Mineração de Dados nos Hábitos de Estudo: Uma Análise para o Desempenho no ENEM 2022

Abílio Nogueira Barros[1], **Danielle Karla Alves da Silva**[2], **Paulo J. L. Adeodato**[3]

[1]Departamento de Computação – Universidade Federal Rural de Pernambuco (UFRPE)

[2]Departamento de Micologia-Universidade Federal de Pernambuco(UFPE)

[3]Centro de Informática-Universidade Federal de Pernambuco(UFPE)

abilionbarros@gmail.com, daniellekarlas@yahoo.com.br,pjla@cin.ufpe.br

Abstract. *This article addresses the knowledge discovery process through data mining applied to the study habits of participants in the 2022 National High School Examination (ENEM). The primary objective of this research is to employ data mining techniques to identify and highlight which study practices are most effective in achieving a positive performance in the exam. The analysis aims to provide valuable insights that can contribute to optimizing candidates' preparation methods, offering informed guidance for improved performance in future ENEM exams. Preliminary results indicate that the frequent organization of study material and the consistent practice of summarizing video classes and/or podcasts are important factors for better student performance in ENEM. Therefore, these findings can guide educational practices and improve preparation strategies for the ENEM, improving the understanding of the factors that can influence student performance.*

Resumo. *Este artigo aborda o processo de descoberta de conhecimento por meio da mineração de dados aplicada aos hábitos de estudo dos participantes do Exame Nacional do Ensino Médio (ENEM) de 2022. O objetivo principal desta pesquisa é empregar técnicas de mineração de dados para identificar e destacar quais práticas de estudo são mais eficazes na obtenção de um desempenho positivo no exame. A análise visa fornecer insights valiosos que podem contribuir para a otimização dos métodos de preparação dos candidatos, oferecendo orientações fundamentadas para um melhor desempenho em exames do ENEM no futuro. Os resultados iniciais indicam que a organização frequente do material de estudo e a prática consistente de resumir videoaulas e/ou podcasts são importantes fatores para o melhor desempenho dos estudantes no ENEM. Sendo assim, esses achados podem orientar práticas educacionais e aprimorar estratégias de preparação para o ENEM, melhorando o entendimento dos fatores que podem influenciar o desempenho dos estudantes.*

1. Introdução

O exame nacional do ensino médio (ENEM) é uma avaliação realizada anualmente no Brasil pelo Instituto Nacional de Estudos e Pesquisas Educacionais (INEP). Realizado

DOI: 10.5753/sbie.2024.244620

desde 1998, o ENEM fornece dados valiosos para avaliação do sistema educacional brasileiro. A prova abrange áreas como linguagens, códigos e suas tecnologias, ciências humanas e suas tecnologias, matemática e suas tecnologias, ciências da natureza e suas tecnologias, além da redação.

Desde a sua implementação em 1998, o ENEM tem sido uma fonte valiosa de dados, tornando-se uma importante via de ingresso ao ensino superior. Em 2004, tornou-se critério para o Programa Universidade para Todos (ProUni), e em 2009, a nota do ENEM passou a ser adotada como forma de ingresso nas universidades federais por meio do Sistema de Seleção Unificada (SISU) criado pelo Ministério da Educação (MEC) [1]. Além disso, a nota do ENEM é utilizada por instituições privadas para acesso ao Fundo de Financiamento Estudantil (FIES) e para ingresso em universidades portuguesas.

Os dados dos participantes do ENEM incluem não apenas as notas nas provas, mas também informações sobre o perfil socioeconômico, o que permite uma análise abrangente das características dos participantes em relação às questões socioeconômicas. Em 2022, o INEP introduziu um questionário específico sobre os hábitos de estudo durante a pandemia para os participantes do ENEM. A pesquisa, realizada por meio de um questionário com perguntas de resposta única e múltipla, contou com a participação de quase 1 milhão de estudantes (aproximadamente 29% do total de inscritos). O questionário abordou diversos temas, incluindo a situação de matrícula, a percepção do processo de aprendizagem, estratégias de gestão do tempo e planejamento dos estudos, práticas de estudo e pesquisa, tipos de acesso e uso de tecnologia, problemas enfrentados na rotina de estudo, dificuldades de infraestrutura, ajuda de terceiros, e uma autoavaliação da experiência de estudo durante o segundo ano da pandemia.

Identificar os fatores que afetam o desempenho dos estudantes é um desafio complexo, dada a diversidade dos perfis estudantis. No entanto, os dados provenientes do questionário de hábitos de estudo, aliados aos dados socioeconômicos dos participantes, são de grande valor para interpretar os fatores que influenciam o desempenho dos alunos. Nesse sentido, a Descoberta de Conhecimento em Banco de Dados (DCBD, ou *Knowledge Discovery in Databases – KDD*) é uma abordagem que envolve as etapas de seleção, preparação e limpeza dos dados, além da mineração de dados (MD, ou *Data Mining – DM*), considerada uma fase do processo de KDD, na qual se aplicam algoritmos específicos para extrair padrões dos dados [Fayyad et al. 1996].

Diversas abordagens podem ser utilizadas no processo de mineração de dados, e neste estudo, adotamos o Processo Padrão para Mineração de Dados (*Cross-Industry Standard Process for Data Mining – CRISP-DM*), que se divide em seis etapas: compreensão do negócio, compreensão dos dados, preparação dos dados, modelagem, avaliação e implantação [Shearer 2000]. Essas ferramentas são úteis para extrair informações de grandes volumes de dados e são especialmente valiosas para compreender dados educacionais. Com o aumento da geração de dados na área educacional, a mineração de dados tornou-se conhecida como ”mineração de dados educacionais” (MDE, ou *Educational Data Mining – EDM*) [Romero et al. 2010].

Pesquisas têm sido conduzidas para compreender o impacto do perfil socioeconômico dos participantes no desempenho no ENEM. Nesse contexto,

[1] http://portal.mec.gov.br

[Guardieiro et al. 2022] investigaram se a prova do ENEM favorece ou desfavorece grupos específicos com base em gênero, raça e nível de renda. Os autores constataram que as questões de ciências humanas e matemática não apresentavam preferência ou desvantagem para nenhum grupo. No entanto, identificaram que alguns grupos se destacam nas questões de ciências naturais, linguagens e códigos, e língua estrangeira. Em geral, participantes brancos, do sexo masculino e com alto nível de renda apresentaram melhor desempenho nessas áreas.

Ao analisar uma década de dados do ENEM (2009-2019), [Silva Filho et al. 2023] constataram que a renda per capita, a raça, o nível educacional dos pais, a idade do estudante foram os fatores mais preponderantes sobre o desempenho dos alunos. Além desses aspectos, os autores identificaram tendências que sugerem uma certa influência de variáveis como acesso a computador, formação educacional e treinamento adequado do corpo docente, bem como a carga de trabalho destes profissionais. Essas descobertas proporcionam uma base sólida para uma compreensão mais profunda de como esses fatores impactam o desempenho dos estudantes no ENEM.

[Franco et al. 2020], ao analisarem os fatores que influenciam o desempenho dos estudantes com base em anos de dados do ENEM, evidenciaram variáveis como renda familiar, escolha da língua estrangeira, sexo, acesso à internet e computador, tipo de escola frequentada no ensino médio influenciam a atuação dos participantes. Vale destacar que este estudo é o primeiro a analisar a base de dados relacionada aos hábitos de estudos dos participantes do ENEM que começou a ser preenchida em 2022.

No cenário educacional contemporâneo, a compreensão dos fatores que impactam o desempenho dos estudantes em exames como o ENEM tornou-se essencial. Nesse contexto, a aplicação da mineração de dados surge como uma ferramenta poderosa, possibilitando a análise de grandes volumes de informações para identificar padrões, correlações e realizar predições. Este artigo busca explorar como os hábitos de estudo dos estudantes influenciam diretamente seus resultados no ENEM, utilizando técnicas de mineração de dados para revelar *insights* significativos. O objetivo é não apenas entender as relações existentes, mas também fornecer subsídios para a implementação de estratégias educacionais mais eficazes, visando aprimorar o desempenho acadêmico e promover a equidade no acesso ao ensino superior. Especificamente, este trabalho pretende identificar quais hábitos de estudo mais contribuem para o melhor desempenho dos estudantes no ENEM.

2. Levantamento e prepararação dos dados

Os dados utilizados para a realização desta pesquisa foram os microdados do Exame Nacional do Ensino Médio 2022, o ENEM. Essa base de dados é composta por dois arquivos *CSV* e seu dicionário de dados, ambos disponibilizados pelo INEP[2].

O primeiro arquivo traz resultados da prova do ENEM realizada por cada candidato, contendo algumas informações sobre localidade, tipo de ensino e respostas sobre o mapeamento do perfil socioeconômico dos participantes e sua família.

A segunda base de dados é referente aos hábitos de estudo e traz as informações principais para esse estudo, que foram os hábitos de estudo vivenciados pelos participantes do exame durante o segundo ano da pandemia, descrevendo como foi sua

[2]https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem

preparação, maiores dificuldades, maiores problemas e os meios utilizados para contornar os desafios educacionais vivenciados no ano de 2021.

2.1. Preparação dos dados

A base completa de respostas do ENEM contém quase 4 milhões de registros do respondente, entretanto, as informações necessárias para realizar a análise completa são as do hábito de estudo que possuem cerca de 1 milhão de registros, criando assim um viés que limita a aplicabilidade deste estudo apenas para os alunos que optaram preencher voluntariamente e que podem ser cruzados para o experimento.

Sendo assim, foram realizados os recortes para garantir que os registros sejam aderentes ao resultado que buscamos, como garantir que todos os registros na base correspondam aos critérios buscados por este trabalho: preenchimento total do questionário, possuir notas em todos os componentes do exame e não estar prestando o exame como experiência (popularmente conhecido como *treineiro*) Com isso, após o cruzamento das bases de resultados do ENEM e de hábitos de estudo, a base combinada foi finalizada com pouco mais de 793 mil registros que correspondiam aos critérios definidos. Por esse estudo se tratar diretamente dos hábitos de estudos, o questionário sócio-econômico não foi utilizado.Dentre as 84 colunas na base de dados sobre hábitos de estudo do ENEM, apenas 29 foram incluídas na análise. Algumas questões foram excluídas devido ao elevado número de dados faltantes e/ou por apresentarem informações inconsistentes que não permitiam interpretação confiável.

2.2. Caracterização do Problema de Decisão Binária

A nota de cada aluno no ENEM é calculada pela média das notas em cada prova, produzindo um valor contínuo na avaliação de cada candidato.

Para simplificar a abordagem e após uma análise preliminar e exploratória dos fatores, optou-se pela classificação binária devido à disponibilidade de técnicas e à necessidade de explicabilidade. Assim, o problema foi caracterizado como uma decisão binária, exigindo a aplicação de um limiar sobre as notas para categorizar os alunos em dois grupos. Devido à falta de consenso entre educadores sobre o valor ideal desse limiar, os autores adotaram o critério sugerido por [Adeodato 2015], utilizando o quartil superior para atribuir o rótulo 1 aos alunos de alto desempenho.

Dessa forma, o objetivo do problema tornou-se quantificável e inequivocamente algorítmico. Na amostra de dados, o quartil superior, correspondente aos melhores alunos, foi definido por um limiar de média igual a 498. Alunos com notas inferiores a esse valor receberam o rótulo “0”. No entanto, é importante destacar que, mesmo com notas acima desse limiar, os alunos podem não garantir o ingresso na universidade, pois a admissão depende da concorrência na opção escolhida pelo candidato, independentemente de suas boas notas.

2.3. Utilização do KNIME

Para a próxima fase será utilizada a ferramenta open-source KNIME, esse software possibilita a criação de pipelines de mineração de dados, onde possamos abstrair etapas mais complexas como acoplamento de bases entre processos e a possibilidade de uma nova execução parcial ou total do pipeline criada.

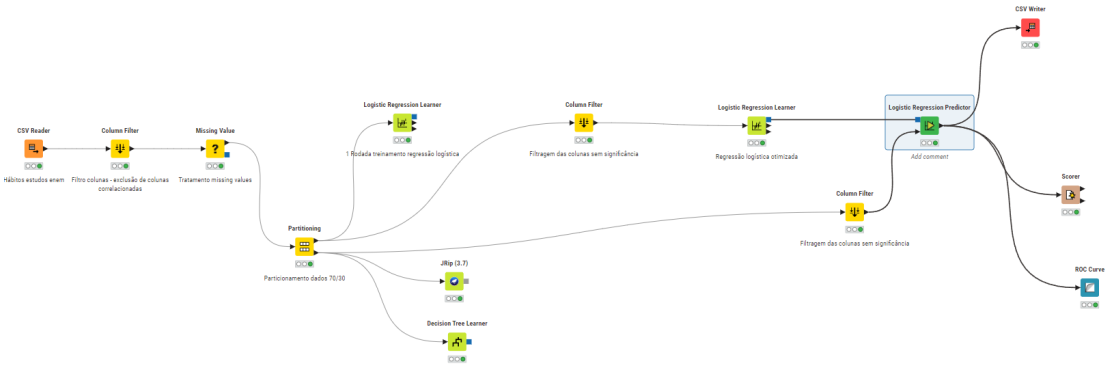

Figure 1. Pipeline do KNIME.

Com a utilização do KNIME foi possível criar os artefatos necessários para o desenvolvimento deste trabalho. Como pode ser ilustrado na figura os passos aplicados foram:

1. Carregamento dos dados: Os dados foram carregados após o pré-processamento do CSV;
2. Processamento de valores ausentes. Os valores ausentes nas colunas foram substituídos pelo mais frequente. Colunas que apresentassem 80% de valores ausentes ou mais.
3. Particionamento dos dados: Os dados foram divididos de forma que o conjunto de treino possuísse 70% das observações.
4. A Priori + filtragens: O a priori foi utilizado como indutor de regras, dentre as quais apenas as de classificação foram selecionadas e filtradas para ficarem apenas as consistentes para comporem a base de regras.
5. *DecisionTreeRegressor*: Uma árvore de decisão para a regressão que visa apontar quais fatores foram determinantes para determinada classe.
6. *Logistic Regression*:O algoritmo *backward stepwise logistic regression* foi utilizado para descarte gradual das variáveis com p-valor menor que 0,05, até que todas as variáveis remanescentes tivessem coeficientes com significância estatística.

3. Resultados

A árvore de decisão na figura 2 destacou variáveis de maior influência no desempenho no ENEM, como a organização frequente do material de estudo, o resumo consistente de videoaulas e/ou podcasts, a ocorrência de distrações durante as aulas (em diferentes intensidades) e a prática de estruturar ideias para o treino de redação. Esses resultados mostram a importância da organização do material, resumos eficazes e estruturação de ideias para o treino da redação. Notavelmente, mesmo a presença de possíveis distrações não pareceu impactar negativamente o rendimento, indicando que momento de relaxamento ou pausas na rotina de estudo são relevantes e contribuem para um melhor desempenho.

Quanto a indução de regras, dado o elevadíssimo custo computacional do algoritmo a priori [Agrawal and Srikant 1994] usado para a indução de regras, tomamos uma subamostra aleatória estratificada pela variável-alvo com 20 mil registros. Adicionalmente, em várias variáveis, agrupamos categorias com percentual muito baixo de respondentes à categoria de semântica mais similar, num processo Domain-Driven Data Mining

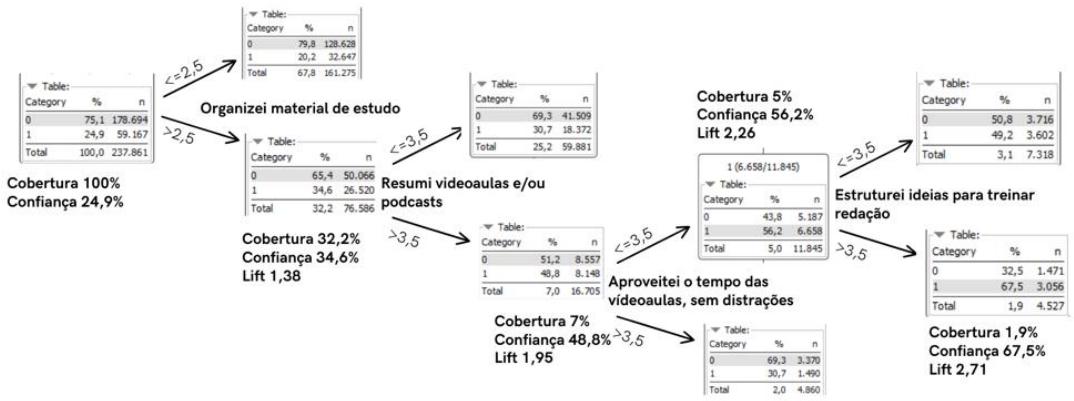

Figure 2. Arvore de Decisão

– D3M [Cao et al. 2007]. Então, o algoritmo a priori induziu mais de 1 milhão de regras de associação e de classificação misturadas. Após o descarte das regras de associação, eliminamos a maioria das regras de classificação, mantendo apenas aquelas com cobertura acima de 2,5% do volume da amostra. Em seguida, eliminamos a pequena parcela de regras com mais de 3 cláusulas no antecedente, para facilitar a interpretabilidade humana (eXplainable-AI – X-AI). Finalmente, eliminamos as regras com inconsistências e ficamos com a base de regras final.

A base final contém 49 regras de 1 cláusula, 168 de 2 cláusulas e 159 de 3 cláusulas, totalizando 376 regras. Essa base é completa e só contém regras de classificação consistentes e com cobertura acima de 2,5% da amostra. O objetivo de ter uma base de regras consolidadas é poder colocá-la num Sistema de Suporte à Decisão para justificar recomendações geradas por escores de propensão de alta qualidade, como por Deep Learning, poder elaborar e validar políticas públicas ou campanhas de marketing por nichos e poder validar o conhecimento extraído por outras técnicas explicáveis.

As 3 regras abaixo mostram "bons hábitos" de estudo (para ter sucesso na prova do ENEM) em que destacamos aquelas de maior cobertura e maior confiança, escolhendo pares de cláusulas o mais ortogonais possível. Naturalmente, na base também há as regras de "maus hábitos" de estudo, mas que preferimos não destacar quantificando-as na figura.

1. Q029=Não (sobre ter tido dificuldade de infraestrutura) Cobertura=52,6%, Confiança=32,1%, Lift=1,28
2. Q006=4 $\Leftrightarrow$ Todas as vezes alocou tempos proporcionais à dificuldade de cada matéria: Cobertura=6,7%, Confiança=49,3%, Lift=1,97
3. Q029=Não e Q006=4: Cobertura=4,3%, Confiança=60,9%, Lift=2,43

Elas mostram que o acesso à infraestrutura e a alocação de tempo de estudo proporcional à dificuldade da matéria são fatores relevantes para o sucesso do aluno. Quando combinadas, essas condições aumentam muito mais as chances de sucesso.

Após análise minuciosa das regras com expertise do domínio e uso de ferramenta de consultas analíticas (OnLine Analytical Processing – OLAP), percebemos algumas condições que fogem às expectativas do senso comum, como, por exemplo:

1. A resposta à Q034 tem níveis crescentes de autoconfiança (1 a 5, escala de Likert) quanto ao aluno estar preparado para o ENEM. Os alunos que se julgam totalmente

preparados (nível 5) têm menor chance de estar no quartil superior que os alunos que se consideram bem preparados (nível 4).
2. A Q001 fala do vínculo escolar durante a pandemia. Os alunos de escolas com Ensino do Jovem Adulto (EJA) têm metade da chance daqueles que não concluíram o ensino médio ou não estavam matriculados durante a pandemia. Ambos são péssimos.
3. A Q014 fala do grau de distração dos alunos nas aulas online ou atividades de reforço. Essa variável não apareceu como discriminante nas chances dos alunos.

3.1. Regressão Logística

A avaliação de desempenho do modelo de regressão logística, observado pela curva ROC, que demonstrou uma área sob a curva de 0,711 indicando que o modelo aprendeu com os dados utilizados e alcançou uma acurácia de 73,8%.

Na tabela abaixo, detalhamos métricas como precisão, sensibilidade, especificidade e o *F-measure* 1. Esses dados evidenciam que o modelo conseguiu aprender com as duas condições, no entanto, a especificidade para os dados de baixa desempenho foi baixa. Esses resultados podem estar relacionados ao desbalanceamento dos dados entre as classes. Precisão, sensibilidade, especificidade e *F-measure* do modelo gerado pelo regressor logístico gerado a partir dos dados dos hábitos de estudo no ENEM 2022.

Desempenho	Precisão	Sensibilidade	Especificidade	F-measure
0 - Baixa performance	0,75	0,936	0,285	0,833
1 - Alta performance	0,66	0,285	0,936	0,398

Table 1. Desempenho do modelo

Analisando possíveis vieses no modelo em relação às notas de alta desempenho, ao comparar os dados reais 3 e os dados preditos 4 dos 10% melhores notas, observamos que essas notas se concentravam em Minas gerais e São Paulo, e o modelo predito pôde prever com sucesso as maiores notas nesses estados, evidenciaram a sensibilidade do modelo.

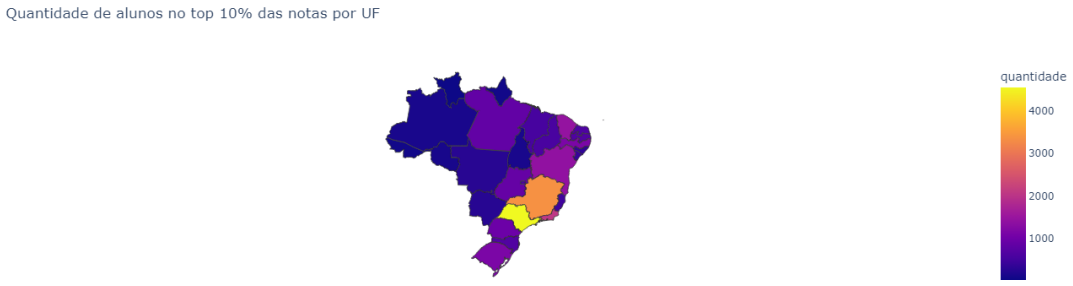

Figure 3. Valores reais.

A quantidade de alunos com as 10% melhores notas, a partir dos dados reais 3 e a partir dos dados preditos pelo regressor logístico 4 gerado a partir dos dados dos hábitos de estudo no ENEM 2022.

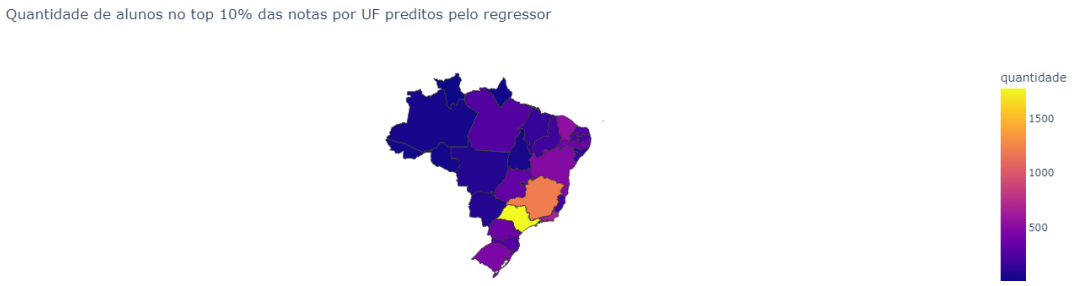

Figure 4. Valores descritos pelo regressor.

4. Conclusões

A árvore de decisão identificou que a organização frequente do material de estudo e a prática consistente de resumir videoaulas e/ou podcasts são fatores cruciais para o desempenho dos estudantes no ENEM. Esses achados destacam a importância de abordagens organizadas e métodos eficientes de revisão. Curiosamente, a presença de distrações durante as aulas não teve impacto significativo no desempenho dos alunos, sugerindo que momentos de relaxamento ou pausas durante a rotina de estudos podem ser benéficos, desde que bem equilibrados.

O conjunto de regras gerado pelo indutor a priori resultou em uma base de alta qualidade para propor estratégias que estimulem o sucesso dos alunos (quartil superior no ENEM). Assim como as regras da árvore de decisão, as regras induzidas não consideraram o grau de distração durante as atividades online e de reforço como um fator de relevância significativa. Em geral, as boas práticas aumentam as chances de sucesso dos alunos, conforme esperado.

Em síntese, os resultados apontam para a importância de estratégias organizacionais nos hábitos de estudo e destacam um perfil específico associado ao sucesso no ENEM. Essas conclusões podem orientar práticas educacionais e aprimorar estratégias de preparação para o ENEM, promovendo um entendimento mais profundo dos fatores que influenciam o desempenho dos estudantes.

Limitações quanto a redundância na base de dados, a qual apresentou grande colinearidade entre os dados, impossibilidade de conexão com outras bases educacionais e baixa adesão ao questionário, menos de 1/3 dos estudantes responderam ao questionário foram grandes entraves dessa pesquisa.

Como trabalho futuro, planejamos validar esses resultados junto a especialistas na área como curso de preparatórios de universidades públicas, apresentado-os os resultados já obtidos. Pretendemos também avaliar as regras obtidas em sub-grupos de perfis de participante aliado a outros dados do questionário sócio-econômico, buscando a possibilidade de maior aderência de terminados hábitos para a maximização do resultado estudantil.

References

Adeodato, P. J. L. (2015). Variable transformation for granularity change in hierarchical databases in actual data mining solutions. In Jackowski, K., Burduk, R., Walkowiak, K., Wozniak, M., and Yin, H., editors, *Intelligent Data Engineering and Automated Learning – IDEAL 2015*, pages 146–155, Cham. Springer International Publishing.

Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules. In *Proceedings of the 20th International Conference on Very Large Data Bases, VLDB*, pages 487–499. Morgan Kaufmann.

Cao, L., Zhang, C., and Yu, P. S. (2007). Domain driven data mining: Challenges and prospects. In *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, volume 37, pages 767–772.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37–37.

Franco, J., Miranda, F., Stiegler, D., Dantas, F., Brancher, J., and Nogueira, T. (2020). Usando mineração de dados para identificar fatores mais importantes do enem dos Últimos 22 anos. In *Anais do XXXI Simpósio Brasileiro de Informática na Educação*, pages 1112–1121, Porto Alegre, RS, Brasil. SBC.

Guardieiro, V., Raimundo, M. M., and Poco, J. (2022). Analyzing the equity of the brazilian national high school exam by validating the item response theory's invariance.

Romero, C., Ventura, S., Pechenizkiy, M., and Baker, R. S. (2010). *Handbook of educational data mining*. CRC press.

Shearer, C. (2000). The crisp-dm model: the new blueprint for data mining. *Journal of data warehousing*, 5(4):13–22.

Silva Filho, R. L. C., Brito, K., and Adeodato, P. J. L. (2023). A data mining framework for reporting trends in the predictive contribution of factors related to educational achievement. *Expert Systems with Applications*, 221:119729.