UNIVERSIDADE FEDERAL DO CEARÁ

CENTRO DE CIÊNCIAS

DEPARTAMENTO DE FÍSICA

CURSO DE GRADUAÇÃO EM FÍSICA

LUÍS EDUARDO BINO SOUZA

ANÁLISE UTILIZANDO TEORIA DE REDES COMPLEXAS APLICADA AOS MICRODADOS DO ENEM

FORTALEZA

2024

LUÍS EDUARDO BINO SOUZA

ANÁLISE UTILIZANDO TEORIA DE REDES COMPLEXAS APLICADA AOS MICRODADOS DO ENEM

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Física do Centro de Ciências da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de licenciado em Física.

Orientador: Prof.Dr. Carlos Lenz Cesar

FORTALEZA

2024

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Sistema de Bibliotecas
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

S239a Souza, Luis Eduardo Bino.
Análise utilizando teoria de redes complexas aplicada aos microdados do ENEM / Luis Eduardo Bino Souza. – 2023.
84 f. : il. color.

Trabalho de Conclusão de Curso (graduação) – Universidade Federal do Ceará, Centro de Ciências, Curso de Física, Fortaleza, 2023.
Orientação: Prof. Dr. Carlos Lenz Cesar.

1. Minimum Spanning Tree. 2. Sistemas Complexos. 3. Microdados do ENEM. I. Título.
CDD 530

LUÍS EDUARDO BINO SOUZA

ANÁLISE UTILIZANDO TEORIA DE REDES COMPLEXAS APLICADA AOS MICRODADOS DO ENEM

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Física do Centro de Ciências da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de licenciado em Física.

Aprovada em: 24 de setembro de 2024

BANCA EXAMINADORA

Prof.Dr. Carlos Lenz Cesar (Orientador)
Universidade Federal do Ceará (UFC)

Prof. Dr. Francisco Nepomuceno Filho
Universidade Federal do Ceará (UFC)

Prof. Dr. Juvêncio Santos Nobre
Universidade Federal do Ceará (UFC)

À minha avó, minha mãe, à meu tio Francisco das Chagas, à minha irmã e irmão. Sem vocês nada disso teria sido possível, eterna gratidão.

AGRADECIMENTOS

Agradeço às mulheres da minha vida minha avó, minha mãe e minha irmã, por todo apoio, puxões de orelha e por acreditarem no meu potencial, o que possibilitou chegar até a universidade e concluir meus objetivos iniciais, amos vocês.

Ao meu Tio Francisco das Chagas, por todo apoio, incentivo e valorização da minha educação indenpendente de qualquer coisa, ao meu irmão por todo apoio e conversas.

Agradeço a uma pessoa mais que especia, Enya Ludmilla, por todo apoio e paciência que teve comigo, em especial nessa reta final.

Aos meus amigos de graduação, Davi, Lucas Sievers, Paulo, Lucas, José, Sabrina, Matheus, Pedro,Ricardo e Márcio por todo apoio, dicas e lazer, com certeza por conta de vocês a graduação foi mais leve, além de todos os membros do entropia.

Ao meu orientador Prof. Dr. Carlos Lenz Cesar, por me orientar durante 2 anos e por ter sido um excelente orientador, me mostrando e ensinando uma área de pesquisa em que eu almejava me inserir, além de disponibilizar um laboratório e computadores que foram essenciais para esse trabalho.

Ao Ludwing, por todas as dicas de programação e por disponibilizar códigos que ajudaram e facilitaram o desenvolvimento deste trabalho.

Aos meus companheiros no laboratório, Eliezer por disponibilizar o código de criação das redes em python, e Gustavo, por todas as descontrações e ajudas no laboratório, tornando o trabalho mais leve.

Ao Prof. Dr. Francisco Nepomuceno Filho, por aceitar fazer parte da banca e por compartilhar o laboratório para colocar os computadores que desenvolveu esse trabalho.

Ao Prof, Dr. Juvêncio Santos Nobre, por aceitar fazer parte da banca e por todas as dicas e motivações durante as aulas, o Sr. é um exemplo e referência de profissional.

Aos meus amigos Roberto Costa, Dyego, Martina, Fabrício, Huncas, Johnie, Weslley Rodrigues, Bel Maia, por terem sido presentes em vários momentos de minha vida, trazendo lazer e leveza para os dias.

Ao grupo das Mangas por todo momentos de lazer e descontração, vocês também tornaram o dia a dia mais leve.

A pessoas que com certeza não sabem o impacto positivo durante esses anos, como Tia Ana Lúcia, PC, John Lenon, Seu Luís, Timóteo.

Ao CNPQ e a UFC, aonde fui bolsista durante toda a graduação.

“Se pude enxergar mais longe, foi porque me apoiei em ombros de gigantes.”

(Isaac Newton)

RESUMO

A ferramenta de *Minimum Spanning Tree* (MST), em português Árvore de Abrangência Mínima, tem sido usada para encontrar cluster de similaridades em bolsas de valores, votações de deputados etc. Grande vantagem da técnica é analisar a rede de correlações entre todas as variáveis. A distância de correlação $d(i, j)$ é definida por $d(i, j) = \sqrt{2(1 - \rho(i, j))}$, onde $\rho(i, j)$ é o coeficiente de correlação de Pearson. Os clusters de similaridade são extraídos pela MST obtida com algoritmo de Prim. No ENEM um conjunto de estudantes acima de 100.000 no Ceará respondem às mesmas 180 questões. Uma limpeza inicial, em relação aos dados faltantes, restringiu esse número para aproximadamente 80.000, sobre os quais realizamos uma amostragem de 25.000 alunos estratificada proporcional aos municípios. Os microdados foram obtidos na plataforma Kaggle para os anos de 2018 e 2019. Colocamos os gabaritos dos 4 tipos de provas presenciais para um gabarito padrão discretizado que usamos para obter a matriz de correlação aluno por aluno. Nosso objetivo é extrair os clusters dos estudantes apenas dos vetores respostas, com os quais analisamos o papel das variáveis socioeconômicas para entender a natureza das correlações. As correlações positivas vieram das melhores notas, como esperado. Também analisamos papel da renda, escolaridade dos pais, município e escola.

Palavras-chave: minimum spanning tree; sistemas complexos; microdados do ENEM.

ABSTRACT

The *Minimum Spanning Tree* (MST) tool, in Portuguese Minimum Spanning Tree, has been used to find clusters of similarities in stock exchanges, deputy votes, etc. all variables. The brightness distance $d(i, j)$ is defined by $d(i, j) = \sqrt{2(1 - \rho(i, j))}$, where $\rho(i, j)$ is the Pearson radiance coefficient. Similarity clusters are extracted by MST obtained with Prim's algorithm. In ENEM, a group of over 100,000 students in Ceará answered the same 180 questions. An initial cleaning, in relation to missing data, restricted this number to approximately 80,000, from which we sampled 25,000 students stratified proportionally by municipality. The microdata were obtained from the Kaggle platform for the years 2018 and 2019. We placed the templates for the 4 types of face-to-face tests into a discretized standard template that we used to obtain the student-by-student brightness matrix. Our objective is to extract student clusters from response vectors only, with which we analyze the role of socioeconomic variables to understand the nature of correlations. The positive correlations came from the best grades, as expected. We also analyzed the role of income, parental education, municipality and school.

Keywords: minimum spanning tree; complex systems; ENEM microdata.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

EDM	Educational Data Mining
ENEM	Exame Nacional do Ensino Médio
FIES	Fundo de Financiamento Estudantil Do Ensino Superior
FUNDEB	Fundo de Manutenção e Desenvolvimento da Educação Básica e de Valorização dos Profissionais da Educação
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
LDB	Lei de Diretrizes e Bases da Educação Nacional
MEC	Ministério Da Educação
MST	Minimum Spanning Tree
PNE	Plano Nacional de Educação
PROUNI	Programa Universidade para Todos
SAT	Scholastic Aptitude Test
SISU	Sistema de Seleção Unificada
TRI	Teorema da Resposta ao Item

LISTA DE SÍMBOLOS

G	Grafo Qualquer
$V(G)$	Conjunto de Vértices do Grafo
$E(G)$	Conjunto de Arestas do Grafo
$\|V(G)\|$	Módulo ou cardinalidade do conjunto de Vértices
$\|E(G)\|$	Módulo ou cardinalidade do conjunto de Vértices
ε	Experimento Aleatório
Ω	Espaço Amostral
$\emptyset$	Conjunto Vazio
M_i	i-ésimo Momento não centrado
m_i	i-ésimo Momento centrado
$cov(X,Y)$	Coeficiente de Covariância entre Variáveis aleatórias X e Y
E	Esperança de uma Variável aleatória
σ^2	Variância Populacional
σ	Desvio Padrão Populacional
ρ	Coeficiente de Correlação
λ	Valor Real qualquer
$\mathbb{N}$	Conjunto dos Números Naturais
$\mathbb{R}_+$	Conjunto dos Números Reais Positivos
$\bar{x}$	Estimador da Média/Esperança
$\sum$	Somatório
s^2	Estimador de Variância
$\forall$	Para todo
$d(X,Y)$	Distância de Correlação entre as Variáveis X e Y

SUMÁRIO

1 INTRODUÇÃO

A Árvore Geradora Mínima (MST, Minimum Spanning Tree) é uma poderosa ferramenta da teoria dos grafos, amplamente utilizada em diversos campos do conhecimento, como o mercado financeiro, as ciências sociais e a biologia, devido à sua capacidade de otimizar redes e identificar padrões em grandes conjuntos de dados. A MST conecta todos os pontos de uma rede (nós) com o menor custo possível, sem formar ciclos, o que a torna essencial para solucionar problemas de otimização e estruturação de redes complexas. No mercado financeiro, a MST tem um papel central na análise de correlação entre ativos e na clusterização de investimentos (RN., 1999). Ao representar ações ou títulos como nós e as correlações entre eles como arestas com pesos, a MST permite identificar grupos de ativos que apresentam comportamentos semelhantes. Esse agrupamento (ou clustering) é fundamental para a criação de carteiras de investimento diversificadas e para a análise de risco sistêmico. Durante períodos de crise, por exemplo, a MST ajuda a revelar como certos ativos se conectam e se comportam em conjunto, permitindo uma melhor gestão do risco e a antecipação de crises no mercado. Além disso, outras aplicações da MST foram implementadas em análise de correlação e clusterização de votação de deputados no Brasil e em outros países (Camacho, 2017), de maneira que é possível identificar as complexidades e alinhamentos politícos entre os membros dos partidos. Dessa forma, devido ao avanço tecnológico e a disponibilidade de dados educacionais fornecidos pelo governo, foi idealizado criar uma rede complexa de relação entre os estudantes que estão finalizando o ensino médio. Assim, será utilizado a matriz de correlação dos estudantes em relação ao vetor de respostas dentre as 175 questões respondidas, uma vez que não foi considerado as questões de língua estrangeira, e assim fazer uma clusterização através do algoritmo de Prim e da distância de correlação, de maneira a ordenar a matriz de correlação através da MST e assim associar os resultados com variáveis socio-econômicas, afim de entender sobre os parâmetro educacionais e verificar a tendência da educação brasileira, visto que o ENEM é o principal exame de admissão para Universidades Públicas em todo país, de maneira que serve como um dos principais indicadores da educação no Brasil.

2 ASPECTOS TEÓRICOS

Neste capítulo, abordaremos aspectos teóricos essenciais, começando pela teoria dos grafos, que será introduzida com uma visão geral histórica e sua importância, conforme descrito na seção 2.1. Em seguida, na seção 2.1.1, exploraremos as definições fundamentais e nomenclaturas associadas aos grafos, acompanhadas de exemplos para ilustrar melhor os conceitos apresentados. A seção 2.1.2 dará sequência ao tema, detalhando conexões e caminhos em grafos, explicando como "varrer" um grafo para entender as interações entre seus vértices e arestas. A seção 2.1.3 focará nas representações computacionais de grafos, discutindo as diferentes abordagens, como listas e matrizes de adjacência, e suas implementações práticas em Python. Na seção 2.1.4, serão abordadas as árvores e florestas, destacando suas propriedades e aplicações. A seção 2.1.5 apresentará a estrutura para construção de subgrafos que revelam o caminho mínimo necessário para conectar todos os pontos, essencial para resolver problemas de otimização. Finalmente, na seção 2.1.6, discutiremos um algoritmo específico desenvolvido para este trabalho, conforme a proposta do Prof. Dr. Carlos Lenz Cesar, e sua implementação em Python, detalhando sua aplicação prática e relevância para o trabalho. Na seção 2.2, será abordado a Teoria da Probabilidade, com uma introdução histórica. Na sequência na seção 2.2.1, irá ser definido o contexto probabilístico, abordando conceitos como experimentos aleatórios, espaço amostral e eventos. Experimenta-se a probabilidade de eventos por meio de definições clássica, frequentista e axiomática, destacando a necessidade de uma abordagem mais precisa e formal. A seção 2.2.2 explora a função densidade de probabilidade e momentos, definindo a função distribuição de probabilidade e suas propriedades, e introduz o conceito de momentos e covariância, enfatizando a importância dessas medidas para analisar dependência linear entre variáveis. A covariância é detalhada, incluindo suas propriedades e cálculo, e segue-se na seção 2.2.3 com a definição do coeficiente de correlação, uma medida adimensional de linearidade entre variáveis. Finalmente, a seção 2.2.4 elenca sobre espaço métrico e apresenta a definição de uma métrica e a utilização da distância euclidiana para calcular a distância de correlação, fornecendo uma base para análise estatística no contexto de espaços métricos e variáveis aleatórias.

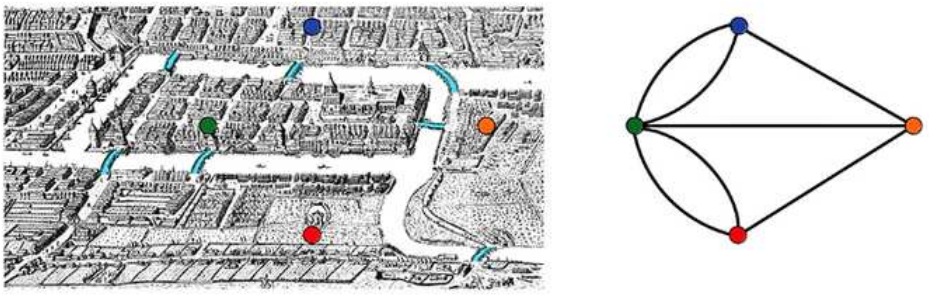
2.1 Teoria dos Grafos

Segundo Camacho (2017), a teoria dos grafos é uma área da matemática discreta que estuda as propriedades e relações de estruturas conhecidas como grafos, representadas

informalmente por pontos conectados por arestas.

A primeira representação de um grafo tem origem na publicação de Leonhard Euler em 1736, onde ele apresentou uma solução negativa para o famoso problema das sete pontes de Königsberg (atual Kaliningrado, Rússia). O problema consistia em atravessar todas as pontes uma única vez, sem repetir nenhuma, e retornar ao ponto de partida. Euler (1968) demonstrou que isso não era possível devido à estrutura específica de conexões das pontes e que seria possível somente se, houvesse zero ou dois vértices de onde saísse um número ímpar de caminhos. Pode-se ver logo abaixo na 1, a imagem das pontes (em azul-claro) e sua representação em grafo.

Figura 1 – Ponte de Konisberg e sua representação em grafos



Fonte: Retirado de Schulz (2020)

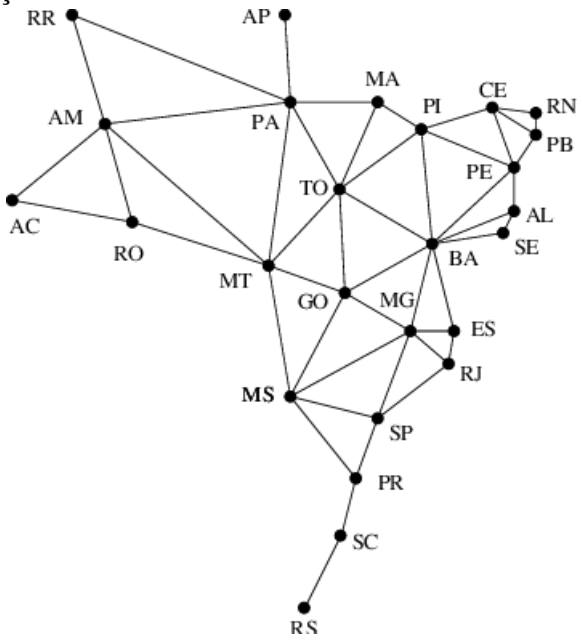
De acordo com Melo (2014), Boaventura e Jurkiewicz (2017) durante o século XX, a teoria dos grafos foi formalizada e se desenvolveu, impulsionada por contribuições oriundas de problemas práticos em otimização de processos, teoria da computação e teoria da informação, como o problema do caminho mínimo. Essas aplicações demonstraram a utilidade dos grafos na modelagem e resolução de uma variedade de problemas complexos de diversas áreas, uma vez que a topologia ou estrutura dos grafos, permitem estudar a forma em que se relacionam os vértices e arestas de um grafo.

2.1.1 Definição e Terminologias

Por conta de sua estrutura, diversas situações são modeladas por grafos, por exemplo, as fronteiras entre os estados do Brasil, podem ser representadas por um grafo, em que os vértices são os estados e as arestas que ligam dois estados estabelece uma fronteira, conforme a Figura 2.

Para Gersting (2017), de maneira informal e baseada em visualização, temos que um grafo pode ser definido como um conjunto não vazio de vértices e um conjunto de arestas, de maneira que, uma aresta conecta dois vértices .

Figura 2 – Representação dos estados brasileiros e suas fronteiras em grafos



Fonte: Retirado de Feofiloff *et al.* (2004)

Apesar de termos uma definição que representa de maneira satisfatória, certas situações, elas podem não se aplicar a situações que necessitem de uma abstração matemática, dessa forma, de maneira formal e geral, para englobar todas as situações, segundo Bondy e Murty (1976), um grafo G é uma tripla ordenada **(V(G), E(G), g)**, sendo **V(G)** um conjunto não vazio de vértices ou nós, **E(G)** um conjunto disjunto de **V(G)** que representa as arestas ou arcos e **g** uma função que associa cada aresta de $E(G)$ a um par não ordenado de vértices de **V(G)** e que representa as extremidades das arestas. Logo, temos que um grafo é:

$$G = (V(G), E(G), g) \tag{2.1}$$

em que:

$$V(G) = (v_1, v_2, v_3, ..., v_N), \text{ com } N \in \mathbb{N}$$

$$E(G) = (e_1, e_2, e_3, ..., e_k), \text{ com } K \in \mathbb{N}$$

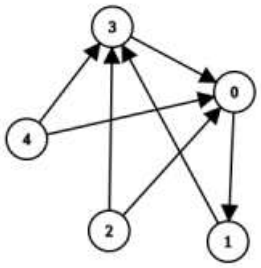
$$g(e_k) = (v_i, v_j), \text{ com } v_i, v_j \in V(G) \text{ e } K \in \mathbb{N}$$

Com isso, podemos estabelecer as principais terminologias de grafos e utilizar um exemplo para ratificar as ideias.

Ao observar as relações entre as arestas, podemos distiguir os grafos em dois tipo:

- **Grafos direcionados**: São grafos que suas representações do conjunto de arestas E(G) são flechas que indicam o sentido de orientação entre os vértices, conforme mostra a Figura 3 abaixo.

Figura 3 – Exemplo de um Grafo Direcionado



Fonte: Retirado de Wikimath-Usp (2023)

- **Grafos não-direcionados**: São grafos que sua representação do conjunto de arestas E(G) não implica em sentido, indicando somente a conexão entre os vértices, conforme a figura 4.

Logo, se G_D for um grafo direcionado, entao:

$$g_D(e_k) = (v_i, v_j) \neq (v_j, v_i) \text{ , com } v_i, v_j \in V(G) \text{ e } K \in \mathbb{N}$$

ou

$$g_D(e_k) = (v_i, v_j) = (v_j, v_i) \text{ , com } v_i, v_j \in V(G) \text{ e } K \in \mathbb{N}$$

enquanto, se G_N for um grafo não-direcional, então, temos somente:

$$g_N(e_k) = (v_i, v_j) = (v_j, v_i) \text{ , com } v_i, v_j \in V(G) \text{ e } K \in \mathbb{N}$$

Além disso, segundo Gersting (2017) a **ordem** de um grafo G é o número de elementos presente no conjunto de vértices $V(G)$, ou seja é quantidade de vérices de um grafo ($|V(G)|$) enquanto o **tamanho** de um grafo é definido como o número de elementos do conjunto de arestas $E(G)$, ou seja a quantidade de arestas presentes no grafo ($|E(G)|$).

De acordo com Gersting (2017) ,em relação ao seus vértices, os grafos apresentam relações com vértices vizinhos, chamados de **vértices adjacentes**, que necessariamente devem ser ligados através de uma aresta, ou seja se $g(e_K) = (v_i, v_j)$, logo v_i e v_j são adjacentes e ligados por e_K, com $i, j, k \in \mathbb{N}$, também podemos ter **vértices isolados**, ou seja, um vértice que não é

adjacente a nenhum outro, enquanto o **grau de um vértice** é quantidade de vértices adjacentes a ele. Com ênfase nas arestas, temos **laços** que são definidos como arestas que se liga a si mesma e **arestas paralelas**, que são arcos que ligam as mesmas extremidades . Algumas característcas dos grafos ainda são importantes de salientar(Gersting, 2017):

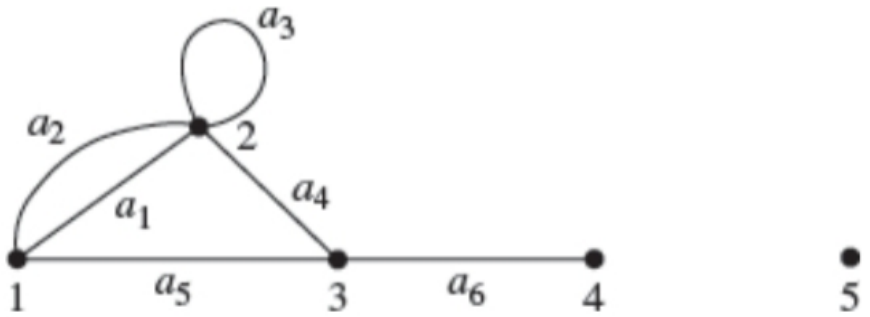
- **Grafos Simples**: São grafos que não possuem arestas paralelas e laços.
- **Grafos Completo**: São grafos que, dado quaisquer dois vértices, eles são adjacentes, ou seja, todos os vértices tem ligações entre si.
- **Sub-Grafo**: São grafos, que o conjunto de vértices e arestas são subconjuntos de nós e arcos, de um grafo original.

Dessa forma, tomamos como exemplo o seguinte grafo. G é um grafo não direcionado, tal que os vértices são: $V(G) = (1,2,3,4,5)$, enquato as arestas: $E(G) = (e_1, e_2, e_3, e_4, e_5, e_6)$ e a função g: $g(e_1) = (1,2); g(e_2) = (2,1); g(e_3) = (2,2); g(e_4) = (2,3); g(e_5) = (3,1)$ e $g(e_6) = (3,4)$

- **Ordem**: 5, cardinalidade do conjunto V(G).
- **Tamanho**: 6, cardinalidade do conjunto E(G)
- **Laços**: e_3, visto que $g(e_3)$ = (2,2), liga o vértice a ele mesmo.
- **Arestas paralelas**:e_1 e e_2, uma vez que o grafo é não-direcionado e $g(e_1) = (1,2)$ e $g(e_2) = (2,1)$
- **Vértice isolado**: (5), uma vez que ele não se conecta a nenhum outro vértice por g, ou seja não há vértices adjacentes, logo seu grau é 0.

A representacação gráfica do grafo, pode ser a seguinte:

Figura 4 – Representação do grafo utilizado como exemplo



Fonte: Retirado de Gersting (2017)

2.1.2 Conectividade e caminhos

Como foi visto na seção anterior, se dois vértices são conectados por uma aresta, logo esses vértices são adjacentes ou vizinhos, assim como, o grau de um nó foi definido como a quantidade de vértices adjacentes a ele, dessa maneira, pode-se afirmar que se dois vértices são adjacentes, logo eles são conectados e a quantidade de conexões de um vértice é o seu **grau**. Dessa forma, se um grafo tem vértices conectados, podemos estabelecer, ou não, um caminho que vai de um vértice a outro, assim, segundo Gersting (2017), **caminho** é definido como uma sequência de nós e arestas, de maneira que conecta dois vértices, e o comprimento do caminho é o número total de arestas, assim:

$$n_0, a_0, n_1, a_1, ..., n_{k-1}, a_{k-1}, n_k \tag{2.2}$$

em que: para cada i, as extremidades do arco a_i são $n_i - n_{i+1}$.

Além disso, para Gersting (2017) pode-se ter caminhos que iniciam em um nó e terminam nele mesmo, a esses tipos de caminhos denominaremos **ciclo**.

Assim, surge a seguinte definição, segundo Gersting (2017), um **grafo é conexo**, se existir um caminho em quaisquer entre dois nós. Podemos verificar que o grafo da figura 4, não é conexo, uma vez que não há um caminho que liga qualquer vértice ao nó 5.

Para o grafo da Figura 4, temos que um caminho que liga o vértice 1 ao vértice 4, pode ser:

$$1, a_1, 2, a_4, 3, a_6, 4$$

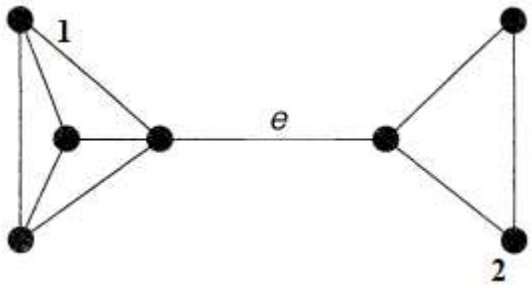
ou

$$1, a_2, 2, a_4, 3, a_6, 4$$

Com isso, surge um questionamento natural, de saber qual caminho é o menor, para isso, devemos verificar o comprimento do caminho, que no caso da análise é o mesmo, e igual a 3. Dessa forma, temos que não há menor caminho, porém, em diversas situações do mundo real, consideramos as arestas dos grafos com um peso associado, e a estes grafos denominamos como **grafos ponderados** (Cormen et al., 2012) . Assim, as arestas dos grafos podem recebem valores numéricos, por exemplo a distância entre cidades ou valores associados a métrica do problema em questão.

Em certos grafos, há a arestas que se forem removidas, torna o **grafo desconexo**, há essas arestas chamaremos de **pontes**(Wilson, 1996). Segue abaixo, uma representação de uma ponte.

Figura 5 – Exemplo de Grafo com uma ponte



Fonte: Retirado de Wilson (1996)

Podemos observar, que se removermos a aresta **e**, nosso grafo se torna desconexo, uma vez que não existe caminho entre o conjunto de vértices do lado direito, com o lado esquerdo, como exemplo, se removermos a aresta **e**, não consegue-se estabelecer um caminho entre os vértices 1 e 2, logo **e** é uma ponte.

2.1.3 Principais Representações

A representação computacional de grafos é fundamental na solução de problemas complexos(Cormen et al., 2012) e otimização de processos, sendo amplamente utilizada para modelar diversas situações. Uma das formas mais comuns de representação de grafos é através de matrizes, que oferecem uma estrutura organizada e eficiente para armazenar dados sobre as relações entre os vértices. Esta escolha não é arbitrária, as matrizes permitem, utilizar conceitos e propriedades da álgebra linear(Cormen et al., 2012) e a representação clara das conexões entre os elementos do grafo, mas também facilitam operações computacionais como busca de caminhos, identificação de ciclos, e determinação de propriedades estruturais. Assim, exploraremos o papel das matrizes na representação de grafos e dois tipos de representações.

2.1.3.1 Matriz de adjacência

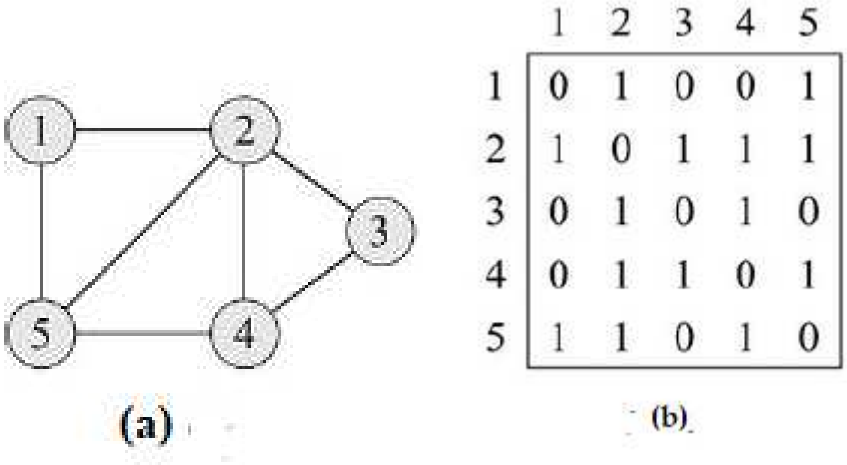
Segundo Cormen *et al.* (2012), a representação através de matrizes de adjacência de um grafo $G = (V(G), E(G), g)$, com vértices numerados de 1,2,3,..., $|V(G)|$, é uma maneira utilizada quando temos um grafo onde o número de arestas é próximo do quadrado do número

de vértices **(grafo denso)**, logo a matriz de adjacência é uma matriz quadrada A = (a_{ij}), tal que:

$$a_{ij} = \begin{cases} 1, & se \quad (i,j) \in g \\ 0, & c.c \end{cases} \tag{2.3}$$

Por se tratar de uma matriz simétrica, uma vez que o grafo não é dirigido, temos que a matriz de adjacência é igual sua transposta , $A = A_T$, com isso, podemos reduzir a matriz, utilizando somente a sua diagonal principal e os elementos, ou acima ou abaixo da diagonal(Cormen et al., 2012), para grafos muito grandes, a matriz pode conter muitos zeros, isso representa um grande consumo de memória, logo podendo se mostrar desvantajosa nessa situação. Logo abaixo na Figura 6, podemos verificar um grafo e sua matriz de adjacência.

Figura 6 – Exemplo de Grafo e sua matriz de adjacência



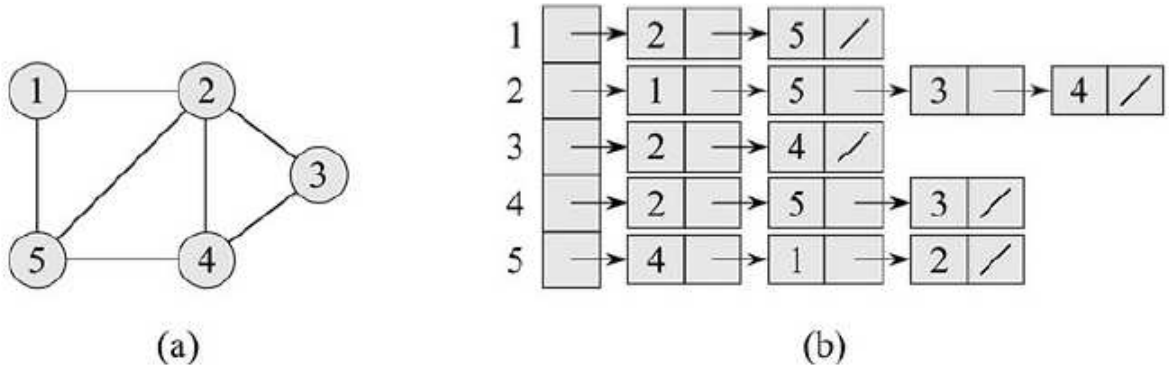
	1	2	3	4	5
1	0	1	0	0	1
2	1	0	1	1	1
3	0	1	0	1	0
4	0	1	1	0	1
5	1	1	0	1	0

Fonte: Retirado de Cormen *et al.* (2012) a) Grafo G não direcionado utilizado de exemplo. b) Matriz de adjacência do grafo G.

2.1.3.2 Lista de adjacência

Uma outra maneira de representar os grafos, bem mais utilizado quando o número de arestas é muito menor que o quadrado do número de vértices **(grafo esparso)**, é a representação por listas de adjacência(Cormen et al., 2012). Dessa maneira, cada vértice representa uma lista e em cada uma delas é utilizado ponteiros para indicar os vértices adjacentes, um por vez, assim, em certos casos pode ser mais eficiente pois representa somente valores não nulos (Gersting, 2017). Pode-se observar, na Figura 7, um exemplo dessa representação.

Figura 7 – Exemplo de Grafo e sua lista de adjacência



Fonte: Retirado de Cormen *et al.* (2012) a) Grafo G não direcionado utilizado de exemplo. b)Lista de adjacência do grafo G, as barras invertidas ao final de cada lista, informa o último vértice adjacente.

2.1.4 Árvores

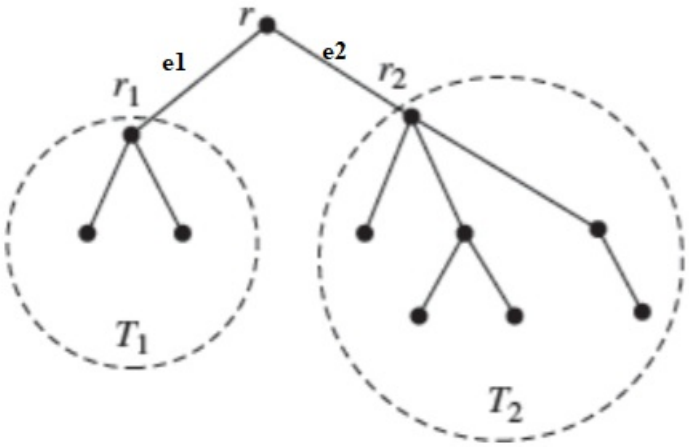
Para Bondy e Murty (1976), uma árvore é grafo simples, conexo e acíclico, ou seja sem laços ou arestas paralelas, há um caminho entre quaisquer dois vértices e não há um caminho que ligue o vértice a ele mesmo. Pode-se afirmar, que o caminho é único entre dois vértices, uma vez que a árvore é acíclica (Bondy; Murty, 1976; Gersting, 2017). Além disso, algumas definições são importantes de serem mencionadas, as árvores podem ter ou não uma raiz, que é um nó especial, no qual a árvore se origina, caso a árvore não tenha uma **raíz**, chamamos de raíz **árvore** livre, temos também as **florestas**, que são grafos acíclicos, conexo ou não, ou seja, um conjunto de árvores disjuntas (Gersting, 2017). Assim, podemos estabelecer, o seguinte teorema, conforme Wilson (1996):

Teorema: Seja T um grafo com n vértices, as seguintes afirmações são válidas

- **i): T é uma árvore.**
- **ii): T não apresenta ciclos e o número de arestas é: n-1.**
- **iii): T é conexo e o número de arestas é: n-1.**
- **iv):T é conexo e cada aresta é uma ponte.**
- **v): Quaisquer dois vértices de T são conectados através de somente um caminho.**
- **vi): T não contém ciclos, mas a adição de uma nova aresta cria extamente um ciclo.**

Logo abaixo, podemos ver a diferença entre árvores e floresta.

Figura 8 – Representação de uma árvore.



Fonte: Retirado de Gersting (2017)

Na Figura 8, podemos esclarecer a distinção entre árvores e florestas. Tomamos de exemplo remover a ponte **e1**, dessa maneira o conjunto de vértices **T1**, fica desconexo do conjunto formado pelo vértice **r** e pelos vértices de **T2**, logo teriamos duas árvores disjuntas, ou simplismente, uma floresta.

Dessa forma, podemos estabelecer o que é uma **árvore geradora** de um grafo conexo. Segundo Gersting (2017), uma **árvore geradora** pode ser definida como uma árvore sem raiz, cujo o conjunto de vértices é igual a do grafo e o conjunto das arestas é um subconjunto dos conjunto de arestas do grafo, logo podemos considerar as árvores geradoras, como sub-grafos.

2.1.5 Árvore Geradora Mínima (Minimum Spanning Tree)

Muitos problemas reais, envolve há busca por otimização em processos, de maneira a facilitar ou tornar economicante vantajosa certas situações. Dessa forma, a árvore geradora mínima, é um processo de otimização, que consiste em estabelecer uma árvore geradora cujo peso é o menor possível (Gersting, 2017).

O peso de grafo valorado, G = (V(G),E(G),g), com n vértices, definido como:

$$P(G) = \sum_{i=1}^{\mathrm{n}} e_i, \quad onde : g(e_i) = (v_k, v_j) \quad v_i, v_j \in V(G); i, j, k \in \mathbb{N} \tag{2.4}$$

Além disso, temos importantes algoritmos para gerar uma MST em um grafo, como, o de Kruskal, de Boruvka´s, Dijkstra e o de Prim. Neste trabalho, utilizaremos o algoritmo de Prim, uma vez que o algoritmo permite sequenciar a ordem da MST, assim a ordem dos vértices retornadas pelo algortimo é utilizado como reordenação de uma matriz.

2.1.6 Algoritmo de Prim

O algoritmo de Prim, busca gerar uma Minimum Spanning Tree (MST), selecionando um vértice inicial e a cada passo busca o menor valor de aresta que liga um vértice já selecionado a um não selecionado, dessa forma, a cada ciclo do algoritmo, com base em cada vértices já selecionados, busca-se entre os vértices não selecionados o que tem menor menor peso em sua aresta. Assim, podemos considerar o algoritmo de Prim, guloso(Cormen et al., 2012) , uma vez que ele faz a escolha dos vértices a cada passo do algoritmo, não considerando o quadro geral do problema.

Dessa forma, o algoritmo de Prim, pode ser expresso da seguinte maneira:

Seja **G** um grafo ponderado, podemos obter uma Minimum Spanning Tree **T**, com o seguinte processo:

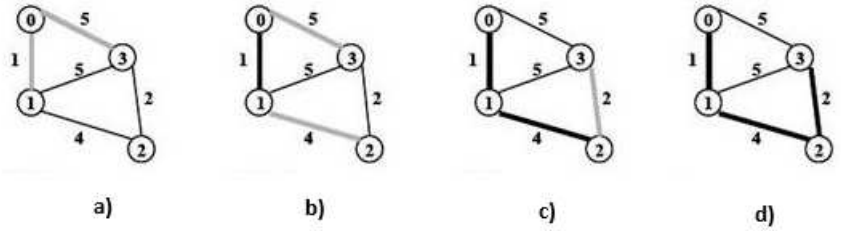
Passo 1: Definir o conjunto de vértices visitados, não visitados e as arestas que ligam os vértices.

Passo 2: Escolher dentro conjunto de arestas, a aresta de menor valor e inserir dentro do conjunto dos vértices visitados, os vértices ligados pela aresta escolhida (menor valor), remover os vértices do conjunto dos vértices não visitados e manter a ligação entre esses vértices.

Passo 3: Escolher entre os elementos do conjunto dos vértices visitados, qual tem a aresta de menor peso que ligue a um vértice do conjunto de vértices não visitados e manter a ligação entre os vértices da aresta escolhida e em seguida remover o vértice do conjunto de vértices não visitados e inserir no conjunto de vértices visitados.

Passo 4: Repetir o passo 3 até todos os vértices terem sido selecionados.

Figura 9 – Processo do Algoritmo de Prim



Fonte: Retirado de Wordpress-blog (2022) a) Grafo inicial. b) Escolhido primeiros vértices. c) Desenvolvimento do Algoritmo. d) MST encontrada.

Na Figura 9, pode-se ver o funcionamento do Algoritmo de Prim. Inicialmente na

Figura 9 a), temos o grafo inicial, logo em seguida na Figura 9 b) temos os primeiros vértices escolhidos, que são (0) e (1), uma vez que a aresta de menor peso é a que conecta-os, mais adiante na imagem c) temos que escolher dentro os vértices ja visitados, o que tem um vértice adjacente de menor peso, logo foi escolhido o vértice (2), uma vez que dentre os vértices adjacentes de (0) e (1), a aresta que conecta (1) e (2), é a de menor peso, na imagem d) o processo é repetido, de maneira que todos os vértices foram visitados.

2.2 Teoria da Probabilidade

A teoria da probabilidade é um ramo da matemática que estuda fenômenos aleatórios e suas regularidades. Suas origens remontam ao século XVII, com os trabalhos de Blaise Pascal e Pierre de Fermat, que discutiram problemas relacionados a jogos de azar. Segundo Viali (2020) esses estudos iniciais foram fundamentais para a formalização do conceito de probabilidade.

Historicamente, a probabilidade surgiu da necessidade de entender e prever eventos incertos. De acordo com Vasconcelos *et al.* (2022) antes de Pascal e Fermat, já havia registros de jogos de azar na antiguidade, mas foi com esses matemáticos que a probabilidade começou a ser tratada de forma científica. A correspondência entre Pascal e Fermat em 1654 é considerada um marco, pois eles desenvolveram métodos para calcular probabilidades em jogos de dados e cartas (Gneri, 2023).

No século XVIII, a teoria da probabilidade foi ampliada por Jakob Bernoulli e Abraham de Moivre. Bernoulli introduziu a Lei dos Grandes Números, que descreve a tendência de resultados de experimentos aleatórios se aproximarem de um valor esperado à medida que o número de experimentos aumenta. De Moivre, por sua vez, contribuiu com a fórmula de aproximação para distribuições binomiais, conhecida como Teorema do Limite Central.

O século XIX viu a consolidação da probabilidade como uma disciplina matemática rigorosa, com contribuições significativas de Carl Friedrich Gauss e Pierre-Simon Laplace. De Moivre também desenvolveu a distribuição normal, também conhecida como curva de Gauss, que descreve a distribuição de muitos fenômenos naturais. Segundo Gneri (2023) Laplace, por sua vez, publicou a "Teoria Analítica das Probabilidades", que sistematizou e expandiu os métodos probabilísticos.

No século XX, a teoria da probabilidade foi formalizada através da abordagem axiomática de Andrey Kolmogorov, que estabeleceu os fundamentos modernos da probabilidade. Kolmogorov definiu a probabilidade como uma medida em um espaço de probabilidade, propor-

cionando uma base sólida para o desenvolvimento de teorias subsequentes (Vasconcelos et al., 2022).

Atualmente, a teoria da probabilidade é aplicada em diversas áreas, como Estatística, Física, Economia, Biologia e Ciências Sociais, por exemplo. Sua importância reside na capacidade de modelar e prever comportamentos em sistemas complexos e incertos.

2.2.1 Aspectos Prévios e definição

No contexto probabilístico, antes de começar a definição formal, é importante entender algumas nomenclaturas básicas que são fundamentais para modelar diversas situações matematicamente. Essas nomenclaturas geralmente estão relacionadas à teoria dos Conjuntos Numéricos e ajudam a estabelecer uma ligação entre o mundo observável e o formalismo matemático, além de produzir o contexto para a definição.

- **Experimento aleatório (ε)**: Segundo Dantas (2008), um experimento aleatório é qualquer processo que, quando repetido sob as mesmas condições, não produzem o mesmo resultado, logo os experimentos podem ser entendidos como determinísticos ou não-determinísticos, o primeiro se refere a experimentos que dada as condições iniciais, podemos prever o resultado final, visto que o experimento não está sujeito a incertezas, equanto os experimentos não-determinísticos, se referem a experimentos sujeito a incertezas, mesmo sabendo as condições iniciais, o resultado final pode ser o qualquer um dentro dos resultados previstos. Para Meyer (1987) dessa forma, em experimentos não-determinísticos, temos aqueles que são ditos aletórios, uma vez que cada experimento pode ser repetido indefinidamente sob as mesmas condições, pode-se saber o conjunto de possibilidades daquele experimento e caso o experimento for repetido muitas vezes, por um grande número de tentativas, surgirá uma tendência matemática, plausível de ser modelada, mas ainda não teremos toda certeza do resultado final.
- **Espaço Amostral (Ω)**: Há todo experimento aleatório, podemos associar a seus possíveis resultados um conjunto, dentro do contexto desse trabalho podemos que o conjunto de todos os possível resultados é chamado de espaço amostral (Meyer, 1987).
- **Eventos (E)**: Em relação a um experimento, temos o espaço amostral e em relação ao espaço amostral e ao experimento temos um evento, que é um conjunto de resultados possíveis (Meyer, 1987), logo podemos estabelecer que um evento é um sub conjunto do espaço amostral, enquanto o complementar de um evento é o subconjunto de eventos

restantes, com exceção do evento de análise. Se A é um evento, seu complementar é A^C

Assim, podemos estabelecer o seguinte exemplo: seja E, um experimento, definido como, lançar duas moedas e esperar o resultado da face voltada para cima. Em relação a esse experimento, o espaço amostral Ω = (cara,cara), (cara,coroa), (coroa,cara), (coroa, coroa), um evento A, poderia ser A = Obter coroa no primeiro lançamento, logo A = (cara,cara),(cara,coroa), ou seja, um subconjunto do espaço amostral.

Dessa forma, podemos estabelecer algumas definições de probabilidade. Primeiramente, temos a definição clássica, definida para subconjuntos unitários e igualmente prováveis, segundo Cesar (2013), se temos um espaço amostral Ω finito com N resultados equiprováveis e A um evento com N_A elementos, entao a probabilidade pode ser definida como:

$$P(A) = \frac{N_A}{N} \tag{2.5}$$

Uma vez que não é aplicada a todos os casos e por se tratar de uma definição que usa a própria ideia de probabilidade para ser definida, precisaremos de uma definição mais precisa.

Com isso, temos a definição frequêntista, come essa definição utilizaremos a frequência com que um dado evento ocorra, não se limitando, a quantidade de elementos de um espaço amostral e de eventos equiprováveis, segundo Magalhães (2006), seja n_A o número de vezes que um evento A ocorre em n repetições, logo a probabilidade do evento ocorrer é:

$$P(A) = \lim_{n\to\infty} \frac{n_A}{n} \tag{2.6}$$

Dessa forma, chegou-se a conclusão que a probabilidade deve ser definida com base em axiomas (Cesar, 2013). Assim, a definição axiomática, utiliza os axiomas de Kolmogorov para uma função de conjuntos, seja P(A) = f : $\Omega \to \mathbb{R}$, com $P(A) \in [0,1]$, então:

- $P(A) \geq 0, \forall A \in \Omega$
- $P(\Omega) = 1, \quad \Omega$ é chamado de evento certo.
- Se $\{A_n\}$ é uma sequeência monótona, com n ≥ 1, então $P(\cup A_n) = \sum P(A_n)$.

Contudo, agora pode-se estabelecer uma função numérica que associa uma probabilidade a ocorrência de um evento, de maneira a modelar diversas situações, inclusive o tema deste trabalho.

2.2.2 Função densidade de Probabilidade, momentos e covariância

Na estatística, temos uma função que associa um evento, há uma probabilidade, a partir de uma variável aleatória, chamada de função distribuição de probabilidade. As variáveis aleatórias podem ser entendidas como "Uma função do espaço amostral nos reais, para o qual é possi´vel calcular a probabilidade de ocorrência de seus valores"(Magalhães, 2006) Seja, A, um evento, tal que A = $\{x_V \leq x\}$, podemos calcular P(A), então:

$$F(x) = P(\{x_V \leq x\}), x \in R \tag{2.7}$$

Podemos estabelecer algumas propriedades da função distribuição de probabilidade:

- A função F_X é monótona não decrescente, i.e $\forall x < y,\ F_X(x) \leq\ F_Y(y)$.
- Contínua a direita, i.e $\lim_{x \to a^+} F_X(x) = F_X(a)$.
- Tem as seguintes assíntotas, $\lim_{n \to -\infty} F_X(x) = 0\ e\ \lim_{n \to \infty} F_X(x) = 1$.

Enquanto no caso multivariado, a função recebe o nome de distribuição conjunta, sendo $\{X_1, X_2, ..., Xn\}$ um conjunto de variáveis aleatórias onde pode-se estabelecer, cada uma a um evento $\{X_i < x_i\}$e pertencentes ao R^n, entao:

$$F(X_1, X_2, ..., X_n) = P\{X_1 < x_1, X_2 < x_2, ..., X_n < X_n\} \tag{2.8}$$

Segue as seguintes propriedades:

- $F(-\infty, X_2, ..., X_n) = F(X_1, -\infty, ..., V_n) = F(X_1, X_2, ..., -\infty) = 0$
- $F(+\infty, +\infty, ..., +\infty) = 1$
- $P(x_x < X_1 \leq x_y, X_2 \leq x_2, ...X_n \leq x_n) = F(x_y, X_2, ..., X_n) - F(x_x, X_2, ..., X_n)$

Agora, iremos utilizar somente o caso multivariado, uma vez que no desenvolvimento deste trabalho foi o tipo de análise utilizada. Assim, podemos definir a função densidade de probabilidade, entendida como a derivada da função probabilidade, logo:

$$f(X1, X2, ..., Xn) = \frac{\partial^n F(X_1, X_2, ..., X_n)}{\partial X_1 \partial X_2 ... \partial X_n} \tag{2.9}$$

Assim, para ser uma função densidade de probabilidade, a função acima deve satisfazer:

- $f(x) \geq 0, \forall\ x \in \mathbb{R}^n$
- $\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} ... \int_{-\infty}^{+\infty} f(X_1, X_2, ..., X_n)\, dX_1 dX_2 ... dX_n = 1$

Dessa forma, é de grande interesse, estabelecer medidas que possibilitem análises, com isso, podemos definir o valor esperado ou esperança. Seja, $Z = g(X_1, X_2, X_3, ..., X_n)$ uma função escalar das variáveis aleatórias, $\{X_1, X_2, X_3, ..., X_n\}$ e função densidade de probabilidade $f(X_1, X_2, X_3, ..., X_n)$, a operação esperança, pode ser definida como:

$$E[Z] = E[g(X_1, X_2, X_3, ..., X_N)] \tag{2.10}$$

$$E[Z] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} ... \int_{-\infty}^{+\infty} g(X_1, X_2, ..., X_n) f(X_1, X_2, ..., X_n)\, dX_1 dX_2 ... dX_n \tag{2.11}$$

Algumas propriedades notórias são:

- $E[K] = K$, sendo K uma constante.
- $E[\alpha X_1 + \beta X_2] = \alpha E[X_1] + \beta E[X_2]$

Com isso, podemos construir os momentos, de maneira geral, os momentos centrados em relação a origem, podem ser definidos da seguinte maneira:

$$M_{k_1,k_2,,...,k_n} = E[X_1^{k_1} X_2^{k_2} ... X_n^{k_n}] \tag{2.12}$$

$$M_{k_1,k_2,...,k_n} = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} ... \int_{-\infty}^{+\infty} (X_1^{k_1}, X_2^{k_2}, ..., X_n^{k_n}) f(X_1, X_2, ..., X_n)\, dX_1 dX_2 ... dX_n \tag{2.13}$$

Assim, cada k_n, indica o momento de cada variável X_n, logo queremos, encontrar o primeiro momento centrado na origem de cada variável, que será $M_{100...0}$ para a variável X_1, $M_{010...0}$ para a variável X_2, $M_{000...1}$ para a variável X_n.Dessa maneira, o primeiro momento é conhecido como média (μ_n) ou valor esperado, como definimos antes, assim:

$$M_{100...0} = E[X_1] = \mu_1$$
$$M_{010...0} = E[X_2] = \mu_2$$
$$M_{000...1} = E[X_n] = \mu_n$$

Dessa forma, podemos estabelecer os momentos centrados em relação a média, da seguinte maneira:

$$m_{k_1,k_2,k_3,...,k_n} = E[(X_1 - \mu_1)^{k_1} (X_2 - \mu_2)^{k_2} ... (X_n - \mu_n)^{k_n}] \tag{2.14}$$

$$m_{k_1,k_2,...,k_n} = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} ... \int_{-\infty}^{+\infty} (X_1 - \mu_1)^{k_1} (X_2 - \mu_2)^{k_2} ... (X_n - \mu_n)^{k_n} f(X_1, X_2, ..., X_n)\, dX_1 dX_2 ... dX_n \tag{2.15}$$

Dessa maneira,podemos calcular os momentos centrados em relação a média,de interesse deste trabalho, estudaremos a covariância e variância. A covariância, é uma medida estatística que possibilita calcular a dependência linear entre duas variáveis aleatórias (X_1, X_n), podemos definir da seguinte maneira:

$$m_{100...1} = cov(X_i, X_j) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (X_1 - \mu_1)^1 (X_n - \mu_n)^1 f(X_1, X_n)\, dX_1 dX_n \tag{2.16}$$

$$m_{100...1} = cov(X_i, X_j) = E[(X_1 - \mu_{X_1})(X_n - \mu_{X_n})] \tag{2.17}$$

Logo, podemos afirmar que a covariância entre duas variáveis aleatórias pode ser positiva, negativa ou nula, se $cov(X_i, X_j) > 0$, dizemos que há uma relação de depedência positiva, caso $cov(X_i, X_j) < 0$, dizemos que há uma relação de depedência negativa e se $cov(X_i, X_j) = 0$, dizemos que as duas variáveis são independentes, em determinados casos. Algumas propriedades da covariância são:

- $cov(X_i, X_j) = cov(X_j, X_i)$
- $cov(X_i + X_j, X_k) = cov(X_i, X_k) + cov(X_j, X_k)$
- $cov(X_i, X_j) = E[X_i X_j] - E[X_i]E[X_j]$

  Prova:

$$m_{100...1} = cov(X_i, X_j) = E[(X_i - \mu_{X_i})(X_j - \mu_{X_j})], \text{ logo:}$$
$$cov(X_i, X_j) = E[X_i X_j - (V_i \mu_{X_j}) - (X_j \mu_{X_i}) + (\mu_{X_j} \mu_{X_i})]$$
$$cov(X_i, X_j) = E[X_i X_j] - \mu_{X_j} E[X_i] - \mu_{X_i} E[X_j] + (\mu_{X_j} \mu_{X_i}), \text{ sendo: } E[X_j] = \mu_{X_j} \text{ e } E[X_i] = \mu_{X_i}$$
$$cov(X_i, X_j) = E[X_i X_j] - \mu_{X_j} \mu_{X_i} - \mu_{X_i} \mu_{X_j} + \mu_{X_j} \mu_{X_i}$$
$$cov(X_i, X_j) = E[X_i X_j] - E[X_i]E[X_j]$$

  Logo, se duas variáveis são independentes, então: $E[X_i X_j] = E[X_i]E[X_j]$ e $cov(X_i, X_j) = 0$.
- $cov(\alpha X_i, \beta X_j) = \alpha\beta\, cov(X_i, X_j)$
- $cov(X_i, k) = 0$, sendo k, uma constante.

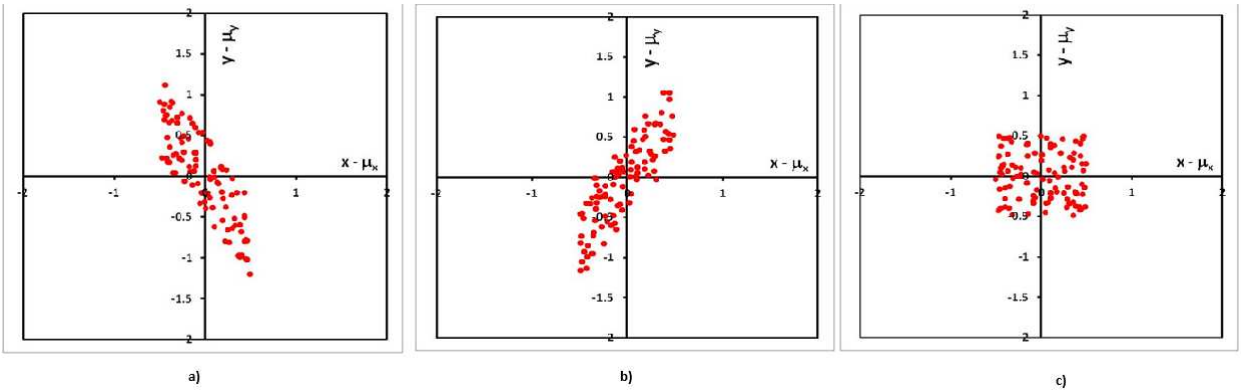
  Enquanto as variâncias, conhecidas como segundo momento centrado, de cada variável pode ser definida como, Seja $Var(X_n)$, a variância denotado por σ_n^2:

$$m_{000...2} = \sigma_n^2 = E[(V_n - \mu_n)^2] = E[V_n^2 - 2V_n\mu_n + \mu_n^2]$$
$$= E[V_n^2] - 2\mu_n E[V_n] + \mu_n^2, \text{ sendo: } E[V_n] = \mu_n.$$
$$= E[V_n^2] - 2\mu_n^2 + \mu_n^2$$
$$= E[V_n^2] - \mu_n^2$$

Sendo a variância σ_n^2, podemos definir o desvio padrão, outra importante medida, que dimensionaliza o problema para uma dimensão coerente, uma vez que diversos problemas não fazem sentido suas dimensões ficarem ao quadrado, como: $\sigma_n = \sqrt{\sigma_n^2}$. É importante frisar que: $V(X_i) = cov(X_i, X_i) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(X_i - \mu_i)^2 f(X_i, X_i)\, dX_i dX_i$. Algumas propriedades da variância :

- $V(X_i) = E[X_i^2] - (E[X_i])^2$
- $V(kX_i) = k^2 V[X_i]$
- $V(\alpha + \beta X_i) = \beta^2 V[X_i]$
- $V(\alpha X_i + \beta X_j) = \alpha^2 V[X_i]\beta^2 V[X_j] + 2\alpha\beta cov(X_i, X_j)$

Figura 10 – Visualização Gráfica da Covariância



Fonte: Retirado de Cesar (2013) a) Covariância menor que zero, logo as variáveis são negativamente correlacionadas. b) Covariância maior que zero, logo as variáveis são positivamente correlacionadas. c) Covariância igual a zero, não há correlação entre as variáveis.

2.2.3 Coeficiente de correlação

Apesar da covariância verificar o estabelecimento ou não de uma dependência entre as variáveis, surgem problemas devido a dimensionalidade das variáveis, uma vez que não podemos atribuir um sentido as dimensões. Dessa forma, precisamos de uma maneira de verificar o estabelecimento da correlação entre as variáveis, sem nós preocupar com os efeitos da dimensionalidade, para isso vamos definir, o coeficiente de correlação de Pearson, uma grandeza

adimensional, que funciona como medida de linearidade entre as variáveis e é obtida apartir da covariância. Assim, segundo Magalhães (2006) o coeficiente de correlação entre duas variáveis aleatórias X_i e X_j é:

$$\rho(X_i, X_j) = \frac{cov(X_i, X_j)}{\sigma_{X_i}\sigma_{X_j}} \tag{2.18}$$

Dessa maneira, ao dividirmos a covariância, pelo desvião padrão das variáveis obtemos uma grandeza adimensional. Porém, falta verificar quais os possíveis valores para o coeficiente de correlação, assim, iremos utilizar a Desigualdade de Cauchy-Schwarz para uma esperança de uma quantidade positiva, dessa forma, seja: $E\{[\lambda(X-\mu_x)-(Y-\mu_y)]^2\} \geq 0$, onde $\lambda \in$ R. Primeiramente, desenvolvendo a esperança, temos:

$$\begin{aligned} E\{[\lambda(X-\mu_x)-(Y-\mu_y)]^2\} &= E\{[\lambda^2(X-\mu_x)^2 - 2\lambda(X-\mu_x)(Y-\mu_y) + (Y-\mu_y)^2] \\ &= \lambda^2 E[(X-\mu_x)^2] - 2\lambda E[(X-\mu_x)(Y-\mu_y)] + E[(Y-\mu_y)^2] \\ &= \lambda^2\sigma_x^2 - 2\lambda cov(X,Y) + \sigma_y^2 \end{aligned}$$

Logo, temos:

$$\lambda^2\sigma_x^2 - 2\lambda cov(X,Y) + \sigma_y^2 \geq 0 \tag{2.19}$$

Podemos observar, que se trata de uma desigualdade de uma equação quadrática. Para equações do tipo $A\lambda^2 + B\lambda + C$, só pode ser satisfeita, se $A\lambda^2 + B\lambda + C = 0$, não admitir raizes reais, logo $B^2 - 4AC \leq 0$, subistituindo os valores, temos que:

$$\begin{aligned} B^2 - 4AC &\leq 0 \\ (2cov(X,Y))^2 - 4\sigma_x^2\sigma_y^2 &\leq 0 \\ 4cov^2(X,Y) - 4\sigma_x^2\sigma_y^2 &\leq 0 \\ 4cov^2(X,Y) &\leq 4\sigma_x^2\sigma_y^2 \\ \frac{cov^2(X,Y)}{\sigma_x^2\sigma_y^2} &\leq 1 \\ \frac{cov^2(X,Y)}{\sigma_x^2\sigma_y^2} &\leq 1 \end{aligned}$$

Dessa maneira, temos:

$$-1 \leq \frac{cov(X,Y)}{\sigma_x\sigma_y} \leq 1 \tag{2.20}$$

Logo podemos concluir, que o coeficiente de correlação, varia de -1 a 1, logo podendo ser zero, o que indica que as variáveis não possui dependência linear, se o valor for negativo ou próximo a -1, indica que as variáveis possui uma depedência linear negativa, e quando o valor for positivo ou próximo a 1, a dependência linear entre as variáveis é positiva. Além disso, podemos construir uma matriz quadrada KxK, onde K é o número de variáveis, onde os elementos da matriz ρ_{ij} são os respectivos coeficientes de correlação entre as variáveis i e j, assim, construímos uma matriz simétrica, da seguinte forma:

$$\rho_{ij} = \begin{pmatrix} \rho_{11} & \rho_{12} & \rho_{13} & \cdots & \rho_{1k} \\ \rho_{21} & \rho_{22} & \rho_{23} & \cdots & \rho_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{k1} & \rho_{k2} & \rho_{k3} & \cdots & \rho_{kk} \end{pmatrix}$$

Vale ressaltar que a diagonal principal é toda igual a 1, uma vez que o coeficiente de correlação de uma variável com ela mesmo é 1. Dessa forma:

$$\rho(i,i) = \frac{cov(i,i)}{\sigma_i \sigma_i}$$
$$\rho(i,i) = \frac{\sigma_i^2}{\sigma_i^2}$$
$$\rho(i,i) = 1$$

2.2.4 Amostragem Estatística e Estimadores

A amostragem é uma especialidade na estatística que permite extrair informações representativas de uma população maior através da seleção de uma parte dela. De acordo com Morettin P.A.; Bussab (2006), em estudos estatísticos e científicos, a escolha adequada da amostra é crucial para garantir que os resultados sejam generalizáveis e aplicáveis à população de interesse, evitando vieses e distorções nos dados.

Dessa forma, algumas nomenclaturas são importantes.**População** refere-se ao conjunto completo de elementos que compartilham uma característica comum e são objeto de estudo ou interesse. Por outro lado, **Amostra** é uma parte selecionada dessa população, escolhida de forma a representar suas características principais de maneira precisa e confiável. Enquanto a população é o universo completo que queremos estudar, a amostra é uma fração dele que permite inferências sobre toda a população sem necessariamente examiná-la por completo. Assim, de

acordo com Bolfarine H.; Bussab (2005) podemos entender que a amostra é um subconjunto da população, que pode ser interpretada como o espaço amostral. Em muitos casos, é impraticável ou impossível estudar todos os membros de uma população devido a limitações de recursos, tempo ou logística. Para Silva (2015) A amostragem resolve esse problema ao selecionar uma parcela significativa e diversificada da população, garantindo que os resultados obtidos da amostra possam ser extrapolados com confiança para o conjunto total.

Existem duas principais abordagens para a amostragem em estudos estatísticos: probabilística e não probabilística. A amostragem probabilística se destaca por garantir que todos os elementos da população tenham uma chance conhecida e não nula de serem selecionados para a amostra, o que permite inferências mais confiáveis sobre a população maior, em determinados casos. Entre os métodos probabilísticos estão a amostragem aleatória simples, onde cada membro tem igual probabilidade de ser escolhido, e a amostragem estratificada, que divide a população em grupos homogêneos (estratos) antes da seleção.. Dentro da amostragem estratificada, há a variação da estratificação proporcional, onde o tamanho de cada estrato na amostra é proporcional à sua presença na população total. Isso garante que grupos significativos dentro da população sejam representados de maneira adequada na amostra, aumentando a precisão das estimativas (Silva, 2015). Em contrapartida, a amostragem não probabilística depende de critérios subjetivos ou de conveniência na seleção dos elementos da amostra, o que pode introduzir vieses e limitar a generalização dos resultados (Morettin; Bussab, 2006).

Neste trabalho, será adotada a amostragem probabilística, com ênfase na amostragem estratificada proporcional. Essa escolha visa assegurar que os dados coletados sejam representativos da população estudada, permitindo inferências robustas e confiáveis para os objetivos específicos do estudo estatístico em questão.

Os estimadores são ferramentas fundamentais na estatística para inferir características da população a partir da amostra. Dois dos principais estimadores são a média e a variância. A média amostral $(\bar{x})$ é um estimador não viesado da média populacional (μ), o que significa que, em média, a média amostral é igual à média da população. A fórmula para a média amostral é:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2.21}$$

em que x_i são os valores da amostra e **n** é o tamanho da amostra. A variância amostral (S^2) é um estimador da variância populacional (σ^2). A variância amostral é calculada como:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{2.22}$$

em que x_i são os valores da amostra, $\bar{x}$ é a média amostral e **n** é o tamanho da amostra. A variância amostral é um estimador não viesado da variância populacional, o que significa que, em média, a variância amostral é igual à variância da população.

2.2.5 Espaço Métrico e Distância de Correlação

É possível definir uma distância entre as variáveis a partir da correlação, para isso, devemos definir o que é um espaço métrico, de maneira que poderemos encontrar uma distância de correlação utilizando o espaço métrico euclidiano, .

Segundo Berni (2021), um espaço métrico é um par (M,d), onde M é um conjunto qualquer e d é uma função, chamada de métrica ou norma, que dado um conjunto M:

$$\begin{aligned} d : M\ X\ M &\rightarrow \mathbb{R}_+ \\ (x_i, x_j) &\rightarrow d(x_i, x_j) \end{aligned}$$

E satfifaz os seguintes axiomas, $\forall\ x_i,\ x_j,\ x_z\ \in\ M$:

- Propriedade positividade definida: $Se\ d(x_i,x_j) = 0 \Leftrightarrow x_i\ = x_j$
- Propriedade simétrica: $d(x_i,x_j) = d(x_j,x_i)$
- Desigualdade Triangular: $d(x_i,x_z) \leq d(x_i,x_j) + d(x_j,x_z)$

Com isso, podemos definir uma distância de correlação, a partir da distância euclidiana e verificar, se o espaço métrico encontrado satisfaz os axiomas para ser utilizado, assim a métrica do espaço euclidiano é definida como $d(x,y) : \mathbb{R}^n\ X\ \mathbb{R}^n \rightarrow \mathbb{R}$:

$$d(\vec{x},\vec{y}) = \sqrt{\sum_{i=1}^{n}(x_i, -y_i)^2} \tag{2.23}$$

Assim, definimos os estimadores da média ($\bar{x}\ e\ \bar{y}$) e da variância ($S_x\ e\ S_y$), sendo n o número de variáveis, então:

$$\begin{aligned} \bar{x} &= \frac{1}{n}\sum_i x_i, \qquad {S_x}^2 = \frac{1}{n}\sum_i (x_i - \bar{x}) \\ \bar{y} &= \frac{1}{n}\sum_i y_i, \qquad {S_y}^2 = \frac{1}{n}\sum_i (y_i - \bar{y}) \end{aligned}$$

Logo a covariância e o coeficiente de correlação mostral são dados, respectivamente, por:

$$cov(x,y) = \frac{1}{n}\sum_i (x_i - \bar{x})(y_i - \bar{y})\ \text{ e }\ \rho(x,y) = \frac{cov(x,y)}{S_x S_y}$$

Agora, vamos definir duas variáveis padronizadas:

$$z_i = \frac{x_i - \overline{x}}{\sqrt{n}S_x} \quad \text{e} \quad k_i = \frac{y_i - \overline{y}}{\sqrt{n}S_y}$$

Pode-se verificar acima que a padronização, cria variáveis com esperança ou média igual a 0 e variância igual a 1, dessa forma, as variáveis ficam independente da escala. Assim, z_i *e* k_i podem ser interpretadas como vetores unitários, logo a distância euclidiana entre dois vetores unitários:

$$d^2(x,y) = \sum_i (z_i - k_i)^2$$
$$d^2(x,y) = \sum_i (z_i^2 - 2z_iK_i + k_i^2)$$
$$d^2(x,y) = \sum_i z_i^2 + \sum_i + k_i^2 - 2\sum_i z_i k_i$$
$$d^2(x,y) = \sum_i z_i^2 + \sum_i + k_i^2 - 2\sum_i z_i k_i$$

Devemos calcular: $\sum_i z_i^2$ *e* $\sum_i k_i^2$. Assim:

$$\sum_i z_i^2 = \frac{1}{n\ S_x^2}\sum_i (x_i - \overline{x})^2 = \frac{1}{S_x^2}[\frac{1}{n}\sum_i (x_i - \overline{x})^2] = \frac{1}{S_x^2}S_x^2 = 1$$

O mesmo se aplica para $\sum_i k_i^2$, de maneira similar podemos verificar que $\sum_i z_i = \sum_i k_i = 0$. Logo, temos que:

$$d^2(x,y) = 2(1 - \sum_i z_i k_i)$$

Desenvolvendo $\sum_i z_i k_i$, temos:

$$\sum_i z_i k_i = \sum_i (\frac{x_i - \overline{x}}{\sqrt{n}S_x})(\frac{y_i - \overline{x}}{\sqrt{n}S_y}) = \frac{1}{S_x S_y}\frac{1}{n}\sum_i (x_i - \overline{x})(y_i - \overline{y}) = \frac{cov(x,y)}{S_x S_y}$$

Substituindo na equação, temos:

$$d(x,y) = \sqrt{2(1 - \rho(x,y))} \tag{2.24}$$

Assim, podemos verificar que o valor da distância de correlação varia entre 0 e 2, uma vez que se $\rho(x,y) = -1$, valor mínimo, implica que $d(x,y) = 2$ e se $\rho(x,y) = 1$, valor máximo, implica que $d(x,y) = 0$

3 ENEM E MINERAÇÃO DE DADOS EDUCACIONAIS

Neste capítulo será abordardado sobre a principal prova para avaliar o ensino médio no Brasil, o Exame Nacional o Exame Nacional do Ensino Médio (ENEM), na seção 3.1 será descrito a origem e o desenvolvimento do ENEM desde sua criação em 1998, destacando sua transformação de uma ferramenta de avaliação do desempenho dos estudantes em uma ferramenta de inclusão social e um critério de seleção para universidades públicas, bem como as mudanças políticas e educacionais que moldaram seu papel. Na seção 3.2, é abordada a Matriz de Referência detalha a estrutura do exame, incluindo suas áreas de conhecimento e a evolução das competências e habilidades avaliadas ao longo do tempo, refletindo a mudança de foco de conteúdos específicos para uma abordagem mais prática e contextualizada. Em seguida na seção 3.3, é explicado a Estrutura do Exame explica as modificações no formato do ENEM, como a divisão em dois dias, além de abordar a Teoria de Resposta ao Item (TRI), por último na seção 3.4, será destinado a Mineração de Dados educacionais, mostrando sua origem e sua importância, além das áreas utilizadas neste trabalho

3.1 Exame Nacional do Ensino Médio (ENEM)

A política educacional brasileira tem suas raízes na Constituição Federal de 1988 (Brasil, 1988), que estabeleceu a educação como um direito social fundamental. Segundo Brasil (1988) o artigo 205, a educação é um direito de todos e um dever do Estado e da família, promovida e incentivada com a colaboração da sociedade, visando ao pleno desenvolvimento da pessoa, seu preparo para o exercício da cidadania e sua qualificação para o trabalho. A partir deste marco, a legislação educacional do Brasil passou por diversas reformas e a criação de políticas públicas para garantir acesso à educação de qualidade para todos os cidadãos, independentemente de sua origem socioeconômica, etnia, gênero ou qualquer outra condição.

Uma das principais legislações que amparam o ensino de qualidade no Brasil é a Lei de Diretrizes e Bases da Educação Nacional (LDB), instituída pela Lei nº 9.394, de 1996. Assim, segundo Brasil (1996), a LDB estabelece os princípios, objetivos e normas gerais para a organização do sistema educacional brasileiro, abrangendo desde a educação infantil até o ensino superior. Ela enfatiza a importância da equidade e da qualidade na educação, determinando que os currículos escolares contemplem uma formação integral, promovendo o desenvolvimento das competências e habilidades necessárias para a vida em sociedade. Além da LDB, outro marco

significativo na política educacional brasileira é o Plano Nacional de Educação (PNE), atualizado pela Lei nº 13.005, de 2014. Segundo Brasil (2014), o PNE estabelece diretrizes, metas e estratégias para a política educacional a serem cumpridas em um período de dez anos. Entre suas metas, destacam-se a universalização do ensino fundamental e médio, a ampliação do acesso à educação infantil e ao ensino superior, a melhoria da qualidade do ensino, e a valorização dos profissionais da educação. O PNE serve como um guia estratégico para os governos federal, estadual e municipal, além de orientar o financiamento da educação no país. Outra medida de suma importância para educação brasileira é o Fundo de Manutenção e Desenvolvimento da Educação Básica e de Valorização dos Profissionais da Educação (FUNDEB) é outra importante política de financiamento educacional. Estabelecido inicialmente em 2007 e recentemente prorrogado e ampliado pela Emenda Constitucional nº 108 de 2020, o Fundeb assegura recursos financeiros destinados à manutenção e desenvolvimento da educação básica pública, contribuindo para a redução das desigualdades regionais e a promoção da equidade. Além desses instrumentos legais e financeiros, a política educacional brasileira tem buscado incorporar temas como a inclusão, a diversidade e a sustentabilidade. A Lei Brasileira de Inclusão da Pessoa com Deficiência (Lei nº 13.146, de 2015) é um exemplo, garantindo o acesso e a permanência de estudantes com deficiência nas escolas regulares, promovendo a educação inclusiva. Já as Diretrizes Curriculares Nacionais incentivam a inserção de conteúdos relacionados à educação ambiental, aos direitos humanos e à diversidade étnico-racial.

Segundo Silva M. R.;Ribeiro (2008), a década de 1990 e o início dos anos 2000, portanto, marcaram um período de redefinição e expansão das políticas educacionais brasileiras, com um foco crescente na avaliação e na qualidade do ensino, com início na Conferência Mundial de Educação para Todos, em 1990 realizada na Tailândia, que estabeleceu diretrizes para os países com os piores indicadores educacionais do mundo. O contexto político e econômico da época, caracterizado por reformas no Estado e por uma busca por maior eficiência e equidade no setor público, além dos altos índices de analfabetismo, evasão escolar e repetência (Silva M.R.; Ribeiro, 2008), ligado aos planos determinados na conferência Mundial de Educação para Todos, influenciou diretamente a criação do Exame Nacional do Ensino Médio (ENEM) como um instrumento essencial na análise dos indicadores educacionais. Além disso, o exame passou a desempenhar um papel importante na inclusão social, proporcionando acesso à educação superior a segmentos da população historicamente excluídos desse nível de ensino.

Segundo Silveira F. L. (2015) o ENEM foi criado em 1998 pelo Ministério Da

Educação (MEC) durante o governo do presidente Fernando Henrique Cardoso. Inicialmente, o exame tinha o objetivo principal de avaliar a qualidade do ensino médio no Brasil, servindo como uma ferramenta de diagnóstico das competências e habilidades adquiridas pelos estudantes ao longo dessa etapa da educação básica. Para Oliveira (2016), a criação do ENEM no contexto da avaliação educacional ocorreu em um período de intensas transformações na política educacional brasileira, marcadas por um esforço de universalização do ensino fundamental e pela tentativa de melhorar a qualidade da educação pública no país, através dos resultados obtidos nos exames. Na década de 1990, o Brasil enfrentava desafios significativos relacionados à desigualdade de acesso à educação e à necessidade de modernização do sistema educacional, visando preparar os jovens para um mercado de trabalho em rápida transformação devido ao avanço da globalização e das tecnologias. A inspiração para a criação do ENEM veio de experiências internacionais, como o Scholastic Aptitude Test (SAT) dos Estados Unidos e outros modelos europeus de avaliação que buscavam ir além da simples memorização de conteúdos e estavam mais voltados para a avaliação de competências amplas. Para Garcia F. M. (2021) no Brasil, havia um reconhecimento crescente da necessidade de desenvolver um sistema de avaliação mais alinhado com as demandas contemporâneas da sociedade e do mercado de trabalho, que exigiam habilidades críticas, analíticas e uma visão integrada do conhecimento, de maneira que se pudesse analisar a qualidade das escolas, assim visualizando todo o panorama educacional, desde a relação aluno professor, até a propria gestão escolar.

Inicialmente, o ENEM tinha caráter voluntário e servia como um referencial para estudantes e escolas. No entanto, sua relevância começou a aumentar rapidamente, em parte devido à introdução do Programa Universidade para Todos (PROUNI) em 2004, que passou a utilizar as notas do ENEM como critério de seleção para bolsas de estudo em instituições privadas de ensino superior. Com isso, o exame passou a ganhar um caráter mais competitivo e uma importância estratégica para os alunos que buscavam uma vaga no ensino superior. De acordo com Silveira F. L. (2015) em 2009, o ENEM sofreu uma reformulação significativa, transformando-se em um dos principais instrumentos de seleção para o ingresso em universidades públicas brasileiras, através do Sistema de Seleção Unificada (SISU). Essa mudança foi um marco na democratização do acesso ao ensino superior, pois permitiu que estudantes de diferentes regiões do país tivessem oportunidades iguais de concorrer a vagas em instituições públicas, independentemente de sua localização geográfica. Segundo Silvana M. (2014), a prova também passou a ser um dos critérios para obtenção de financiamento estudantil pelo Fundo de

Financiamento Estudantil Do Ensino Superior (FIES).

O ENEM, desde sua criação, tem sido um dos pilares das políticas de acesso à educação no Brasil, refletindo um compromisso contínuo com a inclusão e a melhoria da qualidade educacional. A partir de sua expansão e consolidação, tornou-se um dos maiores exames de ingresso universitário do mundo, com milhões de estudantes participando anualmente, o que demonstra sua importância e relevância no cenário educacional brasileiro.

3.1.1 Matriz de Referência

De acordo com Inep (2009b), Inep (2009, p.8), a Matriz de Referência do Exame Nacional do Ensino Médio (ENEM) é um documento essencial que orienta a elaboração das questões do exame e a preparação dos estudantes. Desenvolvida pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), a matriz é dividida em áreas de conhecimento específicas: Linguagens, Códigos e suas Tecnologias; Ciências Humanas e suas Tecnologias; Ciências da Natureza e suas Tecnologias; e Matemática e suas Tecnologias. Cada área de conhecimento é estruturada em torno de competências e habilidades. As competências são conjuntos de conhecimentos, habilidades, atitudes e valores que os estudantes devem desenvolver. Já as habilidades são ações específicas que demonstram a aplicação prática dessas competências. Por exemplo, na área de Ciências da Natureza e suas Tecnologias, uma das competências é "Compreender as ciências naturais e as tecnologias a elas associadas como construções humanas, percebendo seus papéis nos processos de produção e no desenvolvimento econômico e social da humanidade.", e uma das habilidades associadas é "Reconhecer características ou propriedades de fenômenos ondulatórios ou oscilatórios, relacionando-os a seus usos em diferentes contextos." (Inep, 2009, p.8), pode-se verificar as competências e habilidades da prova de ciências da natureza, no apêndice tal.

Desde a criação do ENEM em 1998, a matriz de referência passou por diversas atualizações para se adaptar às mudanças educacionais e às necessidades dos estudantes. Inicialmente, o ENEM tinha como objetivo principal avaliar o desempenho dos estudantes ao final da educação básica e servir como um instrumento de autoavaliação. Com o tempo, o exame ganhou importância e passou a ser utilizado como critério de seleção para o ingresso em instituições de ensino superior, através do Sistema de Seleção Unificada (SISU), do Programa Universidade para Todos (ProUni) e do Fundo de Financiamento Estudantil (FIES), como visto anteriormente. É perceptível que a evolução da matriz de referência reflete essas mudanças. De acordo com

Inep (2018), as primeiras versões da matriz focavam mais na avaliação de conteúdos específicos, enquanto as versões mais recentes enfatizam a aplicação prática desses conhecimentos em situações do cotidiano. Isso se alinha com a abordagem interdisciplinar e contextualizada do conhecimento, que é uma característica marcante do ENEM. A Lei de Diretrizes e Bases da Educação Nacional (LDB), Lei nº 9.394, de 20 de dezembro de 1996, também desempenhou um papel crucial na definição das diretrizes educacionais no Brasil, influenciando diretamente a estrutura e os objetivos do ENEM.

A partir de Inep (2018), Inep (2014), comparando a matriz de referência original com a atual, observa-se uma ampliação das competências e habilidades avaliadas. Por exemplo, na área de Ciências Humanas e suas Tecnologias, a matriz atual inclui competências relacionadas à compreensão e análise de fenômenos sociais, políticos, econômicos e culturais, enquanto as versões anteriores eram mais restritas a conteúdos históricos e geográficos. Além disso, a matriz atual incorpora uma maior diversidade de habilidades, como a capacidade de interpretar gráficos e tabelas, resolver problemas matemáticos aplicados a situações reais e compreender a importância da produção cultural em diferentes contextos. Essas mudanças visam preparar os estudantes não apenas para o ingresso no ensino superior, mas também para a vida em sociedade, promovendo o desenvolvimento de um pensamento crítico e analítico.

Em resumo, a Matriz de Referência do ENEM evoluiu significativamente desde a sua criação, refletindo as mudanças nas políticas educacionais e nas necessidades dos estudantes. A matriz atual é mais abrangente e contextualizada, enfatizando a aplicação prática dos conhecimentos e o desenvolvimento de competências e habilidades essenciais para a vida acadêmica e profissional.

3.1.2 Estrutura do Exame

O Exame Nacional do Ensino Médio (ENEM), criado em 1998 pelo Ministério da Educação (MEC), teve um impacto significativo na avaliação educacional brasileira. Inicialmente, o exame visava fornecer uma certificação de conclusão do ensino médio e avaliar o desempenho dos alunos ao final desse ciclo escolar. A estrutura e a aplicação do ENEM passaram por várias mudanças desde sua criação, refletindo o desenvolvimento das políticas educacionais e as necessidades emergentes dos candidatos.

Durante seus primeiros anos, o ENEM era realizado em um único domingo, e a prova compreendia 63 questões objetivas e uma redação. Este formato inicial, embora funcional,

apresentava algumas limitações. De Inep (1998), a carga de trabalho para os candidatos em um único dia era considerável, o que muitas vezes resultava em cansaço e, consequentemente, em um impacto negativo no desempenho dos participantes. Além disso, a estrutura inicial não permitia uma avaliação detalhada e abrangente das competências dos candidatos, o que motivou mudanças significativas no formato do exame.

Em 2009, o ENEM passou por uma reforma significativa que alterou sua estrutura e formato e que passou a ser chamado de Novo Enem. Além da mudança na matriz de referência, outra mudança foi a divisão da prova em dois dias consecutivos, um sábado e um domingo, essa alteração teve como objetivo aliviar a carga horária do exame e proporcionar uma avaliação mais equilibrada e eficaz. De acordo com Inep (2009a) no primeiro dia, os candidatos realizam as provas de Ciências da Natureza e suas Tecnologias e Ciências Humanas e suas Tecnologias, enquanto no segundo dia, são aplicadas as provas de Linguagens, Códigos e suas Tecnologias, mais a Redação, e Matemática e suas Tecnologias. No primeiro dia de prova possui uma duração de 4 horas e 30 minutos, enquanto o segundo dia com duração de 5 horas e 30 minutos por conta da redação. Segundo Mec (2017), em 2017, houve outra mudança na prova do Enem, as provas passaram a ser aplicadas em dois domingos consecutivos e a redação passou a ser no primeiro dia junto com as provas de Linguagens e Códigos e suas Tecnologias e Ciências Humanas e suas Tecnologias, assim alterando o tempo de prova para ter 1 hora a mais no primeiro dia.

A mudança para dois domingos consecutivos trouxe vários benefícios. Reduzir a quantidade de questões por dia (mudança de 2009) ajudou a diminuir a fadiga dos candidatos, permitindo que eles se concentrassem melhor em cada seção do exame. Além disso, essa divisão melhorou a logística do exame, facilitando a aplicação e a correção das provas. A nova estrutura (2009) também proporcionou uma avaliação mais justa e detalhada das competências dos participantes, refletindo um compromisso com a qualidade da avaliação. Assim, a decisão de realizar o ENEM aos domingos, mesmo após a reforma, está relacionada à necessidade de minimizar conflitos com o calendário escolar e outras atividades dos candidatos. Os domingos foram escolhidos para garantir a disponibilidade dos alunos e evitar a sobreposição com o horário escolar regular. Isso também ajuda a garantir que o exame não interfira nas atividades acadêmicas dos participantes e que todos tenham a oportunidade de realizar a prova em condições adequadas

Outro aspecto importante da reforma, de acordo com Mec (2013) foi a introdução da Teorema da Resposta ao Item (TRI), um modelo estatístico que avalia a habilidade dos candidatos com base na dificuldade das questões e no padrão de respostas. A TRI permite uma avaliação

mais precisa do desempenho dos participantes, levando em consideração a dificuldade das questões e a probabilidade de acerto de acordo com o nível de conhecimento do candidato. Esse modelo representa uma inovação significativa no sistema de avaliação, aprimorando a precisão e a justiça dos resultados. Essa abordagem estatística substituiu métodos tradicionais e trouxe uma nova perspectiva probabilística para a correção das provas. Segundo Mec (2011), a TRI é fundamental para entender como a dificuldade das questões e o desempenho dos candidatos são avaliados de forma mais precisa e justa, onde baseia-se em modelos estatísticos que consideram três parâmetros principais: dificuldade, discriminação e acerto casual, o parâmetro de dificuldade, refere-se ao nível de habilidade necessário para que um candidato responda corretamente a uma questão, o parâmetro de discriminação, mede a capacidade da questão em distinguir entre candidatos com diferentes níveis de habilidade, o parâmetro de acerto casual, representa a probabilidade de um candidato acertar a questão por sorte, sem possuir o conhecimento necessário.

Desde sua criação, o ENEM tem se adaptado às mudanças no cenário educacional e às necessidades dos candidatos, refletindo um compromisso contínuo com a melhoria da avaliação educacional. A mudança para dois domingos consecutivos e a introdução da TRI são exemplos claros desse esforço para aprimorar o exame e garantir uma experiência mais justa e eficaz para todos os participantes. Essas transformações demonstram a evolução do ENEM de uma avaliação básica para uma ferramenta complexa e abrangente, voltada para a promoção da qualidade educacional no Brasil.

3.2 Mineração de dados Educacionais

De acordo com Guimarães Júnior *et al.* (2023), Baker e Yacef (2009), a Mineração de Dados Educacionais (Educational Data Mining (EDM)) é uma área emergente no campo da análise de dados e ciência da computação, focada na extração de padrões úteis e informativos a partir de dados coletados em ambientes educacionais, de maneira que utiliza-se as análises para a melhoria da qualidade de ensino e aprendizagem. Esses dados podem incluir registros acadêmicos, interações em plataformas de ensino online, avaliações de desempenho, e até mesmo dados comportamentais dos estudantes, como participação em aulas e uso de recursos digitais. O objetivo principal da mineração de dados educacionais é identificar *insights* que possam melhorar os processos de ensino e aprendizagem, personalizar a educação para atender melhor às necessidades individuais dos alunos, e apoiar a tomada de decisões baseada em evidências por

parte de educadores e gestores escolares.

Segundo Baker *et al.* (2011) a área de EDM vem desde o ano de 2004, com uma sequência de workshops que culminaram em 2008 na Conferência Internacional sobre Mineração de Dados (International Conference on Educational Data Mining)e desde então o número de pesquisa e artigos aumentaram, além da criação de uma revista em 2009 e de dois livros sobre o assunto um em 2006 (Data Mining in e-learning) e outro em 2010 (Handbook of Educational Data Mining). Assim, utilizar técnicas de mineração de dados em dados educacionais, como algoritmos de aprendizado de máquina, análise estatística, e visualização de dados, é possível identificar tendências e padrões que não seriam evidentes em análises tradicionais. Por exemplo, com o estudo de Romero C.; Ventura (2017) pode-se prever quais estudantes estão em risco de evadir da escola, em relação ao trabalho de Riggan M.; Chen (2018) pode-se identificar alunos em risco de reprovação, assim permintindo a intervenção da escola no auxílio aos estudantes. Dessa forma, a EDM pode ser utilizada para avaliar a eficácia de políticas educacionais e currículos, permitindo ajustes que visam à melhoria contínua da qualidade da educação.

Para Baker *et al.* (2011), as principais sub-áreas da Mineração de Dados educacionais, seguem a seguinte taxonomia:

Predição englobando os algoritmos de classificação, regressão e estimação de densidade. A área de **Agrupamento** (*clustering*), também tem-se a área de **Mineração de Relações** que engloba a Mineração de Regras de Associação, Mineração de Correlações, Mineração de Padrões Sequênciais, Mineração de Causas, e por último as áreas de **Destilação de dados para facilitar decisões** e **Descobrimento de Modelos**.

Para este trabalho será considerada duas áreas, a de mineração de correlações e de agrupamento, dessa maneira, para Baker *et al.* (2011), a área de agrupamento ou clustering é a de achar dados que se agrupem e formem grupos similares e assim possam ser separados em categorias, de maneira que as regiões de agrupamento reflitam sobre informações inerentes aos dados, como exemplo verificar regiões de agrupamentos entre alunos devido a fatores que podem variar desde escola até a gestão escolar, região de moradia ou notas. Enquanto a Mineração de correlações, busca achar linearidades entre um conjunto de variáveis, assim pode-se achar correlações positivas, negativas ou netras, como visto na seção 2.2.

Com o avanço da tecnologia e a crescente digitalização das práticas educacionais, o volume de dados disponível está aumentando rapidamente, tornando a mineração de dados educacionais uma ferramenta essencial para lidar com esses recursos informacionais de maneira

eficiente e produtiva. Essa área, portanto, oferece um potencial significativo para transformar a educação, proporcionando uma base mais sólida para intervenções educacionais, promovendo o sucesso acadêmico e preparando melhor os estudantes para os desafios do futuro, além de ser um excelente suporte aos gestores escolares e docentes. Dessa forma, no contexto deste trabalho iremos verificar se há existência de correlações positivas ou negativas entre os concluintes do Ensino Médio do Estado do Ceará que realizaram a prova do Enem, com base no vetor de respostas dos estudantes e dessa forma clusterizar a partir da distância de correlação e o algoritmo de Prim, a rede de alunos analisados e inferir sobre os resultados.

4 METODOLOGIA E PROCESSO DE ANÁLISE

4.1 Obtenção e Base de dados

Inicialmente, serão definidos os aspectos do sistema abordado neste trabalho, que se baseia na análise dos microdados do ENEM. O objetivo é identificar correlações positivas entre os participantes. Na próxima seção, serão detalhados os dados utilizados e a forma como as correlações foram calculadas. A análise focará em identificar clusters de alta correlação para obter informações sobre os participantes e suas interações. Portanto, o sistema em questão é o ENEM, e os membros são os participantes do exame.

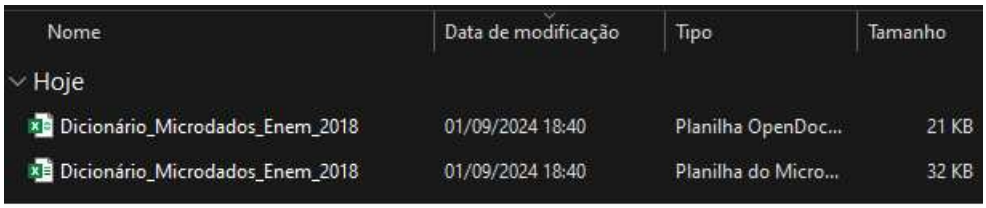
A base de dados foi obtida no site oficial do governo Inep (2024). Os microdados disponíveis cobrem desde o primeiro ENEM em 1998 até o ENEM do ano anterior a este trabalho, 2023. Após o download, um arquivo .zip é gerado. Ao descompactá-lo, cria-se uma pasta contendo cinco subpastas, conforme ilustrado na figura abaixo. Para este trabalho os dados de 2018 e 2019 foram obtidos no site da plataforma Kaggle (2024).

Figura 11 – Pastas de Arquivos dos Dados

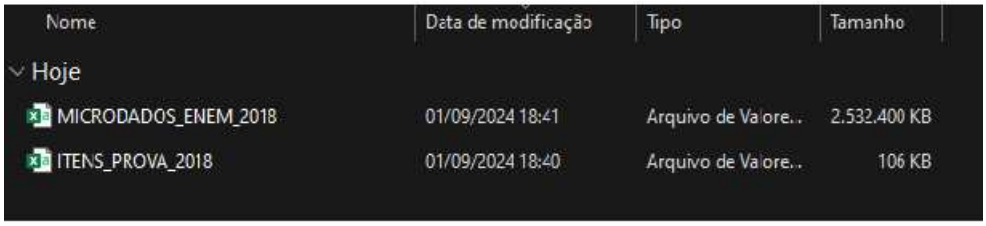
Fonte: Elaborado pelo Autor

Para esta análise, foram utilizadas duas pastas específicas. A primeira é a pasta "dados", que inclui dois arquivos .csv. Um arquivo contém os microdados para análise, cuja visualização está disponível no Anexo A, e o outro apresenta os itens das provas, detalhando a posição de cada código de prova e a habilidade necessária, como descrito no capítulo anterior. A segunda é a pasta "dicionário", que contém dois arquivos com o dicionário dos microdados, um em formato .xlsx e outro em .odp. Os microdados servem como a base da análise, onde cada linha representa um participante e cada coluna oferece informações relevantes, como local de prova, notas e questionário socioeconômico. A análise inicial será focada no vetor de respostas dos participantes, que são variáveis categóricas, para cada prova, assumindo que candidatos com boas notas terão vetores de respostas semelhantes. Isso permitirá identificar os alunos com maior correlação.

Figura 12 – Pastas utilizadas no Trabalho



a)



b)

Fonte: Elaborado pelo Autor. a) Arquivos do Dicionário. b) Arquivos contendo os Microdados e as informações dos itens das provas.

É importante destacar que, segundo Okbr (2024), a partir de 2022, houve uma redução na transparência das informações devido à Lei de Proteção de Dados Pessoais. Essa mudança afetou não apenas os dados do ENEM, mas também várias bases de dados educacionais, resultando na remoção de informações como o código das escolas dos participantes. Por essa razão, o trabalho utilizará os dados do ENEM de 2018 e 2019, que ainda contêm os códigos INEP das escolas.

4.2 Proposta da Análise

Partindo da premissa, que a rede complexa é o ENEM e os membros são os participantes, podemos obter a partir do processo visto na subsecção anterior os vetores de resposta dos alunos e assim ordernar os itens tomando uma cor de prova como base e verificar se há correlação de aluno para aluno, uma vez que os alunos que obtiveram as melhores notas, terão um vetor de respostas parecidos, assim terão correlação positivas. Assim é esperado que haja um número maior de correlações negativas ou não correlação entre os alunos, uma vez que a variação de opção de questões marcadas diferentes é maior que a de marcadas iguais . O mesmo processo será aplicado para verificar correlação prova por prova, uma vez que pretende-se observar questões mais correlacionadas e verificar os motivos das correlações.

4.3 Mineração dos Dados

Toda a análise foi realizada utilizando a linguaguem de programação Python, assim podemos dividir o processos e manipulação e tratamento dos dados a partir de alguns passos, que estão descritos logo abaixo. Além disso, o foco da análise são os alunos concluintes do ensino médio no estado do Ceará, uma vez que, são esses os alunos que pleiteiam as vagas nas universidades de ensino superior. Todo o processo de análise está no apêndice A.

Passo 1: Upload da base de dados e tratamento das informações.

Em primeira etapa, foi feito o download da base de dados para o ambiente de programação em Python, em seguida, através da coluna de local da escola e de situação de conclusão, seleciona-se os alunos do estado do Ceará e que estão concluindo o ensino médio, por fim, faz-se uma limpeza inicial nos dados, a partir da coluna de presença no dia da prova, exlcui-se os alunos que não foram aos dois dias de prova, em relação ao vetor de respostas, foram excluídos os participantes que marcaram somente uma única opção no gabarito nos dois dias, ou então que marcaram uma única opção em cada dia e os alunos que em relação as 180 questões tiverem mais de 20 questões com erros na leitura que são apontadas, como dupla marcação ou resposta em branco. Por fim, foram selecionados somente os alunos em que os códigos das provas foram somente de primeira aplicação e presencial (provas de cores rosa, amarela, branco ou cinza e azul).

Passo 2: Reordenamento das questões e criação da tabela de respostas.

A segunda etapa é iniciada removendo dos vetores de repostas, as questões pertinentes a lingua estrangeira, pois há duas opções de linguas, de maneira que não seria possível verificar a correlação de questões diferentes, em seguida foi realizado a reordenação das questões dos vetores de repostas, uma vez que, a ordem das questões das provas diferem em cada código de prova, logo o vetor de respostas entre os alunos de código de provas diferentes não trazem a mesma ordem das questões. Assim, tomando como base a prova de código de cor azul, foi verificado as questões correspondentes em cada código de prova selecionado, de maneira que foi criado um script que faz essa reordenação, colocando os vetores de respostas dos participantes, todos na mesma ordem, com base na prova de código azul.

Passo 3: Matriz de reposta dos alunos

Nesta etapa, tem-se o vetor de respostas dos alunos ordenados, agora é realizado a mudança das opções marcads para variáveis aleatórias de maneira que será possível calcular a correlação, que pode ser vista, logo abaixo.

$$\begin{cases} -1, \ se \text{ ``.''} \ (Resposta \ em \ Branco) \\ 0, \ se \text{ ``}*\text{''} \ (Dupla \ Marcacao) \\ 1, \ se \ item \ marcado \ for \ A \\ 2, \ se \ item \ marcado \ for \ B \\ 3, \ se \ item \ marcado \ for \ C \\ 4, \ se \ item \ marcado \ for \ D \\ 5, \ se \ item \ marcado \ for \ E \end{cases}$$

Dessa forma, obtemos uma matriz m x n, onde m é o número de questões analisadas e n é o número de participantes, com a seguinte configuração:

Figura 13 – Matriz de Resposta dos Alunos Ordenada

	0	1	2	3	4	5	6	7	8	9	...	84983	84984	84985	84986	84987	84988	84989	84990	84991	84992
0	2	5	5	1	4	1	4	5	1	1	...	1	2	4	1	3	5	2	2	3	3
1	3	3	1	3	1	5	3	2	3	2	...	2	4	4	5	1	1	1	3	4	4
2	5	2	2	4	4	4	4	4	4	4	...	4	4	1	4	1	3	4	4	3	1
3	3	1	3	1	1	5	3	2	2	3	...	3	2	3	1	2	3	3	1	3	1
4	5	2	4	1	5	4	2	1	1	5	...	3	1	2	3	1	3	1	1	4	4
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
170	1	2	2	2	2	2	2	4	2	2	...	4	4	2	2	1	0	4	4	5	1
171	2	2	2	5	1	4	3	5	2	1	...	2	2	3	1	2	0	1	1	4	3
172	2	2	2	1	3	2	2	5	3	4	...	3	1	4	2	1	0	4	5	2	4
173	2	3	5	3	1	4	4	4	3	5	...	3	4	5	2	2	0	5	1	3	4
174	5	2	3	2	4	5	2	1	3	2	...	4	4	1	3	3	0	1	4	4	3

175 rows × 84993 columns

Fonte: Elaborado pelo Autor.

Passo 4: Amostragem.

Uma vez, que a quantidade de alunos para traçar a correlação é muito alta, surge um problema de alocação dos dados na memória do computador, o que impede de calcular a matriz de correlação, nos casos analisados, a quantidade de alunos gira em torno de 75 mil alunos, assim, fazemos uma amostragem com 25 mil participantes, cerca de um terço da população. De maneira a obter uma amostra representativa da população, foi realizada cem uma amostragem estratificada proporcional em relação ao número de participantes por escola, nos casos em que tem-se as escolas dos participantes e por município, quando não há escola dos participantes

na base de dados, assim, foi gerada cem amostras e escolhida a que obteve valores de média e desvio padrão amostral mais próximos da média populacional e do desvio padrão populacional, com base na média da nota das quatro áreas realizadas, para essa amostragem foi escrito um código.

Passo 5: Cálculo da Matriz de Correlação, Distância e MST.

Após realizada a amostragem, foi calculado a matriz de correlação, aluno por aluno e em seguida a matriz de distância de correlação, que é a matriz de adjacência do grafo da rede analisada, dessa forma, obtemos a distância entre os alunos, com isso, podemos obter a MST, e verificar os alunos mais correlacionados e realizar análises a partir de suas informações na base de dados. Também foi calculada a matriz de correlação entre as questões, realizando somente uma transposição na matriz de respostas dos alunos, organizando as linhas como os alunos as colunas como as questões das provas.

Passo 6: Análise das Variáveis Sócio-Econômicas e da Média de Notas

Por fim, com a ordem da MST, podemos relacionar com a média de notas e com as variáveis sócio-econômico dentro da matriz ordenada pela MST, pega-se a nota do aluno i com a nota do aluno j e divide-se por dois, a diagonal principal é a nota do candidato, assim pode-se verificar as regiões com notas mais altas e baixas. Em relação as variáveis sócio-econômicas, foi realizado o seguinte, verifica a variável do aluno i e a variável do aluno j, se for a mesma é pintada a região de acordo com a cor escolhida para variável,

5 RESULTADOS

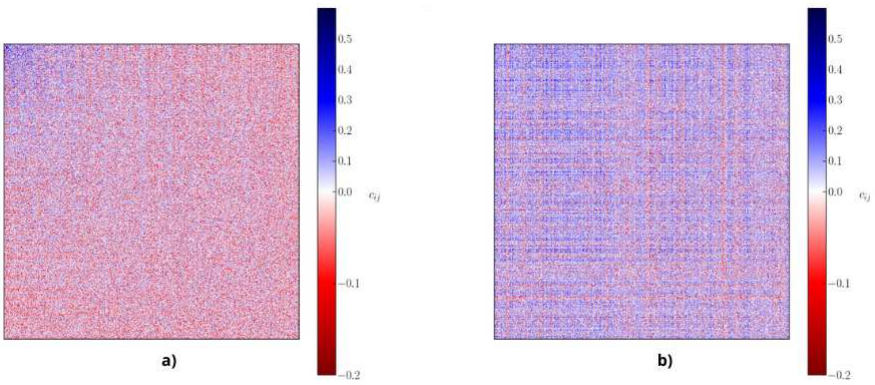
Neste capítulo será abordado as considerações em relação aos resultados, verificando a influência das variáveis, como escolaridade da mãe, escolaridade do pai, renda, raça, sexo, nota, tipo de escola e acesso a internet. Dessa maneira, depois de obtida a matriz de correlação e utilizar o algoritmo de Prim, para obter a MST, verifica-se alguns agrupamentos (clusters), com isso analisar a natureza dessas correlações entre os alunos é de suma importância para entender sobre a qualidade da educação no momento, além de servir como suporte na tomada de decisão. Os resultados obtidos serão mostrados de duas maneiras, através de gráficos como heatmap e na perspectiva de um grafo, aonde a cor dos nós representa as variáveis.

Na primeira Secção será apresentado a matriz de correlação dos anos de 2018 e 2019, a matriz de correlação reordenada pela MST e verificar a influência da nota na correlação. Em seguinda da Secção 5.2, será analisado a influência das variáveis mencionadas anteriormente e discutido apesctos relevantes e esperados dentre os clusters de alunos com notas acima de 500. Na Secção 5.3 é analisado a correlação entre as questões do exame com excecção das questões de lingua estrangeira. E por fim, será mostrado a visualização das variáveis através da rede complexa.

5.1 Resultados obtidos com a Matriz de correlação e MST

Inicialmente, obtemos uma matriz de correlação com a ordem aleatória, devido a amostragem, assim, para os anos de 2018 e 2019, temos as seguintes Matrizes de correlação Aluno por Aluno:

Figura 14 – Matriz de Correlação Aluno por Aluno

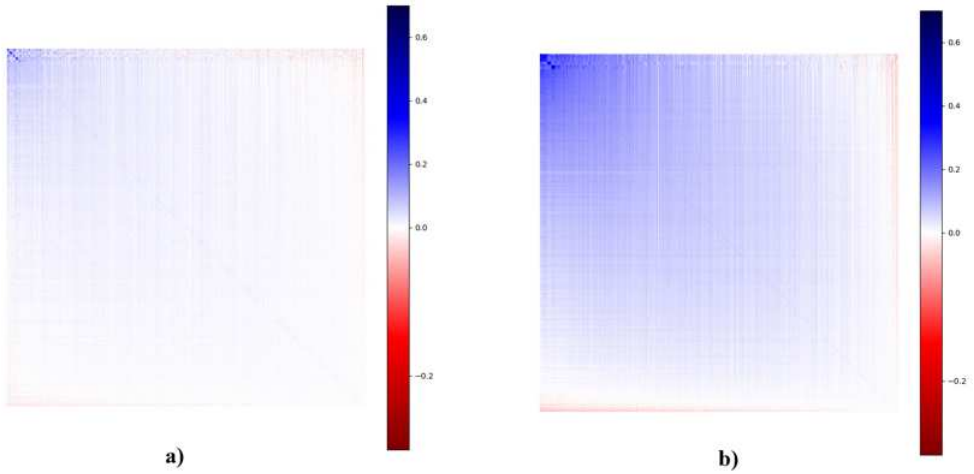


Fonte: Elaborado pelo Autor. a) Ano de 2018. b) Ano de 2019.

Podemos verificar que não há informações úteis a primeira vista, somente os valores da correlação entre os alunos, assim verifica-se valores de correlação positiva (Azul), de descorrelação (Branco) e de correlação negativa (Vermelho).

Em seguida, reordenamos a Matriz de correlação com o algoritmo de Prim e obtém-se os seguintes resultados, na Figura 15.

Figura 15 – Matriz De Correlação ordenada pela MST



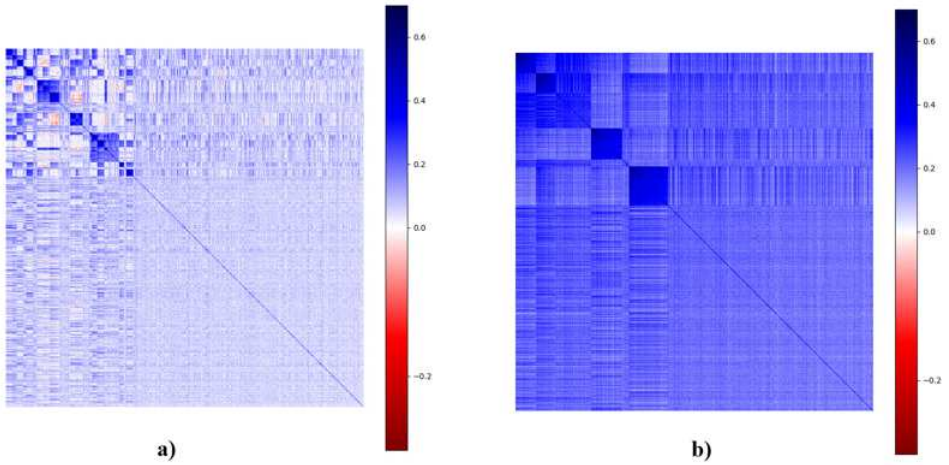
Fonte: Elaborado pelo Autor. a) Matriz de Correlção ordenada de 2018. b) Matriz de Correlção ordenada de 2019.

Assim, obtivemos os clusters, que podem ser vistos no canto superior esquerdo de cada gráfico. Devido a correlação ser calculada apartir do vetor de respostas, verificamos uma grande quantidade de alunos que estão descorrelacionados ou uma correlação fraca, visto que para obter correlação positiva, temos que ter respostas parecidas, e quando se trata de uma questão, temos que a chance de está correlacionada é de uma para cinco itens. Assim, podemos investigar a natureza desses clusters, primeiramente vamos ampliar a imagem para uma melhor verificação dos Clusters.

O primeiro palpite para entender as correlações é através da nota dos alunos, assim calculamos a média de notas entre os alunos de cada eixo e construímos o gráfico na figura 16, logo abaixo. O resultado esperado é que tenhamos uma concentração de notas altas no canto superior esquerdo e a medida que se a afasta a média das notas dos alunos vão diminuindo, assim podemos determinar o motivo das correlação e analisar a influência de variáveis socio-econômicas.

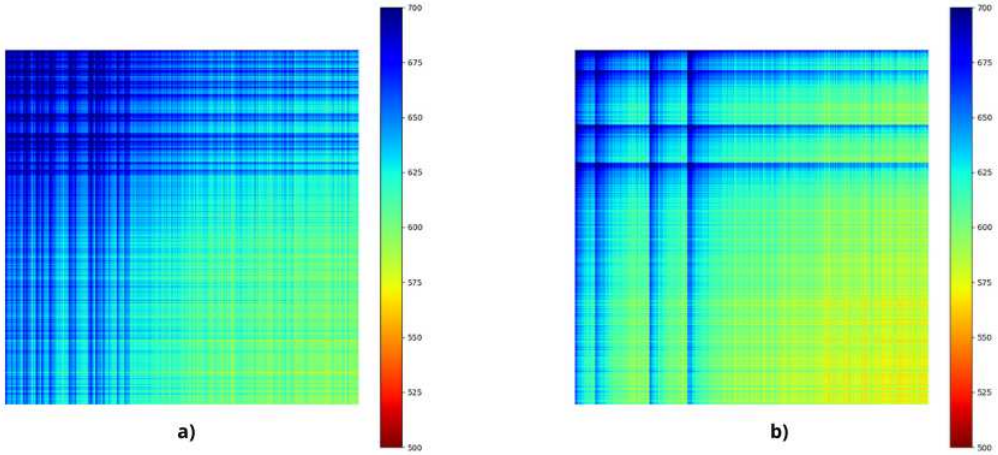
Com a imagem abaixo, podemos entender que a correlação através do vetor de respostas está sendo através das notas, o que era esperado.

Figura 16 – Matriz De Correlação ordenada pela MST Ampliada



Fonte: Elaborado pelo Autor. a) Matriz de Correlação ordenada de 2018 Ampliada. b) Matriz de Correlção ordenada de 2019 Ampliada.

Figura 17 – Matriz De Correlação ordenada pela MST da Média de Notas



Fonte: Elaborado pelo Autor. a) Matriz de Correlação ordenada de 2018 da Média da nota. b) Matriz de Correlção ordenada de 2019 Ampliada da Média da nota.

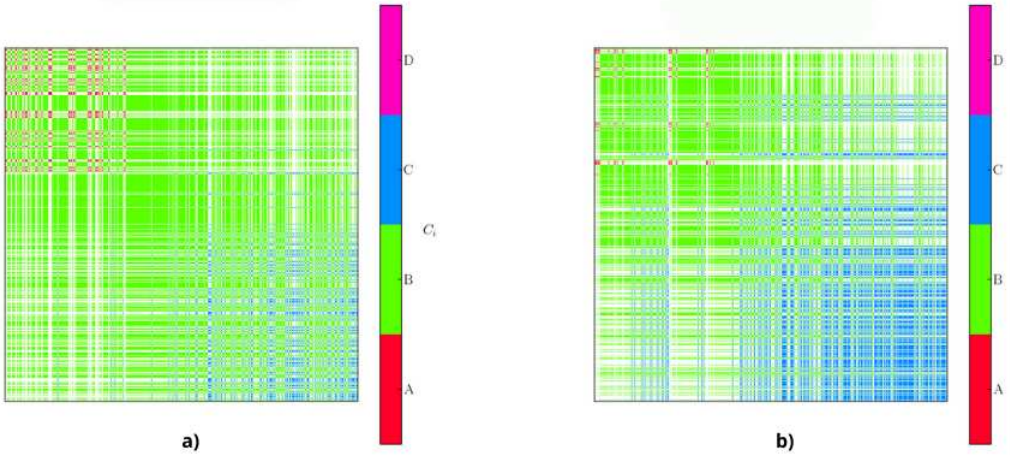
5.2 Análise da Influência das Variáveis Sócio-econômicas

Após entender a natureza das correlações, iremos analisar as variáveis respondidas pelos alunos no questionário do Enem, será utilizado a matriz ampliada, uma vez que as informações pertinentes aos clusters estão localizadas nessa região. Inicialmente, vamos criar um intervalo de notas e associá-los a uma variável. Temos os seguintes intervalos:

(**Alunos com Nota maior ou igual que 700 - A, Alunos com Nota menor que 700 e maior ou igual a 600 - B, Alunos com nota menor que 600 e maior ou igual a 500 - C,**

Alunos com Nota menor que 500 - D).

Figura 18 – Máscara de Intervalo de Notas

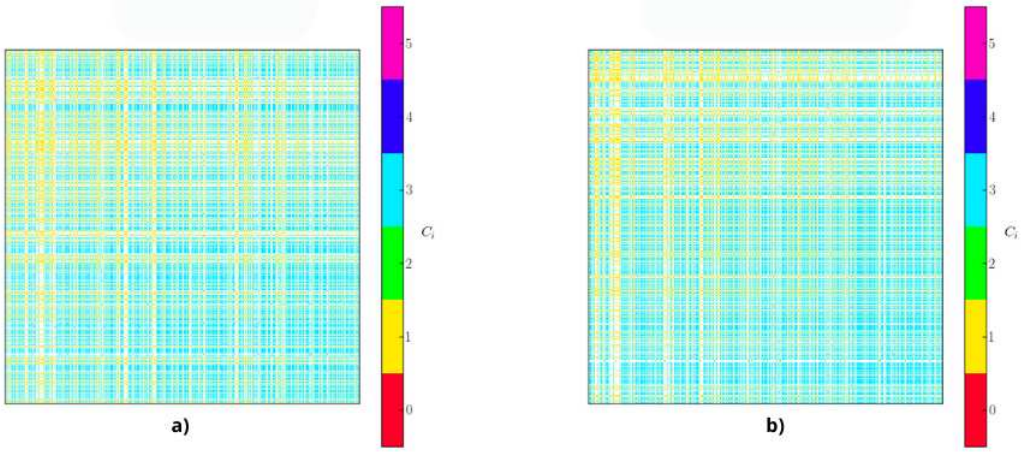


Fonte: Elaborado pelo Autor. a) Intervalo de Notas para 2018. b) Intervalo de Notas para 2019.

Assim fica nítido da influência de notas postivas na correlação com base no vetor de resposta, uma vez que existe regiões de agrupamento de Notas A no canto superior esquerdo e a medida que se afasta o valor da Nota vai decaindo. Ademais, iremos verificar os alunos em relação a sua raça, seguindo as seguintes variáveis:

(**0 - Não Declarado / 1 - Branco/ 2- Preto / 3 - Pardo / 4 - Amarelo / 5 - Indígenas**)

Figura 19 – Máscara Raça dos Participantes

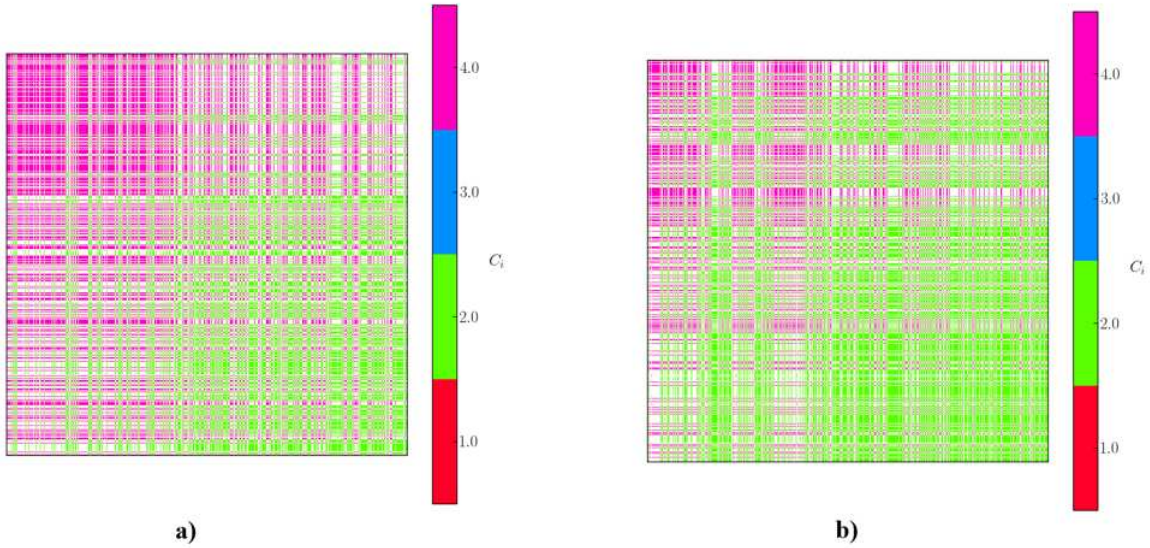


Fonte: Elaborado pelo Autor. a) Máscara para o ano de 2018. b) Máscara para o ano de 2019.

Esses clusters são predominados por alunos pardos e brancos (regiões amarelas e cianos, respectivamente), há pouca influência de outras raças, o que levanta questões em relação a democratização e acesso ao ensino superior, além de evidenciar uma problemática que segue desde os tempos do Brasil colônia, e assim segue o questionamento, como o Brasil um país de maioria da população negra, não há influência maior ou igual de participantes negros nos clusters de melhores notas.

Em relação ao tipo da depêndencia administrativa da escola, obtemos o seguinte:

Figura 20 – Máscara Tipo da Escola

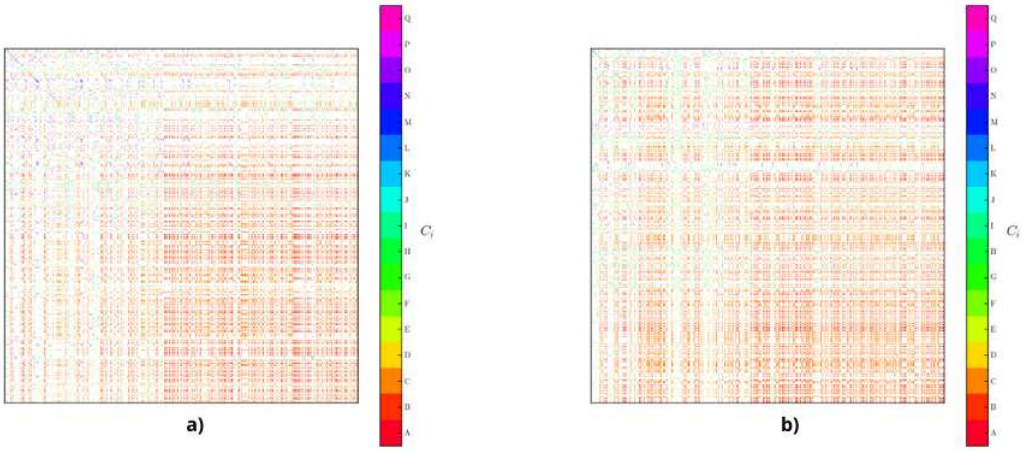


Fonte: Elaborado pelo Autor. a) Máscara para o ano de 2018. b) Máscara para o ano de 2019.

(**1 - Federal / 2 - Estadual / 3 - Municipal / 4 - Privado**). Paro o ano de 2018, verificamos uma grande influência de escolas privadas (regiões rosas) nos Clusters enquanto no ano de 2019, há uma maior influência de escolas Públicas Estaduais (regiões verdes). Porém, é nítida a tendência de no canto superior esquerdo, ficar concentrado de alunos de escolas privadas, evidenciando, uma maior parcela desses alunos em cursos que precisam de notas mais altas para serem cursados, assim podemos destacar outra discrepância no ensino brasileiro, sendo privilegiado o ensino privado.

Verificando agora, a influência da Renda, estamos interessados, em saber se rendas mais altas levam a melhores notas, uma vez que pessoas com maior poder aquisitivo, podem pagar as melhores escolas e os melhores materiais, ou se basta a vontade de estudar, como afirmam discursos meritocráticos assim, pode-se verificar na figura 21, logo abaixo, a máscara de renda.

Figura 21 – Máscara Renda



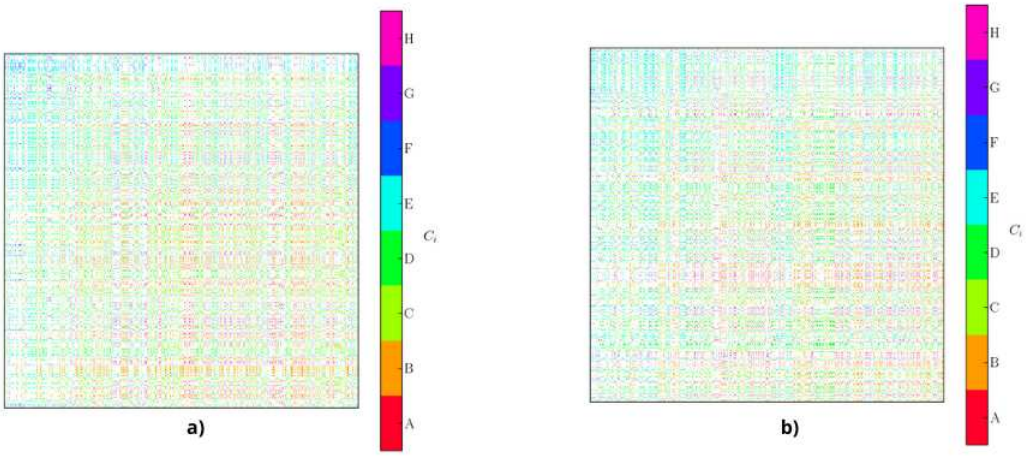
Fonte: Elaborado pelo Autor. a) Máscara para o ano de 2018. b) Máscara para o ano de 2019.

Tabela 1 – Variável Tipo de Renda

Variável	Renda
A	Nenhuma renda.
B	Até R 998,00 .
C	De R 998,01 até R 1.497,00.
D	De R 1.497,01 até R 1.996,00.
E	De R 1.996,01 até R 2.495,00.
F	De R 2.495,01 até R 2.994,00.
G	De R 2.994,01 até R 3.992,00.
H	De R 3.992,01 até R 4.990,00.
I	De R 4.990,01 até R 5.988,00.
J	De R 5.988,01 até R 6.986,00.
K	De R 6.986,01 até R 7.984,00.
L	De R 7.984,01 até R 8.982,00.
M	De R 8.982,01 até R 9.980,00.
N	De R 9.980,01 até R 11.976,00.
O	De R 11.976,01 até R 14.970,00.
P	De R 14.970,01 até R 19.960,00.
Q	Mais de R 19.960,00.

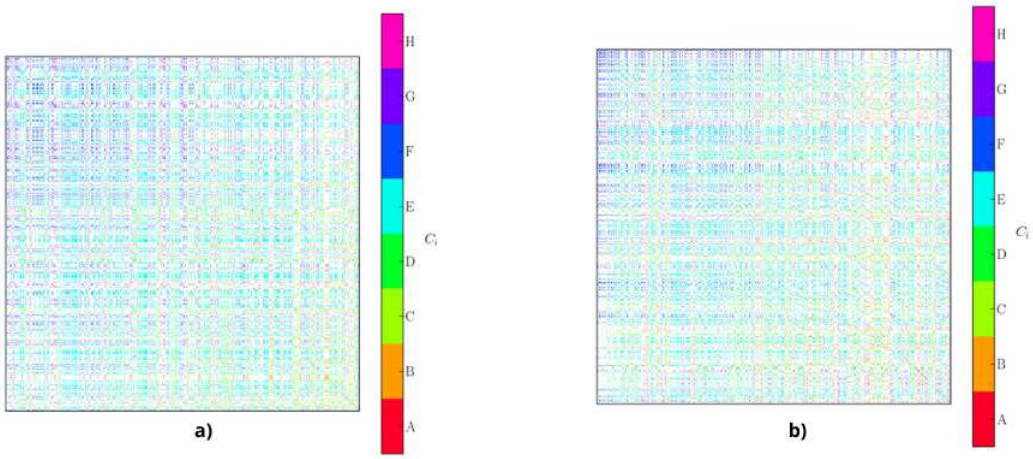
É notório que rendas mais altas estão concentradas nos cantos superiores esquerdos (cores azuis para cima), assim alunos com melhores rendas, obtém melhores notas, outro resultado esperado, porém rendas mais baixas são evidentes e com maiores influências, devido a concentração de renda no Brasil, poucas pessoas tem salários superiores a três salários mínimos. Porém, posso afirmar com experiência em escolas públicas, pode-se entender que pais com rendas mais baixas, tendem a incentivar os filhos a estudarem, visando uma melhor qualidade de vida.

Figura 22 – Máscara Escolaridade da Mãe



Fonte: Elaborado pelo Autor. a) Máscara para o ano de 2018. b) Máscara para o ano de 2019.

Figura 23 – Máscara Escolaridade do Pai



Fonte: Elaborado pelo Autor. a) Máscara para o ano de 2018. b) Máscara para o ano de 2019.

É nítido a influência da Escolaridade da mãe e do pai, uma vez que está sendo predominada escolaridades apartir de concluído o Ensino Médio (E) para escolaridades mais altas e a medida que se afasta do canto superior esquerdo essa tendência diminui, outro ponto interessante é que no gráfico de escolaridade da mãe tende a aparecer, escolaridades mais baixas, isso envolve questões dos moldes da sociedade, uma vez que historicamente as mulheres eram em sua maioria dona de casa, se limitando a cuidar da família, isso fez com que muitas mulheres não dessem continuidade aos estudos, se limitando a cuidar de suas casas, ou tendo que interrempor estudo e trabalho por conta de gravidez. Algo importante de salientar, é que por hipótese, se

Tabela 2 – Variável Escolaridade Da Mãe e do Pai

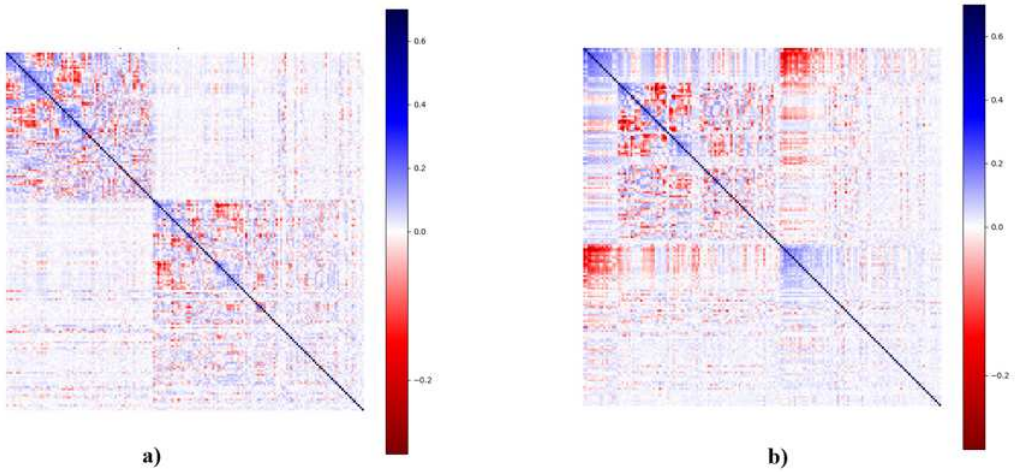
Variável	Escolaridade
A	Nunca estudou.
B	Não completou a 4ª série/5º ano do Ensino Fundamental.
C	De Completou a 4ª série/5º ano, mas não completou a 8ª série/9º ano do Ensino Fundamental.
D	Completou a 8ª série/9º ano do Ensino Fundamental, mas não completou o Ensino Médio.
E	Completou o Ensino Médio, mas não completou a Faculdade.
F	Completou a Faculdade, mas não completou a Pós-graduação.
G	Completou a Pós-graduação.
H	Não sei.

pegarmos base de dados mais antigas, devemos ver escolaridade da mãe cada vez mais baixa, uma vez que a inserção da mulher no mercado de trabalho e métodos anticoncepcionais, são relativamentes novos.

5.3 Resultados obtidos com a Clusterização das questões

Iremos agora, verificar a relação entre as questões e verificar os clusters a partir da área do conhecimento das questões, de maneira a verificar a natureza dos clusters.

Figura 24 – Matriz de Correlação Questão por Questão ordenada pela MST



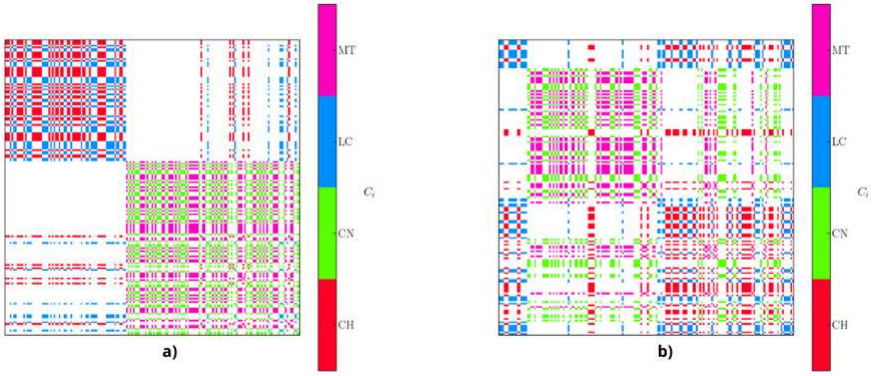
Fonte: Elaborado pelo Autor. a) Ano de 2018. b) Ano de 2019.

É notório que para o ano de 2018, temos duas regiões bastantes delimitadas, enquanto no ano de 2019, vemos clusters mais agrupados e menores. No ano de 2018 a nota média dos alunos da amostragem foi de 508.4, enquanto para o ano de 2019 foi de 482.9. Em relação a média das notas por dia de prova, temos: Dia 1 - Ciências Humanas e Linguaguens e Códigos, temos uma média de 522.4 para 2018 e 489.0 para 2019, enquanto o Dia 2 - Ciências Da Natureza

e Matemática, a média foi de 494.4 para 2018 e 476.8 para 2019.

Podemos verificar os seguintes clusters em relação a Área: **LC- Linguaguens e Códigos/ CH - Ciências Humanas / CN - Ciências da Natureza / MT - Matemática**

Figura 25 – Máscara por área da MST de correlação entre as questões



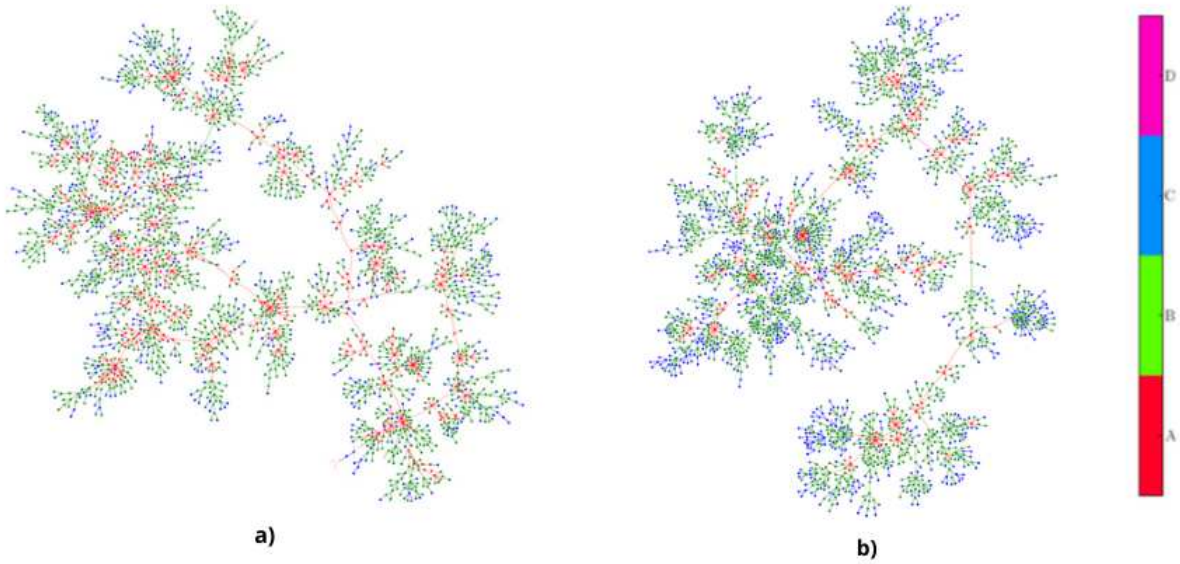
Fonte: Elaborado pelo Autor. a) Ano de 2018. b) Ano de 2019.

Fica nítido que em 2018 as questões estão se agrupando em relação a área de conhecimentos parecidos, como por exemplo maior correlação entre áreas como Matemática e Ciências da natureza, de mesma maneira verifica-se clusters entre Ciências Humanas e Linguagens e Códigos, enquanto em 2019, ainda é possível verificar essa separação, porém há questões de Ciências da Natureza, se relacionando com questões de Ciências Humanas e Linguagens e Códigos (canto inferior direito), logo vemos que uma diminuição na média das notas deixa os clusters mais desorganizados.

5.4 Imagens Das Redes Geradas

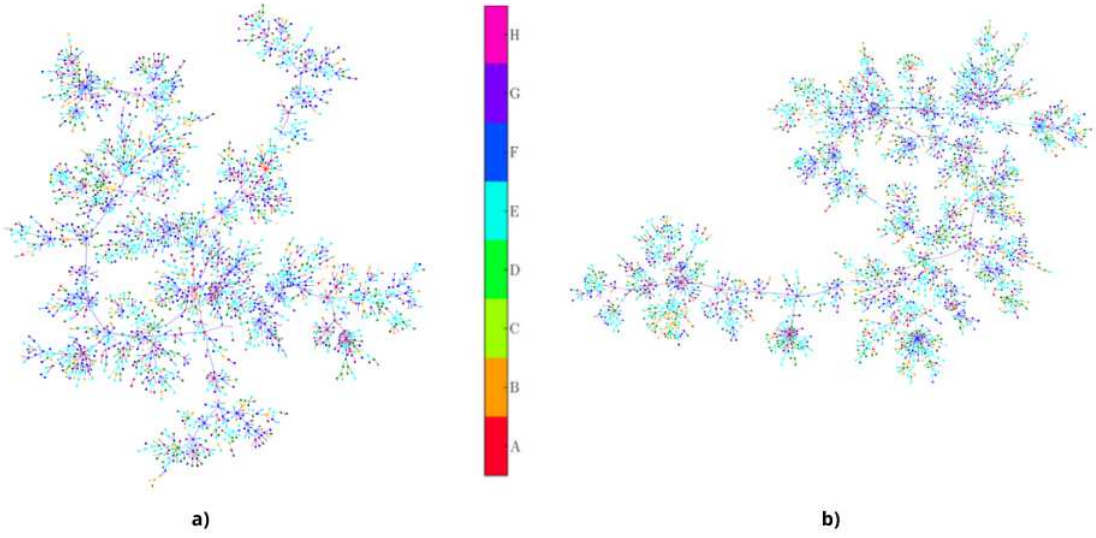
Nesta seção será mostrado as redes geradas, analisando as variáveis mencionadas na seção anterior. A graduação de cores segue o mesmo dos heatmaps da seção anterior, com exceção da Variável Renda, em que as cores mais parecidas, o que indica uma renda parecida, foram agrupudas de maneira a verificar melhor a rede complexa. A visualização da rede permite verificar de melhor maneira os clusters

Figura 26 – Rede Colorida pela Variável Nota



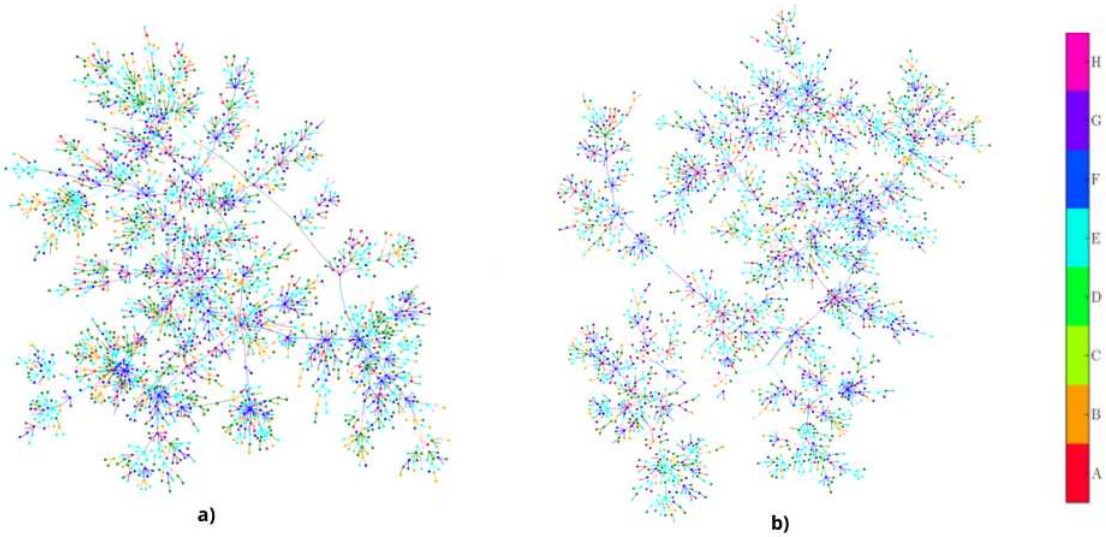
Fonte: Elaborado pelo Autor. a) Ano de 2018. b) Ano de 2019.

Figura 27 – Rede Colorida pela Variável Escolaridade da Mãe



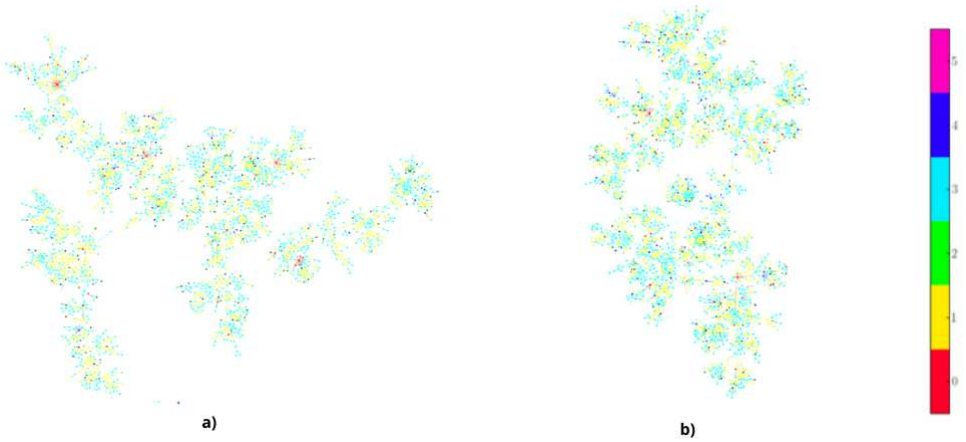
Fonte: Elaborado pelo Autor. a) Ano de 2018. b) Ano de 2019.

Figura 28 – Rede Colorida pela Variável Escolaridade do Pai



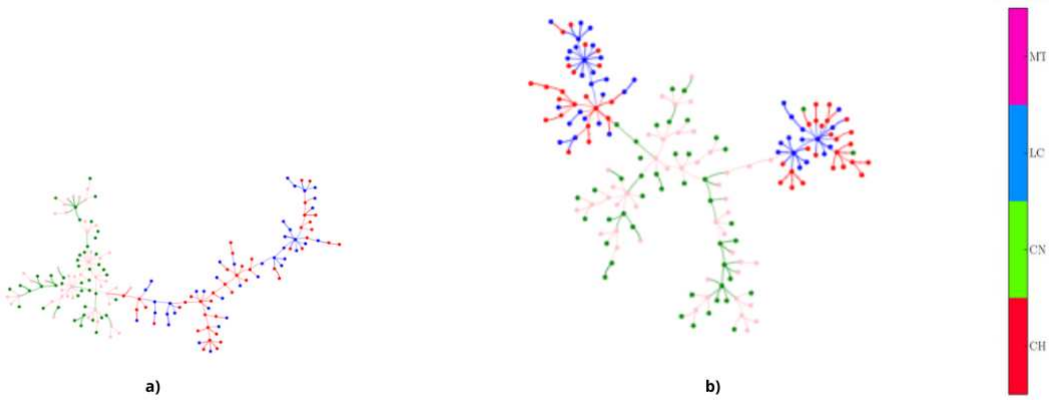
Fonte: Elaborado pelo Autor. a) Ano de 2018. b) Ano de 2019.

Figura 29 – Rede Colorida pela Variável Raça



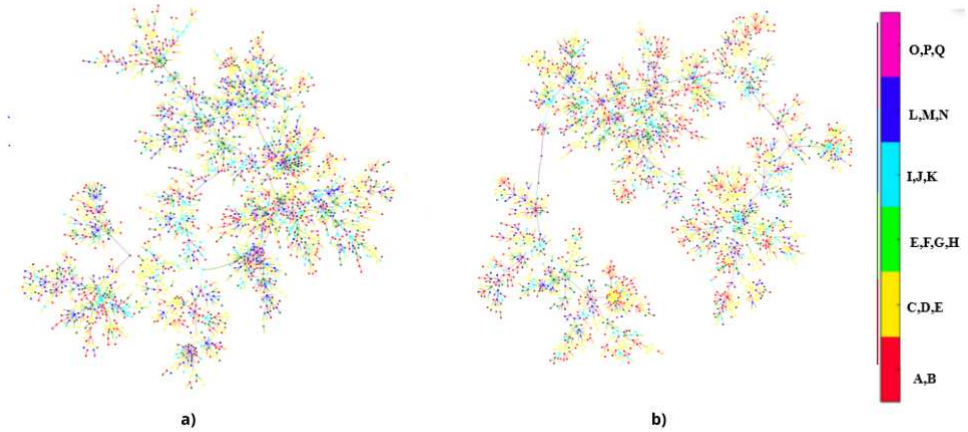
Fonte: Elaborado pelo Autor. a) Ano de 2018. b) Ano de 2019.

Figura 30 – Rede da MST de Correlação entre as Questões Colorida pela Variável Area



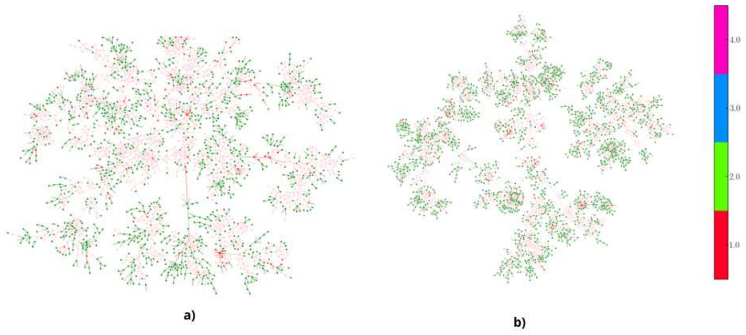
Fonte: Elaborado pelo Autor. a) Ano de 2018. b) Ano de 2019.

Figura 31 – Rede Colorida pela Variável Renda



Fonte: Elaborado pelo Autor. a) Ano de 2018. b) Ano de 2019.

Figura 32 – Rede Colorida pela Variável Tipo da Escola



Fonte: Elaborado pelo Autor. a) Ano de 2018. b) Ano de 2019.

6 CONCLUSÃO

Diante dos resultados, podemos comparar e verificar as tendências da educação no ensino médio para os anos de 2018 e 2019. Dessa forma, o objetivo deste trabalho foi propor uma análise para verificar as tedências dos indicadores educacionais, apartir da principal prova realizada no Brasil (ENEM), assim diante das correlações terem sido motivadas pelas melhores notas, como havia sido suposto, podemos traçar os perfis dos clusters encontrados entre esses participantes e assim entender a educação no Brasil.

Dessa maneira, vimos que é possível fazer a clusterização, entre os alunos com melhores notas, apartir do vetor de respostas, além disso, vimos que a tedência dentro do heatmap é de notas maiores no canto superior esquerdo e a medida que se afasta as notas vão diminuindo. Foi possível relacionar esses dois mil e quinhentos alunos mais bem correlacionados, dentro dos vinte e cinco mil selecionados na amostragem de um total na ordem de oitenta mil alunos.

Em relação às variáveis socio-econômicas, temos ainda grandes disparidades sociais e econômicas evidentes, uma vez que dentre os alunos mais bem correlacionados, as variáveis tendem a não mudar os seus valores, como por exemplo na máscara de raça, vemos um predomínio de alunos brancos e pardos, o que realça desigualdades dentro de um país tão miscigenado quanto o Brasil mostrando a importância da lei de cotas e da democratização da entrada ao ensino superior, além disso, em relação ao tipo da escola, vimos uma maioria de participantes de escolas privadas em 2018, porém, a situação mudou em 2019 pois houve um maior incremento de escolas públicas, mas ainda assim, quem está presente de maneira preponderante nos clusters, são alunos de escolas privadas, o motivo de não aparecer alunos de escolas municipais, uma vez que o número de escolas municipais que contam com ensino médio é baixa, assim como a proporção de alunos de escolas federais (escolas militares e institutos federais).

Com ênfase na renda, vimos um maior número de alunos com rendas maiores nos cantos superiores esquerdos e a medida que afasta a renda diminui, é importante salientar que há um maior número de rendas mais baixas no ano de 2019, isso é natural de pensar, uma vez que houve uma maior inserção de alunos de escolas públicas. Em relação a escolaridades dos pais, vimos uma predominância de escolaridades apartir do ensino médio conluído até completar pós-graduação, com essa se inserindo nos cantos superiores esquerdos, de ambos os anos analisados. Outra constatação interessante é que a escolaridade da mãe tende a ser mais baixa que a do pai, devido aos moldes sociais que tendem a mudar.

Por fim, analisamos as questões do exame, e que elas se relacionam a medida que a média das notas mudam, o que indica a dificuldade em determinada área, uma vez que no ano de 2019, tivemos uma maior relação entre áreas diferentes, mostrando que esssas correlações devem ser por alunos que foram boas em ambas, enquanto no ano de 2018, as áreas parecidas formaram blocos bem determinados evidenciando uma maior facilidade com a prova, além disso, há regiões onde as correlações entre duas áreas distintas se correlacionam, de maneira que podemos intepretar essas questões como fáceis.

Contudo, devido a limitações computacionais não se pôde realizar a análise para todo o estado ou então para todo o Brasil, visto que, trabalhando com uma matriz de correlação temos um grande número de dados devido a quantidade de alunos que realizam o exame.

Diante do PNE instaurado em 2014, é interessante realizar uma análise temporal do comportamento das variáveis sócio-econômicas nos clusters num período de dez anos antes de ser instaurado o PNE e no período de 2014 a 2024, de maneira que, torna-se possível fazer uma validação dos planos estabelecidos além de analisar as tendências educacionais para entender a educação no ensino médio no estado do Ceará e utilizar a análise para propor melhorias no ensino básico, levando em conta as disparidades regionais.

REFERÊNCIAS

BAKER, R. S.; ISOTANI, S.; CARVALHO, A. M. J. B. de. Mineração de dados educacionais: Oportunidades para o brasil. **Revista Brasileira de Informática na Educação (RBIE)**, *[S.I.]*,v.19, n. 2, p. 3–13, 2011.

BAKER, R. S.; YACEF, K. The state of educational data mining in 2009: A review and future visions. **Journal of Educational Data Mining**, *[S.I.]*, v.1, n. 1, p. 3–17, 2009.

BERNI, J. C. **Espaços Métricos**. 2021. Disponível em: https://www.ime.usp.br/~jeancb/EM2021.pdf. Acesso em: 05 jul. 2024.

BOAVENTURA, P. O. N.; JURKIEWICZ, S. **Grafos**: introdução e prática. 2. ed. São Paulo: Blucher, 2017.

BOLFARINE H.; BUSSAB, W. O. **Elementos de Amostragem**. 1.ed.São Paulo: Blucher, 2005.

BONDY, J.; MURTY, U. **Graph theory with applications**. 1. ed. New York: Elsevier Science Publishing, 1976.

BRASIL. **Constituição da República Federativa do Brasil de 1988**. 1988. Disponível em: https://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm. Acesso em: 21 ago. 2024.

BRASIL. **Lei de Diretrizes e Bases da Educação Nacional (LDB), Lei nº 9.394**. 1996. Disponível em: http://www.planalto.gov.br/ccivil_03/leis/l9394.htm. Acesso em: 21 ago. 2024.

BRASIL. **Plano Nacional de Educação (PNE) 2014-2024**. 2014. Disponível em: http://www.pne.mec.gov.br/. Acesso em: 21 ago. 2024.

CAMACHO, L. F. M. **Brazilian House of Representatives Analysis from Network Theory Perspective.** Dissertação (Mestrado) — Instituto de Física Gleb Wataghin, Programa de Pós-Graduação em Física, Unicamp, Campinas, 2017.

CESAR, C. L. **Teoria da Probabilidade II**. 2013. Disponível em: https://www.ifi.unicamp.br//~lenz/Econofisica/. Acesso em: 01 jul. 2024.

CORMEN, T.; LEILERSON, C.; RIVEST, R. L.; STEIN, C. **Algoritmos**: teoria e prática. 3. ed. Rio de Janeiro: Elsevier, 2012.

DANTAS, C. A. B. **Probabilidade**: um curso introdutório. 3. ed. São Paulo: Edusp, 2008.

EULER, L. **Mathematics and the modern world**: the koenigsberg bridges. San Francisco: W.H. Freeman and Company, 1968.

FEOFILOFF, P.; KOHAYAKAWA, Y.; WAKABAYASHI, Y. **Uma introdução suscinta à teoria dos grafos.** 2. ed. São Paulo: IMI, 2004.

GARCIA F. M., C. R. S. M. T. G. C. T. O enem como política de avaliação e as contradições ao processo de democratização educacional. **Perspectiva**, Florianópolis, v.39, n. 3, p. 01–21, 2021.

GERSTING, J. L. **Mathematical Structures for Computer Science**. 7. ed. New York: W. H. Freeman, 2017.

GNERI, M. A. **Apostila de Probabilidade**. 2023. Disponível em: https://www.ime.unicamp.br/~veronica/ME203ME414/apostilaprob.pdf. Acesso em: 01 ago. 2024.

GUIMARÃES JÚNIOR, J. C.; FORTALEZA, I.; POLAK, A.; CHAGAS, L. Análise de dados educacionais: como a tecnologia pode ser usada para obter insights sobre o desempenho dos alunos. **Contemporânea**, São Paulo, 3, n. 8, p. 11056–11072, 2023.

INEP. **Exame nacional do ensino médio**: Relatório final 98. 1998. Disponível em: https://download.inep.gov.br/publicacoes/institucionais/avaliacoes_e_exames_da_educacao_basica/enem_exame_nacional_do_ensino_medio_relatorio_final_1998.pdf. Acesso em: 22 ago. 2024.

INEP. **Enem 2009 acontece neste fim de semana**: saiba tudo. 2009. Disponível em: https://www.gov.br/inep/pt-br/assuntos/noticias/enem/enem-2009-acontece-neste-fim-de-semana-saiba-tudo#:~:text=Aplica%C3%A7%C3%A3o%20das%20provas,17h30%20no%20dia%205%2C%20s%C3%A1bado. Acesso em: 22 ago. 2024.

INEP. **Matriz de referência Enem**. 2009. Disponível em: https://download.inep.gov.br/download/enem/matriz_referencia.pdf. Acesso em: 21 ago. 2024.

INEP. **Matriz de referência Enem**. 2009, p.8. Disponível em: https://download.inep.gov.br/download/enem/matriz_referencia.pdf. Acesso em: 21 ago. 2024.

INEP. **Interpretação pedagógica das escalas de proficiência**. 2014. Disponível em: https://download.inep.gov.br/publicacoes/institucionais/avaliacoes_e_exames_da_educacao_basica/interpretacao_pedagogica_das_escalas_de_proficiencia.pdf. Acesso em: 21 ago. 2024.

INEP. **Exame nacional do ensino médio**: Enem escalas de proficiência 1998/2008. 2018. Disponível em: https://download.inep.gov.br/publicacoes/institucionais/avaliacoes_e_exames_da_educacao_basica/enem_escala_de_proficiencia_1998_2008.pdf. Acesso em: 21 ago. 2024.

INEP. **Microdados do Enem**. 2024. Disponível em: https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem. Acesso em: 30 ago. 2024.

KAGGLE. **Microdados do Enem**. 2024. Disponível em: https://www.kaggle.com/search?q=microdados+enem+. Acesso em: 30 ago. 2024.

MAGALHÃES, M. N. **Probabilidade e Variáveis Aleatórias**. 2. ed. São Paulo: Edusp, 2006.

MEC. **Teoria de resposta ao item avalia habilidade e minimiza o "chute" de candidatos**. 2011. Disponível em: http://portal.mec.gov.br/ultimas-noticias/389-ensino-medio-2092297298/17319-teoria-de-resposta-ao-item-avalia-habilidade-e-minimiza-o-chute. Acesso em: 22 ago. 2024.

MEC. **Entenda a sua nota no Enem: guia do participante**. 2013. Disponível em: https://download.inep.gov.br/educacao_basica/enem/guia_participante/2013/guia_do_participante_notas.pdf. Acesso em: 22 ago. 2024.

MEC. **Com mudanças positivas e mais seguro, Enem 2017 está pronto para as provas neste domingo**. 2017. Disponível em: http://portal.mec.gov.br/ultimas-noticias/418-enem-946573306/56631-aplicacao-do-exame-mais-barata-que-a-de-2016-envolvera-mais-de-600-mil-pessoas#:~:text=Novidades%20da%20edi%C3%A7%C3%A3o%20%E2%80%93%20A%20prova,cinco%20horas%20e%2030%20minutos. Acesso em: 22 ago. 2024.

MELO, S. M. **Introdução à Teoria dos Grafos.** Dissertação (Mestrado Profissional em matemática) — Centro de Ciências exatas e da natureza, Mestrado profissional em matemática em rede nacional: PROFMAT-CCEN-UFPB, Universidade Federal da Paraíba, João Pessoa, 2014.

MEYER, P. L. **Probabilidade**: aplicações à estatística. 2. ed. Rio de Janeiro: LTC, 1987.

MORETTIN P.A.; BUSSAB, W. **Estatística Básica**. 5. ed. São Paulo: Saraiva, 2006.

OKBR. **De portas fechadas, Inep discute microdados da educação sem envolver ampla sociedade civil**. 2024. Disponível em: https://ok.org.br/noticia/de-portas-fechadas-inep-discute-microdados-da-educacao-sem-envolver-ampla-sociedade-civil/#:~:text=Em%20fevereiro%20de%202022%2C%20o,do%20Ensino%20M%C3%A9dio%20(Enem). Acesso em: 30 ago. 2024.

OLIVEIRA, T. S. O enem: breves considerações sobre importância avaliativa e refroma educacional. **Educação por escrito**, Porto Alegre, v.7, n. 2, p. 278–288, 2016.

RIGGAN M.; CHEN, L. N. R. Predictive analytics for improving student outcomes in high schools. journal of learning analytics. **Journal of Learning Analytics**, *[S.I.]*, v.5, n. 1, p. 96–121, 2018.

ROMERO C.; VENTURA, S. Z. A. B. A. Predicting students'dropouts in moocs with anonymized forum data. **IEEE Transactions on Learning Technologies**, *[S.I.]*, v.10, n. 1, p. 3–16, 2017.

SCHULZ, P. **Ligando os pontos na pandemia**. 2020. Disponível em: https://unicamp.br/unicamp/ju/artigos/peter-schulz/ligando-os-pontos-da-pandemia/#:~:text=No%20meio%20da%20pandemia%20de,duas%20vezes%20pelo%20mesmo%20caminho. Acesso em: 09 jul. 2024.

SILVA M. R.;RIBEIRO, A. C. B. M. Reformas para quê? as políticas educacionais nos anos de 1990, o "novo projeto de formação" e os resultados das avaliações nacionais. **Perspectiva**, São Paulo, v.26, n. 2, p. 523–550, 2008.

SILVA, N. **Amostragem Probabilística**: um curso introdutório. 3. ed. São Paulo: Edusp, 2015.

SILVANA M., O. L. L. M. R. E. Enem: pontos positivos para a educação brasileira. **Revista Educação e Políticas em Debate**, Uberlândia, v.3, n. 2, 2014.

SILVEIRA F. L., B. M. C. d. S. R. Exame nacional do ensino médio (enem): Uma análise críitica. **Revista Brasileira de Ensino de Física**, São Paulo, v.37, n. 1, p. 1101, 2015.

VASCONCELOS, V. B. de.; VASCONCELOS, G. B. de.; CHAQUIAM, M. **Um percurso pela história da probabilidade**. 2022. Disponível em: https://www.researchgate.net/publication/360985894_Um_percurso_pela_historia_da_probabilidade. Acesso em: 01 ago. 2024.

VIALI, L. **Algumas considerações sobre a origem da teoria da probabilidade**. 2020. Disponível em: https://rbhm.org.br/index.php/RBHM/article/view/177. Acesso em: 01 ago. 2024.

WIKIMATH-USP. **Grafos**: introdução e algumas aplicações. 2023. Disponível em: https://sites.icmc.usp.br/aurichi/wikimat/doku.php?id=grafos:definicaografos#:~:text=Um%20grafo%20G%3D(V%2C,suas%20arestas%20(ou%20linhas). Acesso em: 30 ago. 2024.

WILSON, R. J. **Introduction to Graph Theory**. England: Prentice Hall, 1996.

WORDPRESS-BLOG. **Grafos**: introdução e algumas aplicações. 2022. Disponível em: https://estructurasite.wordpress.com/algoritmo-de-prim/. Acesso em: 30 ago. 2024.

7 APÊNDICE A - CÓDIGOS-FONTES UTILIZADOS

Código-fonte 1 – Processo de Análise usando como exemplo 2018 para discretizar o vetor de respostas e ordena-lós

```
1
2  import pandas as pd
3  import numpy as np
4  import matplotlib.pyplot as plt
5
6
7  dados_i = []
8  for dados in pd.read_csv("/content/drive/MyDrive/
       MICRODADOS_ENEM_2018.csv",
9      sep=';',
10     encoding="ISO-8859-1", chunksize=500000):
11      dados_i.append(dados
12                     .loc[dados['TP_ST_CONCLUSAO'] == 2]
13                     .loc[dados['NU_ANO'] == 2018]
14                     .loc[dados['SG_UF_ESC'] == 'CE'])
15
16
17
18 dados = pd.concat(dados_i)
19 #Removendo os dados faltantes
20 dados.dropna(inplace = True)
21 dados.index = range(len(dados))
22
23 #Criando vari vel vetor de respostas dos dois dias
24
25 dados['GABARITO_AL_D1'] = 0 * len(dados)
26 dados['GABARITO_AL_D2'] = 0 * len(dados)
27
```

```
28 for i in range(len(dados)):
29   print(i)
30   dados['GABARITO_AL_D1'][i] = (dados['TX_RESPOSTAS_LC'][i
        ][10:51] + dados['TX_RESPOSTAS_CH'][i])
31   dados['GABARITO_AL_D2'][i] = (dados['TX_RESPOSTAS_CN'][i]
         + dados['TX_RESPOSTAS_MT'][i])
32
33
34 # Ordena  o do primeiro dia de prova e do segundo.
35
36 dados['GABARITO_ORD_AL_D1'] = 0 * len(dados)
37 dados['GABARITO_ORD_AL_D2'] = 0 * len(dados)
38
39 for i in range(len(dados)):
40   print(i)
41   al_ord = []
42
43   if (dados.CO_PROVA_LC[i]) == 455:
44     for j in range(85):
45       al_ord.append(dados['GABARITO_AL_D1'][i][d1['AMARELO'
           ][j]])
46
47   elif (dados.CO_PROVA_LC[i]) == 456:
48     for j in range(85):
49       al_ord.append(dados['GABARITO_AL_D1'][i][d1['ROSA'][j
           ]])
50
51   elif (dados.CO_PROVA_LC[i]) == 457:
52     for j in range(85):
53       al_ord.append(dados['GABARITO_AL_D1'][i][d1['CINZA'][
           j]])
54
```

```
55
56    elif (dados.CO_PROVA_LC[i]) == 458:
57      for j in range(85):
58        al_ord.append(dados['GABARITO_AL_D1'][i][j])
59
60
61    dados['GABARITO_ORD_AL_D1'][i] = al_ord
62
63
64  for i in range(len(dados)):
65    print(i)
66    al_ord2 = []
67
68    if (dados.CO_PROVA_CN[i]) == 447:
69      for j in range(90):
70        al_ord2.append(dados['GABARITO_AL_D2'][i][d2['AMARELO
            '][j]])
71
72    elif (dados.CO_PROVA_CN[i]) == 448:
73      for j in range(90):
74        al_ord2.append(dados['GABARITO_AL_D2'][i][d2['CINZA'
            ][j]])
75
76
77    elif (dados.CO_PROVA_CN[i]) == 449:
78      for j in range(90):
79        al_ord2.append(dados['GABARITO_AL_D2'][i][d2['ROSA'][
            j]])
80
81    elif (dados.CO_PROVA_CN[i]) == 450:
82      for j in range(90):
83        al_ord2.append(dados['GABARITO_AL_D2'][i][j])
```

```
84
85    dados['GABARITO_ORD_AL_D2'][i] = al_ord2
86
87
88  # Criando a tabela de respostas
89
90  qd1 = []
91  qd2 = []
92  for i in range(85):
93    qd1.append([])
94
95  for j in range(90):
96    qd2.append([])
97
98  for i in range(len(dados)):
99    for j in range(85):
100     qd1[j].append(dados['GABARITO_ORD_AL_D1'][i][j])
101
102 for i in range(len(dados)):
103   for j in range(90):
104     qd2[j].append(dados['GABARITO_ORD_AL_D2'][i][j])
105
106 quest = pd.DataFrame()
107 for i in range(85):
108   quest[('Q{}'.format(6+i))] = qd1[i]
109
110 for i in range(90):
111   quest[('Q{}'.format(91+i))] = qd2[i]
112
113 #Aplicando a discretiza  o
114
```

```
115 sub = {'A':1, 'B':2, 'C' :  3, 'D' : 4, 'E': 5, '.': 0, '*'
        : -1}
116
117 for i in range(6,181):
118   quest['Q{}'.format(i)] = quest[('Q{}'.format(i))].map(sub
        )
119
120 #Remo   o de alunos que responderiam somente um item nos
       gabaritos de cada dia
121
122 ind = []
123 for i in range(quest.shape[1]):
124     print(i)
125     if len(quest[i][0:40].unique()) == 1:
126         quest.drop(i,axis = 1,inplace = True)
127         dados_ce.drop(i,axis = 0,inplace = True)
128     elif len(quest[i][40:85].unique()) == 1:
129         quest.drop(i,axis = 1,inplace = True)
130         dados_ce.drop(i,axis = 0,inplace = True)
131     elif len(quest[i][85:130].unique()) == 1:
132         quest.drop(i,axis = 1,inplace = True)
133         dados_ce.drop(i,axis = 0,inplace = True)
134     elif len(quest[i][130:175].unique()) == 1:
135         quest.drop(i,axis = 1,inplace = True)
136         dados_ce.drop(i,axis = 0,inplace = True)
```

Código-fonte 2 – Código para realização da amostragem

```
1
2 import pandas as pd
3 import numpy as np
4
```

```
5  dados_ce['NOTA'] = [0]*len(dados_ce)
6  for i in range(len(dados_ce)):
7      dados_ce['NOTA'][i] = round((dados_ce['NU_NOTA_CN'][i]+
           dados_ce['NU_NOTA_CH'][i]+dados_ce['NU_NOTA_LC'][i]+
           dados_ce['NU_NOTA_MT'][i])/4,2)
8
9
10 escolas_frame = []
11 for i in range(len(dados_ce.CO_MUNICIPIO_RESIDENCIA.unique
       ())):
12   escolas_frame.append(dados_ce.loc[dados_ce['
         CO_MUNICIPIO_RESIDENCIA'] == dados_ce.
         CO_MUNICIPIO_RESIDENCIA.unique()[i]])
13   escolas_frame[i] = escolas_frame[i][['NOTA','NU_NOTA_LC',
         'NU_NOTA_CH','NU_NOTA_CN','NU_NOTA_MT','
         NU_NOTA_REDACAO','NO_MUNICIPIO_RESIDENCIA']]
14
15 amostras = []
16
17 for i in range(100):
18   amostras.append([])
19
20 medias_amostrais = []
21 desvios_amostrais = []
22
23 for i in range(100):
24   for j in range(len(escolas_frame)):
25     amostras[i].append(escolas_frame[j].sample(round(len(
           escolas_frame[j])/len(dados_ce)*25000), random_state
            = i))
26
27   sample = pd.concat(amostras[i])
```

```
28    medias_amostrais.append(sample['NOTA'].mean())
29    desvios_amostrais.append(sample['NOTA'].std())
```

Código-fonte 3 – Código para remoção de Ruídos

```
1
2  import numpy as np
3  import pandas as pd
4  from sympy import *
5
6  #Fun   es para remo   o de ru dos
7
8  def autovalores(matriz):
9      valorc, vetorc = np.linalg.eig(matriz)
10     valorc.sort()
11     valorc = valorc[::-1]
12     return valorc,vetorc
13 def corte(k,n):
14     y = k / n
15     b = (1 + (y) ** (0.5)) ** (2)
16     return b
17 def matrizes_De_autovetores_diag(matriz,ordem):
18     valorc,vetorc = autovalores(matriz)
19     m = []
20     for k in range(ordem):
21         m.append(np.zeros(ordem))
22     for i in range(ordem):
23         for j in range(ordem):
24             if i == j:
25                 m[i][j] = valorc[i]
26             else:
27                 pass
```

```
28      # Matriz S de autovetores
29      matriz_de_autovetores = vetorc
30      # Matriz transposta de autovetores
31      matriz_t_autovetores = np.transpose(
            matriz_de_autovetores)
32      # Matriz Diagonal com os autovalores
33      matriz_diagonalizada = m
34      # Para usar futuramente
35      matriz_diagonalizada_fixa = matriz_diagonalizada.copy()
36      return matriz_de_autovetores, matriz_t_autovetores,
            matriz_diagonalizada
37  def matriz_diagonal_limpa(matriz,corte,ordem):
38      # Obtendo a matriz Diagonal clean
39      matriz_clean = []
40      for i in range(ordem):
41          for j in range(ordem):
42              if matriz[i][j] >= corte:
43                  if i == j:
44                      matriz[i][j] = matriz[i][j] * 1
45                  else:
46                      matriz[i][j] = 0
47              else:
48                  matriz[i][j] = 0
49
50      matriz_diagonalizada_clean = matriz
51      return matriz_diagonalizada_clean
52  def matriz_De_correlacao_limpa(matriz_autov, matriz_autovt,
        matriz_diag_clean,ordem):
53      # Matriz clean
54      # matriz_de_autovetores*matriz_diagonalizada_clean*
            matriz_de_autovetores_t
55      matriz_clean1 = np.dot(matriz_autov, matriz_diag_clean
```

```
        )
    matriz_clean = np.dot(matriz_clean1, matriz_autovt)
    matriz_clean_fixa = matriz_clean.copy()

    # Matriz de correla  o clean sem mudan a na diagonal
          a seginte
    delta = 0
    for i in range(ordem):
        for j in range(ordem):
            if i == j:
                matriz_clean[i][j] = 1
            else:
                matriz_clean[i][j] = ((delta + (1 - delta))
                    * matriz_clean_fixa[i][j])
    m = []
    for k in range(ordem):
        m.append(np.zeros(ordem))
    for i in range(ordem):
        for j in range(ordem):
            m[i][j] = matriz_clean[i][j].real

    matriz_correla  o_clean = m

    return matriz_correla  o_clean
def desvios(matriz,ordem):
    # Calculando os desvios para obter a matriz de
       covari ncia clean
    # Para calcular rapidamente os desvios
    desvios = []
    for i in range(ordem):
        for j in range(ordem):
            if i == j:
```

```
84                  desvios.append((matriz[i][j]) ** 0.5)
85      desvios_T = np.transpose(desvios)
86
87      return desvios,desvios_T
88
89
90  #Remo   o de ru dos
91  o = int(input('Digite a quantidade de vezes que deseja
        fazer a limpeza: '))
92  ordem = mst_correl.shape[0]
93  lista_matrizes_corr_clean =[]
94  for n in range(o):
95      print('{},'.format(n), end =' ')
96
97      #Obtendo os autovalores ordenados
98      autovalor = autovalores(mst_correl)
99
100     #Definindo o corte
101     b = corte(ordem,mst_correl.shape[1])
102
103
104     #MAtrizes de autovetores e a diagonalizada pelos
            autovalores
105     matriz_de_autovetores, matriz_t_autovetores,
            matriz_diagonalizada = matrizes_De_autovetores_diag(
            mst_correl,ordem)
106
107     #Matriz Diagonal Clean
108     matriz_diagonal_clean = matriz_diagonal_limpa(
            matriz_diagonalizada,b,ordem)
109
110     matriz_correl_limpa = matriz_De_correlacao_limpa(
```

```
        matriz_de_autovetores, matriz_t_autovetores,
        matriz_diagonalizada,ordem)

    lista_matrizes_corr_clean.append(matriz_correl_limpa)

    matriz_corr_fixa = matriz_correl_limpa
```

Código-fonte 4 – Código para Algoritmo de Prim (MST)

```

def prim(matriz,ordem):

  mst = []


  node = np.array(matriz)

  x = np.unravel_index(node.argmin(),node.shape)

  mst.append(x[0])
  mst.append(x[1])
  print(mst)
  for j in range(ordem):
    node[j][x[0]] = 3
    node[j][x[1]]= 3


  lista_cand = []
  lista_mst = []
  for i in range(ordem- 2):
    print(i, end= ',')
    min = 10
```

```
24      cand = (0,0)
25      for val in mst:
26        if np.min(node[val]) < min:
27          val_min = val
28          min = np.min(node[val])
29          cand = (val,node[val].argmin())
30        lista_mst.append((val_min,cand[1]), cand[0])
31
32      for l in range(ordem):
33        node[l][cand[1]] = 3
34
35      mst.append(cand[1])
36
37    return mst
38
39  def correl(data):
40      matriz_corr = np.corrcoef(data,rowvar = False)
41      matriz_corr = matriz_corr.astype(np.float32)
42
43      return matriz_corr
44
45  def dist(ordem,matriz_corr):
46
47      matriz_dist = (2*abs(1-np.array(matriz_corr,dtype = np.
            float32)))**(0.5)
48
49
50      #Colocando diagonal principal igual a 10
51      np.fill_diagonal(matriz_dist,10)
52
53      return matriz_dist
54
```

```
55 def mst_dist(ordem,matriz_dist,mst):
56
57     matriz_dist_mst = np.zeros((ordem,ordem))
58
59     for i in range(ordem):
60         for j in range(ordem):
61
62             matriz_dist_mst[i][j] = matriz_dist[mst[i]][mst
                   [j]]
63
64     matriz_dist_mst = pd.DataFrame(matriz_dist_mst)
65
66     return matriz_dist_mst
67
68 def mst_corr(ordem,matriz_correl,mst):
69
70     matriz_correl_mst = np.zeros((ordem,ordem))
71
72     for i in range(ordem):
73         for j in range(ordem):
74             matriz_correl_mst[i][j] = matriz_correl[mst[i
                   ]][mst[j]]
75
76     matriz_correl_mst = pd.DataFrame(matriz_correl_mst)
77
78     return matriz_correl_mst
```

8 ANEXO A - MICRODADOS ENEM

Figura 33 – Visualização dos Microdados

	NU_INSCRICAO	NU_ANO	TP_FAIXA_ETARIA	TP_SEXO	TP_ESTADO_CIVIL	TP_COR_RACA	TP_NACIONALIDADE	TP_ST_CONCLUSAO	TP_ANO_CONCLUIU	TP_ESCOLA	...	Q020	Q021	Q022	Q023	Q024	Q025
0	210056717205	2022	5	F	1	3	1	2	0	2	...	A	A	C	A	A	B
1	210055098334	2022	3	F	1	3	1	2	0	2	...	A	A	C	A	A	B
2	210056381096	2022	3	M	1	3	1	2	0	2	...	A	A	B	A	B	B
3	210055099914	2022	3	F	1	3	1	2	0	2	...	A	A	C	A	A	A
4	210057315267	2022	3	M	1	3	1	2	0	2	...	A	A	D	A	A	B
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
70151	210055423279	2022	2	M	1	1	1	2	0	3	...	B	A	C	A	B	B
70152	210055583593	2022	3	F	1	1	1	2	0	2	...	B	A	D	A	A	B
70153	210055186902	2022	4	M	2	3	1	2	0	2	...	A	B	E	A	C	B
70154	210055242439	2022	2	M	1	3	1	2	0	2	...	B	A	E	A	A	B
70155	210056141525	2022	4	M	1	3	1	2	0	3	...	B	B	C	A	B	B

Fonte: Elaborado pelo Autor.