INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DE SÃO PAULO

CÂMPUS GUARULHOS

GABRIELA SANTANA CAMILO

GABRIEL VINÍCIUS ROCHA BARBOZA

Visualização de Padrões em Óbitos por Alzheimer no Brasil (2012–2022) com Mapas Auto-Organizáveis (SOMs)

GUARULHOS

2025

GABRIELA SANTANA CAMILO
GABRIEL VINÍCIUS ROCHA BARBOZA

Visualização de Padrões em Óbitos por Alzheimer no Brasil (2012–2022) com Mapas Auto-Organizáveis (SOMs)

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

Orientador: Professor Dr. Cleber Silva de Oliveira

Coorientador: Professor Dr. Thiago Schumacher Barcelos

GUARULHOS

2025

Catalogação na fonte
Biblioteca IFSP Câmpus Guarulhos
Dados fornecidos pelo(a) autor(a)

ERROCOD , IGOCUTTE R
_ / _. Guarulhos: [s.n.], 2024.
100 f. il.

Orientador: _
Co-orientador: _

Trabalho de Conclusão de Curso (Tecnologia em Análise e Desenvolvimento de Sistemas) - Instituto Federal de Educação, Ciência e Tecnologia de São Paulo, IFSP, 2024.

Inclui bibliografia.

1. _. 2. _. 3. _. 4. _. 5. _. I. Instituto Federal de Educação, Ciência e Tecnologia de São Paulo II. Título.

CDD 004.21

ATA N.º /2024 - DAE-GRU/DRG/GRU/IFSP

ATA DE DEFESA DE MONOGRAFIA

Na presente data realizou-se a sessão pública de defesa da Monografia intitulada

Membros	IES	Presença (Sim/Não)	Aprovação/Conceito *(Quando Exigido)*

Observações:

A banca examinadora, tendo terminado a apresentação do conteúdo da monografia, passou à arguição do candidato. Em seguida, os examinadores reuniram-se para avaliação e deram o parecer final sobre o trabalho apresentado pelo aluno, tendo sido atribuído o seguinte resultado:

[ ] Aprovado [ ] Reprovado

Proclamados os resultados pelo presidente da banca examinadora, foram encerrados os trabalhos e, para constar, eu lavrei a presente ata que assino juntamente com os demais membros da banca examinadora.

Campus Guarulhos

Documento assinado eletronicamente.

RESUMO

Este estudo analisou 211.658 óbitos por Doença de Alzheimer (DA) no Brasil (de 2012 à 2022) com foco em perfis de comorbidades, utilizando Mapas Auto-Organizáveis (SOMs). Após processo de ETL e pré-processamento de dados do SIM/DATASUS, aplicou-se um SOM com grade 7×7 e distância de cosseno. Três clusters foram identificados: (1) dominado pela DA como causa básica; (2) com alta frequência de causas mal definidas (R99X, R98X) e senilidade (R54X), sugerindo lacunas no preenchimento da Declaração de Óbito; e (3) marcado por comorbidades cardiovasculares (I10X) e complicações de imobilidade (L89X). Os achados evidenciam o potencial dos SOMs na identificação de padrões clínicos relevantes, destacando a importância de integrar o manejo de comorbidades cardiovasculares ao cuidado em DA e de aprimorar a vigilância epidemiológica para reduzir subnotificações.

Palavras-chave: Doença de Alzheimer; Comorbidades; Mapas Auto-Organizáveis (SOM); Mortalidade; Vigilância epidemiológica; Declaração de Óbito.

ABSTRACT

This study analyzed 211,658 deaths from Alzheimer's Disease (AD) in Brazil (2012–2022), focusing on comorbidity patterns using Self-Organizing Maps (SOMs). After an ETL process and preprocessing of data from the SIM/DATASUS system, a SOM with a 7×7 grid and cosine distance was applied. Three clusters were identified: (1) dominated by AD as the underlying cause; (2) with a high frequency of ill-defined causes (R99X, R98X) and senility (R54X), suggesting gaps in the completion of Death Certificates; and (3) marked by cardiovascular comorbidities (I10X) and immobility-related complications (L89X). The findings highlight the potential of SOMs to identify clinically relevant patterns, emphasizing the importance of integrating cardiovascular comorbidity management into AD care and improving epidemiological surveillance to reduce underreporting.

Keywords: Alzheimer's Disease; Comorbidities; Self-Organizing Maps (SOM); Mortality; Epidemiological Surveillance; Death Certificate.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

CID	Classificação Internacional de Doenças
DATASUS	Departamento de Informação e Informática do Sistema Único de Saúde
DA	Doença de Alzheimer
DO	Declaração de Óbito
ETL	Extração, Transformação e Carga
KDD	Knowledge Discovery in Databases ou Descoberta de Conhecimento em Bases de Dados
RF	Requisitos Funcionais
RNF	Requisitos Não Funcionais
SIM	Sistema de Informações sobre Mortalidade
SOM	Self Organizing-map ou Mapa Auto-organizável
SQL	Structured Query Language
SUS	Sistema Único de Saúde

SUMÁRIO

1 INTRODUÇÃO

A doença de Alzheimer é caracterizada pela perda progressiva da memória, da fala e da capacidade de planejamento, o que compromete a autonomia e a qualidade de vida dos pacientes (Teixeira et al., 2015). No Brasil, de 2000 a 2019, registraram-se 211 658 óbitos por Alzheimer, com aumentos médios anuais de 4,3% (60–69 anos), 8,1% (70–79 anos) e 11,3% (≥ 80 anos) em todas as regiões e gêneros (Paschalidis et al., 2023). Globalmente, a demência afetava 55 milhões de pessoas em 2019 e deverá alcançar 139 milhões até 2050, devido ao envelhecimento populacional.

Diante desse crescimento, torna-se imprescindível contar com sistemas de informação capazes de coletar e disponibilizar dados de mortalidade para orientar políticas públicas. Nesse sentido, o Departamento de Informação e Informática do Sistema Único de Saúde (DATASUS) e o Sistema de Informações sobre Mortalidade (SIM) inserem-se como ferramentas essenciais para a vigilância epidemiológica no Brasil. Criado em 1991 o Departamento de Informática do SUS (DATASUS) passou a gerenciar e aprimorar a informatização de sistemas essenciais como o Sistema de Informações sobre Mortalidade (SIM), que já reunia e padronizava os registros de óbito em todo o país desde sua implementação nacional em 1975 (BRASIL. Ministério da Saúde, 2025). Para assegurar a confiabilidade, os dados do SIM passam por três etapas de qualificação (prévio, preliminar e final) em parceria com Estados e Municípios, permitindo ao Ministério da Saúde publicar estatísticas confiáveis, o que torna o DATASUS uma fonte vital de informações que subsidiam análises objetivas da situação sanitária e embasam políticas públicas eficazes.

Com toda essa base de dados validada e disponível, destacam-se informações além da causa principal de óbito, investigando as doenças associadas que somam risco. Em registros de óbito do DATASUS, além da causa principal (Alzheimer), anotam-se também as comorbidades — ou seja, as condições médicas que contribuiram para a morte. Estudos de Wang et al. (2025) mostram que, entre 2 618 pacientes com Alzheimer, 55,1% tinham hipertensão, 38,2% osteoartrite, 32,3% depressão, 25,7% diabetes e 22,7% doenças cerebrovasculares, elevando o risco de morte. Além disso, infecções como pneumonia chegam a dobrar esse risco em pacientes com demência Manabe et al. (2019). Ao identificar essas associações, é possível planejar cuidados que tratem não só os sintomas de memória, mas também as demais condições clínicas associadas.

Para lidar com esse conjunto de variáveis complexas, a epidemiologia tradicional se beneficia de complementos analíticos. Assim, surge a necessidade de empregar

métodos de machine learning capazes de explorar relações complexas entre múltiplos indicadores. Enquanto os métodos supervisionados utilizam dados rotulados para tarefas como diagnóstico e prognóstico, os métodos não supervisionados identificam padrões sem categorias pré-definidas (Ono; Goto, 2022). Entre os algoritmos não supervisionados, é fundamental escolher técnicas que lidem bem com dados de alta dimensionalidade e que permitam interpretar visualmente os resultados, de modo a tornar os achados acessíveis a pesquisadores.

Nesse contexto, os Mapas Auto-organizáveis (Self-Organizing Maps – SOMs) destacam-se como uma técnica adequada, pois utilizam aprendizagem competitiva para organizar dados em um mapa bidimensional, preservando as relações topológicas originais (Petersen et al., 2024) e proporcionam uma visualização intuitiva de agrupamentos, além de reduzirem a dimensionalidade dos dados (Holt et al., 2023). Por essas razões, SOMs são úteis em saúde pública, onde os conjuntos de dados englobam variáveis clínicas, demográficas e epidemiológicas complexas, e já foram usados para identificar perfis de idosos (Parra-Rodríguez et al., 2022) e subgrupos de Alzheimer em exames de ressonância magnética (Petersen et al., 2024). Contudo, não há estudos que apliquem SOMs especificamente para analisar comorbidades em óbitos por Alzheimer.

Por isso, este estudo tem como objetivo demonstrar como os Mapas Auto-Organizáveis podem ser utilizados para identificar e visualizar os padrões mais frequentes de doenças contribuintes em óbitos por Alzheimer no Brasil, entre 2012 e 2022. Para isso, extrairemos e prepararemos dados de mortalidade do DATASUS em SQL, aplicaremos SOMs em Python para gerar mapas de comorbidades e validaremos os agrupamentos encontrados com estudos epidemiológicos anteriores.

Contudo, essa abordagem apresenta limitações: a mais notável é a ausência de informações clínicas detalhadas, como estágio da doença e tratamentos utilizados. Isso obriga os pesquisadores a formular hipóteses apenas a partir dos códigos de CID e dos registros de óbito, sem poder estabelecer a ordem temporal em que as doenças surgiram, o que dificulta a inferência de relações de causa e efeito.

2 OBJETIVOS

A seguir são apresentados os objetivos deste trabalho.

2.1 Objetivo Geral

Avaliar a aplicabilidade e a eficácia dos Mapas Auto-Organizáveis (SOMs) como ferramenta de análise exploratória para identificação e visualização de padrões de comorbidades em registros do Sistema de Informação sobre Mortalidade (SIM) do DATASUS referentes a óbitos por Alzheimer no Brasil (2012–2022), contribuindo para o avanço metodológico em análise de dados em saúde.

2.2 Objetivos Específicos

- Desenvolver e automatizar (por script) um processo de extração dos registros de óbitos do SIM de 2012 à 2022, assegurando a integridade dos arquivos baixados.
- Padronizar formatos, normalizar caractéres (UTF-8) e indexar campos-chave; integrar fontes auxiliares para aprimorar a decodificação de variáveis sociodemográficas e de comorbidades (CID-10) para preparar e enriquecer a base de dados.
- Filtrar registros de interesse, selecionando registros com códigos correspondentes a Alzheimer (F00.0, F00.1, F00.2, F00.9, G30.0, G30.1, G30.8, G30.9) e aplicar critérios de limpeza (remoção de duplicatas e inconsistências).
- Implementar pré-processamento nos dados para tratar valores ausentes, agrupar faixas etárias e regiões geográficas, e codificar variáveis de CID das comorbidades.
- Definir topologia, hiperparâmetros (tamanho da grade, taxa de aprendizado, raio de vizinhança) e treinar a rede neural de forma iterativa, documentando escolhas técnicas com o intuito de configurar e treinar o SOM.
- Produzir U-matrix, planos de componentes e mapas de clusters para facilitar a interpretação dos agrupamentos de comorbidades e gerar visualizações exploratórias.
- Confrontar os padrões identificados com achados de pesquisas epidemiológicas existentes, avaliando consistência e divergências para validação dos resultados encontrados.

- Documentar e disseminar o fluxo de trabalho, elaborando um Jupyter Notebook comentado com código, resultados e instruções de execução para garantir a transparência e a reprodutibilidade do estudo.

3 MÉTODO DE DESENVOLVIMENTO

A seguir são abordados os principais pontos para realização na análise de dados nesse projeto.

3.1 Visão Geral do Projeto

Este projeto de pesquisa e análise de dados foi concebido com o objetivo central de investigar e visualizar padrões de comorbidades associadas a óbitos por Doença de Alzheimer no Brasil, utilizando dados do Sistema de Informações sobre Mortalidade (SIM) do DATASUS, abrangendo o período de 2012 a 2022. Adotando o paradigma de Descoberta de Conhecimento em Bases de Dados (KDD), o estudo emprega uma abordagem iterativa e exploratória, com foco na aplicação de técnicas de aprendizado não supervisionado, especificamente os Mapas Auto-Organizáveis (SOM), para identificar correlações e agrupamentos entre a Classificação Internacional de Doenças (CID) referente ao Alzheimer e as demais CIDs registradas como causas associadas nos óbitos.

O fluxo de trabalho analítico foi estruturado em etapas metodológicas sequenciais e interconectadas. Iniciou-se com um processo robusto de Extração, Transformação e Carregamento (ETL), que envolveu a coleta automatizada dos dados brutos do SIM, sua conversão para o padrão UTF-8, transformação para o formato SQL e carregamento em um banco de dados MySQL. Seguiu-se uma fase de otimização e preparação da base de dados, com a indexação de colunas chave para ganho de performance em consultas e a filtragem específica dos registros de óbito cuja causa básica estava relacionada à Doença de Alzheimer, utilizando os códigos CID pertinentes (F00.0-F00.9, G30.0-G30.9).

Posteriormente, os dados filtrados foram submetidos a um tratamento e limpeza detalhados no ambiente Jupyter Notebook. Esta fase incluiu a gestão de dados ausentes através de descarte ou imputação criteriosa, a complexa tarefa de separação de múltiplos códigos CID presentes em um mesmo campo e a conversão destes para um formato numérico adequado aos algoritmos de aprendizado de máquina. Com os dados devidamente preparados e normalizados (utilizando MinMaxScaler), procedeu-se à aplicação do algoritmo SOM. Esta etapa envolveu uma experimentação sistemática com diferentes hiperparâmetros, como métricas de distância (Manhattan e Cosseno), tamanho da grade do mapa e número de iterações, utilizando visualizações como a U-Matrix e mapas de frequência de ativação para guiar o refinamento do modelo até a obtenção de uma configuração ótima (grade 7x7, distância de Cosseno).

Finalmente, os agrupamentos (clusters) de padrões de comorbidades gerados pelo SOM foram analisados em profundidade. Esta análise envolveu a segmentação dos registros de acordo com os clusters identificados, a decodificação dos códigos CID para sua forma original e a investigação da frequência e composição dos CIDs e capítulos CID em cada cluster, utilizando gráficos de pizza e de barras para facilitar a interpretação clínica dos padrões emergentes. O projeto visa, assim, não apenas aplicar uma técnica de inteligência artificial a dados de saúde, mas também demonstrar um processo completo de KDD, desde a aquisição dos dados até a geração de insights potencialmente valiosos para a compreensão das comorbidades na Doença de Alzheimer.

3.2 Requisitos

De acordo com Sommerville (2019), os requisitos são essenciais para descrever o que deve ser alcançado (requisitos funcionais) e quais restrições devem ser observadas (requisitos não funcionais). Em contextos de KDD (Knowledge Discovery in Databases), os requisitos funcionais correspondem às perguntas de pesquisa que guiarão a análise exploratória, enquanto os requisitos não funcionais definem atributos de qualidade — como desempenho do processo, confiabilidade, integridade e segurança dos dados — e se aplicam ao método e às rotinas adotadas pelos analistas.

Diante desse cenário, abaixo estão descritos os requisitos estruturados para esse projeto:

3.2.1 Requisitos Funcionais

RF1 Há alguma outra doença que é comum entre os registros de óbitos por Alzheimer?

RF2 Quais comorbidades tendem a ocorrer em conjunto (por exemplo, hipertensão e diabetes) entre os pacientes que morreram de Alzheimer?

RF3 Certas combinações de comorbidades (por exemplo, hipertensão + diabetes) estão associadas a maior probabilidade de óbito por Alzheimer do que outras combinações?

RF4 Há diferença no perfil de comorbidades por sexo (masculino versus feminino)?

RF5 Existem padrões de comorbidades? Eles são consistentes com achados de pesquisas epidemiológicas existentes sobre Alzheimer?

RF6 Qual faixa etária é predominante entre os óbitos causados por Alzheimer?

RF7 A ocupação exercida pelo indivíduo pode ter influenciado no óbito por Alzheimer?

RF8 Há relação entre a idade e as ocorrências de óbitos por Alzheimer?

RF9 Existe alguma relação entre o sexo do indivíduo e a incidência da doença ao longo dos anos?

RF10 Mudanças ao longo do tempo sugerem um aumento ou diminuição de óbitos por Alzheimer?

RF11 A ecolaridade tem relação com os óbitos por Alzheimer?

RF12 Quais grupos apresentam maior vulnerabilidade à mortalidade por Alzheimer?

RF13 Os casos de Alzheimer têm mostrado uma tendência de crescimento ao longo dos últimos anos?

3.2.2 Requisitos Não Funcionais

Categoria	Código	Descrição
Performance	RNF01	A normalização dos dados deve ser realizada em um tempo aceitável, garantindo que os dados estejam disponíveis para análise de forma ágil.
Compatibilidade	RNF02	Todos os dados devem ser compatíveis com o SGBD MySQL.
Segurança	RNF03	Os dados devem ser tratados com confidencialidade e respeitar as leis de privacidade e proteção de dados pessoais (LGPD).
Volumetria	RNF04	O sistema deve suportar grandes volumes de dados provenientes do SIM, garantindo a manipulação e processamento sem perdas ou falhas.
Confiabilidade	RNF05	Os resultados gerados pelas análises devem ser consistentes e reproduzíveis, garantindo a credibilidade dos dados e das informações extraídas.
Integridade	RNF07	Os dados devem ser preservados em sua totalidade durante todas as etapas de processamento, garantindo que não ocorram alterações ou perdas indevidas. Para isso, deve se utilizar a integridade da entidade, que depende de chaves e valores exclusivos criados para identificar dados, integridade referencial que garante que os registros em tabelas relacionadas estejam consistentes, mantendo a coerência entre os dados armazenados.

Tabela 1 – Requisitos Não Funcionais - Performance, Compatibilidade, Segurança, Volumetria, Confiabilidade, Integridade

Fonte: Elaborado pelos autores, 2025.

3.2.3 Fluxo de Inteligência Analítica Aplicada a Dados de Mortalidade

Esse fluxo apresenta as etapas, automatizadas ou realizadas manualmente, para extração, transformação, análise e validação de dados de mortalidade do SIM, com foco na identificação de padrões de comorbidades atreladas ao Alzheimer.

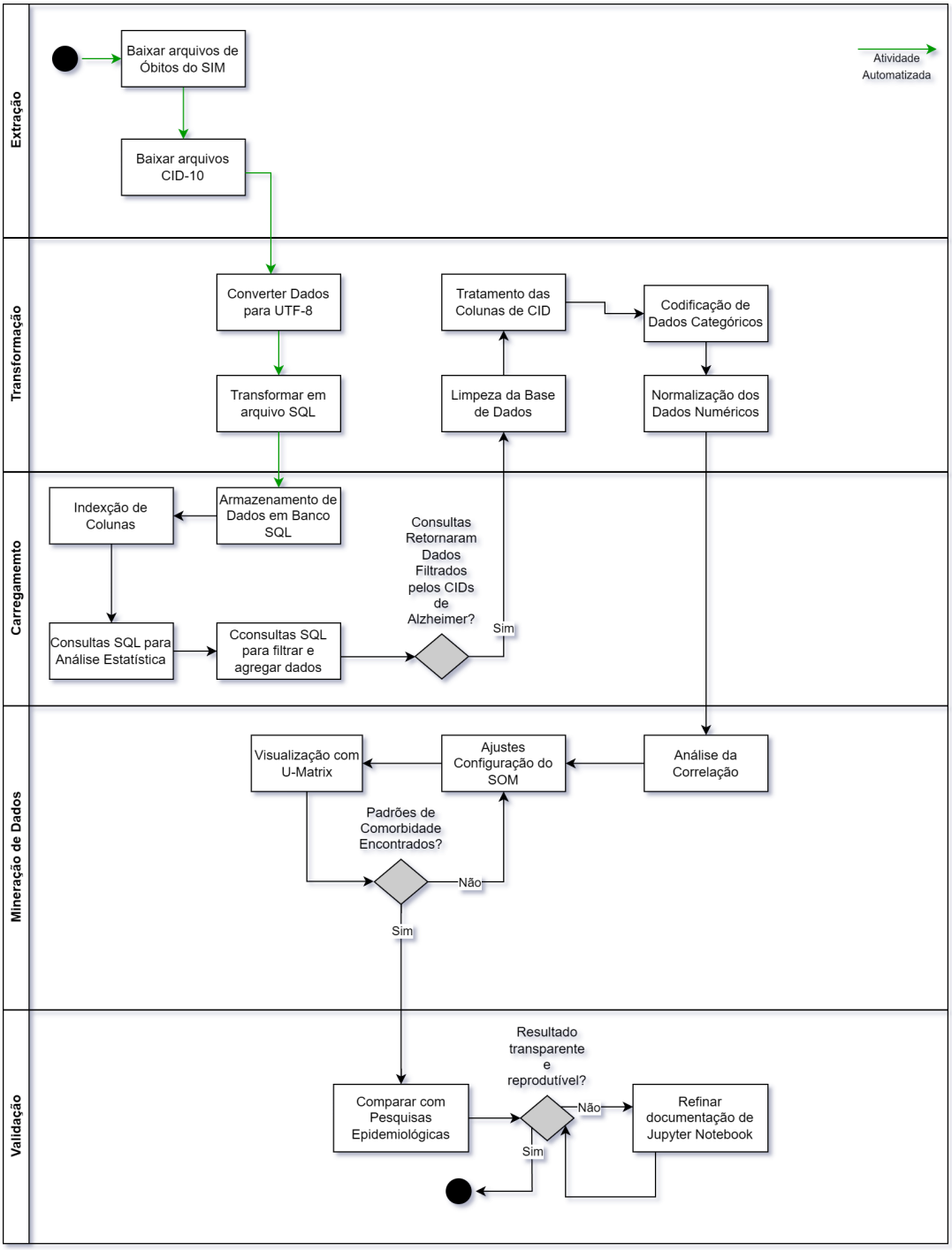

Figura 1 – Diagrama de Atividades do Fluxo de Inteligência Analítica Aplicada a Dados de Mortalidade

3.2.4 Interações entre Jupyter, ETL e Banco de Dados

O diagrama de sequência abaixo representa como o analista coordena as interações com as principais entidades (Script de ETL, Banco de dados e Jupyter Notebook) do fluxo completo de extração, tratamento e validação de dados de mortalidade com foco em padrões de comorbidades associadas ao Alzheimer.

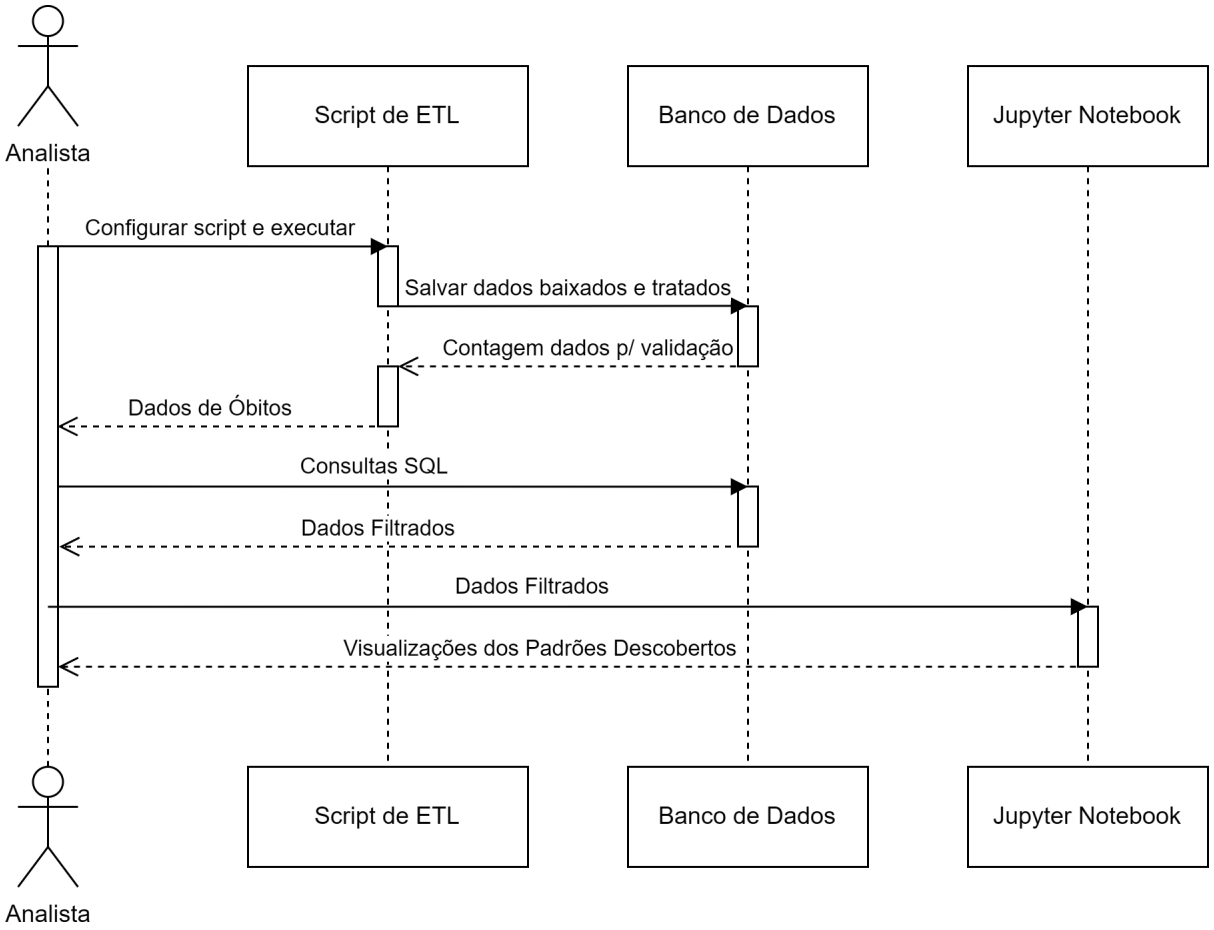

Figura 2 – Diagrama de Sequência de Interações entre Jupyter, ETL e Banco de Dados

3.3 Tecnologias

Nas próximas subseções são apresentadas as tecnologias aplicadas no projeto.

3.3.1 Linguagem C

A linguagem C é uma linguagem de programação de propósito geral, criado por Dennis Ritchie em 1970. A linguagem é utilizada no desenvolvimento de sistemas operacionais, drivers de dispositivos e pilhas de protocolos, onde o código fonte é compilado antes da execução, contribuindo, dessa forma, na eficiência no processamento de informações. Análises de desempenho demonstraram que essa tecnologia é

a mais eficiente para processar 10 anos de dados do Sistema de Informações sobre Mortalidade (SIM) e conversão para SQL, portanto, o script que automatiza o processo de extração dos registros de óbito do SIM será desenvolvido utilizando linguagem C pura.

3.3.2 SQL

Segundo a Oliveira (2023), SQL (Structured Query Language) é uma linguagem usada para lidar com bancos de dados, permitindo criar tabelas, inserir, atualizar ou remover dados e fazer consultas para encontrar informações específicas. Surgiu nos anos 1970 e se tornou uma ferramenta padrão para trabalhar com dados em sistemas como MySQL e Oracle. Usado tanto para gerenciar dados quanto para análises, o SQL facilita o uso de grandes quantidades de informações de forma prática e organizada, ajudando a extrair insights importantes do que está armazenado nos bancos de dados. Por essas razões, a linguagem SQL se tornou ideal para realizar a análise dos dados do Sistema de Informações sobre Mortalidade (SIM) do DATASUS nesse projeto.

3.3.3 MySQL Workbench

Para a Oracle (2024), MySQL Workbench é um Sistema Gerenciador de Banco de Dados (SGBD) usado para modelagem de dados, desenvolvimento de SQL e administração de bancos de dados. Ele facilita a criação de diagramas, que representam a estrutura do banco de dados, com recursos de visualização e ferramentas para a criação e edição de scripts SQL. Dessa forma, o MySQL Workbench se torna muito útil para a análise de dados nesse projeto com SQL.

3.3.4 Jupyter

O Jupyter Notebook é uma aplicação web que permite criar e compartilhar documentos interativos contendo código, equações, visualizações e texto narrativo. Ele suporta execução passo a passo, permitindo registrar toda a Análise Exploratória de Dados de forma reprodutível e documentada, facilitando a revisão e o compartilhamento entre pesquisadores de saúde (Pérez; Granger, 2015). Nesse estudo, é usado para construir um notebook com toda a análise exploratória detalhada de forma a contribuir com o avanço da análise de dados em saúde e ser facilmente executado com o kernel de Python3.

3.3.5 Python

Python é uma linguagem de alto nível, interpretada e muitas bibliotecas *open-source* (Foundation, 2024). No projeto, é usada para pré-processamento, transformação

e mineração de dados (incluindo Self-Organizing Maps via bibliotecas como `MiniSom`), permitindo pipelines automatizados que vão além do SQL. Além de apresentar um ecossistema maduro de pacotes científicos (NumPy, Pandas, Scikit-learn) para manipulação e análise estatística de grandes volumes de dados e a capacidade de gerar visualizações avançadas (Matplotlib, Seaborn, Altair).

3.4 Detalhes dos Dados Utilizados

O Ministério da Saúde implementou, desde 1976, um modelo padrão de Declaração de Óbito (DO) a ser utilizado em todo o país, como documento base do Sistema de Informação sobre Mortalidade (SIM) (Ministério da Saúde, 2009). A DO foi criada visando padronizar os registros coletados para pesquisas acerca de estatísticas vitais e epidemiológicas no Brasil, e cumprir o caráter jurídico da Certidão de Óbito para formalidades legais do sepultamento.

Dentro desta seção, serão detalhadas as características dos dados utilizados na presente pesquisa, abrangendo sua origem, período de abrangência, volume e a estrutura dos atributos relevantes para a análise.

3.4.1 Origem dos Dados

Os dados de mortalidade utilizados nesta pesquisa foram obtidos do Sistema de Informação sobre Mortalidade (SIM), uma iniciativa do Ministério da Saúde desenvolvida em 1975. O SIM fornece informações fundamentais para compreender os aspectos da mortalidade no Brasil e as causas de adoecimento que levam ao óbito, sendo um dos principais instrumentos de apoio para a elaboração de políticas públicas de saúde e seguridade social mais efetivas visando à prevenção, promoção e cuidado em saúde. (Ministério da Saúde, 2025).

Especificamente, as bases de dados sobre mortalidade foram acessadas e baixadas diretamente do site Open DataSUS, plataforma que disponibiliza os registros originários do sistema SIM. Esses registros possuem diversas informações, incluindo dados socioeconômicos, local de residência e ocorrência do óbito, categorização de óbitos fetais e não fetais, além da codificação das causas de óbito, segundo a Classificação Internacional de Doenças (CID).

Adicionalmente, para complementar as informações e facilitar a extração de variáveis específicas como naturalidade, idade e ocupação, que estão codificadas nos dados de mortalidade, utilizamos o software Tabwin. O Tabwin, ferramenta desenvolvida pelo DATASUS, permitiu a tabulação e o download dessas variáveis de forma pré-processada, agilizando a etapa de preparação dos dados para análise.

3.4.2 Período e Volume dos Dados

Para esta pesquisa, foram considerados os dados de mortalidade referentes ao período de 10 anos, abrangendo os registros entre 2012 e 2022. A volumetria total baixada da plataforma do Ministério da Saúde (Open DataSUS) corresponde a 15,1 milhões de registros de óbitos.

3.4.3 Estrutura dos Dados e Atributos

A Tabela 2 apresenta um dicionário de dados detalhado para os dados brutos desta pesquisa, obtidos em formato .csv. Esses dados são compostos por diversas variáveis que descrevem o óbito e o indivíduo. Para cada atributo, o dicionário fornece seu nome, descrição, a ação que tomamos no escopo da nossa pesquisa, o tipo e a porcentagem de dados faltantes. Essa análise foi conduzida no ambiente Jupyter, que será detalhada na Subseção 3.5.3, visando uma compreensão abrangente da estrutura dos dados.

Coluna	**Descrição**	**Ação**	**Tipo**	**% Dados Faltantes**
Contador	ID ordenado dos registros na tabela	Descartar	-	-
Origem	Banco de dados de Origem da Informação	Descartar	-	-
TIPOBITO	Óbito Fetal - Morte antes do nascimento	Descartar	-	-
DtObito	Data em que ocorreu o óbito	Descartar	-	-
HORAOBITO	Hora do óbito	Descartar	Str	1%
Natural0	País e Unidade de Federação onde o falecido nasceu	Manter	Int	17%
DTNASC	Data de Nascimento	Descartar	-	-
Idade	Idade do Falecido codificada	Manter	Int	-
Sexo	Sexo do Falecido	Manter	Int	-
RacaCor	Cor informada pelo responsável pelas informações	Manter	Int	2%
EstCiv	Estado Civil do falecido	Manter	Int	2%
Esc	Escolaridade em anos	Manter	Int	5%
OCUP	Tipo de trabalho predominante na vida do falecido	Manter	Int	11%
CodMunRes	Código do município de residência	Manter	Int	-
LOCOCOR	Local de ocorrência do óbito	Manter	Int	-
CODESTAB	Código do Estabelecimento	Descartar	-	37%
CODMUNOCOR	Código do município onde ocorreu o óbito	Manter	Int	-

Coluna	**Descrição**	**Ação**	**Tipo**	**% Dados Faltantes**
IdadeMae	Idade da mãe	Descartar	-	100%
ESCMAE	Escolaridade da mãe	Descartar	-	100%
SERIESCMAE	Última série escolar concluída pela mãe	Descartar	-	100%
OCUPMAE	Trabalho da mãe	Descartar	-	100%
QTDFILVIVO	Número de filhos vivos	Descartar	-	100%
QTDFILMORT	Número de filhos mortos	Descartar	-	100%
GRAVIDEZ	Óbito na gravidez	Descartar	-	100%
SEMAGESTAC	Semanas de gestação	Descartar	-	100%
GESTACAO	Faixas de semanas de gestação	Descartar	-	100%
PARTO	Tipo de parto	Descartar	-	100%
OBITOPARTO	Óbito no parto	Descartar	-	100%
PESO	Peso ao nascer	Descartar	-	100%
TPMORTEOCO	Situação gestacional/pósgestacional do óbito	Descartar	-	99%
OBITOGRAV	Óbito na gravidez	Descartar	-	99%
OBITOPUERP	Óbito no puerpério	Descartar	-	99%
ASSISTMED	Atendimento médico durante a enfermidade	Manter	Int	28%
EXAME	Realização de exame	Descartar	-	99%
CIRURGIA	Realização de cirurgia	Descartar	-	99%
NECROPSIA	Execução ou não de necropsia	Manter	Int	28%

Coluna	Descrição	Ação	Tipo	% Dados Faltantes
LINHAA	Causa terminal - doença ou estado mórbido que causou diretamente a morte	Manter	Str	4%
LINHAB	Causa antecedente ou consequencial - estado mórbido, se existir, que produziu a causa direta da morte registrada na linha A	Manter	Str	17%
LINHAC	Causa antecedente ou consequencial - estado mórbido, se existir, que produziu a causa direta da morte registrada na linha A	Manter	Str	50%
LINHAD	Causa básica – estado mórbido, se existir, que produziu a causa direta da morte registrada na linha A	Manter	Str	77%
LINHAII	Causa contribuinte	Manter	Str	45%
CAUSABAS	Causa básica	Manter	Str	-
COMUNSVOIM	Código do município do SVO ou IML	Descartar	-	92%
DTATESTADO	Data do atestado	Descartar	-	-
CIRCOBITO	Tipo de morte violenta	Descartar	-	99%
ACIDTRAB	Óbito relacionado ao trabalho	Descartar	-	99%
FONTE	Fonte da informação	Descartar	-	-
TPPOS	Óbito investigado	Descartar	-	34%
DTINVESTIG	Data da investigação	Descartar	-	87%
CAUSABAS_O	Causa básica informada antes da resseleção	Manter	Str	-
DTCADASTRO	Data do cadastro	Descartar	-	-

Coluna	Descrição	Ação	Tipo	% Dados Faltantes
ATESTANTE	Médico atendeu o paciente?	Manter	Int	10%
FONTEINV	Fonte da investigação	Descartar	-	86%
DTRECEBIM	Data de recebimento	Descartar	-	-
CAUSAMAT	CID causa materna	Descartar	-	100%
ESC2010	Escolaridade última série concluída (falecido)	Manter	Int	7%
ESCMAE2010	Escolaridade última série concluída (mãe)	Descartar	-	100%
DIFDATA	Diferença entre datas	Descartar	-	-
STDOEPIDEM	Status DO Epidemiológica	Descartar	Int	-
STDONOVA	Status DO Nova	Descartar	Int	-
DTCADINV	Data do cadastro de investigação	Manter	Str	-
TPOBITOCOR	Momento da ocorrência do óbito	Descartar	Int	-
DTCONINV	Data da conclusão da investigação	Manter	Str	-
DTCADINF	Indica se houve investigação	Manter	Str	-
MORTEPARTO	Falta descrição na fonte	Descartar	-	-
DTCONCASO	Data da conclusão do caso	Descartar	Str	-
NUDIASOBIN	Dias entre óbito e conclusão da investigação	Descartar	-	99%
ANO	Ano que a morte ocorreu	Manter	Str	-

Tabela 2 – Resumo das colunas do dataset e ações adotadas durante o pré-processamento

Fonte: Elaborado pelos autores, 2025.

3.5 Processo de Análise dos Dados

Esta pesquisa foi conduzida no estilo iterativo de análise exploratória de dados (KDD), abordagem que utiliza técnicas para compreender, resumir e extrair informações relevantes de um conjunto de dados (Mukhiya; Ahmed, 2020). Com foco em uma abordagem computacional, foi utilizado um algoritmo de aprendizado não super-

visionado, o Mapa Auto-Organizável (SOM), com o objetivo de identificar padrões e correlações entre o CID (Classificação Internacional de Doenças) de Alzheimer e as demais colunas de CID relacionadas ao óbito por essa doença.

A seguir, são apresentadas as etapas realizadas durante o processo de análise dos dados.

3.5.1 Extração, Transformação e Carregamento - ETL

A etapa do ETL contou com a automatização dos processos, utilizando um script desenvolvido em Shell Script e Linguagem C, responsável por realizar o download dos arquivos de dados referentes à mortalidade no Brasil.

Esse script também realiza a conversão da codificação para o padrão UTF-8, garantindo a leitura correta de caracteres especiais e acentuação. Além disso, realiza a transformação dos arquivos .csv em .sql, permitindo a carga automatizada para o banco de dados. Posteriormente, foi feita a comparação de volumetria entre os dados no banco e os arquivos originais, visando identificar perdas ou falhas durante a importação.

A estrutura desse script foi adaptada para realizar também a conversão de charset e o carregamento de arquivos auxiliares no banco de dados, como os dicionários de idade, nacionalidade e CID. Essas informações, em conformidade com a Lei Geral de Proteção de Dados (LGPD) foram disponíveis no dados de mortalidade de forma codificada. Com essa adaptação, foram criadas tabelas auxiliares no banco, que ofereceram suporte ao processo de análise e filtragem dos dados.

Optou-se pelo uso de Shell Script e C para esse processo devido a eficiência computacional e a agilidade na execução dessas tarefas, principalmente no que diz respeito à manipulação de grandes volumes de dados. Essa linguagens apresentaram menor tempo de execução dessas tarefas quando comparadas a outras, como Java e Python, tornando o processo mais eficiente.

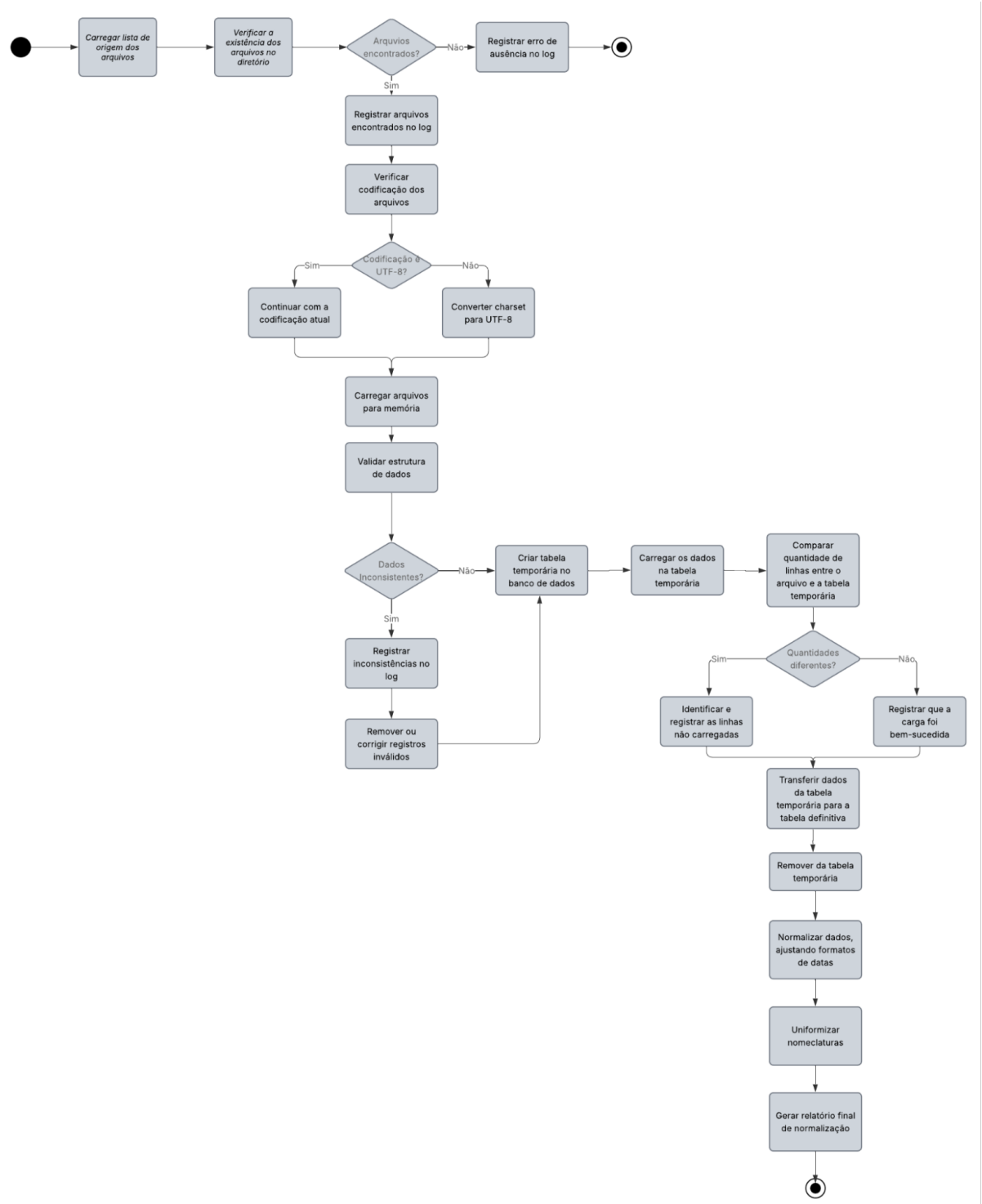

Figura 3 – Diagrama de atividades do processo ETL.

Fonte: Elaborado pelos autores, 2025.

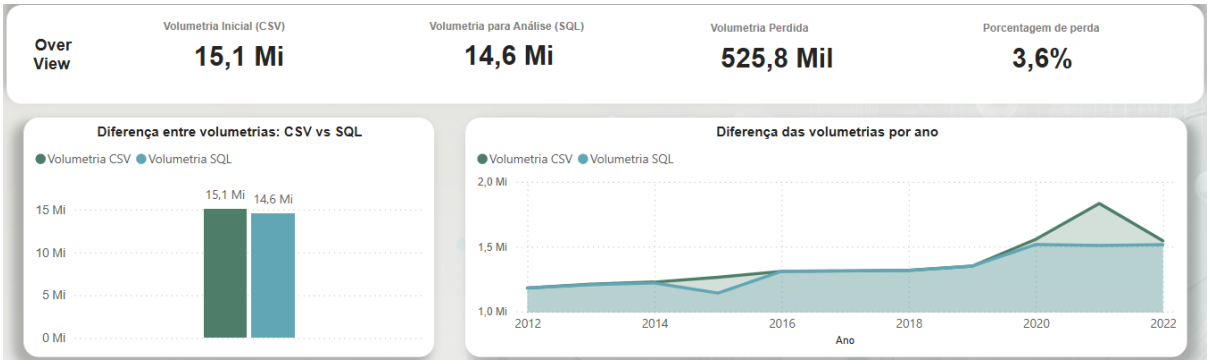

Figura 4 – Análise da volumetria de dados após o processo de ETL.

Fonte: Elaborado pelos autores, 2025.

A Figura 4 detalha o impacto do processo de Extração, Transformação e Carga (ETL) na volumetria total dos dados de mortalidade. Conforme ilustrado, a base de dados inicial, obtida em formato CSV, apresentava uma volumetria de 15,1 milhões de registros. Após as etapas do ETL, a volumetria final carregada no ambiente SQL para análise foi de 14,6 milhões de registros. Essa diferença representa uma perda controlada de 525,8 mil registros, equivalente a apenas 3,6% do total inicial.

3.5.2 Indexação e Filtragem por CID

Com a conclusão da carga automatizada dos dados no banco de dados MySQL, iniciou-se o processo de indexação. Considerando o grande volume de dados, superior a 1 milhão de registros a cada ano, e a finalidade exclusiva para consultas, a criação de índices foi aplicada com o objetivo de aumentar a eficiência na filtragem dos dados. A indexação fornece caminhos de acesso secundário, permitindo a recuperação eficiente dos registros por meio de campos específicos, sem alterar sua disposição física no arquivo. Esses índices oferecem rotas alternativas de busca, aumentando significativamente a velocidade das consultas ao banco de dados (Elmasri; Navathe, 2005).

O público-alvo da análise são idosos que faleceram em decorrência da Doença de Alzheimer, evidenciando a necessidade de filtragem da coluna “causabas”, correspondente a causa básica da mortalidade. Portanto, para a filtragem, foram considerados os CIDs F00.0, F00.1, F00.2, F00.9, G30.0, G30.1, G30.8 e G30.9.

CID	Descrição
F00.0	Demência na doença de Alzheimer de início precoce
F00.1	Demência na doença de Alzheimer de início tardio
F00.2	Demência na doença de Alzheimer, tipo atípico ou misto
F00.9	Demência na doença de Alzheimer, não especificada
G30.0	Doença de Alzheimer de início precoce
G30.1	Doença de Alzheimer de início tardio
G30.8	Outras formas de doença de Alzheimer
G30.9	Doença de Alzheimer, não especificada

Tabela 3 – Classificação dos CIDs relacionados à doença de Alzheimer segundo a CID-10.

Fonte: Elaborado pelos autores, 2025.

Antes de aplicar a indexação, foi utilizado o comando `EXPLAIN ANALYZE`, recurso disponível no MySQL, que executa uma consulta e detalha o custo computacional, visando validar a necessidade de indexação e comparar os impactos no desempenho, com e sem o uso de índices.

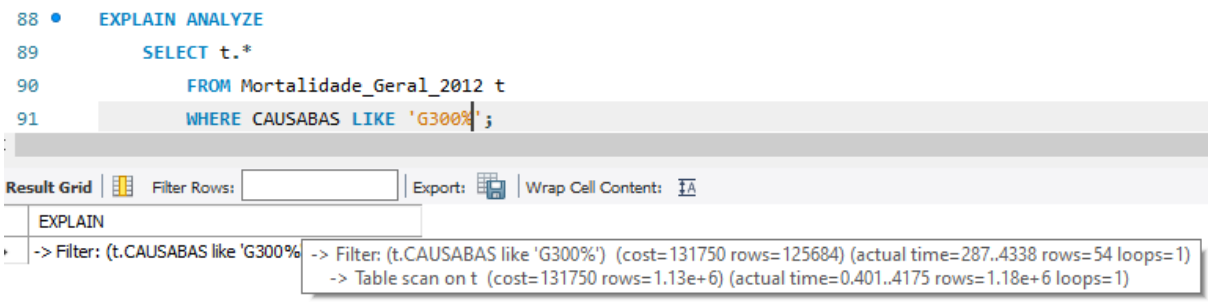

Figura 5 – Plano de execução da consulta sem índice aplicado na coluna CAUSABAS.

Fonte: Elaborado pelos autores, 2025.

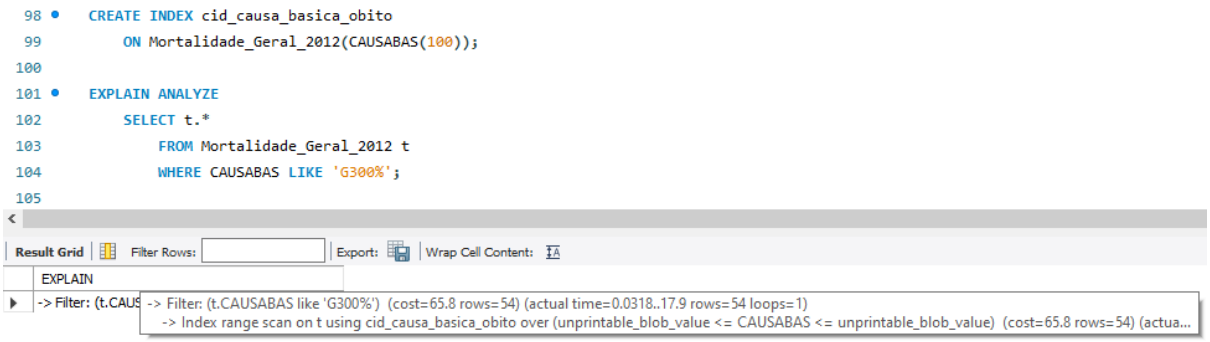

Figura 6 – Plano de execução da consulta com índice aplicado na coluna CAUSABAS.

Fonte: Elaborado pelos autores, 2025.

É possível observar que na Figura 5, o plano de execução de uma consulta sem índice resulta em varredura completa da tabela, ocasionando no alto custo computacional e maior tempo de resposta. Enquanto que na Figura 6, com o índice criado, a busca é otimizada, reduzindo significativamente o custo e o tempo de execução da query.

Para ilustrar essa diferença de desempenho entre as consultas, a Figura 7 apresenta a complexidade computacional em ambos os cenários, ou seja, a quantidade de recursos, como tempo e memória, necessários para que o algoritmo execute essas buscas. No cenário sem índice, a complexidade é O(n), o que significa que o algoritmo precisa percorrer todas as linhas da tabela para retornar os dados filtrados, resultando em um crescimento linear no tempo de execução. Enquanto que com o uso de índice, a complexidade é O(Log n), uma vez que o algoritmo percorre uma estrutura reduzida, os índices, para localizar os registros desejados, demonstrando maior eficiência da consulta à medida que o volume de dados aumenta.

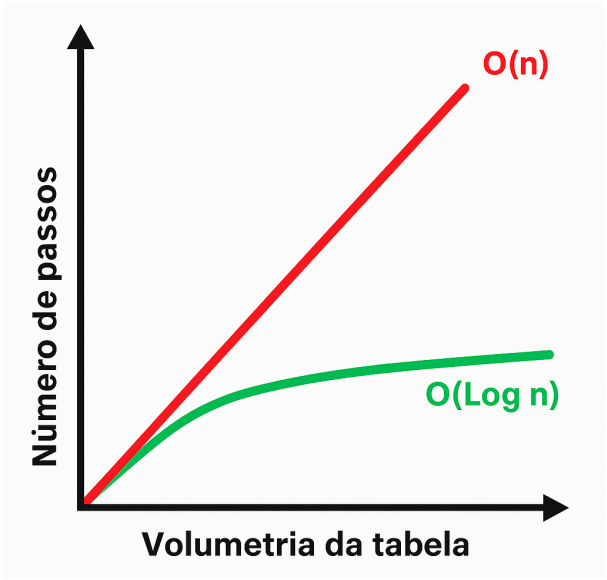

Figura 7 – Comparação da complexidade computacional entre buscas sem e com índice.

Fonte: Elaborado pelos autores, 2025.

Após a criação dos índices nas colunas que seriam aplicado filtros, foi desenvolvido um Procedimento Armazenado Dinâmico (ou *Stored Procedure*), que consiste em um conjunto de instruções SQL estruturadas de forma semelhante a uma função. As procedures tem como finalidade automatizar tarefas repetitivas e executar comandos SQL de forma dinâmica, em tempo de execução, com base no parâmetro informado, no caso foram as tabelas correspondente a cada ano.

```
DELIMITER $$
CREATE PROCEDURE sim_data.SPR_Enriquece_Analise(IN tabela VARCHAR(100))
BEGIN
    DECLARE CIDs VARCHAR(100);
    DECLARE TabelaFiltrada VARCHAR(100);
    DECLARE nomeTabelaFinal VARCHAR(100);

    SET CIDs := '''F000'', ''F001'', ''F002'', ''F009'', ''G300'', ''G301'', ''G308'', ''G309'''; -- CIDs de Alzheimer
    SET TabelaFiltrada := CONCAT(tabela, '_filtrada');
    SET nomeTabelaFinal := CONCAT(tabela, '_analise');
```

Figura 8 – Trecho da Procedure `SPR_Enriquece_Analise`

Fonte: Elaborado pelos autores, 2025.

A *procedure* foi implementada com o objetivo de filtrar o público-alvo da pesquisa, armazenando os registros em uma nova tabela. Em seguida, realiza-se o enriquecimento desses dados filtrados, decodificando campos como idade. Essa abordagem permite reutilizar a lógica de filtragem e enriquecimento dos dados para qualquer tabela informada, facilitando o processamento e garantindo consistência nas operações realizadas em diferentes anos.

Todas as tabelas anuais importadas de forma automatizadas, referentes ao período de 2012 a 2022, passaram por essa *procedure*. Por fim, os dados filtrados de cada ano foram unificados com o comando `UNION ALL`, resultando em uma única tabela com os dados de todos os anos, a qual foi posteriormente exportada para uso nas etapas seguintes da análise.

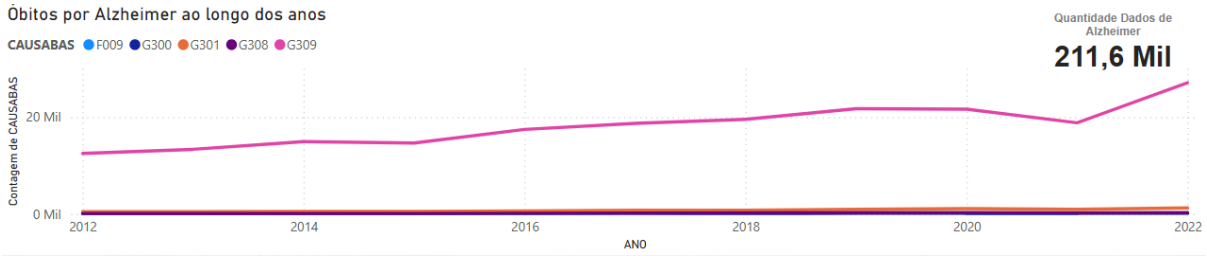

Figura 9 – Óbitos por Alzheimer entre 2012 e 2022 no Brasil.

Fonte: Elaborado pelos autores, 2025.

O gráfico apresentado na Figura 9, construído a partir dos dados da tabela final unificada, mostra a evolução da mortalidade por Alzheimer ao longo dos anos (2012-2022). É notório o crescimento contínuo no número de óbitos por essa doença, totalizando uma volumetria de 211,6 mil registros que serão utilizados para essa análise. Uma observação relevante é a predominância do código CID G30.9, Doença de Alzheimer não especificada. Com a frequência superior dos demais, esse registro indica uma limitação da disponibilidade de informações que permitam especificar a forma da doença nas DOs, resultando em uma classificação genérica para a maior parte das ocorrências de óbito por Alzheimer.

3.5.3 Tratamento e Limpeza de dados no Jupyter

Antes da aplicação de qualquer algoritmo de aprendizado de máquina, é importante garantir que os dados estejam limpos e tratados. Essa é uma etapa essencial em qualquer processo de análise, uma vez que remove inconsistências, trata valores ausentes e transforma dados categóricos em formato adequado para o treinamento do modelo.

Tendo em vista a importância dessa etapa, o processo de tratamento e limpeza foi iniciado no ambiente do *Jupyter Notebook*. Para isso, foram importadas as principais bibliotecas Python para a manipulação e visualização de dados, como `pandas`, `numpy`, `matplotlib` e `seaborn`, que auxiliaram na preparação dos dados. As bibliotecas específicas para a aplicação do algoritmo de aprendizado de máquina foram importadas em etapas posteriores do desenvolvimento. Em seguida, foi realizado a leitura do arquivo "dados_filtrados.csv", previamente exportado do banco de dados, sendo carregado no ambiente como um *DataFrame*, estrutura semelhante a uma tabela.

```
[2]: # Importando os pacotes

import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

```
[3]: # Importando o dataFrame - todos como str

df = pd.read_csv('dados_filtrados.csv', sep=';', dtype=str)
```

```
[4]: df
```

[4]:

	CONTADOR	ORIGEM	TIPOBITO	DTOBITO	HORAOBITO	NATURAL0	DTNASC	IDADE	SEXO	RACACOR	...	STDOEPIDEM	STDONOVA	DTCADINV	TPOBITO
0	42081	1	2	28042012	1800	NaN	18071914	497	2	4	...	0	1	NaN	
1	52106	1	2	12072012	1440	NaN	07101944	467	1	1	...	0	1	NaN	
2	25135	1	2	30072012	1930	NaN	03091938	473	1	1	...	0	1	NaN	
3	10487	1	2	29012012	1745	NaN	23061928	483	1	1	...	0	1	NaN	
4	10560	1	2	27042012	1300	NaN	13041920	492	2	1	...	0	1	NaN	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
211588	1514999	1	2	22122022	0857	843	15041939	483	2	1	...	0	1	NaN	
211589	1515062	1	2	23082022	0920	831	01031953	469	2	4	...	0	1	NaN	
211590	1515229	1	2	05072022	2000	833	24071938	483	2	1	...	0	1	NaN	
211591	1515340	1	2	06072022	2358	190	26101930	491	2	1	...	0	1	NaN	
211592	1515396	1	2	06072022	0655	831	26111929	492	2	4	...	0	1	NaN	

211593 rows × 68 columns

Figura 10 – DataFrame inicial da base de dados.

Fonte: Elaborado pelos autores, 2025.

No canto inferior esquerdo da Figura 10, é possível identificar que o *DataFrame* é composto por 211.593 linhas e 68 colunas. Como a quantidade de colunas é elevada, não foi possível visualizar todas de uma vez, sendo necessário executar uma função para obter uma descrição geral dos dados.

```
[5]: df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 211593 entries, 0 to 211592
Data columns (total 68 columns):
 #   Column      Non-Null Count   Dtype
---  ------      --------------   -----
 0   CONTADOR    211593 non-null  object
 1   ORIGEM      211593 non-null  object
 2   TIPOBITO    211593 non-null  object
 3   DTOBITO     211593 non-null  object
 4   HORAOBITO   208415 non-null  object
 5   NATURAL0    174587 non-null  object
 6   DTNASC      211484 non-null  object
 7   IDADE       211593 non-null  object
 8   SEXO        211593 non-null  object
 9   RACACOR     205564 non-null  object
 10  ESTCIV      206211 non-null  object
 11  ESC         200701 non-null  object
 12  OCUP        187751 non-null  object
 13  CODMUNRES   211593 non-null  object
 14  LOCOCOR     211593 non-null  object
 15  CODESTAB    132717 non-null  object
 16  CODMUNOCOR  211593 non-null  object
 17  IDADEMAE    0 non-null       object
 18  ESCMAE      0 non-null       object
 19  SERIESCMAE  0 non-null       object
 20  OCUPMAE     0 non-null       object
 21  QTDFILVIVO  0 non-null       object
 22  QTDFILMORT  0 non-null       object
 23  GRAVIDEZ    0 non-null       object
 24  SEMAGESTAC  0 non-null       object
 25  GESTACAO    0 non-null       object
```

Figura 11 – Parte da descrição do *DataFrame*

Fonte: Elaborado pelos autores, 2025.

A função `.info()` retorna uma lista de todas as colunas presentes no *DataFrame*, juntamente com a contagem de dados não nulos e dos respectivos tipos de dados. A princípio, é possível identificar a presença de colunas com dados faltantes, entretanto não é apropriado descartá-las sem antes realizar uma breve análise de seu conteúdo e, por esse motivo, foi calculada a porcentagem de valores ausentes em cada coluna. Dependendo da proporção de ausência e da relevância da coluna para a análise, a abordagem de imputação será adotada, isto é, preencher valores faltantes com valores plausíveis.

```
[6]: # Calculando o percentual dos dados que estão em branco

     percentual_nulos = (df.isnull().sum() / len(df)) * 100
     percentual_nulos[percentual_nulos > 0].sort_values(ascending=False)
```

```
[6]: PESO          100.000000
     OCUPMAE       100.000000
     CAUSAMAT      100.000000
     OBITOPARTO    100.000000
     PARTO         100.000000
     GESTACAO      100.000000
     SEMAGESTAC    100.000000
     GRAVIDEZ      100.000000
     QTDFILMORT    100.000000
     QTDFILVIVO    100.000000
     SERIESCMAE    100.000000
     ESCMAE        100.000000
     IDADEMAE      100.000000
     ACIDTRAB      100.000000
     DTCADINF      100.000000
     MORTEPARTO    100.000000
     DTCONCASO     100.000000
     ESCMAE2010    100.000000
     DTCADINV       99.999055
     TPOBITOCOR     99.999055
     DTCONINV       99.999055
     NUDIASOBIN     99.998582
     FONTE          99.983931
     CIRCOBITO      99.982986
     TPMORTEOCO     99.779766
     OBITOPUERP     99.776930
     OBITOGRAV      99.776458
     CIRURGIA       99.180030
     EXAME          99.167269
     COMUNSVOIM     93.338627
```

Figura 12 – Trecho do cálculo de percentual dos dados faltantes em *Python*.

Fonte: Elaborado pelos autores, 2025.

A Figura 12 apresenta o cálculo da porcentagem de valores ausentes em cada coluna do *DataFrame* em formato de lista. Devido à quantidade de colunas com dados ausentes, o retorno foi extenso. Para facilitar a visualização e a interpretação dos resultados, foi gerado um gráfico que mostra o percentual de dados faltantes por coluna.

```
[7]: # Cálculo da quantidade e percentual de colunas com dados faltantes no DataFrame

qtd_colunas_valores_faltantes = percentual_nulos[percentual_nulos > 0].count()
print(f'Total de colunas com dados faltantes: {qtd_colunas_valores_faltantes}')
print(f'Percentual das colunas com dados faltantes: {(qtd_colunas_valores_faltantes / 68) * 100:.2f}%')
```

```
Total de colunas com dados faltantes: 57
Percentual das colunas com dados faltantes: 83.82%
```

```
[8]: # Plot dos valores nulos

plt.figure(figsize=(15, 4))
sns.barplot(x=percentual_nulos.index, y=percentual_nulos.values, color='lightgreen')
plt.title('Percentual de Dados Faltantes')
plt.xlabel('Colunas')
plt.ylabel('Percentual de Dados Faltantes')
plt.xticks(rotation=90) # deixa os labels em pé, melhor p ver

plt.show()
```

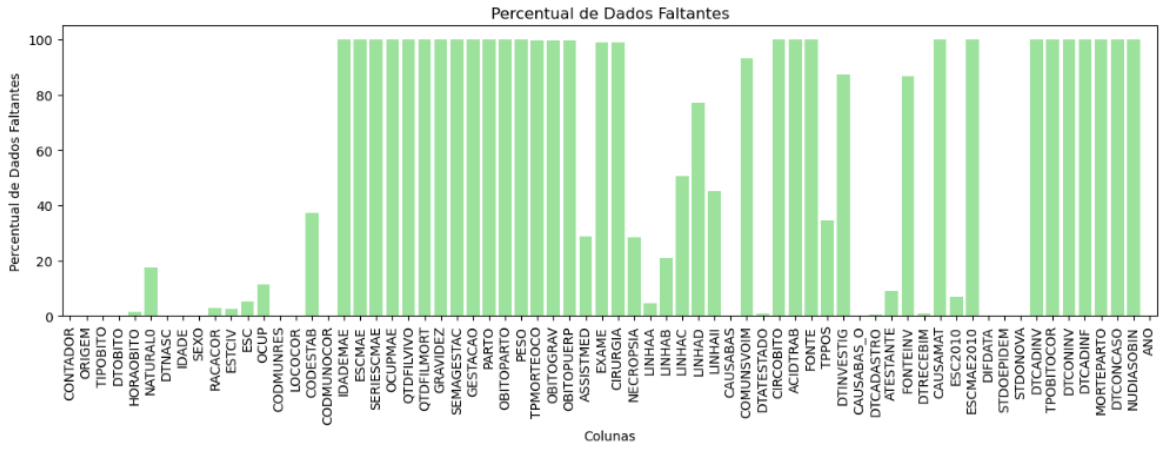

Figura 13 – Gráfico do percentual de valores ausentes em cada coluna.

Fonte: Elaborado pelos autores, 2025.

O gráfico representado na Figura 13 mostra a presença de colunas com 100% de valores ausentes, que serão descartadas para garantir a integridade dos dados. Em contrapartida, colunas com baixo percentual de dados faltantes e relevantes para a análise serão mantidas e passarão por um processo de imputação de valores.

Visando identificar as colunas relevantes para a análise, criamos um dicionário de dados, que foi apresentado na Subseção 3.4.3 na Tabela 2. Este dicionário, apresentado anteriormente na visão geral dos dados, inclui o nome da coluna, o tipo de dado, a porcentagem de valores ausentes e a ação tomada (descartar a coluna ou realizar imputação). Ele foi desenvolvido após a análise do percentual de dados faltantes por coluna.

Com as colunas a serem utilizadas já identificadas, iniciou-se a etapa de processamento dos dados, que os prepara para o treinamento do algoritmo. Primeiramente, as colunas classificadas como do tipo objeto foram convertidas para valores inteiros, conforme o dicionário estruturado anteriormente. Em seguida, os dados considerados irrelevantes para a análise foram descartados.

Figura 14 – Alteração dos tipos de dados e exclusão de colunas.

Fonte: Elaborado pelos autores, 2025.

Após a limpeza do *DataFrame*, as colunas relevantes com dados faltantes foram identificadas para o processo de imputação, considerando alguns critérios específicos.

Figura 15 – Percentual de valores ausentes por coluna nos dados que serão utilizados na análise.

Fonte: Elaborado pelos autores, 2025.

As colunas `assistmed`, `necropsia`, `esc2010`, `esc` e `estciv` possuem, de acordo com a estrutura do SIM, o valor 9 para indicar que a informação não foi declarada. Para as demais colunas com dados ausentes, foi inserido o valor 0, representando “não se aplica”. Embora a imputação ideal seja realizada por meio da média, mediana ou moda,

dependendo do tipo de dado, utilizamos zero para não comprometer a integridade da análise nem o desempenho do algoritmo utilizado.

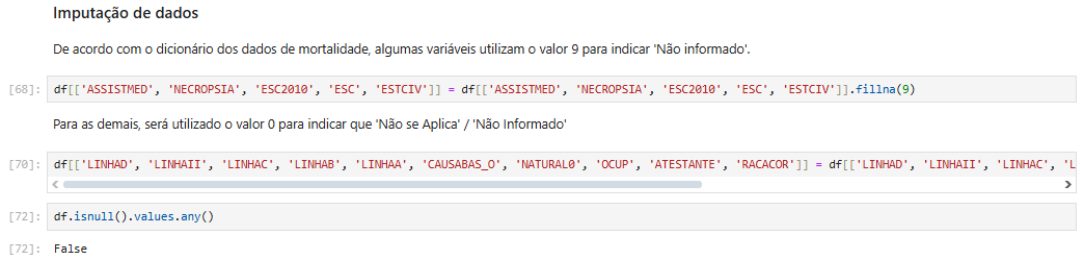

Figura 16 – Preenchimento dos dados ausentes com valores definidos.

Fonte: Elaborado pelos autores, 2025.

Na etapa seguinte, foi realizada a análise das colunas de CID. Durante o processo, foi identificado que algumas células continham mais de um código CID, o que contraria o padrão previsto para esses campos, já que cada campo deveria conter apenas um valor. Além disso, os algoritmos de aprendizado de máquina esperam que as variáveis estejam em formato numérico, preferencialmente do tipo inteiro. Dessa forma, as colunas de CID foram tratadas para atender ao padrão esperado.

Para o tratamento das colunas de CID que continham mais de um código na célula, foi utilizado a função `split()` do Python, considerando o caractere asterisco como delimitador, uma vez que antecede cada código CID. Esse tratamento foi aplicado às colunas `LINHAA`, `LINHAB`, `LINHAC`, `LINHAD`, `LINHAII`, criando novas colunas com sufixo númericos como, por exemplo, `LINHAA_1`. As células vazias geradas após a separação foram preenchidas com o valor 0 também, mantendo a consistência dos dados. Essa etapa resultou em uma estrutura mais organizada, com os códigos separados, facilitando a análise.

```
[31]: # Separar os CID's, criando novas colunas

# Lista das colunas de CID para realizar a separação
df_linhas = ['LINHAA', 'LINHAB', 'LINHAC', 'LINHAD', 'LINHAII']

# Vetor auxiliar
df_splited = []

# Percorre as linhas do df realizando o split, renomeando as colunas e preenchendo como 0 colunas sem informações
for linhas in df_linhas:
    try:
        # Separa os CID's em colunas
        df_split = df[linhas].str.split('*', expand=True)

        # realiza o split no primeiro *, fazendo com que a primeira coluna seja nula, sendo necessário dropar
        df_split = df_split.drop(0, axis=1)

        # renomeia as colunas criadas para linha_i com base na quantidade de colunas
        df_split.columns = [f'{linhas}_{i+1}' for i in range(df_split.shape[1])]

        # linhas que ficaram sem valor recebem 0
        df_split = df_split.replace('', 0).fillna(0)

        # adiciona esse novo "df" na lista aux para realizar o outro loop com a prox linha
        df_splited.append(df_split)

    except:
        pass
```

Figura 17 – Função utilizada para separar valores de CID.

Fonte: Elaborado pelos autores, 2025.

Em seguida, os dados foram convertidos para o tipo inteiro (int). Como os códigos CID são compostos por letras e números como, por exemplo, G300, foi necessário transformar as letras em valores numéricos. Embora existam técnicas de codificação de variáveis categóricas disponíveis na biblioteca `scikit-learn`, como o *One hot encoding* e *Label encoding*, implementamos uma abordagem mais direta para controlar e preservar a dimensionalidade dos dados, evitando a criação de colunas adicionais, o que poderia impactar no desempenho do algoritmo. Para isso, aplicamos a função `ord()` do Python, que converte cada letra em seu respectivo valor na tabela ASCII. Todas as colunas contendo códigos CID passaram por essa transformação.

```
[33]: # Transformando o CID's em colunas numéricas:

def substituir_letras(cid):
    # Verifica se é string
    if isinstance(cid, str):
        aux = ""

        for digito in cid:
            # verifica se é letra (a - z)
            if digito.isalpha():
                # ord -> converte caracteres em inteiros
                numero = ord(digito.upper()) - ord('A') + 1
                # concatena à string auxiliar criada
                aux += str(numero)

            # se for numero, continua
            else:
                aux += digito
        return aux
    return cid
```

Figura 18 – Função utilizada para codificação das letras presentes nos códigos CID.

Fonte: Elaborado pelos autores, 2025.

A função da Figura 18 percorre cada caractere da célula, convertendo letras para números conforme sua posição no alfabeto, enquanto os dígitos numéricos são mantidos. Assim, o código CID "G300"é transformado em "7300", permitindo a conversão da coluna para o tipo inteiro. Após essas transformações, as colunas contendo CIDs tratados foram concatenadas ao *DataFrame* original.

Com o pré-processamento concluído, a etapa seguinte consiste na verificação de *outliers*, que são valores que se distanciam da distribuição geral das variáveis. A presença desses valores pode afetar o desempenho de muitos algoritmos de aprendizado de máquina, especialmente aqueles sensíveis à variação de dados.

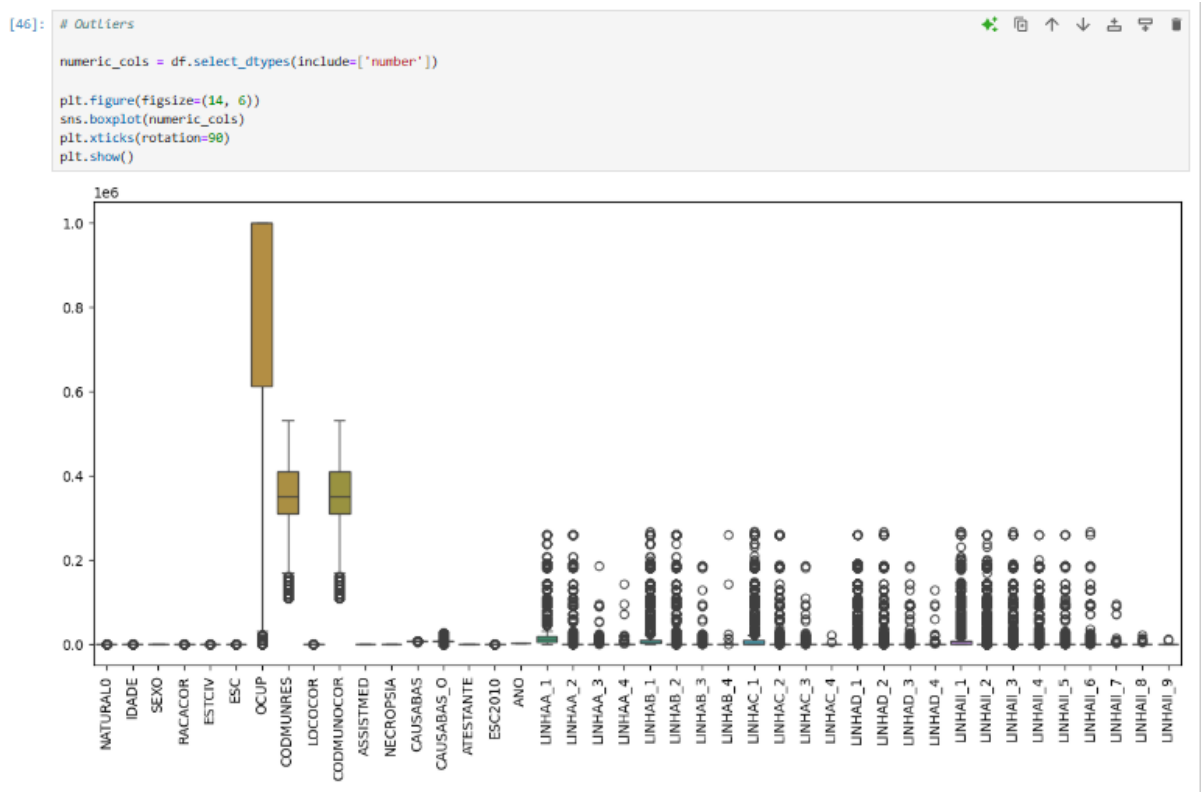

Figura 19 – Boxplot para identificação de *outliers* nas variáveis analisadas.

Fonte: Elaborado pelos autores, 2025.

Com o *boxplot* apresentado na Figura 19, é possível identificar uma presença significativa de *outliers*. Entretanto, considerando que os dados são registros de saúde, mais especificamente de mortalidade, é esperado que haja uma maior variabilidade entre os casos, justificando essa dispersão. Por esse motivo, os dados não foram removidos, priorizando a utilização de algoritmos mais robustos à presença de *outliers*.

Posteriormente, analisamos a correlação entre as variáveis do *DataFrame*, com o objetivo de selecionar as variáveis a serem utilizadas nas etapas de aprendizado de máquina, conforme apresentado no *heatmap* da Figura 20. É possível notar que a maioria das variáveis apresenta baixa correlação linear entre si, indicando a ausência de relações lineares.

```
[48]: target_columns = df.drop(columns=['LINHAA', 'LINHAB', 'LINHAC', 'LINHAD', 'LINHAII'])
numeric_columns = target_columns.select_dtypes(include=['number']).columns

# Compute and plot the heatmap for numeric columns
plt.figure(figsize=(12,8))
sns.heatmap(df[numeric_columns].corr(), cmap='coolwarm')
plt.title("Correlation Heatmap")
plt.show()
```

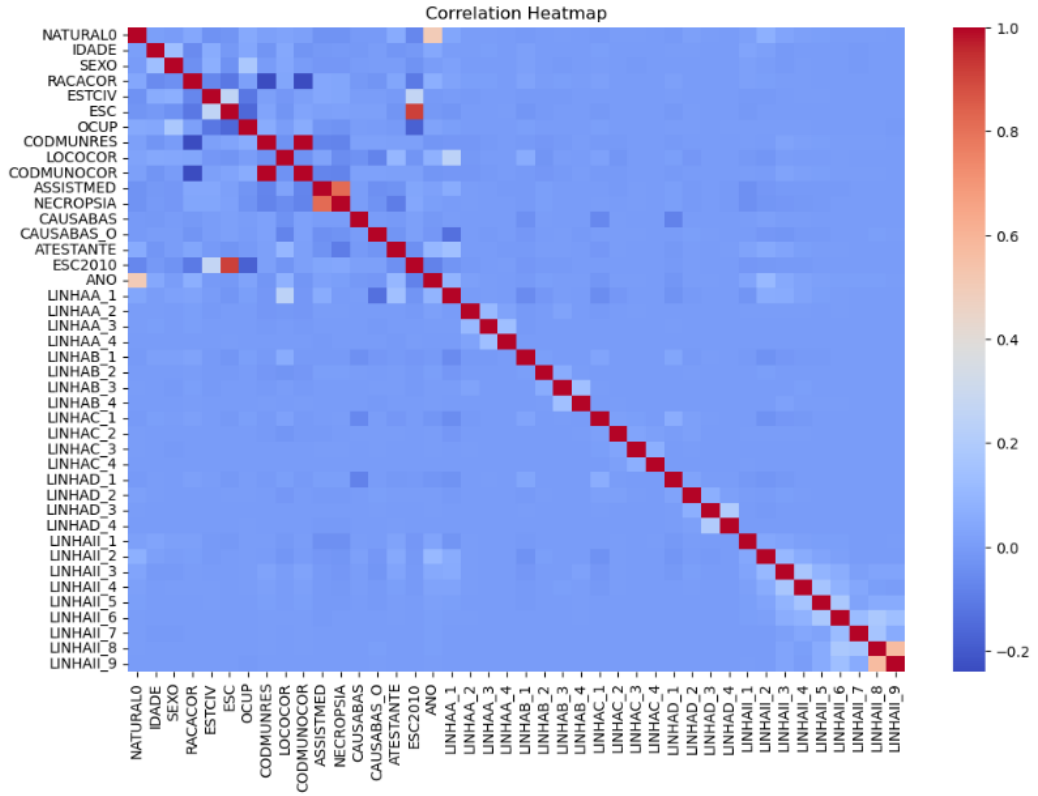

Figura 20 – Heatmap de Correlação entre as variáveis.

Fonte: Elaborado pelos autores, 2025.

Diante desse cenário, surge a necessidade de utilizar algoritmos capazes de capturar padrões não lineares nos dados. Um exemplo é o Mapa Auto-Organizável (SOM) que consegue identificar agrupamentos e estruturas complexas nos dados, mesmo na ausência de correlações lineares evidentes.

3.5.4 Aplicação do algoritmo SOM

O Mapa Auto-Organizável (SOM) é um algoritmo de rede neural capaz de transformar relações complexas e não-lineares em dados com alta dimensionalidade, ou seja, com muitas variáveis, em uma representação gráfica uni ou bidimensional, conhecida como mapa. As redes neurais são um modelo de aprendizado de máquina inspirados no cérebro humano, capazes de obter conhecimento a partir de experiências e armazená-los por meio de conexões ajustáveis (pesos sinápticos) entre unidades chamadas neurônios (Haykin, 2001).

No SOM, os neurônios estão em nós de uma grade que é normalmente unidi-

mensional (linha) ou bidimensional (matriz). Cada neurônio possui um vetor de pesos, com um valor correspondente a cada variável dos dados de entrada. Durante o treinamento, o algoritmo recebe os dados de entrada, que são as linhas do *dataframe* com todas as variáveis (colunas). O algoritmo compara esse vetor de entrada com todos os neurônios da grade, utilizando uma medida de distância para identificar qual neurônio tem os pesos mais parecidos com o vetor apresentado. Esse neurônio é chamado de *BMU* (*Best matching unit*). Após encontrar o *BMU*, o SOM ajusta os valores (pesos) desse neurônio e de seus vizinhos na grade para se aproximarem ainda mais do vetor de entrada. Esse processo vai se repetindo com outros vetores de entrada, e forma a grade para representar a estrutura dos dados, organizando os neurônios de forma que padrões semelhantes fiquem próximos entre si no mapa.

De acordo com Haykin (2001), o principal objetivo do mapa auto-organizável (SOM) é transformar um padrão implícito nos dados, mesmo que de alta dimensão, em um mapa discreto bidimensional, preservando a organização topológica das informações durante essa transformação, ou seja, os elementos semelhantes nos dados de entrada permanecem próximos no mapa resultante, facilitando a identificação de padrões e agrupamentos. Essa característica torna o SOM adequado para a análise exploratória de dados complexos, com alta dimensionalidade e com baixas correlações lineares, tais como os dados analisados nesta pesquisa. Além disso, o SOM realiza treinamentos não supervisionados, ou seja, não exige uma variável-alvo pré-definida, permitindo uma exploração ampla visando encontrar possíveis padrões nos dados.

Iniciamos essa etapa realizando a importação da biblioteca `MiniSom`, ferramenta Python baseada no NumPy que possibilita a implementação do algoritmo SOM e, em seguida, criamos um novo *dataframe* selecionando apenas as variáveis que seriam utilizadas no treinamento. A princípio, optamos por utilizar apenas as colunas de CID relacionadas ao óbito: `LINHAA`, `LINHAB`, `LINHAC`, `LINHAD`, `LINHAII`, `CAUSABAS` e `CAUSABAS_O`. O objetivo dessa seleção foi identificar possíveis padrões entre as causas básicas e secundárias da mortalidade e as doenças associadas, com base no resultado gerado pelo SOM.

4. Aplicando o MAPASOM

Os Mapas Auto-Organizáveis (Self-Organizing Maps – SOM) são um tipo de Rede Neural Artificial capaz de converter dados complexos e com correlações não lineares em uma representação de baixa dimensionalidade. Essa técnica permite visualizar e identificar padrões nos dados, mesmo quando relações lineares não são evidentes, como observado na matriz de correlação.

52]:
```
# Importando a biblioteca MiniSom
from minisom import MiniSom
```

Selecionando as variáveis que iremos utilizar para aplicar no MAPASOM

54]:
```
df_training = df[
    [
        'LINHAA_1', 'LINHAA_2', 'LINHAA_3', 'LINHAA_4', 'LINHAB_1', 'LINHAB_2', 'LINHAB_3', 'LINHAB_4',
        'LINHAC_1', 'LINHAC_2', 'LINHAC_3', 'LINHAC_4', 'LINHAD_1', 'LINHAD_2', 'LINHAD_3', 'LINHAD_4',
        'LINHAII_1', 'LINHAII_2', 'LINHAII_3', 'LINHAII_4', 'LINHAII_5', 'LINHAII_6', 'LINHAII_7', 'LINHAII_8', 'LINHAII_9',
        'CAUSABAS', 'CAUSABAS_O'
    ]
]
```

55]:
```
df_training.head()
```

55]:

	LINHAA_1	LINHAA_2	LINHAA_3	LINHAA_4	LINHAB_1	LINHAB_2	LINHAB_3	LINHAB_4	LINHAC_1	LINHAC_2	...	LINHAII_2	LINHAII_3	LINHAII_4	LINHAII_5	LI
0	7300	0	0	0	185424	0	0	0	0	0	...	0	0	0	0	
1	7300	0	0	0	0	0	0	0	0	0	...	0	0	0	0	
2	10969	0	0	0	54624	0	0	0	7300	0	...	0	0	0	0	
3	7300	0	0	0	0	0	0	0	0	0	...	0	0	0	0	
4	7300	0	0	0	0	0	0	0	0	0	...	0	0	0	0	

5 rows × 27 columns

56]:
```
df_training.shape
```

56]:
```
(211593, 27)
```

Figura 21 – Início do treinamento do SOM

Fonte: Elaborado pelos autores, 2025.

Em seguida, importamos a biblioteca `MinMaxScaler`, do `scikit-learn`, para realizar a normalização dos dados, etapa importante do pré-processamento. Essa função transforma os valores das variáveis para um intervalo padronizado, entre 0 e 1 ou -1 e 1, quando há valores negativos. Aplicar a normalização é essencial antes do treinamento de dados em algoritmos de aprendizado de máquina, uma vez que os dados originais podem apresentar escalas diferentes entre si. Por exemplo, o algoritmo poderia interpretar que o CID `7300` tem menor importância do que o CID `185424`, apenas por conta da magnitude numérica, e não pelo significado clínico. Com a normalização, garantimos que todas as variáveis tenham o mesmo peso no treinamento do SOM.

Normalização dos dados com MinMaxScaler

As features ficam em um intervalo de [0,1], diminuindo a dimensionalidade dos dados

[59]:
```
from sklearn.preprocessing import MinMaxScaler
```

[60]:
```
df_normalized_min = MinMaxScaler().fit_transform(df_training)
```

[61]:
```
df_normalized_min.shape
```

[61]:
```
(211593, 27)
```

Figura 22 – Etapa de normalização dos dados.

Fonte: Elaborado pelos autores, 2025.

Como apontado anteriormente, o SOM utiliza medidas de distância para comparar os pesos entre as variáveis. Nesta pesquisa, testam-se duas métricas: a distância de Manhattan e a distância de Cosseno.

A distância de Manhattan, também conhecida como métrica do táxi, é uma métrica de norma L1 baseada na soma dos valores absolutos das diferenças entre as variáveis. Essa norma tende a ser mais robusta e menos sensível a *outliers*, ou seja, é pouco afetada por valores extremos. Em cenários de alta dimensionalidade, essa métrica pode ser eficaz.

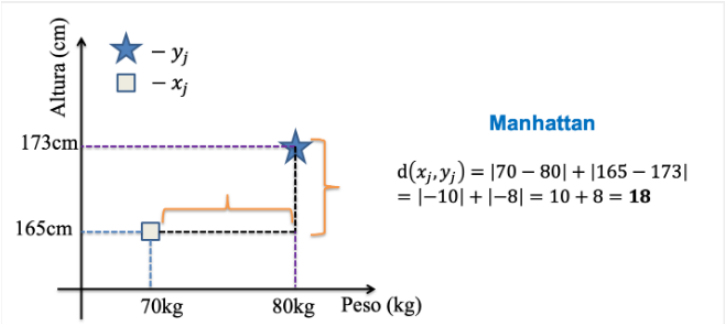

Figura 23 – Distância Manhattan.

(Escovedo; Koshiyama, 2020)

Já a distância do cosseno, ou similaridade do cosseno, é uma métrica angular que avalia o ângulo entre dois vetores no espaço (Harrington, 2012). Diferente de métricas baseadas na magnitude, ela considera apenas a direção dos vetores, ignorando os tamanhos absolutos. Essa abordagem permite medir a similaridade relativa entre os dados, identificando se eles seguem um padrão semelhante de variação. Essa métrica é eficaz em cenários onde a orientação dos dados é mais relevante do que a magnitude, como no agrupamento de vetores com padrões de variação semelhantes.

Ao iniciar o treinamento do SOM utilizando a métrica de distância Manhattan, mantiveram-se os mesmos parâmetros principais no `MiniSom`, alterando apenas a métrica de distância, taxa de aprendizado e iteração.

O `MiniSom` possui diversos hiperparâmetros que influenciam diretamente o desempenho do modelo. Na configuração apresentada na Figura 24, define-se:

- O tamanho da grade `(x, y)` como 4x4, que determina a quantidade de neurônios no mapa;
- `input_len`, que corresponde à quantidade de variáveis utilizadas como entrada;
- `sigma`, que representa o raio da vizinhança inicial — quanto maior o valor, maior a área de influência dos neurônios vizinhos no ajuste dos pesos;

- `learning_rate`, que controla a taxa de aprendizagem inicial, ou seja, o quanto os pesos dos neurônios são ajustados a cada iteração;
- `activation_distance`, que define a métrica utilizada para calcular a distância entre os vetores (neste caso, Manhattan);
- `random_seed`, que garante a reprodutibilidade dos resultados ao inicializar os pesos com uma mesma semente aleatória.

Em seguida, foi realizado o treinamento do modelo com `train_random`, utilizando 5000 iterações e inicialização aleatória dos pesos.

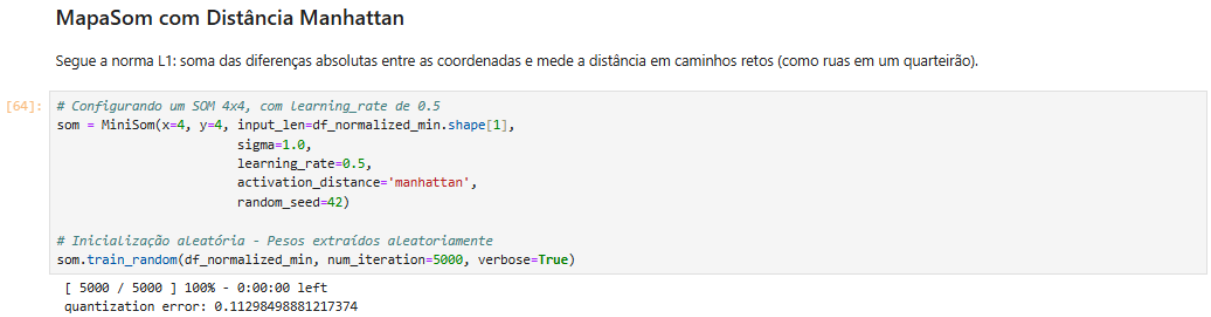

Figura 24 – Inicio do treinamento SOM com Manhattan.

Fonte: Elaborado pelos autores, 2025.

O `Quantization Error`, ou erro de quantização, é uma métrica de desempenho retornada após a finalização do treinamento, indica um erro sistemático resultante da diferença entre um valor contínuo da entrada e a saída quantizada. Quanto menor esse valor, maior a fidelidade do mapa em representar os dados de entrada e, consequentemente, melhor o desempenho do modelo em agrupar os padrões. Na figura, nota-se que o retorno foi de 0,11 — um valor considerado satisfatório. Entretando, é indispensável realizar a análise dos resultados obtidos a partir desse treinamento para compreender o comportamento do modelo.

Para essa análise, utilizou-se dois recursos visuais: o Mapa U-Matrix (Matriz de Distâncias Unificadas) e o Mapa de Frequência de Ativação.

A U-Matrix representa o SOM exibindo a distância entre os neurônios adjacentes. Nesse mapa, as cores mais claras evidenciam possíveis agrupamentos (*clusters*), enquanto as áreas mais escuras indicam maiores distâncias entre os neurônios, sugerindo separações entre regiões distintas do espaço de dados.

Já o Mapa de Frequência de Ativação mostra a frequência com que cada neurônio foi ativado ao longo do treinamento. Regiões mais escuras indicam áreas com maior concentração de ativações, enquanto regiões claras podem apontar neurônios pouco utilizados, sugerindo dados menos representativos.

```
[65]: # Visualização dos Resultados

# dividindo a área para plotar 2 gráficos
fig, axes = plt.subplots(1, 2, figsize=(12, 6))

# Mapa de distância (U-Matrix)
u_matrix = som.distance_map().T
grafico1 = axes[0].pcolor(u_matrix, cmap='bone_r')
fig.colorbar(grafico1, ax=axes[0])
axes[0].set_title("Mapa U-Matrix")

# Mapa de Frequência de Ativação
activation_map = np.zeros((4, 4))
for sample in df_normalized_min:
    winner = som.winner(sample)
    activation_map[winner] += 1

grafico2 = axes[1].pcolor(activation_map.T, cmap='bone_r')
fig.colorbar(grafico2, ax=axes[1], label='Frequência de ativação')
axes[1].set_title('Mapa de Frequência de Ativação')

plt.suptitle("Gráficos com os resultados 4x4 Manhattan")
plt.tight_layout()
plt.show()
```

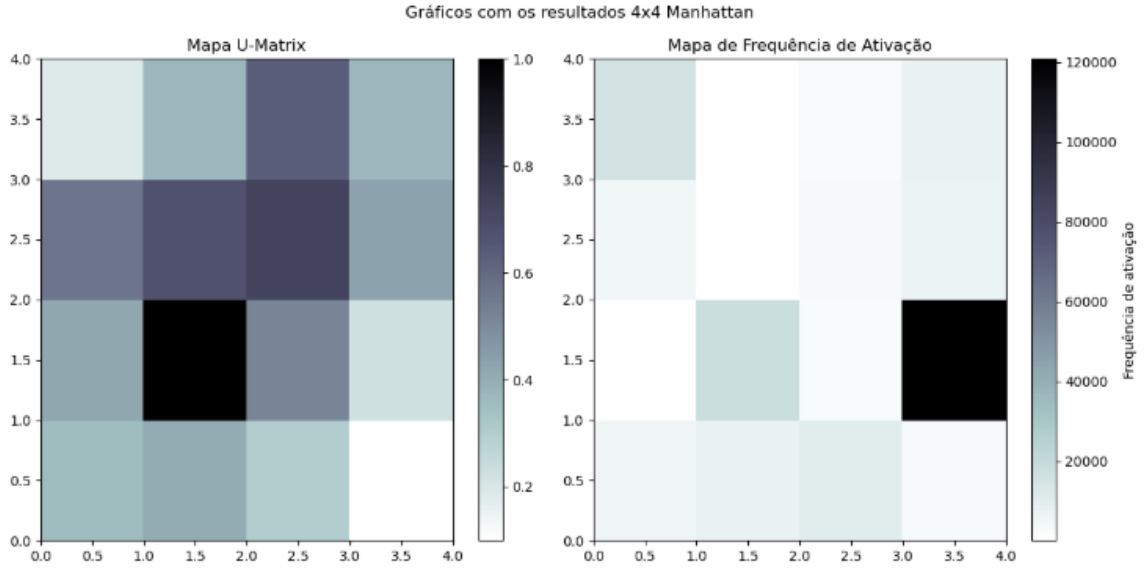

Figura 25 – Mapas do treinamento 4x4 Manhattan.

Fonte: Elaborado pelos autores, 2025.

Ao analisar a Figura 25, resultado do treinamento da Figura 24, observa-se que no Mapa U-Matrix há a presença de um nó com coloração mais clara, a princípio um possível agrupamento. Entretanto, ao comparar essa mesma posição com o Mapa de Frequência de Ativação, verifica-se que houve baixa ativação nesse ponto, indicando que poucos dados foram mapeados para esse neurônio. Dessa forma, considera-se como um falso *cluster*, visto que a região não possui a concentração de dados no treinamento atual.

Após esse resultado, foi realizado um novo treinamento, desta vez com o aumento do número de iterações para 50.000.

```
[67]: # Passando 50.000 iterações
som.train_random(df_normalized_min, num_iteration=50000, verbose=True)

# Visualização dos Resultados

fig, axes = plt.subplots(1, 2, figsize=(12, 6))

# Mapa de distância (U-Matrix)
u_matrix = som.distance_map().T
grafico1 = axes[0].pcolor(u_matrix, cmap='bone_r')
fig.colorbar(grafico1, ax=axes[0])
axes[0].set_title("Mapa U-Matrix")

# Mapa de Frequência de Ativação
activation_map = np.zeros((4, 4))
for sample in df_normalized_min:
    winner = som.winner(sample)
    activation_map[winner] += 1

grafico2 = axes[1].pcolor(activation_map.T, cmap='bone_r')
fig.colorbar(grafico2, ax=axes[1], label='Frequência de ativação')
axes[1].set_title('Mapa de Frequência de Ativação')

plt.suptitle("Gráficos com os resultados 4x4 Manhattan com mais iterações")
plt.tight_layout()
plt.show()
```

```
[ 50000 / 50000 ] 100% - 0:00:00 left
quantization error: 0.10580878008195543
```

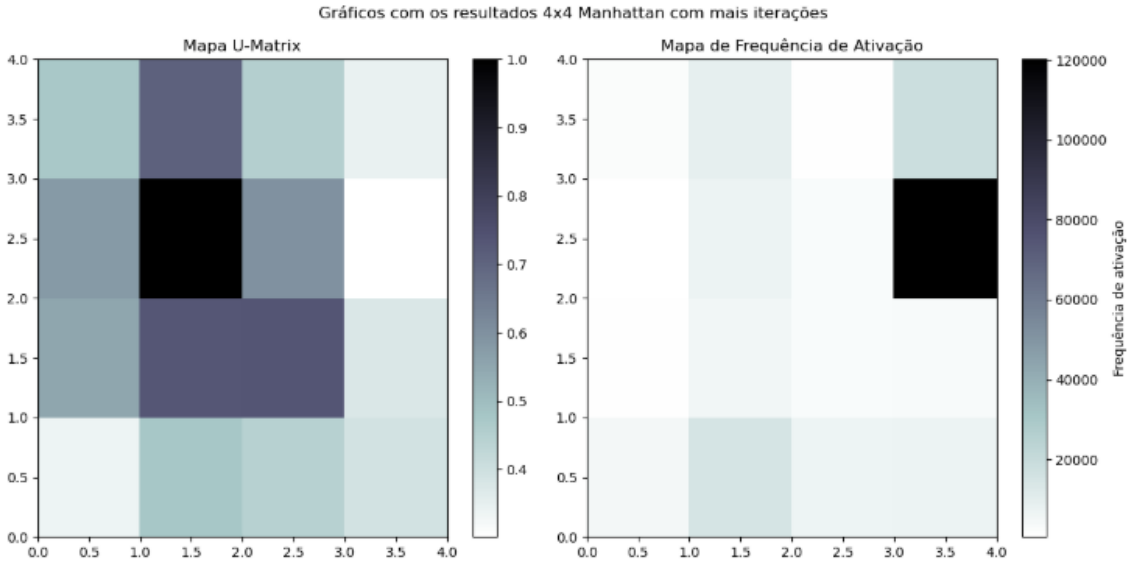

Figura 26 – Mapas do treinamento 4x4 Manhattan com 50 mil iterações.

Fonte: Elaborado pelos autores, 2025.

Com o aumento das iterações, o erro de quantização foi reduzido. A Figura 26 mostra a formação de um possível cluster, evidenciado por uma região clara no Mapa U-Matrix e uma área escura correspondente no Mapa de Frequência de Ativação. Entretanto esse agrupamento é isolado, não sendo suficiente para uma segmentação clara dos dados. Diante desses resultados, testamos diferentes tamanhos de grade para o mapa SOM, variando de 2x2 até 14x14, com o intuito de identificar em qual configuração a rede apresenta melhores resultados de agrupamento.

Após a análise dos gráficos gerados para os diferentes tamanhos de grade, não foi possível identificar a formação de clusters utilizando a distância de Manhattan, mesmo com a redução do erro de quantização. Diante disso, testou-se a utilização da distância cosseno, a fim de verificar se essa métrica apresenta melhor desempenho

no agrupamento dos dados.

Seguindo o mesmo processo realizado com a distância Manhattan, iniciou-se o treinamento com uma grade 4x4 e taxa de aprendizado de 0,5, conforme a Figura 27.

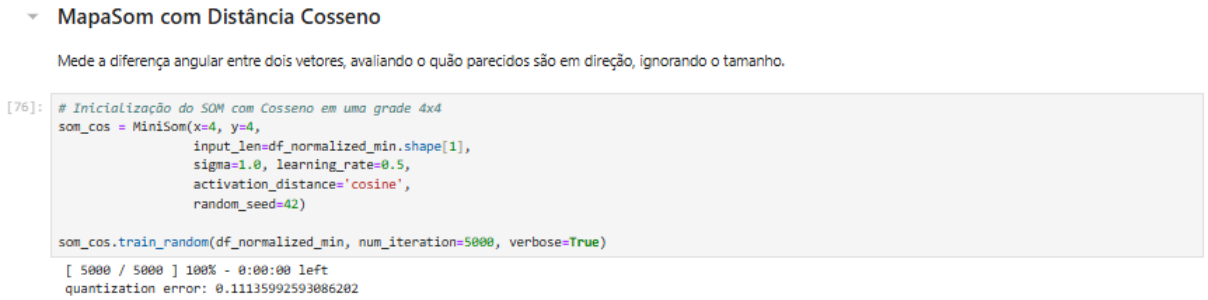

Figura 27 – Início do treinamento SOM com Cosseno.

Fonte: Elaborado pelos autores, 2025.

O erro de quantização obtido utilizando a distância cosseno foi de 0,11, semelhante ao observado com a métrica de Manhattan. Por isso, é relevante analisar os mapas gerados para verificar a eficiência do agrupamento obtido.

Figura 28 – Mapas do treinamento 4x4 Cosseno.

Fonte: Elaborado pelos autores, 2025.

Com um resultado semelhante ao obtido no treinamento com distância Manhattan e 50.000 iterações, a distância cosseno também indicou a formação de um

possível cluster isolado, o que ainda não é suficiente para afirmar a existência de um agrupamento. Diante desse cenário, optou-se por realizar novos treinamentos, testando diferentes tamanhos de grade do mapa SOM, visando identificar uma melhor formação de *cluster*.

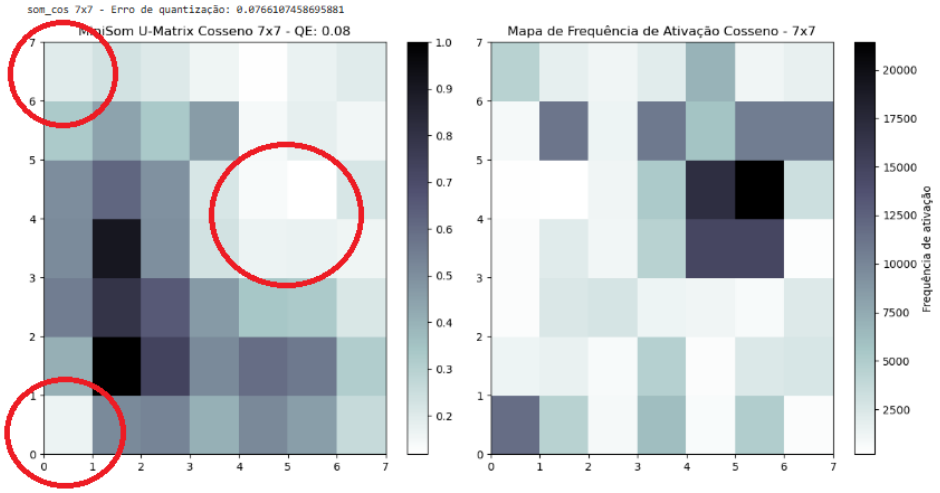

Figura 29 – Mapas do treinamento 7x7 Cosseno.

Fonte: Elaborado pelos autores, 2025.

A Figura 29 mostra três possíveis clusters. O cluster localizado no centro do mapa teve alta frequência de ativação, além de estar cercado por regiões de alta distância no Mapa U-Matrix, indicando uma separação dos demais dados, características de um agrupamento bem definido. Já os demais pontos destacados, apesar de apresentarem baixa distância entre neurônios, não estão tão ativados, mas ainda podem ser considerados possíveis *clusters*.

Com base nos resultados obtidos, sobrepõem-se gráficos de pizza aos nós da grade, com o objetivo de visualizar a distribuição dos pesos das variáveis em cada neurônio do mapa, facilitando a identificação de padrões e aglomerações.

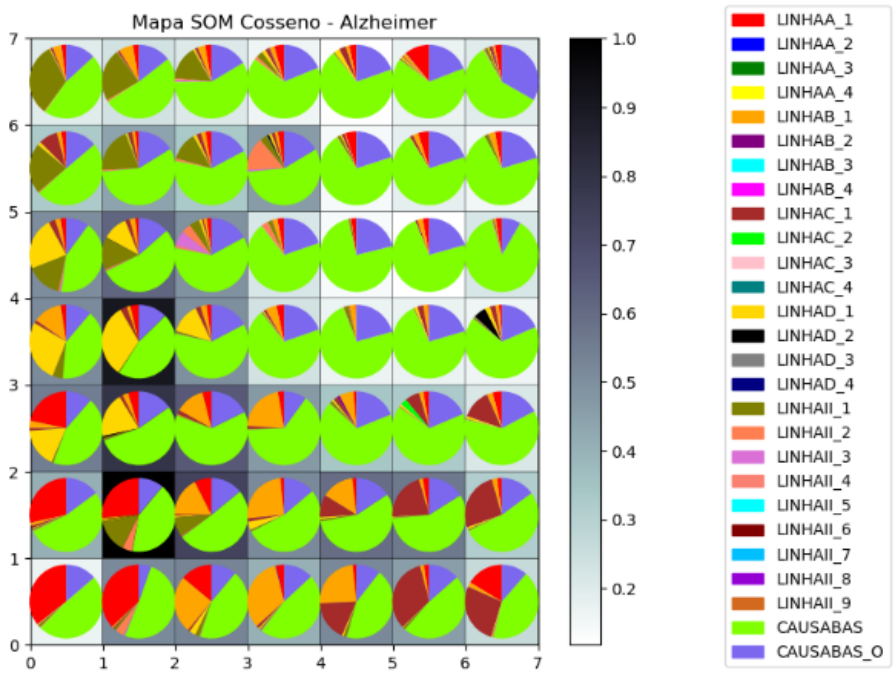

Figura 30 – Sobreposição com Gráfico de Pizza.

Fonte: Elaborado pelos autores, 2025.

Ao analisar a Figura 30, é possível observar que o cluster central apresenta uma forte predominância das colunas CAUSABAS e CAUSABAS_O, indicando uma região do mapa com alta densidade em relação a essas causas. Essa concentração está relacionada ao pré-processamento realizado, uma vez que os dados utilizados foram previamente filtrados para incluir apenas registros associados à doença de Alzheimer, representados justamente por essas duas categorias. Esse comportamento evidencia a capacidade do SOM de organizar adequadamente os dados filtrados, agrupando-os de forma coerente em uma região bem definida do mapa. No canto inferior esquerdo, identifica-se outro agrupamento com predominância da categoria LINHAA_1. Já no canto superior esquerdo, a categoria LINHAII_1 se destaca como principal componente, caracterizando um terceiro cluster com perfil bem definido.

A partir dessa análise e dos resultados obtidos com a clusterização na matriz 7x7, mantém-se essa configuração nas próximas etapas. O tamanho da grade no mapa auto-organizável (SOM) varia de acordo com a quantidade de dados e os objetivos da análise. Como nesta pesquisa são usados cerca de 200 mil registros, a grade 7x7 se mostrou suficiente para revelar padrões e agrupamentos relevantes, sem deixar o mapa nem muito fragmentado, nem genérico demais. Por isso, segue-se a utilização dessa estrutura nas análises seguintes, com alterações apenas nos hiperparâmetros.

Os hiperparâmetros controlam o comportamento do algoritmo de aprendizado e têm impacto significativo no desempenho do modelo final. Diante desse cenário, opta-

se por aumentar o número de iterações para 50.000 e ajustar a taxa de aprendizado para 0,05, com o objetivo de capturar mais detalhes e particularidades presentes nos dados, permitindo uma organização mais refinada no mapa.

A partir desses ajustes, o mapa 7x7 apresentou um erro de quantização de 0,06 e resultou na formação de três clusters, conforme apresentado na Figura 31.

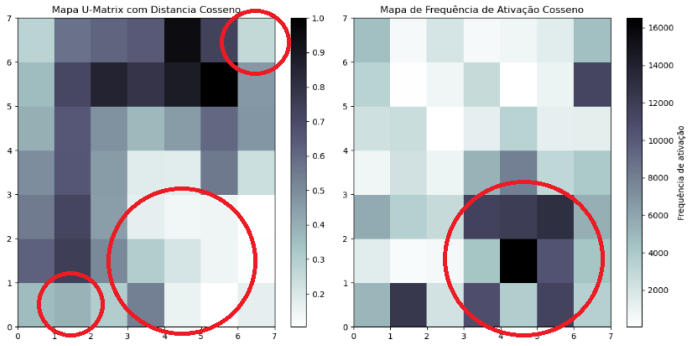

Figura 31 – Clusters do mapa 7x7 com diferentes hiperparâmetros.

Fonte: Elaborado pelos autores, 2025.

Na etapa seguinte, realiza-se a sobreposição do mapa com gráficos de pizza, conforme a Figura 32.

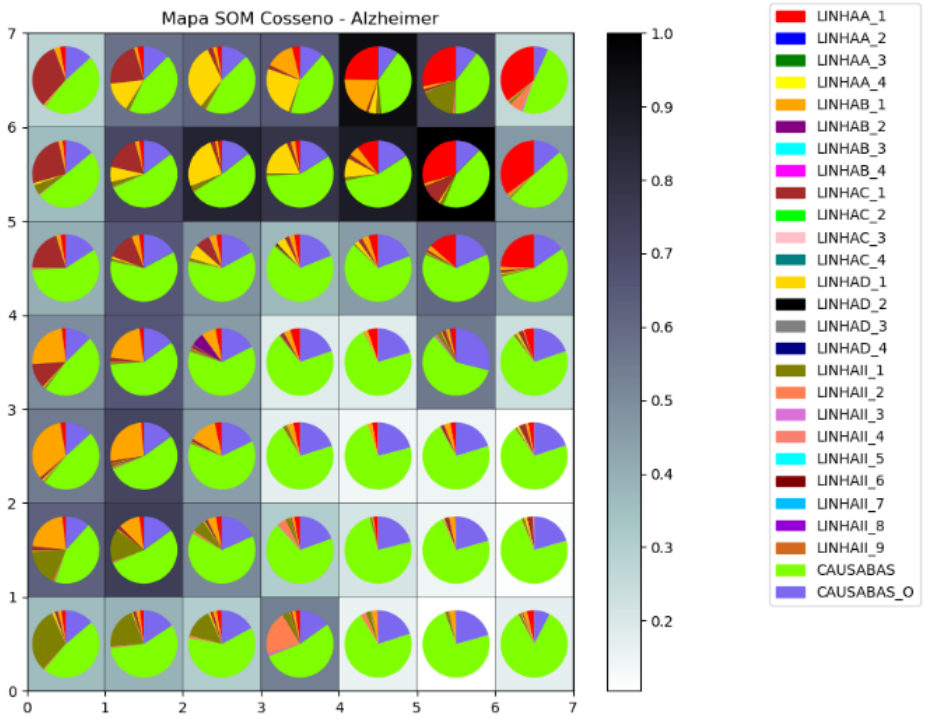

Figura 32 – Sobreposição com gráfico de pizza do mapa 7x7 com diferentes hiperparâmetros.

Fonte: Elaborado pelos autores, 2025.

Após esse resultado, inicia-se a análise dos agrupamentos formados no treinamento do mapa.

3.5.5 Análise dos Padrões gerados

A partir da sobreposição dos mapas de distância (U-Matrix) e de frequência de ativação, foi possível identificar três regiões com características de agrupamentos definidos. Esses *clusters* apresentaram padrões distintos em relação à ativação dos neurônios e à similaridade entre os dados mapeados.

- O *Cluster* 1, localizado na região central do mapa, corresponde aos neurônios nas posições (3,0), (3,1), (3,2), (4,0), (4,1), (4,2), (5,0), (5,1) e (5,2). Esse agrupamento apresentou a maior frequência de ativação entre todos os nós da rede, o que indica forte presença de padrões recorrentes nos dados. A análise da sobreposição com os gráficos de pizza revela uma predominância das colunas CAUSABAS e CAUSABAS_O, indicando que esses registros se concentram predominantemente nesse agrupamento.
- O *Cluster* 2, localizado na extremidade superior direita do mapa, na coordenada (6,6), destacou-se pela sua ativação elevada e pelo relativo isolamento em relação aos neurônios vizinhos. Os gráficos de pizza mostram predominância da coluna LINHAA_1, sugerindo que esse neurônio teve maior concetração para representar dados com esse perfil específico.
- O *Cluster* 3 foi identificado na região inferior esquerda, no neurônio (1,0). Essa área também apresentou uma frequência de ativação relevante, além de baixa distância entre os nós. Nesse agrupamento, observa-se predominância da coluna LINHAII_1.

Com a definição de três *clusters* principais a partir da análise dos mapas U-Matrix, de frequência de ativação e da sobreposição dos gráficos de pizza, inicia-se a extração e análise dos registros pertencentes a cada agrupamento. Para isso, utiliza-se a lista de coordenadas dos neurônios que compõem cada *cluster*.

A partir da iteração sobre os dados já normalizados, identifica-se o neurônio vencedor para cada amostra e, com base em sua posição no mapa SOM, os registros foram alocados na lista correspondente ao seu respectivo *cluster*. Esse processo permite segmentar o conjunto de dados original em subconjuntos, de acordo com os padrões aprendidos pelo modelo. Em seguida, os registros de cada *cluster* são convertidos em novos *DataFrames* (df_cluster_1, df_cluster_2 e df_cluster_3), permitindo análises posteriores e específicas sobre cada agrupamento.

Em seguida, realiza-se a decodificação dos CIDs presentes em cada *DataFrame* resultante da segmentação. Essa etapa teve como objetivo tornar os códigos interpretáveis, facilitando a compreensão dos tipos de causas agrupadas em cada *cluster*. Para isso, utiliza-se uma função que converte os valores previamente codificados em formato numérico durante o treinamento para sua forma textual original, no padrão alfanumérico dos códigos CID.

A função implementada para a decodificação avalia o comprimento da *string* codificada e, conforme a estrutura identificada, reconstrói o código CID original, combinando o capítulo (representado por uma letra) e os números correspondentes. Os valores decodificados são, então, aplicados aos *DataFrames* de cada *cluster*, gerando novas versões com o sufixo _descodificado. Essa etapa possibilita análises mais aprofundadas sobre os agrupamentos gerados, permitindo interpretar com maior clareza os padrões presentes em cada grupo.

A fim de compreender quais capítulos do CID estão mais presentes nos *clusters*, os códigos decodificados foram transformados em uma lista unificada. Em seguida, extraiu-se a primeira letra de cada código, que representa o capítulo ao qual ele pertence. Com isso, realiza-se a contagem da frequência de ocorrência de cada capítulo.

Posteriormente, os dados foram organizados e visualizados por meio de um gráfico de barras, no qual é possível observar a distribuição dos CIDs por capítulo. Essa etapa é fundamental para identificar quais grupos de doenças são mais associados ao agrupamento identificado, permitindo uma análise qualitativa mais aprofundada dos padrões encontrados pelo modelo.

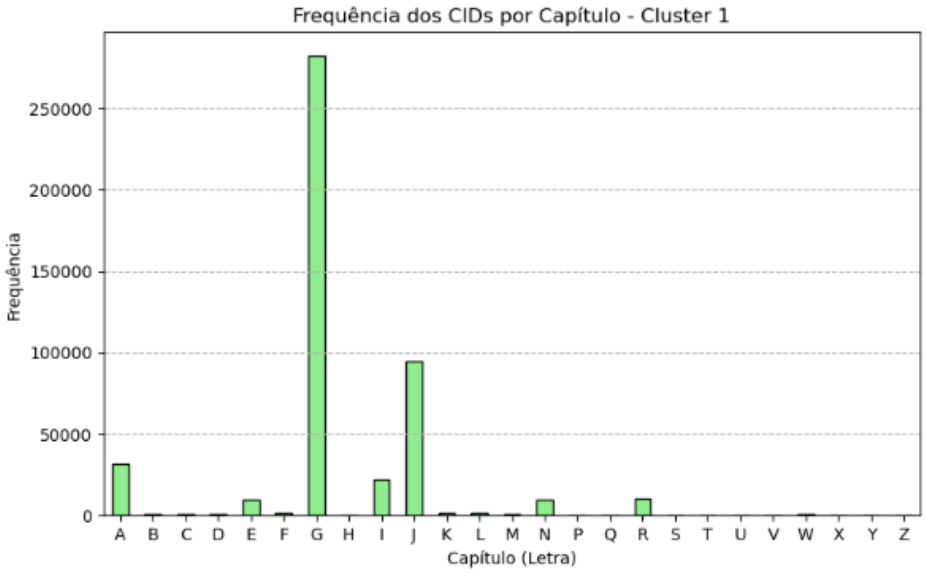

Figura 33 – Gráfico de barras dos capítulos de CID presentes no *Cluster* 1.

Fonte: Elaborado pelos autores, 2025.

A Figura 33 indica que o *Cluster* 1 apresentou maior frequência de CIDs pertencentes aos capítulos G, J e A.

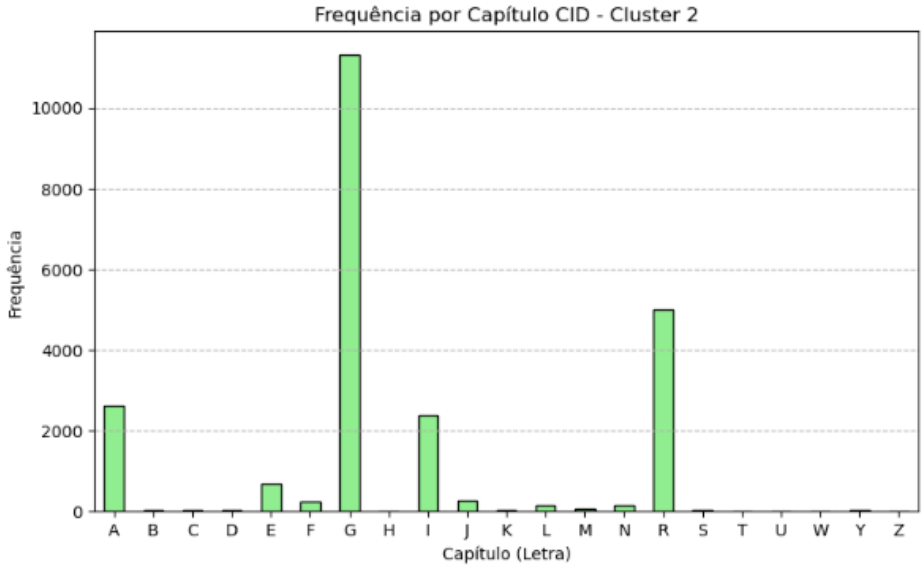

Figura 34 – Gráfico de barras dos capítulos de CID presentes no *Cluster* 2.

Fonte: Elaborado pelos autores, 2025.

A Figura 34 indica que o *Cluster* 2 apresentou maior frequência de CIDs pertencentes aos capítulos G, R, A e I.

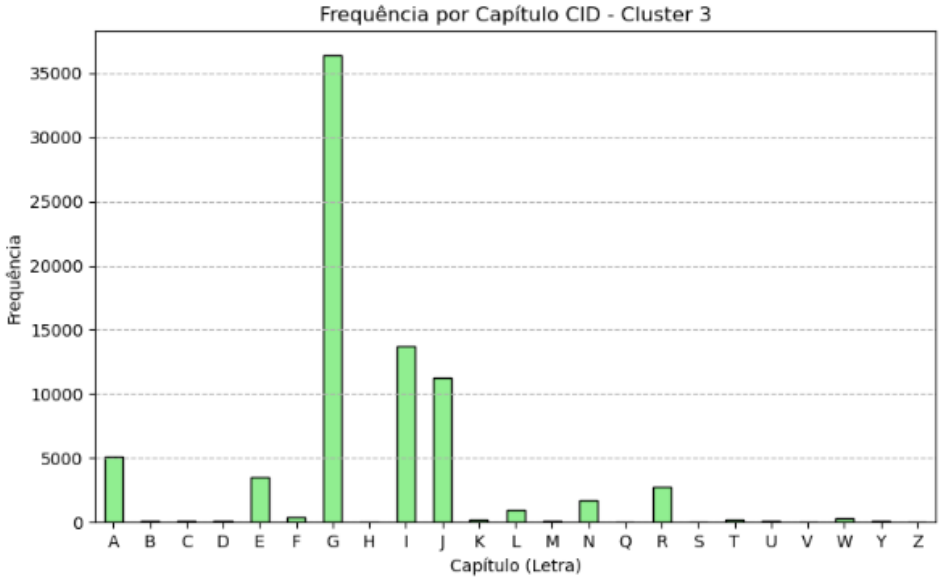

Figura 35 – Gráfico de barras dos capítulos de CID presentes no *Cluster* 3.

Fonte: Elaborado pelos autores, 2025.

A Figura 35 indica que o *Cluster* 3 apresentou maior frequência de CIDs pertencentes aos capítulos G, I, J, A e R.

Em seguida, foi realizada uma análise mais aprofundada dos CIDs presentes a cada *cluster*. Para isso, os dados das colunas referentes às linhas A, B, C, D e II foram unificados, permitindo identificar os CIDs mais frequentes em cada uma dessas categorias.

A partir disso, foram gerados gráficos de pizza que representam as dez ocorrências mais frequentes em cada linha, destacando a proporção relativa de cada CID dentro de seu respectivo conjunto. Essa visualização permite identificar de forma intuitiva os CIDs mais recorrentes em cada campo do *cluster* analisado, contribuindo para uma compreensão mais aprofundada da composição das causas associadas.

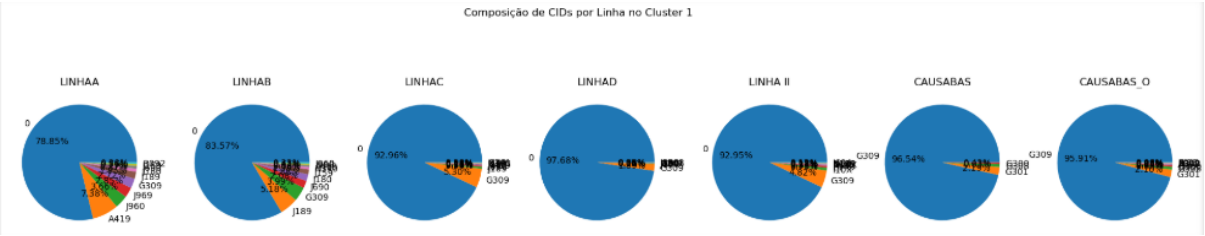

Figura 36 – Gráficos de Pizza - Composição de CIDs por Linha no *Cluster* 1

Fonte: Elaborado pelos autores, 2025.

Conforme ilustrado na Figura 32, o *Cluster* 1 se caracteriza pela predominância da variável CAUSABAS, indicando que é majoritariamente composto por registros cuja causa básica de morte já é a Doença de Alzheimer. Uma vez que o foco da pesquisa são os óbitos por Alzheimer, e considerando que os dados já foram previamente filtrados com base nessa característica, é possível concluir que esse agrupamento, embora seja um cluster válido, não oferece um significado clínico adicional relevante para a identificação de padrões. Ele representa a população geral de óbitos por Alzheimer que constitui a base da nossa análise.

Figura 37 – Gráficos de Pizza - Composição de CIDs por Linha no *Cluster* 2

Fonte: Elaborado pelos autores, 2025.

Como o *Cluster* 2 se caracterizou pela predominância da variável LINHAA_1, conforme ilustrado na Figura 32, foi elaborado um gráfico de barras com os CIDs mais frequentes nessa coluna, dentro do respectivo cluster.

O objetivo é destacar quais causas contribuíram de forma mais expressiva para a ativação desse neurônio, revelando possíveis padrões ou agrupamentos de mortalidade relacionados a essas classificações.

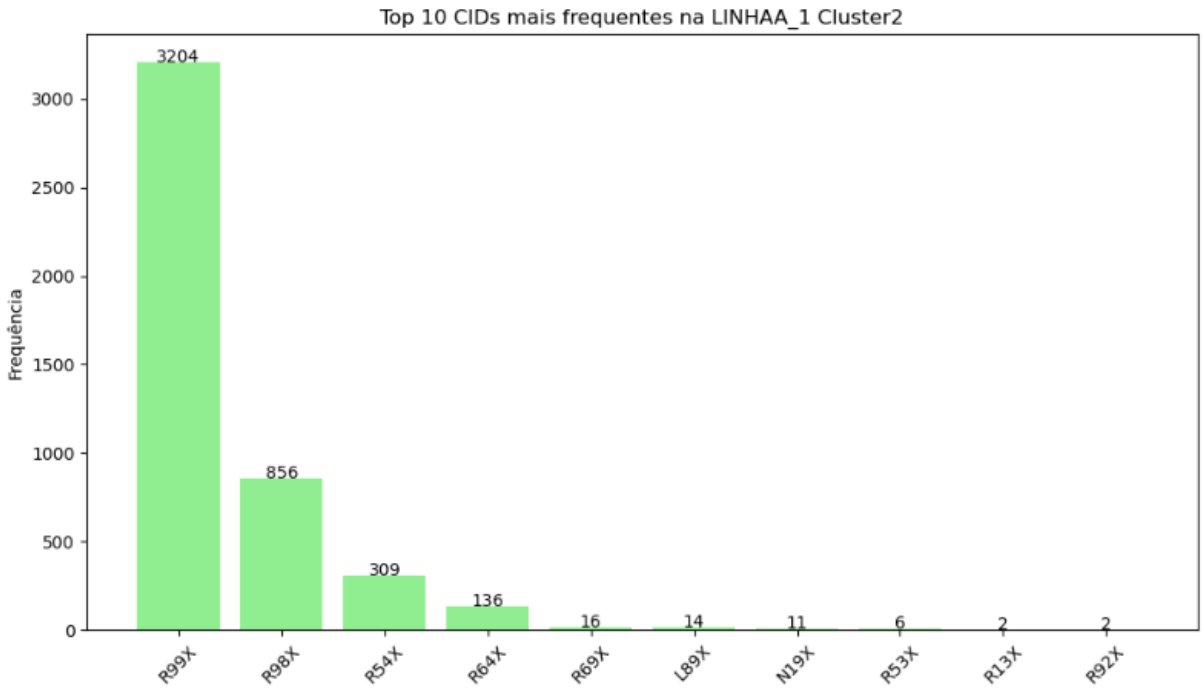

Figura 38 – Gráficos de Barra - Top 10 CIDs mais frequentes da LinhaA Cluster 2.

Fonte: Elaborado pelos autores, 2025.

Como observado na Figura 38, os CIDs com maior presença na LINHAA_1 foram os CIDs R99X, R98X, R54X, R64X.

A mesma análise foi realizada para o *Cluster* 3.

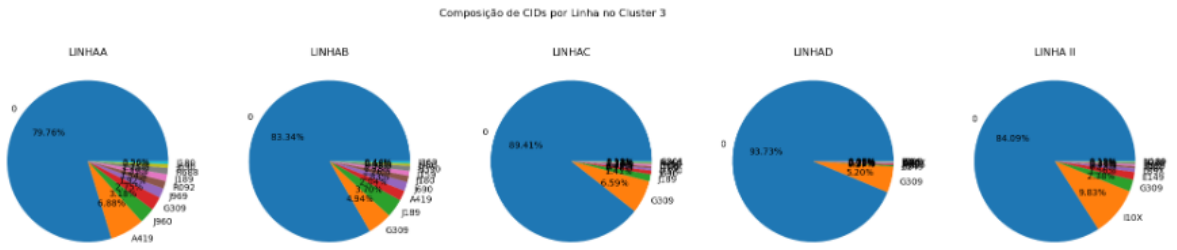

Figura 39 – Gráficos de Pizza - Composição de CIDs por Linha no *Cluster* 3

Fonte: Elaborado pelos autores, 2025.

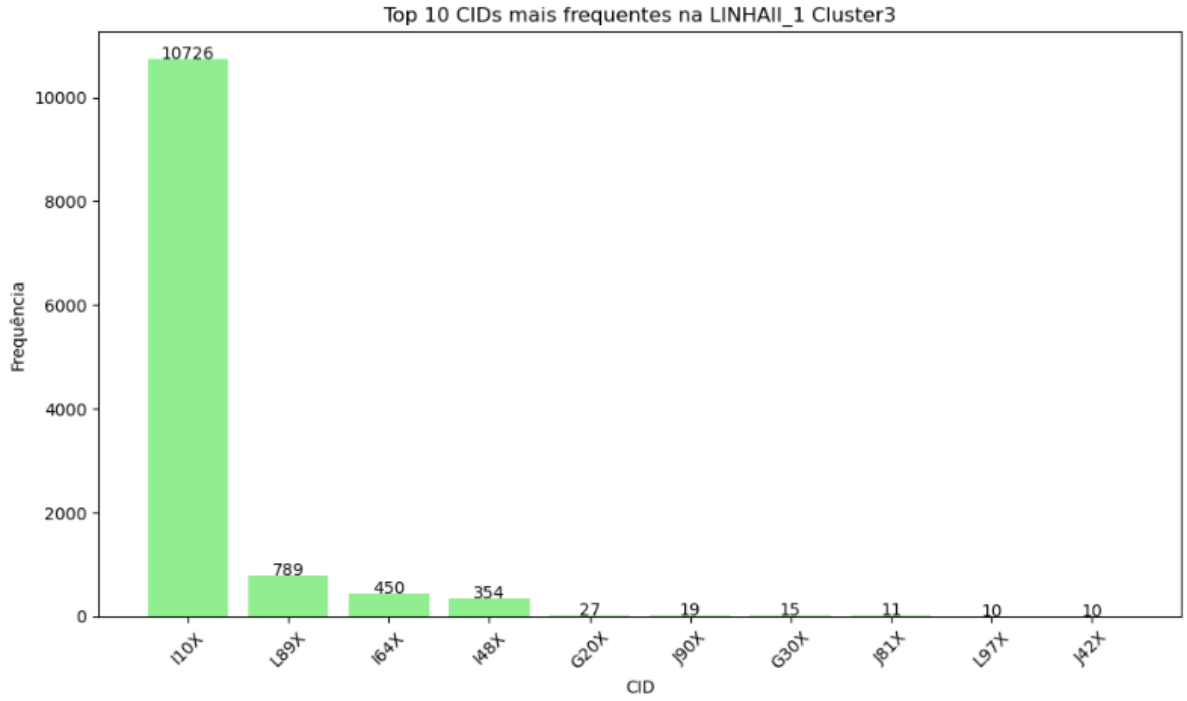

Figura 40 – Gráficos de Barra - Top 10 CIDs mais frequentes da LinhaII *Cluster* 3.

Fonte: Elaborado pelos autores, 2025.

Como observado na Figura 40, os CIDs com maior presença na LINHAII_1 foram I10X, L89X, I64X e I48X.

4 RESULTADOS E DISCUSSÃO

A aplicação do algoritmo SOM ao conjunto de dados de óbitos por Alzheimer, após pré-processamento e ajuste de hiperparâmetros (grade 7×7, distância de cosseno, 50 000 iterações), revelou três clusters distintos com base na U-Matrix e nos mapas de ativação.

4.1 Identificação dos Agrupamentos

Os agrupamentos encontrados pela aplicação do SOM foram:

Cluster 1 (região central do mapa): ativado principalmente pelas variáveis CAUSABAS e CAUSABAS_O, confirmando a Doença de Alzheimer como causa básica dominante. A análise dos capítulos CID mostrou predominância dos capítulos G (Doenças do sistema nervoso), J (Doenças do aparelho respiratório) e A (Algumas doenças infecciosas e parasitárias).

Cluster 2 (neurônio 6,6, canto superior direito): ativado pela variável LINHAA_1. Os CIDs mais frequentes foram R99X (“Outras causas mal definidas e não especificadas de mortalidade”; 70,3 %), R98X (“Morte sem assistência”; 18,8 %), R54X (“Senilidade”; 6,8 %) e R64X (“Caquexia”; 3,0 %) (Tabela 4). Os capítulos CID predominantes foram G, R (Sintomas, sinais e achados anormais), A e I (Doenças do aparelho circulatório).

CID	Descrição	Frequência (%)
R99X	Outras causas mal definidas e não especificadas de mortalidade	70,3%
R98X	Morte sem assistência	18,8%
R54X	Senilidade (envelhecimento associado a doenças e alterações patológicas que afetam a saúde.)	6,8%
R64X	Caquexia (estado de definhamento e perda de massa muscular, geralmente causado por doença crônica grave)	3,0%
Outros	-	1,1%

Tabela 4 – CIDs com maior presença na LINHAA_1 (Cluster 2), conforme mostrado na Figura 38.

Fonte: Elaborado pelos autores, 2025.

Cluster 3 (neurônio 1,0, canto inferior esquerdo): ativado pela variável LINHAII_1. Os principais CIDs contribuidores foram I10X (“Hipertensão essencial”; 86,4 %), L89X (“Úlcera de decúbito”; 6,4 %), I64X (“Acidente vascular cerebral, não especificado”; 3,6 %) e I48X (“Fibrilação e flutter atrial”; 2,9 %) (Tabela 5). Os capítulos CID mais frequentes foram G, I, J, A e R.

CID	Descrição	Frequência (%)
I10X	Hipertensão essencial (primária)	86,4%
L89X	Úlcera de decúbito (escaras)	6,4%
I64X	Acidente vascular cerebral, não especificado	3,6%
I48X	Fibrilação e flutter atrial	2,9%
Outros	-	0,7%

Tabela 5 – CIDs com maior presença na LINHAII_1 (Cluster 3), conforme mostrado na Figura 40.

Fonte: Elaborado pelos autores, 2025.

Os Clusters 2 e 3 exibiram perfis de comorbidades distintos. O Cluster 2 concentrou óbitos com causas terminais frequentemente mal definidas ou sem assistência, sugerindo lacunas no registro de óbitos avançados pela doença, enquanto o Cluster 3 revelou forte associação com comorbidades cardiovasculares e complicações de imobilidade, destacando a hipertensão e as úlceras de decúbito como fatores contribuintes importantes.

4.2 Interpretação e Implicações

Os três clusters identificados pelo algoritmo SOM revelam perfis distintos de mortalidade por Alzheimer, cada um com implicações clínicas e epidemiológicas específicas. O Cluster 1, predominantemente composto por registros com causa básica de morte codificada como Alzheimer (CAUSABAS e CAUSABAS_O), representa o grupo esperado de óbitos diretamente atribuídos à doença, funcionando como uma validação interna da metodologia aplicada.

O Cluster 2 apresenta um perfil preocupante caracterizado por causas terminais mal definidas, com destaque para "outras causas mal definidas e não especificadas de mortalidade"(R99X - 70,3%) e "morte sem assistência"(R98X - 18,8%). Este padrão sugere deficiências significativas no acompanhamento de pacientes em estágios avançados de Alzheimer, possivelmente relacionadas à morte fora do ambiente hospitalar ou sem assistência médica adequada (Morais et al., 2023). A presença expressiva de códigos como "senilidade"(R54X - 6,8%) e "caquexia"(R64X - 3,0%) reforça a hipótese de que esses óbitos ocorrem em contextos de menor suporte clínico.

O Cluster 3 revela um perfil de comorbidades cardiovasculares e complicações da imobilidade, com hipertensão essencial (I10X) representando 86,4% dos casos, seguida por úlceras de decúbito (L89X - 6,4%), acidente vascular cerebral (I64X - 3,6%) e fibrilação atrial (I48X - 2,9%). Este agrupamento alinha-se consistentemente com estudos que associam comorbidades cardiovasculares à mortalidade em demência (Wang et al., 2025), destacando hipertensão e AVC como fatores críticos. Além disso, úlceras de decúbito evidenciam complicações da imobilidade em estágios avançados da doença.

Os achados do presente estudo corroboram substancialmente a literatura internacional sobre comorbidades em Alzheimer. Estudos sistemáticos recentes reportam que a hipertensão afeta entre 30,2% a 73,9% dos pacientes com Alzheimer, enquanto diabetes mellitus varia de 6,0% a 24,3% (Lanctôt et al., 2024). A predominância de hipertensão no Cluster 3 (86,4%) situa-se no limite superior dessa faixa, sugerindo que esta comorbidade pode ser particularmente relevante como fator contributivo para a mortalidade. Segundo estudos de Carey e Fossati (2023), a hipertensão em meio de

vida aumenta significativamente o risco de demência, enquanto a tardia pode exacerbar a progressão da doença de Alzheimer.

As úlceras de decúbito identificadas no Cluster 3 encontram respaldo em estudos que demonstram que pacientes com demência avançada e úlceras de pressão apresentam sobrevida mediana significativamente menor (96 dias, em contraste com os 863 dias observados naqueles sem essas complicações) (Jaul; Meiron; Menczel, 2016). Este achado reforça a importância das complicações da imobilidade como marcadores de prognóstico reservado.

A presença de causas respiratórias em todos os clusters, embora não dominante nos principais CIDs, alinha-se com evidências de Brunnström e Englund (2009) em que doenças respiratórias são causa de morte em 45-55% dos pacientes com demência, comparado a apenas 7% na população geral da mesma faixa etária. Broncopneumonia e pneumonia aspirativa são particularmente prevalentes, refletindo as dificuldades de deglutição e coordenação respiratória características dos estágios avançados da doença.

4.3 Impacto Metodológico nos Resultados

A escolha metodológica de utilizar Mapas Auto-Organizáveis (SOM) com distância de cosseno e configuração 7x7 influenciou significativamente os resultados obtidos. O SOM demonstrou capacidade de identificação de padrões não-lineares complexos em dados de alta dimensionalidade, superando limitações de técnicas tradicionais de clustering quando aplicadas a dados com baixa correlação linear.

A sensibilidade do algoritmo aos hiperparâmetros, particularmente o número de iterações (50.000) e taxa de aprendizagem (0,05), permitiu um refinamento adequado dos agrupamentos, evidenciado pela redução do erro de quantização de 0,11 para 0,06. A métrica de distância de cosseno mostrou-se mais apropriada que a distância Manhattan para capturar similaridades nos padrões de comorbidades.

No entanto, as limitações metodológicas devem ser reconhecidas. A qualidade dos dados constitui um dos principais desafios, uma vez que há variação no detalhamento das declarações de óbito, o que pode influenciar os padrões identificados. Além disso, a ausência de variáveis clínicas, como o estágio da doença e os tratamentos realizados, impede a formulação de inferências causais mais precisas. Outro ponto crítico é a dependência de parâmetros: os resultados obtidos mostraram-se sensíveis à configuração dos hiperparâmetros do algoritmo de Mapas Auto-Organizáveis (SOM) e à escolha da métrica de distância, o que pode impactar na estabilidade e na reprodutibilidade dos clusters formados.

4.4 Recomendações e Pesquisas Futuras

Com base nos achados deste estudo, é possível propor recomendações que abrangem tanto o campo das políticas públicas de saúde quanto o das investigações futuras. No que diz respeito à gestão pública, destaca-se a importância de integrar o manejo de comorbidades cardiovasculares e a prevenção de úlceras de decúbito aos planos de cuidado voltados a pacientes com Alzheimer. Paralelamente, o desenvolvimento de sistemas específicos de investigação epidemiológica para óbitos domiciliares em indivíduos com demência pode contribuir para a redução do uso de códigos inespecíficos, como R99X e R98X, aumentando a acurácia das estatísticas vitais.

Os perfis de comorbidade identificados também oferecem subsídios para o aprimoramento das estratégias de vigilância epidemiológica, favorecendo uma alocação mais eficiente de recursos em saúde. À luz da evidência sobre a hipertensão como fator de risco modificável, recomenda-se ainda a implementação de políticas de prevenção primária da demência, especialmente por meio do controle de fatores cardiovasculares durante a meia-idade. Programas de rastreamento e controle de hipertensão, diabetes e dislipidemias devem, portanto, incorporar explicitamente a prevenção de demência entre seus objetivos.

No campo da pesquisa, futuras investigações podem se beneficiar da validação dos perfis de comorbidade por meio do pareamento de bases de dados com prontuários eletrônicos, o que aumentaria a precisão e a confiabilidade das análises. Adicionalmente, seria pertinente examinar variações regionais e sociodemográficas nos padrões de agrupamento identificados, bem como comparar o desempenho do algoritmo SOM com outras técnicas de clusterização, a fim de testar a robustez dos achados.

Outras linhas de pesquisa relevantes incluem a análise de tendências temporais entre 2012 e 2022, tanto na prevalência dos clusters quanto no impacto da pandemia de COVID-19 sobre esses padrões. Finalmente, recomenda-se o desenvolvimento de modelos preditivos baseados nos clusters identificados, com o objetivo de apoiar a estratificação de risco e o planejamento de cuidados. Técnicas de aprendizado de máquina podem ser empregadas para prever a qual cluster um paciente recém-diagnosticado com Alzheimer tem maior probabilidade de pertencer, permitindo a aplicação de intervenções preventivas mais precisas e personalizadas.

Dessa forma, os Mapas Auto-Organizáveis demonstraram ser uma ferramenta eficaz na exploração de padrões de comorbidade em óbitos por Alzheimer no Brasil, fornecendo evidências relevantes para a prática clínica, a formulação de políticas públicas e a agenda de pesquisa na área de saúde pública.

5 CONCLUSÃO

Este estudo avaliou a aplicabilidade dos Mapas Auto-Organizáveis (SOMs) na identificação e visualização de padrões de comorbidades em 211.658 óbitos por Doença de Alzheimer (DA) no Brasil, registrados no Sistema de Informações sobre Mortalidade (SIM) entre 2012 e 2022. A análise, em resposta à questão central da pesquisa, demonstrou a eficácia do método ao revelar três agrupamentos com perfis distintos e clinicamente relevantes. Um dos clusters evidenciou uma forte associação entre DA e comorbidades cardiovasculares, bem como complicações decorrentes da imobilidade, notadamente hipertensão (CID-10 I10X) e úlceras de decúbito (L89X), corroborando achados prévios da literatura. Outro agrupamento destacou-se pela alta frequência de causas terminais inespecíficas e mal definidas (R99X, R54X), sugerindo contextos de fragilidade avançada e possíveis déficits no preenchimento da Declaração de Óbito.

Com base nesses achados, recomenda-se a integração do manejo de comorbidades cardiovasculares e da prevenção de úlceras aos planos de cuidado de pessoas com Alzheimer, bem como o aprimoramento da vigilância epidemiológica para óbitos domiciliares. Além disso, os perfis identificados fornecem subsídios para a construção de modelos preditivos voltados à estratificação de risco e ao planejamento de cuidados individualizados.

A metodologia implementada — com ênfase em pré-processamento padronizado, codificação transparente e documentação em notebooks reprodutíveis — contribui para a disseminação de boas práticas em ciência de dados aplicados à saúde pública. Essa abordagem técnica fortalece a confiabilidade dos achados e serve como referência para estudos semelhantes baseados em dados secundários. Futuras pesquisas podem avançar na validação dos clusters por meio do pareamento com prontuários eletrônicos, explorar variações regionais e comparar o desempenho dos SOMs com outras técnicas de clusterização, a fim de avaliar a robustez dos resultados.

Conclui-se, portanto, que os Mapas Auto-Organizáveis se mostraram uma ferramenta eficaz na decodificação da complexidade dos óbitos por Alzheimer no Brasil, promovendo avanços tanto no entendimento epidemiológico da doença quanto na incorporação de práticas metodológicas replicáveis e de alto rigor técnico no campo da saúde coletiva.

6 CRONOGRAMA E GESTÃO DE PESSOAS

Segundo Dutra (2002), gestão de pessoas pode ser compreendida como “um conjunto de políticas e práticas que permitem a conciliação de expectativas entre a organização e as pessoas para que ambas possam realizá-las ao longo do tempo”. Esse alinhamento de expectativas é fundamental para que todos trabalhem em harmonia em direção a um mesmo objetivo.

Nesse contexto, optou-se por adotar um modelo de planejamento inspirado no KDD, que funciona de maneira incremental e iterativa: a cada etapa, revisita-se o que já foi feito, ajusta-se os passos seguintes e mantém-se espaço para mudanças conforme novas informações surgem. Essa flexibilidade permite organizar as tarefas de forma a responder rapidamente a eventuais imprevistos ou descobertas no próprio processo de análise.

Com base nessa abordagem flexível, estruturou-se um cronograma dos processos e atividades a serem realizadas por cada integrante do projeto, promovendo uma análise produtiva e direcionada para a pesquisa, conforme será descrito a seguir.

Gabriela Santana exerce o papel de analista de dados e inicia seu trabalho extraindo e validando os registros brutos do SIM, assegurando que cada arquivo de óbito esteja completo e isento de falhas. Somente depois de confirmar a qualidade dos dados é que ela passa para a filtragem dos registros de Alzheimer, removendo duplicatas e corrigindo inconsistências para que apenas as informações relevantes sejam mantidas na análise.

Com o conjunto de dados limpo e confiável, Gabriela avança para a fase de modelagem: ela configura o mapa auto-organizável (SOM), definindo a forma da rede e os parâmetros iniciais. Essa etapa é registrada detalhadamente, de modo que qualquer pessoa possa compreender as escolhas técnicas feitas. Em seguida, ela produz as visualizações exploratórias do SOM, como a U-matrix, para apresentar de forma simples como os casos de comorbidade se agrupam. A partir desses gráficos, Gabriela identifica padrões iniciais e ajusta o próprio SOM, refinando a rede neural de acordo com as métricas de desempenho obtidas.

Por fim, para garantir a reprodução completa do estudo, ela consolida todo o fluxo de trabalho em um Jupyter Notebook comentado, reunindo o código, os resultados e as instruções de execução em um único documento organizado.

Gabriel Barboza também exerce a função de analista de dados e começa preparando e enriquecendo a base de dados: ele integra fontes auxiliares que ajudam

a decodificar variáveis sociodemográficas e de CIDs. Graças a esse enriquecimento, o pré-processamento torna-se mais eficiente, pois ele consegue tratar valores ausentes e codificar todos os campos necessários com maior precisão.

Após as visualizações geradas pela colega — como a U-matrix e os mapas de clusters — Gabriel valida os padrões encontrados, comparando-os com resultados de pesquisas epidemiológicas publicadas e avaliando possíveis consistências ou divergências. Somente depois dessa validação é que ele avança para a etapa de documentação final.

Nessa fase, Gabriel redige o relatório ou artigo científico, apresentando de forma clara a discussão dos resultados e as conclusões do estudo. Por fim, tanto na revisão cruzada do notebook quanto na preparação da apresentação final do projeto, a atividade é colaborativa, em equipe, para garantir que tudo esteja claro e completo, facilitando a compreensão de qualquer leitor.

Dessa forma, a divisão de tarefas é feita de maneira estratégica, permitindo que as competências individuais sejam aproveitadas ao máximo, ao mesmo tempo em que se promove a colaboração e o compromisso com os objetivos específicos definidos para o projeto.

A seguir, apresenta-se o cronograma de atividades detalhado em formato de tabela, seguido por uma versão em gráfico de Gantt para melhor visualização.

Semana	Atividades	Responsável
01–04	Extração e validação dos dados brutos do SIM (2012–2022)	Gabriela Santana
05–08	Preparação e enriquecimento da base de dados com fontes auxiliares	Gabriel Barboza
09–10	Filtragem de registros de Alzheimer e limpeza de inconsistências	Gabriela Santana
11–13	Pré-processamento: valores ausentes, faixas etárias, codificações	Gabriel Barboza
14–18	Configuração e treinamento inicial do SOM (testes de topologia e ajustes)	Gabriela Santana
19–22	Geração de visualizações exploratórias (U-matrix, planos, clusters)	Gabriela Santana
23–25	Ajustes e refinamento do SOM conforme resultados e métricas	Gabriela Santana
26–30	Validação dos padrões encontrados com literatura epidemiológica	Gabriel Barboza
31–34	Consolidação do fluxo no Jupyter Notebook comentado	Gabriela Santana
35–36	Revisão cruzada do notebook e ajustes de clareza e reprodutibilidade	Ambos
37–40	Redação de artigo ou relatório final com discussão e conclusões	Gabriel Barboza
41–44	Preparação da apresentação final e entrega do projeto	Ambos

Tabela 6 – Cronograma de Atividades do Projeto

Fonte: Elaborado pelos autores, 2025.

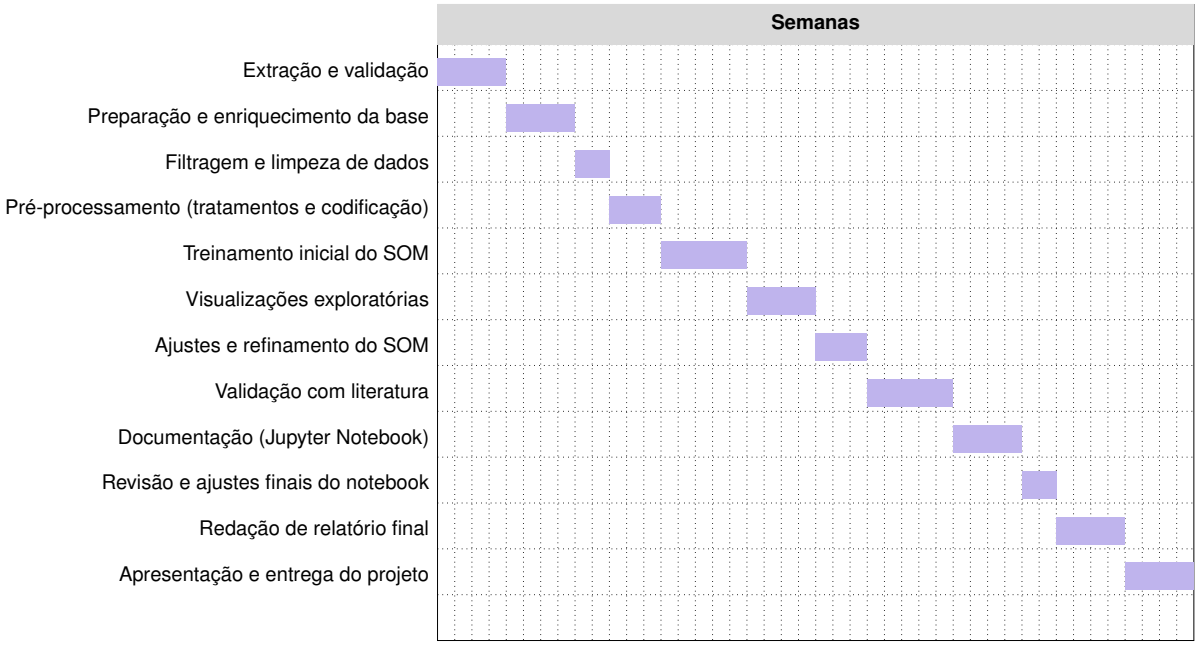

Figura 41 – Gráfico de Gantt para o Cronograma de Atividades

Fonte: Elaborado pelos autores, 2025.

REFERÊNCIAS

BRASIL. Ministério da Saúde. **Sistema de Informações sobre Mortalidade (SIM)**. Brasília, DF, 2025. Disponível em: <https://www.gov.br/saude/pt-br/composicao/svsa/sistemas-de-informacao/sim>. 17

BRUNNSTRöM, H. R.; ENGLUND, E. M. Cause of death in patients with dementia disorders. **European Journal of Neurology**, v. 16, n. 4, p. 488–492, abr. 2009. 70

CAREY, A.; FOSSATI, S. Hypertension and hyperhomocysteinemia as modifiable risk factors for alzheimer's disease and dementia: New evidence, potential therapeutic strategies, and biomarkers. **Alzheimer's & Dementia**, v. 19, n. 2, p. 671–695, fev. 2023. Epub 2022 Nov 19. 69

DUTRA, J. S. **Gestão de pessoas – modelo, processos, tendências e perspectivas**. São Paulo: Editora Atlas, 2002. 210 p. 75

ELMASRI, R.; NAVATHE, S. B. **Sistemas de banco de dados**. 4. ed. São Paulo: Pearson, 2005. 36

ESCOVEDO, T.; KOSHIYAMA, A. **Introdução a Data Science: Algoritmos de Machine Learning e métodos de análise**. Cidade, País: Editora X, 2020. 53

FOUNDATION, P. S. **Python Language Reference, version 3.x**. 2024. Disponível em: <https://docs.python.org/3/>. Acesso em: 20 de maio de 2025. 27

HARRINGTON, P. **Machine Learning in Action**. Greenwich, CT: Manning Publications Co., 2012. 53

HAYKIN, S. **Redes Neurais: Princípios e Prática**. Porto Alegre: Bookman, 2001. 50, 51

HOLT, J. M. et al. Artificial neural network approaches to identify maternal and infant risk and asset factors using peridata.net: a wi-mios study. **JAMIA Open**, v. 6, n. 3, p. ooad080, 09 2023. ISSN 2574-2531. Disponível em: <https://doi.org/10.1093/jamiaopen/ooad080>. 18

JAUL, E.; MEIRON, O.; MENCZEL, J. The effect of pressure ulcers on the survival in patients with advanced dementia and comorbidities. **Experimental Aging Research**, v. 42, n. 4, p. 382–389, jul. 2016. 70

LANCTôT, K. L. et al. Burden of illness in people with alzheimer's disease: A systematic review of epidemiology, comorbidities and mortality. **Journal of Prevention of Alzheimer's Disease**, v. 11, n. 1, p. 97–107, 2024. 69

MANABE, T. et al. Pneumonia-associated death in patients with dementia: A systematic review and meta-analysis. **PLOS ONE**, v. 14, p. e0213825, 03 2019. 17

Ministério da Saúde. **A Declaração de Óbito: documento necessário e importante**. 3. ed. Brasília: Ministério da Saúde, 2009. 28

Ministério da Saúde. **Sistema de Informação sobre Mortalidade – SIM**. 2025. <https://opendatasus.saude.gov.br/dataset/sim>. Acesso em: 08 jun. 2025. Conjunto de dados públicos, licenciados sob Creative Commons Atribuição. 28

MORAIS, G. A. Z. et al. Factors associated with the quality of death certification in brazilian municipalities: A data-driven non-linear model. **PLoS One**, v. 18, n. 8, p. e0290814, ago. 2023. 69

MUKHIYA, S. K.; AHMED, U. **Hands-On Exploratory Data Analysis with Python: Perform EDA techniques to understand, summarize, and investigate your data**. Birmingham: Packt Publishing, 2020. 33

OLIVEIRA, D. **Saiba tudo sobre SQL - A linguagem padrão para trabalhar com banco de dados relacionais!** 2023. Disponível em: <https://www.alura.com.br/artigos/o-que-e-sql>. Acesso em: 15 de out. de 2024. 27

ONO, S.; GOTO, T. Introduction to supervised machine learning in clinical epidemiology. **Annals of Clinical Epidemiology**, v. 4, n. 3, p. 63–71, 2022. 18

ORACLE. **MySQL Workbench Docs**. 2024. Disponível em: <https://www.mysql.com/products/workbench/>. Acesso em: 15 de out. de 2024. 27

PARRA-RODRíGUEZ, L. et al. Self-organizing maps to multidimensionally characterize physical profiles in older adults. **International Journal of Environmental Research and Public Health**, v. 19, n. 19, 2022. ISSN 1660-4601. Disponível em: <https://www.mdpi.com/1660-4601/19/19/12412>. 18

PASCHALIDIS, M. et al. Tendência de mortalidade por doença de alzheimer no brasil, 2000 a 2019. **Epidemiologia e Serviços de Saúde**, v. 32, n. 2, p. e2022886, 2023. 17

PETERSEN, K. K. et al. Mri-guided clustering of patients with mild dementia due to alzheimer's disease using self-organizing maps. **NeuroImage: Reports**, v. 4, n. 4, p. 100227, 2024. ISSN 2666-9560. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2666956024000333>. 18

PéREZ, F.; GRANGER, B. E. Project jupyter: Computational narratives as the engine of collaborative data science. **Proceedings of the 14th Python in Science Conference**, p. 87–90, 2015. 27

SOMMERVILLE, I. **Engenharia de Software**. 10ª. ed. São Paulo: Pearson Universidades, 2019. 22

TEIXEIRA, J. B. et al. Doença de alzheimer: estudo da mortalidade no brasil, 2000-2009. **Cadernos de Saúde Pública**, v. 31, n. 4, p. 850–860, 2015. 17

WANG, J.-H. et al. Medical comorbidity in alzheimer's disease: A nested case-control study. **Journal of Alzheimer's Disease**, v. 63, p. 773–781, 02 2025. 17, 69

Apêndices

APÊNDICE A – DIAGRAMA DE CASO DE USO ETL

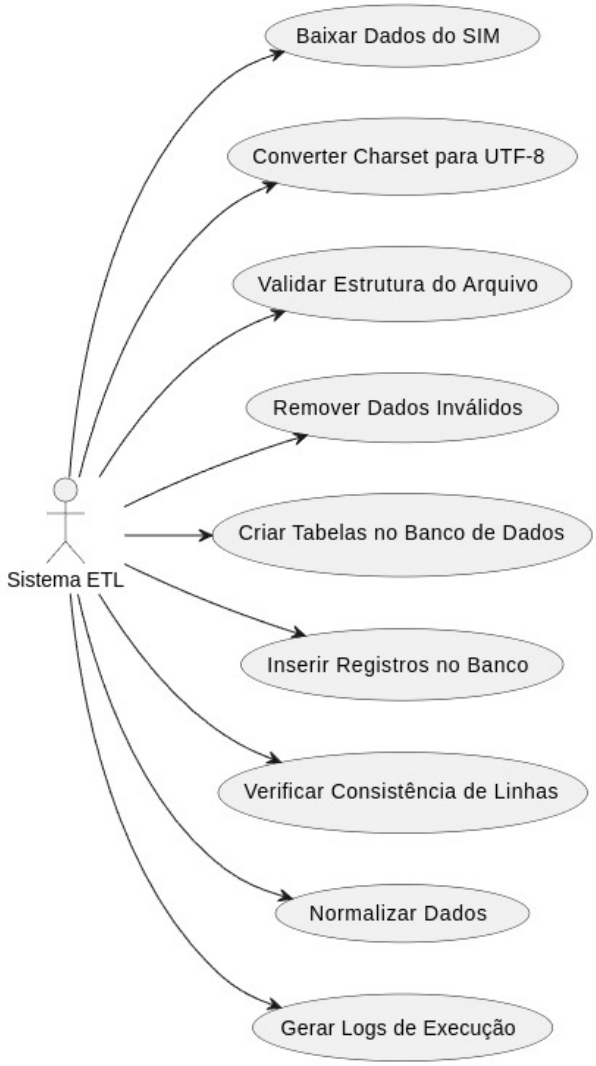

Figura 42 – Diagrama de Caso de Uso ETL

APÊNDICE B – ESTRUTURA DO SIM

Coordenação-Geral de Informações e Análise Epidemiológicas – CGIAE
Departamento de Análise em Saúde e Vigilância de Doenças Não Transmissíveis – DASNT
Secretaria de Vigilância em Saúde – SVS
Ministério da Saúde – MS

Estrutura do SIM

Posição	Nome do Campo	Tipo	Tamanho Final	Descrição	Ano Inicial	Tipo / Tamanho Inicial	Ano Alteração (1)	Tipo / Tamanho alteração (1)	Ano Alteração (2)	Tipo / Tamanho alteração (2)	Ano Alteração (3)	Tipo / Tamanho alteração (3)
42	ACIDTRAB	C	8	Indica se o evento que desencadeou o óbito está relacionado ao processo de trabalho. (1 – sim; 2 – não; 9 – ignorado)	1979	(C8)						
132	ALTCAUSA	C	8	Indica se houve correção ou alteração da causa do óbito após investigação. (1 - Sim; 2 – Não)	2014	(C8)						
16	AREARES	C	7		1979	(C7)						
31	ASSISTMED	C	9	Se refere ao atendimento médico continuado que o paciente recebeu, ou não, durante a enfermidade que ocasionou o óbito. (1 – sim; 2 – não; 9 – ignorado)	1979	(C9)						
125	ATESTADO	C	50	CIDs informados no atestado. (Códigos CID 10)	2014	(C33)	2014	(C 43)	2014	(C 48)	2014	(C 50)
32	ATESTANTE	C	9	Indica se o medico que assina atendeu o paciente 1: Sim 2: Substituto 3: IML 4: SVO 5: Outros	1979	(C9)						
15	BAIRES	C	6	Bairro de residencia. Este codigo e atribuido por cada estado e/ou municipio, nao fazendo parte da base nacional	1979	(C6)						
2	CARTORIO	C	8	Código do cartório onde o óbito foi registrado	1979	(C8)						
38	CAUSABAS	C	8	Causa básica da DO. (Códigos CID 10)	1979	(C8)						
83	CAUSABAS_O	C	10	Causa básica informada antes da resseleção. (Códigos CID 10)	2006	(C10)						
97	CAUSAMAT	C	8	CID da causa externa associada a uma causa materna. (Códigos CID 10)	2011	(C8)						
87	CB_PRE	C	6	Causa básica informada antes da resseleção (localidade). (Código CID 10)	2006	(C6)						
74	CIRCOBITO	C	9	Tipo de morte violenta ou circunstâncias em que se deu a morte não natural. (1 – acidente; 2 – suicídio; 3 – homicídio; 4 – outros; 9 – ignorado)	1996	(C9)						
34	CIRURGIA	C	8	Realização de cirurgia. (1 – sim; 2 – não; 9 – ignorado)	1979	(C8)						
78	CODBAIOCOR	C	10	Código do bairro de ocorrência.	2006	(C10)						
56	CODBAIRES	C	9	Código do Bairro de residência	1996	(C9)						
103	CODCART	C	7		2012	(C7)						
76	CODESTAB	C	8	Código do estabelecimento.	2001	(C8)						
122	CODIFICADO	C	10	Informa se formulario foi codificado. (Se estiver codificado (valor: S) ou não (valor: N))	2014	(C10)						
12	CODIGO	C	7	Codigo do estabelecimento onde ocorreu o óbito, se LOCOCOR = 1. Este codigo e atribuido por cada estado e/ou municipio, nao fazendo parte da base nacional	1979	(C6)	1995	(C 7)				
102	CODMUNCART	C	10	Código do município do cartório	2012	(C10)						
117	CODMUNNATU	C	10	Código do município de naturalidade do falecido.(Números)	2014	(C10)						
58	CODMUNOCOR	C	10	Código relativo ao município onde ocorreu o óbito. (Números)	1996	(C10)						

Figura 43 – Estrutura do SIM - 1.

57	CODMUNRES	C	6	Código do município de residência. Em caso de óbito fetal, considerar o município de residência da mãe. (Números)				1996	(C 6)					
94	COMUNSVOIM	C	10	Código do município do SVO ou do IML				2011	(C 10)					
1	CONTADOR	C	8					1979	(C 8)					
44	CRITICA	C	7	Dado para controle interno				1979	(C 7)					
119	CRM	C	15	Número de inscrição do Médico atestante no Conselho Regional de Medicina				2014	(C 8)		2014	(C 9)	2018	(C 15)
46	CRSOCOR	C	7	Regional de Saúde de ocorrencia. Este codigo e atribuido por cada estado e/ou município, não fazendo parte da base nacional				1979	(C 7)					
47	CRSRES	C	6	Regional de Saúde de residencia. Este codigo e atribuido por cada estado e/ou município, não fazendo parte da base nacional				1979	(C 6)					
9	DATAINASC	C	8	Data do nascimento do falecido. Em caso de óbito fetal as datas de óbito e nascimento deverão ser iguais.(Data no padrão ddmmaaaa)				1979	(C 8)					
6	DATAOBITO	C	9	Data em que ocorreu o óbito. (Data no padrão ddmmaaaa)				1979	(C 9)					
4	DATAREG	C	7	Data do recebimento. (Data no padrão ddmmaaaa)				1979	(C 7)					
113	DIFDATA	C	8	Diferença entre a data de óbito e data do recebimento original da DO ([DTOBITO] – [DTRECORIG]). (Números)				2012	(C 7)		2018	(C 8)		
80	DTATESTADO	C	10	Data do Atestado .				2006	(C 10)					
84	DTCADASTRO	C	10	Data do cadastro do óbito. (Data no padrão ddmmaaaa)				2006	(C 10)					
92	DTCADINF	C	8	Quando preenchido indica se a investigação foi realizada. (Data no padrão ddmmaaaa)				2010	(C 8)					
93	DTCADINV	C	8	Data do cadastro de investigação. (Data no padrão ddmmaaaa)				2010	(C 8)					
115	DTCONCASO	C	9	Data de conclusão do caso. (Data no padrão ddmmaaaa)				2012	(C 9)					
114	DTCONINV	C	8	Data da conclusão da investigação. (Data no padrão ddmmaaaa)				2012	(C 8)					
82	DTINVESTIG	C	10	Data da investigação do óbito. (Data no padrão ddmmaaaa)				2006	(C 10)					
52	DTNASC	C	8	Data do nascimento do falecido. Em caso de óbito fetal as datas de óbito e nascimento deverão ser iguais.(Data no padrão ddmmaaaa)				1996	(C 8)					
51	DTOBITO	C	8	Data em que ocorreu o óbito. (Data no padrão ddmmaaaa)				1996	(C 8)					
86	DTRECEBIM	C	9	Data do recebimento. (Data no padrão ddmmaaaa)				2006	(C 9)					
95	DTRECORIG	C	9	Data do recebimento original. (Data no padrão ddmmaaaa)				2011	(C 9)					
96	DTRECORIGA	C	10	Campo Criado no Tratamento para Data do recebimento original. (Data no padrão ddmmaaaa)				2011	(C 10)					
105	DTREGCART	C	9	Data do registro do cartório (Data no padrão ddmmaaaa)				2012	(C 9)					
54	ESC	C	3	Escolaridade em anos. (1 – Nenhuma; 2 – de 1 a 3 anos; 3 – de 4 a 7 anos; 4 – de 8 a 11 anos; 5 – 12 anos e mais; 9 – Ignorado)				1996	(C 3)					
98	ESC2010	C	7	Escolaridade 2010. Nível da última série concluída pelo falecido. (0 – Sem escolaridade; 1 – Fundamental I (1ª a 4ª série); 2 – Fundamental II (5ª a 8ª série); 3 – Médio (antigo 2º Grau); 4 – Superior incompleto; 5 – Superior completo; 9 – Ignorado)				2011	(C 7)					
108	ESCFALAGR1	C	10	Escolaridade do falecido agregada (formulário a partir de 2010). (00 – Sem Escolaridade; 01 – Fundamental I Incompleto; 02 – Fundamental I Completo; 03 – Fundamental II incompleto; 04 – Fundamental II Completo; 05 – Ensino Médio Incompleto; 06 – Ensino Médio Completo; 07 – Superior Incompleto; 08 – Superior Completo; 09 – Ignorado; 10 – Fundamental I Incompleto ou Inespecífico; 11 – Fundamental II Incompleto ou Inespecífico; 12 – Ensino Médio Incompleto ou Inespecífico)				2012	(C 10)					
59	ESCMAE	C	6	Escolaridade da mãe em anos. (1 – Nenhuma; 2 – de 1 a 3 anos; 3 – de 4 a 7 anos; 4 – de 8 a 11 anos; 5 – 12 anos e mais; 9 – Ignorado)				1996	(C 6)					
99	ESCMAE2010	C	10	Escolaridade 2010. Nível da última série concluída pela mãe. (0 – Sem escolaridade; 1 – Fundamental I (1ª a 4ª série); 2 – Fundamental II (5ª a 8ª série); 3 – Médio (antigo 2º Grau); 4 – Superior incompleto; 5 – Superior completo; 9 – Ignorado)				2011	(C 10)					

Figura 44 – Estrutura do SIM - 2.

107	ESCMAEAGR1	C	10	Escolaridade da mãe agregada (formulário a partir de 2010). (00 – Sem Escolaridade; 01 – Fundamental I Incompleto; 02 – Fundamental I Completo; 03 – Fundamental II Incompleto; 04 – Fundamental II Completo; 05 – Ensino Médio Incompleto; 06 – Ensino Médio Completo; 07 – Superior Incompleto; 08 – Superior Completo; 09 – Ignorado; 10 – Fundamental I Incompleto ou Inespecífico; 11 – Fundamental II Incompleto ou Inespecífico; 12 – Ensino Médio Incompleto ou Inespecífico)	2012	(C 10)				
118	ESTABDESCR	C	40		2014	(C 10)	2014	(C 19)	2014	(C 40)	
53	ESTCIV	C	6	Situação conjugal do falecido informada pelos familiares. (1 – Solteiro; 2 – Casado; 3 – Viúvo; 4 – Separado judicialmente/divorciado; 5 – União estável; 9 – Ignorado)	1996	(C 6)					
7	ESTCIVIL	C	8	Estado civil, conforme a tabela: 1: Solteiro 2: Casado 3: Viúvo 4: Separado judicialmente 5: União consensual (versões anteriores) 9: Ignorado	1979	(C 8)					
50	ETNIA	C	5		1995	(C 5)					
33	EXAME	C	5	Realização de exame. (1 – sim; 2 – não; 9 – ignorado)	1979	(C 5)					
112	EXPOSIFDATA	C	10		2012	(C 10)					
26	FILHMORT	C	8	Numero de filhos mortos ignorados, nao incluindo o proprio. O valor zero corresponde a Nenhum filho morto e está codificado como XX.	1979	(C 8)					
25	FILHVIVOS	C	9	Numero de filhos vivos. O valor zero corresponde a Ignorado. Nenhum filho vivo esta codificado como XX.	1979	(C 9)					
75	FONTE	C	5	fonte de informação utilizada para o preenchimento dos campos 48 e 49. (1 – ocorrência policial; 2 – hospital; 3 – família; 4 – outra; 9 – ignorado)	1996	(C 5)					
85	FONTEINV	C	8	Fonte de investigação (1 – Comitê de Morte Materna e/ou Infantil; 2 – Visita domiciliar / Entrevista família; 3 – Estabelecimento de Saúde / Prontuário; 4 – Relacionado com outros bancos de dados; 5 – S V O; 6 – I M L; 7 – Outra fonte; 8 – Múltiplas fontes; 9 – Ignorado)	2006	(C 8)					
127	FONTES	C	6	Combinado de caracteres conforme o preenchimento dos campos de fontes (FONTENTREV, FONTEAMBUL, FONTEPRONT, FONTESVO, FONTEIML, FONTEPROF): se preenchido caractere "S", se o campo estiver vazio caractere "X". (Letras)	2014	(C 6)					
131	FONTESINF	C	9		2014	(C 9)					
41	FONTINFO	C	8		1979	(C 8)					
63	GESTACAO	C	8	Faixas de semanas de gestação (1 - Menos de 22 semanas; 2 - 22 a 27 semanas; 3 - 28 a 31 semanas; 4 - 32 a 36 semanas; 5 - 37 a 41 semanas; 6 - 42 e + semanas)	1996	(C 8)					
62	GRAVIDEZ	C	8	Tipo de gravidez (1 – única; 2 – dupla; 3 – tripla e mais; 9 – ignorada)	1996	(C 8)					
77	HORAOBITO	C	9	Horário do óbito. (Números (padrão 24 horas 00:00))	2006	(C 9)					
10	IDADE	C	5	Idade do falecido em minutos, horas, dias, meses ou anos. (Idade: composto de dois subcampos. - O primeiro, de 1 dígito, indica a unidade da idade (se 1 = minuto, se 2 = hora, se 3 = mês, se 4 = ano, se = 5 idade maior que 100 anos). - O segundo, de dois dígitos, indica a quantidade de unidades: Idade menor de 1 hora: subcampo varia de 01 e 59 (minutos); De 1 a 23 Horas: subcampo varia de 01 a 23 (horas); De 24 horas e 29 dias: subcampo varia de 01 a 29 (dias); De 1 a menos de 12 meses completos: subcampo varia de 01 a 11 (meses); Anos - subcampo varia de 00 a 99; - 9 - Ignorado)	1979	(C 5)					
23	IDADEMAE	C	8	Idade da mãe. (Números)	1979	(C 8)					
24	INSTRMAE	C	8	Instrução da mae, conforme codificação de INSTRUCAO	1979	(C 8)					
21	INSTRPAI	C	8	Instrução do pai, conforme codificação de INSTRUCAO	1979	(C 8)					
19	INSTRUCAO	C	9	Instrução, conforme a tabela:	1979	(C 9)					

Figura 45 – Estrutura do SIM - 3.

				0: Ignorado 1: Nenhuma 2: Primeiro grau 3: Segundo grau 4: Superior								
69	LINHAA	C	20	CIDs informados na Linha A da DO referente ao diagnóstico na Linha A da DO (causa terminal - doença ou estado mórbido que causou diretamente a morte). (Códigos CID 10)	1996	(C 6)	1999	(C15)	1999	(C 20)		
70	LINHAB	C	20	CIDs informados na Linha B da DO referente ao diagnóstico na Linha B da DO (causa antecedente ou consequencial - estado mórbido, se existir, que produziu a causa direta da morte registrada na linha A). (Códigos CID 10)	1996	(C 6)	1999	(C10)	1999	(C 15)	1999	(C 20)
71	LINHAC	C	20	CIDs informados na Linha C da DO referente ao diagnóstico na Linha C da DO (causa antecedente ou consequencial - estado mórbido, se existir, que produziu a causa direta da morte registrada na linha A). (Códigos CID 10)	1996	(C 6)	1999	(C20)				
72	LINHAD	C	20	CIDs informados na Linha D da DO referente ao diagnóstico na Linha D da DO (causa básica – estado mórbido, se existir, que produziu a causa direta da morte registrada na linha A). (Códigos CID 10)	1996	(C 6)	1999	(C10)	1999	(C 15)	1999	(C 20)
73	LINHAII	C	30	CIDs informados na Parte II da DO referente ao diagnóstico na Parte II da DO (causa contribuinte - outras condições significativas que contribuíram para a morte e que não entraram na cadeia definida na Parte I). (Códigos CID 10)	1996	(C 7)	1999	(C20)	2006	(C 25)	2006	(C 30)
43	LOCAOCID	C	7	Indica o local do acidente, se cabível, conforme a tabela: 0: Ignorado 1: Via Publica 2: Domicilio 3: Outro 4: Local de trabalho	1979	(C 7)						
11	LOCOCOR	C	7	Local de ocorrência do óbito. (1 – hospital; 2 – outros estabelecimentos de saúde; 3 – domicílio; 4 – via pública; 5 – outros; 6 - aldeia indígena; 9 – ignorado).	1979	(C 7)						
89	MORTEPARTO	C	10	Momento do óbito em relação ao parto. (1 - antes; 2- durante; 3-depois; 9-Ignorado)	2009	(C 10)						
13	MUNIOCOR	C	8	Município de ocorrencia do obito, conforme codificação do IBGE. Veja mais adiante, neste documento, as observações sobre a codificação de municípios	1979	(C 8)						
14	MUNRES	C	7	Município de residencia, em codificação idêntica a de MUNIOCOR. Os óbitos de residentes no exterior ou de residencia completamente ignorada foram desprezados nos anos de 1979 a 1992	1979	(C 7)						
18	NATURAL	C	7	País e Unidade da Federação onde faleceido nascou. Se estrangeiro informar País. (Números)	1979	(C 7)						
35	NECROPSIA	C	9	Refere-se a execução ou não de necropsia para confirmação do diagnóstico. (1 – sim; 2 – não; 9 – ignorado)	1979	(C 9)						
130	NUDIASINF	C	9		2014	(C 9)						
126	NUDIASOBCO	C	10	Diferença entre a data óbito e a data conclusão da investigação, em dias. (Números)	2014	(C 10)						
116	NUDIASOBIN	C	10		2012	(C 10)						
88	NUMERODN	C	8	Número da Declaração de Nascido Vivo. (Números)	2006	(C 8)						
120	NUMEROLOTE	C	10	Número do lote. (Números)	2014	(C 10)						
45	NUMEXPORT	C	9	Dado para controle interno	1979	(C 9)						
104	NUMREGCART	C	10	Número do registro do cartório	2012	(C 10)						

Figura 46 – Estrutura do SIM - 4.

36	OBITOPUERP	C	8	Para obitos femininos em idade fertil, indica se estava gravida no momento da morte, conforme a tabela: 0: Ignorado 1: Sim 2: Não	1979	(C 8)			
37	OBITOFE2	C	8	Para obitos femininos em idade fertil, indica se esteve gravida nos 12 meses anteriores a morte, conforme a tabela: 0: Ignorado 1: Sim 2: Não	1979	(C 8)			
67	OBITOGRAV	C	9	Óbito na gravidez. (1 – sim; 2 – não; 9 – ignorado)	1996	(C 9)			
65	OBITOPARTO	C	10	Momento do óbito em relação ao parto. (1 - antes; 2- durante; 3 -depois; 9- Ignorado)	1996	(C 10)			
68	OBITOPUERP	C	10	Óbito no puerpério. (1 – Sim, até 42 dias após o parto; 2 – Sim, de 43 dias a 1 ano; 3 – Não; 9 – Ignorado)	1996	(C 10)			
55	OCUP	C	6	Tipo de trabalho que o falecido desenvolveu na maior parte de sua vida produtiva. Preenchimento de acordo com Classificação Brasileira de Ocupações – CBO 2002. (Números)	1996	(C 5)	2006	(C6)	
17	OCUPACAO	C	8	Ocupação, conforme a Classificação Brasileira de Ocupações (CBO)	1979	(C 8)			
22	OCUPMAE	C	7	Tipo de trabalho exercido habitualmente pela Mãe, de acordo com Classificação Brasileira de Ocupações – CBO 2002. No caso da mãe do falecido(a) ser "aposentada", preencher com a ocupação habitual anterior.	1979	(C 7)			
20	OCUPPAI	C	7	Ocupação do pai, conforme codificação de OCUPACAO	1979	(C 7)			
91	ORIGEM	C	6	Origem do registro. (1 - Oracle; 2 - Banco estadual disponibilizado via FTP; 3 - Banco SEADE; 9 - Ignorado)	2010	(C 6)			
64	PARTO	C	5	Tipo de parto. (1 – vaginal; 2 – cesareo; 9 – ignorado)	1996	(C 5)			
66	PESO	C	4	Peso ao nascer em gramas. (Número (quatro algarismos))	1996	(C 4)			
30	PESONASC	C	8	Peso ao nascer, em gramas	1979	(C 8)			
61	QTDFILMORT	C	10	Número de filhos mortos. Não incluir a criança cujo óbito se notifica na respectiva DO. (Número; 9 - ignorado)	1996	(C 10)			
60	QTDFILVIVO	C	10	Número de filhos vivos. (Número; 9 - ignorado)	1996	(C 10)			
49	RACACOR	C	7	Cor informada pelo responsável pelas informações do falecido. (1 – Branca; 2 – Preta; 3 – Amarela; 4 – Parda; 5 – Indígena)	1995	(C 7)			
3	REGISTRO	C	8	Numero de registo do óbito	1979	(C 8)			
110	SEMAGESTAC	C	10	Semanas de gestação com dois algarismos. (Números com dois algarismos; 9 - Ignorado)	2012	(C 10)			
27	SEMANGEST	C	9	Semanas de gestação, conforme as tabelas: Para os anos de 1979 a 1994 : 0: Ignorado 1: Menos de 20 semanas 2: 20 a 27 semanas 3: 28 e mais semanas Para os anos a partir de 1995 0: Ignorado 4: Menos de 21 semanas 5: 22 a 27 semanas 6: 28 a 36 semanas 7: 37 a 41 semanas 8: 42 semanas e mais	1979	(C 9)			

Figura 47 – Estrutura do SIM - 5.

106	SERIESCOPAL	C	10	Ultima série escolar concluída pelo falecido. (Números de 1 a 8)	2012	(C 10)	
109	SERIESCMAE	C	10	Última série escolar concluída pela mãe. (Números de 1 a 8)	2012	(C 10)	
8	SEXO	C	4	Sexo do falecido. "Ignorado" selecionada em casos especiais como cadáveres mutilados, em estado avançado de decomposição, genitália indefinida ou hermafroditismo. (M – masculino; F – feminino; I - ignorado)	1979	(C 4)	
121	STCODIFICA	C	10	Status de instalação. (Se codificadora (valor: S) ou não (valor: N))	2014	(C 10)	
100	STDOEPIDEM	C	10	Status de DO Epidemiológica. (1 - Sim; 0 - Não)	2011	(C 10)	
101	STDONOVA	C	8	Status de DO Nova. (1 - Sim; 0 - Não)	2011	(C 8)	
40	TIPOACID	C	8	Indica o tipo de acidente, se cabível: 0: Ignorado 1: Atropelamento 2: Demais acidentes de transito 3: Queda 4: Afogamento 5: Outros tipos de acidente	1979	(C 8)	
5	TIPOBITO	C	8	Tipo do óbito Óbito fetal: morte antes da expulsão ou da extração completa do corpo da Mãe, independentemente da duração da gravidez. Indica o óbito o fato de o feto, depois da expulsão do corpo materno, não respirar nem apresentar nenhum outro sinal de vida, como batimentos do coração, pulsações do cordão umbilical ou movimentos efetivos dos músculos de contração voluntária. (1-Fetal; 2-Não Fetal)	1979	(C 8)	
28	TIPOGRAV	C	8	Tipo de gravidez, conforme a tabela: 0: Ignorado 1: Única 2: Dupla 3: Tríplice 4: Mais de 3	1979	(C 8)	
29	TIPOPARTO	C	9	Tipo de parto, conforme a tabela: 0: Ignorado 1: Espontaneo 2: Operatório 3: Forceps 4: Outro	1979	(C 9)	
39	TIPOVIOL	C	8	Indica o tipo de violencia, se cabível, conforme a tabela: 0: Ignorado 1: Homicidio 2: Suicidio 3: Acidente 4: Outros tipos de violencia	1979	(C 8)	
79	TPASSINA	C	8	-	2006	(C 8)	
111	TPMORTEOCO	C	10	Situação gestacional ou pósgestacional em que ocorreu o óbito. (1 – na gravidez; 2 – no parto; 3 – no abortamento; 4 – até 42 dias após o término do parto; 5 – de 43 dias a 1 ano após o término da gestação ; 8 – não ocorreu nesses períodos; 9 – Ignorado)	2012	(C 10)	
129	TPNIVELINV	C	10	Tipo de nível investigador. (E – estadual; R - regional; M - Municipal)	2014	(C 10)	
90	TPOBITOCOR	C	10	Momento da ocorrência do óbito. (1-Durante a gestação, 2- Durante o abortamento, 3- Após o abortamento , 4- No parto ou até 1 hora após o parto, 5- No puerpério - até 42 dias após o parto, 6- Entre 43 dias e até 1 ano após o parto, 7- A investigação não identificou o momento do óbito, 8- Mais de um ano após o parto , 9- O óbito não ocorreu nas circunstancias anteriores, Branco - Não investigado)	2009	(C 10)	
81	TPPOS	C	5	Óbito investigado. (1 – sim; 2 – não)	2006	(C 5)	
128	TPRESGINFO	C	10	Informa se a investigação permitiu o resgate de alguma causa de óbito não informado, ou a correção de alguma antes informada. (01 - Não acrescentou	2014	(C 10)	

Figura 48 – Estrutura do SIM - 6.

				nem corrigiu informação; 02 - Sim, permitiu o resgate de novas informações; 03 - Sim, permitiu a correção de alguma das causas informadas originalmente)		
48	UFINFORM	C	8	Código da UF que informou o registro.	1979	(C 8)
124	VERSAOSCB	C	9	Versão do seletor de causa básica. (Números)	2014	(C 9)
123	VERSAOSIST	C	10	Versão do sistema. (Números)	2014	(C 10)

Obs 1: As definições contidas nesta planilha têm origem nos arquivos Estrutura_SIM_Anterior.DBF e Estrutura_SIM_para_CD que podem ser obtidos em:

http://www2.datasus.gov.br/DATASUS/index.php?area=0901

Obs 2: As descrições foram atualizadas e homologadas pela área técnica da CGIAE

Figura 49 – Estrutura do SIM - 7.