UNIVERSIDADE FEDERAL
DE MINAS GERAIS

CARLOS HENRIQUE BRITO MALTA LEÃO
VINÍCIUS ALVES DE FARIA RESENDE

MINERAÇÃO DE DADOS

ENEM 2022: O Impacto das Características Socioeconômicas no Desempenho dos Candidatos

Belo Horizonte
2023

CARLOS HENRIQUE BRITO MALTA LEÃO
VINÍCIUS ALVES DE FARIA RESENDE

ENEM 2022: O Impacto das Características Socioeconômicas no Desempenho dos Candidatos

Trabalho Prático 1 apresentado como atividade avaliativa da disciplina de Mineração de Dados pela Universidade Federal de Minas Gerais — UFMG

Professor: Wagner Meira, Jr

Belo Horizonte
2023

Endereço Magnético Google Colab: **TP1_Mineração_de_Dados.ipynb**

1. Entendimento do Negócio

Ao que se refere ao entendimento do negócio, boa parte desta investida foi desenvolvida na primeira parte do trabalho a qual tratava sobre a proposta do projeto. Nesta etapa foram discutidos os objetivos do negócio, seu contexto e relevância e até os critérios de sucesso.

Considerando a natureza única do projeto, a ideia de "negócio" pode ser um pouco abstrata, uma vez que não estar-se-á tratando de dados empresariais. A natureza dos dados considerados, Microdados do ENEM 2022, é de cunho demográfico, com riqueza no contexto socioeconômico dos participantes. Nessa perspectiva, é possível determinar os critérios de sucesso embasados nas qualidades esperadas das políticas públicas relacionadas à educação. No contexto citado acima, destacamos como critérios principais: **melhorar o desempenho dos alunos, promover a equidade na educação, personalização da educação e avaliação das escolas**.

Considerando os critérios supracitados, o projeto visa alguns objetivos perante ao modelo do negócio, esses objetivos tem como função retratar o cenário atual em virtude da perspectiva ideal. Dessa forma, um projeto de mineração de dados tem o poder de identificar relações indiretas e falhas implícitas que porventura estejam presentes no sistema educacional, como também, existe a possibilidade de reforçar ideias que relacionam critérios socioeconômicos com desempenho estudantil, servindo como argumento para políticas como, por exemplo, o sistema de cotas. De forma geral, as faces da investida em direção ao entendimento dos dados foram sumarizadas em: **identificação de padrões de desempenho, equidade na educação, customização de educação e avaliação da eficácia escolar**.

2. Entendimento dos dados

2.1. Descrever os dados

De forma geral, a descrição inicial dos dados já foi dada pelo próprio INEP, isso se dá pelo fato de que a base de dados apresenta muita qualidade, e acompanha um dicionário que descreve em detalhes cada um dos dados presentes. Este dicionário é anexado juntamente com a página e fizemos-o disponível neste endereço magnético: Dicionário_Microdados_Enem_2022 .

Observando o dicionário, é possível ver que os dados são bem divididos entre alfanuméricos e numéricos. Além disso, o prefixo do nome da variável é utilizado para realizar uma "tipagem" dos dados, sendo o prefixo *NU* utilizado para números, *TP* utilizado para valores categóricos (independe da classificação de alfanuméricos e numéricos), *IN* para valores booleanos, *CO* para códigos, *NO* para nomes e *SG* para siglas. Além dos citados, existem as perguntas socioeconômicas que seguem sempre o mesmo padrão de variável categórica com valor alfanumérico.

O dicionário também apresenta outras informações primordiais sobre os dados no contexto da mineração de padrões frequentes. Sendo eles o comprimento(número de caracteres) dos dados e também os

valores possíveis, bem como o significado de cada dado e o de cada um dos valores possíveis. Essas informações se mostram como essenciais no contexto da mineração de padrões frequentes, onde o objetivo é identificar valores para as variáveis que são observados em conjunto, e, a partir daí, extrair informações valiosas que terão de ser mapeadas com o auxílio do dicionário.

2.2. Explorar os dados

Durante a etapa de exploração dos dados, o objetivo principal foi o de ter um entendimento mais valioso sobre a base de dados do ENEM 2022 em si, e não apenas o formato desta base. Entender de forma geral os aspectos dos dados contidos na base se mostra como uma atividade primordial.

A primeira investida neste sentido foi no estudo que relaciona a presença ou eliminação dos participantes, isso pelo fato de o desempenho desses alunos não poderia ser quantizado conforme a nota média. Contudo, é de interesse do projeto verificar a relação dos fatores socioeconômicos com a presença ou não na prova. Foi observado que um número considerável de candidatos não esteve presente em pelo menos uma das provas, mais especificamente, cerca de 32,54% dos candidatos faltaram ou foram eliminados de pelo menos uma das provas.

Posteriormente foi de interesse do projeto verificar algumas características socioeconômicas da base de forma isolada. Essa investida se mostrou muito proveitosa por demonstrar critérios muitas vezes não esperados referente aos participantes. Os dados que foram mantidos no documento *Jupyter* foram escolhidos arbitrariamente conforme a percepção dos integrantes do projeto. São válidos de serem mencionados os dados referentes à **autodeclaração de cor/raça**, onde vimos que 41% dos participantes se auto declararam como brancos, 42,8% se auto declaram como pardos e apenas 11,75% como pretos. Foi feita também uma avaliação conforme o **sexo biológico** dos participantes, onde vimos que 61% dos participantes são do sexo feminino, tamanha disparidade foi motivo de surpresa para os integrantes do projeto. No que se refere ao **tipo de escola** notamos que existe uma grande maioria não respondida mas que, dentre os respondidos, o número de participantes de escola pública supera em mais de 5 vezes o número de escolas privadas. Em relação à **quantidade de carros nas residências**, vimos que mais de 51% dos participantes sequer possuem um carro, e que apenas cerca de 10% possuem mais de um carro. Por último, consideramos um critério de interesse a avaliação da **disponibilidade de internet dos participantes**, onde 90,6% dos participantes declaram ter acesso à internet. De forma geral, todos os dados mencionados neste parágrafo são de grande valia se comparados com o censo mais recente do Instituto Brasileiro de Geografia e Estatística (IBGE), uma vez que pode-se observar a disparidade entre amostras da sociedade brasileira como um todo e os participantes do ENEM.

2.3. Verificar a qualidade dos dados

Por fim, foi de interesse do projeto verificar o percentual de dados nulos, que, até o momento da prática automatizada, não eram bem conhecidos do ponto de vista individual de cada variável. Neste momento observamos que existia um alto número de valores nulos, contudo, boa parte deles eram esperados dado o

número de desistentes da prova. Porém, não era esperado que algumas variáveis como o município da escola onde estudou o participante, teria um percentual de valores nulos proporcionalmente superior, chegando até a 70 pontos percentuais.

Dado o contexto supracitado, foi decidido que estes valores que apresentavam um alto percentual de valores nulos seriam desconsiderados para o projeto, de forma que poderiam se mostrar enviesados dado o volume de dados total.

Considerando os demais critérios de qualidade de dados, pudemos verificar que todos se adequam ao que é descrito no Dicionário_Microdados_Enem_2022, até mesmo por ser uma base tão crucial para o país que é submetida por diversas revisões. Por conta disso, não foram necessárias nenhuma das correções de erros descritos no documento de Guia de Usuário do CRISP.

3. Preparação dos dados

3.1. Selecionar dados

Como primeiro passo da preparação dos dados, é necessário fazer a seleção das variáveis que se mostram como relevantes para o estudo, de forma a reduzir a dimensionalidade considerada para o projeto. Esse processo é necessário para evitar distrações dos dados realmente significativos ao projeto, além de auxiliar na viabilização de algoritmos que são grandemente prejudicados pela dimensionalidade da base estudada.

O critério de seleção das variáveis com maior significado para o projeto foi embasado na percepção dos integrantes do projeto. Essa percepção se deu uma vez que os objetivos desejados com o projeto estavam bem definidos na entrega da primeira parte do trabalho.

Por fim, dado o critério mencionado no último parágrafo, optamos pela seleção de algumas colunas que configurar as variáveis da base de dados, sendo elas: **faixa etária, sexo biológico, estado civil, cor/raça, tipo de escola, situação de presença em cada uma das provas, nota de cada uma das provas, situação da redação, nota da redação e questões do questionário socioeconômico que foram consideradas como mais relevantes.**

3.2. Limpar dados

A etapa de limpeza de dados se caracterizou mais por uma análise dos integrantes do grupo sobre o significado dos dados. Porém, a priori, não foi identificada a necessidade de uma ação sobre os dados nesta etapa. Contudo, foi levantada a necessidade de um critério mais cuidadoso com os dados do tipo categórico, uma vez que o seu significado não era homogêneo entre todas as variáveis categóricas, mudando o espaço amostral para cada uma.

Dito isto, durante o processo de análise vimos que seria crucial manter sempre uma referência à variável juntamente com seu valor no momento da identificação dos padrões. Tal descoberta motivou o processo descrito na seção **3.4 (Formatar dados)**. Mantendo a referência da variável atrelada ao valor de cada

dado, conseguiríamos fazer a correta avaliação de resultados. Concluindo assim o processo de limpeza dos dados.

3.3. Construir dados

Nesta etapa mostrou-se necessário o desenvolvimento de novas variáveis que tenham a função de sumarizar várias, como a nota média. Ademais, por conta do contexto de mineração de padrões frequentes do projeto, se fez necessária a geração de valores categóricos para variáveis numéricas, como realizado para a nota média que classificou os participantes em intervalos.

Atributos derivados

Confeccionamos um atributo derivado, sendo este a nota média. Este atributo se mostrou conveniente para a mineração dos padrões frequentes uma vez que sumariza todas as variáveis relacionadas à nota. Uma vez que o objetivo do projeto é fazer a relação entre quesitos socioeconômicos e o desempenho no exame em si, a utilização de uma nota média é conveniente pelo fato de simplificar o conceito de "desempenho" em um único valor.

Esta variável foi criada por meio da média simples de todas as notas, somando seus valores e dividindo pelo total de avaliações distintas (cinco). Esta nota foi acrescentada como uma nova coluna na representação dos dados (*Data Frame*). Complementarmente, tivemos a curiosidade de realizar uma visualização deste dado como um gráfico visual. O resultado da visualização foi interessante porém esperado, ao realizar o *plot* dos dados arredondados em gráfico de barras, podemos ver o formato claro de distribuição gaussiana/normal. Essa visualização nos possibilitou ter a certeza da hipótese sobre a distribuição dos dados, motivando o *insight* sobre a utilização de quartis para a geração de registros.

Registros gerados

No que tange à confecção de registros, se mostrou necessária a criação de uma representação categórica para o valor numérico da nota média. Considerando que, por conta do que foi supracitado, já conhecíamos a distribuição dos dados, foi intuitivo pensar na utilização de quartis para segmentar as notas em valores categóricos.

A divisão foi feita seguindo o modelo clássico, dividindo as observações em intervalos onde a população representa 25% da base. Ao fazer a segmentação do *top* 25% dos alunos pela perspectiva de desempenho no exame, notamos que o "valor de corte" era de aproximadamente 602. Após uma pesquisa rápida nas notas de corte para os cursos da UFMG, vimos que este valor não era suficiente para ingressar na grande maioria dos cursos da universidade, o que nos confirmou a hipótese de que a grande maioria dos participantes do ENEM sequer têm a chance de ingressar em uma universidade, mesmo considerando os 25% com melhor performance.

Considerando o fenômeno observado, decidimos por adicionar uma segmentação além dos quartis, a qual consideramos como "alunos com poder de barganha", isto é, alunos que, dado seu desempenho, têm a possibilidade de escolher dentre a maioria dos cursos, qual irá cursar e até mesmo ter o poder de escolha entre

as universidades. Sendo excluídos apenas os cursos mais concorridos como Medicina, Ciência da Computação, Engenharia Aeroespacial entre outros. Após uma exploração dos dados, ficou decidido que esse grupo seleto de alunos representaria o *top* 1% da base de dados, sendo separados por um "valor de corte" igual a 750.

Nesse contexto, a nova variável "Tipo Nota Média" foi criada, segmentando os "alunos com poder de barganha" com o valor A (corte: 750), os alunos no *top 25%* com o valor B (corte: 602), os alunos no *top* 50% com o valor C (corte: 541), os alunos no *top* 75% com o valor D (corte: 485), os alunos com nota válida com o valor E (corte 56) e os alunos faltantes, desistentes e ou com nota igual a 0 com o valor F.

Após o término das ações supracitadas, no contexto da mineração de padrões frequentes, não tinha mais sentido manter os valores referentes às notas em termos numéricos. Portanto a nota de todas as avaliações do ENEM e também a nota média numérica foram removidas da base.

3.4. Formatar dados

Na etapa de formatação dos dados uma investida foi necessária considerando a natureza da modelagem a qual pretende-se utilizar. Para a mineração de padrões frequentes os dados precisam estar bem segmentados e com o significado explícito. Trazendo para o contexto da base, é crucial que, durante a confecção dos conjuntos frequentes, esteja explícito que a resposta "A" à pergunta 3 do questionário seja diferente a uma resposta "A" da pergunta 22, por exemplo. Ou seja, as variáveis categóricas precisam ser "etiquetadas" para que um algoritmo de mineração frequente funcione de maneira apropriada.

Neste contexto, utilizamos uma iteração sobre os dados para alterar todos os valores da base, transformando todos em strings e adicionando um prefixo pré determinado a cada um dos valores. Este prefixo é constituído das iniciais do nome da variável em si, que é adicionado como um prefixo do valor. Para facilitar a visualização, um valor da coluna *TP_COR_RACA* que seja igual à 1, será transformado para a string "TCR:1". Utilizando esta estratégia, conseguimos sanar a problemática e concluir o processo de **Preparação dos Dados**, possibilitando-nos prosseguir com a modelagem.

4. Modelagem

Nesta seção, abordaremos o processo de modelagem, que envolve a aplicação de técnicas de mineração de dados à base de dados do ENEM 2022. O objetivo principal desta fase é extrair informações significativas e identificar padrões relevantes relacionados ao desempenho dos participantes com base em fatores socioeconômicos.

4.1. Técnica de modelagem

Para a modelagem inicial, optamos por utilizar a técnica de Mineração de Conjuntos Frequentes, mais especificamente, a utilização do algoritmo FP-Growth. A escolha por essa técnica se deve à sua capacidade de identificar conjuntos de itens frequentes em grandes conjuntos de transações, o que é fundamental para entender as associações entre os atributos da base de dados. Essa técnica nos permitirá descobrir conjuntos de

atributos frequentes de cada candidato que podem ser posteriormente analisados em busca de insights relevantes.

O algoritmo FP-Growth é uma técnica eficiente para a mineração de conjuntos frequentes, especialmente em grandes conjuntos de dados. Ele se destaca por sua capacidade de gerar uma estrutura de dados compacta chamada de "Árvore de Prefixo" (ou FP-Tree) que permite uma mineração de padrões frequentes rápida e eficaz.

A aplicação do algoritmo à base de dados do ENEM 2022 tem o objetivo de identificar conjuntos frequentes de atributos socioeconômicos e de desempenho dos participantes. Isso nos permitirá responder a perguntas como:

- Quais combinações de características socioeconômicas estão associadas a um desempenho excepcional no ENEM?
- Existem padrões frequentes que indicam a influência de variáveis como cor/raça, tipo de escola, ou acesso à internet no desempenho dos alunos?
- Podemos descobrir grupos específicos de alunos que compartilham características socioeconômicas comuns e têm um desempenho semelhante?

Dessa forma, estamos explorando a estrutura subjacente dos dados do ENEM 2022 para identificar associações e padrões que podem ser valiosos para a compreensão do sistema educacional brasileiro e o impacto das políticas socioeconômicas na educação. Este é um passo crucial na nossa análise de dados, pois nos ajuda a descobrir conexões significativas entre os atributos dos candidatos, fornecendo uma base sólida para futuras análises e tomada de decisões baseadas em dados.

4.2. Construção do modelo

Antes de executar o algoritmo FP-Growth, é fundamental definir os parâmetros iniciais que afetarão a construção do modelo. Nesse sentido, é necessário definir o **Suporte Mínimo**. Esse parâmetro define a frequência mínima com a qual um conjunto de itens deve aparecer na base de dados para ser considerado frequente. O valor escolhido para o **Suporte Mínimo** é crucial, pois afeta a quantidade e a qualidade dos conjuntos frequentes identificados. Nesse caso, escolhemos um valor de 30%, ou seja, para um conjunto ser considerado frequente, ele deve ocorrer em, pelo menos, 30% de todas as transações do banco de dados.

Com as configurações de parâmetros definidas, procedemos à execução do algoritmo no conjunto de dados já previamente preparado. Ao executarmos, encontramos, aproximadamente, 93 conjuntos frequentes para cada tipo de nota média, totalizando 557 conjuntos. Cada um apresenta seu próprio suporte relativo, variando de 0,30 até 1, que representa a frequência do conjunto no banco de dados.

4.3 Avaliação do modelo

A partir dos conjuntos frequentes, foi possível gerar diversas regras de associação para cada tipo de nota. Nesse sentido, foi utilizada a função *association_rules*, da biblioteca *mlxtend.frequent_patterns*. Essa

função foi utilizada para gerar regras de associação com base no critério lift, em que foi aplicado um filtro, em relação a essa métrica de 1.0. Isso ajudou a identificar relações relevantes entre os itens do conjunto de dados. Nesse sentido, foram encontradas, em média, 216 regras de associação para cada tipo de nota média, totalizando 1294 regras.

Nesse contexto, seria extremamente complicado realizar uma análise de todo esse conjunto de regras. Por isso, utilizando de algumas métricas de avaliação, foram aplicados mais filtros, para analisarmos apenas as regras mais relevantes para o contexto trabalhado. Dessa forma, as regras foram filtradas para terem uma confiança acima de 70%, *lift* maior que 1.005, *leverage* maior que 0.005 e *conviction* maior que 1.2.

Por fim, encontramos um total de 187 regras de associação, que estão divididas para cada tipo de nota, em que todas estão filtradas de acordo com as métricas supracitadas. Nesse contexto, encontramos 14 regras para a nota A, 24 para a B, 34 para a nota C, 35 para a D, 47 para a E e 33 para a F. Estas regras serão analisadas e avaliadas na sessão seguinte.

5. Avaliação

Nesta seção, avaliaremos o projeto de mineração de dados em relação aos objetivos estabelecidos no início do projeto. Relembrando, o objetivo principal é a identificação de padrões de desempenho, equidade na educação, customização de educação e avaliação da eficácia escolar, no contexto específico da prova do ENEM 2022 e de seus participantes.

Nessa etapa é crucial conduzir uma análise individualizada para cada tipo de nota (A, B, C, D, E, F), uma vez que cada um deles está associado a um conjunto exclusivo de regras de associação. Essa abordagem permitirá uma avaliação minuciosa e específica do desempenho da mineração de dados, oferecendo interpretações detalhadas sobre os resultados obtidos.

5.1. NOTA A - Nota média maior que 750

Encontramos no banco de dados um total de 23.766 candidatos que obtiveram a nota A, representando cerca de 0,7% do total de participantes. Esses são considerados os participantes que, dado seu desempenho, têm a possibilidade de escolher dentre a grande maioria dos cursos, qual irá cursar e até mesmo ter o poder de escolha entre diversas universidades. Sendo excluídos apenas alguns dos cursos mais concorridos, como Medicina, Ciência da Computação, Engenharia Aeroespacial, entre outros.

Durante a análise, foi possível identificar 14 regras de associação relevantes para esse grupo específico de candidatos. Algumas dessas regras forneceram informações substanciais. No entanto, a regra de índice 248 chamou atenção, uma vez que tem como antecedente o candidato ser autodeclarado branco, e possuir três quartos em sua residência, enquanto o consequente se refere ao candidato ser solteiro e também ter internet em casa . Essa regra apareceu com um suporte do antecedente de 40% e uma confiança de quase 98%.

Outras regras que também se destacaram foram as regras de índice 285 e 305. Ambas apresentam como antecedente a profissão do pai pertencer ao Grupo 4, determinado pelo próprio questionário

socioeconômico como: Professor (de ensino fundamental ou médio, idioma, música, artes etc.), técnico (de enfermagem, contabilidade, eletrônica etc.), policial, militar de baixa patente (soldado, cabo, sargento), corretor de imóveis, supervisor, gerente, mestre de obras, pastor, microempresário (proprietário de empresa com menos de 10 empregados), pequeno comerciante, pequeno proprietário de terras, trabalhador autônomo ou por conta própria. Essas regras apareceram com um suporte do antecedente de 42%.

5.2. NOTA B - Nota média entre 602 e 750

Encontramos no banco de dados um total de 566.996 candidatos que obtiveram a nota B, representando cerca de 16,3% do total de participantes.

Durante a análise, foi possível encontrar 24 regras de associação relevantes para esse grupo de candidatos. Assim como nas regras para o grupo A, algumas regras forneceram informações muito substanciais e simples. Enquanto isso, algumas regras mais interessantes, de certa forma, são muito semelhantes às regras encontradas e descritas na seção anterior, representando uma possível semelhança entre os dados dos participantes de nota A e B. Nesse sentido, ao observar regras que apresentam a cor branca como antecedente, identificamos que 61% dos candidatos que obtiveram a nota B são autodeclarados brancos.

Além disso, o vínculo empregatício do pai pertencer ao Grupo 4, detalhado na seção anterior, aparece com um suporte de 36%. Ademais, também encontramos que o emprego da mãe do participante também pertence ao Grupo 4 em 42% dos casos observados. Por fim, também verificamos que 45% dos candidatos que obtiveram a nota B apresentam três quartos em casa e 34% apresentam dois banheiros.

5.3. NOTA C - Nota média entre 541 e 602

Encontramos no banco de dados um total de 582.062 candidatos que obtiveram a nota C, representando cerca de 16,7% do total de participantes. Durante a análise, foi possível encontrar 34 regras de associação relevantes para esse grupo de candidatos. Nesse sentido, após uma avaliação encontramos algumas regras mais interessantes para o objetivo do projeto.

Entre essas, está definido que 34,6% dos pais e 39,7% das mães desses candidatos completaram o Ensino Médio, mas não completaram a faculdade. Além disso, é possível verificar uma queda na frequência de candidatos brancos que obtiveram a nota C. Nesse contexto, cerca de 47,6% desses participantes são autodeclarados brancos, uma queda significativa quando comparado com os 61% encontrados para a nota B.

5.4. NOTA D - Nota média entre 485 e 541

Encontramos no banco de dados um total de 585.817 candidatos que obtiveram a nota D, representando cerca de 16,8% do total de participantes. Durante a análise, foi possível encontrar 35 regras de associação relevantes para esse grupo de candidatos. Assim como todos os outros grupos, muitas regras não agregaram muitas informações para a análise. Porém, também encontramos algumas regras interessantes, dados os objetivos do projeto.

Uma regra interessante que encontramos foi o antecedente de que o candidato é autodeclarado pardo, e como consequente, possui apenas um banheiro em sua residência. Nesse sentido, essa regra apresenta uma confiança de 75%. A partir disso, verificamos também que, desse grupo de candidatos, 45% são pardos e 69% possuem somente um banheiro em casa. Uma regra semelhante, com o mesmo consequente, apresenta o antecedente, com 44% de frequência, de que a mãe do participante pertence ao Grupo 2: Diarista, empregado doméstico, cuidador de idosos, babá, cozinheiro (em casas particulares), motorista particular, jardineiro, faxineiro de empresas e prédios, vigilante, porteiro, carteiro, office-boy, vendedor, caixa, atendente de loja, auxiliar administrativo, recepcionista, servente de pedreiro, repositor de mercadoria.

Nesse conjunto de participantes, verificamos novamente uma diminuição da frequência de brancos, em que apenas 37,5% desses candidatos apresentam essa autodeclaração. Além disso, 54% não apresentam nenhum carro em sua residência e 38,6% das mães desses candidatos concluíram o Ensino Médio mas não completaram a faculdade.

5.5. NOTA E - Nota média entre 56 e 485

Encontramos um total de 586.160 candidatos que obtiveram a nota E, representando cerca de 16,9% de todos os participantes do banco de dados. Após a aplicação dos filtros, encontramos 47 regras de associação relevantes para esse grupo. Apesar de muitas apresentarem informações muito simplórias, outras apresentaram informações extremamente engrandecedoras para o contexto do presente trabalho.

Nesse contexto, temos a regra de índice 207, com 36% de suporte relativo e 83,5% de confiança. Esta apresenta como antecedente, com 43% de frequência, a renda mensal total da família do participante chegando até R$1.212,00, que representa o valor do salário mínimo em 2022. Além disso, como consequente, com 65,5% de frequência, que a família do candidato não possua nenhum carro em casa.

Além disso, observando outras regras, identificamos também que 52,6% desses candidatos são autodeclarados pardos. Além disso, também encontramos que 43,2% das mães desses participantes pertencem ao Grupo 2 de trabalhadores. Por fim, 76,5% desses candidatos possuem apenas um banheiro em suas residências.

5.6. NOTA F - Nota média entre 0 e 56

Para terminar, existem 1.131.304 candidatos que obtiveram a nota F, representando, impressionantes, 32,5% de todos os participantes da prova. A grande maioria desses participantes, apresentam a nota 0, por conta de não terem realizado a prova. Após a filtragem das regras de associação, encontramos 33 relevantes para esse grupo. Assim como nas outras análises, diversas são informações simples, que não agregam informações para o contexto desse projeto, dessa forma, explicitamos aqui somente as regras interessantes para o presente trabalho.

No grupo de nota F, encontramos informações muito semelhantes aos dados do grupo E. Identificando esses padrões, temos a mãe de 45,6% dos participantes atuando profissionalmente no Grupo 2 de trabalhadores.

Além disso, encontramos que 34,6% das famílias desses participantes possuem uma renda mensal familiar de até um salário mínimo no ano de 2022. Além disso, encontramos também que 35,5% desses indivíduos são autodeclarados brancos, um número muito menor quando comparado com os 61% encontrados para o grupo de participantes brancos de nota B e até mesmo para 41% totais, incluindo todas as notas juntas. Por outro lado, encontramos que 46% desses são autodeclarados como pardos.

6. Conclusões e perspectivas

Nesta seção, apresentamos as principais conclusões derivadas da análise dos microdados do ENEM 2022 neste projeto de mineração de dados. Ao longo deste trabalho, conduzimos uma análise completa dos dados, abrangendo desde a compreensão inicial até a modelagem e avaliação, com o objetivo de identificar tendências e padrões de desempenho, equidade na educação, customização de ensino e avaliação da eficácia escolar.

Uma das descobertas mais significativas foi a identificação de padrões de desempenho dos candidatos. Através da análise de regras de associação, foi possível estabelecer que diferentes grupos demográficos apresentam médias de notas distintas. Candidatos autodeclarados brancos e com acesso a recursos econômicos tendem a obter notas mais altas, enquanto outros grupos, como autodeclarados pardos e com menor renda, têm notas mais baixas. Essas associações ressaltam a presença de desigualdades educacionais que precisam ser abordadas. Isso sublinha a necessidade urgente de políticas educacionais voltadas para a promoção da igualdade de oportunidades.

Em síntese, este projeto de mineração de dados proporcionou uma análise detalhada dos microdados do ENEM 2022, oferecendo conhecimentos valiosos sobre desempenho, equidade, customização e eficácia escolar. Essas conclusões podem servir como ponto de partida para políticas educacionais mais eficazes e personalizadas, visando à redução das desigualdades e ao aprimoramento da educação no Brasil.

Este trabalho não representa o fim, mas sim o início de uma jornada contínua na compreensão e no aprimoramento da educação. Os resultados obtidos aqui devem ser usados como base para análises mais aprofundadas, estudos complementares e ações práticas que visem a um sistema educacional mais igualitário e de alta qualidade para todos os estudantes brasileiros. A mineração de dados continuará a desempenhar um papel crucial nesse processo.

Referências Bibliográficas

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA. **Microdados do Enem 2022.** Brasília: Inep, 2023. Disponível em: <https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem>. Acesso em: 19 abr. 2023.

UFMG. **Sisu UFMG 2022.** Minas Gerais: UFMG, 2022. Disponível em: <https://www.ufmg.br/sisu/repositorio/?edicao=sisu-ufmg-2022&repositorio_tipo=nota-corte>. Acesso em: 1 out. 2023.