INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA
DE SÃO PAULO
CÂMPUS GUARULHOS

BIANCA TOLEDO
GISELLE SILVA

Ataques Cardíacos em Mulheres Jovens: Uma Análise Baseada no SIM/DATASUS utilizando Mapas Auto-Organizáveis (SOM)

GUARULHOS
2024

BIANCA TOLEDO
GISELLE SILVA

Ataques Cardíacos em Mulheres Jovens: Uma Análise Baseada no SIM/DATASUS utilizando Mapas Auto-Organizáveis (SOM)

Relatório Técnico apresentado ao Instituto Federal de Educação, Ciência e Tecnologia de São Paulo, como parte dos requisitos para a obtenção do grau de Técnolo em Análise e desenvolvimento de sistemas.

Orientador: Prof. Cleber Silva de Oliveira

GUARULHOS
2024

SUMÁRIO

1 INTRODUÇÃO

As doenças cardiovasculares representam a principal causa de mortalidade no Brasil, com o Infarto Agudo do Miocárdio (IAM) respondendo por grande parte dos óbitos relacionados a condições cardíacas. Em 2022, as cardiopatias foram responsáveis por aproximadamente 400 mil mortes no país, um número que ressalta a magnitude da ameaça para a saúde pública brasileira, comparando-se, inclusive, ao número de óbitos registrados durante o pico da pandemia de COVID-19 (Cardiology, 2023).

O Infarto Agudo do Miocárdio, dada sua gravidade e alta taxa de mortalidade, é um dos principais focos de atenção das políticas de saúde pública. Anualmente, o Brasil registra entre 300 mil e 400 mil casos de infarto, com uma taxa de mortalidade significativa, de aproximadamente um óbito a cada 5 a 7 casos (Saúde, 2024).

Tradicionalmente, as estratégias de prevenção e tratamento focam em grupos mais velhos, porém, as mulheres jovens também estão em risco e frequentemente são negligenciadas nas abordagens clínicas. Essa subnotificação e subestimação do risco fazem com que essas mulheres permaneçam em um grupo vulnerável, sem a devida atenção em termos de diagnóstico precoce e prevenção (Associação Paulista de Medicina, 2023).

Diante dessa lacuna, este projeto tem como objetivo realizar uma análise aprofundada da prevalência de óbitos por ataque cardíaco entre mulheres jovens no Brasil, utilizando a vasta base de dados do Sistema de Informações sobre Mortalidade (SIM). Para superar as limitações das análises estatísticas tradicionais na identificação de padrões complexos, empregar-se-á a metodologia de Mapas Auto-Organizáveis (SOM). Através do treinamento do modelo, busca-se agrupar os dados em neurônios que representam padrões de características comuns entre os pacientes. A subsequente análise e interpretação desses neurônios e dos padrões presentes em cada cluster permitirá identificar possíveis características que se manifestam de forma coesa, revelando padrões de dados. A pesquisa visa, assim, trazer à luz estruturas e relações nos dados que podem não ser evidentes por meio de abordagens exploratórias convencionais, fornecendo bases para futuras investigações sobre perfis de risco ao Infarto Agudo do Miocárdio.

Com base na identificação e interpretação detalhada desses padrões de clusterização, este estudo busca contribuir significativamente para a produção de conhecimento na área de saúde pública e epidemiologia, especificamente sobre a mortalidade por infarto agudo do miocárdio em mulheres jovens. A expectativa é que os resultados gerados por este trabalho possam fornecer informações valiosas que subsidiem futuras pesquisas e discussões para a formulação de estratégias mais eficazes, adaptadas às características específicas dessa população vulnerável.

1.1 Justificativa Teórica

O Brasil é um dos poucos países que coleta dados de saúde de forma sistemática em todo o território nacional, o que confere uma riqueza ímpar à base de dados do Sistema de Informações sobre Mortalidade (SIM). A análise desses dados é fundamental para identificar tendências e padrões de saúde, contribuindo para o aprimoramento das políticas públicas, especialmente em relação às doenças cardiovasculares, que continuam sendo a principal causa de morte no país (Ciências, 2024).

Para o estudo das doenças cardiovasculares, a compreensão de conceitos como prevalência e incidência é fundamental. Segundo o Centro de Atenção Psicossocial, a prevalência é uma medida que representa o número total de casos de uma doença em uma população em um determinado momento, enquanto a incidência se refere à quantidade de novos casos que surgem durante um período específico. Ambas as métricas são cruciais para dimensionar o impacto das doenças cardíacas e entender sua evolução (Psicossocial, 2024).

- **Prevalência** = (número de casos existentes em uma população / número total de indivíduos da população) x 100 (%).
- **Incidência** = (número de casos novos em um intervalo de tempo / número total de indivíduos expostos ao risco) x 100 (%).

Essas métricas permitem não apenas medir o impacto das doenças cardiovasculares na população, mas também identificar grupos específicos em risco, como mulheres jovens, que podem ser afetadas de maneira diferente pela condição.

Para garantir a confiabilidade dos dados analisados, é importante destacar que, de acordo com o Conselho Regional de Medicina do Estado do Paraná, desde 1976, o Ministério da Saúde adota um modelo de Declaração de Óbito (DO) padrão em todo o território nacional, alimentando o Sistema de Informações sobre Mortalidade (SIM). As estatísticas geradas a partir desses dados são essenciais para o monitoramento da saúde pública, incluindo políticas de vigilância e avaliação. Além disso, a DO possui caráter jurídico, conforme a Lei dos Registros Públicos (Lei nº 6.015/1973), sendo essencial para a lavratura da Certidão de Óbito. (Transmissíveis, 2011).

A escolha do SIM como base de dados para esta pesquisa se deve à sua abrangência e à qualidade das informações coletadas. O SIM fornece informações abrangentes sobre os óbitos no Brasil, incluindo dados demográficos, geográficos e causas de morte. Esses dados são fundamentais para análises epidemiológicas e são validados conforme critérios rigorosos, garantindo a confiabilidade dos resultados. Para realizar uma análise aprofundada e identificar padrões complexos não evidentes em abordagens estatísticas tradicionais, este estudo emprega ferramentas de programação em Python, utilizando o ambiente Jupyter Notebook, e a metodologia de Mapas

Auto-Organizáveis (SOM), implementada com a biblioteca MiniSOM e outras bibliotecas Python relevantes. Essa abordagem de inteligência artificial possibilita uma exploração mais robusta e uma apresentação dos dados de forma estruturada, facilitando a interpretação e a comunicação dos padrões encontrados.

Este estudo focará na análise de padrões de mortalidade de infarto agudo do miocárdio em mulheres jovens, um grupo muitas vezes negligenciado nas abordagens clínicas. A expectativa é que os resultados obtidos, através da clusterização dos dados por SOM, gerem insights valiosos sobre as características dessa condição em uma população específica, contribuindo para a produção de conhecimento que possa subsidiar futuras estratégias mais eficazes de prevenção e intervenção.

2 OBJETIVOS

2.1 Objetivos gerais

Este projeto tem como objetivo identificar e analisar padrões de vulnerabilidade em óbitos por infarto agudo do miocárdio em mulheres jovens no Brasil, utilizando a metodologia de Mapas Auto-Organizáveis (SOM) a partir de dados do Sistema de Informações sobre Mortalidade (SIM) do DATASUS.

2.2 Objetivos específicos

1. Pré-processar os dados: Organizar e preparar a base de dados do Sistema de Informações sobre Mortalidade (SIM) do DATASUS, filtrando os registros de óbitos por sexo feminino, faixa etária de 18 a 40 anos e causa de morte relacionada a ataque cardíaco e realizar o tratamento necessário para a aplicação dos Mapas Auto-Organizáveis (SOM) em ambiente Python.

2. Aplicar o SOM: Treinar um Mapa Auto-Organizável (SOM) com os dados pré-processados, visando agrupar os casos de óbitos por infarto agudo do miocárdio em neurônios que concentrem registros com características semelhantes, permitindo a identificação de agrupamentos intrínsecos nos dados.

3. Analisar os padrões intrínsecos: Interpretar os agrupamentos (clusters) formados nos neurônios do SOM, identificando e descrevendo as características predominantes que emergem desses agrupamentos, visando compreender os perfis de óbitos por infarto agudo do miocárdio em mulheres jovens.

4. Discutir os achados com apoio especializado: Apresentar e discutir os padrões identificados nos dados com o apoio de um profissional da saúde, buscando validações e observações clínicas sobre as características emergentes dos perfis de óbitos.

3 MÉTODO DE DESENVOLVIMENTO

3.1 Metodologia

O processo de metodologia foca em um modelo incremental, no qual as atividades foram divididas em ciclos de desenvolvimento, permitindo retorno aos ciclos anteriores para ajustes e melhorias, garantindo flexibilidade e otimização contínua do processo de pesquisa. Essa estrutura se alinha com princípios de desenvolvimento ágil, comuns em projetos de ciências de dados (PRESSMAN Roger S.; MAXIM, 2020).

- **Extração**: Extração e organização dos dados obtidos da base DATASUS para o MySQL através de um sistema ETL projetado em C.
- **Planejamento**: Determinação de escopo, hospedagem e acessibilidade dos dados e seleção de ferramentas.
- **Análise e Interpretação**: Elaboração de consultas SQL e aplicação de Mapas Auto-Organizáveis (SOM) para análise dos dados. Envolve o desenvolvimento, teste e validação de hipóteses relevantes ao contexto estudado.
- **Visualização**: Geração de gráficos e representações visuais para facilitar a compreensão e comunicação dos resultados obtidos.

3.2 Tecnologias

A seleção das tecnologias que foram utilizadas no projeto visa segurança, eficiência, praticidade e performance.

- ETL: Implementação de um sistema ETL em C que visou uma boa performance para a extração, transformação e carregamento dos dados para o MySQL, com processos de limpeza e normalização adequados para a grande quantidade de dados que podem variar em formato entre diferentes anos.
- SQL: Tecnologia utilizada para manipulação e consultas para os dados extraídos, proporcionando uma estrutura organizada e eficiente. A capacidade de manipular grandes quantidades de dados foi o motivo da escolha dessa ferramenta.
- Python: A escolha do python basea-se na sua simplicidade e vasta coleção de bibliotecas especializadas para analíse de dados, como numpy e pandas (também usadas no projeto) (MCKINNEY, 2017).

- Jupyter: Ferramenta para visualização interativa de dados, selecionada por sua flexibilidade e execução de comandos passo a passo, além da facilidade na criação de gráficos personalizados, , sendo uma ferramenta poderosa para a exploração de conjuntos de dados complexos.
- Minisom: : Biblioteca escolhida por sua estabilidade e por permitir a implementação de Mapas Auto-Organizáveis (SOM - Self-Organizing Maps), uma técnica de aprendizado não supervisionado utilizada para redução de dimensionalidade e descoberta de padrões ou agrupamentos ocultos em grandes volumes de dados (KOHONEN, 2001).

3.3 Sistema ETL

O sistema ETL foi desenvolvido em C para processar dados do Sistema de Informações sobre Mortalidade (SIM) e coloc-los no MySQL de forma eficiente, dividindo-se em três etapas principais:

3.3.1 Extração e Normalização

Na primeira etapa, os dados brutos em formato CSV são extraídos e convertidos para a codificação UTF-8, garantindo a integridade dos arquivos e a correta interpretação dos caracteres.

3.3.2 Transformação

Nesta etapa, os dados extraídos são preparados para o banco de dados de destino. As colunas dos arquivos CSV geram tabelas no banco, e os dados são limpos e formatados conforme as exigências.

3.3.3 Carga no Banco de Dados

Os dados transformados são carregados no banco de dados MySQL. Inserções agrupadas e transações são utilizadas para otimizar a performance, especialmente com grandes volumes de dados.

3.3.4 Funcionalidades do Sistema

- Automação Completa: Todos os processos são automatizados, minimizando erros manuais.
- Normalização e Estruturação: Conversão e limpeza automáticas dos dados.
- Inserções Otimizadas: Uso de inserções agrupadas para melhorar a performance.

3.3.5 Benefícios do Sistema

- Escalabilidade: Capacidade de processar volumes crescentes de dados.
- Flexibilidade: Adaptação a mudanças nos dados ou regras de validação.
- Confiabilidade: Garantia de consistência e integridade dos dados.

3.3.6 Tecnologia e Desempenho

O sistema foi desenvolvido utilizando Shell Script e C. O Shell Script automatiza o download e a organização dos dados, enquanto C é utilizado para as etapas críticas de transformação e carga, resultando em um processamento concluído em um único dia — muito mais rápido que soluções em Java, PHP ou Python.

Este sistema é escalável e flexível, sendo ideal para o processamento, preparo e download de grandes volumes de dados para o ínicio da analíse.

3.3.7 Base de Dados

Como mencionado anteriormente, os dados utilizados neste trabalho são provenientes da base de mortalidade do Sistema Único de Saúde (SUS), disponibilizada pelo DATASUS. Essa base foi desenvolvida com o objetivo de subsidiar análises sobre a situação sanitária da população brasileira, possibilitando a formulação de políticas públicas, a tomada de decisões baseadas em evidências e o planejamento de ações de saúde.

Foram utilizados os registros do Sistema de Informações sobre Mortalidade (SIM) referentes ao período de 2012 a 2020, abrangendo quase uma década de dados sobre os óbitos registrados em todo o território nacional. A base de dados, em suas versões anuais, apresenta em média 68 colunas, contendo informações demográficas, clínicas e administrativas.

Abaixo, destacamos alguns dos principais campos disponíveis:

- **ANO_OBITO**: ano em que ocorreu o óbito;
- **SEXO**: sexo da pessoa falecida;
- **IDADE**: idade da pessoa no momento do óbito;
- **CAUSABAS**: causa básica da morte, codificada segundo a Classificação Internacional de Doenças (CID-10);
- **LINHAA, LINHAB, LINHAC, LINHAD**: causas múltiplas registradas nas linhas da declaração de óbito, também codificadas em CID-10;

- **RACACOR**: raça/cor da pessoa falecida;
- **UF**: unidade federativa onde ocorreu o óbito;
- **NATURALIDADE**: local de nascimento da pessoa falecida;
- **ESC**: escolaridade da pessoa falecida;
- **LOCOCOR**: local de ocorrência do óbito (hospital, domicílio, via pública, entre outros);
- **DT_OBITO**: data do óbito;
- **DT_NASC**: data de nascimento da pessoa falecida;
- **CODMUNRES**: código do município de residência.

Os dados contidos na base apresentam diferentes tipos, como:

- **Categóricos**: como `SEXO`, `RACACOR`, `UF`, `LOCOCOR`, que representam informações qualitativas;
- **Numéricos**: como `IDADE`, representando valores quantitativos discretos;
- **Datas**: como `DT_OBITO` e `DT_NASC`, no formato dia/mês/ano;
- **Códigos alfanuméricos**: como `CAUSABAS`, `LINHAA` até `LINHAD`, que representam códigos CID-10.

Para esta etapa inicial da pesquisa, os campos de maior relevância foram `CAUSABAS`, `LINHAA`, `LINHAB`, `LINHAC` e `LINHAD`, uma vez que contêm as informações referentes às causas básicas e associadas dos óbitos, fundamentais para a análise com o algoritmo SOM.

3.3.8 Algoritmo Self-Organizing Map (SOM)

O mapa SOM foi escolhido como o primeiro algoritmo de treinamento neste trabalho. O *Self-Organizing Map* (SOM) é um tipo de rede neural que aprende de forma **não supervisionada**. Isso quer dizer que ele é capaz de encontrar padrões nos dados **sem precisar de rótulos ou categorias já conhecidas**.

O principal objetivo do SOM é organizar os dados em um mapa de duas dimensões, agrupando informações parecidas próximas umas das outras. Dessa forma, é possível visualizar como os dados estão distribuídos e identificar grupos com características semelhantes.

Ao inserir os dados de entrada no SOM, é possível definir o tamanho do *grid* (malha) que representará o mapa. Cada ponto dessa malha representa um **neurônio**,

que é uma unidade computacional associada a um vetor de **pesos** de mesma dimensão que os dados de entrada.

Pesos são valores numéricos ajustáveis atribuídos a cada neurônio, que representam a "posição"do neurônio no espaço dos dados. Inicialmente definidos de forma aleatória, esses pesos são responsáveis por determinar como cada neurônio "responde"a uma entrada. Durante o treinamento, os pesos são atualizados para se tornarem mais semelhantes às amostras apresentadas, permitindo que o neurônio aprenda a representar uma determinada região dos dados.

Durante o treinamento, para cada instância do conjunto de dados, o algoritmo calcula a **distância entre os dados e os pesos de todos os neurônios**. O neurônio cuja distância for a menor é chamado de **Best Matching Unit (BMU)**, ou seja, o neurônio mais semelhante à entrada. Os pesos do BMU e de seus vizinhos no mapa são então ajustados para se tornarem ainda mais parecidos com a entrada apresentada.

As **métricas de distância** utilizadas nesse processo influenciam diretamente na definição do BMU e, portanto, na forma como os dados são organizados no mapa. Neste trabalho, testamos três métricas:

- **Distância Euclidiana**: mede a raiz quadrada da soma das diferenças ao quadrado entre os elementos dos vetores;
- **Distância Manhattan**: mede a soma dos módulos das diferenças absolutas entre os vetores;
- **Distância do Cosseno**: mede o cosseno do ângulo entre dois vetores, refletindo sua orientação e não necessariamente sua magnitude.

Ao final do treinamento, o SOM forma **clusters de dados semelhantes**, agrupando entradas que ativam os mesmos neurônios ou neurônios vizinhos no mapa. Esses agrupamentos podem ser visualizados por meio de gráficos como o mapa de ativação e a **U-Matrix**, facilitando a identificação de padrões ou grupos com características semelhantes dentro do conjunto de dados.

3.4 Visão Geral do Projeto

Este projeto tem como objetivo identificar e analisar padrões de vulnerabilidade em óbitos por infarto agudo do miocárdio em mulheres jovens no Brasil. Utilizando a vasta base de dados do Sistema de Informações sobre Mortalidade (SIM) do DATASUS, a pesquisa empregará a metodologia de Mapas Auto-Organizáveis (SOM) para agrupar os casos de óbitos com características semelhantes. A análise e interpretação desses agrupamentos permitirá compreender os perfis que emergem dos dados, revelando estruturas e relações que podem indicar maior vulnerabilidade. A condução da pesquisa

será realizada em ambiente Python, utilizando Jupyter Notebook e a biblioteca MiniSOM, entre outras ferramentas. O intuito é gerar conhecimento específico sobre a mortalidade cardiovascular em mulheres jovens, fornecendo informações valiosas que poderão subsidiar futuras pesquisas e discussões na área da saúde pública.

3.5 **Processos**

Os dados passaram por um processo de ETL (Extração, Transformação e Carga) para garantir que estejam prontos para análise.

O diagrama mostra o fluxo do sistema ETL, detalhando cada etapa do processamento dos dados do Sistema de Informações sobre Mortalidade (SIM). O processo começa com a busca e verificação dos arquivos de origem, seja de um diretório local ou remoto.

Após localizar os arquivos, o sistema verifica se todos estão disponíveis e se estão codificados em UTF-8. Caso contrário, converte-os para garantir compatibilidade. Em seguida, os arquivos são carregados na memória e validados para garantir que estejam organizados corretamente, com o número certo de colunas e registros.

Depois de validados, os dados são carregados em uma tabela temporária no banco de dados. O sistema compara as linhas do arquivo com os registros inseridos e gera logs se houver erros. Se tudo estiver certo, a carga é registrada como bem-sucedida.

O próximo passo é garantir a integridade dos dados, verificando novamente e gerando um relatório de carga. Após isso, os dados são transferidos para a tabela definitiva no banco de dados e a tabela temporária é removida.

Por fim, os dados são normalizados, ajustando formatos de data e padronizando nomenclaturas. Um relatório final é gerado e o processo é concluído, com os dados prontos para análise.

Nessa fase, foi realizada a filtragem específica dos registros para delimitar o escopo da pesquisa. Foram selecionados apenas os registros de óbitos que atendiam aos seguintes critérios:

- **Sexo:** = Feminino.
- **Faixa Etária:** = Entre 18 e 40 anos.
- **Causa Básica do Óbito:** = Relacionada a ataques cardíacos (Infarto Agudo do Miocárdio - IAM), conforme códigos da Classificação Internacional de Doenças (CID-10).

Do banco de dados, os dados já filtrados e com tratamento inicial foram então extraídos e carregados para o ambiente Python, utilizando o Jupyter Notebook. Nessa

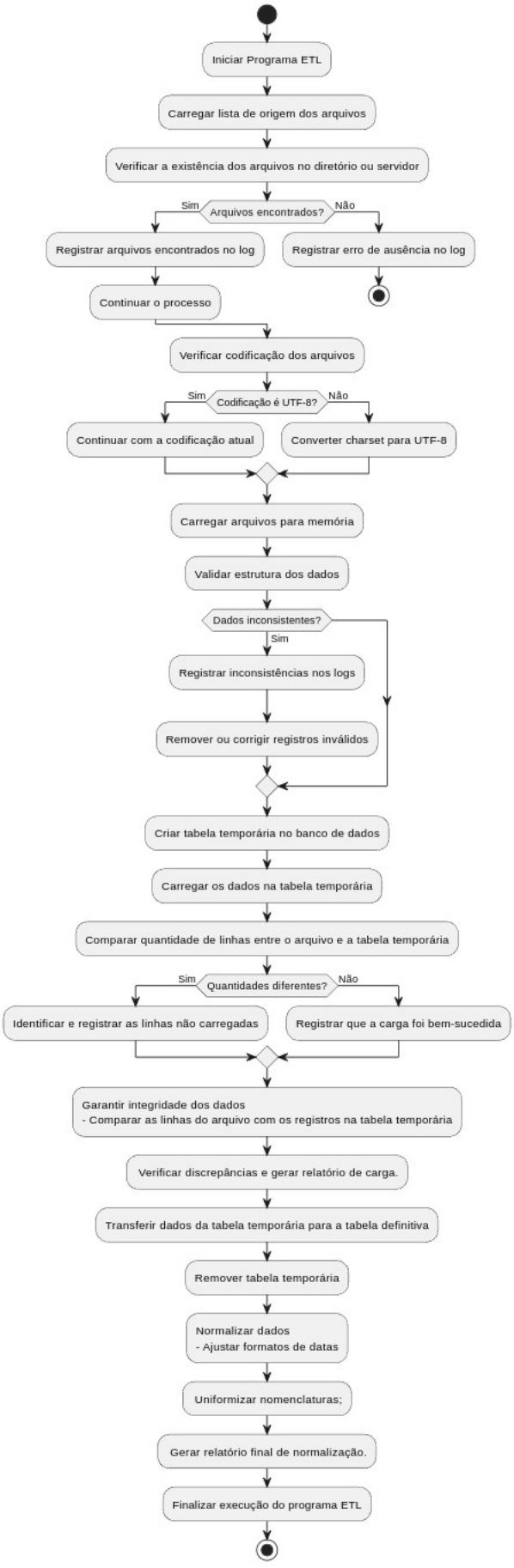

Figura 1 – Diagrama de Atividades do Processo ETL

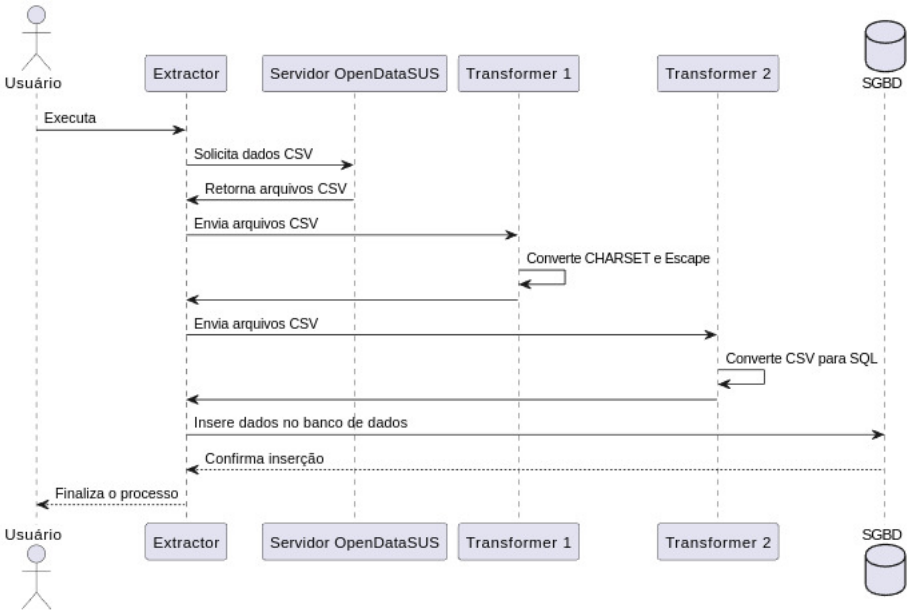

Figura 2 – Diagrama de sequência

etapa, as variáveis (colunas) relevantes para a análise com o SOM foram selecionadas para o escopo final do modelo. Foi realizado um tratamento e uma análise superficial inicial desses dados para identificar a distribuição e a qualidade das informações.

Crucialmente, os códigos da Classificação Internacional de Doenças (CID-10), que são variáveis categóricas, foram transformados em representações numéricas para viabilizar sua utilização como entrada no Mapa Auto-Organizável (SOM). Observou-se que algumas colunas da base de dados continham múltiplos códigos CID em um único campo. Para lidar com essa situação, esses códigos foram separados e distribuídos em colunas derivadas, como `LINHAA1`, `LINHAB1`, entre outras, de forma a preservar a informação de maneira estruturada. Após esse processo, os dados foram devidamente pré-processados e carregados em DataFrames do Pandas, ficando prontos para o treinamento do Mapa Auto-Organizável.

3.6 Implementando o SOM

Com a base de dados já preparada e transformada inteiramente em formato numérico, iniciamos a seleção dos parâmetros mais relevantes para a aplicação do Self-Organizing Map (SOM). No total, como citado anteriormente, testamos três tipos de métricas de distância para avaliar sua eficácia na clusterização dos dados:

- **Distância Euclidiana:** é a métrica mais comum, que calcula a raiz quadrada da soma das diferenças ao quadrado entre os elementos correspondentes de dois

vetores. Representa a “distância em linha reta” entre dois pontos no espaço.

$$d_{\text{euclidiana}}(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

- **Distância Manhattan:** também conhecida como distância de “manhattam” ou “city-block”, calcula a soma dos valores absolutos das diferenças entre os elementos correspondentes. Representa a distância percorrida em caminhos ortogonais (como em ruas de uma cidade).

$$d_{\text{manhattan}}(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

- **Distância Cosseno:** mede a similaridade entre dois vetores com base no cosseno do ângulo entre eles, ao invés da diferença absoluta. É útil para comparar a orientação dos vetores, sendo muito usada em problemas de alta dimensionalidade.

$$d_{\text{cosseno}}(x, y) = 1 - \frac{x \cdot y}{\|x\|\|y\|}$$

Inicialmente, utilizamos a métrica Euclidiana aplicando todas as variáveis (colunas) da base de dados. No entanto, o elevado número de instâncias comprometeu a eficiência da clusterização, resultando em um desempenho insatisfatório. Diante disso, optamos por testar outras métricas — Manhattan e Cosseno — além de reduzir o número de variáveis utilizadas.

A partir desse ponto, limitamos os testes às seguintes variáveis: **Causa Básica**, **Linha A**, **Linha B**, **Linha C** e **Linha D**, incluindo também as variáveis derivadas dessas após o processo de transformação dos dados.

Para analisar os resultados, empregamos as métricas Manhattan e Cosseno, a Frequência de Ativação e a U-Matrix.

A **Frequência de Ativação** mede quantas vezes um neurônio específico do SOM é o "melhor correspondente"para um determinado conjunto de dados, indicando a frequência com que ele representa um padrão nos dados.

Já a **U-Matrix** é um gráfico que visualiza a organização dos neurônios do SOM e suas diferenças. Pontos mais escuros na U-Matrix indicam maiores distâncias entre neurônios vizinhos, o que ajuda a identificar clusters de dados similares e separações entre esses grupos.

Com base nisso, observamos:

A análise do gráfico demonstra que a **métrica Cosseno** obteve o melhor desempenho, apresentando um menor erro de quantização e, consequentemente, uma clusterização mais assertiva. Diante disso, optamos por prosseguir os testes utilizando exclusivamente a métrica Cosseno.

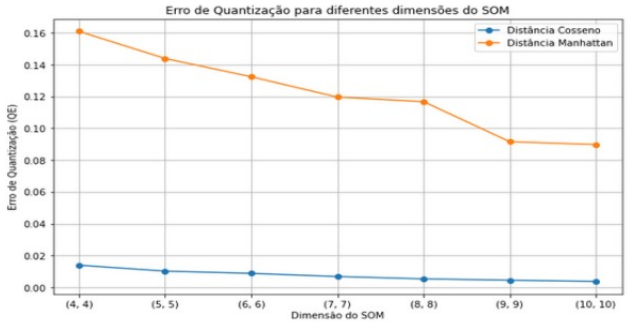

Figura 3 – Erro de Quantização — Manhattan e Cosseno

Em seguida, passamos à etapa de ajuste fino dos parâmetros do SOM com o objetivo de otimizar ainda mais os resultados. Ajustamos os seguintes parâmetros principais:

- **Sigma:** controla o raio de vizinhança dos neurônios durante o processo de treinamento. Valores maiores promovem uma influência mais ampla no início do treinamento, sendo progressivamente reduzidos.
- **Learning Rate:** define a taxa de aprendizado, ou seja, a intensidade com que os pesos dos neurônios são atualizados a cada iteração.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.preprocessing import MinMaxScaler
from minisom import MiniSom

dados = pd.read_csv('/home/giselle/Documentos/Mortalidade_dados_selecionados - dados.csv')

# Lista das variáveis específicas
variaveis_desejadas = [
    'LINHAA', 'LINHAA_1', 'LINHAA_2', 'LINHAA_3', 'LINHAB', 'LINHAB_1', 'LINHAC', 'LINHAC_1', 'LINHAC_2',
    'LINHAD', 'LINHAD_1', 'LINHAD_2','LINHAII', 'LINHAII_1', 'LINHAII_2', 'LINHAII_3', 'LINHAII_4', 'LINHAII_5',
    'CAUSABAS', 'CAUSABAS_O'
]

df_filtrado = dados[variaveis_desejadas].select_dtypes(include=['int64', 'float64'])
scaler = MinMaxScaler()
dados_normalizados = scaler.fit_transform(df_filtrado)

sigma = 1.0
learning_rate = 0.2
num_iteration = 50000
random_seed = 42

som1 = MiniSom(x=4, y=4, input_len=dados_normalizados.shape[1],
               sigma=sigma, learning_rate=learning_rate,
               random_seed=random_seed, activation_distance='cosine')
som1.random_weights_init(dados_normalizados)
som1.train_random(dados_normalizados, num_iteration=num_iteration)

som2 = MiniSom(x=4, y=4, input_len=dados_normalizados.shape[1],
               sigma=sigma, learning_rate=learning_rate,
               random_seed=random_seed, activation_distance='cosine')
som2.random_weights_init(dados_normalizados)
som2.train_random(dados_normalizados, num_iteration=num_iteration)

plt.figure(figsize=(20, 10))

plt.subplot(1, 2, 1)
frequencia = som1.activation_response(dados_normalizados)
plt.pcolor(frequencia.T, cmap='Blues')
plt.colorbar(label='Frequência de ativação')
plt.title('Mapa SOM - Frequência de ativação por neurônio')

plt.subplot(1, 2, 2)
plt.pcolor(som2.distance_map().T, cmap='bone_r')
plt.colorbar(label='Distância entre neurônios')
plt.title('Mapa SOM - U-Matrix com distância cosseno (variáveis selecionadas)')
plt.tight_layout()
plt.show()
```

Figura 4 – Configuração com métrica Cosseno

Ajustamos iterativamente os valores desses parâmetros, analisando os resultados de cada execução, até alcançarmos um modelo ideal, com a melhor qualidade de clusterização observada.

Como podemos observar nos gráficos apresentados, o gráfico da direita, conhecido como U-Matrix, foi essencial para confirmar a formação de **dois clusters**

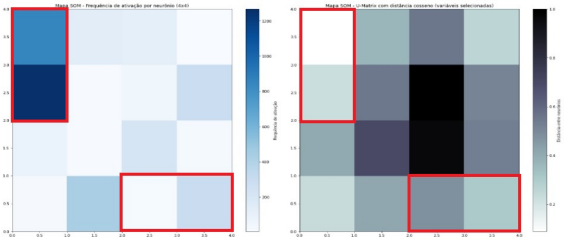

Figura 5 – Resultados finais - Frequência de ativação X U-Matrix

compostos pelos neurônios: (0,0), (0,1), (1,3) e (3,3). No gráfico de ativação (à esquerda), identificamos dois neurônios com forte ativação na região extrema esquerda. Essa mesma região, na U-Matrix, exibe uma coloração clara, indicando a semelhança entre esses neurônios e, consequentemente, sua pertinência ao mesmo grupo de dados com características próximas.

Com base nessa observação, aprofundamos a análise dos dados presentes nesses dois clusters.

4 RESULTADOS

Para aprofundar a análise das instâncias, foi desenvolvido um código para análise das intâncias de cada neurônio do mapa auto-organizável (SOM) selecionado.

O objetivo principal foi identificar as variáveis com maior peso nos neurônios que compõem os clusters formados, além de analisar os códigos da Classificação Internacional de Doenças (CIDs) associados às colunas de maior influência. Para isso, foram construídos gráficos de pizza, cujos resultados são apresentados a seguir:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.preprocessing import MinMaxScaler
from minisom import MiniSom

np.random.seed(42)

caminho_dataset = '/home/giselle/Documentos/Mortalidade_dados_selecionados - dados.csv'

dados = pd.read_csv(caminho_dataset)

variaveis_desejadas = [
    'LINHAA', 'LINHAA_1', 'LINHAA_2', 'LINHAA_3', 'LINHAB', 'LINHAB_1', 'LINHAC', 'LINHAC_1', 'LINHAC_2','LINHAD',
    'LINHAD_1', 'LINHAD_2', 'LINHAII', 'LINHAII_1', 'LINHAII_2', 'LINHAII_3', 'LINHAII_4', 'LINHAII_5',
    'CAUSABAS', 'CAUSABAS_O'
]

df_original = dados[variaveis_desejadas].select_dtypes(include=['int64', 'float64'])

scaler = MinMaxScaler()
dados_normalizados = scaler.fit_transform(df_original)

x, y = 4, 4
som = MiniSom(x=x, y=y,
              input_len=dados_normalizados.shape[1],
              sigma=1.0,
              learning_rate=0.2,
              activation_distance='cosine',
              random_seed=42)
som.random_weights_init(dados_normalizados)
som.train_random(dados_normalizados, num_iteration=50000)

pesos = som.get_weights()
cores = plt.cm.tab20(np.linspace(0, 1, dados_normalizados.shape[1]))
labels = df_original.columns.tolist()

plt.figure(figsize=(10, 10))
plt.pcolor(som.distance_map().T, cmap='bone_r', edgecolors='k')
plt.title('Composição dos pesos das variáveis em cada neurônio (SOM 4x4)')

for i in range(x):
    for j in range(y):
        neuron_weights = pesos[i, j]
        neuron_weights = neuron_weights / np.sum(neuron_weights)
        cx, cy = j, i
        pie_ax = plt.axes([cx/y, 1 - (cy+1)/x, 1/y, 1/x])
        pie_ax.pie(neuron_weights, colors=cores, startangle=90)
        pie_ax.set_xticks([])
        pie_ax.set_yticks([])

fig_legenda = plt.figure(figsize=(10, 2))
plt.legend(handles=[plt.Line2D([0], [0], color=c, lw=4) for c in cores],
           labels=labels, loc='center', ncol=4)
plt.axis('off')
plt.title("Legenda - Variáveis representadas por cor")
plt.show()
```

Figura 6 – Algoritmo para análise das instâncias de cada neurônio

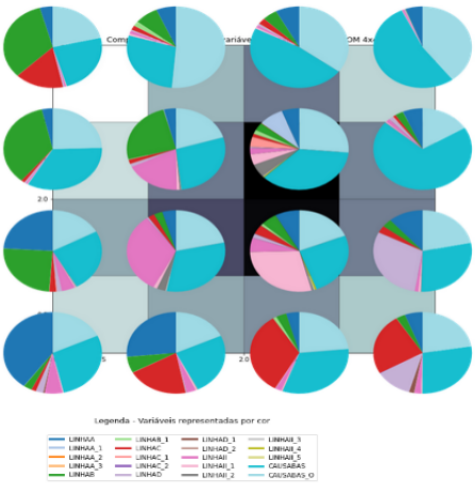

Figura 7 – Resultado do algoritmo para análise das instâncias de cada neurônio

Para uma melhor representação visual dos dados de cada instância, foram utilizados histogramas. O algoritmo de análise das instâncias revelou que, no primeiro neurônio (0,0), a coluna **LINHAB** exerceu maior influência. A LINHAB refere-se às

causas antecedentes ou consequenciais da morte. Ao consultar a LINHAB no neurônio (0,0), os seguintes códigos CID foram obtidos:

Código CID-10	Descrição
I219	Infarto agudo do miocárdio não especificado
I210	Infarto agudo transmural da parede anterior do miocárdio
I312	Hemopericárdio não classificado em outra parte
I119	Cardiopatia hipertensiva sem insuficiência cardíaca
I469	Parada cardíaca, não especificada
I212	Infarto agudo transmural do miocárdio em outras localizações
I259	Doença isquêmica crônica do coração, não especificada
I110	Cardiopatia hipertensiva com insuficiência cardíaca
I211	Infarto agudo transmural da parede inferior
R072	Dor precordial
R092	Parada respiratória
I674	Encefalopatia hipertensiva
I213	Infarto agudo transmural do miocárdio, localização não especificada
J189	Pneumonia, agente não especificado
I509	Insuficiência cardíaca, não especificada
I500	Insuficiência cardíaca congestiva

Figura 8 – Tabela de CID´s - Neurônio 0,0 LINHAB

A LINHAB também demonstrou maior influência no neurônio (0,1). Os CIDs correspondentes são:

CID-10	Descrição da Doença
R51X	Cefaleia (dor de cabeça)
R58X	Hemorragia não especificada
I10X	Hipertensão essencial (primária)
J81X	Edema pulmonar
J90X	Derrame pleural não classificado em outra parte
I48X	Fibrilação e flutter atrial
I21X	Infarto agudo do miocárdio

Figura 9 – Tabela de CID´s - Neurônio 0,1 LINHAB

Já no neurônio (1,3), a LINHAC apresentou maior influência, com os seguintes resultados:

Por fim, no neurônio (3,3), a LINHAC também se destacou como a de maior influência:

CID-10	Descrição da Doença
I219	Infarto agudo do miocárdio não especificado
I519	Doença não especificada do coração
I517	Cardiomegalia (aumento do coração)
I119	Doença cardíaca hipertensiva sem insuficiência cardíaca (congestiva)
J969	Transtorno respiratório não especificado
I229	Infarto do miocárdio recorrente de localização não especificada
I251	Doença aterosclerótica do coração

Figura 10 – Tabela de CID´s - Neurônio 0,1 LINHAB

CID-10	Descrição da Doença
J81X	Edema pulmonar
I10X	Hipertensão essencial (primária)
E86X	Depleção de volume (ou desidratação)

Figura 11 – Tabela de CID´s - Neurônio 1,3 LINHAB

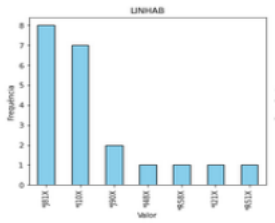

Figura 12 – Histograma da coluna LINHAB do neurônio 0,0

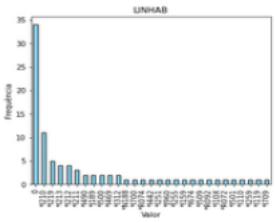

Figura 13 – Histograma da coluna LINHAB do neurônio 0,1

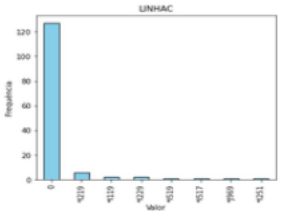

Figura 14 – Histograma da coluna LINHAC do neurônio 1,3

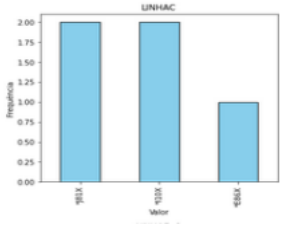

Figura 15 – Histograma da coluna LINHAC do neurônio 3,3

5 CONCLUSÕES

A análise dos neurônios gerados pelo mapa auto-organizável (SOM) permitiu identificar padrões relevantes nos dados de mortalidade por doenças cardíacas, com destaque para códigos CID-10 como I21X (Infarto Agudo do Miocárdio), I48X (Fibrilação e flutter atrial) e I10X (Hipertensão essencial). Esses agrupamentos sugerem possíveis relações clínicas entre causas de óbito, reforçando a capacidade do SOM em detectar estruturas latentes nos dados.

Contudo, a interpretação médica desses agrupamentos é essencial. A presença de condições como edema pulmonar (J81X) e derrame pleural (J90X), comumente associadas a insuficiências cardíacas, indica a necessidade de aprofundar a análise em parceria com profissionais da saúde, como cardiologistas e epidemiologistas.

Como grupo de pesquisa, reforçamos o compromisso de dar continuidade ao estudo, expandindo a base de dados, incorporando novas variáveis (como idade, sexo e região) e validando os achados com especialistas. O objetivo é contribuir com o entendimento dos padrões de mortalidade cardiovascular e, futuramente, apoiar políticas públicas e estratégias de prevenção.

REFERÊNCIAS

Associação Paulista de Medicina. **Problemas cardíacos nas mulheres são negligenciados**. 2023. <https://www.apm.org.br/problemas-cardiacos-nas-mulheres-sao-negligenciados/>. Acesso em: 24 Mai. 2025. 5

CARDIOLOGY, J. of the American College of. **Carga Global de Doenças e Fatores de Risco Cardiovasculares**. 2023. <https://revistapesquisa.fapesp.br/cerca-de-400-mil-pessoas-morreram-em-2022-no-brasil-por-problemas-cardiovasculares/>. Acesso em: 26 nov. 2024. 5

CIêNCIAS, A. B. de. **A importância da análise de dados na saúde**. 2024. Disponível em: <https://bvsms.saude.gov.br/bvs/publicacoes/manualdeclaracaoobitos.pdf>. 6

KOHONEN, T. **Self-Organizing Maps**. 2001. Livro. Springer: Berlin. 10

MCKINNEY, W. **Python for Data Analysis: Data Wrangling with Pandas, Numpy, and Python**. 2017. Livro. O´Reilly Media: Sabastopol, CA. 9

PRESSMAN ROGER S.; MAXIM, B. R. **Software Engineering: A Practitioner´s Approach**. 2020. Livro. McGraw-Hill Education: New York. 9

PSICOSSOCIAL, C. de A. **Qual a diferença entre prevalência e incidência?** 2024. <http://www.capcs.uerj.br/qual-a-diferenca-entre-prevalencia-e-incidencia/>. Acesso em: 27 nov. 2024. 6

SAúDE, B. M. da. **Infarto agudo do miocárdio**. 2024. <https://www.gov.br/saude/pt-br/assuntos/saude-de-a-a-z/i/infarto>. Acesso em: 26 nov. 2024. 5

TRANSMISSíVEIS, B. M. da Saúde. Secretaria de Vigilância em Saúde. Departamento de Análise em Saúde e Vigilância de D. N. **Manual de declaração de óbitos**. [S.l.]: Ministério da Saúde, 2011. <https://bvsms.saude.gov.br/bvs/publicacoes/manualdeclaracaoobitos.pdf>. Acesso em: 27 nov. 2024. 6