estudantes, e fatores mais impactantes ao desempenho, e à QP2, para encontrar associações que ratifiquem os perfis de estudantes de modo geral e individualmente por cada grupo. Para isso, as seções seguintes descrevem os principais aspectos considerados na aplicação dos algoritmos escolhidos de agrupamento e de regras de associação.

[10] Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler. Acesso em 25 ago. 2023.

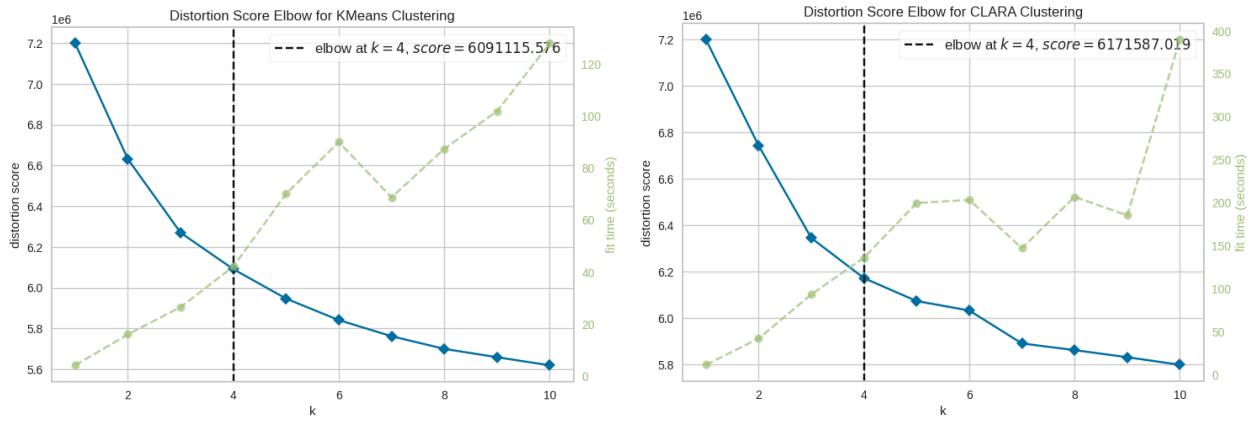
4.4.1. Algoritmos de agrupamento

Na abordagem, algoritmos partitivos e baseados em densidade são usados na tarefa de agrupamento.

No caso dos algoritmos partitivos, o K-means foi escolhido por ser o mais aplicado na literatura e, principalmente, por ser escalável para conjuntos de dados de alta dimensionalidade, como é o caso do dataset ENEM_HE. Outro fato importante é que ele é mais rápido para grandes conjuntos de dados (HAN; KAMBER; PEI, 2012). A definição dos medóides iniciais é uma das limitações do K-means. O algoritmo CLARA foi usado como alternativa a ser verificada tendo em vista sua característica de melhorar a eficiência de obtenção dos medóides, porém as análises dos resultados dos algoritmos foram realizadas de maneira independente. Internamente o algoritmo CLARA usa outro algoritmo partitivo, o K-medoides, onde obtém amostras aleatórias de instâncias e recupera os melhores medóides a cada iteração realizada pelo K-medoides. O uso desses algoritmos permitiu avaliar uma possível similaridade e dissimilaridades entre os grupos formados, no entanto convergem para resultados semelhantes, não destacando melhorias significativas. O K-means foi avaliado como aquele que obteve um resultado na identificação dos perfis de estudantes mais consistente com relação aos perfis de desempenho. Para identificar o número ideal de clusters a serem formados normalmente é feita a calibração de parâmetros e medidas de avaliação podem ser usadas para auxílio. Exemplos de medidas são aquelas obtidas por meio do método elbow, coeficiente de silhueta, Índice Davies-Bouldin (DBI) e Índice Calinski and Harabasz (ICH).

Para definição do número de clusters mais adequado à identificação dos perfis, foram testados inicialmente valores de K variando de 2 a 10 por meio do método elbow. A Figura 4.3 ilustra o método elbow indicando K=4 para os algoritmos K-Means e CLARA.

Figura 4.3 - Resultado do número ideal de clusters definido pelo método elbow.



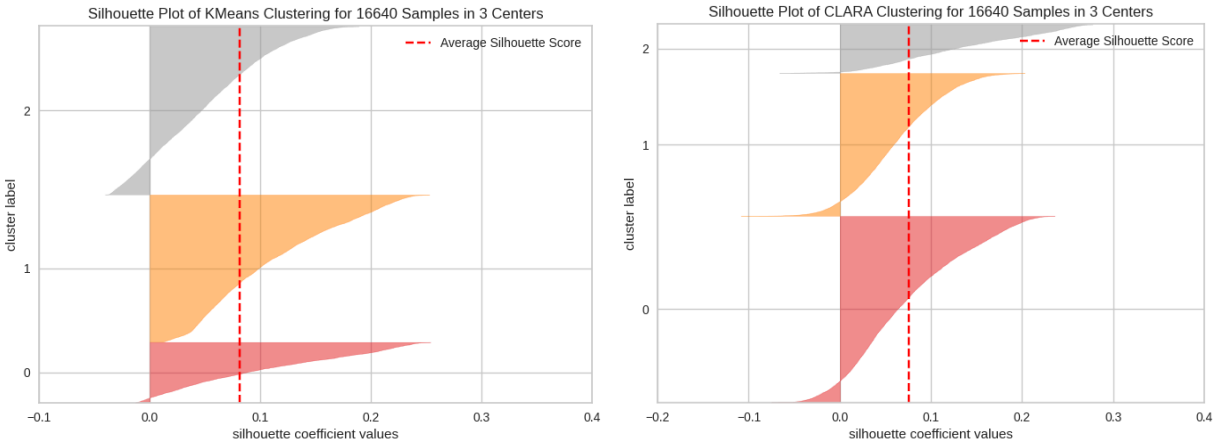
(a) (b)

Fonte: Elaborado pelo autor.

Para uma melhor avaliação da indicação do número de clusters, para cada algoritmo é verificada a qualidade da formação de clusters por meio do coeficiente de silhueta, IHC e IDB. A

Figura 4.4 mostra um gráfico de silhueta cujos coeficientes obtidos foram os mais próximos de 1, para o K=3 considerando os algoritmos K-Means e CLARA.

Figura 4.4 - Resultado da formação de clusters e cálculo do coeficiente de silhueta.



(a) (b)

Fonte: Elaborado pelo autor.

Para os dois algoritmos verificados, observa-se que a qualidade da atribuição das instâncias aos clusters tem uma certa equivalência. A média da silhueta para os dois resultados indica que os grupos formados possuem uma similaridade intracluster não tão bem definida, tendo em vista o valor estar próximo a 0.1. Os grupos formados também possuem muitas instâncias com alocação inconclusiva, principalmente para os clusters com maior quantidade de instâncias.

Para consolidação do número ideal de grupos, as medidas de IHC e IDB foram avaliadas como forma de consolidar a indicação do número ideal de clusters. O Apêndice C contém os resultados obtidos por meio desses índices de qualidade, bem como apresenta detalhes sobre a distribuição de estudantes nos cenários obtidos com os algoritmos divisivos testados.

A escolha do número ideal depende também do entendimento dos dados e do problema a ser solucionado. O dataset possui três categorias de desempenho (baixo, médio e alto), logo com base nessa variável e nos resultados apontados pelas medidas de qualidade, foi definido k=3 como número ideal nas discussões deste experimento com o K-means. Os perfis de estudantes foram avaliados conforme o nível de similaridade presente nos grupos gerados. Para isso, todas as instâncias presentes em cada grupo (perfil) são consideradas na discussão da avaliação.

Os resultados obtidos com o CLARA foram semelhantes ao obtido com o K-means (K=3). Logo, esse resultado reforça a confirmação acerca dos perfis de estudantes encontrados. Outros detalhes dos resultados de cada algoritmo estão disponíveis no Apêndice D e no material suplementar[11].

Após a execução do algoritmo partitivo, uma nova variável (CLUSTER) é inserida para identificar o grupo ao qual cada instância (estudante) pertence. Os grupos gerados são avaliados

[11] Sumário de resultados obtidos com todos os algoritmos na tarefa de agrupamento, disponível em: https://docs.google.com/spreadsheets/d/17YKBQGvxn_sdva_zr18MiRgsDRnyjrCs

com base nas variáveis do ENEM e dos hábitos de estudo, principalmente, em relação ao desempenho obtido no ENEM.

Como os algoritmos partitivos, em geral, são sensíveis a outliers, buscou-se adicionalmente avaliar a presença de perfis outliers. Assim a abordagem inclui esta verificação por meio da aplicação de um método de agrupamento baseado em densidade. Pressupõe-se que algoritmos dessa categoria permitem detectar instâncias (outliers) em relação ao desempenho obtido no ENEM. Nessa abordagem foi usado o algoritmo DBSCAN.

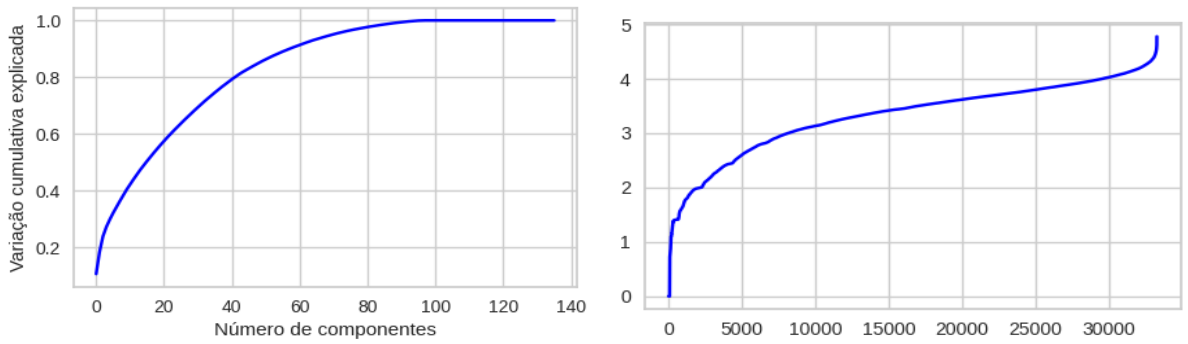
Para identificar os parâmetros ideais para o "melhor resultado", tendo em vista que o próprio algoritmo define a quantidade de clusters a partir dos parâmetros indicados, testes de execução com intervalos de valores são realizados para identificá-los. Ao final, a distribuição de grupos permite avaliar o perfil de desempenho de estudantes identificados como outliers. Os resultados são avaliados principalmente em relação ao desempenho de estudantes (outliers).

Para o DBSCAN, inicialmente, para obtenção dos melhores valores para os parâmetros eps e min_samples, foram feitos vários testes com atribuições de valores variados. As execuções foram realizadas por meio de amostras de 5% a 40%, no máximo, pois acima dessa porcentagem o ambiente não suportou a execução em termos de memória. Por isso dedicou-se a realizar com o conjunto completo em máquina local.

Para determinar o valor de min_samples foi realizado o cálculo de distância média entre cada ponto no dataset com os K vizinhos mais próximos. A literatura indica o dobro da dimensionalidade do dataset, para casos de dataset de alta dimensionalidade. Foi testado de 136 a 272. Como não houve impacto significativo do número de clusters gerados. Assim definiu-se o valor máximo 136 que foi atribuído a min_samples.

Como passo de verificação, foram realizados processamentos considerando a redução de dimensionalidade por meio da técnica de *Principal Component Analysis* (PCA). A Figura 4.7(a) mostra que com menos da metade da dimensão do dataset (60/136), é possível manter cerca de 90% da variação cumulativa dos dados explicada. Usando um número menor de componentes, a precisão dos dados originais poderia ser perdida. A Figura 4.7(b) ilustra um gráfico de método elbow para identificação de eps ideal considerando uma amostra de 10% do número de instâncias do dataset.

Figura 4.7 - Testes de verificação para definição de eps com PCA.

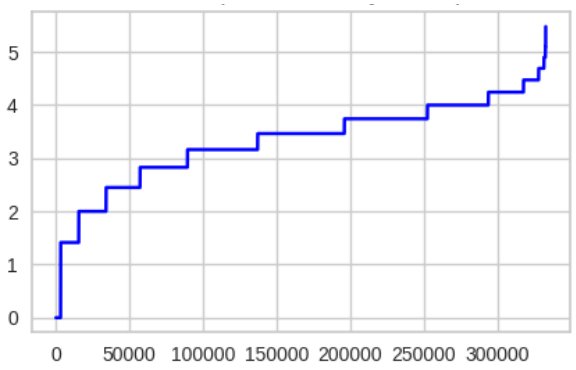


(a) (b)

Fonte: Elaborado pelo autor.

O cálculo de distância indicou que o valor ideal para eps seria entre ~4.5 e ~5.2. Testou-se diversas combinações com 4.5|4.6|..|5.1|5.2. Tentou-se com 0.1 a 2.0, porém os clusters não eram gerados. A Figura 4.8 representa um gráfico de método elbow para identificação de eps ideal considerando o dataset completo. Os resultados de indicação do valor de eps foram equivalentes quando usado 10% ou 100% de instâncias do dataset.

Figura 4.8 - Método elbow para identificação de eps ideal.



Fonte: Elaborado pelo autor.

Os seguintes resultados foram observados:

- A cada valor maior de eps, a quantidade de outliers reduz gradativamente, e o número de clusters aumenta até do valor 4.0, com o máximo de 4 clusters, porém com maior quantidade de outliers;
- Nos valores de 4.1. a 4.5 são formados 2 clusters. Em 4.1 ocorre uma inversão na quantidade de outliers, passando estes a serem minoria em relação a soma dos 2 clusters; de 4.6 a 5.3 só é gerado 1 cluster de modo que a quantidade de outliers é reduzida e o cluster aumenta a quantidade de pontos de dados.
- A partir de 5.7 não há outliers. A proporção de redução de outliers é mais evidenciada no valor de eps=5. Esse motivo também reforça a escolha por eps=5 já que a intenção é avaliar os outliers.

4.4.2. Algoritmos de regras de associação

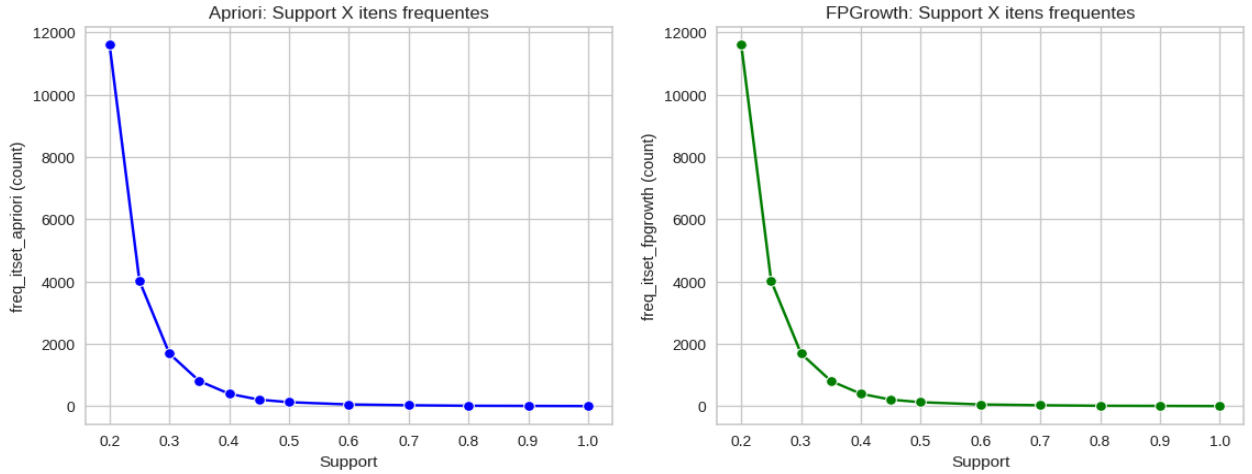
Para a tarefa de regras de associação considera-se o uso dos algoritmos Apriori e FP-Growth.

Considera-se usar dois algoritmos de regras de associação como forma de validação das regras obtidas e por questões relacionadas ao tempo de execução, tendo em vista que o dataset ENEM_HE se trata de um conjunto de alta dimensionalidade. O FP-Growth, diferentemente do Apriori, tem um tempo de execução muito menor e consome menos recursos computacionais (SILVA; PERES; BOSCARIOLI, 2020; CASTRO; FERRARI, 2016).

Na primeira etapa da tarefa de regras de associação deve ser obtido o conjunto de itens frequentes. Para isso, foram verificados intervalos de medida de suporte para, em seguida, gerar as regras de associações por meio dos itemsets selecionados.

A Figura 4.9 mostra a quantidade de itemsets frequentes em relação às medidas de suporte, variando entre 20% e 100%, para os algoritmos Apriori e FPGrowth. Naturalmente, o valor de 100% de medida de suporte anula a quantidade de itemsets, sendo colocado aqui apenas como verificação de um valor máximo.

Figura 4.9 - Verificação da medida de suporte por quantidade de itens frequentes.



(a) (b)

Fonte: Elaborado pelo autor.

Os valores de itens frequentes por medidas de suporte indicados por ambos algoritmos é exatamente o mesmo. Percebe-se que o valor de 25% de suporte contém uma boa quantidade de itens mais frequentes (4.291) em relação ao total de transações (332.793) do dataset ENEM_HE. Escolher um valor de suporte maior levaria à redução significativa de itens frequentes, eliminando muitas variáveis consideradas importantes à análise de perfis de estudantes.

A segunda parte se refere à geração de regras de associação. Como o dataset ENEM_HE caracteriza-se por ser de alta dimensionalidade, consequentemente torna-se inviável explorar um grande número de regras, assim os limites mínimos (*minimum threshold*) para as medidas de avaliação (suporte, confiança e lift) foram aplicadas e avaliadas para ambos cenários. Para discussão dos resultados é usado o lift maior que 1, visto que essa medida de correlação com valor positivo mostrará as ocorrências em que antecedentes implicam nos consequentes. Os resultados são apresentados considerando o dataset completo e por cada um dos grupos obtidos com o padrão de 3 perfis identificados e explorados no Experimento 2.

A seleção das principais regras obtidas para discussão dos resultados é realizada por meio de observação, considerando principalmente as categorias de atributos relevantes ao desempenho do estudante. Naturalmente, também é considerada a calibração e resultados das principais medidas de avaliação (suporte, confiança e lift) aplicadas sobre a definição de itens frequentes e regras de associação geradas.

Neste sentido, são identificadas e avaliadas as principais regras de associação que possam estar relacionadas ao desempenho do estudante de modo geral e individualmente por cada grupo, bem como favorece ao entendimento de quais os principais fatores que permitem ao estudante ter

um melhor ou pior desempenho. Os resultados obtidos com os métodos e a avaliação com os experimentos realizados por meio da abordagem são descritos e discutidos no Capítulo 5.

4.5. ABORDAGEM PROPOSTA *versus* TRABALHOS RELACIONADOS

O Quadro 4.2 sumariza características dos trabalhos relacionados, descritos no Capítulo 3, neste momento, de modo comparativo à abordagem proposta neste trabalho.

Quadro 4.2 - Comparativo entre os trabalhos relacionados e a abordagem proposta.

TR/ edição dos dados	**Tipo de análise/ Algoritmo(s) de AM**	**Atributos relevantes identificados**	**Perfis de estudantes**	**Associações de características**
- Weber Neto et al. (2022a) - 2019 e 2020	- Exploratória	SG_UF_ESC, TP_PRESENCA (nas provas), TP_DEPENDENCIA_ADM_ESC, TP_ESCOLA, TP_COR_RACA, RENDA_FAMILIAR, NOTA_MEDIA.	- - -	- - -
- Weber Neto et al. (2022b) - 2019 e 2020	- Exploratória	TP_SEXO, TP_DEPENDENCIA_ADM_ESC, TP_ESCOLA, TP_COR_RACA, RENDA_FAMILIAR, Possui Computadores, Possui Celulares, Acesso à Internet, Escolaridade Pai, Escolaridade Mãe, NOTA_MEDIA.	- - -	- - -
- Weber Neto et al. (2023) - 2017, 2018, 2019 - 2020 e 2021	- Exploratória	SG_UF_ESC, TP_PRESENCA (nas provas), TP_DEPENDENCIA_ADM_ESC, TP_ESCOLA, RENDA_FAMILIAR, NOTA_MEDIA.	- - -	- - -
- Pereira Junior et al. (2021) - 2020 (gerou dataset próprio)	- Análise de clusters/ SOM e K-means	foco_para_ead, sanar_duvidas_ead, nivel_perfil_aluno_ead, lugar_calmo, uso_software_educativo, desvantagem_suspensao_aulas, sexo, animo_para_atividades_ead, nivel_academico, situacao_emprego, nivel_estimulo_negativo_estresse, nivel_perfil_aluno.	Favoráveis / Não favoráveis ao ensino remoto.	- - -
- Lima et al. (2020) - 2012 a 2017	- Análise de clusters/ K-means	TP_PRESENCA (nas provas), NU_IDADE, TP_SEXO, TP_ESTADO_CIVIL, TP_COR_RACA, TP_ESCOLA, SG_UF_ESC, TP_DEPENDENCIA_ADM_ESC, TP_LOCALIZACAO_ESC, NOTAS (de cada área), Atributos relacionados a tipos de deficiências,	Baixo / Médio / Alto desempenho.	- - -

		NOTA_MEDIA.		
- Leoni e Sampaio (2017) - 2015	- Análise de clusters/ K-means	TP_ESCOLA, TAXA_PARTICIPACAO, IND_FORM_DOCENTE, TAXA_APROVACAO, MEDIA_CH, MEDIA_CN, MEDIA_LC, MEDIA_MT, MEDIA_RED.	Pior / Melhor desempenho por tipo de escola.	- - -
- Silva et al. (2020) - 2019	- Análise de clusters/ K-means. - Análise de regras de associação / Apriori	TP_COR_RACA, TemPC, RENDA_FAMILIAR, Escolaridade Mãe, NOTAS (de cada área), NOTA_MEDIA.	Melhor / Pior desempenho.	Atributos socioeconômicos.
- Gomes et al. (2017) - 2013 e 2014	- Análise de regras de associação/ Apriori	NU_IDADE, TP_ESCOLA, TP_LOCALIZACAO_ESC, TP_ENSINO, RENDA_FAMILIAR, NOTAS (de cada área).	- - -	Atributos socioeconômicos, de escola e de local de residência.
- De Moraes et al. (2021) - 2017	- Exploratória	Escolaridade Mãe, Renda familiar, Escola possui biblioteca, Escola possui sala de leitura, Faixa etária, Nota em Matemática.	- - -	- - -
- Este trabalho - 2022	- Análise de clusters/ K-means, CLARA e DBSCAN. - Análise de regras de associação/ Apriori e FP-Growth	TP_FAIXA_ETARIA, TP_SEXO, TP_COR_RACA, NIVEL_ESC_PAI, NIVEL_ESC_MAE, RENDA_FAMILIAR, NU_CELULAR, NU_COMPUTADOR, ACESSO_INTERNET, NOTA_MEDIA, **Dados de hábitos de estudo.**	Baixo / Médio / Alto.	Atributos socioeconômicos relevantes ao desempenho e hábitos de estudo durante a pandemia.

Fonte: Elaborado pelo autor.

O presente trabalho traz uma visão mais abrangente do desempenho de estudantes no ENEM em nível nacional ao considerar a integração com dados de hábitos de estudo, podendo também ser aplicada tanto em análises baseadas em contextos regionais quanto estaduais.

Durante o desenvolvimento desta dissertação, não foi encontrado nenhum estudo que tivesse trabalhado com os dados de hábitos de estudo durante a pandemia de modo associado aos dados padrões do ENEM. O diferencial principal desta dissertação em relação aos trabalhos supracitados está pautado no foco em analisar os dados atuais de hábitos de estudo integrados a fatores relevantes de desempenho de estudantes no ENEM em nível nacional. Para isso são utilizadas tarefas de AM não supervisionado, a fim de se obter um melhor entendimento dos perfis de estudantes durante a pandemia considerando seus hábitos de estudo.

Esta dissertação, por outro lado, propõe uma abordagem para integrar atributos relevantes identificados a partir da RSL descrita por Dutra, Firmino Júnior, Souza (2023), com os dados de hábitos de estudo com o intuito de avaliar o desempenho de estudantes na prova do ENEM.

No contexto da abordagem proposta, este trabalho gerou o dataset ENEM_HE, que pode ser uma importante contribuição para outros trabalhos e outras análises. O conjunto de dados ENEM_HE em está disponibilizado sob licença *Creative Commons By Attribution 4.0 International* em: https://bit.ly/ENEM_HE_2022_BR.

5. AVALIAÇÃO E RESULTADOS

Este capítulo apresenta a avaliação da abordagem proposta, onde são mostrados dois experimentos com o objetivo de responder às questões de pesquisa.

Os resultados do primeiro experimento respondem à QP1, onde são discutidos os perfis obtidos de cada grupo de estudantes a partir dos cenários destacados por meio da tarefa de agrupamento. Quanto aos resultados do segundo experimento, em resposta à QP2, são discutidas as principais regras de associações de modo geral e específicas por grupos de estudantes.

5.1. EXPERIMENTO 1

O objetivo do Experimento 1 é avaliar a identificação de perfis dos estudantes de acordo com a abordagem proposta. Para isso, dois cenários são considerados: (i) a identificação de três perfis; e (ii) a identificação de perfis outliers, ambos conforme as definições explicadas na Seção 4.4.1.

5.1.1. Perfis de estudantes com respeito ao desempenho em três categorias

A Figura 5.1 mostra a distribuição de cada cluster considerando todos os atributos do conjunto de dados ENEM_HE. Os atributos estão organizados conforme aqueles identificados como relevantes e por grupos temáticos de hábitos de estudo. Os valores nos Clusters 0, 1 e 2 correspondem às respostas mais frequentes e indicam os perfis de estudantes, em resposta à QP1, ao passo que é evidenciado o desempenho no ENEM.

Figura 5.1 - Distribuição de grupos obtidos.

K-MEANS (K=3)				
RSL + GRUPOS TEMÁTICOS DE HE	**VARIÁVEL**	**CLUSTER 0**	**CLUSTER 1**	**CLUSTER 2**
Atributos relevantes da RSL	TP_SEXO	femino	femino	femino
	TP_COR_RACA	pretaPardaIndígena	pretaPardaIndígena	pretaPardaIndígena
	NIVEL_ESC_PAI	nãoCompletouMédio	nãoCompletouMédio	nãoCompletouMédio
	NIVEL_ESC_MAE	completouAtéMédio	completouAtéMédio	completouAtéMédio
	RENDA_FAMILIAR	até1818	até1818	até1818
	NU_CELULAR	tem	tem	tem
	NU_COMPUTADOR	tem	tem	nãoTem
	ACESSO_INTERNET	tem	tem	tem
Situação de matrícula escolar e percepção da própria aprendizagem	APR_MATR_SIT_MEDIO	ensinoRegular	ensinoRegular	ensinoRegular
	APR_MATR_VINCULO	nãoInterrompeu	nãoInterrompeu	nãoInterrompeu
	APR_MATR_TP_ESTUDO	híbrido	híbrido	híbrido
	APR_MATR_APREND_PANDEMIA	aprendeuMaisPresen	aprendeuMaisPresen	aprendeuMaisPresen
Gestão do tempo e planejamento de estudos (1ª parte)	GEST_TEMP_ATV_CRONOGR	muitasVezes	nenhumaVez	poucasVezes
	GEST_TEMP_ATV_TEMPO	muitasVezes	nenhumaVez	poucasVezes
	GEST_TEMP_ATV_MATERIAL	muitasVezes	nenhumaVez	poucasVezes
	GEST_TEMP_ATV_HORA_PROG	muitasVezes	nenhumaVez	poucasVezes
Práticas de estudo e pesquisa (1ª parte)	PRAT_EST_LER	muitasVezes	nenhumaVez	poucasVezes
	PRAT_EST_RESUM_TEXTO	muitasVezes	nenhumaVez	poucasVezes
	PRAT_EST_RESUM_VIDEO	muitasVezes	nenhumaVez	poucasVezes
	PRAT_EST_ATV_FIXACAO	muitasVezes	nenhumaVez	poucasVezes
	PRAT_EST_ATV_AVALIACAO	muitasVezes	nenhumaVez	poucasVezes
	PRAT_EST_DISTRACOES	muitasVezes	nenhumaVez	poucasVezes
	PRAT_EST_ANOT_DUV_VIDEO	muitasVezes	nenhumaVez	poucasVezes
	PRAT_EST_ANOT_DUV_VIDEO_COMPL	muitasVezes	nenhumaVez	poucasVezes
	PRAT_EST_ANOT_DUV_PROF	muitasVezes	nenhumaVez	poucasVezes
	PRAT_EST_ESTRUT_IDEIA_REDACAO	muitasVezes	nenhumaVez	poucasVezes
	PRAT_EST_TREINAR_REDACAO	muitasVezes	poucasVezes	poucasVezes
	PRAT_EST_PARTICIPAR_FORUM	poucasVezes	nenhumaVez	poucasVezes
Gestão do tempo e planejamento de estudos (2ª parte)	GEST_TEMP_PONTUAL_AULA_ONLINE	muitasVezes	nenhumaVez	poucasVezes
	GEST_TEMP_ASSID_AULA_ONLINE	muitasVezes	nenhumaVez	poucasVezes
Práticas de estudo e pesquisa (2ª parte)	PRAT_EST_REV_ANOT	muitasVezes	nenhumaVez	poucasVezes
	PRAT_EST_REV_VIDEOAULA	muitasVezes	nenhumaVez	poucasVezes
Problemas na rotina de estudos	PROB_ROT_EST_PANDEMIA	sim	sim	sim
Dificuldades de infraestrutura	DIF_INFRA_PANDEMIA	não	não	não
Ajuda de terceiros	AJUD_TERC_PANDEMIA	sim	ninguémAuxiliou	sim
Avaliação sobre a própria experiência	AUTOAV_PREPARACAO_APRENDIZ	bemPreparado	poucoPreparado	poucoPreparado
DESEMPENHO	NOTA_MEDIA	alto	médio	baixo

Fonte: Elaborado pelo autor.

A avaliação dos perfis de estudantes foi realizada considerando também cada resposta específica de cada grupo formado para dados do ENEM e de hábitos de estudo. Os dados estão disponíveis no Apêndice B.

Com base no cenário apresentado, os Clusters (Grupos) formados indicam os seguintes perfis de estudantes:

Grupo 0: do ponto de vista de desempenho, contém estudantes com melhores notas. Em relação ao perfil socioeconômico, destacam-se estudantes com as seguintes características: têm maior acesso a computadores e celulares em casa; a mãe completou até o ensino médio, já o pai não completou o ensino médio; a renda familiar é mais elevada e são em maioria do sexo feminino. Sobre as práticas de estudos e gestão do tempo para as atividades, observa-se o seguinte: participaram e realizaram por muitas vezes de tais atividades, com exceção da participação em fóruns, nas quais fizeram poucas vezes. A participação em fóruns foi pouco frequente em todos os grupos de estudantes. Os estudantes procuravam ter maior pontualidade e assiduidade para participar dos momentos de aulas remotas. A maior disponibilidade de tempo pode ser justificada pelo relato de que não apresentaram dificuldades de infraestrutura e por terem recebido ajuda de terceiros durante a pandemia. Sobre a avaliação da própria experiência, eles se auto avaliam como bem preparados.

Grupo 1: em geral, contém estudantes com desempenho considerado médio. Evidenciam-se neste grupo principalmente estudantes em que: a faixa etária predominante é a profissional; possuem renda familiar de até 1 salário mínimo e meio; aprenderam mais na modalidade de estudo híbrida; por nenhuma vez organizam o tempo, o material ou a programação de horas para os estudos; realizam práticas de estudos aplicadas por nenhuma vez como, por exemplo, leituras, resumo e anotações sobre textos e vídeos; dedicaram pouco tempo para treinar a redação, sendo este um fator importante que os diferencia dos estudantes com melhor desempenho e os assemelham com aqueles com baixo desempenho; sobre a ajuda de terceiros, ninguém os auxiliou durante a pandemia; além disso, na autoavaliação, indicam estar pouco preparados.

Grupo 2: contém estudantes, em geral, com notas baixas. Observa-se que: a relação de ter ou não ter computador e celular é mais discrepante neste grupo; tem a maioria de estudantes com renda familiar até R$ 1.818,00; não preferem a modalidade de estudo híbrida; há predominância das raças cotistas (PretaPardaIndígena); declaram que poucas vezes realizaram gerenciamento do tempo e planejamento durante as atividades realizadas durante a pandemia, bem como por poucas vezes se dedicaram às atividades práticas de estudo; quanto à ajuda de terceiros referem que foram auxiliados e que tiveram muitos problemas de rotina de estudos durante a pandemia; se auto avaliam como pouco preparados para o exame.

Em relação aos fatores que mais influenciam na caracterização dos perfis de estudantes estão os hábitos de estudo dos grupos temáticos referentes à gestão do tempo, planejamento e práticas de estudo e pesquisa. As variáveis desses grupos temáticos são as que mais favorecem a interpretação dos Clusters obtidos, o que indica que tais fatores são mais relevantes para definir o desempenho obtido pelo estudante na prova do ENEM.

Quanto às características que pouco distinguem os perfis estão as variáveis presentes no grupo temático de situação de matrícula escolar e percepção da própria aprendizagem. Já em

relação às características socioeconômicos estão a faixa etária e o sexo como fatores que menos impactam na definição dos grupos formados.

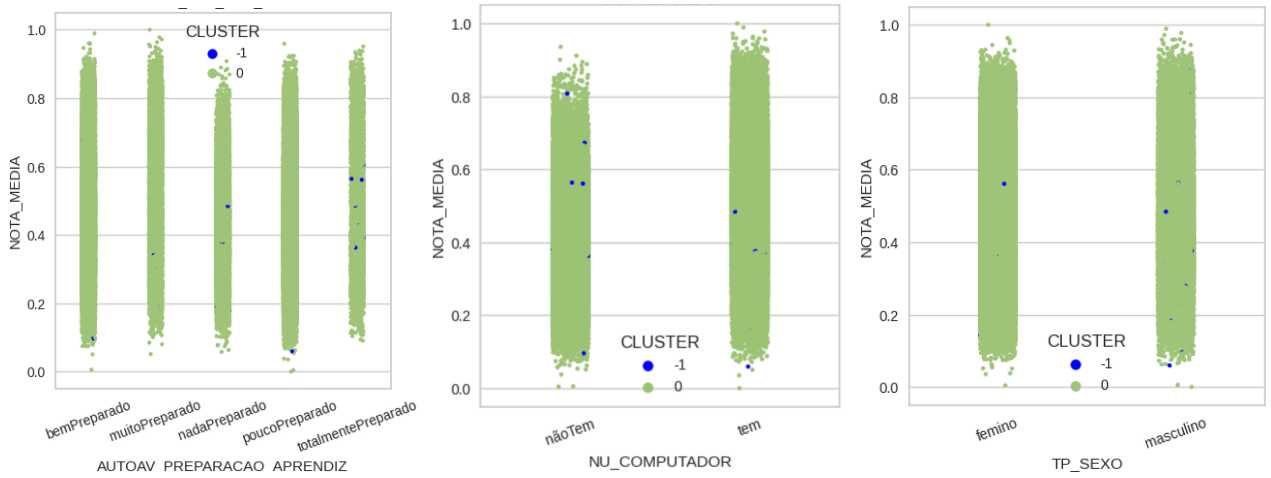
5.1.2. Perfis outliers de estudantes

O objetivo do Experimento 1 também é avaliar a identificação de perfis outliers de estudantes de acordo com a abordagem proposta.

Conforme descrito na Seção 4.4.1, após a calibração dos parâmetros para a aplicação do DBSCAN e sua validação, obteve-se a identificação de 387 estudantes com perfis outliers, o que representa 0,001163% das instâncias do dataset ENEM_HE.

Os gráficos indicados na Figura 5.2 apontam uma amostra de estudantes com base em três atributos em relação a nota média (desempenho). Os pontos de cor azul são os perfis outliers.

Figura 5.2- Amostra de variáveis com identificação de outliers.



Fonte: Elaborado pelo autor.

Essa amostra de resultados mostra que os outliers com relação ao desempenho são em geral representados por instâncias de estudantes predominatemente com as seguintes características: do sexo masculino; em maioria não possuem computador e se auto avaliam como totalmente preparados.

A Figura 5.3 ilustra a distribuição de grupos destacando um perfil de outliers de estudantes e um outro perfil considerado padrão.

Figura 5.3 - Distribuição de grupos com DBSCAN.

DBSCAN (outliers)			
RSL + GRUPOS TEMÁTICOS DE HE	**VARIÁVEL**	**CLUSTER -1 (outliers)**	**CLUSTER 0**
Atributos relevantes da RSL	TP_FAIXA_ETARIA	faixaProfissional	faixaProfissional
	TP_SEXO	masculino	femino
	TP_COR_RACA	pretaPardaIndígena	pretaPardaIndígena
	NIVEL_ESC_PAI	nãoCompletouMédio	nãoCompletouMédio
	NIVEL_ESC_MAE	nãoCompletouMédio	completouAtéMédio
	RENDA_FAMILIAR	até1818	até1818
	NU_CELULAR	tem	tem
	NU_COMPUTADOR	tem	tem
	ACESSO_INTERNET	tem	tem
Situação de matrícula escolar e percepção da própria aprendizagem	APR_MATR_SIT_MEDIO	ensinoRegular	ensinoRegular
	APR_MATR_VINCULO	nãoInterrompeu	nãoInterrompeu
	APR_MATR_TP_ESTUDO	apenasRemoto	híbrido
	APR_MATR_APREND_PANDEMIA	aprendeuMaisRemoto	aprendeuMaisPresencial
Gestão do tempo e planejamento de estudos (1ª parte)	GEST_TEMP_ATV_CRONOGR	poucasVezes	poucasVezes
	GEST_TEMP_ATV_TEMPO	muitasVezes	poucasVezes
	GEST_TEMP_ATV_MATERIAL	todasAsVezes	poucasVezes
	GEST_TEMP_ATV_HORA_PROG	poucasVezes	poucasVezes
Práticas de estudo e pesquisa (1ª parte)	PRAT_EST_LER	todasAsVezes	poucasVezes
	PRAT_EST_RESUM_TEXTO	todasAsVezes	poucasVezes
	PRAT_EST_RESUM_VIDEO	todasAsVezes	poucasVezes
	PRAT_EST_ATV_FIXACAO	todasAsVezes	poucasVezes
	PRAT_EST_ATV_AVALIACAO	muitasVezes	poucasVezes
	PRAT_EST_DISTRACOES	muitasVezes	poucasVezes
	PRAT_EST_ANOT_DUV_VIDEO	todasAsVezes	poucasVezes
	PRAT_EST_ANOT_DUV_VIDEO_COMPLEMENTAR	todasAsVezes	poucasVezes
	PRAT_EST_ANOT_DUV_PROF	muitasVezes	poucasVezes
	PRAT_EST_ESTRUT_IDEIA_REDACAO	todasAsVezes	poucasVezes
	PRAT_EST_TREINAR_REDACAO	poucasVezes	poucasVezes
	PRAT_EST_PARTICIPAR_FORUM	nenhumaVez	nenhumaVez
Gestão do tempo e planejamento de estudos (2ª parte)	GEST_TEMP_PONTUAL_AULA_ONLINE	nenhumaVez	muitasVezes
	GEST_TEMP_ASSID_AULA_ONLINE	todasAsVezes	muitasVezes
Práticas de estudo e pesquisa (2ª parte)	PRAT_EST_REV_ANOT	todasAsVezes	poucasVezes
	PRAT_EST_REV_VIDEOAULA	nenhumaVez	poucasVezes
Problemas na rotina de estudos	PROB_ROT_EST_PANDEMIA	sim	sim
Dificuldades de infraestrutura	DIF_INFRA_PANDEMIA	sim	não
Ajuda de terceiros	AJUD_TERC_PANDEMIA	não	sim
Avaliação sobre a própria experiência	AUTOAV_PREPARACAO_APRENDIZ	bemPreparado	poucoPreparado
DESEMPENHO	NOTA_MEDIA	baixo	alto

Fonte: Elaborado pelo autor.

Em geral, o perfil outliers indicam que esses estudantes realizaram atividades de prática e gestão de estudos, com forte evidência a terem realizado por todas as vezes. A maioria não frequentou o ensino regular e destaca que não aprendeu muito no modelo de aula presencial. A maioria relata ter problemas na rotina de estudos e dificuldade de infraestrutura, sem que houvesse recebido ajuda de terceiros e que, mesmo diante das dificuldades destacadas, se sente bem preparada. Mesmo assim, este grupo teve um desempenho baixo.

O perfil outliers representa estudantes que se esforçaram nas atividades de estudos durante a pandemia, porém obtiveram um desempenho muito abaixo do esperado, ainda que se auto avaliem como bem preparados. No entanto, pode-se destacar a possibilidade dos outliers aparecerem porque os estudantes não responderam o questionário de modo fidedigno e consistente, ou seja, possivelmente nem leram as questões.

5.2. EXPERIMENTO 2

O objetivo do Experimento 2 é encontrar quais regras de associações que podem ajudar a confirmar os perfis de estudantes quando se une os dados do ENEM com os hábitos de estudo, tanto de modo geral quanto individualmente por cada grupo, auxiliando no entendimento sobre a implicação das características de cada perfil de estudante no desempenho.

Os itens mais frequentes em relação aos dados socioeconômicos estão relacionados ao uso de celulares e computadores, acesso à internet e ao nível de renda familiar. Já em relação aos hábitos de estudos, predominam como mais frequentes os itens sobre aproveitamento de estudos e tipo de vínculo de matrícula na pandemia, problemas na rotina de estudos e dificuldades de infraestrutura.

Considerando cada atributo do dataset ENEM_HE, basicamente prevalecem como mais frequentes os itens de:

- ***dados socioeconômicos***: TP_FAIXA_ETARIA=faixaProfissional, TP_SEXO=femino, TP_COR_RACA=pretaPardaIndígena, NIVEL_ESC_PAI=nãoCompletouMédio, NIVEL_ESC_MAE=completouAtéMédio, RENDA_FAMILIAR=até1818, NU_CELULAR=tem, NU_COMPUTADOR=tem, ACESSO_INTERNET=tem;
- ***situação da matrícula***: APR_MATR_SIT_MEDIO=ensinoRegular, APR_MATR_VINCULO=nãoInterrompeu, APR_MATR_TP_ESTUDO=híbrido, APR_MATR_APREND_PANDEMIA=aprendeuMaisPresencial;
- ***gestão de tempo***: GEST_TEMP_ATV_CRONOGR=poucasVezes, GEST_TEMP_ATV_TEMPO=poucasVezes, GEST_TEMP_ATV_MATERIAL=poucasVezes, GEST_TEMP_ATV_HORA_PROG=poucasVezes, GEST_TEMP_PONTUAL_AULA_ONLINE=muitasVezes, GEST_TEMP_ASSID_AULA_ONLINE=muitasVezes;
- ***práticas de estudo e pesquisa***: PRAT_EST_LER=poucasVezes, PRAT_EST_RESUM_TEXTO=poucasVezes, PRAT_EST_RESUM_VIDEO=poucasVezes, PRAT_EST_ATV_FIXACAO=poucasVezes, PRAT_EST_ATV_AVALIACAO=poucasVezes, PRAT_EST_DISTRACOES=poucasVezes, PRAT_EST_ANOT_DUV_VIDEO=poucasVezes, PRAT_EST_ANOT_DUV_VIDEO_COMPLEMENTAR=poucasVezes, PRAT_EST_ANOT_DUV_PROF=poucasVezes, PRAT_EST_ESTRUT_IDEIA_REDACAO=poucasVezes, PRAT_EST_TREINAR_REDACAO=poucasVezes, PRAT_EST_PARTICIPAR_FORUM=nenhumaVez, PRAT_EST_REV_ANOT=poucasVezes, PRAT_EST_REV_VIDEOAULA=poucasVezes;
- ***problemas na rotina de estudos***: PROB_ROT_EST_PANDEMIA=sim;
- ***dificuldades de infraestrutura***: DIF_INFRA_PANDEMIA=não;
- ***ajuda de terceiros***: AJUD_TERC_PANDEMIA=sim;
- ***autoavaliação***: AUTOAV_PREPARACAO_APRENDIZ=poucoPreparado;

- ***desempenho***: NOTA_MEDIA=alto.

O Apêndice E contém exemplos itens frequentes, após aplicação de 25% de suporte em todas as transações (instâncias de estudantes) do dataset ENEM_HE. Conforme apresentado na abordagem proposta, essa é uma boa porcentagem para representar os itens mais frequentes em relação ao total de transações, visto que um valor de suporte maior levaria à redução significativa de itens frequentes. No Apêndice E também estão disponíveis amostras de regras de associação obtidas considerando o conjunto de dados completo e, de modo individual, com as instâncias pertencentes a cada um dos clusters/perfis (0, 1 e 2). Essas amostras estão classificadas em ordem decrescente conforme o valor de lift (maior que 1).

5.2.1. Regras de associação obtidas com o conjunto de dados completo

Considerando o cenário com todo o dataset, prevalecem as regras que envolvem as variáveis com valores mais frequentes como, por exemplo, presença de celulares, computadores e celulares na residência. Em relação aos dados de hábitos de estudo, em geral são destacados principalmente as questões relacionadas à gestão de tempo, organização de estudos e como foi o aproveitamento dos estudos na preparação para a prova do ENEM durante a pandemia.

O Quadro 5.1 contém um fragmento de regras obtidas considerando o dataset ENEM_HE completo. São indicadas regras, na forma de "X → Y", e os valores de Confiança (Conf.) e Lift.

Quadro 5.1 - Exemplos de regras de associação obtidas com o conjunto de dados completo.

X (antecedente) → Y (consequente)	Conf.	Lift
GEST_TEMP_ATV_MATERIAL=poucasVezes, NU_CELULAR=tem ➔ CLUSTER=2	75,4%	1.94
PRAT_EST_REV_ANOT=poucasVezes, APR_MATR_VINCULO=nãoInterrompeu ➔ CLUSTER=0	79,9%	1.78
APR_MATR_VINCULO=nãoInterrompeu, NU_CELULAR=tem, NU_COMPUTADOR=nãoTem ➔ RENDA_FAMILIAR=até1818	75,8%	1.51
TP_COR_RACA=branca, APR_MATR_SIT_MEDIO=ensinoRegular, NU_COMPUTADOR=tem, NU_CELULAR=tem ➔ APR_MATR_VINCULO=nãoInterrompeu	96,9%	1.10
RENDA_FAMILIAR=até1818, APR_MATR_SIT_MEDIO=ensinoRegular ➔ TP_FAIXA_ETARIA=faixaProfissional, NU_CELULAR=tem	60,4%	1.10
APR_MATR_APREND_PANDEMIA=aprendeuMaisPresencial, GEST_TEMP_ATV_TEMPO=nenhumaVez ➔ CLUSTER=1	63,4%	1.63
NIVEL_ESC_MAE=completouSuperior ➔ APR_MATR_VINCULO=nãoInterrompeu, NU_CELULAR=tem, ACESSO_INTERNET=tem	93,5%	1.08
TP_SEXO=femino, PROB_ROT_EST_PANDEMIA=sim, APR_MATR_VINCULO=nãoInterrompeu, DIF_INFRA_PANDEMIA=não ➔ ACESSO_INTERNET=tem	96,9%	1.05
APR_MATR_TP_ESTUDO=apenasRemoto, NU_CELULAR=tem, ACESSO_ INTERNET=tem ➔ PROB_ROT_EST_PANDEMIA=sim	80,5%	1.03

Fonte: Elaborado pelo autor.

A regra mais forte indica que estudantes que realizaram poucas vezes atividade de gestão de tempo para atividades de preparação de material, tendo o celular à disposição, têm praticamente o dobro de chance de pertencerem ao Cluster 2, ou seja, com baixo desempenho.

Estudantes que realizaram por muitas vezes práticas de revisão de anotações sobre as aulas e que não interromperam os estudos na pandemia, tem uma forte indicação a pertencerem ao Cluster 0 (alto desempenho). Já aqueles estudantes que indicaram ter aprendido mais na modalidade presencial e realizaram nenhuma vez tarefas de gestão de tempo para as atividades pertencem ao Cluster 1 (médio desempenho). Estudantes que pertencem ao Cluster 2 (baixo desempenho) estão entre os que indicaram ter aprendido mais na modalidade presencial e foram, em maioria, aqueles que nenhuma vez executaram tarefas de gestão de tempo para as atividades. A influência da prática dos hábitos de estudo pode ser explicada, por exemplo, pelos aspectos socioeconômicos, como o menor nível de renda familiar, baixo nível de escolaridade dos pais, sexo, faixa de idade e meios de acesso às aulas remotas.

Como ilustração, uma regra mostra que quando mães completaram o Ensino Superior, isso implicaria em estudantes que não interromperam os estudos, assim como possuem celular e acesso à internet. Outra regra mostra que estudantes de cor/raça branca estão em maioria entre os que não interromperam os estudos durante a pandemia. Estudantes com renda familiar de até R$ 1.818,00 e que fizeram todo o ensino médio na modalidade regular, pertencem, em maioria, à faixa etária profissional e têm celular.

Considerando aspectos socioeconômicos, as regras indicam que os estudantes que não interromperam os estudos, têm acesso a celulares, computador e internet podem obter melhores desempenhos no ENEM. Estudantes com esse perfil socioeconômico de terem acesso aos recursos e ferramentas que facilitam o acompanhamento de aulas remotas têm, em geral, um desempenho melhor, enquanto que a falta desses recursos levam à redução do desempenho.

As regras ilustradas ajudam a ratificar que os hábitos de estudo, associados às questões socioeconômicas, quando realizados com mais frequência, podem influenciar no desempenho, levando o estudante a ter notas melhores na prova do ENEM. Do contrário, quando praticados por poucas vezes, tende a levar o estudante a obter notas mais baixas.

Na seção seguinte são mostradas outras regras identificadas considerando os grupos.

5.2.2. Avaliação de regras de associação por cluster/perfil de estudante

O Quadro 5.2 contém amostras de regras obtidas considerando os cenários de instâncias de cada Cluster (0, 1 e 2).

Quadro 5.2 - Exemplos de regras de associação obtidas por cluster/perfil de estudante.

X (antecedente) ➔ Y (consequente)	Conf.	Lift
CLUSTER 0		
APR_MATR_SIT_MEDIO=ensinoRegular, APR_MATR_VINCULO=nãoInterrompeu, NU_CELULAR=tem, NU_COMPUTADOR=tem, GEST_TEMP_ASSID_AULA_ONLINE=muitasVezes ➔ ACESSO_INTERNET=tem	99,3%	1.06
DIF_INFRA_PANDEMIA=não, APR_MATR_SIT_MEDIO=ensinoRegular, PRAT_EST_TREINAR_REDACAO=muitasVezes ➔APR_MATR_VINCULO=nãoInterrompeu	99,3%	1.06
TP_COR_RACA=pretaPardaIndígena, NU_CELULAR=tem ➔ APR_MATR_VINCULO=nãoInterrompeu, ACESSO_INTERNET=tem, PRAT_EST_RESUM_VIDEO=muitasVezes	53,5%	1.02
APR_MATR_VINCULO=nãoInterrompeu, ACESSO_INTERNET=tem, PRAT_EST_TREINAR_REDACAO=muitasVezes, APR_MATR_SIT_MEDIO=ensinoRegular ➔ PRAT_EST_ESTRUT_IDEIA_REDACAO=muitasVezes	73,1%	1.31
ACESSO_INTERNET=tem, PRAT_EST_ANOT_DUV_VIDEO_COMPLEMENTAR =muitasVezes, PROB_ROT_EST_PANDEMIA=sim ➔ PRAT_EST_ATV_FIXACAO=muitasVezes	59,5%	1.17
CLUSTER 1		
NU_CELULAR=tem, ACESSO_INTERNET=tem, GEST_TEMP_ATV_TEMPO=nenhumaVez, GEST_TEMP_ATV_MATERIAL=nenhumaVez, PRAT_EST_ANOT_DUV_PROF =poucasVezes ➔ GEST_TEMP_ATV_HORA_PROG=nenhumaVez	95,0%	1.37
GEST_TEMP_ATV_HORA_PROG=nenhumaVez, PRAT_EST_RESUM_TEXTO=poucasVezes, PRAT_EST_RESUM_VIDEO=poucasVezes ➔ PRAT_EST_LER=nenhumaVez, PRAT_EST_ANOT_DUV_PROF=nenhumaVez	81,6%	1.56
PRAT_EST_PARTICIPAR_FORUM=poucasVezes, AJUD_TERC_PANDEMIA=ninguémAuxiliou, NU_CELULAR=tem, ACESSO_INTERNET=tem ➔ PROB_ROT_EST_PANDEMIA=sim	94,7%	1.13
GEST_TEMP_ATV_TEMPO=nenhumaVez, PRAT_EST_LER=nenhumaVez, PRAT_EST_RESUM_TEXTO=nenhumaVez, PRAT_EST_RESUM_VIDEO=nenhumaVez, PRAT_EST_ATV_FIXACAO=nenhumaVez, PRAT_EST_REV_ANOT=nenhumaVez ➔ GEST_TEMP_ATV_HORA_PROG=nenhumaVez	94,8%	1.37
NIVEL_ESC_MAE=nãoCompletouMédio, RENDA_FAMILIAR=até3636, APR_MATR_TP_ESTUDO=híbrido ➔ GEST_TEMP_PONTUAL_AULA_ONLINE=nenhumaVez, GEST_TEMP_ASSID_AULA_ONLINE=nenhumaVez	88,1%	1.33
CLUSTER 2		
NU_CELULAR=tem, ACESSO_INTERNET=tem, APR_MATR_SIT_MEDIO=ensinoRegular, APR_MATR_VINCULO=nãoInterrompeu, PROB_ROT_EST_PANDEMIA=sim, PRAT_EST_ANOT_DUV_VIDEO=poucasVezes ➔ APR_MATR_APREND_PANDEMIA=aprendeuMaisPresencial	83,1%	1.03
NU_CELULAR=tem, ACESSO_INTERNET=tem, PRAT_EST_ATV_FIXACAO=poucasVezes, PRAT_EST_RESUM_TEXTO=poucasVezes, PRAT_EST_RESUM_VIDEO=poucasVezes ➔ PRAT_EST_LER=poucasVezes	90,4%	1.23

GEST_TEMP_ATV_CRONOGR=poucasVezes, EST_TEMP_ATV_TEMPO=poucasVezes, PRAT_EST_DISTRACOES=poucasVezes, PRAT_EST_ANOT_DUV_VIDEO=poucas Vezes ➔ RENDA_FAMILIAR=até1818, APR_MATR_SIT_MEDIO=ensinoRegular, DIF_INFRA_PANDEMIA=sim	77,1%	1.09
NU_CELULAR=tem, ACESSO_INTERNET=tem, PROB_ROT_EST_PANDEMIA=sim, APR_MATR_VINCULO=nãoInterrompeu, PRAT_EST_ANOT_DUV_PROF=poucasVezes ➔ AUTOAV_PREPARACAO_APRENDIZ=poucoPreparado, APR_MATR_APREND_PANDEMIA =aprendeuMaisPresencial	67,8%	1.02
TP_SEXO=masculino, APR_MATR_SIT_MEDIO=ensinoRegular, PRAT_EST_TREINAR_REDACAO=poucasVezes, GEST_TEMP_ATV_TEMPO=poucasVezes, PRAT_EST_LER=poucasVezes ➔ DIF_INFRA_PANDEMIA=sim, AUTOAV_PREPARACAO_APRENDIZ=poucoPreparado	80,4	1.06

Fonte: Elaborado pelo autor.

Para o grupo com o **Perfil de alto desempenho** foram geradas no total 89.491 regras de associação, já considerando o filtro aplicado pela medida de lift. As regras referentes a esse perfil envolvem, em maioria, respostas positivas (muitas vezes) sobre as atividades desempenhadas quanto aos hábitos de estudo.

As regras confirmam, por exemplo, que estudantes com perfil de alto desempenho relatam que não apresentaram dificuldades de infraestrutura, têm matrícula no ensino regular e que treinaram a redação por muitas vezes, implicando que não interromperam os estudos durante a pandemia. Tal regra ratifica a importância de que a disponibilidade de meios tecnológicos, infraestrutura e bons hábitos de estudos favoreceram para que o estudante permanecesse matriculado durante a pandemia.

A prova de redação tende a ser decisiva no desempenho final obtido pelo estudante. Assim, uma característica importante revelada para esse grupo de estudantes está no fato de que eles praticaram por muitas vezes a redação. Outra característica importante foi terem dedicado mais atenção nas atividades de organização e estruturação de ideias durante os treinamentos da redação. As regras também revelaram que, mesmo apresentando problemas na rotina de estudos, os hábitos de estudo de práticas de resumo de videoaulas, vídeos complementares, anotações de dúvidas, atividades de fixação, quando realizados por muitas vezes, favoreceram a bons resultados no exame.

Para este grupo de alto desempenho fica evidente que as atividades de hábitos de estudo, em sendo realizadas com maior frequência, endossam a possibilidade do estudante obter um bom desempenho na prova.

Para o grupo com o **Perfil de médio desempenho** foram geradas 3.603.659 regras de associação. A amostra de regras envolvem principalmente respostas negativas (nenhuma vez) sobre as atividades desempenhadas nos hábitos de estudo.

As regras em destaque nesse grupo mostram estudantes que relatam terem feito por nenhuma vez atividades de gerenciamento de tempo, material, dúvidas para esclarecer com os professores, implicando em pouca dedicação na sua programação de estudo por dia e participação

nas aulas remotas. Além destas características, as principais atividades, em que estudantes indicam não terem realizado, estão as práticas essenciais ao entendimento dos conteúdos (leitura, gestão de tempo para atividades, resumo de textos e vídeos, anotação de dúvidas e atividades de revisão/fixação). A falta de pontualidade e assiduidade nas aulas remotas são hábitos que mostram um contraste com preferência pelo tipo de ensino híbrido por estudantes deste grupo.

O resultado em que estudantes de desempenho médio relatam, em geral, que por nenhuma vez realizaram as atividades de hábitos de estudo é um pouco inesperado. Entretanto, esse perfil é observado com características muito semelhantes às que foram obtidas considerando os demais algoritmos de agrupamento. Para o grupo de médio desempenho fica evidente que as atividades de hábitos de estudo, em sendo realizadas com pouca ou nenhuma frequência, favorecem à queda de desempenho na prova do ENEM.

Por fim, no **Perfil de baixo desempenho** foram geradas 1.098.373 regras de associação. Neste grupo, estudantes de baixo desempenho predominantemente indicam respostas que denotam pouca realização das atividades de hábitos de estudo (poucas vezes).

As regras, neste grupo, indicam uma quantidade maior de estudantes sem acesso a celular, computador e internet, que refletem de modo frequente características de maiores dificuldades de infraestrutura e problemas na rotina de estudos na pandemia. A renda familiar de até R$ 1.818,00 e menor escolaridade dos pais refletem o perfil dos hábitos de estudo de estudantes de baixo desempenho, que por poucas vezes fizeram, por exemplo, práticas de anotações de dúvidas, gerenciamento do tempo para atividades de fixação, realização de resumos de vídeos e textos das matérias, leituras sobre assuntos diversos e treinamento de redação.

Como observado, as regras obtidas para o perfil de baixo desempenho elucidam, em geral, que por poucas vezes estudantes deste grupo realizaram as atividades de hábitos de estudo. Esses resultados reforçam a ideia de que a baixa frequência de realização de atividades de hábitos de estudo proporciona uma queda muito acentuada de desempenho na prova do ENEM.

No Apêndice E encontram-se informações sobre a quantidade de regras de associações geradas, estando organizadas em ordem decrescente por valor de lift, ao passo que são ilustradas amostras de 30 regras obtidas em cada cenário.

6. CONSIDERAÇÕES FINAIS

O ENEM é um dos principais exames realizados no país e, provavelmente, o mais importante para estudantes que almejam uma vaga no ensino superior. A partir dos dados de hábitos de estudo, importante recurso complementar aos microdados do ENEM, este trabalho teve como propósito desenvolver uma abordagem para identificação de perfis de desempenho de estudantes e associações de características de hábitos de estudo durante um período da pandemia, por meio métodos de aprendizado de máquina não supervisionados. Para avaliação da abordagem, foram avaliados resultados obtidos nos modelos de agrupamento e de regras de associação e suas análises nas respostas às questões de pesquisa.

Os resultados mostraram que os estudantes que estabelecem melhores gestão e práticas em sua rotina de estudos, mesmo diante de dificuldades de infraestrutura e pouco acesso a meios tecnológicos, podem almejar bons resultados no exame. Não obstante, ainda sim, reforça a ideia de que quanto mais condições favoráveis o estudante tem a sua disposição, maiores são as chances de se ter um melhor desempenho. Os dados de hábitos de estudo revelaram aspectos importantes relacionados ao desempenho dos estudantes e quanto aos impactos causados pela pandemia.

Alguns atributos de hábitos de estudo não foram usados no trabalho, a exemplo de: tecnologias e meios de acesso (rádio, televisão, podcasts, grupos de mensagens, etc.); tipo de ajuda que recebeu (financeira, alimentação, acesso a equipamentos, etc.); e quem o ajudou (pais, irmãos, tios, amigos, etc.). Essas informações não consideradas neste trabalho podem ser futuramente usadas para um maior detalhamento dos perfis obtidos. Devido ao preenchimento do questionário de hábitos de estudo ser opcional, muitos dados de estudantes com relação a estes aspectos não estavam disponíveis.

Os resultados obtidos foram gerados com base em dados recém divulgados sobre os hábitos de estudo. Foram usados dados somente de 2022. Logo, uma investigação considerando dados de anos seguintes à pandemia poderão indicar um panorama a longo prazo, com maior precisão e riqueza de detalhes a respeito das consequências e impactos causados aos estudantes. Para se ter uma visão mais abrangente, considerando estudantes que responderam parcialmente o questionário de hábitos de estudo, uma estratégia de imputação de valores poderá ser aplicada em novos experimentos.

O desempenho acadêmico no ENEM poderá ser melhor acompanhado por educadores com base nos resultados indicados como fatores mais relevantes para o desenvolvimento da aprendizagem dos estudantes. Ações que busquem potencializar os hábitos de estudo considerados mais importantes poderão ser desenvolvidas por professores e estudantes.

Foram encontradas evidências de que estudantes com perfil de organização do seu próprio espaço de estudo apresentam melhores desempenho no exame. Isso reforça a necessidade de que as escolas possam motivá-los a realizarem práticas de estudo que favoreçam à compreensão de conteúdos, maior capacidade de concentração e, em consequência, a elevação da qualidade de ensino. Com base na atenção às necessidades dos estudantes, algumas ações podem ser realizadas

como, por exemplo, no trabalho com estudantes de baixo rendimento na redação. Outro exemplo que pode ser citado é o trabalho junto aos professores no sentido de que estes possam incentivá-los a realizarem com maior frequência boas práticas de hábitos de estudo. Tal ação poderá contribuir, principalmente para aqueles estudantes cujo perfil é de baixo desempenho.

As regras de associação encontradas sugerem que fornecer melhores condições de permanência, infraestrutura e de acesso, principalmente para estudantes menos favorecidos financeiramente, pode contribuir na redução do problema de desigualdade no desempenho do exame, em meio ao contexto de aulas remotas durante a pandemia.

Gestores de escolas poderão agir em prol de melhores condições no processo de ensino. Tais melhorias podem envolver, por exemplo: (i) planejamento acadêmico que considere questões de vulnerabilidade dos estudantes com menor renda; (ii) melhoria da infraestrutura física e pedagógica das escolas, principalmente daquelas em que os estudantes não obtiveram um bom desempenho na prova; (iii) elaboração de projetos que deem suporte à formação complementar do estudante e assegure a permanência destes até a conclusão da Educação Básica; (iv) desenvolvimento de políticas públicas que ampliem o acesso às boas práticas de hábitos de estudo.

Os dados socioeconômicos e demais atributos relevantes podem ser utilizados por professores como um meio de diagnóstico do desempenho. Os professores poderão: (i) identificar questões sensíveis que influenciam no desempenho como, por exemplo, um baixo desempenho em uma área de conhecimento, como em Matemática; (ii) planejar melhores estratégias de ensino e aprendizagem para suas turmas, ou mesmo considerar um planejamento com base nos hábitos de estudo.

Para os estudantes, os perfis identificados por este trabalho podem ser usados como uma referência à indicação sobre o seu desempenho baseado no seu próprio perfil, de modo que ele possa refletir e verificar sob quais hábitos de estudo poderia se comprometer mais para melhorar o seu desempenho. O fato de se manter uma rotina, por muitas vezes, de revisões e anotações de dúvidas sobre as matérias, treinamento de redação, pode levá-lo a ter um bom êxito no ENEM. Tal conduta poderá motivá-lo a estudar mais e tentar ingressar no curso que almeja, bem como para prepará-lo para outros tipos de processo de avaliação de conhecimento.

6.1. PRINCIPAIS CONTRIBUIÇÕES

As principais contribuições deste trabalho são elencadas em cinco pontos:

- Fornece uma revisão abrangente e atualizada acerca de fatores associados ao desempenho de estudantes quando atuam no ENEM.
- Uso de dados de hábitos de estudo recém publicados a fim de evidenciar um panorama de características que afetam positiva ou negativamente o desempenho de estudantes na prova do ENEM durante a pandemia;
- Identificação de perfis e associações de características que podem ajudar a entender o desempenho de estudantes no ENEM;

- Subsídios para que gestores de instituições de ensino possam investir em novas ferramentas que auxiliem na melhoria do desempenho de estudantes e na qualidade da educação. Supõe-se ainda que é possível entender como os resultados poderão refletir em situações futuras de dificuldades educacionais de larga escala, como uma nova pandemia;
- Caracterização de perfis de estudantes que possam favorecer à criação de diretrizes e políticas públicas de inclusão e apoio àqueles que pretendem ingressar no ensino superior por meio do ENEM.

6.2. TRABALHOS FUTUROS

Como trabalhos futuros nesta linha de pesquisa, podem ser destacadas as seguintes sugestões:

- Avaliar como desempenho, questões socioeconômicas, hábitos de estudo e escolas estão associados considerando uma escala temporal de diversas edições do exame ou mesmo comparando perfis de estudantes de regiões/estados/cidades do Brasil;
- Realizar experimentos a partir de outros algoritmos de agrupamento e regras de associação em busca de possíveis melhorias da abordagem;
- Incrementar novas atividades e tarefas de AM. A abordagem pode incluir, por exemplo, tarefas de AM supervisionado para predição de desempenho ou de perfis de estudantes a partir da sua caracterização socioeconômica e conduta frente aos hábitos de estudo;
- Avaliar um período maior de edições dos dados. Isso permite que os perfis de estudantes possam ser avaliados em escala temporal em outros períodos da pandemia.

REFERÊNCIAS

AGRAWAL, Rakesh et al. Fast discovery of association rules. **Advances in knowledge discovery and data mining**, v. 12, n. 1, p. 307-328, 1996. Citado 2 vezes nas páginas 32 e 45.

AGRAWAL, Rakesh; IMIELIŃSKI, Tomasz; SWAMI, Arun. Mining association rules between sets of items in large databases. In: Proceedings of the 1993 ACM SIGMOD international conference on Management of data. 1993. p. 207-216. Citado 3 vezes nas páginas 32, 42 e 43.

ALDINO, Ahmad Ari et al. Comparison of market basket analysis to determine consumer purchasing patterns using fp-growth and apriori algorithm. In: **2021 International Conference on Computer Science, Information Technology, and Electrical Engineering (ICOMITEE)**. IEEE, 2021. p. 29-34. Citado na página 52.

ALPAYDIN, E. **Introduction to machine learning**. 2nd ed. Cambridge, Mass: MIT Press, 2010. Citado 4 vezes nas páginas 28 e 45.

AOUEDI, Ons et al. Network Traffic Analysis using Machine Learning: an unsupervised approach to understand and slice your network. **Annals of Telecommunications**, p. 1-13, 2021. Citado na página 37.

BENESTY, J. et al. Pearson Correlation Coefficient. Em: COHEN, I. et al. (Eds.). **Noise Reduction in Speech Processing**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009. v. 2p. 1–4. Citado na página 55.

BONACCORSO, G. Machine Learning Algorithms: Popular Algorithms for Data Science and Machine Learning, 2nd Edition. 2nd ed. ed. Birmingham: Packt Publishing Ltd, 2018. Citado 4 vezes nas páginas 32, 33, 35 e 36.

BRASIL. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep). **Política e Plano de Dados Abertos do Inep (Biênio – 2020-2021)**. Brasília, 2020. Disponível em: <https://download.inep.gov.br/publicacoes/institucionais/gestao_e_governanca/politica_e_plano_de_dados_abertos.pdf>. Acesso em: 14 set. 2023. Citado na página 22.

BRASIL. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. **Apresentação do Enem - Inep**. Brasília, 2023. Disponível em: <https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enem>. Acesso em: 10 set. 2023. Citado 2 vezes nas páginas 21 e 22.

BRASIL. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep). **A redação no Enem 2022: cartilha do participante**. Brasília, 2022. Disponível em: <https://download.inep.gov.br/download/enem/cartilha_do_participante_enem_2022.pdf>. Acesso em: 15 set. 2023. Citado na página 22.

BRASIL. Ministério da Educação. **ENEM: documento básico 2002**. Brasília, 2002. Disponível em: <https://download.inep.gov.br/publicacoes/institucionais/avaliacoes_e_exames_da_educacao_basica/enem_exame_nacional_do_ensino_medio_documento_basico_2002.pdf>. Acesso em: 09 ago. 2023. Citado 2 vezes nas páginas 15 e 20.

BRUCE, Peter; BRUCE, Andrew; GEDECK, Peter. **Practical statistics for data scientists: 50+ essential concepts using R and Python**. O'Reilly Media, 2020. Citado na página 35.

CASTRO, L. N.; FERRARI, D. G. **Introdução à mineração de dados: conceitos básicos, algoritmos e aplicações**. São Paulo, BR: Saraiva, 2016. Citado 7 vezes nas páginas 30, 34, 36, 37, 41, 43 e 76.

CALIŃSKI, Tadeusz; HARABASZ, Jerzy. A dendrite method for cluster analysis. **Communications in Statistics-theory and Methods**, v. 3, n. 1, p. 1-27, 1974. Citado na página 38.

CHAPMAN, P. et al. **CRISP-DM 1.0: Step-by-step data mining guide**. 2000. Disponível em: https://api.semanticscholar.org/CorpusID:59777418. Acesso em 24 ago. 2023. Citado 2 vezes na página 26.

DAVIES, D. L.; BOULDIN, D. W. A Cluster Separation Measure. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. PAMI-1, n. 2, p. 224–227, abr. 1979. Citado na página 37.

DE CASTRO, M. H. G.; TIEZZI, S. A reforma do ensino médio e a implantação do Enem no Brasil. Desafios, v. 65, n. 11, p. 46-115, 2004. Citado na página 21.

DE MORAES, C. P.; PERES, R. T.; PEDREIRA, C. E. Eficácia escolar e variáveis familiares em tempos de pandemia: um estudo a partir de dados do ENEM. **INTERFACES DA EDUCAÇÃO**, v. 12, n. 35, p. 635–658, 2 nov. 2021. Citado 3 vezes nas páginas 62, 65 e 80.

DUTT, A.; ISMAIL, M. A.; HERAWAN, T. A systematic review on educational data mining. **IEEE Access**, v. 5, p. 15991–16005, 2017. Disponível em: <https://ieeexplore.ieee.org/abstract/document/7820050>. Acesso em: 16 out. 2023. Citado na página 32.

ELMASRI, R.; NAVATHE, S. B. **Sistemas de Banco de Dados**. 2019. Citado na página 30.

ESTER, Martin et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In: **kdd**. 1996. p. 226-231. Citado 2 vezes nas páginas 31 e 39.

EZUGWU, Absalom E. et al. A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. **Engineering Applications of Artificial Intelligence**, v. 110, p. 104743, 2022. Citado na página 35.

FACELI, Katti et al. **Inteligência artificial: uma abordagem de aprendizado de máquina**. 2021. Citado 8 vezes nas páginas 28, 29, 42, 44 e 51.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery in Databases. **AI Magazine**, [S. l.], v. 17, n. 3, p. 37, 1996. Disponível em: https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/1230. Acesso em: 18 set. 2023. Citado 3 vezes na página 24.

GARCÍA, Enrique et al. A collaborative educational association rule mining tool. **The Internet and Higher Education**, v. 14, n. 2, p. 77-88, 2011. Citado 2 vezes nas páginas 25 e 26.

GÉRON, Aurelien. **Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems**. 2019. Citado 7 vezes nas páginas 31, 35, 36, 37 e 40.

GOMES, T.; GOUVEIA, R.; BATISTA, M. C. **Dados Educacionais Abertos: Associações em dados dos inscritos do Exame Nacional do Ensino Médio**. Anais do XXIII Workshop de Informática na Escola (WIE 2017). **Anais**...Brasil: Sociedade Brasileira de Computação - SBC, 27 out. 2017. Disponível em:<https://sol.sbc.org.br/index.php/wie/article/view/16325>. Acesso em: 17 jul. 2023. Citado 3 vezes nas páginas 62, 65 e 80.

GRAHNE, Gösta; ZHU, Jianfei. Efficiently using prefix-trees in mining frequent itemsets. In: FIMI. 2003. Disponível em: <https://ceur-ws.org/Vol-90/grahne.pdf>. Acesso em: 19 out. 2023. Citado na página 32.

HAMERLY, Greg; ELKAN, Charles. Learning the k in k-means. **Advances in neural information processing systems**, v. 16, 2003. Citado na página 35.

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. Data mining concepts and techniques third edition. **University of Illinois at Urbana-Champaign Micheline Kamber Jian Pei Simon Fraser University**, 2012. Citado 7 vezes nas páginas 29, 34, 35, 45, 49 e 73.

HAN, J.; PEI, J.; YIN, Y. **Mining frequent patterns without candidate generation**. ACM SIGMOD Record, v. 29, n. 2, p. 1–12, jun. 2000. Citado na página 32.

HIPP, J.; GÜNTZER, U.; NAKHAEIZADEH, G. Algorithms for association rule mining - a general survey and comparison. **ACM SIGKDD Explorations Newsletter**, v. 2, n. 1, p. 58–64, jun. 2000. Citado 2 vezes nas páginas 43 e 45.

INEP. **Leia-me Enem 2022**. Brasília: Inep, 2022. Disponível em: < https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem>. Acesso em: Acesso em: 11 jul. 2023. Citado 4 vezes nas páginas 15, 17, 23 e 70.

INEP. **3,9 milhões estão inscritos no Enem 2023**. Brasília, DF: Ministério da Educação, 2021. Disponível em: <https://www.gov.br/inep/pt-br/assuntos/noticias/enem/3-9-milhoes-estao-inscritos-no-enem-2023>. Acesso em: 17 set. 2023. Citado na página 21.

INEP. **Painel apresenta pesquisa sobre estudo na pandemia**. Brasília, DF: Ministério da Educação, 2023. Disponível em: <https://www.gov.br/inep/pt-br/assuntos/noticias/enem/painel-apresenta-pesquisa-sobre-estudo-na-pandemia>. Acesso em: 17 set. 2023. Citado 2 vezes nas páginas 23 e 24.

JARDIM, V. C.; BUCKERIDGE, M. S. Análise sistêmica do município de São Paulo e suas implicações para o avanço dos casos de Covid-19. Estudos Avançados, v. 34, n. 99, p. 157–174, ago. 2020. Citado na página 16.

KAUFMAN, L.; ROUSSEEUW, P. J. **Finding Groups in Data: An Introduction to Cluster Analysis**. 1. ed. [s.l.] Wiley, 1990. Citado 2 vezes nas páginas 38 e 39.

KELLEHER, John D.; MAC NAMEE, Brian; D'ARCY, Aoife. **Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies**. MIT press, 2020. Citado na página 29.

KISSLER, Stephen et al. Social distancing strategies for curbing the COVID-19 epidemic. MedRxiv, p. 2020.03. 22.20041079, 2020. Citado na página 16.

KÖCHE, J. C. **Fundamentos de metodologia científica**. Petrópolis: Editora Vozes, 2016. Disponível em: <http://www.adm.ufrpe.br/sites/ww4.deinfo.ufrpe.br/files/Fundamentos_de_ Metodologia_Cienti%CC%81fica.pdf>. Acesso em: 21 jul. 2023. Citado na página 18.

KOEDINGER, Kenneth et al. An open repository and analysis tools for fine-grained, longitudinal learner data. In: **Educational data mining**. 2008. Citado na página 25.

LEAL, Arthur. Enem: especialistas analisam queda de inscritos nos últimos anos. **O Globo**. Disponível em: <oglobo.globo.com/brasil/educacao/enem-e-vestibular/noticia/2022/11/enem-especialistas-analisam-queda-de-inscritos-nos-ultimos-anos-bolsonaro-tratou-universidades-como-inimigas.ghtml>. Acesso em: 21 set. 2023. Citado na página 20.

LEONI, R. C.; SAMPAIO, N. A. D. S. Desempenho das escolas públicas e privadas da região do Vale do Paraíba: uma aplicação da técnica de agrupamentos kmeans com base nas variáveis do Enem 2015. **Cadernos do IME - Série Estatística**, v. 42, n. 0, p. 31, 23 out. 2017. Citado 3 vezes nas páginas 59, 65 e 80.

LIMA, A. M. S. et al. Analysis of ENEM's attendants between 2012 and 2017 using a clustering approach. **Journal of Information and Data Management**, v. 11, n. 2, 2020. Citado 3 vezes nas páginas 58, 64 e 79.

MACQUEEN, James et al. Some methods for classification and analysis of multivariate observations. In: **Proceedings of the fifth Berkeley symposium on mathematical statistics and probability**. 1967. p. 281-297. Citado 2 vezes nas páginas 31 e 33.

MARIANO, Marcos Alves. **Comparação de Algoritmos Paralelos para a Extração de Regras de Associação no Modelo de Memória Distribuída**. 2011. Disponível em: <https://repositorio.ufms.br/handle/123456789/1674>. Acesso em 23 out. 2023. Citado 2 vezes nas páginas 47 e 50.

MATTE, Marcelo Kuchar; DO CARMO NICOLETTI, Maria. Revisão de Estratégias para a Aceleração do Algoritmo k-Means. In: **Anais do Workshop em Computação da FACCAMP**. 2019. p. 1-6. Citado na página 34.

MITCHELL, T. M. **Machine Learning**. New York: McGraw-Hill, 1997. Citado na página 28.

MONARD, Maria Carolina; BARANAUSKAS, José Augusto. Conceitos sobre aprendizado de máquina. Sistemas inteligentes-Fundamentos e aplicações, v. 1, n. 1, p. 32, 2003. Disponível em: <https://dcm.ffclrp.usp.br/~augusto/publications/2003-sistemas-inteligentes-cap4.pdf>. Acesso em: 15 out. 2023. Citado na página 28.

MÜLLER, Andreas C.; GUIDO, Sarah. **Introduction to machine learning with Python: a guide for data scientists**. " O'Reilly Media, Inc.", 2016. Citado na página 39.

NG, Raymond T..; HAN, Jiawei. CLARANS: A method for clustering objects for spatial data mining. **IEEE transactions on knowledge and data engineering**, v. 14, n. 5, p. 1003-1016, 2002. Citado na página 39.

OZDEMIR, Sinan. Principles of data science. Packt Publishing Ltd, 2016. Citado na página 43.

PATLOLLA. R. Understanding the concept of Hierarchical clustering Technique. Disponível em: <https://towardsdatascience.com/understanding-the-concept-of-hierarchical-clustering-technique-c6e8243758ec>. Acesso em: 21 out. 2023). Citado na página 42.

PYTHON, R. K-Means Clustering in Python: A Practical Guide – Real Python. Disponível em: <https://realpython.com/k-means-clustering-python/>. Acesso em: 21 out. 2023. Citado na página 42.

PELLEG, Dan et al. X-means: Extending k-means with efficient estimation of the number of clusters. In: **Icml**. 2000. p. 727-734. Citado na página 35.

PEREIRA JUNIOR, L.; NASSER MATOS, S.; BRONOSKI BORGES, H. Análise dos Perfis de Alunos do Ensino Superior sobre a Realização de Aulas na Modalidade a Distância Durante Pandemia da Covid-19 Usando Algoritmos de Aprendizagem de Máquina. **RENOTE**, v. 18, n. 2, p. 336–345, 4 jan. 2021. Citado 5 vezes nas páginas 56, 57, 64, 66 e 79.

PRODANOV, C. C.; DE FREITAS, E. C.. **Metodologia do trabalho científico: métodos e técnicas da pesquisa e do trabalho acadêmico**-2ª Edição. Editora Feevale, 2013. Citado na página 42.

RENJITH, Shini; SREEKUMAR, A.; JATHAVEDAN, M. An empirical research and comparative analysis of clustering performance for processing categorical and numerical data extracts from social media. **Acta Scientiarum: Technology**, v. 44, 2022. Citado 2 vezes nas páginas 38 e 39.

RODRIGUES, A. Presidente do Inep anuncia cancelamento do Enem Digital. Agência Brasil, Brasília, 08 de mar. de 2023. Disponível em: <https://agenciabrasil.ebc.com.br/educacao/noticia/2023-03/presidente-do-inep-anuncia-cancelamento-do-enem-digital>. Acesso em: 06 ago. 2023. Citado na página 20.

RODRIGUES, Marcos Wander; ISOTANI, Seiji; ZARATE, Luiz Enrique. Educational Data Mining: A review of evaluation process in the e-learning. Telematics and Informatics, v. 35, n. 6, p. 1701-1717, 2018. Citado na página 25.

ROMERO, C.; VENTURA, S. Data mining in education. **WIREs Data Mining and Knowledge Discovery**, v. 3, n. 1, p. 12–27, jan. 2013. Disponível em: <https://dl.acm.org/doi/10.1002/widm.1075>. Acesso em: 19 out. 2023. Citado 2 vezes nas páginas 25 e 32.

ROUSSEEUW, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. **Journal of Computational and Applied Mathematics**, 20, 1987. p. 53-65. Citado na página 36.

SAMMUT, C.; WEBB, G. I. (EDS.). **Encyclopedia of machine learning and data mining**. Second edition ed. New York, NY: Springer, 2017. Citado na página 40.

SHEIKH, A.; GHANBARPOUR, T.; GHOLAMIANGONABADI, D. A Preliminary Study of Fintech Industry: A Two-Stage Clustering Analysis for Customer Segmentation in the B2B Setting. **Journal of Business-to-Business Marketing**, v. 26, n. 2, p. 197–207, 2019. Citado na página 37.

SILVEIRA, F. L.; BARBOSA, M. C. B.; SILVA, R. Exame Nacional do Ensino Médio (ENEM): uma análise crítica. **Revista Brasileira de Ensino de Física**, v. 37, p. 1101, 2015. Citado na página 21.

SOARES, R. D. A.; SILVA, G. A. Regulamentos da EaD no Brasil e o Impacto da Portaria No 343/2020 no Ensino Superior. **EaD em Foco**, v. 10, n. 3, 5 ago. 2020. Citado na página 16.

SILVA, V. A. A. D. et al. **Identificação de Desigualdades Sociais a partir do desempenho dos alunos do Ensino Médio no ENEM 2019 utilizando Mineração de Dados**. In: Anais do XXXI Simpósio Brasileiro de Informática na Educação. SBC, 2020. p. 72-81. Citado 4 vezes nas páginas 60, 61, 65 e 80.

SILVA, L. A.; PERES, S. M.; BOSCARIOLI, C. **Introdução à Mineração de Dados: com Aplicações em R**. São Paulo: Elsevier, 2016. Citado 4 vezes nas páginas 30, 31 e 76.

SRIVASTAVA, Ms S.; JOSHI, Ms N.; GAUR, M. A review paper on feature selection methodologies and their applications. **IJCSNS**, v. 14, n. 5, p. 78, 2014. Disponível em: <http://paper.ijcsns.org/07_book/201405/20140514.pdf>. Acesso em: 15 ago. 2023. Citado na página 28.

TIBSHIRANI, R.; WALTHER, G.; HASTIE, T. Estimating the Number of Clusters in a Data Set Via the Gap Statistic. Journal of the Royal Statistical Society Series B: Statistical Methodology, v. 63, n. 2, p. 411–423, 1 jul. 2001. Citado na página 59.

WEBER NETO, N. et al. **Análise Exploratória de Dados para Identificar o Impacto da Pandemia da COVID-19 no ENEM dos Estados do Ceará, Maranhão e Piauí**. Anais da X Escola Regional de Computação do Ceará, Maranhão e Piauí (ERCEMAPI 2022). **Anais**...Brasil: Sociedade Brasileira de Computação - SBC, 28 set. 2022a. Disponível em: <https://sol.sbc.org.br/index.php/ercemapi/article/view/21957>. Acesso em: 14 jan. 2023. Citado 5 vezes nas páginas 53, 55, 64, 65 e 79.

WEBER NETO, N.; C. SOARES, R.; REIS COUTINHO, L.; SOARES TELES, A. **A Pandemia da COVID-19 impactou o ENEM? Uma Análise Comparativa de Dados dos Anos de 2019 e 2020**. Revista Novas Tecnologias na Educação, Porto Alegre, v. 20, n. 1, p. 223–232, 2022b. Disponível em: https://seer.ufrgs.br/index.php/renote/article/view/126655. Acesso em: 19 jul. 2023. Citado 5 vezes nas páginas 54, 55, 64, 66 e 79.

WEBER NETO, N. et al. Data Analysis to Identify the Impact of the COVID-19 Pandemic on ENEM in 3 States of Northeast Brazil. **Revista de Sistemas e Computação - RSC**, v. 12, n. 3, 1 jun. 2023. Citado 4 vezes nas páginas 55, 64, 66 e 79.

WORLD HEALTH ORGANIZATION. **Coronavirus disease 2019 (COVID-19)**: situation report, 43. [s.l.] World Health Organization, 3 mar. 2020. Disponível em: <https://apps.who.int/iris/handle/10665/331354>. Acesso em: 27 jul. 2023. Citado na página 16.

YAN, Fuwu et al. Driving style recognition based on electroencephalography data from a simulated driving experiment. **Frontiers in psychology**, v. 10, p. 1254, 2019. Citado na página 38.

ZAKI, M. J. Scalable algorithms for association mining. **IEEE transactions on knowledge and data engineering**, v. 12, n. 3, p. 372-390, 2000. Citado na página 32.

ZHENG, A.; CASARI, A. **Feature engineering for machine learning: principles and techniques for data scientists**. First edition ed. Beijing : Boston: O'Reilly, 2018. Citado na página 25.

APÊNDICES

APÊNDICE A – LISTA COMPLETA DE ATRIBUTOS CATEGORIZADOS

O Quadro 4.4 contém a lista de atributos completa do dataset ENEM_HE que foi usado como entrada para a geração dos modelos de AM não supervisionados por meio da RSL e dos dados de Hábitos de Estudo (HE), organizados por Grupos Temáticos (GT).

Quadro 4.4 - Lista de atributos do dataset ENEM_HE.

RSL + GRUPOS TEMÁTICOS DE HE	**VARIÁVEL**
rsl_var (Enem: atributos relevantes RSL)	TP_FAIXA_ETARIA=faixaAdultosIdosos
	TP_FAIXA_ETARIA=faixaMédio
	TP_FAIXA_ETARIA=faixaProfissional
	TP_SEXO=femino
	TP_SEXO=masculino
	TP_COR_RACA=amarela
	TP_COR_RACA=branca
	TP_COR_RACA=nãoDeclarado
	TP_COR_RACA=pretaPardaIndígena
	NIVEL_ESC_PAI=completouAtéMédio
	NIVEL_ESC_PAI=completouSuperior
	NIVEL_ESC_PAI=nãoCompletouMédio
	NIVEL_ESC_PAI=nãoSabe
	NIVEL_ESC_MAE=completouAtéMédio
	NIVEL_ESC_MAE=completouSuperior
	NIVEL_ESC_MAE=nãoCompletouMédio
	NIVEL_ESC_MAE=nãoSabe
	RENDA_FAMILIAR=acimaDe3636
	RENDA_FAMILIAR=até1818
	RENDA_FAMILIAR=até3636
	NU_CELULAR=nãoTem
	NU_CELULAR=tem
	NU_COMPUTADOR=nãoTem
	NU_COMPUTADOR=tem
	ACESSO_INTERNET=nãoTem
	ACESSO_INTERNET=tem
apr_matr_sit_var (SITUAÇÃO DE MATRÍCULA ESCOLAR E PERCEÇÃO DA PRÓPRIA APRENDIZAGEM)	APR_MATR_SIT_MEDIO=ensinoEJA
	APR_MATR_SIT_MEDIO=ensinoProfissional
	APR_MATR_SIT_MEDIO=ensinoRegular
	APR_MATR_VINCULO=interrompeu
	APR_MATR_VINCULO=nãoInterrompeu
	APR_MATR_TP_ESTUDO=apenasPresencial
	APR_MATR_TP_ESTUDO=apenasRemoto
	APR_MATR_TP_ESTUDO=híbrido
	APR_MATR_TP_ESTUDO=semMatricula
	APR_MATR_APREND_PANDEMIA=aprendeuContaPrópria
	APR_MATR_APREND_PANDEMIA=aprendeuMaisHíbrido
	APR_MATR_APREND_PANDEMIA=aprendeuMaisPresencial
	APR_MATR_APREND_PANDEMIA=aprendeuMaisRemoto

	APR_MATR_APREND_PANDEMIA=aprendeuRemotoEHíbrido
	APR_MATR_APREND_PANDEMIA=semMatriculaNãoEstudou
gest_temp_var (GESTÃO DO TEMPO E PLANEJAMENTO DE ESTUDOS)	GEST_TEMP_ATV_CRONOGR=muitasVezes
	GEST_TEMP_ATV_CRONOGR=nenhumaVez
	GEST_TEMP_ATV_CRONOGR=poucasVezes
	GEST_TEMP_ATV_CRONOGR=todasAsVezes
	GEST_TEMP_ATV_TEMPO=muitasVezes
	GEST_TEMP_ATV_TEMPO=nenhumaVez
	GEST_TEMP_ATV_TEMPO=poucasVezes
	GEST_TEMP_ATV_TEMPO=todasAsVezes
	GEST_TEMP_ATV_MATERIAL=muitasVezes
	GEST_TEMP_ATV_MATERIAL=nenhumaVez
	GEST_TEMP_ATV_MATERIAL=poucasVezes
	GEST_TEMP_ATV_MATERIAL=todasAsVezes
	GEST_TEMP_ATV_HORA_PROG=muitasVezes
	GEST_TEMP_ATV_HORA_PROG=nenhumaVez
	GEST_TEMP_ATV_HORA_PROG=poucasVezes
	GEST_TEMP_ATV_HORA_PROG=todasAsVezes
prat_est_var (PRÁTICAS DE ESTUDO E PESQUISA)	PRAT_EST_LER=muitasVezes
	PRAT_EST_LER=nenhumaVez
	PRAT_EST_LER=poucasVezes
	PRAT_EST_LER=todasAsVezes
	PRAT_EST_RESUM_TEXTO=muitasVezes
	PRAT_EST_RESUM_TEXTO=nenhumaVez
	PRAT_EST_RESUM_TEXTO=poucasVezes
	PRAT_EST_RESUM_TEXTO=todasAsVezes
	PRAT_EST_RESUM_VIDEO=muitasVezes
	PRAT_EST_RESUM_VIDEO=nenhumaVez
	PRAT_EST_RESUM_VIDEO=poucasVezes
	PRAT_EST_RESUM_VIDEO=todasAsVezes
	PRAT_EST_ATV_FIXACAO=muitasVezes
	PRAT_EST_ATV_FIXACAO=nenhumaVez
	PRAT_EST_ATV_FIXACAO=poucasVezes
	PRAT_EST_ATV_FIXACAO=todasAsVezes
	PRAT_EST_ATV_AVALIACAO=muitasVezes
	PRAT_EST_ATV_AVALIACAO=nenhumaVez
	PRAT_EST_ATV_AVALIACAO=poucasVezes
	PRAT_EST_ATV_AVALIACAO=todasAsVezes
	PRAT_EST_DISTRACOES=muitasVezes
	PRAT_EST_DISTRACOES=nenhumaVez
	PRAT_EST_DISTRACOES=poucasVezes
	PRAT_EST_DISTRACOES=todasAsVezes
	PRAT_EST_ANOT_DUV_VIDEO=muitasVezes
	PRAT_EST_ANOT_DUV_VIDEO=nenhumaVez
	PRAT_EST_ANOT_DUV_VIDEO=poucasVezes
	PRAT_EST_ANOT_DUV_VIDEO=todasAsVezes
	PRAT_EST_ANOT_DUV_VIDEO_COMPLEMENTAR=muitasVezes
	PRAT_EST_ANOT_DUV_VIDEO_COMPLEMENTAR=nenhumaVez
	PRAT_EST_ANOT_DUV_VIDEO_COMPLEMENTAR=poucasVezes
	PRAT_EST_ANOT_DUV_VIDEO_COMPLEMENTAR=todasAsVezes

	PRAT_EST_ANOT_DUV_PROF=muitasVezes
	PRAT_EST_ANOT_DUV_PROF=nenhumaVez
	PRAT_EST_ANOT_DUV_PROF=poucasVezes
	PRAT_EST_ANOT_DUV_PROF=todasAsVezes
	PRAT_EST_ESTRUT_IDEIA_REDACAO=muitasVezes
	PRAT_EST_ESTRUT_IDEIA_REDACAO=nenhumaVez
	PRAT_EST_ESTRUT_IDEIA_REDACAO=poucasVezes
	PRAT_EST_ESTRUT_IDEIA_REDACAO=todasAsVezes
	PRAT_EST_TREINAR_REDACAO=muitasVezes
	PRAT_EST_TREINAR_REDACAO=nenhumaVez
	PRAT_EST_TREINAR_REDACAO=poucasVezes
	PRAT_EST_TREINAR_REDACAO=todasAsVezes
	PRAT_EST_PARTICIPAR_FORUM=muitasVezes
	PRAT_EST_PARTICIPAR_FORUM=nenhumaVez
	PRAT_EST_PARTICIPAR_FORUM=poucasVezes
	PRAT_EST_PARTICIPAR_FORUM=todasAsVezes
	GEST_TEMP_PONTUAL_AULA_ONLINE=muitasVezes
	GEST_TEMP_PONTUAL_AULA_ONLINE=nenhumaVez
	GEST_TEMP_PONTUAL_AULA_ONLINE=poucasVezes
	GEST_TEMP_PONTUAL_AULA_ONLINE=todasAsVezes
	GEST_TEMP_ASSID_AULA_ONLINE=muitasVezes
	GEST_TEMP_ASSID_AULA_ONLINE=nenhumaVez
	GEST_TEMP_ASSID_AULA_ONLINE=poucasVezes
	GEST_TEMP_ASSID_AULA_ONLINE=todasAsVezes
	PRAT_EST_REV_ANOT=muitasVezes
	PRAT_EST_REV_ANOT=nenhumaVez
	PRAT_EST_REV_ANOT=poucasVezes
	PRAT_EST_REV_ANOT=todasAsVezes
	PRAT_EST_REV_VIDEOAULA=muitasVezes
	PRAT_EST_REV_VIDEOAULA=nenhumaVez
	PRAT_EST_REV_VIDEOAULA=poucasVezes
	PRAT_EST_REV_VIDEOAULA=todasAsVezes
prob_rot_var(PROBLEMAS NA ROTINA DE ESTUDOS)	PROB_ROT_EST_PANDEMIA=não
	PROB_ROT_EST_PANDEMIA=sim
dif_infra_var (DIFICULDADES DE INFRAESTRUTURA)	DIF_INFRA_PANDEMIA=não
	DIF_INFRA_PANDEMIA=sim
ajud_terc_var (AJUDA DE TERCEIROS)	AJUD_TERC_PANDEMIA=ninguémAuxiliou
	AJUD_TERC_PANDEMIA=não
	AJUD_TERC_PANDEMIA=sim
autoav_var (AVALIAÇÃO SOBRE A PRÓPRIA EXPERIÊNCIA)	AUTOAV_PREPARACAO_APRENDIZ=bemPreparado
	AUTOAV_PREPARACAO_APRENDIZ=muitoPreparado
	AUTOAV_PREPARACAO_APRENDIZ=nadaPreparado
	AUTOAV_PREPARACAO_APRENDIZ=poucoPreparado
	AUTOAV_PREPARACAO_APRENDIZ=totalmentePreparado
desempenho_var (DESEMPENHO)	NOTA_MEDIA=baixo
	NOTA_MEDIA=médio
	NOTA_MEDIA=alto
RSL + GRUPOS TEMÁTICOS DE HE	**VARIÁVEL**

Fonte: Elaborado pelo autor, com base em dados do INEP (2022).

APÊNDICE B – PERFIS DE ESTUDANTES IDENTIFICADOS

O Quadro 5.3 contém a lista de atributos completa do dataset ENEM_HE e a identificação de perfis de estudantes por clusters obtidos por meio do K-means, com cenário de três grupos.

Quadro 5.3 - Perfis de estudantes identificados por clusters formados com K-means (K=3).

K-MEANS (K=3)				
RSL + GRUPOS TEMÁTICOS DE HE	**VARIÁVEL**	**CLUSTER 0**	**CLUSTER 1**	**CLUSTER 2**
rsl_var (Enem: atributos relevantes RSL)	TP_FAIXA_ETARIA=faixaAdultosIdosos	Não	Não	Não
	TP_FAIXA_ETARIA=faixaMédio	Não	Não	Não
	TP_FAIXA_ETARIA=faixaProfissional	Sim	Sim	Sim
	TP_SEXO=femino	Sim	Sim	Sim
	TP_SEXO=masculino	Não	Não	Não
	TP_COR_RACA=amarela	Não	Não	Não
	TP_COR_RACA=branca	Não	Não	Não
	TP_COR_RACA=nãoDeclarado	Não	Não	Não
	TP_COR_RACA=pretaPardaIndígena	Não	Não	Sim
	NIVEL_ESC_PAI=completouAtéMédio	Não	Não	Não
	NIVEL_ESC_PAI=completouSuperior	Não	Não	Não
	NIVEL_ESC_PAI=nãoCompletouMédio	Não	Não	Não
	NIVEL_ESC_PAI=nãoSabe	Não	Não	Não
	NIVEL_ESC_MAE=completouAtéMédio	Não	Não	Não
	NIVEL_ESC_MAE=completouSuperior	Não	Não	Não
	NIVEL_ESC_MAE=nãoCompletouMédio	Não	Não	Não
	NIVEL_ESC_MAE=nãoSabe	Não	Não	Não
	RENDA_FAMILIAR=acimaDe3636	Não	Não	Não
	RENDA_FAMILIAR=até1818	Não	Não	Sim
	RENDA_FAMILIAR=até3636	Não	Não	Não
	NU_CELULAR=nãoTem	Não	Não	Não
	NU_CELULAR=tem	Sim	Sim	Sim
	NU_COMPUTADOR=nãoTem	Não	Não	Sim
	NU_COMPUTADOR=tem	Sim	Sim	Não
	ACESSO_INTERNET=nãoTem	Não	Não	Não
	ACESSO_INTERNET=tem	Sim	Sim	Sim
apr_matr_sit_var (SITUAÇÃO DE MATRÍCULA ESCOLAR E PERCEÇÃO DA PRÓPRIA APRENDIZAGEM)	APR_MATR_SIT_MEDIO=ensinoEJA	Não	Não	Não
	APR_MATR_SIT_MEDIO=ensinoProfissional	Não	Não	Não
	APR_MATR_SIT_MEDIO=ensinoRegular	Sim	Sim	Sim
	APR_MATR_VINCULO=interrompeu	Não	Não	Não
	APR_MATR_VINCULO=nãoInterrompeu	Sim	Sim	Sim
	APR_MATR_TP_ESTUDO=apenasPresencial	Não	Não	Não
	APR_MATR_TP_ESTUDO=apenasRemoto	Não	Não	Não
	APR_MATR_TP_ESTUDO=híbrido	Sim	Sim	Não
	APR_MATR_TP_ESTUDO=semMatricula	Não	Não	Não
	APR_MATR_APREND_PANDEMIA=aprendeuContaPrópria	Não	Não	Não
	APR_MATR_APREND_PANDEMIA=aprendeuMaisHíbrido	Não	Não	Não
	APR_MATR_APREND_PANDEMIA=aprendeuMaisPresencial	Sim	Sim	Sim
	APR_MATR_APREND_PANDEMIA=aprendeuMaisRemoto	Não	Não	Não
	APR_MATR_APREND_PANDEMIA=aprendeuRemotoEHíbrido	Não	Não	Não
	APR_MATR_APREND_PANDEMIA=semMatriculaNãoEstudou	Não	Não	Não
gest_temp_1_var (GESTÃO DO TEMPO E PLANEJAMENTO DE ESTUDOS)	GEST_TEMP_ATV_CRONOGR=muitasVezes	Não	Não	Não
	GEST_TEMP_ATV_CRONOGR=nenhumaVez	Não	Sim	Não
	GEST_TEMP_ATV_CRONOGR=poucasVezes	Não	Não	Sim
	GEST_TEMP_ATV_CRONOGR=todasAsVezes	Não	Não	Não
	GEST_TEMP_ATV_TEMPO=muitasVezes	Sim	Não	Não
	GEST_TEMP_ATV_TEMPO=nenhumaVez	Não	Sim	Não

(1ª PARTE)	GEST_TEMP_ATV_TEMPO=poucasVezes	Não	Não	Sim
	GEST_TEMP_ATV_TEMPO=todasAsVezes	Não	Não	Não
	GEST_TEMP_ATV_MATERIAL=muitasVezes	Sim	Não	Não
	GEST_TEMP_ATV_MATERIAL=nenhumaVez	Não	Sim	Não
	GEST_TEMP_ATV_MATERIAL=poucasVezes	Não	Não	Sim
	GEST_TEMP_ATV_MATERIAL=todasAsVezes	Não	Não	Não
	GEST_TEMP_ATV_HORA_PROG=muitasVezes	Sim	Não	Não
	GEST_TEMP_ATV_HORA_PROG=nenhumaVez	Não	Sim	Não
	GEST_TEMP_ATV_HORA_PROG=poucasVezes	Não	Não	Sim
	GEST_TEMP_ATV_HORA_PROG=todasAsVezes	Não	Não	Não
prat_est_1_var (PRÁTICAS DE ESTUDO E PESQUISA) (1ª PARTE)	PRAT_EST_LER=muitasVezes	Sim	Não	Não
	PRAT_EST_LER=nenhumaVez	Não	Sim	Não
	PRAT_EST_LER=poucasVezes	Não	Não	Sim
	PRAT_EST_LER=todasAsVezes	Não	Não	Não
	PRAT_EST_RESUM_TEXTO=muitasVezes	Sim	Não	Não
	PRAT_EST_RESUM_TEXTO=nenhumaVez	Não	Sim	Não
	PRAT_EST_RESUM_TEXTO=poucasVezes	Não	Não	Sim
	PRAT_EST_RESUM_TEXTO=todasAsVezes	Não	Não	Não
	PRAT_EST_RESUM_VIDEO=muitasVezes	Sim	Não	Não
	PRAT_EST_RESUM_VIDEO=nenhumaVez	Não	Sim	Não
	PRAT_EST_RESUM_VIDEO=poucasVezes	Não	Não	Sim
	PRAT_EST_RESUM_VIDEO=todasAsVezes	Não	Não	Não
	PRAT_EST_ATV_FIXACAO=muitasVezes	Sim	Não	Não
	PRAT_EST_ATV_FIXACAO=nenhumaVez	Não	Sim	Não
	PRAT_EST_ATV_FIXACAO=poucasVezes	Não	Não	Sim
	PRAT_EST_ATV_FIXACAO=todasAsVezes	Não	Não	Não
	PRAT_EST_ATV_AVALIACAO=muitasVezes	Não	Não	Não
	PRAT_EST_ATV_AVALIACAO=nenhumaVez	Não	Sim	Não
	PRAT_EST_ATV_AVALIACAO=poucasVezes	Não	Não	Sim
	PRAT_EST_ATV_AVALIACAO=todasAsVezes	Não	Não	Não
	PRAT_EST_DISTRACOES=muitasVezes	Não	Não	Não
	PRAT_EST_DISTRACOES=nenhumaVez	Não	Sim	Não
	PRAT_EST_DISTRACOES=poucasVezes	Não	Não	Sim
	PRAT_EST_DISTRACOES=todasAsVezes	Não	Não	Não
	PRAT_EST_ANOT_DUV_VIDEO=muitasVezes	Sim	Não	Não
	PRAT_EST_ANOT_DUV_VIDEO=nenhumaVez	Não	Sim	Não
	PRAT_EST_ANOT_DUV_VIDEO=poucasVezes	Não	Não	Sim
	PRAT_EST_ANOT_DUV_VIDEO=todasAsVezes	Não	Não	Não
	PRAT_EST_ANOT_DUV_VIDEO_COMPLEMENTAR=muitasVezes	Sim	Não	Não
	PRAT_EST_ANOT_DUV_VIDEO_COMPLEMENTAR=nenhumaVez	Não	Sim	Não
	PRAT_EST_ANOT_DUV_VIDEO_COMPLEMENTAR=poucasVezes	Não	Não	Sim
	PRAT_EST_ANOT_DUV_VIDEO_COMPLEMENTAR=todasAsVezes	Não	Não	Não
	PRAT_EST_ANOT_DUV_PROF=muitasVezes	Não	Não	Não
	PRAT_EST_ANOT_DUV_PROF=nenhumaVez	Não	Sim	Não
	PRAT_EST_ANOT_DUV_PROF=poucasVezes	Não	Não	Sim
	PRAT_EST_ANOT_DUV_PROF=todasAsVezes	Não	Não	Não
	PRAT_EST_ESTRUT_IDEIA_REDACAO=muitasVezes	Sim	Não	Não
	PRAT_EST_ESTRUT_IDEIA_REDACAO=nenhumaVez	Não	Sim	Não
	PRAT_EST_ESTRUT_IDEIA_REDACAO=poucasVezes	Não	Não	Sim
	PRAT_EST_ESTRUT_IDEIA_REDACAO=todasAsVezes	Não	Não	Não
	PRAT_EST_TREINAR_REDACAO=muitasVezes	Sim	Não	Não
	PRAT_EST_TREINAR_REDACAO=nenhumaVez	Não	Não	Não
	PRAT_EST_TREINAR_REDACAO=poucasVezes	Não	Não	Sim
	PRAT_EST_TREINAR_REDACAO=todasAsVezes	Não	Não	Não
	PRAT_EST_PARTICIPAR_FORUM=muitasVezes	Não	Não	Não
	PRAT_EST_PARTICIPAR_FORUM=nenhumaVez	Não	Sim	Não
	PRAT_EST_PARTICIPAR_FORUM=poucasVezes	Não	Não	Sim
	PRAT_EST_PARTICIPAR_FORUM=todasAsVezes	Não	Não	Não

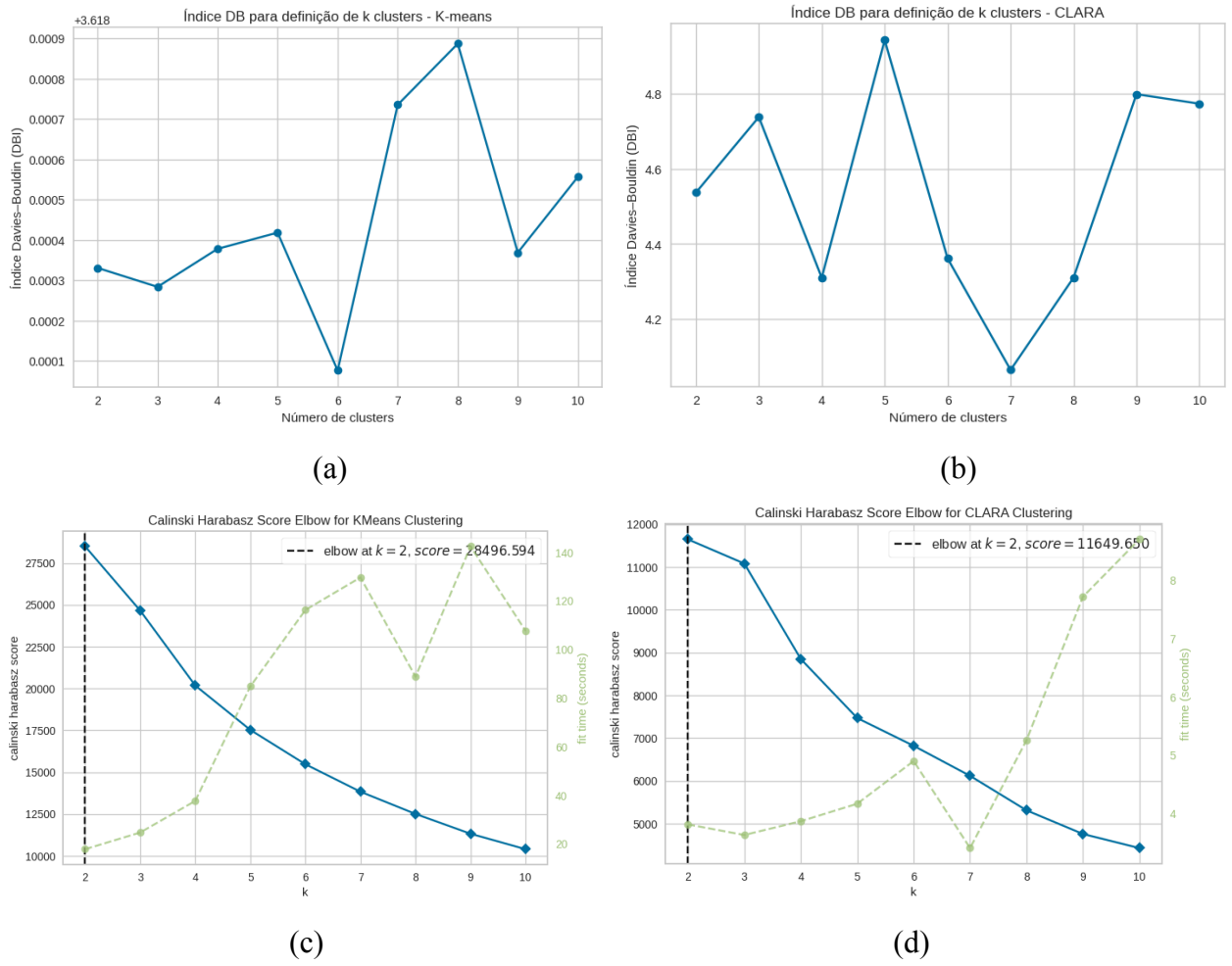
gest_temp_2_var (GESTÃO DO TEMPO E PLANEJAMENTO DE ESTUDOS) (2ª PARTE)	GEST_TEMP_PONTUAL_AULA_ONLINE=muitasVezes	Não	Não	Não
	GEST_TEMP_PONTUAL_AULA_ONLINE=nenhumaVez	Não	Sim	Não
	GEST_TEMP_PONTUAL_AULA_ONLINE=poucasVezes	Não	Não	Não
	GEST_TEMP_PONTUAL_AULA_ONLINE=todasAsVezes	Não	Não	Não
	GEST_TEMP_ASSID_AULA_ONLINE=muitasVezes	Sim	Não	Não
	GEST_TEMP_ASSID_AULA_ONLINE=nenhumaVez	Não	Não	Não
	GEST_TEMP_ASSID_AULA_ONLINE=poucasVezes	Não	Não	Sim
	GEST_TEMP_ASSID_AULA_ONLINE=todasAsVezes	Não	Não	Não
prat_est_2_var (PRÁTICAS DE ESTUDO E PESQUISA) (2ª PARTE)	PRAT_EST_REV_ANOT=muitasVezes	Sim	Não	Não
	PRAT_EST_REV_ANOT=nenhumaVez	Não	Sim	Não
	PRAT_EST_REV_ANOT=poucasVezes	Não	Não	Sim
	PRAT_EST_REV_ANOT=todasAsVezes	Não	Não	Não
	PRAT_EST_REV_VIDEOAULA=muitasVezes	Não	Não	Não
	PRAT_EST_REV_VIDEOAULA=nenhumaVez	Não	Sim	Não
	PRAT_EST_REV_VIDEOAULA=poucasVezes	Não	Não	Sim
	PRAT_EST_REV_VIDEOAULA=todasAsVezes	Não	Não	Não
prob_rot_var (PROB_ROT_EST_PANDEMIA)	PROB_ROT_EST_PANDEMIA=não	Não	Não	Não
	PROB_ROT_EST_PANDEMIA=sim	Sim	Sim	Sim
dif_infra_var (DIFICULDADES DE INFRAESTRUTURA)	DIF_INFRA_PANDEMIA=não	Sim	Sim	Sim
	DIF_INFRA_PANDEMIA=sim	Não	Não	Não
ajud_terc_var (AJUDA DE TERCEIROS)	AJUD_TERC_PANDEMIA=ninguémAuxiliou	Não	Não	Não
	AJUD_TERC_PANDEMIA=não	Não	Não	Não
	AJUD_TERC_PANDEMIA=sim	Não	Não	Não
autoav_var (AVALIAÇÃO SOBRE A PRÓPRIA EXPERIÊNCIA)	AUTOAV_PREPARACAO_APRENDIZ=bemPreparado	Não	Não	Não
	AUTOAV_PREPARACAO_APRENDIZ=muitoPreparado	Não	Não	Não
	AUTOAV_PREPARACAO_APRENDIZ=nadaPreparado	Não	Não	Não
	AUTOAV_PREPARACAO_APRENDIZ=poucoPreparado	Não	Sim	Sim
	AUTOAV_PREPARACAO_APRENDIZ=totalmentePreparado	Não	Não	Não
desempenho_var (DESEMEPENHO)	NOTA_MEDIA=baixo	Não	Não	Sim
	NOTA_MEDIA=médio	Não	Sim	Não
	NOTA_MEDIA=alto	Sim	Não	Não
RSL + GRUPOS TEMÁTICOS DE HE	**VARIÁVEL**	**CLUSTER 0**	**CLUSTER 1**	**CLUSTER 2**

Fonte: Elaborado pelo autor, com base em dados do INEP (2022).

APÊNDICE C – MEDIDAS DE AVALIAÇÃO E CENÁRIOS DE AGRUPAMENTO

Para consolidação do número ideal de grupos, as medidas de IHC e IDB são mostradas na Figura 4.5.

Figura 4.5 - Resultado do número ideal de clusters definido pelos IHC e IDB.



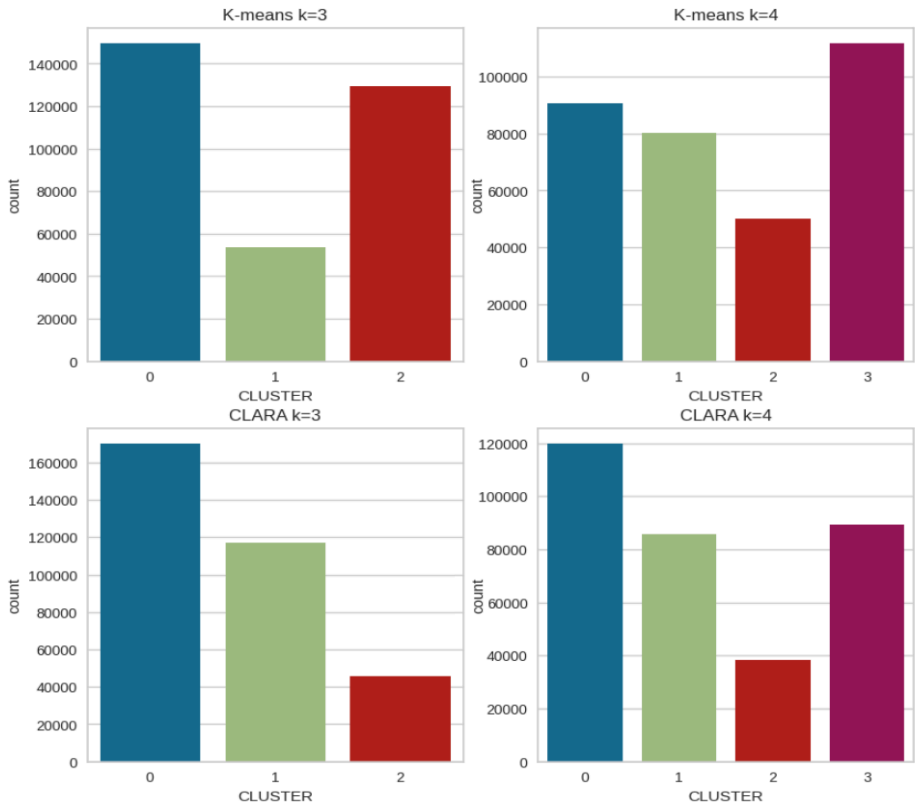
Fonte: Elaborado pelo autor.

O IDB indica que o número ideal (menor índice) de cluster é 6 para o K-means e 7 para o CLARA. Outro ponto relevante considerado é a segunda forte indicação de 3 clusters para o K-means e 4 clusters para o CLARA. Para o IHC a indicação é de 2 clusters. Esses resultados variados podem ser explicados devido à similaridade intraclusters, cujo coeficiente de silhueta indicou de forma mais evidente que muitas instâncias não são tão dissimilares entre os clusters. Também há muitas variáveis no dataset cujos valores se limitam a poucos valores possíveis.

Mediante as indicações apontadas pelas medidas entre cada um dos algoritmos no experimento 1, definiu-se avaliar os resultados dos perfis de estudantes considerando para o

K-means e o CLARA, 3 e 4 clusters. A Figura 4.6 mostra a distribuição de instâncias por cada grupo formado.

Figura 4.6 - Distribuição de instâncias por grupos dos algoritmos divisivos.



Fonte: Elaborado pelo autor.

Após o agrupamento, conforme o cenário de K cluster para cada algoritmo, as instâncias foram rotuladas em uma variável (CLUSTER) conforme o respectivo valor: 0, 1, 2 ou 3. Percebe-se que há um equilíbrio de distribuição de entre os grupos formados quando comparados os algoritmos K-means e CLARA. A Tabela 4.1 apresenta em detalhes a distribuição de estudantes nos cenários obtidos com os algoritmos divisivos, considerando o total de 332.793 instâncias.

Tabela 4.1 - Resultados das atribuições dos estudantes em cada cluster.

Algoritmo	**Total de clusters**	**Cluster 0**	**Cluster 1**	**Cluster 2**	**Cluster 3**
K-means	3	149373 (44,88%)	53857 (16,18%)	129563 (38,93%)	-
	4	90942 (27,33%)	80704 (24,25%)	50057 (15,04%)	111090 (33,38%)
CLARA	3	170160 (51,13%)	117052 (35,17%)	45581 (13,70%)	-
	4	120019 (36,06%)	85154 (25,59%)	38911 (11,69%)	88709 (26,66%)

Fonte: Elaborado pelo autor.

APÊNDICE D – PERFIS DE ESTUDANTES IDENTIFICADOS (CLARA)

A Figura 5.4 (cenário CLARA com K=3) mostra a distribuição dos clusters considerando todas variáveis do dataset ENEM_HE. Os dados de respostas específicas de cada grupo formado estão disponíveis no material suplementar[12].

Figura 5.4 - Distribuição de grupos com CLARA.

CLARA (n_cluster=3)				
RSL + GRUPOS TEMÁTICOS DE HE	**VARIÁVEL**	**CLUSTER 0**	**CLUSTER 1**	**CLUSTER 2**
Atributos relevantes da RSL	TP_FAIXA_ETARIA	faixaProfissional	faixaProfissional	faixaProfissional
	TP_SEXO	femino	femino	femino
	TP_COR_RACA	pretaPardaIndígena	branca	pretaPardaIndígena
	NIVEL_ESC_PAI	nãoCompletouMédio	completouAtéMédio	nãoCompletouMédio
	NIVEL_ESC_MAE	nãoCompletouMédio	completouAtéMédio	completouAtéMédio
	RENDA_FAMILIAR	até1818	acimaDe3636	até1818
	NU_CELULAR	tem	tem	tem
	NU_COMPUTADOR	nãoTem	tem	tem
	ACESSO_INTERNET	tem	tem	tem
Situação de matrícula escolar e percepção da própria aprendizagem	APR_MATR_SIT_MEDIO	ensinoRegular	ensinoRegular	ensinoRegular
	APR_MATR_VINCULO	nãoInterrompeu	nãoInterrompeu	nãoInterrompeu
	APR_MATR_TP_ESTUDO	híbrido	híbrido	híbrido
	APR_MATR_APREND_PANDEMIA	aprendeuMaisPresencial	aprendeuMaisPresencial	aprendeuMaisPresencial
Gestão do tempo e planejamento de estudos (1ª parte)	GEST_TEMP_ATV_CRONOGR	poucasVezes	muitasVezes	nenhumaVez
	GEST_TEMP_ATV_TEMPO	poucasVezes	muitasVezes	nenhumaVez
	GEST_TEMP_ATV_MATERIAL	poucasVezes	muitasVezes	nenhumaVez
	GEST_TEMP_ATV_HORA_PROG	poucasVezes	muitasVezes	nenhumaVez
Práticas de estudo e pesquisa (1ª parte)	PRAT_EST_LER	poucasVezes	muitasVezes	nenhumaVez
	PRAT_EST_RESUM_TEXTO	poucasVezes	muitasVezes	nenhumaVez
	PRAT_EST_RESUM_VIDEO	poucasVezes	muitasVezes	nenhumaVez
	PRAT_EST_ATV_FIXACAO	poucasVezes	muitasVezes	nenhumaVez
	PRAT_EST_ATV_AVALIACAO	poucasVezes	muitasVezes	nenhumaVez
	PRAT_EST_DISTRACOES	poucasVezes	muitasVezes	nenhumaVez
	PRAT_EST_ANOT_DUV_VIDEO	poucasVezes	muitasVezes	nenhumaVez
	PRAT_EST_ANOT_DUV_VIDEO_COMPL	poucasVezes	muitasVezes	nenhumaVez
	PRAT_EST_ANOT_DUV_PROF	poucasVezes	muitasVezes	nenhumaVez
	PRAT_EST_ESTRUT_IDEIA_REDACAO	poucasVezes	muitasVezes	nenhumaVez
	PRAT_EST_TREINAR_REDACAO	poucasVezes	muitasVezes	nenhumaVez
	PRAT_EST_PARTICIPAR_FORUM	poucasVezes	poucasVezes	nenhumaVez
Gestão do tempo e planejamento de estudos (2ª parte)	GEST_TEMP_PONTUAL_AULA_ONLINE	poucasVezes	muitasVezes	nenhumaVez
	GEST_TEMP_ASSID_AULA_ONLINE	poucasVezes	muitasVezes	nenhumaVez
Práticas de estudo e pesquisa (2ª parte)	PRAT_EST_REV_ANOT	poucasVezes	muitasVezes	nenhumaVez
	PRAT_EST_REV_VIDEOAULA	poucasVezes	muitasVezes	nenhumaVez
Problemas na rotina de estudos	PROB_ROT_EST_PANDEMIA	sim	sim	sim
Dificuldades de infraestrutura	DIF_INFRA_PANDEMIA	sim	não	não
Ajuda de terceiros	AJUD_TERC_PANDEMIA	sim	sim	ninguémAuxiliou
Avaliação sobre a própria experiência	AUTOAV_PREPARACAO_APRENDIZ	poucoPreparado	bemPreparado	poucoPreparado
DESEMPENHO	NOTA_MEDIA	baixo	alto	médio

Fonte: Elaborado pelo autor, com base em dados do INEP (2022).

[12] Sumário de resultados obtidos com todos os algoritmos na tarefa de agrupamento, disponível em: https://docs.google.com/spreadsheets/d/17YKBQGvxn_sdva_zr18MiRgsDRnyjrCs

Em relação ao desempenho, semelhante ao cenário com o K-means, o CLARA identificou cada grupo com perfis diferentes, sendo estes identificados como baixo, médio e alto desempenho.

Em síntese, os Clusters (Grupos) formados com o CLARA indicam os seguintes perfis de estudantes:

Grupo 0: contém estudantes de baixo desempenho. Sobre os dados socioeconômicos, os pais e mães não completaram o ensino médio e a renda familiar é de até 1 salário e meio. Sobre as atividades de gestão e práticas de estudo o resultado foi o mesmo obtido com o K-means, onde os estudantes realizaram tais atividades poucas vezes. Sobre a avaliação da própria experiência se auto avaliam como pouco preparados e destacam terem tido dificuldades de infraestrutura na pandemia.

Grupo 2: predominam estudantes com desempenho médio, cujos pais não completaram o ensino médio, porém as mães completaram o até o médio. Sobre as práticas de estudos e gestão do tempo para as atividades, o resultado foi o mesmo do cenário com o K-means, com exceção apenas das práticas de treinamento para redação que destacam terem realizado por nenhuma vez.

Grupo 1: estudantes com desempenho alto, em geral, declaram a raça/cor branca, têm renda familiar acima de 3 salários mínimos, cujos pais e mães completaram até o ensino médio. Se auto avaliam como bem preparados e destacam não terem tido dificuldades de infraestrutura na pandemia.

APÊNDICE E – AMOSTRAS DE REGRAS DE ASSOCIAÇÃO POR CLUSTERS

A Figura 5.5 contém uma pequena amostra da lista de itens frequentes, após aplicação de 25% de suporte em todas as transações (instâncias de estudantes) do dataset ENEM_HE, já incluídos os clusters obtidos com o K-means (K=3) para avaliação geral e por grupos. Para diversificar, na coluna *support*, são indicados itens de intervalos de suporte variados do mínimo ao máximo possível.

Figura 5.5 - Amostra de 20 itens frequentes obtidos com o Apriori.

	support	itemsets	length
864	0.556577	'APR_MATR_VINCULO=nãoInterrompeu', 'APR_MATR_SIT_MEDIO=ensinoRegular', 'TP_SEXO=femino'	3
4077	0.481428	'APR_MATR_APREND_PANDEMIA=aprendeuMaisPresencial', 'ACESSO_INTERNET=tem', 'PROB_ROT_EST_PANDEMIA=sim', 'APR_MATR_VINCULO=nãoInterrompeu', 'APR_MATR_SIT_MEDIO=ensinoRegular'	5
448	0.436394	'APR_MATR_VINCULO=nãoInterrompeu', 'PRAT_EST_REV_VIDEOAULA=poucasVezes'	2
3583	0.355605	'ACESSO_INTERNET=tem', 'NU_CELULAR=tem', 'DIF_INFRA_PANDEMIA=não', 'APR_MATR_VINCULO=nãoInterrompeu', 'TP_SEXO=femino'	5
3415	0.345079	'PROB_ROT_EST_PANDEMIA=sim', 'APR_MATR_SIT_MEDIO=ensinoRegular', 'GEST_TEMP_ATV_CRONOGR=poucasVezes', 'APR_MATR_VINCULO=nãoInterrompeu'	4
1079	0.342240	'NU_COMPUTADOR=nãoTem', 'PROB_ROT_EST_PANDEMIA=sim', 'NU_CELULAR=tem'	3
614	0.329721	'PROB_ROT_EST_PANDEMIA=sim', 'PRAT_EST_ATV_FIXACAO=poucasVezes'	2
1802	0.311710	'APR_MATR_APREND_PANDEMIA=aprendeuMaisPresencial', 'APR_MATR_SIT_MEDIO=ensinoRegular', 'PRAT_EST_ANOT_DUV_PROF=poucasVezes'	3
2433	0.308964	'TP_COR_RACA=branca', 'PROB_ROT_EST_PANDEMIA=sim', 'ACESSO_INTERNET=tem', 'APR_MATR_VINCULO=nãoInterrompeu'	4
393	0.289360	'APR_MATR_SIT_MEDIO=ensinoRegular', 'PRAT_EST_REV_ANOT=muitasVezes'	2
2758	0.287064	'PROB_ROT_EST_PANDEMIA=sim', 'ACESSO_INTERNET=tem', 'NU_CELULAR=tem', 'PRAT_EST_RESUM_VIDEO=poucasVezes'	4
2722	0.282722	'DIF_INFRA_PANDEMIA=não', 'ACESSO_INTERNET=tem', 'NU_CELULAR=tem', 'GEST_TEMP_ATV_CRONOGR=poucasVezes'	4
2711	0.281977	'APR_MATR_APREND_PANDEMIA=aprendeuMaisPresencial', 'ACESSO_INTERNET=tem', 'NU_CELULAR=tem', 'CLUSTER=2'	4
222	0.276307	'DIF_INFRA_PANDEMIA=sim', 'RENDA_FAMILIAR=até1650'	2
1727	0.266099	'PRAT_EST_REV_ANOT=poucasVezes', 'ACESSO_INTERNET=tem', 'CLUSTER=2'	3
2287	0.262782	'NU_CELULAR=tem', 'DIF_INFRA_PANDEMIA=sim', 'TP_SEXO=femino', 'APR_MATR_VINCULO=nãoInterrompeu'	4
2442	0.256279	'PROB_ROT_EST_PANDEMIA=sim', 'NU_CELULAR=tem', 'TP_COR_RACA=pretaPardaIndígena', 'RENDA_FAMILIAR=até1650'	4
4022	0.255540	'APR_MATR_APREND_PANDEMIA=aprendeuMaisPresencial', 'NU_CELULAR=tem', 'GEST_TEMP_ATV_CRONOGR=poucasVezes', 'GEST_TEMP_ATV_MATERIAL=poucasVezes', 'APR_MATR_VINCULO=nãoInterrompeu'	5
4107	0.253320	'PRAT_EST_PARTICIPAR_FORUM=nenhumaVez', 'APR_MATR_APREND_PANDEMIA=aprendeuMaisPresencial', 'ACESSO_INTERNET=tem', 'PROB_ROT_EST_PANDEMIA=sim', 'APR_MATR_SIT_MEDIO=ensinoRegular'	5
3390	0.250312	'PRAT_EST_RESUM_TEXTO=poucasVezes', 'APR_MATR_APREND_PANDEMIA=aprendeuMaisPresencial', 'APR_MATR_SIT_MEDIO=ensinoRegular', 'APR_MATR_VINCULO=nãoInterrompeu'	4

Fonte: Elaborado pelo autor.

Considerando o dataset completo, a Figura 5.6 apresenta uma amostra de 30 regras entre 43.946 que foram obtidas, estando organizadas em ordem decrescente por valor de lift.

Figura 5.6 - Amostra de 30 regras obtidas com o Apriori.

antecedents	consequents	support	confidence	lift
'GEST_TEMP_ATV_HORA_PROG=poucasVezes'	'GEST_TEMP_ATV_CRONOGR=poucasVezes', 'GEST_TEMP_ATV_TEMPO=poucasVezes', 'ACESSO_INTERNET=tem'	0.268867	0.541379	1.504704
'NU_CELULAR=tem', 'APR_MATR_SIT_MEDIO=ensinoRegular', 'APR_MATR_VINCULO=nãoInterrompeu', 'GEST_TEMP_ATV_TEMPO=poucasVezes'	'GEST_TEMP_ATV_CRONOGR=poucasVezes', 'APR_MATR_APREND_PANDEMIA=aprendeuMaisPresencial'	0.253861	0.624990	1.455978
'GEST_TEMP_ATV_CRONOGR=poucasVezes', 'NU_CELULAR=tem'	'GEST_TEMP_ATV_TEMPO=poucasVezes', 'APR_MATR_SIT_MEDIO=ensinoRegular', 'ACESSO_INTERNET=tem'	0.299982	0.577314	1.438446
'PRAT_EST_LER=poucasVezes'	'GEST_TEMP_ATV_CRONOGR=poucasVezes', 'PROB_ROT_EST_PANDEMIA=sim'	0.250342	0.534226	1.245833
'DIF_INFRA_PANDEMIA=não', 'APR_MATR_APREND_PANDEMIA=aprendeuMaisPresencial'	'NU_COMPUTADOR=tem'	0.309850	0.676659	1.215783
'PROB_ROT_EST_PANDEMIA=sim'	'TP_SEXO=femino', 'DIF_INFRA_PANDEMIA=sim', 'APR_MATR_VINCULO=nãoInterrompeu'	0.250318	0.318896	1.177247
'PROB_ROT_EST_PANDEMIA=sim', 'APR_MATR_VINCULO=nãoInterrompeu', 'NU_CELULAR=tem', 'APR_MATR_APREND_PANDEMIA=aprendeuMaisPresencial', 'ACESSO_INTERNET=tem'	'APR_MATR_SIT_MEDIO=ensinoRegular', 'PRAT_EST_DISTRACOES=poucasVezes'	0.253987	0.460943	1.087880
'NU_CELULAR=tem', 'APR_MATR_SIT_MEDIO=ensinoRegular', 'ACESSO_INTERNET=tem'	'TP_FAIXA_ETARIA=faixaMédio', 'APR_MATR_VINCULO=nãoInterrompeu'	0.330410	0.427706	1.066150
'APR_MATR_APREND_PANDEMIA=aprendeuMaisPresencial'	'NU_CELULAR=tem', 'APR_MATR_VINCULO=nãoInterrompeu', 'PROB_ROT_EST_PANDEMIA=sim', 'GEST_TEMP_ATV_HORA_PROG=poucasVezes'	0.307960	0.393970	1.063647
'NU_CELULAR=tem', 'PRAT_EST_PARTICIPAR_FORUM=nenhumaVez'	'ACESSO_INTERNET=tem', 'APR_MATR_APREND_PANDEMIA=aprendeuMaisPresencial'	0.348769	0.768281	1.058552
'NU_CELULAR=tem', 'APR_MATR_VINCULO=nãoInterrompeu', 'PROB_ROT_EST_PANDEMIA=sim', 'PRAT_EST_ATV_AVALIACAO=poucasVezes'	'APR_MATR_APREND_PANDEMIA=aprendeuMaisPresencial'	0.257283	0.823162	1.053062
'GEST_TEMP_ATV_HORA_PROG=poucasVezes'	'PROB_ROT_EST_PANDEMIA=sim', 'APR_MATR_APREND_PANDEMIA=aprendeuMaisPresencial', 'APR_MATR_VINCULO=nãoInterrompeu'	0.314580	0.633425	1.050431
'PROB_ROT_EST_PANDEMIA=sim', 'APR_MATR_VINCULO=nãoInterrompeu', 'DIF_INFRA_PANDEMIA=não', 'NU_CELULAR=tem', 'TP_SEXO=femino'	'ACESSO_INTERNET=tem'	0.257989	0.969752	1.049986
'APR_MATR_SIT_MEDIO=ensinoRegular', 'APR_MATR_APREND_PANDEMIA=aprendeuMaisPresencial', 'APR_MATR_TP_ESTUDO=híbrido'	'PROB_ROT_EST_PANDEMIA=sim', 'ACESSO_INTERNET=tem'	0.290676	0.757779	1.046578
'TP_FAIXA_ETARIA=faixaProfissional', 'PROB_ROT_EST_PANDEMIA=sim', 'APR_MATR_APREND_PANDEMIA=aprendeuMaisPresencial'	'TP_SEXO=femino'	0.252626	0.702861	1.037812
'NU_CELULAR=tem', 'PRAT_EST_ANOT_DUV_PROF=poucasVezes', 'APR_MATR_VINCULO=nãoInterrompeu'	'ACESSO_INTERNET=tem', 'APR_MATR_APREND_PANDEMIA=aprendeuMaisPresencial'	0.318943	0.751442	1.035352
'PROB_ROT_EST_PANDEMIA=sim', 'ACESSO_INTERNET=tem'	'PRAT_EST_ATV_AVALIACAO=poucasVezes', 'NU_CELULAR=tem'	0.307098	0.424137	1.030433
'APR_MATR_APREND_PANDEMIA=aprendeuMaisPresencial'	'TP_SEXO=femino', 'NU_CELULAR=tem', 'ACESSO_INTERNET=tem'	0.489776	0.626565	1.025931
'NU_CELULAR=tem', 'ACESSO_INTERNET=tem', 'PROB_ROT_EST_PANDEMIA=sim'	'NIVEL_ESC_MAE=completouAtéMédio'	0.281827	0.396416	1.022781
'NU_CELULAR=tem', 'APR_MATR_VINCULO=nãoInterrompeu', 'ACESSO_INTERNET=tem', 'GEST_TEMP_ATV_HORA_PROG=poucasVezes'	'DIF_INFRA_PANDEMIA=não'	0.253806	0.594393	1.022733
'PROB_ROT_EST_PANDEMIA=sim'	'NU_CELULAR=tem', 'ACESSO_INTERNET=tem', 'PRAT_EST_LER=poucasVezes'	0.341128	0.434585	1.022222
'TP_SEXO=femino', 'ACESSO_INTERNET=tem'	'NU_CELULAR=tem', 'APR_MATR_VINCULO=nãoInterrompeu', 'PRAT_EST_DISTRACOES=poucasVezes'	0.293468	0.471750	1.021651
'PRAT_EST_ATV_AVALIACAO=poucasVezes', 'ACESSO_INTERNET=tem'	'APR_MATR_SIT_MEDIO=ensinoRegular', 'APR_MATR_APREND_PANDEMIA=aprendeuMaisPresencial'	0.263449	0.679817	1.015428
'GEST_TEMP_ATV_HORA_PROG=poucasVezes', 'APR_MATR_SIT_MEDIO=ensinoRegular', 'APR_MATR_APREND_PANDEMIA=aprendeuMaisPresencial'	'NU_CELULAR=tem', 'APR_MATR_VINCULO=nãoInterrompeu', 'ACESSO_INTERNET=tem'	0.299408	0.876211	1.014999
'ACESSO_INTERNET=tem'	'NU_CELULAR=tem', 'APR_MATR_VINCULO=nãoInterrompeu', 'TP_FAIXA_ETARIA=faixaProfissional', 'APR_MATR_TP_ESTUDO=híbrido'	0.250423	0.271142	1.013821
'APR_MATR_VINCULO=nãoInterrompeu'	'APR_MATR_SIT_MEDIO=ensinoRegular', 'APR_MATR_APREND_PANDEMIA=aprendeuMaisPresencial'	0.643334	0.677818	1.012442
'APR_MATR_SIT_MEDIO=ensinoRegular'	'NU_CELULAR=tem', 'GEST_TEMP_ATV_CRONOGR=poucasVezes', 'APR_MATR_VINCULO=nãoInterrompeu', 'GEST_TEMP_ATV_TEMPO=poucasVezes'	0.309940	0.364418	1.007115
'PRAT_EST_TREINAR_REDACAO=poucasVezes'	'APR_MATR_VINCULO=nãoInterrompeu', 'ACESSO_INTERNET=tem', 'APR_MATR_APREND_PANDEMIA=aprendeuMaisPresencial'	0.291037	0.697506	1.003433

Fonte: Elaborado pelo autor.

A Figura 5.7 apresenta uma amostra de 30 regras obtidas considerando transações (instâncias de estudantes) do Cluster 0, estando organizadas em ordem decrescente por valor de lift>1. No total foram geradas 89.491 regras de associação.

Figura 5.7 - Amostra de 30 regras obtidas com o FPGrowth - CLUSTER 0.

	antecedents	consequents	support	confidence	lift
74626	'PRAT_EST_RESUM_VIDEO=muitasVezes'	'PRAT_EST_ANOT_DUV_VIDEO_COMPLEMENTAR=muitasVezes', 'APR_MATR_VINCULO=nãoInterrompeu', 'PRAT_EST_ATV_FIXACAO=muitasVezes'	0.265647	0.440860	1.218923
94955	'PRAT_EST_ANOT_DUV_VIDEO_COMPLEMENTAR=muitasVezes', 'NU_CELULAR=tem', 'PRAT_EST_REV_ANOT=muitasVezes', 'ACESSO_INTERNET=tem'	'PRAT_EST_ESTRUT_IDEIA_REDACAO=muitasVezes'	0.256018	0.669337	1.197441
85811	'APR_MATR_VINCULO=nãoInterrompeu', 'GEST_TEMP_ATV_TEMPO=muitasVezes'	'GEST_TEMP_ATV_MATERIAL=muitasVezes', 'NU_CELULAR=tem', 'ACESSO_INTERNET=tem', 'APR_MATR_APREND_PANDEMIA=aprendeuMaisPresencial'	0.260127	0.486851	1.193300
14493	'NU_COMPUTADOR=tem', 'ACESSO_INTERNET=tem'	'DIF_INFRA_PANDEMIA=não', 'APR_MATR_VINCULO=nãoInterrompeu', 'APR_MATR_APREND_PANDEMIA=aprendeuMaisPresencial'	0.338777	0.557952	1.183782
98866	'PRAT_EST_RESUM_VIDEO=muitasVezes', 'PRAT_EST_RESUM_TEXTO=muitasVezes'	'GEST_TEMP_ATV_MATERIAL=muitasVezes', 'NU_CELULAR=tem', 'ACESSO_INTERNET=tem'	0.254339	0.627515	1.159241
51954	'NU_CELULAR=tem', 'PRAT_EST_ANOT_DUV_VIDEO=muitasVezes'	'TP_SEXO=femino', 'APR_MATR_VINCULO=nãoInterrompeu', 'PRAT_EST_REV_ANOT=muitasVezes'	0.280166	0.465638	1.135421
84394	'PRAT_EST_ATV_AVALIACAO=muitasVezes', 'APR_MATR_SIT_MEDIO=ensinoRegular', 'ACESSO_INTERNET=tem'	'PRAT_EST_REV_ANOT=muitasVezes'	0.260849	0.666279	1.113493
41450	'ACESSO_INTERNET=tem', 'PRAT_EST_ANOT_DUV_VIDEO=muitasVezes'	'APR_MATR_VINCULO=nãoInterrompeu', 'PRAT_EST_DISTRACOES=muitasVezes'	0.290276	0.508575	1.113203
106938	'PRAT_EST_ANOT_DUV_VIDEO_COMPLEMENTAR=muitasVezes', 'APR_MATR_SIT_MEDIO=ensinoRegular'	'PRAT_EST_RESUM_VIDEO=muitasVezes', 'ACESSO_INTERNET=tem', 'APR_MATR_APREND_PANDEMIA=aprendeuMaisPresencial'	0.251214	0.467548	1.105905
105658	'PRAT_EST_RESUM_VIDEO=muitasVezes'	'TP_SEXO=femino', 'GEST_TEMP_ATV_MATERIAL=muitasVezes', 'APR_MATR_VINCULO=nãoInterrompeu', 'ACESSO_INTERNET=tem'	0.251536	0.417442	1.099171
30477	'NU_COMPUTADOR=tem', 'PROB_ROT_EST_PANDEMIA=sim'	'DIF_INFRA_PANDEMIA=não', 'ACESSO_INTERNET=tem'	0.303423	0.682915	1.097721
96490	'PRAT_EST_ANOT_DUV_VIDEO_COMPLEMENTAR=muitasVezes'	'GEST_TEMP_ATV_HORA_PROG=muitasVezes', 'APR_MATR_VINCULO=nãoInterrompeu', 'APR_MATR_APREND_PANDEMIA=aprendeuMaisPresencial'	0.255356	0.407868	1.071319
101471	'PRAT_EST_ANOT_DUV_VIDEO_COMPLEMENTAR=muitasVezes', 'APR_MATR_SIT_MEDIO=ensinoRegular', 'ACESSO_INTERNET=tem'	'GEST_TEMP_ATV_CRONOGR=muitasVezes', 'APR_MATR_VINCULO=nãoInterrompeu'	0.253389	0.507525	1.059734
72491	'APR_MATR_VINCULO=nãoInterrompeu', 'NU_CELULAR=tem', 'APR_MATR_APREND_PANDEMIA=aprendeuMaisPresencial', 'TP_SEXO=femino', 'APR_MATR_TP_ESTUDO=híbrido'	'APR_MATR_SIT_MEDIO=ensinoRegular', 'ACESSO_INTERNET=tem'	0.266904	0.851062	1.058317
98315	'PROB_ROT_EST_PANDEMIA=sim', 'PRAT_EST_REV_ANOT=muitasVezes'	'GEST_TEMP_ATV_MATERIAL=muitasVezes', 'ACESSO_INTERNET=tem'	0.254567	0.578618	1.051455
57541	'TP_FAIXA_ETARIA=faixaMédio', 'APR_MATR_SIT_MEDIO=ensinoRegular', 'APR_MATR_APREND_PANDEMIA=aprendeuMaisPresencial'	'APR_MATR_VINCULO=nãoInterrompeu', 'ACESSO_INTERNET=tem'	0.275689	0.935859	1.038409
105830	'TP_SEXO=femino', 'NU_CELULAR=tem', 'ACESSO_INTERNET=tem'	'PRAT_EST_REV_ANOT=muitasVezes', 'APR_MATR_VINCULO=nãoInterrompeu', 'APR_MATR_SIT_MEDIO=ensinoRegular', 'APR_MATR_APREND_PANDEMIA=aprendeuMaisPresencial'	0.251475	0.388528	1.032587
76307	'PROB_ROT_EST_PANDEMIA=sim', 'PRAT_EST_LER=muitasVezes'	'TP_SEXO=femino'	0.264937	0.727913	1.026847
45305	'NU_CELULAR=tem', 'PRAT_EST_REV_ANOT=muitasVezes', 'ACESSO_INTERNET=tem', 'APR_MATR_TP_ESTUDO=híbrido'	'APR_MATR_SIT_MEDIO=ensinoRegular'	0.286301	0.883726	1.026453
52743	'APR_MATR_VINCULO=nãoInterrompeu', 'APR_MATR_SIT_MEDIO=ensinoRegular'	'NU_COMPUTADOR=tem', 'GEST_TEMP_ATV_TEMPO=muitasVezes', 'ACESSO_INTERNET=tem'	0.279517	0.334088	1.023261
67911	'ACESSO_INTERNET=tem'	'NU_CELULAR=tem', 'AUTOAV_PREPARACAO_APRENDIZ=bemPreparado', 'APR_MATR_VINCULO=nãoInterrompeu', 'APR_MATR_APREND_PANDEMIA=aprendeuMaisPresencial'	0.269413	0.289026	1.015642
59881	'DIF_INFRA_PANDEMIA=não', 'APR_MATR_SIT_MEDIO=ensinoRegular', 'PRAT_EST_TREINAR_REDACAO=muitasVezes'	'APR_MATR_VINCULO=nãoInterrompeu'	0.273963	0.979382	1.014193
71154	'GEST_TEMP_PONTUAL_AULA_ONLINE=muitasVezes', 'APR_MATR_VINCULO=nãoInterrompeu'	'DIF_INFRA_PANDEMIA=não', 'NU_CELULAR=tem', 'APR_MATR_SIT_MEDIO=ensinoRegular'	0.267674	0.564283	1.013845
4092	'APR_MATR_SIT_MEDIO=ensinoRegular', 'APR_MATR_APREND_PANDEMIA=aprendeuMaisPresencial'	'PRAT_EST_ANOT_DUV_VIDEO_COMPLEMENTAR=muitasVezes'	0.404635	0.628936	1.004569
11217	'NU_CELULAR=tem', 'APR_MATR_VINCULO=nãoInterrompeu'	'APR_MATR_APREND_PANDEMIA=aprendeuMaisPresencial', 'PRAT_EST_LER=muitasVezes'	0.350473	0.370346	1.004176
8816	'NU_CELULAR=tem', 'ACESSO_INTERNET=tem'	'PRAT_EST_ANOT_DUV_VIDEO_COMPLEMENTAR=muitasVezes', 'APR_MATR_VINCULO=nãoInterrompeu', 'APR_MATR_SIT_MEDIO=ensinoRegular', 'APR_MATR_APREND_PANDEMIA=aprendeuMaisPresencial'	0.362557	0.395497	1.004113
7331	'NU_COMPUTADOR=tem', 'APR_MATR_SIT_MEDIO=ensinoRegular', 'ACESSO_INTERNET=tem'	'PROB_ROT_EST_PANDEMIA=sim', 'APR_MATR_VINCULO=nãoInterrompeu'	0.373804	0.704790	1.003253
38644	'APR_MATR_VINCULO=nãoInterrompeu'	'NU_CELULAR=tem', 'GEST_TEMP_ATV_TEMPO=muitasVezes', 'ACESSO_INTERNET=tem', 'PRAT_EST_ESTRUT_IDEIA_REDACAO=muitasVezes'	0.293440	0.303870	1.003200
106090	'NU_CELULAR=tem', 'APR_MATR_APREND_PANDEMIA=aprendeuMaisPresencial', 'TP_SEXO=femino', 'PRAT_EST_RESUM_TEXTO=muitasVezes', 'APR_MATR_SIT_MEDIO=ensinoRegular'	'ACESSO_INTERNET=tem'	0.251395	0.933792	1.001771
1065	'TP_SEXO=femino', 'PROB_ROT_EST_PANDEMIA=sim', 'NU_CELULAR=tem'	'ACESSO_INTERNET=tem'	0.490860	0.933241	1.001179

Fonte: Elaborado pelo autor.

A Figura 5.8 representa uma amostra de 30 regras obtidas com estudantes pertencentes aos Cluster 1 (desempenho médio). No total, para esse grupo, foram geradas 3.603.659 regras de associação.

Figura 5.8 - Amostra de 30 regras obtidas com o FPGrowth - CLUSTER 1.

	antecedents	consequents	support	confidence	lift
2413822	'NU_CELULAR=tem', 'PRAT_EST_RESUM_VIDEO=nenhumaVez', 'PRAT_EST_PARTICIPAR_FORUM=nenhumaVez'	'PRAT_EST_ANOT_DUV_PROF=nenhumaVez', 'GEST_TEMP_ATV_TEMPO=nenhumaVez', 'PRAT_EST_RESUM_TEXTO=nenhumaVez'	0.271100	0.585556	1.587197
1857500	'PRAT_EST_PARTICIPAR_FORUM=nenhumaVez', 'PRAT_EST_ANOT_DUV_VIDEO=nenhumaVez', 'APR_MATR_VINCULO=nãoInterrompeu', 'PRAT_EST_ANOT_DUV_PROF=nenhumaVez'	'PRAT_EST_ANOT_DUV_VIDEO_COMPLEMENTAR=nenhumaVez', 'PRAT_EST_REV_VIDEOAULA=nenhumaVez', 'PRAT_EST_RESUM_TEXTO=nenhumaVez'	0.278341	0.620478	1.586302
1474431	'PRAT_EST_ANOT_DUV_VIDEO=nenhumaVez', 'PRAT_EST_ATV_FIXACAO=nenhumaVez', 'ACESSO_INTERNET=tem'	'PRAT_EST_REV_ANOT=nenhumaVez', 'PRAT_EST_ANOT_DUV_VIDEO_COMPLEMENTAR=nenhumaVez', 'PRAT_EST_RESUM_TEXTO=nenhumaVez'	0.285025	0.549011	1.561785
4605443	'GEST_TEMP_ATV_HORA_PROG=nenhumaVez', 'PRAT_EST_RESUM_VIDEO=nenhumaVez', 'PRAT_EST_REV_VIDEOAULA=nenhumaVez', 'PRAT_EST_RESUM_TEXTO=nenhumaVez'	'NU_CELULAR=tem', 'PRAT_EST_LER=nenhumaVez', 'PRAT_EST_ANOT_DUV_PROF=nenhumaVez'	0.254558	0.816180	1.559235
2941174	'PRAT_EST_LER=nenhumaVez', 'PRAT_EST_ATV_FIXACAO=nenhumaVez', 'PRAT_EST_REV_VIDEOAULA=nenhumaVez'	'NU_CELULAR=tem', 'PRAT_EST_ANOT_DUV_VIDEO=nenhumaVez', 'PRAT_EST_ANOT_DUV_PROF=nenhumaVez', 'ACESSO_INTERNET=tem', 'PRAT_EST_RESUM_TEXTO=nenhumaVez'	0.265920	0.553996	1.528888
2518067	'NU_CELULAR=tem', 'PRAT_EST_ANOT_DUV_VIDEO_COMPLEMENTAR=nenhumaVez', 'PRAT_EST_PARTICIPAR_FORUM=nenhumaVez'	'PRAT_EST_ANOT_DUV_VIDEO=nenhumaVez', 'PRAT_EST_ANOT_DUV_PROF=nenhumaVez', 'ACESSO_INTERNET=tem', 'PRAT_EST_REV_VIDEOAULA=nenhumaVez', 'APR_MATR_VINCULO=nãoInterrompeu', 'PRAT_EST_REV_ANOT=nenhumaVez'	0.270023	0.485804	1.516540
3531251	'PRAT_EST_ATV_FIXACAO=nenhumaVez', 'PRAT_EST_ANOT_DUV_PROF=nenhumaVez'	'PRAT_EST_ANOT_DUV_VIDEO=nenhumaVez', 'PRAT_EST_REV_VIDEOAULA=nenhumaVez', 'APR_MATR_VINCULO=nãoInterrompeu', 'PRAT_EST_PARTICIPAR_FORUM=nenhumaVez', 'PRAT_EST_REV_ANOT=nenhumaVez', 'PRAT_EST_RESUM_TEXTO=nenhumaVez'	0.261242	0.428615	1.441887
4709367	'PRAT_EST_ANOT_DUV_VIDEO=nenhumaVez', 'GEST_TEMP_ASSID_AULA_ONLINE=nenhumaVez'	'NU_CELULAR=tem', 'PRAT_EST_ATV_FIXACAO=nenhumaVez', 'PRAT_EST_ANOT_DUV_PROF=nenhumaVez', 'PRAT_EST_REV_VIDEOAULA=nenhumaVez'	0.254020	0.723915	1.429727
3927136	'PRAT_EST_RESUM_VIDEO=nenhumaVez', 'PRAT_EST_ATV_FIXACAO=nenhumaVez', 'PRAT_EST_ANOT_DUV_PROF=nenhumaVez'	'NU_CELULAR=tem', 'PRAT_EST_ANOT_DUV_VIDEO_COMPLEMENTAR=nenhumaVez', 'APR_MATR_SIT_MEDIO=ensinoRegular'	0.258550	0.675626	1.332597
504575	'PRAT_EST_ANOT_DUV_VIDEO_COMPLEMENTAR=nenhumaVez'	'NU_CELULAR=tem', 'PRAT_EST_ANOT_DUV_VIDEO=nenhumaVez', 'ACESSO_INTERNET=tem', 'APR_MATR_APREND_PANDEMIA=aprendeuMaisPresencial', 'PRAT_EST_PARTICIPAR_FORUM=nenhumaVez', 'PRAT_EST_ATV_FIXACAO=nenhumaVez'	0.318388	0.515977	1.328277
1122645	'PRAT_EST_ANOT_DUV_VIDEO_COMPLEMENTAR=nenhumaVez', 'PRAT_EST_ATV_FIXACAO=nenhumaVez'	'PRAT_EST_PARTICIPAR_FORUM=nenhumaVez', 'PRAT_EST_ESTRUT_IDEIA_REDACAO=nenhumaVez', 'ACESSO_INTERNET=tem', 'PRAT_EST_REV_VIDEOAULA=nenhumaVez'	0.292804	0.549168	1.283823
5263578	'ACESSO_INTERNET=tem', 'APR_MATR_VINCULO=nãoInterrompeu', 'PRAT_EST_PARTICIPAR_FORUM=nenhumaVez', 'APR_MATR_SIT_MEDIO=ensinoRegular', 'PRAT_EST_RESUM_TEXTO=nenhumaVez'	'NU_CELULAR=tem', 'PRAT_EST_ANOT_DUV_VIDEO_COMPLEMENTAR=nenhumaVez', 'PRAT_EST_ATV_FIXACAO=nenhumaVez', 'PRAT_EST_REV_VIDEOAULA=nenhumaVez'	0.251309	0.573608	1.268035
2075770	'NU_CELULAR=tem', 'ACESSO_INTERNET=tem', 'PRAT_EST_DISTRACOES=nenhumaVez', 'PRAT_EST_LER=nenhumaVez', 'PRAT_EST_ATV_FIXACAO=nenhumaVez'	'PRAT_EST_PARTICIPAR_FORUM=nenhumaVez', 'PRAT_EST_REV_ANOT=nenhumaVez'	0.275203	0.733195	1.257286
3087668	'NU_CELULAR=tem', 'APR_MATR_VINCULO=nãoInterrompeu', 'PRAT_EST_LER=nenhumaVez', 'PRAT_EST_RESUM_TEXTO=nenhumaVez', 'GEST_TEMP_ATV_HORA_PROG=nenhumaVez', 'PRAT_EST_ATV_FIXACAO=nenhumaVez'	'PRAT_EST_ANOT_DUV_VIDEO=nenhumaVez', 'ACESSO_INTERNET=tem'	0.264658	0.738180	1.244635
2154360	'PRAT_EST_ANOT_DUV_VIDEO_COMPLEMENTAR=nenhumaVez', 'PROB_ROT_EST_PANDEMIA=sim'	'NU_CELULAR=tem', 'ACESSO_INTERNET=tem', 'PRAT_EST_REV_VIDEOAULA=nenhumaVez', 'PRAT_EST_RESUM_TEXTO=nenhumaVez', 'PRAT_EST_ATV_FIXACAO=nenhumaVez'	0.274219	0.537345	1.219915
69966	'PRAT_EST_PARTICIPAR_FORUM=nenhumaVez', 'PRAT_EST_ANOT_DUV_PROF=nenhumaVez', 'PRAT_EST_DISTRACOES=nenhumaVez'	'PRAT_EST_ANOT_DUV_VIDEO=nenhumaVez'	0.393190	0.775325	1.211680
3265642	'PRAT_EST_ANOT_DUV_PROF=nenhumaVez', 'PRAT_EST_REV_VIDEOAULA=nenhumaVez', 'APR_MATR_VINCULO=nãoInterrompeu', 'PRAT_EST_LER=nenhumaVez', 'PRAT_EST_ATV_FIXACAO=nenhumaVez'	'PRAT_EST_ATV_AVALIACAO=nenhumaVez'	0.263210	0.720230	1.202474
1699666	'APR_MATR_APREND_PANDEMIA=aprendeuMaisPresencial', 'PRAT_EST_LER=nenhumaVez', 'PROB_ROT_EST_PANDEMIA=sim', 'PRAT_EST_PARTICIPAR_FORUM=nenhumaVez'	'GEST_TEMP_ATV_HORA_PROG=nenhumaVez', 'PRAT_EST_ANOT_DUV_PROF=nenhumaVez'	0.280922	0.638008	1.193913
5176852	'NU_CELULAR=tem', 'PROB_ROT_EST_PANDEMIA=sim', 'PRAT_EST_PARTICIPAR_FORUM=nenhumaVez', 'APR_MATR_APREND_PANDEMIA=aprendeuMaisPresencial', 'PRAT_EST_RESUM_TEXTO=nenhumaVez'	'GEST_TEMP_ATV_MATERIAL=nenhumaVez'	0.251717	0.617283	1.161181
305929	'PRAT_EST_ATV_FIXACAO=nenhumaVez', 'GEST_TEMP_ATV_TEMPO=nenhumaVez'	'NU_CELULAR=tem', 'PRAT_EST_ATV_AVALIACAO=nenhumaVez'	0.335747	0.663633	1.134533
1842024	'NU_CELULAR=tem', 'PRAT_EST_PARTICIPAR_FORUM=nenhumaVez', 'ACESSO_INTERNET=tem', 'AJUD_TERC_PANDEMIA=ninguémAuxiliou'	'PROB_ROT_EST_PANDEMIA=sim'	0.278564	0.947043	1.132339
1671773	'PRAT_EST_PARTICIPAR_FORUM=nenhumaVez', 'PRAT_EST_ANOT_DUV_VIDEO=nenhumaVez', 'PRAT_EST_DISTRACOES=nenhumaVez'	'PRAT_EST_ATV_FIXACAO=nenhumaVez', 'APR_MATR_SIT_MEDIO=ensinoRegular', 'PROB_ROT_EST_PANDEMIA=sim'	0.281367	0.622868	1.123766
3263238	'PRAT_EST_ANOT_DUV_VIDEO=nenhumaVez', 'APR_MATR_VINCULO=nãoInterrompeu', 'APR_MATR_SIT_MEDIO=ensinoRegular'	'NU_CELULAR=tem', 'GEST_TEMP_ATV_HORA_PROG=nenhumaVez', 'ACESSO_INTERNET=tem', 'PRAT_EST_RESUM_TEXTO=nenhumaVez'	0.263210	0.524297	1.122806
4255050	'GEST_TEMP_PONTUAL_AULA_ONLINE=nenhumaVez', 'APR_MATR_SIT_MEDIO=ensinoRegular', 'PRAT_EST_ATV_FIXACAO=nenhumaVez'	'PROB_ROT_EST_PANDEMIA=sim', 'PRAT_EST_REV_VIDEOAULA=nenhumaVez'	0.256507	0.740130	1.116350
4349833	'GEST_TEMP_PONTUAL_AULA_ONLINE=nenhumaVez', 'PRAT_EST_ANOT_DUV_PROF=nenhumaVez', 'ACESSO_INTERNET=tem'	'NU_CELULAR=tem', 'GEST_TEMP_ATV_HORA_PROG=nenhumaVez', 'PRAT_EST_PARTICIPAR_FORUM=nenhumaVez'	0.255988	0.686346	1.114735
1535653	'PRAT_EST_ATV_FIXACAO=nenhumaVez', 'ACESSO_INTERNET=tem'	'NU_CELULAR=tem', 'GEST_TEMP_ATV_TEMPO=nenhumaVez', 'PRAT_EST_REV_VIDEOAULA=nenhumaVez', 'APR_MATR_SIT_MEDIO=ensinoRegular', 'GEST_TEMP_ATV_HORA_PROG=nenhumaVez'	0.283855	0.381929	1.110111
1044257	'PRAT_EST_ANOT_DUV_VIDEO=nenhumaVez', 'ACESSO_INTERNET=tem', 'PRAT_EST_RESUM_TEXTO=nenhumaVez'	'NU_CELULAR=tem', 'PRAT_EST_ANOT_DUV_PROF=nenhumaVez', 'APR_MATR_APREND_PANDEMIA=aprendeuMaisPresencial'	0.294939	0.683122	1.109901
3027255	'TP_SEXO=femino', 'PRAT_EST_REV_VIDEOAULA=nenhumaVez'	'APR_MATR_APREND_PANDEMIA=aprendeuMaisPresencial', 'APR_MATR_SIT_MEDIO=ensinoRegular', 'APR_MATR_VINCULO=nãoInterrompeu', 'PRAT_EST_ANOT_DUV_PROF=nenhumaVez'	0.265178	0.536370	1.071706
5041375	'APR_MATR_VINCULO=nãoInterrompeu', 'PRAT_EST_ANOT_DUV_PROF=nenhumaVez', 'PRAT_EST_DISTRACOES=nenhumaVez'	'TP_SEXO=femino', 'APR_MATR_SIT_MEDIO=ensinoRegular', 'PRAT_EST_PARTICIPAR_FORUM=nenhumaVez'	0.252367	0.507448	1.065373
242187	'APR_MATR_APREND_PANDEMIA=aprendeuMaisPresencial', 'NU_COMPUTADOR=tem', 'PRAT_EST_PARTICIPAR_FORUM=nenhumaVez'	'APR_MATR_VINCULO=nãoInterrompeu', 'PRAT_EST_REV_VIDEOAULA=nenhumaVez'	0.343972	0.757874	1.032754

Fonte: Elaborado pelo autor.

A Figura 5.9 ilustra uma amostra de 30 regras obtidas com estudantes do Cluster 2 (desempenho baixo). No total, para esse grupo, foram geradas 1.098.373 regras de associação.

Figura 5.9 - Amostra de 30 regras obtidas com o FPGrowth - CLUSTER 2.

	antecedents	consequents	support	confidence	lift
9654686	'PRAT_EST_DISTRACOES=poucasVezes', 'NU_CELULAR=tem', 'APR_MATR_VINCULO=nãoInterrompeu', 'PRAT_EST_RESUM_TEXTO=poucasVezes', 'PRAT_EST_ANOT_DUV_VIDEO=poucasVezes'	'PRAT_EST_ANOT_DUV_VIDEO_COMPLEMENTAR=poucasVezes', 'PRAT_EST_LER=poucasVezes', 'PRAT_EST_RESUM_VIDEO=poucasVezes'	0.252964	0.672609	1.615641
6495260	'GEST_TEMP_ATV_CRONOGR=poucasVezes', 'PRAT_EST_ATV_FIXACAO=poucasVezes', 'PRAT_EST_RESUM_TEXTO=poucasVezes'	'GEST_TEMP_ATV_HORA_PROG=poucasVezes', 'ACESSO_INTERNET=tem', 'PRAT_EST_LER=poucasVezes', 'PRAT_EST_RESUM_VIDEO=poucasVezes'	0.263074	0.606327	1.430447
6993345	'GEST_TEMP_ATV_HORA_PROG=poucasVezes', 'PRAT_EST_ATV_FIXACAO=poucasVezes', 'PRAT_EST_REV_ANOT=poucasVezes'	'PRAT_EST_REV_VIDEOAULA=poucasVezes', 'PRAT_EST_RESUM_TEXTO=poucasVezes', 'APR_MATR_VINCULO=nãoInterrompeu', 'NU_CELULAR=tem'	0.261143	0.593352	1.386173
6929810	'GEST_TEMP_ATV_MATERIAL=poucasVezes', 'GEST_TEMP_ATV_CRONOGR=poucasVezes', 'PRAT_EST_DISTRACOES=poucasVezes', 'NU_CELULAR=tem', 'APR_MATR_VINCULO=nãoInterrompeu', 'PRAT_EST_RESUM_TEXTO=poucasVezes'	'GEST_TEMP_ATV_TEMPO=poucasVezes', 'PRAT_EST_ATV_FIXACAO=poucasVezes'	0.261360	0.789524	1.348487
1713859	'PRAT_EST_RESUM_TEXTO=poucasVezes'	'PRAT_EST_ATV_FIXACAO=poucasVezes', 'NU_CELULAR=tem', 'PRAT_EST_RESUM_VIDEO=poucasVezes', 'GEST_TEMP_ATV_HORA_PROG=poucasVezes', 'PRAT_EST_ANOT_DUV_PROF=poucasVezes'	0.300402	0.443105	1.287626
8485689	'GEST_TEMP_ATV_TEMPO=poucasVezes', 'PRAT_EST_RESUM_TEXTO=poucasVezes', 'APR_MATR_APREND_PANDEMIA=aprendeuMaisPresencial'	'GEST_TEMP_ATV_HORA_PROG=poucasVezes', 'GEST_TEMP_ATV_MATERIAL=poucasVezes', 'ACESSO_INTERNET=tem', 'PRAT_EST_RESUM_VIDEO=poucasVezes'	0.256246	0.557259	1.280346
8668922	'PRAT_EST_RESUM_TEXTO=poucasVezes'	'PRAT_EST_DISTRACOES=poucasVezes', 'ACESSO_INTERNET=tem', 'APR_MATR_VINCULO=nãoInterrompeu', 'PRAT_EST_RESUM_VIDEO=poucasVezes', 'PRAT_EST_LER=poucasVezes', 'PRAT_EST_ANOT_DUV_VIDEO=poucasVezes'	0.255706	0.377176	1.279250
7952691	'PRAT_EST_REV_VIDEOAULA=poucasVezes', 'PRAT_EST_ANOT_DUV_VIDEO=poucasVezes', 'PRAT_EST_ANOT_DUV_PROF=poucasVezes', 'APR_MATR_VINCULO=nãoInterrompeu'	'GEST_TEMP_ATV_TEMPO=poucasVezes', 'PRAT_EST_ANOT_DUV_VIDEO_COMPLEMENTAR=poucasVezes', 'ACESSO_INTERNET=tem'	0.257876	0.686939	1.279161
5706722	'GEST_TEMP_ATV_MATERIAL=poucasVezes', 'PRAT_EST_ATV_FIXACAO=poucasVezes', 'PROB_ROT_EST_PANDEMIA=sim', 'PRAT_EST_RESUM_TEXTO=poucasVezes'	'GEST_TEMP_ATV_TEMPO=poucasVezes', 'PRAT_EST_LER=poucasVezes', 'NU_CELULAR=tem'	0.266442	0.760365	1.257238
7262297	'GEST_TEMP_ATV_CRONOGR=poucasVezes', 'PRAT_EST_ESTRUT_IDEIA_REDACAO=poucasVezes', 'GEST_TEMP_ATV_TEMPO=poucasVezes', 'ACESSO_INTERNET=tem', 'PRAT_EST_RESUM_TEXTO=poucasVezes'	'GEST_TEMP_ATV_HORA_PROG=poucasVezes', 'GEST_TEMP_ATV_MATERIAL=poucasVezes'	0.260155	0.798743	1.246918
7191941	'PRAT_EST_REV_VIDEOAULA=poucasVezes', 'GEST_TEMP_ATV_TEMPO=poucasVezes', 'ACESSO_INTERNET=tem', 'PRAT_EST_ANOT_DUV_VIDEO_COMPLEMENTAR=poucasVezes', 'PRAT_EST_LER=poucasVezes'	'PRAT_EST_ANOT_DUV_VIDEO=poucasVezes', 'NU_CELULAR=tem'	0.260394	0.856055	1.246552
3029788	'GEST_TEMP_ATV_HORA_PROG=poucasVezes', 'PRAT_EST_RESUM_VIDEO=poucasVezes'	'PRAT_EST_ATV_FIXACAO=poucasVezes', 'APR_MATR_VINCULO=nãoInterrompeu', 'PRAT_EST_LER=poucasVezes', 'PRAT_EST_REV_ANOT=poucasVezes'	0.283789	0.491046	1.243539
8914166	'GEST_TEMP_ATV_MATERIAL=poucasVezes', 'APR_MATR_VINCULO=nãoInterrompeu', 'PRAT_EST_ESTRUT_IDEIA_REDACAO=poucasVezes', 'PRAT_EST_REV_VIDEOAULA=poucasVezes'	'GEST_TEMP_ATV_TEMPO=poucasVezes', 'PRAT_EST_ANOT_DUV_VIDEO=poucasVezes'	0.255011	0.724947	1.240368
5391575	'GEST_TEMP_ATV_MATERIAL=poucasVezes', 'PRAT_EST_ATV_FIXACAO=poucasVezes', 'NU_CELULAR=tem', 'GEST_TEMP_ATV_TEMPO=poucasVezes', 'ACESSO_INTERNET=tem', 'PRAT_EST_RESUM_VIDEO=poucasVezes', 'PRAT_EST_RESUM_TEXTO=poucasVezes'	'PRAT_EST_LER=poucasVezes'	0.267909	0.903660	1.235476
8247640	'GEST_TEMP_ATV_TEMPO=poucasVezes', 'ACESSO_INTERNET=tem', 'PRAT_EST_RESUM_TEXTO=poucasVezes'	'GEST_TEMP_ATV_HORA_PROG=poucasVezes', 'GEST_TEMP_ATV_CRONOGR=poucasVezes', 'APR_MATR_SIT_MEDIO=ensinoRegular', 'PRAT_EST_ANOT_DUV_VIDEO_COMPLEMENTAR=poucasVezes'	0.256965	0.491963	1.229675
6414797	'PRAT_EST_ANOT_DUV_PROF=poucasVezes', 'NU_CELULAR=tem'	'PRAT_EST_ANOT_DUV_VIDEO_COMPLEMENTAR=poucasVezes', 'GEST_TEMP_ATV_HORA_PROG=poucasVezes', 'PRAT_EST_RESUM_TEXTO=poucasVezes', 'PRAT_EST_ANOT_DUV_VIDEO=poucasVezes', 'PRAT_EST_REV_ANOT=poucasVezes'	0.263391	0.376327	1.225334
3822969	'PRAT_EST_ANOT_DUV_VIDEO_COMPLEMENTAR=poucasVezes', 'GEST_TEMP_ASSID_AULA_ONLINE=poucasVezes'	'PRAT_EST_ANOT_DUV_VIDEO=poucasVezes', 'PROB_ROT_EST_PANDEMIA=sim'	0.277286	0.708330	1.223725
4909724	'GEST_TEMP_ATV_MATERIAL=poucasVezes', 'PRAT_EST_LER=poucasVezes', 'PRAT_EST_ANOT_DUV_PROF=poucasVezes', 'PRAT_EST_RESUM_VIDEO=poucasVezes'	'GEST_TEMP_ATV_TEMPO=poucasVezes', 'ACESSO_INTERNET=tem', 'GEST_TEMP_ATV_HORA_PROG=poucasVezes'	0.270450	0.763375	1.207000
6648427	'GEST_TEMP_ATV_MATERIAL=poucasVezes', 'GEST_TEMP_ATV_CRONOGR=poucasVezes', 'PRAT_EST_ATV_FIXACAO=poucasVezes', 'PRAT_EST_TREINAR_REDACAO=poucasVezes'	'GEST_TEMP_ATV_TEMPO=poucasVezes', 'GEST_TEMP_ATV_HORA_PROG=poucasVezes'	0.262449	0.834848	1.202433
5867947	'PRAT_EST_DISTRACOES=poucasVezes', 'PRAT_EST_RESUM_VIDEO=poucasVezes'	'GEST_TEMP_ATV_TEMPO=poucasVezes', 'APR_MATR_VINCULO=nãoInterrompeu', 'PRAT_EST_ATV_AVALIACAO=poucasVezes', 'PROB_ROT_EST_PANDEMIA=sim'	0.265739	0.502255	1.180919
6507511	'GEST_TEMP_ATV_HORA_PROG=poucasVezes', 'PRAT_EST_LER=poucasVezes', 'PRAT_EST_DISTRACOES=poucasVezes', 'APR_MATR_SIT_MEDIO=ensinoRegular'	'PRAT_EST_ANOT_DUV_PROF=poucasVezes', 'PRAT_EST_REV_ANOT=poucasVezes'	0.263028	0.661249	1.180589
7492037	'PRAT_EST_ANOT_DUV_PROF=poucasVezes', 'PRAT_EST_ANOT_DUV_VIDEO=poucasVezes', 'PROB_ROT_EST_PANDEMIA=sim'	'GEST_TEMP_ATV_TEMPO=poucasVezes', 'GEST_TEMP_ATV_MATERIAL=poucasVezes', 'PRAT_EST_DISTRACOES=poucasVezes'	0.259351	0.593667	1.177882
2888871	'GEST_TEMP_ATV_CRONOGR=poucasVezes', 'PRAT_EST_ESTRUT_IDEIA_REDACAO=poucasVezes', 'NU_CELULAR=tem', 'APR_MATR_VINCULO=nãoInterrompeu', 'PRAT_EST_ANOT_DUV_PROF=poucasVezes'	'GEST_TEMP_ATV_HORA_PROG=poucasVezes', 'GEST_TEMP_ATV_MATERIAL=poucasVezes'	0.285117	0.743714	1.161013
7169752	'GEST_TEMP_ATV_TEMPO=poucasVezes', 'PRAT_EST_ANOT_DUV_VIDEO_COMPLEMENTAR=poucasVezes', 'PRAT_EST_ESTRUT_IDEIA_REDACAO=poucasVezes', 'PROB_ROT_EST_PANDEMIA=sim'	'PRAT_EST_RESUM_TEXTO=poucasVezes', 'NU_CELULAR=tem'	0.260471	0.759652	1.149014
8813880	'GEST_TEMP_ATV_TEMPO=poucasVezes', 'PRAT_EST_RESUM_TEXTO=poucasVezes', 'PRAT_EST_DISTRACOES=poucasVezes', 'APR_MATR_SIT_MEDIO=ensinoRegular'	'ACESSO_INTERNET=tem', 'PRAT_EST_ATV_AVALIACAO=poucasVezes'	0.255289	0.686701	1.141414
2588151	'PRAT_EST_ESTRUT_IDEIA_REDACAO=poucasVezes', 'NU_CELULAR=tem'	'GEST_TEMP_ATV_HORA_PROG=poucasVezes', 'GEST_TEMP_ATV_CRONOGR=poucasVezes', 'PRAT_EST_LER=poucasVezes', 'PRAT_EST_RESUM_TEXTO=poucasVezes'	0.288284	0.444345	1.131603
4013696	'GEST_TEMP_ATV_CRONOGR=poucasVezes', 'PRAT_EST_ESTRUT_IDEIA_REDACAO=poucasVezes', 'NU_CELULAR=tem', 'PRAT_EST_RESUM_VIDEO=poucasVezes'	'ACESSO_INTERNET=tem', 'APR_MATR_VINCULO=nãoInterrompeu', 'PRAT_EST_LER=poucasVezes'	0.275965	0.707413	1.130235
8253873	'GEST_TEMP_ATV_TEMPO=poucasVezes', 'PRAT_EST_ATV_FIXACAO=poucasVezes', 'PRAT_EST_ESTRUT_IDEIA_REDACAO=poucasVezes', 'NU_CELULAR=tem'	'GEST_TEMP_ATV_HORA_PROG=poucasVezes', 'ACESSO_INTERNET=tem', 'PROB_ROT_EST_PANDEMIA=sim'	0.256949	0.635844	1.056081
1680573	'GEST_TEMP_ATV_TEMPO=poucasVezes'	'PRAT_EST_TREINAR_REDACAO=poucasVezes', 'APR_MATR_SIT_MEDIO=ensinoRegular', 'GEST_TEMP_ATV_HORA_PROG=poucasVezes', 'APR_MATR_APREND_PANDEMIA=aprendeuMaisPresencial'	0.301028	0.361361	1.049944
5052698	'NU_CELULAR=tem', 'ACESSO_INTERNET=tem', 'APR_MATR_VINCULO=nãoInterrompeu', 'PRAT_EST_ANOT_DUV_VIDEO=poucasVezes', 'PROB_ROT_EST_PANDEMIA=sim', 'APR_MATR_SIT_MEDIO=ensinoRegular', 'PRAT_EST_REV_ANOT=poucasVezes'	'APR_MATR_APREND_PANDEMIA=aprendeuMaisPresencial'	0.269662	0.830633	1.028357

Fonte: Elaborado pelo autor.

ANEXO

ANEXO A – DICIONÁRIO DE DADOS DO DATASET ENEM_HE

O Quadro 4.3 contém a descrição das variáveis e os respectivos valores categorizados, com base em fontes oficiais.

Quadro 4.3 - Categorização das variáveis para aplicação dos algoritmos.

Variável	Descrição	Valores categorizados
TP_FAIXA_ETARIA *	Faixa etária do estudante (a partir da idade do inscrito em 31/12/2022)	1 a 2 = 'faixaMédio' 3 a 10 = 'faixaProfissional' 11 a 20 = 'faixaAdultosIdosos'
TP_SEXO	Sexo do estudante	'M' = 'masculino' 'F' = femino'
TP_COR_RACA **	Cor/raça do estudante	0 = 'nãoDeclarado' 1 = 'branca' 2, 3, 5 = 'pretaPardaIndígena' 4 = 'amarela' 6 = 'semInformação'
NIVEL_ESC_PAI	Até que série seu pai, ou o homem responsável por você, estudou?	A, B, C, D = 'nãoCompletouMédio' E = 'completouAtéMédio' F, G = 'completouSuperior' H = 'nãoSabe'
NIVEL_ESC_MAE	Até que série sua mãe, ou a mulher responsável por você, estudou?	
RENDA_FAMILIAR ***	Qual é a renda mensal de sua família? (Some a sua renda com a dos seus familiares.)	A, B, C = 'até1818' D, E, F = 'até3636' G, H, I, J, K, L, M, N, O, P, Q = 'acimaDe3636'
NU_CELULAR	Na sua residência tem telefone celular?	A = 'nãoTem' B, C, D, E = 'tem'
NU_COMPUTADOR	Na sua residência tem computador?	
ACESSO_INTERNET	Na sua residência tem acesso à Internet?	A = 'nãoTem' B = 'tem'

APR_MATR_SIT_MEDIO	Considerando a etapa de Ensino Médio, qual dessas situações está de acordo com o seu vínculo escolar durante a pandemia?	A = 'ensinoRegular' B = 'ensinoEJA' C = 'ensinoProfissional' D = 'concluiuAntes2021' E = 'nãoConcluiuMedio'
APR_MATR_VINCULO	Considerando a continuidade do vínculo escolar na pandemia, qual dessas situações está de acordo com sua realidade?	A = 'nãoInterrompeu' B = 'interrompeu' C = 'interrompeu'
APR_MATR_TP_ESTUDO	Considerando o ano de 2021 (o segundo ano da pandemia), qual dessas situações está de acordo com sua experiência?	A = 'apenasPresencial' B = 'apenasRemoto' C = 'híbrido' D = 'semMatricula' E = 'semMatricula'
APR_MATR_APREND_PANDEMIA	Como você percebe o seu processo de aprendizagem durante a pandemia?	A = 'aprendeuMaisRemoto' B = 'aprendeuMaisHíbrido' C = 'aprendeuMaisPresencial' D = 'aprendeuRemotoEHíbrido' E = 'aprendeuContaPrópria' F = 'semMatriculaNãoEstudou'
GEST_TEMP_ATV_CRONOGR	Queremos saber quantas vezes você realizou determinadas atividades de estudo para preparar-se para o Enem, mesmo que você não tenha se matriculado regularmente em 2021. Considerando a frequência de atividades de estudo no segundo ano da pandemia: Organizei cronograma de estudos com tempos mais longos e mais curtos para estudar de acordo com a dificuldade das matérias.	A = 'nenhumaVez' B = 'poucasVezes' C = 'muitasVezes' D = 'todasAsVezes'
GEST_TEMP_ATV_TEMPO	Queremos saber quantas vezes você realizou determinadas atividades de estudo para preparar-se para o Enem, mesmo que você não tenha se matriculado regularmente em 2021. Considerando a frequência de atividades de estudo no segundo ano da pandemia: Reservei tempos mais longos e mais curtos para estudar de acordo com a dificuldade das matérias.	
GEST_TEMP_ATV_MATERIAL	Queremos saber quantas vezes você realizou determinadas atividades de estudo para preparar-se para o Enem, mesmo que você não tenha se matriculado regularmente em 2021. Considerando a frequência de atividades de estudo no segundo ano da pandemia: Organizei material para ser estudado.	
GEST_TEMP_ATV_HORA_PROG	Queremos saber quantas vezes você realizou determinadas atividades de estudo para preparar-se para o Enem, mesmo que você não tenha se matriculado regularmente em 2021. Considerando a frequência de atividades de estudo no segundo ano da pandemia: Eu me dediquei aos horários programados de estudo	

	de acordo com a dificuldade das matérias.	
PRAT_EST_LER	Queremos saber quantas vezes você realizou determinadas atividades de estudo para preparar-se para o Enem, mesmo que você não tenha se matriculado regularmente em 2021. Considerando a frequência de atividades de estudo no segundo ano da pandemia: Li os textos indicados em cada matéria antes de assistir as aulas ou videoaulas sobre o assunto dos textos.	A = 'nenhumaVez' B = 'poucasVezes' C = 'muitasVezes' D = 'todasAsVezes'
PRAT_EST_RESUM_TEXTO	Queremos saber quantas vezes você realizou determinadas atividades de estudo para preparar-se para o Enem, mesmo que você não tenha se matriculado regularmente em 2021. Considerando a frequência de atividades de estudo no segundo ano da pandemia: Resumi os textos das matérias, destacando as partes mais importantes.	
PRAT_EST_RESUM_VIDEO	Queremos saber quantas vezes você realizou determinadas atividades de estudo para preparar-se para o Enem, mesmo que você não tenha se matriculado regularmente em 2021. Considerando a frequência de atividades de estudo no segundo ano da pandemia: Resumi as videoaulas ou os podcasts, destacando as partes mais importantes.	
PRAT_EST_ATV_FIXACAO	Queremos saber quantas vezes você realizou determinadas atividades de estudo para preparar-se para o Enem, mesmo que você não tenha se matriculado regularmente em 2021. Considerando a frequência de atividades de estudo no segundo ano da pandemia: Fiz as atividades das matérias para fixação de conteúdo.	
PRAT_EST_ATV_AVALIACAO	Queremos saber quantas vezes você realizou determinadas atividades de estudo para preparar-se para o Enem, mesmo que você não tenha se matriculado regularmente em 2021. Considerando a frequência de atividades de estudo no segundo ano da pandemia: Fiz atividades avaliativas, inclusive simulados, para verificar o quanto aprendi durante a pandemia.	
PRAT_EST_DISTRACOES	Queremos saber quantas vezes você realizou determinadas atividades de estudo para preparar-se para o Enem, mesmo que você não tenha se matriculado regularmente em 2021. Considerando a frequência de atividades de estudo no segundo ano da pandemia: Aproveitei o tempo das aulas online ou atividades de reforço, sem desperdiçá-lo com distrações.	
PRAT_EST_ANOT_DUV_VIDEO	Queremos saber quantas vezes você realizou determinadas atividades de estudo para preparar-se para o Enem, mesmo que você não tenha se matriculado regularmente em 2021. Considerando a frequência de atividades de estudo no segundo ano da pandemia: Anotei as explicações obtidas em videoaulas ou podcasts das matérias.	
PRAT_EST_ANOT_DUV_VIDEO_COMPLEMENTAR	Queremos saber quantas vezes você realizou determinadas atividades de estudo para preparar-se para o Enem, mesmo que você não tenha se matriculado regularmente em 2021. Considerando a frequência de atividades de estudo no segundo ano da pandemia: Anotei as informações que obtive ao assistir vídeos complementares de assuntos do meu interesse.	
PRAT_EST_ANOT_DUV_PROF	Queremos saber quantas vezes você realizou determinadas atividades de estudo para preparar-se para o Enem, mesmo que você não tenha se matriculado regularmente em 2021. Considerando a	

	frequência de atividades de estudo no segundo ano da pandemia: Destaquei as dúvidas que tive ao ler os textos das disciplinas para esclarecer com os professores.	
PRAT_EST_ESTRUT_IDEIA_REDACAO	Queremos saber quantas vezes você realizou determinadas atividades de estudo para preparar-se para o Enem, mesmo que você não tenha se matriculado regularmente em 2021. Considerando a frequência de atividades de estudo no segundo ano da pandemia: Estruturei as principais ideias para produzir redações.	
PRAT_EST_TREINAR_REDACAO	Queremos saber quantas vezes você realizou determinadas atividades de estudo para preparar-se para o Enem, mesmo que você não tenha se matriculado regularmente em 2021. Considerando a frequência de atividades de estudo no segundo ano da pandemia: Treinei redação.	
PRAT_EST_PARTICIPAR_FORUM	Queremos saber quantas vezes você realizou determinadas atividades de estudo para preparar-se para o Enem, mesmo que você não tenha se matriculado regularmente em 2021. Considerando a frequência de atividades de estudo no segundo ano da pandemia: Participei de fóruns de discussão por matéria para tirar dúvidas.	
GEST_TEMP_PONTUAL_AULA_ONLINE	Queremos saber quantas vezes você realizou determinadas atividades de estudo para preparar-se para o Enem, mesmo que você não tenha se matriculado regularmente em 2021. Considerando a frequência de atividades de estudo no segundo ano da pandemia: Entrei nas aulas online por videoconferência sem atraso da minha parte.	A = 'nenhumaVez' B = 'poucasVezes' C = 'muitasVezes' D = 'todasAsVezes'
GEST_TEMP_ASSID_AULA_ONLINE	Queremos saber quantas vezes você realizou determinadas atividades de estudo para preparar-se para o Enem, mesmo que você não tenha se matriculado regularmente em 2021. Considerando a frequência de atividades de estudo no segundo ano da pandemia: Assisti todas as aulas online nas datas programadas para estudo.	
PRAT_EST_REV_ANOT	Queremos saber quantas vezes você realizou determinadas atividades de estudo para preparar-se para o Enem, mesmo que você não tenha se matriculado regularmente em 2021. Considerando a frequência de atividades de estudo no segundo ano da pandemia: Revisei as anotações das aulas, os resumos e anotações dos demais materiais que li ou assisti.	A = 'nenhumaVez' B = 'poucasVezes' C = 'muitasVezes' D = 'todasAsVezes'
PRAT_EST_REV_VIDEOAULA	Queremos saber quantas vezes você realizou determinadas atividades de estudo para preparar-se para o Enem, mesmo que você não tenha se matriculado regularmente em 2021. Considerando a frequência de atividades de estudo no segundo ano da pandemia: Reassisti as videoaulas e os podcasts das matérias.	
PROB_ROT_EST_PANDEMIA	Você vivenciou problemas em sua rotina para estudar ou manter-se informado(a) durante a pandemia?	A = 'sim' B = 'não'
DIF_INFRA_PANDEMIA	Durante a pandemia, há relatos de pessoas que tiveram dificuldade para criar condições para estudar ou manterem-se informadas em casa, por exemplo, dificuldades com a falta de internet,	A = 'sim' B = 'não'

	equipamentos, espaço ou materiais. Você teve dificuldades de infraestrutura para estudar ou manter-se informado(a) durante 2021?	
AJUD_TERC_PANDEMIA	Você precisou da ajuda de alguém para estudar ou manter-se informado(a) em 2021?	A = 'sim' B = 'não' C = 'ninguémAuxiliou'
AUTOAV_PREPARACAO_APRENDIZ	A partir da sua experiência de estudos em 2021, o quanto você se sente preparado(a) para conduzir o seu processo de aprendizagem?	A = 'nadaPreparado' B = 'poucoPreparado' C = 'bemPreparado' D = 'muitoPreparado' E = 'totalmentePreparado'
NOTA_MEDIA	Nota média considerando as provas de: Ciências da Natureza, Ciências Humanas, Linguagens e Códigos, Matemática e Redação.	305.66 a 507.43 = 'baixo' 507.44 a 581.37 = 'médio' 581.38 a 855.82 = 'alto'

* Com base na fonte disponível em: unicef.org/brazil/media/461/file/Panorama_da_distorcao_idade-serie_no_Brasil.pdf
** Com base na fonte: portal.mec.gov.br/cotas/perguntas-frequentes.html
*** Com base na fonte: gov.br/inep/pt-br/assuntos/noticias/enem/aberto-periodo-de-isencao-para-o-enem -2023
**** Com base na fonte: https://download.inep.gov.br/educacao_basica/enem/nota_tecnica/2015/nota_explicativa_enem2015_por_escola.pdf

Fonte: Elaborado pelo autor, com base em dados do INEP (2022).