Mineração de dados educacionais para identificar a influência de fatores socioeconômicos no desempenho dos estudantes no ENEM

Kazuhiro Daiti Kojio
ICT-UNIFESP
Universidade Federal de São Paulo
São José dos Campos - SP, Brasil
kazuhiro.kojio@unifesp.br

***Abstract*—A mineração de dados educacionais é uma ferramenta de extrema importância e pode auxiliar governos no desenvolvimento de políticas públicas efetivas na área da educação. No Brasil uma rica fonte de informações na área de educação são os microdados do Exame Nacional do Ensino Médio (ENEM), disponibilizados anualmente, que contém diversas informações dos candidatos. Portanto, este trabalho explora o processo do *KDD* (*Knowledge Discovery in Databases*) para extrair conhecimentos a partir dos dados do ENEM. Utilizou-se o algoritmo *Apriori* para extrair regras que associam os fatores socioeconômicos com o desempenho dos candidatos. A análise foi feita utilizando os dados do ENEM entre os anos de 2019 a 2022, o que possibilitou verificar quais regras persistiram em todos estes anos. Assim, este trabalho identifica quais fatores socioeconômicos desempenham grande influência no desempenho dos candidatos.**

Index Terms*—mineração de dados educacionais, KDD, regras de associação, fatores socioeconômicos, ENEM, *apriori

I. Introdução e motivação

O Exame Nacional do Ensino Médio (ENEM) surgiu em 1998 com a proposta de avaliar o desempenho escolar dos estudantes que concluíram a educação básica no Brasil [INEP, 2023], e se manteve exclusivamente com este objetivo por uma década. A partir de 2009, o exame passou a ser uma porta de entrada para quem almeja cursar o ensino superior, permitindo ao estudante utilizar sua nota para se candidatar a uma vaga em uma instituição pública de ensino pelo Sistema de Seleção Unificada (SiSU), ou para bolsas de estudo em instituições privadas pelo Programa Universidade para Todos (ProUni), além de outros programas de financiamento que facilitam o ingresso à instituições de ensino superior. O exame como forma de ingresso a instituições de ensino pode ser realizado por qualquer pessoa que tenha concluído o ensino médio ou que o esteja concluindo, e pode ser feito também por treineiros que queiram apenas avaliar seus conhecimentos.

Por ser um evento que abrange todo o território nacional, é natural que se tenha, em todos os anos, milhões de inscritos para o exame. Somente no ano de 2022, foram mais de 3 milhões de inscritos no exame [INEP 2022], mas já chegou a alcançar mais de 8 milhões de inscritos em 2016. Sendo assim, é uma grande fonte de dados para o desenvolvimento de estudos e indicadores educacionais, pois são coletados diversos dados demográficos e socioeconômicos dos inscritos. Estes dados são anonimizados, juntamente com o resultado final do exame, e são disponibilizados anualmente no próprio site do INEP[1], permitindo que qualquer pessoa que tenha interesse em analisá-los possa utilizá-los para fins de estudo e pesquisa.

Com essa quantidade massiva de dados, pode-se explorar qual o perfil dos estudantes que alcançam um ótimo desempenho no exame e dos que estão muito abaixo da média. Estudos mostram que fatores socioeconômicos influenciam diretamente no acesso de jovens ao ensino superior. Andrade [2012] evidencia que há uma grande disparidade no acesso aos diferentes níveis de ensino quando se observa a renda familiar do estudante, mostrando que os jovens com renda familiar baixa acabam abandonando com frequência o ensino básico e poucos conseguem cursar uma graduação. Entretanto, para os jovens com renda familiar mais elevada o cenário se inverte, pois há pouca evasão no ensino básico e alta concentração de estudantes que conseguem ter acesso ao ensino superior. Para complementar, Araújo [2019] ao investigar fatores determinantes para o bom desempenho no ENEM no município de Viçosa - MG entre os anos de 2015 e 2017, notou que o background familiar contribui de forma significante para um bom resultado, como escolaridade dos pais, acesso à internet em casa e renda familiar.

Os dados do ENEM disponibilizados no portal do INEP possuem informações valiosas, como os dados demográficos e socioeconômicos, além do desempenho, de milhões de inscritos que realizam o exame anualmente. Porém, analisar estas informações de forma manual é completamente inviável dado o grande volume de dados disponíveis. Assim, torna-se necessário utilizar conhecimentos relacionados à mineração de dados para conseguir obter resultados confiáveis de forma automatizada. Segundo Baker *et al.* [2011], a mineração de dados tem como objetivo extrair novas informações a partir da análise de um grande volume de dados, ou seja, baseado nas relações entre os dados contidos em um conjunto de

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

dados (*dataset*), são produzidos novos conhecimentos que não estavam explícitos no *dataset* de entrada.

II. Conceitos fundamentais

Nesta seção são apresentados os conceitos fundamentais sobre mineração de dados e os algoritmos utilizados neste trabalho, e também os trabalhos relacionados.

A. Mineração de dados

De acordo com Faceli *et al.* [2021], a mineração de dados (*data mining*) nada mais é que a extração de novos conhecimentos a partir de um grande conjunto de dados. Em alguns estudos, é comum ela ser chamada também de descoberta de conhecimento em bases de dados (*KDD - Knowledge Discovery in Databases*), mas muitos autores consideram que a mineração de dados é apenas parte do processo de *KDD* [Fayyad *et al.*, 1996].

Uma boa definição para *KDD* que insere a mineração de dados como etapa do processo é: *"O processo não trivial de identificar padrões válidos, novos, potencialmente úteis e, em última análise, compreensíveis em dados."* [Fayyad *et al.*, 1996, p. 30, tradução nossa]. E para realizar este processo, são necessárias algumas etapas: seleção, pré-processamento, transformação, mineração de dados e interpretação dos resultados [Fayyad *et al.*, 1996], conforme ilustrado na **Figura 1**.

Fig. 1: Etapas do processo de KDD. Adaptado de Fayyad *et al.* (1996, p. 29)

A seguir é detalhada cada etapa do processo de KDD:

a) Seleção: Nesta etapa, é feita a seleção do *dataset* a ser analisado. Geralmente, neste momento os dados são chamados de dados brutos, ou seja, dados que ainda não foram tratados para serem minerados adequadamente.

b) Pré-processamento: Esta etapa é responsável pela limpeza dos dados selecionados na etapa anterior, garantindo que não haja dados inconsistentes ou faltantes, assim como qualquer outro fator que possa prejudicar a qualidade dos dados analisados. Ao fim desta etapa, garante-se o controle dos tipos de dados presentes, eliminando quaisquer informações indesejadas ou anomalias nos dados.

c) Transformação: Nesta etapa é realizada a transformação dos dados para ficarem de acordo com o desejado para a análise. Pode-se agregar dados, normalizá-los, criar novos atributos baseados nos já existentes, etc. O resultado desta etapa é um *dataset* pronto para ser minerado.

d) Mineração de dados: Neste estágio do processo *KDD* é onde se aplicam as técnicas de mineração de dados. Com o output desse processo, criam-se condições de validar algumas hipóteses e adquirir novos conhecimentos sobre o *dataset* estudado. Os modelos construídos nesta etapa podem ser preditivos ou descritivos.

e) Interpretação e avaliação dos resultados: Por fim, nesta etapa é possível realizar a análise do desempenho do algoritmo utilizado no *dataset* e gerar informações relevantes sobre os resultados obtidos. A clareza sobre os objetivos desejados com a mineração de dados é importante, pois facilita a validação dos novos conhecimentos obtidos, seja para gerar estatísticas ou para orientar tomadas de decisão do projeto ou pesquisa.

B. Mineração de dados educacionais

Dadas as definições de mineração de dados apresentadas anteriormente, existe uma sub-área de pesquisa que visa trazer todo o contexto de mineração de dados para o âmbito educacional, a mineração de dados educacionais. Com o constante crescimento no uso de ambientes educacionais, tais como Sistemas Tutores Inteligentes (STI), Ambientes Virtuais de Aprendizagem (AVA), entre outros, é possível coletar diversos dados, que são produzidos a partir da interação do aluno com o professor, do aluno com o conteúdo, do professor com a turma e da própria interação dos atores com as plataformas [Costa *et al.*, 2013]. Com estes dados em mãos, é preciso elaborar metodologias que favoreçam a extração de conhecimentos valiosos sobre esses dados educacionais, que irão ajudar os profissionais da área da educação a melhorarem seus processos e a compreenderem como se comporta o processo de aprendizagem de seus alunos.

Assim, há particularidades da área da educação que devem ser levadas em consideração ao se aplicar o processo de *KDD* [Baker, 2010], e isso inclui desde o tratamento dos dados, algoritmo utilizado até a avaliação dos resultados. Na **Seção V** são apresentadas as tomadas de decisão necessárias utilizadas neste trabalho, para a aplicação de mineração de dados educacionais.

C. Modelos descritivos

Quando é preciso estimar uma tendência futura com base na análise de um *dataset*, os modelos preditivos funcionam muito bem. No entanto, este trabalho visa associar dados socioeconômicos e demográficos do estudante com o seu desempenho no exame do ENEM, então os modelos descritivos se adequam melhor com o objetivo do trabalho que será desenvolvido. Assim, quando há um *dataset* e se consegue extrair informações valiosas deste sem a interferência de um fator externo que guie o aprendizado, pode-se defini-lo como aprendizado descritivo ou não supervisionado, e esse novo conhecimento surge a partir das propriedades do próprio *dataset* de entrada [Faceli *et al.*, 2021]. Existem três tipos de tarefas descritivas: agrupamento, sumarização e associação.

O agrupamento é capaz de identificar similaridades entre os objetos contidos no *dataset*, possibilitando a classificação em grupos, e são instrumentos importantes para análises exploratórias em diversas áreas [Faceli *et al.*, 2021]. A sumarização busca organizar os dados de forma que possam ser sumarizadas em descrições mais sintetizadas e simples, por exemplo, ao invés de analisar um grande *dataset*, cada um com sua informação detalhada, pode-se usar alguma medida (média,

desvio padrão, mínimo, etc.) para organizá-los de forma que se possa processar uma quantidade menor de informações, garantindo maior eficiência em processamento, mas com perda de informações. A associação permite que, ao analisar um *dataset*, seja possível identificar padrões com um certo grau de frequência, em que a presença de algum atributo pode estar associada a outro.

O foco deste trabalho está na tarefa de associação, com o intuito de gerar regras que associam dados socioeconômicos com o desempenho dos candidatos do ENEM. Para isso, é necessário discutir um tópico de associação chamado de mineração de itens frequentes, que será apresentado na seção seguinte.

1) Mineração de itemsets frequentes: Han *et al.* [2011] chamam de padrões frequentes aqueles padrões que se repetem com frequência em um conjunto de dados, e ainda afirma que há diferentes tipos de padrões, como *itemsets* frequentes, subsequências frequentes e subestruturas frequentes. Um exemplo clássico de cesta de compras explica bem essa definição para *itemsets* frequentes: imagine uma base de dados de um supermercado em que são armazenadas várias transações, e cada transação tem um conjunto de itens (*itemset*) adquiridos. Ao realizar uma breve análise das transações, nota-se que em quase todas as transações que contém café na cesta de compras, também tem leite, então esse par de itens que estão repetidamente presentes em conjunto na transação, são padrões frequentes. Inclusive, os primeiros trabalhos realizados nesta área foram para análise de cestas de compras, utilizando mineração de dados transacionais para entender o comportamento do cliente [Faceli *et al.*, 2021].

Formalmente, seja A = a_1,...,a_m o universo de m itens quaisquer. Um *itemset* representa qualquer subconjunto de itens de A que podem ser adquiridos juntos. Por exemplo, no universo de itens de um mercado U = pão, cerveja, manteiga, fralda, o subconjunto I = pão, manteiga é um *itemset* de U. Seja T = t_1,...,t_n um conjunto de n transações. Cada transação conta com um identificador tid e um *itemset* k_n tal que $k_n \subset A$. Este conjunto T de transações nada mais é que uma listagem de *itemsets*, que pode armazenar, por exemplo, todas as transações de um mercado com os itens comprados por cada cliente. A partir do conjunto de transações, pode-se identificar quais itens são mais frequentes, e com isso, é possível derivar algumas regras de associação. Faceli *et al.* [2021] afirmam que regras de associação são da forma se antecedente então consequente ($consequente \cup antecedente$), ao qual antecedente e consequente são *itemsets*.

Para auxiliar no entendimento do conceito de regras de associação, a Tabela I apresenta transações com alguns produtos, ao qual cada linha é uma transação e, caso esteja com o valor 1, quer dizer que o produto foi comprado na transação em questão, e se estiver com valor 0, que não foi comprado. A transação 1 indica que foram comprados os itens: {Pão, Manteiga, Fralda}.

Como exemplo, observando as transações da tabela, nota-se que toda transação que tem pão, também tem manteiga, logo, é possível derivar uma regra $\{Pão\} \Rightarrow \{Manteiga\}$,

Tabela I: Tabela de transações de produtos

	Produtos			
tid	***Pão***	***Manteiga***	***Cerveja***	***Fralda***
1	1	1	0	1
2	1	1	1	1
3	0	1	1	1
4	0	0	1	1
5	1	1	0	0

ou seja, ao comprar pão (antecedente) é provável que seja comprado manteiga (consequente) também. Para extrair regras conforme o exemplo supracitado, é interessante levar em consideração algumas medidas que indicam a relevância das regras derivadas:

- **Suporte:** dado um conjunto de transações D, a porcentagem do quão frequente $A \cup B$ é contido em D, ou seja, *suporte*($A \Rightarrow B$) = P($A \cup B$) onde P é a função de probabilidade. Assim, podemos afirmar que uma regra de associação é frequente de acordo com o seu suporte. Assim, temos que *suporte*($Pão \Rightarrow Manteiga$) = $\frac{3}{5}$ = 0.6 = 60%. Quando é informado um suporte mínimo, só interessa extrair dados de itens que tiverem o suporte mínimo desejado, ou seja, se for informado um suporte mínimo de 30% para o algoritmo, não serão consideradas regras de associação em que o conjunto de itens (que fazem parte desta regra) não estão presentes em menos de 30% do *dataset*.
- **Confiança:** consiste na probabilidade de ocorrer um conjunto de itens dado que já ocorreu um outro conjunto [Faceli *et al.* 2021]. Assim, seja A o conjunto antecedente, B o item consequente, e P seja a função de probabilidade, define-se que *confiança*($A \Rightarrow B$) = P($A \cup B$) / P(A) = *suporte*($A \cup B$) / *suporte*(A). Assim, *confiança*($Pão \Rightarrow Manteiga$) = ($\frac{3}{5}$) / ($\frac{3}{5}$) = 1 = 100%, ou seja, em todas as vezes que foi comprado pão, a manteiga também foi comprada.
- ***lift***: também chamado de coeficiente de interesse, calcula a chance de B estar presente se o A está presente, considerando a frequência de B. Essa medida da correlação dá mais suporte à informação de confiança. Pode ser expressa na forma *lift*($A \Rightarrow B$) = *confiança*($A \Rightarrow B$) / *suporte*(B) = *suporte*($A \cup B$) / ((*suporte*(A) * *suporte*(B)). Faceli *et al.* [2021] explicam que quando o valor do *lift* é maior que 1, indica que A e B aparecem com frequência juntos, então A tem um impacto positivo sobre a ocorrência de B. Se o *lift* for menor do que 1, indica que A e B aparecem com pouca frequência juntos e que A tem um impacto negativo na ocorrência de B. Enfim, se o *lift* for igual a 1 (ou muito próximo), indica que A e B estão presentes quase sempre em conjunto, logo, a ocorrência de B não tem impacto sobre a ocorrência de A.

Regras que satisfazem um suporte e confiança mínimos dado um limite estipulado, são chamadas de fortes [Han *et al.*, 2011]. O suporte pode ser absoluto ou relativo, em que

o primeiro é a quantidade total de elementos no conjunto de transações, enquanto o segundo é a quantidade total de elementos (suporte absoluto) dividido pelo número de transações [Faceli *et al.*, 2021]. Assim, os principais objetivos para mineração de *itemsets* frequentes é conseguir encontrar itens frequentes, respeitando o suporte relativo mínimo informado pelo usuário e conseguir extrair regras de associação com o grau de confiança mínimo informado.

D. Algoritmos

Diversos algoritmos foram propostos com intuito de extrair regras de associação, a partir de *itemsets* frequentes, de um conjunto de dados. Nesta seção são apresentados dois desses algoritmos, mas com foco especial no *Apriori*, que foi utilizado nos experimentos. Para compreender melhor o racional dos algoritmos abaixo, é preciso ter clareza do que são regras de associação: pense que você é responsável pela conferência das transações de um mercado e começa a perceber um padrão frequente (**Seção 2.A.1**) entre os itens que são adquiridos pelos cliente, que é o fato de que quase todos os clientes que compram pão também compram manteiga. Essa ação de explorar essas relações entre itens em um conjunto de dados pode ser considerada uma regra de associação, que pode ser relevante ou não, de acordo com a confiança de interesse.

1) Apriori: O algoritmo *Apriori* é focado na mineração de *itemsets* frequentes e construção de regras de associação [Agrawal et al. 1993; Agrawal & Srikant 1994]. Seu princípio é baseado na ideia de que qualquer subconjunto de *itemsets* frequentes deve ser um *itemset* frequente [Faceli et al. 2021]. O algoritmo *Apriori* utiliza o suporte como parâmetro para extrair um subconjunto de itens frequentes, e para isso precisa calcular o suporte de todas as combinações previamente. Basicamente, os passos do algoritmo são:

1) Extrair um subconjunto de itens frequentes do itemset que atendam ao suporte mínimo informado;
2) Iterar dentro do subconjunto de itens frequentes para formar combinações com o *itemset*, aplicando o suporte mínimo e acumulando no subconjunto;
3) Continua iterando até não sobrar mais itens ao aplicar o suporte mínimo.

O *output* destes passos serão as regras de associação com o suporte mínimo desejado.

2) FP-Growth: Este algoritmo, diferentemente do *Apriori*, utiliza a estratégia de busca por profundidade e árvores de sufixo. Para encontrar as regras de associação, o *FP-growth* conta com duas etapas: primeiro se percorre a base de dados duas vezes para construir uma árvore de padrões frequentes (*FP-tree*), e a partir desta árvore é possível encontrar as regras de associação. Na primeira varredura dos dados, segundo explica Faceli *et al.* [2021], é definido o conjunto de itens frequentes e seu suporte, que são armazenados em uma matriz em ordem decrescente por suporte. Já na segunda varredura, é construída a *FP-tree* a partir de base de transações. A árvore gerada pelo algoritmo é uma representação das regras que se deseja extrair. É importante ressaltar que o *FP-growth* é muito utilizado em algoritmos de mineração de padrões frequentes quando conta com dados de fluxo contínuo [Chi *et al.*, 2004].

III. Trabalhos relacionados

Silva *et al.* [2020] realizaram, com os microdados do ENEM de 2019, um trabalho ao qual aplicou técnicas de clusterização e mineração de regras de associação para mapear as variáveis determinantes para o desempenho dos estudantes concluintes do ensino médio no estado de Minas Gerais. O processo utilizado nesse trabalho foi *KDD*, aplicando todas as etapas descritas na Seção 2.A. No final da análise, os autores perceberam que a renda familiar e a dependência administrativa da escola são fatores que afetam o desempenho dos estudantes. Para os grupos de alunos com notas mais baixas, percebeu-se a predominância de estudantes de baixa renda familiar e em sua grande maioria de escolas estaduais. Já no grupo com notas mais altas, percebeu-se predominância de alunos das redes particular e federal sobre as demais, podendo ser notado uma semelhança entre os desempenhos destes alunos (federal e particular). Em sua conclusão, levantou-se dois pontos: a necessidade de discutir o modelo das escolas federais, uma vez que são públicas e mostram bom desempenho se comparado às estaduais e municipais, e aprofundar a análise dos aspectos socioeconômicos dos estudantes de escolas privadas e federais para compreender melhor essas variáveis.

Adeodato [2016] explorou também a base de dados do ENEM e do Censo Escolar de 2011 para avaliar a qualidade das escolas secundárias privadas. Em seu trabalho, utilizou o método *Cross Industry Standard Process for Data Mining* (CRISP-DM), que foi criado em 1996 como um conjunto de boas práticas para execução de um projeto de ciência de dados. Não será apresentado detalhes do método, mas vale ressaltar que essa metodologia consiste em seis etapas principais: entendimento do negócio, entendimento dos dados, preparação dos dados, modelagem, avaliação e deployment. Ao analisar os resultados do seu trabalho, Adeodato [2016] notou que fatores econômicos influenciam diretamente o desempenho, como renda familiar e características do domicílio do estudante, e indiretamente dependendo da região em que a escola está localizada. Adeodato [2016] mostrou também que a infraestrutura do laboratório de informática das escolas deixam a desejar. Há um contraponto do autor no trabalho, que levanta a necessidade de se analisar com cautela esses resultados, pois houve indícios de fraudes nas edições do ENEM do período, o que poderia enviesar algumas métricas do lado socioeconômico, e sobre os laboratórios de informática, deve-se entender qual foi o real uso do recurso, pois pode ter sido utilizados sem fins educacionais, como jogos e redes sociais.

Barcellos *et al.* [2020] aplicaram mineração de dados educacionais na base de dados do ENEM de 2018 para avaliar o perfil dos alunos utilizando técnicas de agrupamento. O objetivo era identificar os fatores que influenciam no desempenho do estudante, como escolaridade dos pais, renda familiar, acesso à infraestrutura, entre outros. A metodologia utilizada foi o *KDD*, com ferramentas que auxiliaram a aplicação de

clusterização também. O autor alcançou resultados semelhantes aos dos trabalhos apresentados anteriormente.

Os trabalhos supracitados mostram, por meio de regras de associação, que fatores socioeconômicos afetam diretamente o desempenho do estudante, assim como a escolaridade dos pais e estrutura doméstica, como acesso à internet. No entanto, estes trabalhos realizam análise de desempenho de um ano específico do ENEM ou de um estado brasileiro em especial. O presente trabalho visa analisar associações do desempenho dos inscritos do ENEM com seus dados socioeconômicos, em todo o território nacional, e que persistiram durante o período de 2019 a 2022.

IV. Objetivo

O objetivo deste trabalho foi aplicar conhecimentos acerca de modelos descritivos, que consiste em identificar propriedades e extrair informações a partir de um *dataset* sem a necessidade de um elemento externo para guiar o aprendizado [Faceli *et al.*, 2021], para compreender quais os fatores influenciam no desempenho do estudante ao realizar o exame do ENEM, com ênfase em dados demográficos e socioeconômicos. Para esse trabalho, utiliza-se o algoritmo *Apriori* [Agrawal *et al.* 1993; Agrawal & Srikant 1994] para identificar regras que associam as informações socioeconômicas e demográficas dos inscritos com seus desempenhos. No entanto, o enfoque da análise será em desempenhos que estão muito acima ou muito abaixo da média, e quais regras persistem entre os exames de 2019 até 2022.

V. Metodologia experimental

Para este trabalho, foram utilizados os microdados do ENEM do ano de 2019 até 2022 a fim de identificar, utilizando regras de associação, os fatores socioeconômicos e demográficos que influenciam no desempenho dos inscritos e persistiram neste período de quatro anos. O objeto de interesse deste trabalho são as regras geradas para quem tem o desempenho muito acima ou muito abaixo da média.

Toda a análise e experimentos, apresentados neste trabalho, foram desenvolvidos em linguagem *Python* [*Python Software Foundation 2022*]. O algoritmo escolhido para os experimentos foi o *Apriori*, e utilizou-se da biblioteca *apyori* [PyPi 2019], que é uma implementação do algoritmo *Apriori* em *Python*. O código-fonte do trabalho está disponível em um repositório público[2] do *Github* [*Github* Inc. 2023], onde é possível acessar a implementação dos passos aqui apresentados e as instruções para reprodução dos experimentos apresentados nesta seção[3].

A seguir, são apresentados os detalhes e as tomadas de decisões de projeto conforme as etapas do processo de *KDD*, descritos na **Seção 2.A**.

[2]Disponível em: https://github.com/kazuhirodk/enem-data-mining

[3]Todos os experimentos foram realizados em um computador com as seguintes configurações: MacBook Pro (14-inch, 2023), chip Apple M2 Pro (10-core CPU, 16-core GPU, 16-core Neural Engine), 16GB de memória, sistema operacional macOS Ventura 13.4.

A. Seleção

O primeiro passo prático deste trabalho de mineração de dados educacionais foi buscar os microdados do ENEM de 2019 até 2022, e estes dados são públicos e estão disponíveis no INEP [2023b] para *download*. No portal do INEP é possível encontrar microdados do ENEM a partir do ano de 2010, e também há microdados de outros programas, como Enade, Censo Escolar, Censo de Educação Superior, Encceja, entre outros. Na **Tabela II** foram descritos os tamanhos dos *datasets* selecionados, que totalizam quase 18 milhões de linhas e aproximadamente 7,54 GB de dados.

Tabela II: Informações de arquivo de microdados do ENEM

Ano	*Linhas*	*Colunas*	*Tamanho (GB)*
2019	5.095.170	76	2.42
2020	5.783.108	76	2.03
2021	3.389.831	76	1.51
2022	3.476.104	76	1.58

Cada um destes microdados é um arquivo no formato **.csv** e contam com um dicionário de referência, para que seja possível entender do que se trata cada coluna e dado presente. Por ser um volume grande de dados, houve dificuldade em abrir estes arquivos com algumas ferramentas de planilhas, como *Microsoft Excel* ou *Numbers*. Para manipular estes arquivos, foi utilizado o *Pandas* [*The pandas development team* 2022], uma ferramenta *open source* muito comum para manipular e analisar dados, implementada em *Python*.

B. Pré-processamento

Como esperado, estes microdados carregam muitas informações dos estudantes e, consequentemente, muitas colunas. Para conseguir executar os algoritmos em uma quantidade muito grande de dados e combinações, é necessário muito recurso computacional, além de exigir bastante tempo de execução. Como o objetivo deste trabalho é associar o desempenho do estudante com o seu contexto socioeconômico e demográfico, é interessante extrair apenas as colunas de interesse e ignorar linhas que contém dados que não são interessantes para a análise, como os valores nulos.

Assim, baseado nos dicionários que acompanham os microdados, foram selecionadas as colunas apresentadas no **APÊNDICE A**, onde consta o nome da coluna, seu significado, o formato do dado no *dataset* (item), a descrição do item e o tipo de dado que aquela coluna tem.

Com essa extração, a quantidade de colunas do *dataset* foi reduzida para 19. Para evitar qualquer problema nas análises, foram filtradas apenas as linhas que tenham todos os seus dados válidos, ou seja, que não sejam nulos para estas colunas selecionadas. Após este filtro, a quantidade de linhas do *dataset* reduziu, tornando mais viável a análise com os recursos computacionais disponíveis para este trabalho. A **Tabela III** apresenta a quantidade de linhas antes e depois do pré-processamento descrito.

Percebe-se uma redução significativa no número de inscritos do ENEM com o passar dos anos, especialmente na edição

Tabela III: Redução de linhas após pré-processamento

Ano	*Linhas totais*	*Linhas filtradas*	*Redução em %*
2019	5.095.170	3.701.909	27,34%
2020	5.783.108	2.561.304	55,71%
2021	3.389.831	2.238.106	33,87%
2022	3.476.104	2.344.823	32,54%

de 2021, que pode ser justificada pelo período de pandemia pelo *Covid-19*, que afetou de várias formas o ano letivo dos estudantes. Muitos estudantes tiveram suas aulas presenciais canceladas e muitos não tinham acesso à internet para acompanhar as aulas online, dificultando o acompanhamento do conteúdo e aumentando a insegurança para realizar a prova. Além disso, a evasão escolar pode ter sido um fator importante para a baixa adesão ao ENEM 2021 [Novaes 2021].

C. Transformação

Como apresentado na **APÊNDICE A**, nota-se que as informações do *dataset* estão com os dados brutos ainda, por exemplo, a coluna de TP_COR_RACA tem como dados os valores de 0 a 5, e isso fora de contexto fica inviável de compreender, já que existem outros campos que utilizam o mesmo critério. Então o que foi feito nesta etapa é a tradução de todas as respostas do *dataset* de forma que seja compreensível mesmo que não tenha informações da coluna.

Para este trabalho, foi utilizada a tradução se baseando no dicionário apresentado na **APÊNDICE A** para boa parte das colunas, porém, há alguns dados que foram agregados e/ou agrupados. A seguir, serão apresentados os agrupamentos realizados:

- **SG_UF_PROVA**: foi agrupado por regiões do Brasil (ex: SP se torna **Região Sudeste**). Como o interesse é avaliar todo o território nacional, não era de interesse separar por estado. Assim, foi preferível converter para região cada um dos estados.
- **Q006**: de acordo com o dicionário, esta coluna se refere à renda mensal da família. Assim, foi preferível agrupar em classe social, de acordo com os últimos dados do IBGE (ex: B - Até R$ 1.045,00 se torna **Classe E**).
- **Q022**: de acordo com o dicionário, esta coluna se refere à quantidade de celulares que o candidato tem. Como não é de interesse saber a quantidade, foi agrupado apenas em **Tem/Não tem celular**.
- **Q024**: de acordo com o dicionário, esta coluna se refere à quantidade de computadores que o candidato tem. Como não é de interesse saber a quantidade, foi agrupado apenas em **Tem/Não tem computador**.

Após essa primeira tradução dos dados, é possível compreender melhor o conteúdo de cada linha. Por exemplo:

- **Antes da tradução:** [4, F, 1, 3, 2, 2, 2, 2.0, CE, 497.7, 532.4, 457.6, 582.6, 780.0, D, E, C, E, B, B]
- **Após a tradução:** [Feminino, Solteiro(a), Pardo, Entre 17 e 25 anos, Estou cursando e concluirei o Ensino este ano, Pública, Estadual, Região Nordeste, 497.7, 532.4, 457.6, 582.6, 780.0, Pai Ensino médio incompleto, Mãe Ensino médio completo, Classe Social E, Tem celular, Tem computador, Tem internet em casa]

Quando for aplicado o algoritmo *Apriori*, cada uma dessas listas será considerada uma transação, então é essencial que se tenha clareza dos dados da transação. Alguns campos não estão devidamente transformados ainda, como as notas do inscrito. Para classificar o desempenho dos inscritos baseado na nota, foram necessários alguns passos adicionais:

1) Foi criada uma coluna nova chamada NU_NOTA_GERAL, que é a média aritmética simples de todas as notas incluindo a redação;
2) Para facilitar a comparação das notas entre os candidatos, foi aplicado o *z-score* na coluna **NU_NOTA_GERAL** e gerada uma nova coluna chamada **Z_SCORE_NOTA**. A fórmula básica para calcular o *z-score*, assumindo uma distribuição normal, é $z = \frac{(x-\mu)}{\sigma}$, onde μ é a média da população, σ é o desvio padrão da população e o x é o valor a ser testado. Por exemplo, se calcular o *z-score* de uma nota x = 700, em que a média das notas é $\mu = 600$ e o desvio padrão é $\sigma = 50$, tem-se que $z = \frac{(700-600)}{50} = 2$, ou seja, o *z-score* é 2, que representa a distância, em desvio padrão, que x está distante da média em uma distribuição normal. O *z-score* geralmente varia de -3 a +3, onde 0 é a média. Se um *z-score* está acima de +3 ou abaixo de -3, são considerados *outliers*. A **Figura 2** mostra uma distribuição normal, ilustrando como é distribuído e classificado o *z-score* em relação à média quando ela é 0;
3) Por fim, uma nova coluna chamada **CLASSIFICACAO_NOTA** foi criada, e o desempenho foi classificado da seguinte forma:
   - ***z-score*** ≤ -2, NOTA: MUITO ABAIXO DA MÉDIA
   - $-2 <$ ***z-score*** ≤ -1, NOTA: ABAIXO DA MÉDIA
   - $-1 <$ ***z-score*** ≤ 1, NOTA: MÉDIA
   - $1 <$ ***z-score*** ≤ 2, NOTA: ACIMA DA MÉDIA
   - ***z-score*** > 2, NOTA: MUITO ACIMA DA MÉDIA

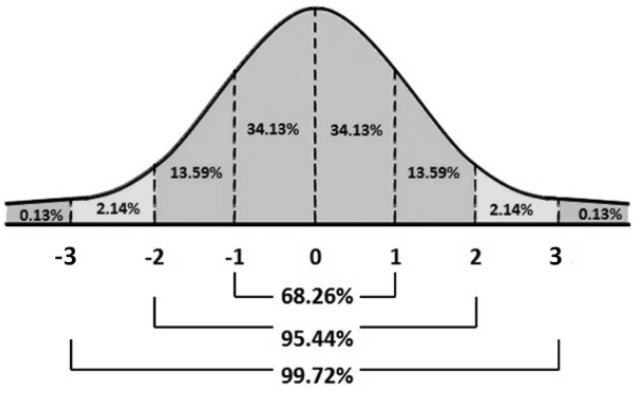


Fig. 2: Curva com distribuição normal com base no *z-score*

Na **Figura 3** é apresentado o histograma com a distribuição das notas do ano de 2019 depois de aplicar o *z-score*. Observa-

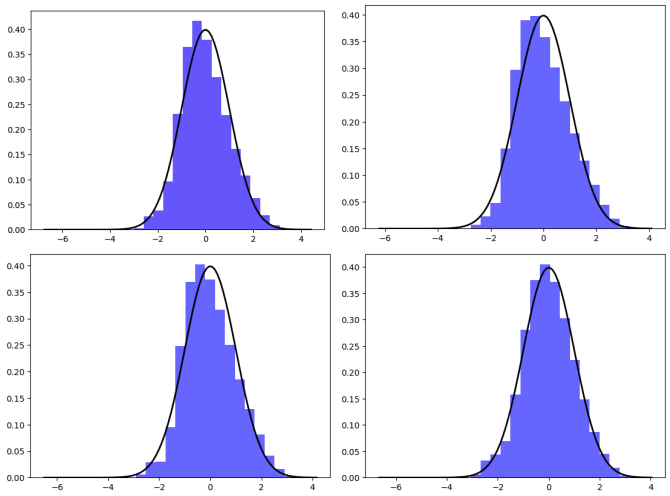


Fig. 3: Distribuição das notas (por *z-score*) do ENEM de 2019 a 2022

se que a quantidade de notas que estão acima de 2 e abaixo de -2, que são o foco deste trabalho, corresponde a uma parcela bem pequena do conjunto total de dados, conforme mostrado na **Figura 2**, este valor é inferior a 5% do valor total de dados. Com isso, é possível compreender como se comportam os dados que estão sendo preparados para a análise.

Após isso, o *dataset* conta com uma nova coluna em que se tem a classificação das notas dos inscritos, porém, como o interesse do estudo é apenas sobre os estudantes que tiveram uma nota muito acima da média ou muito abaixo da média, foram extraídos do *dataset* apenas os dados dos candidatos que têm o *z-score* acima de 2 (muito acima da média) e abaixo de -2 (muito abaixo da média). Assim, a quantidade de linhas foi reduzida novamente, conforme apresentado na **Tabela IV**.

Tabela IV: Redução de linhas após transformação

Ano	***Qtde. inicial***	***Qtde. final***	***Redução em %***
2019	5.095.170	189.427	96,28%
2020	5.783.108	117.225	97,97%
2021	3.389.831	110.263	96,74%
2022	3.476.104	125.412	96,39%

Agora que todos os dados já estão no formato desejado para a análise, foram selecionadas as seguintes colunas listadas na **Tabela V** que serão exploradas pelo processo de mineração de dados:

Note que, ao final desta etapa, está bem definido os tipos de dados contidos nestas colunas, ou seja, mesmo sem contexto do domínio, é possível entender o significado de cada linha, como mostrado no exemplo abaixo:

Exemplo: *[Feminino, Solteiro(a), Pardo, Entre 17 e 25 anos, Estou cursando e concluirei o Ensino este ano, Pública, Estadual, Região Nordeste, Pai Ensino médio incompleto, Mãe Ensino médio completo, Classe Social E, Tem celular, Tem computador, Tem internet em casa, NOTA: MUITO ACIMA DA MÉDIA]*.

Tabela V: Colunas selecionadas para análise

TP_SEXO	Q001
TP_ESTADO_CIVIL	Q002
TP_COR_RACA	Q006
TP_FAIXA_ETARIA	Q022
TP_ST_CONCLUSAO	Q024
TP_ESCOLA	Q025
TP_DEPENDENCIA_ADM_ESC	CLASSIFICACAO_NOTA
SG_UF_PROV	

As etapas de mineração de dados e interpretação dos resultados estão condensadas na **Seção VI**, que será apresentada a seguir.

VI. Resultados

Nesta seção são apresentados os resultados do algoritmo *Apriori* que foi executado sobre os dados processados indicados na **Seção V** para os anos de 2019 a 2022 do ENEM.

O primeiro passo foi transformar os dados em uma lista de transações, que é um *array* de *itemsets* com os dados relacionados a cada inscrito, e aplicou-se o *Apriori* para estas transações. Foram utilizados nos experimentos 10% de suporte mínimo e 50% de confiança mínima como parâmetros do algoritmo *Apriori*. Após a execução do algoritmo, cada *dataset* gerou uma quantidade diferente de regras. A **Tabela VI** apresenta o tempo de execução e a quantidade de regras de associação encontradas para cada ano.

Tabela VI: Regras de associação geradas e tempo de execução

Ano	***Tempo de execução*** (HH:MM:SS)	***Qtde. de itemsets***	***Qtde. de regras de associação***
2019	1:45:41	10.008	291.866
2020	1:52:14	13.073	485.076
2021	1:27:46	12.015	411.157
2022	0:57:03	10.066	272.535

Na **Figura 4** é mostrada a quantidade de regras de associação geradas de acordo com a variação do suporte. Nota-se, de maneira clara, que conforme o suporte aumenta o número de regras diminui. É importante ressaltar que o ano de 2020, partindo de um suporte mínimo de 10%, foi o que gerou mais regras dentre os anos testados. O gráfico mostra que a maior concentração de regras de associação geradas estão em até cerca de 30% de suporte.

Traçou-se o mesmo gráfico para a quantidade de regras de associação geradas de acordo com a confiança mínima obtida, conforme a **Figura 5**. Pode-se notar que há uma redução do número de regras à medida que a confiança mínima aumenta. O ano de 2020 gerou mais regras que os demais anos em todos os graus de confiança mínima. É importante destacar que estas regras de associação que foram geradas envolvem todas as variáveis, e não se restringem às que incluem a nota geral do candidato.

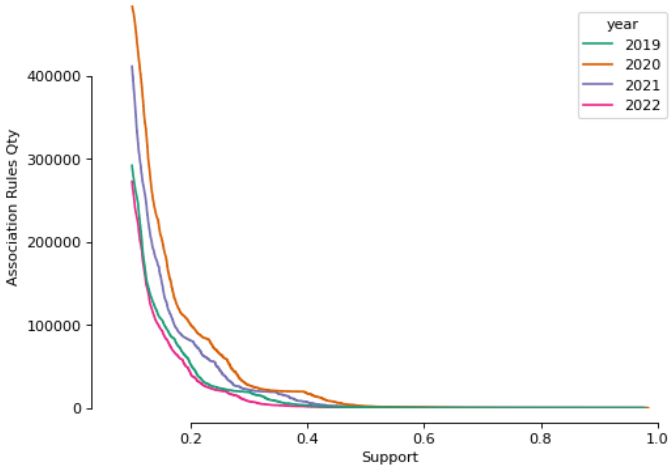


Fig. 4: Gráfico de Suporte x Qtde. Regras de Associação

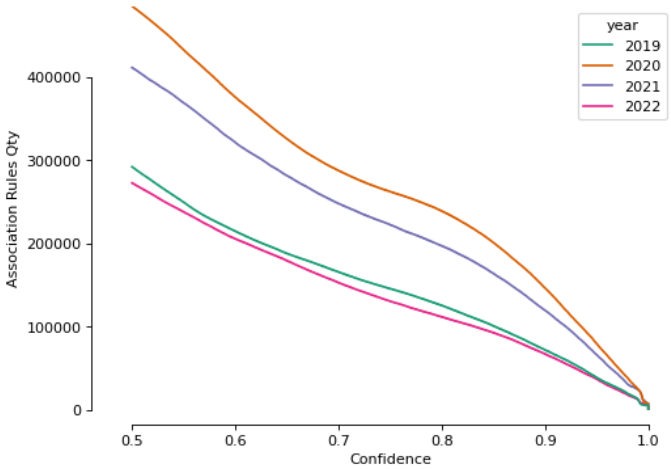
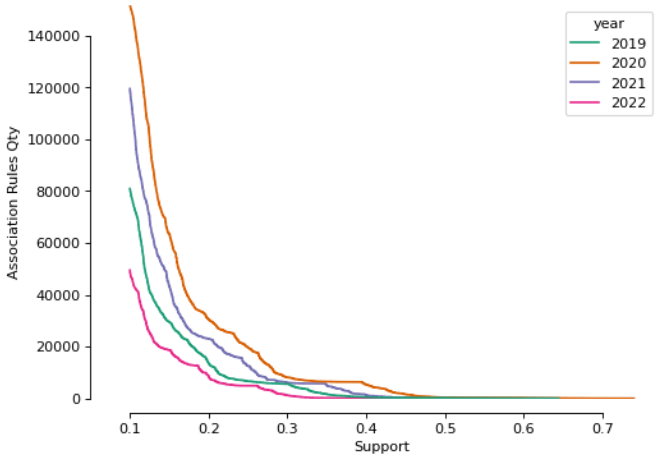
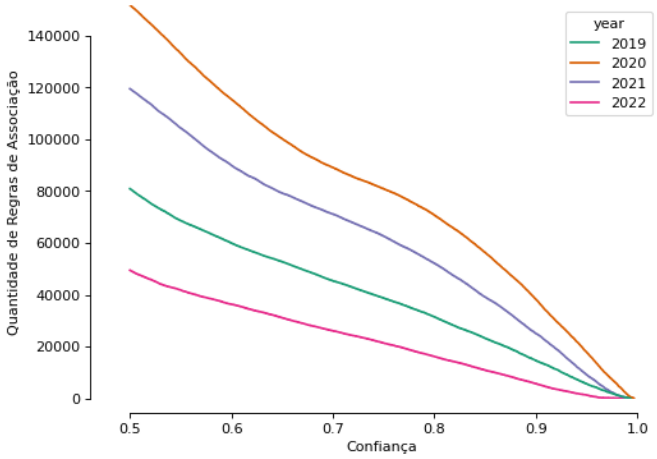


Fig. 5: Gráfico de Confiança x Qtde. Regras de Associação



(a) Regras de Associação x Suporte



(b) Regras de Associação x Confiança

Fig. 6: Quantidade de regras de associação para desempenhos muito acima da média

Pode-se traçar também os mesmos gráficos filtrando apenas por regras que envolvam o desempenho do candidato. Na **Figura 6** são apresentados os gráficos com a quantidade de regras de associação geradas, por ano, com o suporte mínimo de 10% e confiança mínima de 50% e que tenham a nota muito acima da média como consequente.

Na **Figura 7** são apresentados os mesmos gráficos, porém relacionados às regras que envolvam nota muito abaixo da média como item consequente. É possível notar que foram geradas muito mais regras de associação relacionadas ao desempenho muito acima da média do que para desempenhos muito abaixo da média, se for desconsiderado o *lift*.

Para este trabalho, o interesse é identificar regras que persistiram no decorrer dos quatro anos analisados. A **Tabela VII** apresenta a quantidade de regras que persistiram no decorrer dos anos e, além disso, estão relacionadas ao desempenho do candidato.

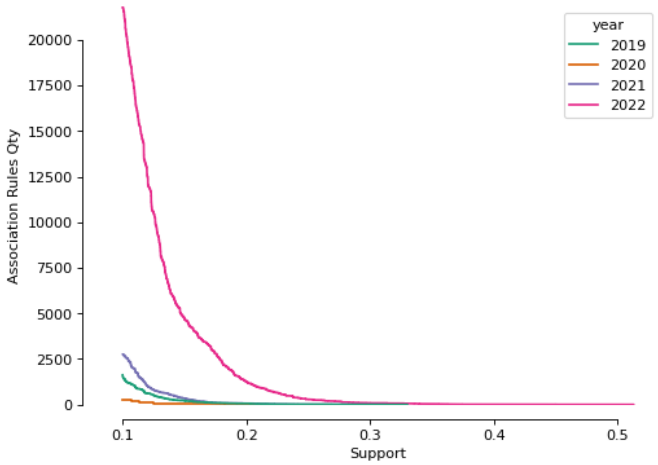
Para ter uma representação mais visual dos resultados, a **Figura 8** mostra, por nuvem de palavras, os itens que acompanham, como antecedentes, as notas muito acima da média de forma frequente, enquanto que a **Figura 9** apresenta os itens que acompanharam, como antecedentes, as notas muito abaixo da média.

Tabela VII: Tabela com a quantidade de regras que persistiram em 2019 a 2022

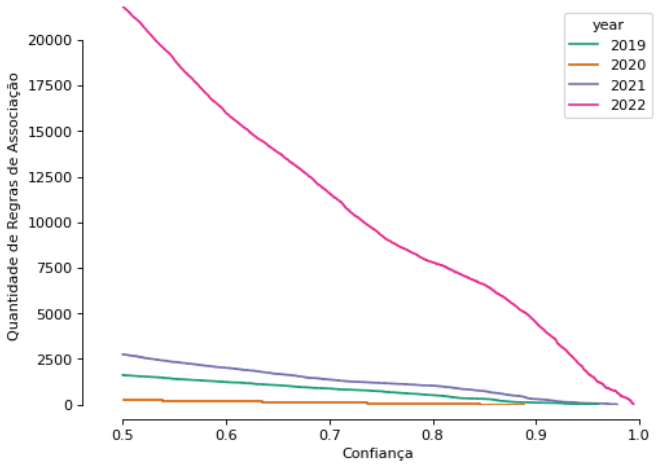
Tipo de regra	***Quantidade (regras)***
Regras totais	150.259
Regras com notas muito acima da média	46.785
Regras com notas muito abaixo da média	251

A partir da leitura dos resultados das regras de associação persistentes, destacam-se os seguintes aspectos sobre os quatro anos de prova para os candidatos que tiveram desempenho muito acima da média:

- São autodeclarados brancos
- Integrantes da classe social C (segundo o IBGE)



(a) Regras de Associação x Suporte



(b) Regras de Associação x Confiança

Fig. 7: Quantidade de regras de associação para desempenhos muito abaixo da média

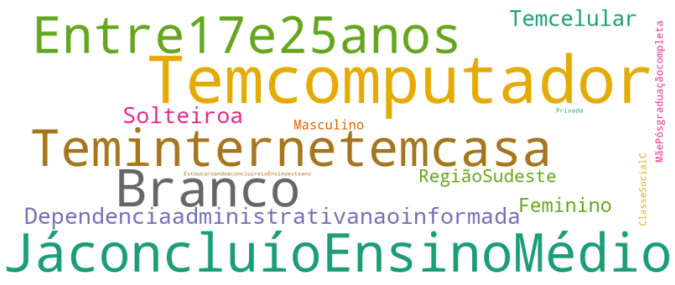


Fig. 8: Nuvem de palavras que acompanham notas muito acima da média

Fig. 9: Nuvem de palavras que acompanham notas muito abaixo da média

- Tem acesso a pelo menos um computador em casa
- Jovens entre 17 e 25 anos
- Tem pelo menos um celular
- Tem acesso à internet em casa
- Mãe com pós-graduação completa
- Estudantes da rede privada de ensino
- Residentes da região sudeste
- Já concluíram o ensino médio

E para os inscritos que tiveram desempenho muito abaixo da média, destacam-se os seguintes aspectos:

- São autodeclarados pardos
- Integrantes da classe Social E (segundo o IBGE)
- Não tem computador em casa
- Jovens entre 17 e 25 anos
- Residentes da região nordeste
- Tem pelo menos um celular
- Tem acesso à internet em casa

Dado os resultados apresentados, nota-se que os conhecimentos extraídos destes experimentos são similares aos trabalhos realizados por outros autores, mostrando que fatores socioeconômicos impactam diretamente no desempenho dos inscritos no ENEM de forma recorrente.

Do ponto de vista familiar, é possível notar que a escolaridade dos pais é um fator que impacta diretamente no bom desempenho do estudante. Isso poderia ser justificado pela possibilidade dos pais poderem contribuir para os estudos dos filhos no dia-a-dia ou promoverem melhores condições para a qualidade do aprendizado.

Do ponto de vista financeiro, nota-se que a renda familiar também é um fator que influencia no desempenho da prova, uma vez que se identificou uma maior ocorrência de inscritos na classe C com uma nota boa, enquanto que os inscritos da classe E estão bastante presentes nos exames com nota muito abaixo da média. Estudantes da rede privada de ensino também são frequentes no grupo com notas muito acima da média, indicando que a qualidade do ensino dessas instituições geralmente é superior.

Do ponto de vista social e demográfico, percebe-se, nos inscritos com bom desempenho, uma presença constante de auto declarados brancos, que pode levantar pontos de atenção em relação à desigualdade social que existe no Brasil, indicando a necessidade de medidas afirmativas mais eficientes. É possível

notar também que os inscritos da região sudeste tendem a ter notas superiores no ENEM.

É importante ressaltar que as análises realizadas no experimento não consideram o contexto do Brasil nos anos correspondentes, e isso pode acarretar em inconsistências nos resultados ano a ano. Por exemplo, nos anos de 2020 a 2022 o mundo passou por uma pandemia de *Covid-19* que gerou crises em diversos setores, e é perceptível a mudança na quantidade de inscritos no ENEM durante este período, então não é possível assumir que o comportamento e os fatores que influenciam no resultado do exame será o mesmo para todos os anos.

Como os dados abrangem parte de cada período (pré-covid nos microdados de 2019 a 2020 e pós-covid nos microdados de 2021 a 2022), é possível realizar a mesma análise separadamente para cada um dos períodos e identificar as principais diferenças entre as regras obtidas.

Nas subseções seguintes, o interesse é analisar quais foram as regras geradas relacionadas ao desempenho dos candidatos durante o período em que não ocorria a pandemia de *Covid-19* (2019 a 2020) e comparar com as regras, sob os mesmos parâmetros, do período em que mesma ocorria (2021 a 2022).

A. Regras de associação no período pré-covid e pós-covid para desempenho muito acima da média

A primeira análise é sobre os candidatos que tiveram resultados muito acima da média. Seguindo o mesmo processo descrito nos experimentos anteriores, foram extraídas as regras que foram recorrentes durante o período pré-covid e pós-covid com os mesmos parâmetros (10% de suporte mínimo e 50% de confiança mínima). Na **figura 10** é possível verificar as regras que cada período gerou.

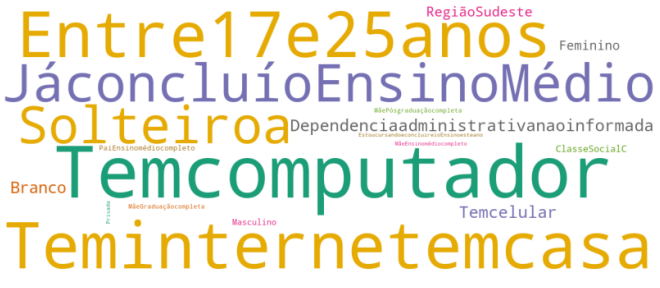
É possível notar, baseado na nuvem de palavras gerada, que os fatores não se alteraram de forma expressiva, comparando-se os dois períodos, para candidatos com desempenho muito acima da média. O perfil de ambos é quase idêntico, porém, há uma pequena diferença entre eles: o pai ter uma pós-graduação completa é um fator relevante para o bom desempenho no cenário pós-covid, pois no período pré-covid apenas a mãe ter uma pós-graduação completa era identificada nas regras.

Uma possível explicação para essa diferença é o fato de, na ausência das aulas presenciais durante a pandemia, os pais com uma formação acadêmica superior são capazes de dar um melhor suporte aos estudos dos dos filhos.

B. Regras de associação no período pré-covid e pós-covid para desempenho muito abaixo da média

A segunda análise é sobre os candidatos que tiveram resultados muito abaixo da média. Seguindo o mesmo processo descrito nos experimentos anteriores, foram extraídas as regras que foram recorrentes durante o período pré-covid e pós-covid com os mesmos parâmetros (10% de suporte mínimo e 50% de confiança mínima). Na **figura 11** é possível verificar as regras que cada período gerou.

É possível notar, baseado na nuvem de palavras gerada, que os fatores não se alteraram de forma expressiva, comparando-se os dois períodos, para candidatos com desempenho muito



(a) Nuvem de palavras para o período pré-covid

(b) Nuvem de palavras para o período pós-covid

Fig. 10: Regras de associação geradas para candidatos muito acima de média no período pré-covid e pós-covid

(a) Nuvem de palavras para o período pré-covid

(b) Nuvem de palavras para o período pós-covid

Fig. 11: Regras de associação geradas para candidatos muito abaixo de média no período pré-covid e pós-covid

abaixo da média. Assim como na primeira análise, há uma pequena diferença entre os dois perfis: o fato do candidato ser estudante da rede pública de ensino é um fator determinante para o desempenho ruim no período de pandemia, o que não ocorre no período que o antecede.

Uma possível explicação para essa diferença é o fato de que a rede pública de ensino não tinha estrutura para lidar com o cenário de pandemia, prejudicando o bom andamento do ano letivo e comprometendo a qualidade de ensino. Outro fato é que muitos estudantes não tem computador em casa, o que dificulta, quando aplicável, a acompanhar aulas *online*.

VII. Conclusão

Com os resultados obtidos nos experimentos, foi possível identificar fatores socioeconômicos que impactam diretamente - e de forma persistente com o passar dos anos - o desempenho dos estudantes que realizam o ENEM. O processo de *KDD* permitiu que este trabalho pudesse ser desenvolvido em etapas bem definidas e claras, desde a seleção dos dados até a análise dos resultados.

Dado o alto volume de dados públicos disponibilizados sobre o ENEM, tornou-se necessário aplicar algumas técnicas de mineração de dados para ser possível extrair informações valiosas que seriam quase impossíveis de identificar de forma manual. Com o uso das ferramentas corretas, como *Pandas* e bibliotecas do *Python*, foi possível manipular e analisar os dados de forma simples e rápida.

Durante a análise dos resultados, observou-se que características como renda familiar, escolaridade dos pais, acesso à recursos digitais, entre outros, impactam o desempenho do estudante. Estudantes ao qual pelo menos um dos pais têm curso superior, por exemplo, tendem a alcançar notas muito acima da média, e o mesmo vale quando o mesmo dispõe de pelo menos um computador em casa, o que pode estar atribuído à facilidade de acesso a mais conteúdos educativos e informativos quando se tem acesso à ferramenta. Por outro lado, estudantes que não têm computador em casa geralmente alcançam notas muito abaixo da média. A renda familiar também mostrou ser um fator determinante para qualidade da performance na prova, pois inscritos com renda familiar muito baixa tendem a alcançar notas muito abaixo da média, enquanto que outros com renda familiar superior acabam frequentemente alcançando notas muito boas.

Se comparar os resultados deste trabalho com os resultados dos trabalhos anteriores desenvolvidos por outros autores e apresentados na **Seção III**, nota-se uma semelhança dos fatores que influenciam desempenho que foram identificados, podendo ser entendido como uma regra para o formato atual do exame, permitindo formular algumas hipóteses e possíveis soluções com políticas públicas, como facilitar o acesso a computadores para estudantes, melhora a qualidade de ensino em instituições públicas para ser possível competir com estudantes de instituições privadas ou até ações afirmativas para que todos os brasileiros possam ter a mesma oportunidade de fazer um curso superior e melhorar seu perfil financeiro e familiar.

É possível também, a partir dos resultados obtidos, refletir sobre a influência dos pais no desempenho dos estudantes. Ao perceber que os candidatos que tiveram nota muito acima da média geralmente vem acompanhado de pais que têm pelo menos uma graduação completa, é possível afirmar que o acesso à educação permite que se possa construir uma base familiar, no contexto de educação, mais predisposta a ter melhores resultados no ENEM. O fato de escolas privadas estarem associadas a excelentes resultados no ENEM também dão força à afirmação de que as escolas públicas carecem de investimentos e melhorias na qualidade de ensino para este tipo de exame

Utilizando-se da mesma base de dados, foi possível também analisar os fatores determinantes de desempenho no período pré-covid e pós-covid, que apesar de não apresentarem uma quantidade significativa de novas regras, foi capaz de encontrar resultados interessantes, como o fato de que a relevância da escolaridade dos pais é maior no período de pandemia nos casos de desempenho muito acima da média, assim como identifica que estudantes da rede pública de ensino foram os mais prejudicados durante o período de pandemia, sendo uma regra forte para candidatos com desempenho muito abaixo da média.

O poder público, em posse destes resultados, deve dar uma atenção especial a dois principais pontos: melhorar a qualidade de ensino nas instituições públicas e facilitar o acesso à internet de qualidade para os brasileiros que ainda não o tem. A melhoria na qualidade de ensino promove uma melhor base para as futuras gerações e o acesso à internet facilita a democratização do acesso à informação. Com isso, espera-se que candidatos de escolas públicas possam compor em maior proporção o grupo de estudantes que tiveram desempenho muito acima da média.

Referência bibliográfica

[1] ADEODATO, P.. *Data Mining* Solution for Assessing Brazilian Secondary School Quality Based on ENEM and Census Data. CONTECSI USP - International Conference on Information Systems and Technology Management - ISSN 2448-1041, Brasil, jun. 2016. Disponível em: http://www.contecsi.tecsi.org/index.php/contecsi/13CONTECSI/paper/view/3818/2521. Acesso em: 29 de Maio de 2023.

[2] AGRAWAL, R. and Srikant, R. (1994). Fast Algorithms for Mining Association Rules in Large Databases. In: Bocca, J. B., Jarke, M. e Zaniolo, C. (Ed.) Proceedings of the 20th International Conference on Very Large Data Bases, p. 487-499, Santiago, Chile.

[3] AGRAWAL, R., Imielinski, T. and Swami, A. (1993). Mining Association Rules between Sets of Items in Large Databases. Proceedings of the ACM SIGMOD International Conference on Management of Data, p. 207-216, Washington D. C.

[4] ANDRADE, Cibele Yahn de. Acesso ao ensino superior no Brasil: equidade e desigualdade social. Revista Ensino Superior Unicamp 6ª edição, pp. 18-27, jun. 2012. Disponível em https://www.revistaensinosuperior.gr.unicamp.br/edicoes/ed06_julho2012/Cibele_Yahn.pdf. Acesso em: 29 de Maio de 2023.

[5] ARAÚJO, Douglas Luís de. Determinantes do desempenho no ENEM dos concluintes do ensino médio no município de Viçosa – MG. 2019. 55 f. Dissertação/(Mestrado em Administração) - Universidade Federal de Viçosa, Rio Paranaíba. 2019.

[6] BAKER, R. S. J., ISOTANI, S., CARVALHO, A. de: Mineração de dados educacionais: Oportunidades para o Brasil. Revista Brasileira de Informática na Educação, 2011, v. 12, n. 2, p. 3 – 13.

[7] BAKER, R.. *Data Mining* for Education. In McGaw, B., Peterson, P., Baker, E. (Eds.) International Encyclopedia of Education (3rd edition), 2010, Elsevier, Oxford, UK.

[8] BARCELLOS, A. & Isotani, S & Damasceno, C. (2020). Mineração de Dados Abertos - ENEM 2018. Anais dos Trabalhos de Conclusão de Curso. Pós-Graduação em Computação Aplicada à Educação Instituto de Ciências Matemáticas e de Computação. Universidade de São Paulo.

[9] Chi, Y., Wang, H., Yu, P. S. e Muntz, R. R. (2004). Moment: Maintaining closed frequent itemsets over a stream sliding window. In: Proceedings of the IEEE International Conference on Data Mining, p. 59–66, Brighton, UK.

[10] COSTA, Evandro *et al.* Mineração de Dados Educacionais: Conceitos, Técnicas, Ferramentas e Aplicações. Jornada de Atualização em Informática na Educação, [S.l.], p. 1-29, fev. 2013. ISSN 23167734. Disponível em: http://ojs.sector3.com.br/index.php/pie/article/view/2341. Acesso em: 29 de Maio de 2023.

[11] Divulgados números dos inscritos no ENEM 2022 por UF (2022). INEP - Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Ministério da Educação, 2022. Disponível em https://www.gov.br/inep/pt-br/assuntos/noticias/enem/divulgados-numeros-dos-inscritos-no-enem-2022-por-uf. Acesso em: 29 de Maio de 2023.

[12] Enem - microdados (2023b). INEP - Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Ministério da Educação, 2022. Disponível em <https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem>. Acesso em: 06 de Junho de 2023.

[13] Exame Nacional do Ensino Médio (Enem) (2023a). INEP - Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Ministério da Educação, 2023. Disponível em https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enem. Acesso em: 29 de Maio de 2023.

[14] FACELI, Katti et. al. Inteligência artificial: uma abordagem de aprendizado de máquina. 2ª edição. Rio de Janeiro: LTC, 2021.

[15] FAYYAD, U., Piatetski-Shapiro, G. e Smyth, P. (1996). The *kdd* process for extracting useful knowledge from volumes of data. Communications of the ACM, p. 27–34.

[16] *Github*, Inc. *Github* Documentation (2023). Página de documentação. Disponível em: https://docs.*github*.com/pt. Acesso em: 29 de Maio de 2023.

[17] HAN, J., KAMBER, M. e PEI, J. (2011). *Data Mining*: Concepts and Techniques, 3ª edição. Morgan Kaufmann.

[18] NOVAES, J. (2021), ENEM 2021 tem menor número de inscritos em 16 anos; o que explica esse fato?. FDR, 04 de ago. de 2021. Disponível em: <https://fdr.com.br/2021/08/04/enem-2021-tem-menor-numero-de-inscritos-em-16-anos-o-que-explica-esse-fato/>. Acesso em: 6 de Junho de 2023.

[19] PyPi. PyPi - apyori 1.1.2. (2019) Disponível em: https://pypi.org/project/apyori/. Acesso em: 29 de Maio de 2023.

[20] Python Software Foundation. Python Language Site: Documentation (2023). Página de documentação. Disponível em: https://www.python.org/doc/. Acesso em: 29 de Maio de 2023.

[21] SILVA, V. A. A. da MORENO, L. L. O. GONÇALVES, L. B. SOARES, S. S. R. F. SOUZA JÚNIOR, R. R.. Identificação de Desigualdades Sociais a partir do desempenho dos alunos do Ensino Médio no ENEM 2019 utilizando Mineração de Dados. In: SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, 31. , 2020, Online. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2020 . p. 72-81. DOI: https://doi.org/10.5753/cbie.sbie.2020.72.

[22] The pandas development team. pandas-dev/pandas: Pandas 1.4.3 Zenodo , 23 Jun. 2022. Disponível em: <https://doi.org/10.5281/zenodo.6702671>. Acesso em: 06 de Junho de 2023.

VIII. APÊNDICES

APENDICE A			
Nome da coluna	**Descrição da Coluna**	**Itens**	**Descrição do item**
TP_FAIXA_ETARIA	Faixa etária	1	Menor de 17 anos
		2	17 anos
		3	18 anos
		4	19 anos
		5	20 anos
		6	21 anos
		7	22 anos
		8	23 anos
		9	24 anos
		10	25 anos
		11	Entre 26 e 30 anos
		12	Entre 31 e 35 anos
		13	Entre 36 e 40 anos
		14	Entre 41 e 45 anos
		15	Entre 46 e 50 anos
		16	Entre 51 e 55 anos
		17	Entre 56 e 60 anos
		18	Entre 61 e 65 anos
		19	Entre 66 e 70 anos
		20	Maior de 70 anos
TP_SEXO	Sexo	M	Masculino
		F	Feminino
TP_ESTADO_CIVIL	Estado Civil	0	Não informado
		1	Solteiro(a)
		2	Casado(a)/Mora com companheiro(a)
		3	Divorciado(a)/Desquitado(a)/Separado(a)
		4	Viúvo(a)
TP_COR_RACA	Cor/raça	0	Não declarado
		1	Branca
		2	Preta
		3	Parda
		4	Amarela
		5	Indígena
TP_ST_CONCLUSAO	Situação de conclusão do Ensino Médio	1	Já concluí o Ensino Médio
		2	Estou cursando e concluirei o Ensino Médio em 2020
		3	Estou cursando e concluirei o Ensino Médio após 2020
		4	Não concluí e não estou cursando o Ensino Médio
TP_ESCOLA	Tipo de escola do Ensino Médio	1	Não Respondeu
		2	Pública
		3	Privada
		4	Exterior
TP_DEPENDENCIA_ADM_ESC	Dependência administrativa	1	Federal
		2	Estadual
		3	Municipal
		4	Privada
SG_UF_PROVA	Sigla da Unidade da Federação da aplicação da prova		
NU_NOTA_CN	Nota da prova de Ciências da Natureza		
NU_NOTA_CH	Nota da prova de Ciências Humanas		
NU_NOTA_LC	Nota da prova de Linguagens e Códigos		
NU_NOTA_MT	Nota da prova de Matemática		
NU_NOTA_REDACAO	Nota da prova de redação		
Q001	Até que série seu pai, ou o homem responsável por você, estudou?	A	Nunca estudou.
		B	Não completou a 4ª série/5º ano do Ensino Fundamental.
		C	Completou a 4ª série/5º ano, mas não completou a 8ª série/9º ano do Ensino Fundamental.
		D	Completou a 8ª série/9º ano do Ensino Fundamental, mas não completou o Ensino Médio.
		E	Completou o Ensino Médio, mas não completou a Faculdade.
		F	Completou a Faculdade, mas não completou a Pós-graduação.
		G	Completou a Pós-graduação.
		H	Não sei.
Q002	Até que série sua mãe, ou o mulher responsável por você, estudou?	A	Nunca estudou.
		B	Não completou a 4ª série/5º ano do Ensino Fundamental.
		C	Completou a 4ª série/5º ano, mas não completou a 8ª série/9º ano do Ensino Fundamental.
		D	Completou a 8ª série/9º ano do Ensino Fundamental, mas não completou o Ensino Médio.
		E	Completou o Ensino Médio, mas não completou a Faculdade.
		F	Completou a Faculdade, mas não completou a Pós-graduação.
		G	Completou a Pós-graduação.
		H	Não sei.
Q006	Qual é a renda mensal de sua família?	A	Nenhuma Renda
		B	Até R$ 1.045,00
		C	De R$ 1.045,01 até R$ 1.567,50
		D	De R$ 1.567,51 até R$ 2.090,00
		E	De R$ 2.090,01 até R$ 2.612,50
		F	De R$ 2.612,51 até R$ 3.135,00
		G	De R$ 3.135,01 até R$ 4.180,00
		H	De R$ 4.180,01 até R$ 5.225,00
		I	De R$ 5.225,01 até R$ 6.270,00
		J	De R$ 6.270,01 até R$ 7.315,00
		K	De R$ 7.315,01 até R$ 8.360,00
		L	De R$ 8.360,01 até R$ 9.405,00
		M	De R$ 9.405,01 até R$ 10.450,00
		N	De R$ 10.450,01 até R$ 12.540,00
		O	De R$ 12.540,01 até R$ 15.675,00
		P	De R$ 15.675,01 até R$ 20.900,00
		Q	Acima de R$ 20.900,00
Q022	Na sua residência tem telefone celular?	A	Não.
		B	Sim, um.
		C	Sim, dois.
		D	Sim, três.
		E	Sim, quatro ou mais.
Q024	Na sua residência tem computador?	A	Não.
		B	Sim, um.
		C	Sim, dois.
		D	Sim, três.
		E	Sim, quatro ou mais.
Q025	Na sua residência tem acesso à Internet?	A	Não.
		B	Sim