

UNIVERSIDADE FEDERAL DE OURO PRETO
INSTITUTO DE CIÊNCIAS EXATAS E BIOLÓGICAS
DEPARTAMENTO DE COMPUTAÇÃO

RUI AUGUSTO DRUMOND MENDONÇA

**OBSERVASRAG-MG: VIGILÂNCIA E TRANSPARÊNCIA DE DADOS
DA SÍNDROME RESPIRATÓRIA AGUDA GRAVE EM MINAS GERAIS**

Ouro Preto
2026

RUI AUGUSTO DRUMOND MENDONÇA

**OBSERVASRAG-MG: VIGILÂNCIA E TRANSPARÊNCIA DE DADOS DA SÍNDROME
RESPIRATÓRIA AGUDA GRAVE EM MINAS GERAIS**

Monografia apresentada ao Curso de Ciência da Computação da Universidade Federal de Ouro Preto como parte dos requisitos necessários para a obtenção do grau de Bacharel em Ciência da Computação.

Orientadora: Prof. Dra. Fernanda Sumika Hojo De Souza

Ouro Preto
2026



MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
INSTITUTO DE CIÊNCIAS EXATAS E BIOLÓGICAS
DEPARTAMENTO DE COMPUTAÇÃO



FOLHA DE APROVAÇÃO

Rui Augusto Drumond Mendonça

OBSERVASRAG-MG: VIGILÂNCIA E TRANSPARÊNCIA DE DADOS DA SÍNDROME RESPIRATÓRIA AGUDA GRAVE EM MINAS GERAIS

Monografia apresentada ao Curso de Ciência da Computação da Universidade Federal de Ouro Preto como requisito parcial para obtenção do título de Bacharel em Ciência da Computação

Aprovada em 27 de Julho de 2026.

Membros da banca

Fernanda Sumika Hojo de Souza (Orientadora) - Doutora - Universidade Federal de Ouro Preto
Aline Norberta de Brito (Examinadora) - Doutora - Universidade Federal de Ouro Preto
Daniel Ludovico Guidoni (Examinador) - Doutor - Universidade Federal de Ouro Preto

Fernanda Sumika Hojo de Souza, Orientadora do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 27/07/2026.



Documento assinado eletronicamente por **Fernanda Sumika Hojo de Souza, PROFESSOR DE MAGISTERIO SUPERIOR**, em 28/07/2026, às 09:53, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).



A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **1142662** e o código CRC **A5A96080**.

Resumo

Esta monografia propõe o desenvolvimento de uma plataforma interativa para aprimorar a transparência e a vigilância epidemiológica da Síndrome Respiratória Aguda Grave (SRAG) no estado de Minas Gerais, utilizando dados do Sistema de Informação de Vigilância Epidemiológica da Gripe (SIVEP-Gripe). A metodologia do projeto está dividida em três etapas principais: seleção dos dados de interesse, preparação dos dados e persistência dos dados. São utilizadas as bases de dados do SIVEP-Gripe referentes ao período de 2021 a 2026, contemplando informações demográficas, clínicas e geográficas dos casos notificados. Os dados coletados são tratados por meio de um *pipeline* ETL e armazenados em formatos otimizados (Parquet e DuckDB). Para a análise e visualização dos dados, são utilizadas as bibliotecas Python, Pandas e Streamlit, resultando em *dashboards* interativos, com filtros por classificação etiológica e município, capazes de auxiliar na vigilância epidemiológica e na tomada de decisões na área da saúde.

Palavras-chave: Saúde pública. Visualização de dados. Vigilância epidemiológica. SRAG. Minas Gerais.

Abstract

This work proposes the development of an interactive platform to enhance transparency and epidemiological surveillance of Severe Acute Respiratory Syndrome (SARS) in the state of Minas Gerais, using data from the Influenza Epidemiological Surveillance Information System (SIVEP-Gripe). The project's methodology is divided into three main stages: data selection, data preparation, and data persistence. SIVEP-Gripe databases from 2021 to 2026 are used, encompassing demographic, clinical, and geographic information of reported cases. The collected data is processed through an ETL pipeline and stored in optimized formats (Parquet and DuckDB). For data analysis and visualization, Python, Pandas, and Streamlit libraries are utilized, resulting in interactive dashboards with filters by etiological classification and municipality, capable of assisting in epidemiological surveillance and health decision-making processes.

Keywords: Public health. Data visualization. Epidemiological surveillance. SARI. Minas Gerais.

Lista de Figuras

Figura 2.1 – Etapas do processo KDD (Knowledge Discovery in Databases), desde os dados brutos até a geração de conhecimento.	8
Figura 3.1 – Arquitetura da aplicação: do dado bruto ao <i>dashboard</i> interativo.	20
Figura 4.1 – Aba de visão geral do ObservaSRAG-MG: indicadores-chave de desempenho e evolução anual dos casos por classificação etiológica.	27
Figura 4.2 – Aba demográfica do ObservaSRAG-MG: perfil dos pacientes por sexo, faixa etária, raça/cor e escolaridade.	28
Figura 4.3 – Aba epidemiológica do ObservaSRAG-MG: curva mensal de casos, comparação sazonal entre anos, taxa de letalidade e prevalência de comorbidades. . .	28
Figura 4.4 – Aba de mapa do ObservaSRAG-MG: distribuição geográfica dos casos de SRAG pelos 853 municípios de Minas Gerais em escala logarítmica de cores. .	29

Lista de Tabelas

Tabela 2.1 – Principais categorias de variáveis do SIVEP-Gripe utilizadas neste trabalho.	11
Tabela 3.1 – Tabelas geradas pelo <i>pipeline</i> ETL.	17
Tabela 4.1 – Volume de registros de SRAG por ano. * Ano incompleto.	22
Tabela 4.2 – Proporção de valores ausentes ou “ignorados” em campos selecionados.	23
Tabela 4.3 – Comparação de desempenho entre o acesso aos dados brutos e aos dados curados.	24
Tabela 4.4 – Casos de SRAG por classificação etiológica e ano. * Ano incompleto.	24
Tabela 4.5 – Letalidade por SRAG por ano em Minas Gerais. * Ano incompleto.	25

Sumário

1	Introdução	1
1.1	Objetivos	3
1.1.1	Objetivo Geral	3
1.1.2	Objetivos Específicos	3
1.2	Justificativa	3
1.3	Organização da Monografia	3
2	Revisão Bibliográfica	5
2.1	Fundamentação Teórica	5
2.1.1	<i>Big Data</i>	5
2.1.2	<i>Data Analytics</i>	6
2.1.3	Processo KDD (Knowledge Discovery in Databases)	7
2.1.4	Visualização de Dados	8
2.2	Vigilância Epidemiológica e SRAG	9
2.2.1	Síndrome Respiratória Aguda Grave (SRAG)	9
2.2.2	SIVEP-Gripe: estrutura e variáveis	10
2.2.3	Contexto epidemiológico em Minas Gerais (2021–2026)	12
2.3	Trabalhos Relacionados	12
3	Metodologia	14
3.1	Seleção dos Dados de Interesse	14
3.2	Preparação dos Dados	15
3.2.1	Leitura e filtragem	15
3.2.2	Limpeza e normalização	15
3.2.3	Construção do modelo de dados	16
3.3	Persistência dos Dados	17
3.4	Construção dos <i>Dashboards</i>	18
3.5	Síntese: Arquitetura da Aplicação	19
3.6	Ferramentas	21
4	Resultados	22
4.1	Mapeamento e qualidade dos dados	22
4.2	<i>Pipeline ETL</i> e ganho de desempenho	23
4.3	Análises demográficas, epidemiológicas e geográficas	24
4.3.1	Evolução temporal e transição etiológica	24
4.3.2	Sazonalidade e curva epidemiológica	25
4.3.3	Letalidade	25
4.3.4	Perfil demográfico	25
4.3.5	Comorbidades	26

4.3.6	Distribuição geográfica	26
4.4	A plataforma ObservaSRAG-MG	26
4.5	Discussão e comparação com a literatura	29
5	Considerações Finais	31
5.1	Trabalhos Futuros	32
 Referências		 33

1 Introdução

O mundo contemporâneo está imerso em um cenário onde a geração e a coleta de dados ocorrem em uma escala sem precedentes. A cada segundo, bilhões de dados são gerados a partir de uma variedade de fontes, como redes sociais, dispositivos móveis, sensores em cidades e, de maneira crescente, serviços de saúde. A quantidade e diversidade desses dados oferecem uma oportunidade única para análises mais profundas e tomadas de decisões baseadas em informações precisas e atualizadas. No entanto, o desafio reside em transformar essa massa de dados brutos em informações valiosas, capazes de informar e guiar políticas públicas eficazes (Marr, 2017).

Nesse contexto, o conceito de cidades inteligentes ganha relevância. Cidades inteligentes utilizam tecnologia e dados para melhorar a eficiência dos serviços, a qualidade de vida dos cidadãos e a sustentabilidade urbana. Em cidades inteligentes, os dados são fundamentais para criar políticas públicas baseadas em evidências, permitindo que os gestores tomem decisões mais informadas e ágeis. A utilização de grandes volumes de dados para monitorar e otimizar o funcionamento das cidades é uma tendência que tem o potencial de transformar o planejamento urbano, incluindo áreas cruciais como a saúde pública (Jonek-Kowalska, 2022).

A evolução da tecnologia na saúde culmina no conceito de Saúde 5.0 (Wazid et al., 2022), que se caracteriza pela integração de sistemas inteligentes para aprimorar o atendimento ao paciente e a gestão dos serviços de saúde. Entretanto, essa transformação traz consigo desafios significativos, como a segurança e a proteção dos dados dos pacientes, além de questões éticas relacionadas ao uso de dados pessoais (Kostkova et al., 2016). A implementação de tecnologias avançadas em saúde exige um equilíbrio cuidadoso entre inovação e a preservação da privacidade, bem como a garantia de que as decisões automatizadas respeitem princípios éticos.

O Sistema Único de Saúde (SUS) (Noronha; Lima; Machado, 2008) foi concebido pela Constituição Federal de 1988 e regulamentado pela Lei nº 8.080 de 19 de setembro de 1990, que estabeleceu os princípios de universalidade, equidade e integralidade no atendimento à saúde pública no Brasil (Brasil, 1990). Desde sua implantação, o SUS representa um marco na saúde pública brasileira, oferecendo acesso universal e gratuito a cuidados médicos para toda a população. O Departamento de Informática do Sistema Único de Saúde (DATASUS) (Lima et al., 2015) desempenha um papel central ao coletar, armazenar e disponibilizar dados que apoiam a tomada de decisões em saúde pública. Dentre os diversos sistemas mantidos pelo DATASUS, o Sistema de Informação de Vigilância Epidemiológica da Gripe (SIVEP-Gripe) (Brasil. Ministério da Saúde, 2024) é responsável pelo registro dos casos de Síndrome Respiratória Aguda Grave (SRAG) em todo o território nacional, constituindo uma das bases mais relevantes para a vigilância de doenças respiratórias no país.

A Síndrome Respiratória Aguda Grave (Ministério da Saúde, 2026a) é uma condição

clínica que pode ser causada por diferentes agentes etiológicos, como os vírus Influenza, o SARS-CoV-2 ([Ministério da Saúde, 2026b](#)) (COVID-19) e outros vírus respiratórios. A pandemia de COVID-19, iniciada em 2020, evidenciou a importância da vigilância da SRAG ao gerar um volume massivo de notificações no SIVEP-Gripe, tornando essa base de dados ainda mais relevante para a compreensão do cenário epidemiológico brasileiro ([Ranzani et al., 2021](#)). No estado de Minas Gerais, com seus 853 municípios e mais de 20 milhões de habitantes, a análise dos dados de SRAG é particularmente desafiadora devido à heterogeneidade territorial e demográfica do estado.

Para lidar com o grande volume de dados gerados pelo SIVEP-Gripe e outras fontes do SUS, áreas de estudo como ciência de dados, *big data*, mineração de dados e inteligência artificial se mostram essenciais. Essas ferramentas permitem a análise de grandes conjuntos de dados de forma eficiente, identificando padrões, tendências e *insights* que não seriam perceptíveis a olho nu. No contexto da vigilância epidemiológica, essas áreas são cruciais para transformar dados brutos em informações acionáveis, otimizando a gestão de recursos e aprimorando o monitoramento de doenças respiratórias ([Galetsi; Katsaliaki; Kumar, 2020](#); [Wu et al., 2013](#)).

Um dos grandes desafios no contexto da vigilância epidemiológica é a transformação de dados brutos, muitas vezes não estruturados, em informações úteis que possam guiar a tomada de decisões. Embora o SIVEP-Gripe disponibilize seus dados publicamente, os arquivos são extensos, com milhões de registros nacionais por ano, e exigem processamento significativo para extração de informações relevantes em nível estadual ou municipal. A falta de ferramentas adequadas e acessíveis para esse processo limita o potencial dos dados em gerar melhorias efetivas na vigilância e na resposta a surtos de doenças respiratórias ([Matheus; Janssen; Maheshwari, 2020](#)).

O estado de Minas Gerais ([Instituto Brasileiro de Geografia e Estatística, 2022](#)), localizado na região Sudeste do Brasil, é a unidade federativa com o maior número de municípios do país. Essa vasta extensão territorial e a diversidade socioeconômica entre suas regiões resultam em realidades epidemiológicas distintas, onde municípios como Belo Horizonte concentram grande volume de notificações, enquanto municípios menores podem apresentar dinâmicas epidemiológicas distintas em termos de incidência, letalidade e perfil demográfico dos casos. Essa heterogeneidade demanda ferramentas de análise que permitam tanto uma visão panorâmica do estado quanto o detalhamento por município e por classificação etiológica da SRAG.

Considerando este breve contexto apresentado, a questão de pesquisa que norteou a monografia foi definida: É possível construir uma plataforma interativa para vigilância e transparência dos dados de SRAG em Minas Gerais, baseada em técnicas de análise e visualização de dados, utilizando dados públicos do SIVEP-Gripe?

1.1 Objetivos

1.1.1 Objetivo Geral

Promover transparência dos dados de Síndrome Respiratória Aguda Grave do estado de Minas Gerais por meio de uma plataforma interativa que combine técnicas de análise e visualização de dados, utilizando dados do SIVEP-Gripe, para aprimorar a vigilância epidemiológica.

1.1.2 Objetivos Específicos

- Mapear e analisar a qualidade dos dados disponíveis no SIVEP-Gripe para o estado de Minas Gerais, no período de 2021 a 2026;
- Desenvolver um *pipeline* de extração, transformação e carga (ETL) para processar e persistir os dados em formatos otimizados para consulta;
- Construir *dashboards* interativos com filtros por classificação etiológica e município, contemplando análises demográficas, epidemiológicas e geográficas dos casos de SRAG.

1.2 Justificativa

Apesar de existirem plataformas nacionais que disponibilizam dados do SIVEP-Gripe, como o InfoGripe da Fiocruz (Bastos et al., 2019), as soluções voltadas para o estado de Minas Gerais são poucas. Nessa perspectiva, identificou-se uma plataforma de visualização mantida pela Secretaria de Estado de Saúde de Minas Gerais (SES-MG), desenvolvida em Microsoft Power BI (Secretaria de Estado de Saúde de Minas Gerais, 2024), que apresenta indicadores de SRAG com granularidade municipal para todo o estado. No entanto, essa ferramenta apresenta instabilidade de disponibilidade, ficando ocasionalmente fora do ar. A solução proposta nesta monografia se diferencia pelo uso exclusivo de ferramentas de código aberto — como Python, DuckDB e Streamlit —, o que favorece a auditabilidade, a reprodutibilidade e a adaptação por outros pesquisadores, além de proporcionar maior desempenho no processamento de grandes volumes de dados.

A plataforma desenvolvida nesta monografia permite a identificação de padrões sazonais, a comparação entre municípios, a análise de fatores de risco como comorbidades e faixas etárias, além do acompanhamento da taxa de letalidade ao longo dos anos. Essas informações podem auxiliar gestores de saúde na alocação de recursos e na resposta a surtos de doenças respiratórias.

1.3 Organização da Monografia

A estrutura desta monografia está organizada da seguinte forma:

No Capítulo 2, é apresentada a revisão bibliográfica e o arcabouço teórico que fundamentam o trabalho, abordando conceitos de Big Data, Data Analytics, processo KDD e visualização de dados, além do contexto epidemiológico da Síndrome Respiratória Aguda Grave (SRAG), da estrutura do SIVEP-Gripe e dos trabalhos relacionados à vigilância epidemiológica e visualização de dados em saúde pública.

O Capítulo 3 descreve a metodologia e os procedimentos utilizados no desenvolvimento da plataforma, incluindo o processo de extração, transformação e carga (ETL), a modelagem e persistência dos dados, as tecnologias empregadas e a construção dos *dashboards* interativos.

O Capítulo 4 apresenta e discute os resultados obtidos, organizados segundo os objetivos específicos: o mapeamento e a análise da qualidade dos dados do SIVEP-Gripe para Minas Gerais, o desempenho do *pipeline* ETL e da persistência otimizada, e as análises demográficas, epidemiológicas e geográficas viabilizadas pelo *dashboard*, encerrando com a comparação frente aos trabalhos relacionados.

Por fim, o Capítulo 5 traz as considerações finais, retomando o atendimento dos objetivos propostos, as limitações identificadas e as possibilidades de trabalhos futuros.

2 Revisão Bibliográfica

2.1 Fundamentação Teórica

Nesta seção, são apresentados conceitos teóricos necessários para entender a proposta desta monografia, como vigilância em saúde pública, análise, tratamento e visualização de dados.

2.1.1 *Big Data*

O conceito de *Big Data* tem sido amplamente discutido na literatura devido à sua capacidade de lidar com grandes volumes de dados provenientes de diversas fontes. Esse conceito começou a emergir na década de 1990, à medida que a tecnologia de armazenamento e processamento de dados evoluía, com o advento da internet e a digitalização de informações. De acordo com (Tsai et al., 2015), *Big Data* é caracterizado pelos “Vs”: Volume, Velocidade e Variedade. Esses aspectos permitem que organizações e pesquisadores processem e analisem grandes quantidades de dados, obtendo *insights* que podem informar a tomada de decisões:

- **Volume** refere-se à quantidade massiva de dados gerados a cada segundo.
- **Velocidade** representa a rapidez com que os dados são gerados e precisam ser processados.
- **Variedade** diz respeito aos diferentes tipos de dados disponíveis, como textos, imagens, vídeos e dados estruturados.

Além dos três Vs clássicos, Veracidade e Valor têm se tornado fundamentais nas discussões contemporâneas sobre *Big Data* (Debattista et al., 2015). Veracidade refere-se à qualidade e confiabilidade dos dados, um aspecto crucial, especialmente na área da saúde, onde decisões podem impactar diretamente a vida dos pacientes. Dados imprecisos ou incompletos podem levar a diagnósticos errôneos e a intervenções inadequadas. Assim, garantir a veracidade dos dados é essencial para a eficácia das análises realizadas.

Valor, por sua vez, diz respeito à capacidade de transformar dados em informações acionáveis que influenciem decisões informadas e estratégias eficazes. No contexto da saúde pública, isso significa não apenas coletar grandes volumes de dados, mas também ser capaz de interpretá-los de maneira a oferecer melhorias concretas na qualidade dos serviços. A criação de *dashboards* interativos, por exemplo, pode ajudar gestores de saúde a monitorar indicadores em tempo real e a identificar áreas que necessitam de atenção urgente.

Esses atributos tornam o *Big Data* uma ferramenta essencial para enfrentar desafios complexos, como os encontrados na saúde pública. A capacidade de analisar dados provenientes

de diversas fontes, como registros eletrônicos de saúde, redes sociais e sensores de dispositivos *wearables*, permite a identificação de padrões que podem prever surtos de doenças ou monitorar a eficácia de campanhas de vacinação. Um exemplo notável é o uso de algoritmos de aprendizado de máquina para prever a disseminação de doenças infecciosas, permitindo uma resposta mais rápida e eficaz (Zillner; Neururer, 2016).

No entanto, a adoção do *Big Data* não está isenta de desafios. Questões de privacidade são particularmente críticas, uma vez que os dados de saúde são considerados sensíveis. A implementação de normas rigorosas para garantir a proteção das informações dos pacientes é vital. Além disso, a segurança dos dados se torna uma preocupação premente, visto que vazamentos de informações podem resultar em consequências graves. Outro desafio importante é a infraestrutura: muitas organizações de saúde ainda carecem de sistemas adequados para armazenar, processar e analisar grandes volumes de dados, o que limita seu potencial de uso (Tole, 2013).

Portanto, enquanto o *Big Data* oferece oportunidades sem precedentes para a melhoria da saúde pública, sua implementação bem-sucedida depende da superação desses desafios, garantindo que os dados não apenas sejam acessíveis, mas também utilizados de forma ética e responsável.

2.1.2 *Data Analytics*

Data Analytics refere-se ao processo de examinar conjuntos de dados para extrair conclusões sobre as informações que eles contêm. Segundo (Marr, 2015), a análise de dados pode transformar dados brutos em informações acionáveis, essenciais para a vigilância em saúde pública. Através de técnicas estatísticas e algoritmos de aprendizado de máquina, *Data Analytics* permite identificar padrões, tendências e correlações que não seriam perceptíveis a olho nu.

Na área da saúde, *Data Analytics* desempenha um papel crucial em várias frentes:

- **Monitoramento de tendências epidemiológicas:** Permite a identificação precoce de surtos de doenças, ajudando a implementar medidas preventivas de forma mais eficaz.
- **Alocação eficiente de recursos de saúde:** Ajuda os gestores a distribuírem recursos onde eles são mais necessários, otimizando o uso dos mesmos.
- **Avaliação da eficácia de intervenções de saúde pública:** Facilita a medição do impacto de programas e políticas de saúde, permitindo ajustes baseados em evidências.

Além disso, a combinação de *Data Analytics* com *Big Data* pode potencializar ainda mais a capacidade de resposta em saúde pública. A integração de dados de diferentes fontes, como registros eletrônicos de saúde, pesquisas de saúde e dados ambientais, oferece uma visão mais

holística da saúde da população, permitindo intervenções mais direcionadas e eficazes (Galetsis; Katsaliaki; Kumar, 2020).

Esses processos são especialmente importantes em um sistema de saúde como o SUS, onde a capacidade de gerenciar e utilizar grandes volumes de dados pode resultar em melhorias significativas na prestação de serviços de saúde e na saúde da população. A implementação eficaz de *Data Analytics* não apenas melhora a eficiência do sistema de saúde, mas também pode contribuir para a redução das desigualdades em saúde, promovendo um atendimento mais equitativo.

2.1.3 Processo KDD (Knowledge Discovery in Databases)

O processo KDD, ou Descoberta de Conhecimento em Bases de Dados, é um método sistemático que inclui a seleção, pré-processamento, transformação, mineração de dados e interpretação/avaliação dos resultados (Fayyad; Piatetsky-Shapiro; Smyth, 1996). Cada uma dessas etapas desempenha um papel crucial na extração de conhecimento útil a partir de grandes volumes de dados:

- **Seleção:** Nesta fase, os dados relevantes são identificados a partir de grandes bancos de dados, garantindo que apenas as informações úteis para o problema em questão sejam consideradas.
- **Pré-processamento:** Essa etapa envolve a limpeza e a preparação dos dados, removendo inconsistências e preenchendo lacunas, o que é essencial para garantir a qualidade da análise.
- **Transformação:** Os dados são convertidos em um formato adequado para a mineração. Isso pode incluir normalização, agregação e redução de dimensionalidade, ajudando a melhorar a eficiência da mineração.
- **Mineração de Dados:** Aqui, técnicas de análise são aplicadas para descobrir padrões e relacionamentos nos dados. Métodos estatísticos e de aprendizado de máquina são frequentemente utilizados nesta fase.
- **Interpretação/Evaluação:** Os resultados da mineração de dados são interpretados, avaliando sua relevância e validade, e integrando-os ao conhecimento existente.

No contexto da saúde pública, o processo KDD pode ser utilizado para identificar padrões epidemiológicos, apoiar a tomada de decisões baseadas em dados e melhorar a qualidade dos dados do SUS. Por exemplo, a análise de dados históricos de doenças pode ajudar a detectar surtos emergentes, permitindo que as autoridades de saúde implementem intervenções precoces e direcionadas. Além disso, o KDD pode revelar *insights* sobre a distribuição e a prevalência de

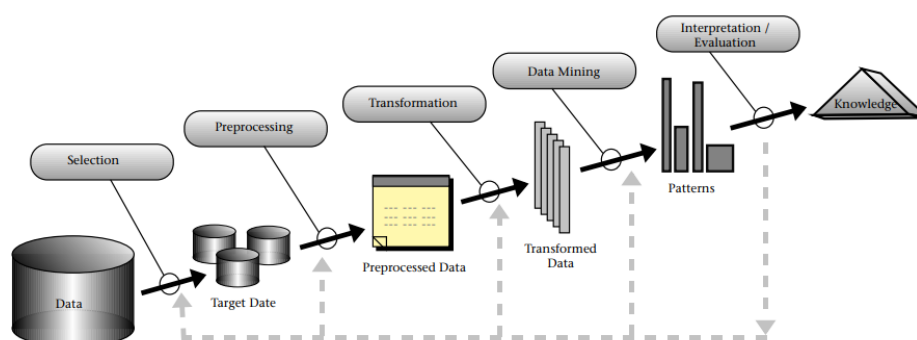


Figura 2.1 – Etapas do processo KDD (Knowledge Discovery in Databases), desde os dados brutos até a geração de conhecimento.

doenças, permitindo intervenções mais eficazes, como a otimização da alocação de vacinas em áreas com maior risco (Malik, 2011).

2.1.4 Visualização de Dados

A visualização de dados é uma técnica fundamental para transformar grandes volumes de informações complexas em representações gráficas que facilitam a interpretação e análise. De acordo com Knaflitz (2018), a visualização de dados não apenas apresenta as informações de forma clara, mas também conta uma história, auxiliando na comunicação eficaz dos dados e no suporte à tomada de decisões estratégicas em diversas áreas, como negócios, saúde e ciência. Por meio de gráficos, mapas, diagramas e outras formas de visualização, é possível identificar padrões, tendências e anomalias que não seriam facilmente perceptíveis em dados brutos (Midway, 2020).

Além de facilitar a compreensão, a visualização de dados desempenha um papel essencial na democratização do acesso à informação, tornando os dados mais acessíveis a públicos com diferentes níveis de especialização técnica (Midway, 2020). Ferramentas como Tableau, Power BI e D3.js são amplamente utilizadas para criar representações gráficas interativas que permitem aos usuários explorar grandes volumes de dados de maneira intuitiva e dinâmica. Esses são exemplos de plataformas que possibilitam a criação de painéis interativos e relatórios visuais, que são essenciais para a análise de dados em tempo real e a tomada de decisões rápidas (Schwabish, 2021).

No contexto de grandes volumes de dados, como no caso do *Big Data*, surgem desafios adicionais relacionados à visualização de dados. Segundo Muniswamaiah, Agerwala e Tappert (2023), a visualização de grandes quantidades de dados requer abordagens especializadas para lidar com a complexidade e o volume massivo de informações. Esses desafios incluem a necessidade de técnicas que possam escalar adequadamente, mantendo a clareza e a precisão das informações representadas. A integração de algoritmos de aprendizado de máquina e inteligência

artificial nas ferramentas de visualização de dados também tem sido uma estratégia para lidar com o crescimento exponencial dos dados (Muniswamaiah; Agerwala; Tappert, 2023).

No setor de saúde pública, a visualização de dados tem um papel crucial. Ela permite que profissionais da saúde e gestores visualizem surtos epidemiológicos, rastreiem a propagação de doenças e identifiquem padrões de saúde populacional. Através da apresentação visual, dados coletados em sistemas de informação de saúde, como o Sistema Único de Saúde (SUS), podem ser utilizados de maneira mais eficiente, gerando *insights* valiosos para intervenções imediatas. Além disso, a visualização clara de dados complexos pode aumentar a transparência e melhorar a comunicação entre diferentes partes interessadas, como gestores, médicos e o público em geral (Skender; Ali, 2022).

Portanto, a visualização de dados desempenha um papel crucial não só na análise e interpretação de grandes volumes de dados, mas também na comunicação desses dados de forma clara, acessível e, acima de tudo, útil para o processo decisório. As boas práticas de visualização, como a escolha apropriada de gráficos, cores e níveis de detalhe, são essenciais para garantir que as visualizações realmente promovam o entendimento e o *insight*, como discutido por Kirk (Kirk, 2016). Ao dominar essas práticas, profissionais podem transformar dados brutos em conhecimento acionável, promovendo intervenções mais eficazes e uma compreensão mais profunda dos fenômenos analisados.

2.2 Vigilância Epidemiológica e SRAG

A vigilância epidemiológica constitui um dos pilares do Sistema Único de Saúde, sendo responsável pela detecção, monitoramento e resposta a agravos de saúde pública. Nesta seção, são apresentados os conceitos fundamentais relacionados à Síndrome Respiratória Aguda Grave, o sistema de informação utilizado para seu registro no Brasil e o contexto epidemiológico do estado de Minas Gerais no período analisado por este trabalho.

2.2.1 Síndrome Respiratória Aguda Grave (SRAG)

A Síndrome Respiratória Aguda Grave (SRAG) é definida pelo Ministério da Saúde como um quadro clínico de síndrome gripal acompanhado de sinais de gravidade, como dispneia, saturação de oxigênio inferior a 95% ou desconforto respiratório (Brasil. Ministério da Saúde. Secretaria de Vigilância em Saúde e Ambiente, 2024). A notificação dos casos de SRAG é compulsória em todo o território nacional, o que significa que todos os serviços de saúde são obrigados a registrar esses casos nos sistemas de vigilância.

A SRAG pode ser causada por diferentes agentes etiológicos. Os principais são os vírus Influenza A e B, responsáveis pela gripe sazonal, e o SARS-CoV-2, agente causador da COVID-19. Além desses, outros vírus respiratórios como o Vírus Sincicial Respiratório (VSR), Parainfluenza e Adenovírus também podem levar a quadros graves. A classificação etiológica

dos casos é registrada no sistema de vigilância por meio de exames laboratoriais, embora uma parcela significativa dos casos permaneça com classificação “não especificada” devido à ausência de coleta de amostras ou resultados inconclusivos.

Do ponto de vista clínico, a notificação de SRAG segue critérios padronizados que incluem febre, tosse ou dor de garganta, acompanhados de pelo menos um dos seguintes sinais: dispneia, saturação de O₂ abaixo de 95%, desconforto respiratório ou óbito por doença respiratória aguda, independentemente de internação ([Brasil. Ministério da Saúde. Secretaria de Vigilância em Saúde e Ambiente, 2024](#)). Essa padronização permite a comparabilidade dos dados ao longo do tempo e entre diferentes unidades federativas.

2.2.2 SIVEP-Gripe: estrutura e variáveis

O Sistema de Informação de Vigilância Epidemiológica da Gripe (SIVEP-Gripe) é o sistema oficial do Ministério da Saúde para o registro dos casos de SRAG hospitalizados no Brasil ([Brasil. Ministério da Saúde, 2024](#)). Criado originalmente para a vigilância sentinela da Influenza, o sistema ganhou atenção a partir de 2009, com a pandemia de H1N1, e teve seu papel ampliado significativamente em 2020 com a pandemia de COVID-19, quando passou a receber um volume expressivo de notificações.

Os dados do SIVEP-Gripe são disponibilizados publicamente pelo Ministério da Saúde através da plataforma OpenDataSUS, em arquivos CSV anuais (denominados INFLUD seguidos do ano, como INFLUD21, INFLUD22, etc.). Cada arquivo contém os registros de notificação de SRAG de todo o território nacional para o respectivo ano epidemiológico. O volume desses arquivos é expressivo: o arquivo de 2021, por exemplo, contém mais de 1,7 milhão de registros nacionais.

As variáveis disponíveis no SIVEP-Gripe abrangem múltiplas dimensões dos casos notificados. A Tabela [2.1](#) apresenta as principais categorias de variáveis utilizadas neste trabalho.

Tabela 2.1 – Principais categorias de variáveis do SIVEP-Gripe utilizadas neste trabalho.

Categoria	Variáveis	Descrição
Identificação	SG_UF_NOT, CO_MUN_NOT	Unidade federativa e código do município de notificação
Demográficas	CS_SEXO, CS_RACA, CS_ESCOL_N, NU_IDADE_N, TP_IDADE	Sexo, raça/cor, escolaridade, idade e tipo de idade
Clínicas	DT_SIN_PRI, EVOLU- CAO, CLASSI_FIN	Data dos primeiros sintomas, desfecho (cura/óbito) e classificação etiológica final
Comorbidades	CARDIOPATI, DIA- BETES, OBESIDADE, NEUROLOGIC, PNEU- MOPATI, entre outras	Presença de condições de risco pré-existent

Um ponto que merece destaque na estrutura do SIVEP-Gripe é o fato de o sistema registrar a localização do caso sob mais de uma perspectiva, com significados distintos. As variáveis SG_UF_NOT e CO_MUN_NOT — esta última também identificada como ID_MUNICIP — referem-se à unidade federativa e ao município “onde está localizada a Unidade que realizou a notificação”, ou seja, ao local do estabelecimento de saúde que atendeu o paciente e registrou o caso. Já as variáveis SG_UF e CO_MUN_RES (ou ID_MN_RESI) correspondem à unidade federativa e ao “município de residência do paciente”. O sistema espelha essa dupla geografia também no nível das regionais de saúde, por meio dos campos CO_REGIONA, para a notificação, e CO_RG_RESI, para a residência ([Brasil. Ministério da Saúde, 2024](#)).

Essa distinção não é meramente formal. Em uma rede de saúde regionalizada, como a de Minas Gerais, casos graves ocorridos em municípios de menor porte são frequentemente encaminhados a hospitais de referência situados em polos regionais. Nessas situações, o município de notificação e o de residência divergem: o caso é contabilizado onde ocorreram a internação e o registro, e não onde o paciente vive. Como consequência, as duas variáveis respondem a perguntas diferentes — o município de notificação descreve onde a demanda assistencial se materializa, ao passo que o município de residência descreve onde reside a população acometida — e a escolha entre elas condiciona a leitura que se pode fazer da distribuição geográfica dos casos.

Um aspecto relevante dos dados do SIVEP-Gripe é a questão da qualidade. Campos como a classificação etiológica final e a evolução do caso apresentam proporções significativas de valores ausentes ou classificados como “ignorado”, o que reflete desafios operacionais na coleta e registro das informações em nível local. Essa característica deve ser considerada na interpretação dos resultados das análises.

2.2.3 Contexto epidemiológico em Minas Gerais (2021–2026)

O estado de Minas Gerais, com seus 853 municípios e mais de 20 milhões de habitantes, é a unidade federativa com o maior número de municípios do Brasil. Essa vasta extensão territorial e a diversidade socioeconômica entre suas regiões resultam em realidades epidemiológicas distintas, o que torna a análise de dados de SRAG no estado particularmente desafiadora e relevante.

O período de 2021 a 2026 abrange transformações significativas no cenário epidemiológico da SRAG em Minas Gerais. O ano de 2021 concentrou o maior volume de notificações de SRAG no estado, refletindo o impacto da circulação do SARS-CoV-2 e suas variantes e o elevado número de hospitalizações associadas à COVID-19 (Ranzani et al., 2021). O início de 2022 foi marcado pela disseminação da variante Ômicron, caracterizada por elevada transmissibilidade, porém associada a menor letalidade em comparação às ondas anteriores, especialmente em razão do avanço da vacinação contra a COVID-19 (Brasil. Ministério da Saúde. Secretaria de Vigilância em Saúde, 2022).

A partir de 2023, observa-se, nos dados processados neste trabalho, uma redução expressiva no número de notificações de SRAG, acompanhada de uma mudança no perfil etiológico: a COVID-19 deixa de ser a causa predominante e os vírus Influenza e outros vírus respiratórios voltam a ganhar relevância no cenário das notificações. Essa transição é particularmente importante para a vigilância, pois demanda atenção contínua mesmo em um contexto pós-pandêmico.

A heterogeneidade territorial do estado também se manifesta na distribuição dos casos. Municípios como Belo Horizonte, Uberlândia e Juiz de Fora, que concentram a maior parte da população urbana do estado (Instituto Brasileiro de Geografia e Estatística, 2022), tendem a registrar maior volume de notificações, enquanto municípios menores podem apresentar dinâmicas epidemiológicas distintas em termos de incidência, letalidade e perfil demográfico dos casos. Essa disparidade reforça a necessidade de ferramentas de análise que permitam tanto uma visão panorâmica do estado quanto o detalhamento por município.

2.3 Trabalhos Relacionados

A análise da literatura existente sobre vigilância epidemiológica da SRAG e o uso de plataformas de visualização de dados em saúde pública revela contribuições relevantes que fundamentam e contextualizam a proposta deste trabalho.

No âmbito da vigilância da SRAG no Brasil, o InfoGripe, desenvolvido pelo Programa de Computação Científica da Fundação Oswaldo Cruz (Fiocruz), é a principal referência. Trata-se de um sistema de monitoramento semanal que utiliza técnicas de *nowcasting* para estimar o número de casos de SRAG em tempo quase real, compensando os atrasos de notificação e digitalização inerentes ao SIVEP-Gripe (Bastos et al., 2019). O InfoGripe opera em nível nacional

e por unidades federativas, apresentando estimativas por faixa etária e por vírus respiratório. No entanto, a plataforma não oferece detalhamento em nível municipal nem permite cruzamentos flexíveis entre variáveis demográficas, clínicas e geográficas, o que limita sua utilidade para análises mais granulares no contexto estadual.

No contexto de Minas Gerais, a Secretaria de Estado de Saúde (SES-MG) mantém um painel de monitoramento de SRAG desenvolvido em Microsoft Power BI ([Secretaria de Estado de Saúde de Minas Gerais, 2024](#)). Essa plataforma apresenta indicadores de SRAG com granularidade municipal para os 853 municípios do estado, constituindo a principal iniciativa oficial de transparência epidemiológica em âmbito estadual. No entanto, a ferramenta apresenta limitações relevantes: instabilidade de disponibilidade, ficando periodicamente fora do ar, e restrições inerentes ao modelo proprietário, que dificultam a auditoria metodológica, a reprodutibilidade das análises e a adaptação por pesquisadores e gestores externos. A proposta desta monografia se diferencia por adotar exclusivamente ferramentas de código aberto — Python, DuckDB e Streamlit —, o que garante auditabilidade, reprodutibilidade e facilidade de adaptação, além de proporcionar desempenho superior no processamento dos grandes volumes de dados do SIVEP-Gripe e maior flexibilidade para cruzamentos entre variáveis demográficas, clínicas e geográficas em nível municipal.

No campo das plataformas de visualização de dados de saúde, [Saldanha et al. \(2016\)](#) propõem a criação de um observatório epidemiológico para o SUS, enfatizando a importância de centralizar e analisar dados epidemiológicos para melhorar a tomada de decisões. O trabalho de [Mendes, Lobato e Jr \(2019\)](#), por sua vez, desenvolve uma ferramenta para o pré-processamento e visualização dos dados do DATASUS, demonstrando como a simplificação do acesso pode facilitar a análise das informações. Ambos os trabalhos oferecem contribuições relevantes, porém operam com dados genéricos do SUS e não abordam especificamente a SRAG.

A proposta desta monografia se distingue dos trabalhos citados em alguns aspectos. Em relação ao InfoGripe, o diferencial está na granularidade municipal e na possibilidade de cruzamentos entre classificação etiológica, variáveis demográficas e comorbidades em nível de município para o estado de Minas Gerais. Por fim, em relação às plataformas genéricas do SUS, o foco específico na SRAG e no SIVEP-Gripe permite uma profundidade de análise que não seria viável com dados de múltiplos sistemas.

3 Metodologia

Este capítulo descreve o caminho seguido para construir o ObservaSRAG-MG, a plataforma proposta nesta monografia. O desenvolvimento foi dividido em três etapas principais, que se conectam diretamente ao processo de KDD apresentado no Capítulo 2: a seleção dos dados de interesse, a preparação dos dados e a sua persistência. A essas três etapas, adiciona-se a criação dos *dashboards* interativos, que é, no final, a interface pela qual os dados tratados chegam ao usuário. Todo o desenvolvimento foi feito em Python, escolhido pela robustez de seu conjunto de bibliotecas para análise de dados e pela facilidade de integrar, em um único fluxo, desde a leitura dos arquivos brutos até a visualização final.

3.1 Seleção dos Dados de Interesse

A fonte de dados utilizada foi o SIVEP-Gripe, disponibilizado publicamente pelo Ministério da Saúde por meio do OpenDataSUS ([Brasil. Ministério da Saúde, 2024](#)). Os dados são publicados em arquivos CSV anuais, nomeados com o prefixo INFLUD seguido do ano de referência (por exemplo, INFLUD21, INFLUD22, e assim por diante) e recebem atualizações periódicas com os novos casos. Foram considerados os arquivos referentes aos anos de 2021 a 2026, totalizando seis bases anuais.

É importante ressaltar a dimensão desses arquivos. Como cada base contém as notificações de SRAG de todo o território nacional, o volume é expressivo: somados, os seis arquivos ocupam cerca de 3,3 GB e reúnem milhões de registros. O arquivo de 2021, isoladamente, possui mais de 1,7 milhão de linhas. Esse volume, por si só, já justifica a necessidade de um tratamento automatizado, uma vez que a abertura direta desses arquivos em ferramentas convencionais de planilha é inviável.

Como o objetivo desta monografia é o recorte estadual, a primeira decisão de seleção foi filtrar apenas os registros notificados em Minas Gerais, identificados pela variável SG_UF_NOT. Após esse filtro, restaram 384.827 casos no período analisado, distribuídos de forma bastante desigual entre os anos, refletindo o cenário epidemiológico descrito no Capítulo 2.

Cabe explicitar que essa escolha recai sobre a geografia de *notificação*, e não sobre a de residência — distinção detalhada no Capítulo 2. O mesmo critério se estende ao recorte municipal: as agregações por município utilizam CO_MUN_NOT, o município da unidade de saúde que notificou o caso. A opção se justifica pelo propósito da plataforma, voltada ao apoio à gestão e à alocação de recursos, pois é no município de notificação que a demanda assistencial efetivamente se materializa, em termos de leitos ocupados e serviços mobilizados. Dela decorrem, contudo, duas consequências que precisam ser consideradas na leitura dos resultados. A primeira

é que o recorte estadual corresponde aos casos *notificados* em Minas Gerais, e não aos casos de *residentes* no estado: um mineiro internado em outra unidade federativa não é contabilizado, ao passo que um residente de outro estado atendido em território mineiro é. A segunda é que municípios que sediam hospitais de referência tendem a acumular notificações originadas de toda a sua região de abrangência, aspecto retomado na análise da distribuição geográfica, no Capítulo 4.

A segunda decisão de seleção envolveu as variáveis. Em vez de carregar todas as dezenas de colunas presentes no SIVEP-Gripe, optou-se por trabalhar apenas com aquelas relevantes para as análises demográficas, epidemiológicas e geográficas pretendidas. As variáveis selecionadas podem ser agrupadas em quatro conjuntos, já apresentados na Tabela 2.1: as de identificação geográfica (SG_UF_NOT e CO_MUN_NOT); as demográficas (CS_SEXO, CS_RACA, CS_ESCOL_N, NU_IDADE_N e TP_IDADE); as clínicas (DT_SIN_PRI, EVOLUCAO e CLASSI_FIN); e as comorbidades, representadas por doze variáveis binárias, como CARDIOPATI, DIABETES, OBESIDADE e NEUROLOGIC, entre outras. Essa seleção criteriosa reduz o consumo de memória durante o processamento e mantém o foco nas informações que efetivamente alimentam o *dashboard*.

3.2 Preparação dos Dados

Os dados do SIVEP-Gripe, em sua forma bruta, não estão prontos para análise. As variáveis são codificadas numericamente, as datas precisam ser interpretadas, a idade é registrada em diferentes unidades e há campos com valores ausentes ou ignorados. Essa etapa, portanto, é o núcleo da metodologia e foi desenvolvida em um *pipeline* de extração, transformação e carga (ETL), implementado no *script* `build_marts.py`.

3.2.1 Leitura e filtragem

Os arquivos CSV são lidos com a biblioteca Pandas (Pandas, 2024), considerando o separador de campos ponto e vírgula e a codificação *latin1*, características do formato distribuído pelo Ministério da Saúde. Para contornar a limitação imposta pelo tamanho dos arquivos, a leitura é feita em blocos (*chunks*) de duzentas mil linhas por vez. Em vez de carregar uma base inteira na memória de uma só vez, o programa processa um bloco, acumula os resultados parciais e descarta o bloco antes de seguir para o próximo. Essa estratégia permite processar arquivos de mais de um gigabyte mesmo em máquinas com recursos modestos. À medida que cada bloco é lido, aplica-se imediatamente o filtro por Minas Gerais, de modo que apenas os registros de interesse seguem para as etapas seguintes.

3.2.2 Limpeza e normalização

Com os registros de interesse em mãos, realiza-se a normalização dos valores. Os códigos de município (CO_MUN_NOT) são padronizados para o formato de seis dígitos do IBGE, o que é

fundamental para o posterior cruzamento com a malha geográfica. As variáveis categóricas, por sua vez, são traduzidas de seus códigos numéricos para rótulos legíveis, seguindo os dicionários oficiais do SIVEP-Gripe. O sexo, por exemplo, passa de M, F e I para “Masculino”, “Feminino” e “Ignorado”; a classificação final é mapeada de modo que o código 1 corresponda a “Influenza”, 4 a “Não especificado” e 5 a “COVID-19”, e assim por diante. O mesmo tratamento é dado à raça/cor, à escolaridade e à evolução do caso.

Duas variáveis exigiram um tratamento mais elaborado. A primeira é a idade. No SIVEP-Gripe, ela é registrada em conjunto com um campo que indica sua unidade (TP_IDADE), que pode ser dias, meses ou anos. O *pipeline* converte todos os valores para uma idade em anos completos e, a partir dela, classifica cada caso em uma faixa etária — 0 a 4, 5 a 19, 20 a 39, 40 a 59, 60 a 79 e 80 anos ou mais —, agrupamento que facilita a leitura dos perfis demográficos. A segunda é a data dos primeiros sintomas (DT_SIN_PRI), a partir da qual se extrai o mês de ocorrência, informação essencial para a construção da curva epidemiológica. Por fim, as doze comorbidades são convertidas em indicadores binários, registrando a presença ou a ausência de cada condição.

Quanto à qualidade dos dados, optou-se por não descartar os registros incompletos, mas por tratar explicitamente as categorias “Ignorado” e “Não especificado”. Essa escolha é coerente com a discussão feita no Capítulo 2: a alta proporção de casos sem classificação etiológica definida é, ela própria, uma característica relevante dos dados, e ocultá-la distorceria o retrato do cenário epidemiológico. As agregações que dependem de um valor válido, contudo, desconsideram os ausentes para não introduzir viés.

3.2.3 Construção do modelo de dados

O *pipeline* organiza os dados tratados em um pequeno modelo dimensional, composto por uma tabela de fatos em nível de caso, uma dimensão de municípios e um conjunto de tabelas agregadas, chamadas de *marts*. Cada *mart* é pré-calculado para responder rapidamente a uma necessidade específica do *dashboard*, evitando que agregações pesadas sejam refeitas a cada interação do usuário. A Tabela 3.1 resume esse modelo.

Tabela	Conteúdo
fato_caso	Tabela em nível de caso (uma linha por notificação), com os valores já traduzidos. Sustenta os filtros dinâmicos por classificação e município.
dim_municipio_mg	Dimensão dos municípios de Minas Gerais, com código IBGE e nome, derivada da malha geográfica.
mart_kpi_ano	Total de registros por ano.
mart_categoria_ano	Contagens por ano e categoria (sexo, raça, escolaridade, classificação, evolução e faixa etária).
mart_mapa_ano_municipio	Casos por município e por ano, base do mapa coroplético.
mart_curva_mensal	Casos por mês, base da curva epidemiológica.
mart_comorbidades_ano	Prevalência de cada comorbidade por ano.

Tabela 3.1 – Tabelas geradas pelo *pipeline* ETL.

A integração com a dimensão geográfica foi realizada como descrito a seguir. A malha dos 853 municípios de Minas Gerais foi obtida em formato GeoJSON e é cruzada com os dados de SRAG por meio do código IBGE de seis dígitos. Esse cruzamento é o que permite colorir cada município no mapa de acordo com o número de casos. Além das tabelas de dados, o *pipeline* ainda registra metadados de proveniência — como o instante de execução, os arquivos de origem e um *hash* que os identifica —, o que torna o processo rastreável e reproduzível.

3.3 Persistência dos Dados

Devido à natureza da aplicação ser analítica e não transacional, o *dashboard* não insere ou atualiza registros individuais, mas executa consultas que agregam grandes quantidades de linhas — somas, contagens e médias por ano, município ou categoria. Esse perfil de carga de trabalho, conhecido como OLAP (*Online Analytical Processing*), é mais bem atendido por soluções orientadas a colunas do que por bancos relacionais orientados a linhas.

Por essa razão, a camada de persistência foi apoiada em duas tecnologias que atuam em níveis distintos e complementares: um formato de armazenamento e um motor de consulta.

O formato de armazenamento é o Apache Parquet ([Apache Software Foundation, 2025](#)), um formato de arquivo colunar e comprimido. Por organizar os dados coluna a coluna, e não linha a linha, o Parquet permite que uma consulta leia apenas as colunas de que necessita, ignorando fisicamente as demais — exatamente o padrão de acesso das agregações do *dashboard*, que envolvem poucas colunas de muitas linhas. Soma-se a isso o efeito da compressão, favorecida pela homogeneidade dos valores vizinhos dentro de uma mesma coluna. O ganho de espaço é expressivo: o conjunto curado em Parquet — que contém apenas os dados de Minas Gerais, já filtrados, com as colunas selecionadas e as tabelas pré-agregadas — ocupa cerca de 2,9 MB, contra os 3,3 GB dos arquivos CSV nacionais lidos como entrada.

O motor de consulta é o DuckDB ([DuckDB, 2025](#)), um banco de dados analítico embar-

cado, frequentemente descrito como o “SQLite da análise de dados”. A relação entre as duas tecnologias é de complementaridade, e não de alternativa: o Parquet define *como os dados são gravados* em disco, ao passo que o DuckDB define *como eles são consultados*, oferecendo uma interface SQL completa sobre o mesmo conjunto curado. Essa combinação é viável porque o DuckDB lê arquivos Parquet de forma nativa, sem etapa de importação — um arquivo pode ser referenciado diretamente em uma cláusula FROM, como se fosse uma tabela do banco. Em outros termos, o Parquet cumpre o papel que, em uma arquitetura de banco de dados tradicional, caberia aos arquivos internos de dados, e o DuckDB cumpre o do processador de consultas, com a diferença de que o formato permanece aberto e legível por outras ferramentas.

Ao final da execução, o *pipeline* materializa tanto os arquivos Parquet quanto um banco DuckDB (`curated.duckdb`, com cerca de 12 MB), este último contendo as mesmas tabelas do modelo dimensional. Os arquivos Parquet constituem a fonte de dados do *dashboard*, ao passo que o banco DuckDB fica disponível como camada de acesso SQL para exploração *ad hoc* e para auditoria por terceiros, que podem reproduzir qualquer indicador exibido na plataforma por meio de uma consulta. Cabe registrar que, no volume de dados aqui tratado, essa escolha não é ditada pelo desempenho: os 384.827 registros de Minas Gerais cabem com folga na memória de uma máquina comum, e o custo de execução das consultas deixa de ser o fator limitante. O valor da camada SQL está na auditabilidade e na escalabilidade — caso o recorte venha a ser ampliado para outras unidades federativas ou para o nível nacional, o DuckDB permite consultar os mesmos arquivos Parquet sem carregá-los integralmente na memória.

Na prática, é essa camada de dados curados que viabiliza um *dashboard* responsivo, em que o usuário troca de ano ou aplica filtros sem esperas perceptíveis.

Por fim, é importante observar questões éticas relativas aos dados. Os dados utilizados são públicos e já disponibilizados de forma anonimizada pelo Ministério da Saúde, não contendo informações que permitam identificar pacientes. Dessa forma, não houve necessidade de aprovação em comitê de ética nem de aplicação de técnicas adicionais de anonimização.

3.4 Construção dos *Dashboards*

A camada de visualização foi construída com a biblioteca Streamlit ([Streamlit, 2024](#)), que permite criar aplicações web interativas diretamente em Python. Ao iniciar, a aplicação carrega os dados a partir dos arquivos Parquet e mantém esse resultado em *cache*, de modo que a leitura ocorra apenas uma vez, e não a cada interação.

A interface foi organizada em torno de indicadores-chave exibidos no topo — como o total de registros e a variação em relação ao ano anterior — e de quatro abas temáticas. A aba de **visão geral** apresenta a evolução anual das notificações e a composição dos casos por classificação etiológica. A aba **demográfica** detalha o perfil dos pacientes por sexo, raça/cor, faixa etária e escolaridade. A aba **epidemiológica** concentra as análises mais específicas: a curva mensal de

casos, a comparação sazonal entre os anos, a taxa de letalidade e a prevalência de comorbidades. A aba de **mapa**, por sua vez, traz a distribuição geográfica dos casos pelos municípios mineiros.

Dois recursos de interatividade devem ser destacados, por atenderem diretamente ao diferencial proposto no Capítulo 1. O primeiro são os filtros por classificação etiológica e por município, que permitem ao usuário estratificar as análises com a granularidade que desejar — de uma visão estadual a um único município, de todos os casos a apenas os de COVID-19, por exemplo. O segundo é o mapa coroplético, gerado com as bibliotecas GeoPandas ([GeoPandas, 2025](#)) e Folium ([Folium, 2025](#)), em que a intensidade da cor de cada município reflete o número de casos. Como esse número varia em ordens de grandeza entre a capital e os pequenos municípios, adotou-se uma escala logarítmica, evitando que poucos municípios de grande volume achatassem visualmente todos os demais. Os gráficos das demais abas foram produzidos com a biblioteca Plotly ([Plotly, 2025](#)), que oferece visualizações interativas com recursos de *zoom*, seleção e exibição de valores ao passar o cursor.

3.5 Síntese: Arquitetura da Aplicação

As seções anteriores trataram cada etapa do desenvolvimento separadamente. Esta seção fecha o capítulo mostrando como elas se encaixam como um sistema único: quais módulos compõem o ObservaSRAG-MG e como eles interagem entre si. Como mostra a Figura 3.1, a aplicação é dividida em dois blocos desacoplados, que só se comunicam por meio de uma camada intermediária de dados curados.

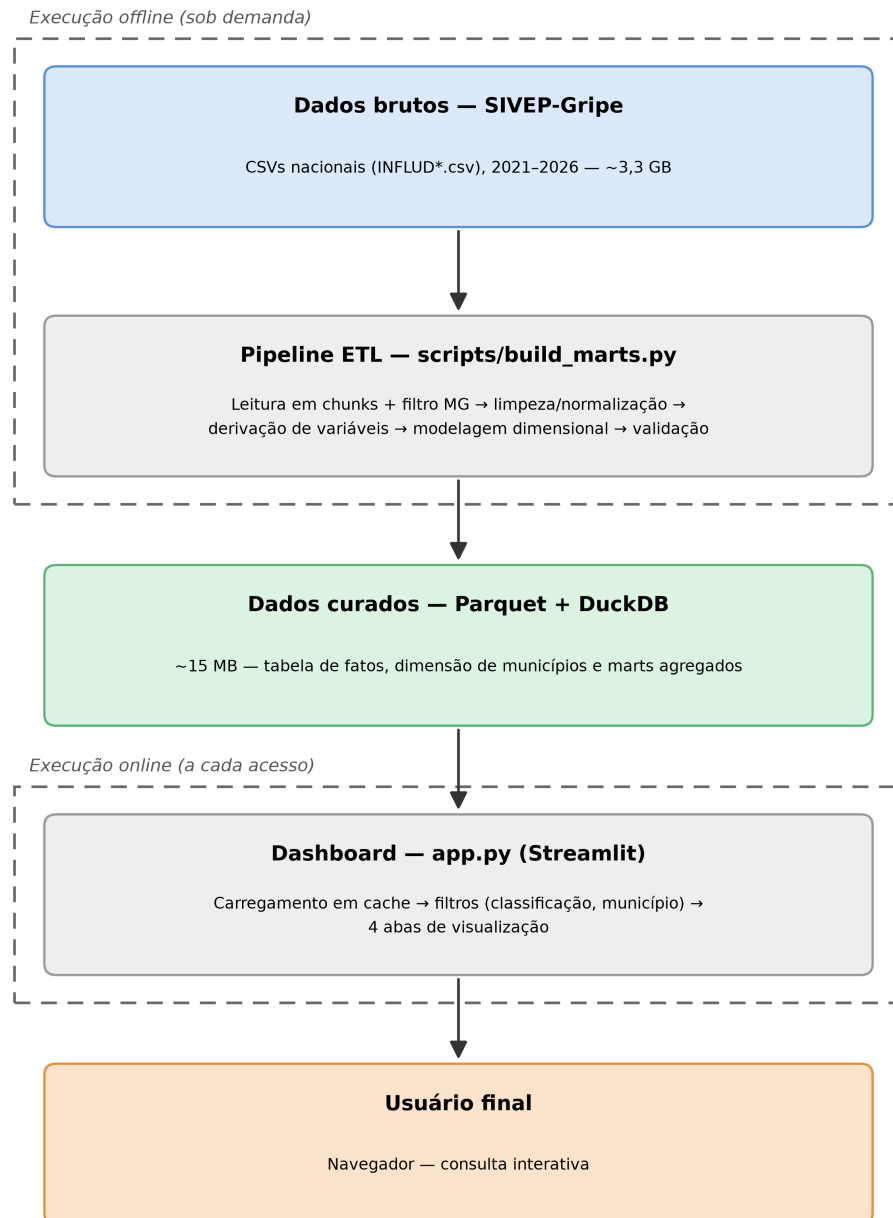


Figura 3.1 – Arquitetura da aplicação: do dado bruto ao *dashboard* interativo.

O primeiro bloco é o *pipeline* ETL (`scripts/build_marts.py`), de execução **offline**: ele é executado manualmente pelo desenvolvedor, sempre que se deseja atualizar a base com uma nova versão publicada pelo SIVEP-Gripe, e nunca de forma automática ou durante o uso do *dashboard*. Internamente, ele encadeia as etapas já descritas neste capítulo — leitura em blocos, filtragem por UF, limpeza e normalização, derivação de variáveis e construção do modelo dimensional — inteiramente sobre estruturas *DataFrame* da biblioteca Pandas, mantidas em memória. Somente ao final dessa cadeia os dados tratados são gravados em disco, nos formatos Parquet e DuckDB apresentados na seção anterior.

O segundo bloco é o *dashboard* (`app.py`), de execução **online**, acionado a cada acesso

do usuário. Diferentemente do *pipeline* ETL, esse módulo nunca lê os arquivos CSV brutos nem executa qualquer limpeza ou normalização: sua única fonte de dados são os arquivos já curados, carregados uma única vez e mantidos em *cache* pelo Streamlit. É essa separação que garante o desempenho relatado no Capítulo 4 — o *dashboard* nunca paga o custo de processar 3,3 GB de dados brutos, porque esse custo já foi pago antes, offline, pelo *pipeline*.

Os dois blocos, portanto, só se encontram na camada de dados curados, ou seja, nos arquivos Parquet e no banco DuckDB gerados ao final do ETL. Essa camada funciona como um contrato entre eles: o *pipeline* pode mudar por dentro — uma nova regra de limpeza, uma nova variável derivada — e o *dashboard* nem precisa saber disso, desde que as tabelas continuem com as mesmas colunas descritas na Tabela 3.1. O mesmo vale ao contrário, para mudanças na interface do *dashboard*.

3.6 Ferramentas

A seguir, resumem-se as principais ferramentas empregadas no desenvolvimento, todas pertencentes ao ecossistema da linguagem Python.

- **Python** — linguagem de programação adotada em todas as etapas, da extração à visualização, escolhida por sua versatilidade e pelo amplo suporte a bibliotecas de análise de dados.
- **Pandas** (Pandas, 2024) — biblioteca central para a manipulação dos dados, responsável pela leitura dos CSVs em blocos, pelas operações de limpeza e normalização e pela construção das tabelas agregadas.
- **Apache Parquet** (Apache Software Foundation, 2025) — formato de armazenamento colunar e comprimido, utilizado para persistir os dados curados de forma compacta e eficiente para leitura.
- **DuckDB** (DuckDB, 2025) — banco de dados analítico embarcado, que disponibiliza uma interface SQL sobre os dados persistidos.
- **Streamlit** (Streamlit, 2024) — *framework* para a construção do *dashboard* interativo diretamente em Python.
- **Plotly** (Plotly, 2025) — biblioteca de gráficos interativos empregada nas análises demográficas e epidemiológicas.
- **GeoPandas** (GeoPandas, 2025) e **Folium** (Folium, 2025) — bibliotecas de geoprocessamento e de geração de mapas, responsáveis pela leitura da malha municipal e pela construção do mapa coroplético de Minas Gerais.

4 Resultados

Neste capítulo são apresentados e discutidos os resultados obtidos com o desenvolvimento do ObservaSRAG-MG. A exposição está organizada de modo a evidenciar o atendimento dos três objetivos específicos definidos no Capítulo 1: inicialmente, discute-se o mapeamento e a qualidade dos dados do SIVEP-Gripe para Minas Gerais; em seguida, apresentam-se os resultados do *pipeline* ETL e o ganho de desempenho obtido com a persistência otimizada; por fim, exploram-se as análises demográficas, epidemiológicas e geográficas viabilizadas pelo *dashboard*, encerrando com uma comparação com a literatura. É importante ressaltar que os dados de 2026 foram extraídos ainda no primeiro semestre daquele ano, esse último ano corresponde a um período incompleto e deve ser interpretado com a devida ressalva sempre que comparado aos demais.

4.1 Mapeamento e qualidade dos dados

O primeiro objetivo específico consistia em mapear e analisar a qualidade dos dados disponíveis no SIVEP-Gripe para Minas Gerais. Após o filtro estadual, foram consolidados 384.827 registros de SRAG no período de 2021 a 2026. A Tabela 4.1 apresenta a distribuição desses casos por ano, contrastando o total nacional bruto de cada arquivo com o volume efetivamente referente a Minas Gerais.

Ano	Registros nacionais (bruto)	Registros em MG
2021	1.745.672	208.464
2022	560.577	69.090
2023	279.453	25.339
2024	267.986	27.049
2025	336.188	45.474
2026*	75.587	9.411
Total	3.265.463	384.827

Tabela 4.1 – Volume de registros de SRAG por ano. * Ano incompleto.

Os números evidenciam tanto a magnitude do desafio computacional — mais de 3,2 milhões de registros nacionais — quanto a relevância do recorte estadual, que isola pouco mais de 384 mil casos de interesse direto para Minas Gerais. A distribuição entre os anos não é homogênea: 2021 concentra mais da metade de todo o período, refletindo o auge da pandemia de COVID-19, ao passo que os anos seguintes apresentam volumes substancialmente menores.

A análise da qualidade dos dados revelou um aspecto que merece atenção e que já havia sido antecipado na revisão bibliográfica, ao se discutir a veracidade como atributo do *Big Data*. Diversos campos apresentam proporções expressivas de valores ausentes ou registrados como

“ignorado”, como sintetiza a Tabela 4.2. O caso mais crítico é o da escolaridade, cuja informação é desconhecida em cerca de dois terços dos registros, mas a própria classificação etiológica final — variável central para a vigilância — permanece como “não especificada” em pouco mais de 42% dos casos.

Variável	Ausência de informação
Escolaridade (ignorado/ausente)	63,9%
Classificação final (não especificado/ausente)	45,3%
Raça/cor (ignorado)	10,6%
Evolução do caso (ignorado/ausente)	5,0%

Tabela 4.2 – Proporção de valores ausentes ou “ignorados” em campos selecionados.

Esse dado não invalida a análise, mas impõe cautela à sua interpretação. A decisão metodológica de preservar e exibir explicitamente as categorias “ignorado” e “não especificado”, em vez de descartá-las, mostra-se importante: ela torna a incompletude visível ao usuário do *dashboard*, em vez de mascará-la sob percentuais aparentemente precisos. Trata-se, em si, de uma contribuição à transparência, já que evidencia ao gestor de saúde onde o registro da informação na ponta precisa ser aprimorado.

4.2 Pipeline ETL e ganho de desempenho

O segundo objetivo específico era o desenvolvimento de um *pipeline ETL* capaz de processar e persistir os dados em formatos otimizados. O *pipeline* implementado, descrito no Capítulo 3, transforma os arquivos brutos no conjunto de tabelas apresentado na Tabela 3.1. O resultado mais imediato dessa transformação é a redução de tamanho: os 3,3 GB dos arquivos CSV nacionais de entrada são convertidos em cerca de 2,9 MB de arquivos Parquet — a fonte efetivamente consumida pelo *dashboard* —, aos quais se acrescentam cerca de 12 MB do banco DuckDB que replica as mesmas tabelas, totalizando aproximadamente 15 MB de dados curados. Essa redução é proveniente de dois fatores combinados: o filtro estadual, que retém apenas os 384.827 registros de Minas Gerais dos mais de 3,2 milhões nacionais, e a persistência em formato colunar comprimido, com seleção das variáveis de interesse e pré-agregação nas tabelas de *mart*. A título de comparação, estimando-se proporcionalmente o volume dos dados de Minas Gerais — que representa 11,8% das linhas nacionais — chega-se a aproximadamente 400 MB de CSV equivalente; isolado o efeito do filtro estadual, o ganho atribuível à combinação de formato colunar, seleção de colunas e pré-agregação é de cerca de 135 vezes sobre os arquivos Parquet, ou de cerca de 27 vezes quando computadas as duas cópias persistidas.

A Tabela 4.3 compara o tempo de carregamento e de geração do mapa entre duas abordagens: a leitura e o processamento dos arquivos brutos a cada execução e a leitura dos dados já curados.

Operação	Dados brutos	Dados curados	Ganho
Carregamento dos dados	98,8 s	0,2 s	99,8%
Geração do mapa coroplético	4,2 s	1,9 s	56,0%

Tabela 4.3 – Comparação de desempenho entre o acesso aos dados brutos e aos dados curados.

O carregamento dos dados, que levava perto de cem segundos a partir dos arquivos brutos, passa a ocorrer em uma fração de segundo. Na prática, é esse ganho que viabiliza a experiência interativa pretendida, na qual o usuário troca de ano ou aplica filtros sem esperas perceptíveis. Confirma-se, assim, a adequação da escolha por uma arquitetura de persistência orientada à análise, baseada em Parquet e DuckDB, em detrimento de um banco relacional tradicional.

4.3 Análises demográficas, epidemiológicas e geográficas

O terceiro objetivo específico — a construção de *dashboards* interativos com análises demográficas, epidemiológicas e geográficas — é o que melhor evidencia o valor da plataforma. As subseções a seguir destacam os principais padrões revelados pelas visualizações.

4.3.1 Evolução temporal e transição etiológica

A análise da classificação etiológica ao longo dos anos revela uma das transições mais significativas do período, apresentada na Tabela 4.4. Em 2021, a COVID-19 era responsável pela maioria dos casos classificados, com 137.903 notificações. Esse predomínio se desfaz progressivamente nos anos seguintes: em 2025, a COVID-19 cai para 1.765 casos, ao passo que os “outros vírus respiratórios” (10.115) e a Influenza (5.006) tornam-se as principais classificações definidas.

Classificação	2021	2022	2023	2024	2025	2026*
COVID-19	137.903	29.046	4.689	3.387	1.765	336
Influenza	553	700	479	1.487	5.006	737
Outro vírus respiratório	1.126	1.982	3.416	4.840	10.115	1.565
Outro agente etiológico	608	258	99	100	391	65
Não especificado	61.581	35.854	16.279	16.576	26.966	5.470

Tabela 4.4 – Casos de SRAG por classificação etiológica e ano. * Ano incompleto.

Esse resultado retrata, em nível estadual, a transição do cenário pandêmico para um cenário pós-pandêmico, em que os vírus respiratórios sazonais voltam a ganhar protagonismo. É

um achado relevante para a vigilância, pois reforça a necessidade de atenção contínua à SRAG mesmo após o arrefecimento da COVID-19.

4.3.2 Sazonalidade e curva epidemiológica

A curva mensal de casos torna visíveis os grandes surtos do período. O pico absoluto ocorre em março de 2021, com 38.661 casos em um único mês, seguido por uma sustentação elevada entre abril e julho daquele ano — comportamento compatível com a circulação do SARS-CoV-2 e com o auge da pandemia em Minas Gerais. Um segundo pico, de menor amplitude, aparece em janeiro de 2022, com 17.556 casos, coincidindo com a chegada da variante Ômicron. A capacidade de visualizar e comparar essas ondas, inclusive sobrepondo a sazonalidade de diferentes anos, é um dos recursos analíticos centrais do *dashboard*.

4.3.3 Letalidade

A taxa de letalidade, calculada como a razão entre os óbitos por SRAG e o total de casos de cada ano, apresenta uma queda consistente ao longo do período, conforme a Tabela 4.5. De um patamar de 24,3% em 2021, a letalidade recua para 4,5% em 2026.

Ano	Casos	Óbitos por SRAG	Letalidade
2021	208.464	50.554	24,3%
2022	69.090	11.188	16,2%
2023	25.339	2.368	9,3%
2024	27.049	2.227	8,2%
2025	45.474	2.860	6,3%
2026*	9.411	423	4,5%

Tabela 4.5 – Letalidade por SRAG por ano em Minas Gerais. * Ano incompleto.

Essa redução acentuada é coerente com o avanço da vacinação contra a COVID-19 e com a circulação de variantes de menor virulência, fatores discutidos no Capítulo 2. A magnitude elevada da letalidade em 2021 é alinhada com a literatura que documentou a gravidade das internações por COVID-19 no início da pandemia no Brasil (Ranzani et al., 2021).

4.3.4 Perfil demográfico

Quanto ao perfil dos pacientes, observa-se uma leve predominância do sexo masculino, que responde por 51,9% dos casos. A distribuição por faixa etária confirma o caráter mais grave da SRAG entre os idosos: a faixa de 60 a 79 anos concentra 30,0% dos casos, e os pacientes com 80 anos ou mais somam outros 16,6%. Chama atenção, contudo, a parcela de 15,9% de casos na faixa de 0 a 4 anos, evidenciando que as crianças de baixa idade constituem um grupo igualmente sensível, o que é relevante para o direcionamento de medidas de prevenção. No que

se refere à raça/cor, a categoria parda é a mais frequente, com 50,3% dos registros, seguida pela branca, com 32,0%.

4.3.5 Comorbidades

A análise das comorbidades reforça o papel das condições de risco pré-existentes. No conjunto dos casos, 58,4% apresentam ao menos uma comorbidade registrada. As mais prevalentes são a cardiopatia, presente em 26,9% dos casos, e o diabetes, em 17,1%. A possibilidade de filtrar essas informações por ano, município e classificação etiológica permite investigações mais finas sobre o perfil de risco da população acometida.

4.3.6 Distribuição geográfica

A dimensão geográfica evidencia a forte concentração das notificações nos grandes centros urbanos do estado. Belo Horizonte, sozinha, responde por 28,1% de todas as notificações, seguida por Uberlândia (5,9%), Betim (3,7%) e Contagem (3,7%). Essa concentração, com poucos municípios acumulando volumes muito superiores aos demais, foi precisamente o que motivou a adoção de uma escala logarítmica no mapa coroplético: sem ela, a predominância da capital tornaria praticamente invisíveis as variações entre os municípios de menor porte. O mapa interativo permite, assim, tanto a visão panorâmica do estado quanto a inspeção individual de cada município, atendendo ao diferencial de granularidade municipal proposto nesta monografia.

Esses percentuais, entretanto, exigem uma leitura cuidadosa. Como discutido nos Capítulos 2 e 3, o mapa é construído sobre o município de notificação, e não sobre o de residência do paciente. Os 28,1% atribuídos a Belo Horizonte não significam, portanto, que 28,1% dos mineiros acometidos por SRAG residam na capital: parcela expressiva desse volume corresponde a pacientes encaminhados de outros municípios para os hospitais de referência ali situados. O mapa deve ser lido, assim, como uma representação da carga assistencial — onde os casos graves foram atendidos e registrados —, e não como uma medida da incidência da doença no território. Uma leitura de risco populacional exigiria, simultaneamente, a geografia de residência e a normalização pela população de cada município.

4.4 A plataforma ObservaSRAG-MG

O conjunto dessas análises é entregue ao usuário por meio de um *dashboard* interativo, organizado nas quatro abas descritas no Capítulo 3 e provido de filtros por classificação etiológica e por município. Para que a plataforma cumpra seu propósito de promover transparência, ela foi disponibilizada publicamente na Internet e pode ser acessada em <<https://observa-srag-mg.streamlit.app/>> (Mendonça, 2026), por qualquer interessado a partir de um navegador, sem necessidade de instalação ou de conhecimentos técnicos. Essa publicação concretiza, na prática, o

objetivo geral desta monografia: tornar acessíveis e interpretáveis dados que, em sua forma bruta, permaneceriam restritos a quem dispusesse de ferramentas e conhecimento para processá-los.

As Figuras 4.1 a 4.4 apresentam capturas de tela das quatro abas da plataforma em execução, ilustrando os recursos de visualização disponibilizados ao usuário. Cada aba expõe um recorte temático dos dados e pode ser filtrada de forma independente por classificação etiológica e por município, permitindo análises que vão de uma visão estadual consolidada até o perfil individual de um único município.

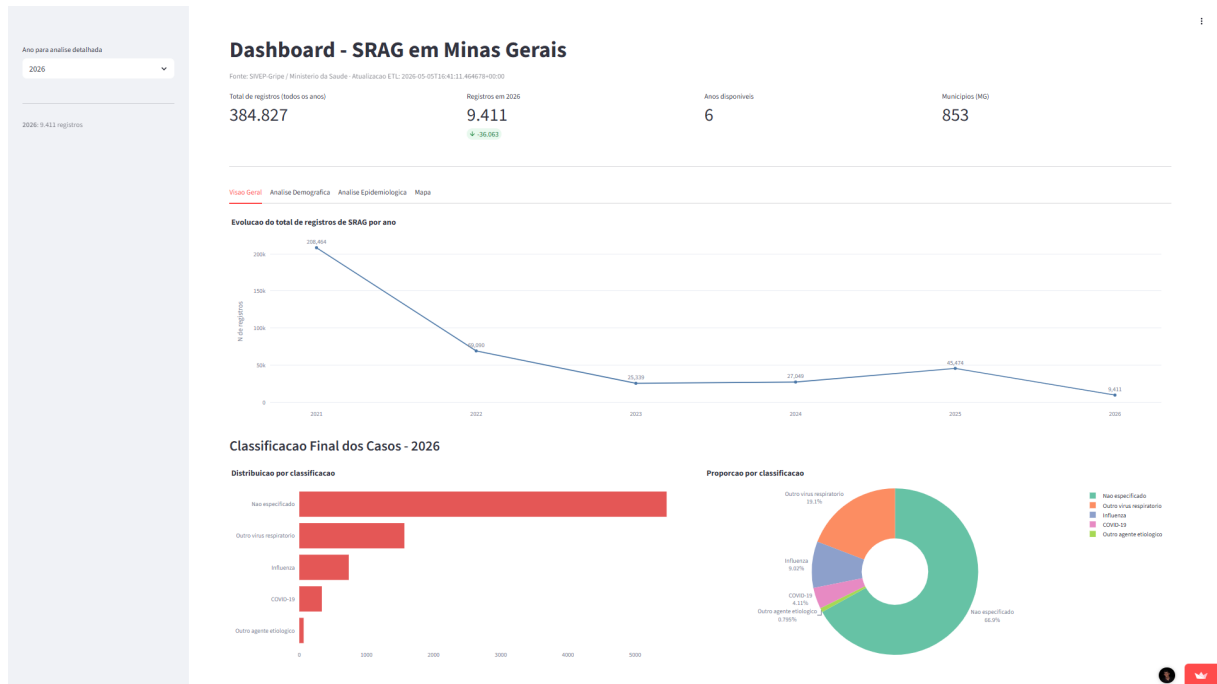


Figura 4.1 – Aba de visão geral do ObservaSRAG-MG: indicadores-chave de desempenho e evolução anual dos casos por classificação etiológica.

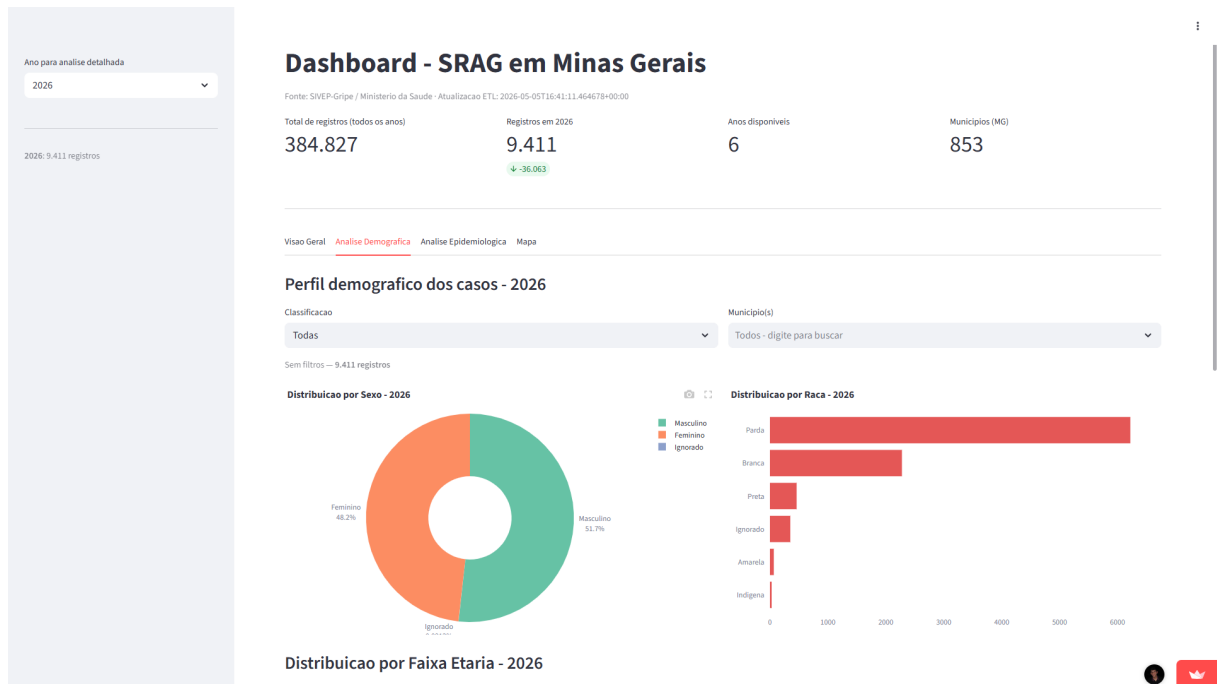


Figura 4.2 – Aba demográfica do ObservaSRAG-MG: perfil dos pacientes por sexo, faixa etária, raça/cor e escolaridade.

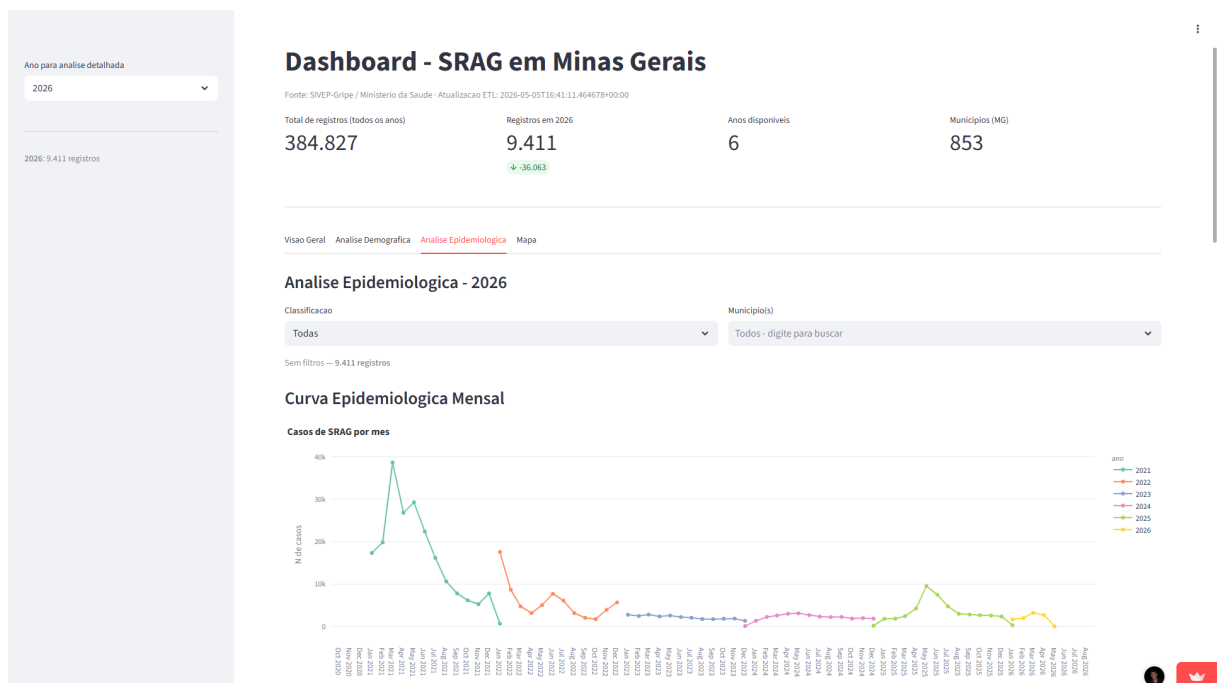


Figura 4.3 – Aba epidemiológica do ObservaSRAG-MG: curva mensal de casos, comparação sazonal entre anos, taxa de letalidade e prevalência de comorbidades.

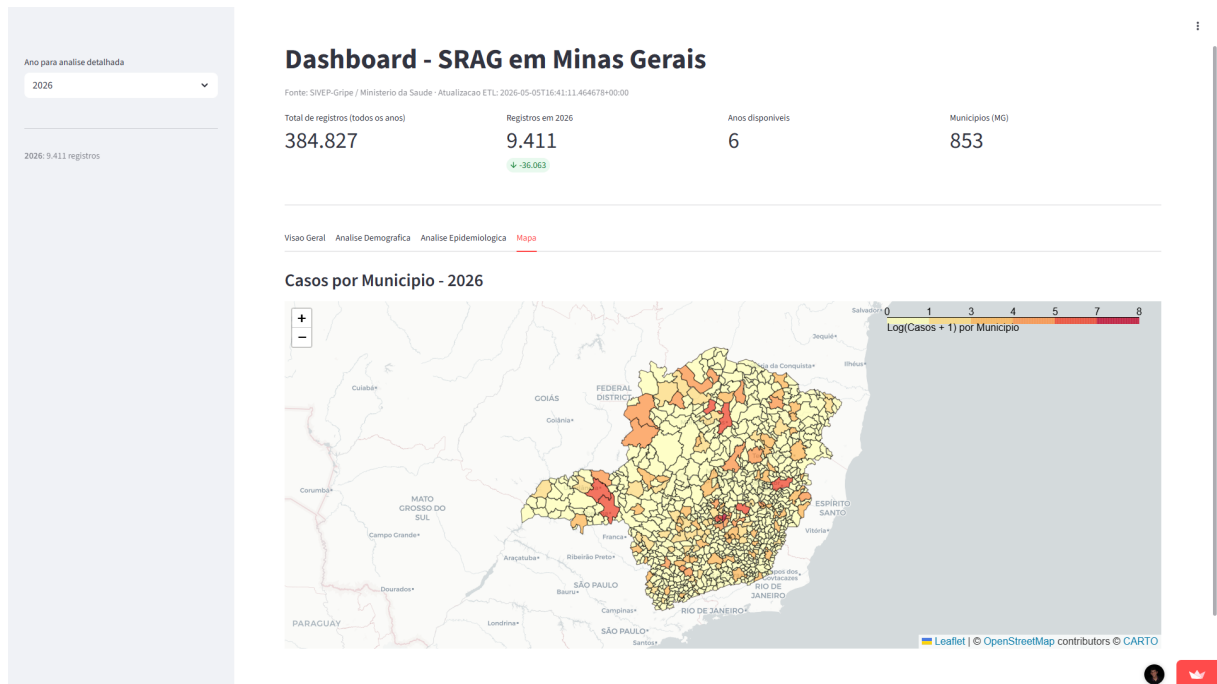


Figura 4.4 – Aba de mapa do ObservaSRAG-MG: distribuição geográfica dos casos de SRAG pelos 853 municípios de Minas Gerais em escala logarítmica de cores.

4.5 Discussão e comparação com a literatura

Os resultados obtidos permitem situar a contribuição desta monografia frente aos trabalhos relacionados apresentados no Capítulo 2. Em relação ao InfoGripe (Bastos et al., 2019), principal referência nacional na vigilância da SRAG, o ObservaSRAG-MG não pretende competir em análise estatística — o InfoGripe emprega técnicas de *nowcasting* para corrigir atrasos de notificação —, mas oferece um diferencial complementar: a granularidade municipal e a possibilidade de cruzar livremente classificação etiológica, variáveis demográficas e comorbidades para o estado de MG, recortes que a plataforma da Fiocruz não disponibiliza.

O trabalho mais próximo desta monografia é o painel da Secretaria de Estado de Saúde de Minas Gerais (Secretaria de Estado de Saúde de Minas Gerais, 2024), desenvolvido em Microsoft Power BI, pois compartilha o mesmo recorte geográfico e a mesma granularidade municipal para a SRAG. No entanto, primeiro, a disponibilidade: o painel da SES-MG apresenta instabilidade recorrente, ficando periodicamente fora do ar, ao passo que o ObservaSRAG-MG, por ser executado a partir de código aberto e de dados publicamente disponíveis, pode ser reproduzido e mantido de forma independente. Segundo, a auditabilidade: a natureza proprietária do Power BI impede que pesquisadores examinem ou reproduzam os cálculos por trás dos indicadores exibidos; o ObservaSRAG-MG, ao contrário, disponibiliza integralmente o *pipeline* ETL e o código-fonte em repositório público (Mendonça, 2025), tornando cada etapa rastreável. Terceiro, o escopo das variáveis demográficas: a aba demográfica do painel da SES-MG contempla

sexo e faixa etária, mas não inclui raça/cor nem escolaridade. O ObservaSRAG-MG incorpora ambas as dimensões, relevantes para a identificação de iniquidades em saúde — valor alinhado ao princípio de equidade do SUS. Embora a escolaridade apresente elevada proporção de valores ausentes nos dados do SIVEP-Gripe (Tabela 4.2), sua exibição explícita na plataforma, com a devida sinalização de incompletude, contribui para evidenciar uma lacuna no próprio registro epidemiológico. Esses três aspectos — disponibilidade, auditabilidade e amplitude das variáveis demográficas — representam os principais diferenciais do ObservaSRAG-MG em relação à iniciativa oficial do estado.

Em relação às plataformas genéricas de dados do SUS (Saldanha et al., 2016; Mendes; Lobato; Jr, 2019), o foco específico na SRAG e no SIVEP-Gripe permitiu uma profundidade de análise — letalidade, sazonalidade, transição etiológica e perfil de comorbidades — que dificilmente seria alcançada por soluções que precisam abarcar múltiplos sistemas de informação.

Em síntese, os resultados demonstram que os três objetivos específicos foram atendidos: os dados foram mapeados e tiveram sua qualidade analisada, um *pipeline* ETL eficiente foi construído e validado em desempenho, e os *dashboards* interativos revelaram padrões epidemiológicos relevantes e coerentes com o cenário descrito na literatura.

5 Considerações Finais

Esta monografia teve como objetivo promover transparência aos dados de Síndrome Respiratória Aguda Grave do estado de Minas Gerais por meio de uma plataforma interativa que combinasse técnicas de análise e visualização de dados, a partir dos registros públicos do SIVEP-Gripe. Ao final do desenvolvimento, é possível afirmar que a questão de pesquisa formulada no Capítulo 1 foi respondida de forma afirmativa: demonstrou-se ser viável construir uma ferramenta acessível que transforma milhões de registros brutos em informações epidemiológicas interpretáveis, com granularidade municipal e capacidade de cruzamento entre variáveis.

Cada um dos três objetivos específicos foi contemplado. O primeiro, de mapear e analisar a qualidade dos dados do SIVEP-Gripe para Minas Gerais, foi atendido com a consolidação de 384.827 registros no período de 2021 a 2026 e com a identificação de aspectos relevantes de incompletude — como a elevada proporção de casos sem classificação etiológica definida ou sem informação de escolaridade —, que passaram a ser exibidos de forma explícita, e não ocultados. O segundo objetivo, de desenvolver um *pipeline* ETL com persistência otimizada, materializou-se em um fluxo capaz de transformar os 3,3 GB de arquivos CSV nacionais em cerca de 2,9 MB de arquivos Parquet para Minas Gerais, acrescidos de 12 MB do banco DuckDB que os replica — resultado do filtro estadual combinado com a persistência em formato colunar comprimido e com a pré-agregação das tabelas de *mart* — e de reduzir o tempo de carregamento dos dados de 98,8 s para 0,2 s, uma redução de 99,8%, viabilizando a interatividade da plataforma. O terceiro objetivo, de construir *dashboards* interativos, resultou em uma aplicação organizada em quatro frentes analíticas — visão geral, perfil demográfico, análise epidemiológica e distribuição geográfica —, dotada de filtros por classificação etiológica e por município.

Para além do cumprimento dos objetivos, as análises produzidas revelaram padrões de valor epidemiológico. A transição do predomínio da COVID-19 para o ressurgimento da Influenza e de outros vírus respiratórios, a queda consistente da taxa de letalidade ao longo dos anos, a sazonalidade marcada pelas grandes ondas de 2021 e 2022 e a forte concentração das notificações nos principais centros urbanos do estado são achados coerentes com o cenário descrito na literatura e que ilustram o potencial da ferramenta como apoio à vigilância. A plataforma torna os dados acessíveis a públicos com diferentes níveis de especialização técnica, o que constitui sua principal contribuição no sentido da transparência.

Cabe, contudo, reconhecer as limitações do trabalho. A análise depende inteiramente da qualidade do preenchimento dos dados na origem, e a alta proporção de campos ignorados impõe cautela à interpretação de certos recortes. Os números são apresentados em valores absolutos, sem normalização pela população de cada município, o que dificulta comparações diretas de risco entre localidades de portes distintos. Além disso, a plataforma reflete os dados tal como

notificados, sem corrigir os atrasos de notificação e digitalização inerentes ao SIVEP-Gripe, e a atualização dos dados ainda depende de execução manual do *pipeline*. Essas limitações delimitam o escopo da proposta e apontam para os próximos passos.

5.1 Trabalhos Futuros

A arquitetura construída neste trabalho é extensível, e diversos caminhos de continuidade se mostram promissores.

- **Normalização por população:** integrar os dados do SIVEP-Gripe às estimativas populacionais do IBGE permitiria calcular taxas de incidência e de mortalidade por cem mil habitantes, tornando as comparações entre municípios de portes diferentes muito mais justas e informativas do que as contagens absolutas atuais.
- **Atualização automatizada:** a ingestão dos dados poderia ser automatizada por meio de rotinas agendadas que baixassem periodicamente os arquivos do OpenDataSUS e reexecutassem o *pipeline*, mantendo a plataforma continuamente atualizada sem intervenção manual.
- **Correção de atrasos de notificação:** a incorporação de técnicas de *nowcasting*, a exemplo do que faz o InfoGripe (Bastos et al., 2019), permitiria estimar o número real de casos recentes, compensando o atraso entre a ocorrência e o registro dos casos.
- **Modelagem preditiva:** com base na série histórica consolidada, modelos de aprendizado de máquina poderiam ser explorados para a previsão de surtos e para a antecipação de demandas sobre o sistema de saúde, agregando uma camada prospectiva às análises hoje descritivas.
- **Integração de novas fontes:** dados complementares, como cobertura vacinal, ocupação de leitos de UTI e a divisão por regiões e macrorregiões de saúde, enriqueceriam as análises e permitiriam correlações adicionais, como a relação entre vacinação e letalidade.
- **Generalização para outras unidades federativas:** como o *pipeline* é parametrizado por UF, a mesma solução poderia ser estendida a outros estados ou ao nível nacional, com adaptações pontuais, servindo de modelo para iniciativas semelhantes.
- **Validação com gestores de saúde:** por fim, um estudo de usabilidade com profissionais e gestores de saúde permitiria avaliar a efetividade da plataforma no apoio à tomada de decisão e orientar melhorias na interface e nos indicadores apresentados.

Em suma, o ObservaSRAG-MG cumpre o objetivo geral proposto e estabelece uma base sólida sobre a qual essas extensões podem ser construídas, reforçando a contribuição do trabalho tanto como produto quanto como ponto de partida para investigações futuras.

Referências

Apache Software Foundation. *Apache Parquet Documentation*. 2025. Acessado em: maio 2026. Disponível em: <<https://parquet.apache.org/docs/>>.

BASTOS, L. S.; ECONOMOU, T.; GOMES, M. F. C.; VILLELA, D. A. M.; COELHO, F. C.; CRUZ, O. G.; STONER, O.; BAILEY, T.; CODEÇO, C. T. A modelling approach for correcting reporting delays in disease surveillance data. *Statistics in Medicine*, Wiley, v. 38, n. 22, p. 4363–4377, 2019. Base metodológica do sistema InfoGripe/Fiocruz.

Brasil. *Lei nº 8.080, de 19 de setembro de 1990*. 1990. Dispõe sobre as condições para a promoção, proteção e recuperação da saúde, a organização e o funcionamento dos serviços correspondentes. Diário Oficial da União, Brasília, DF, 20 set. 1990. Disponível em: <https://www.planalto.gov.br/ccivil_03/leis/18080.htm>.

Brasil. Ministério da Saúde. Secretaria de Vigilância em Saúde. *Boletim Epidemiológico Especial – COVID-19: monitoramento das variantes do SARS-CoV-2 no Brasil*. 2022. Secretaria de Vigilância em Saúde, Ministério da Saúde. Acesso em: jun. 2026. Disponível em: <<https://www.gov.br/saude/pt-br/centrais-de-conteudo/publicacoes/boletins/epidemiologicos/covid-19>>.

Brasil. Ministério da Saúde. *SIVEP-Gripe – Sistema de Informação de Vigilância Epidemiológica da Gripe*. 2024. Dados disponibilizados via OpenDataSUS. Acesso em: maio 2026. Disponível em: <<https://opendatasus.saude.gov.br/dataset/srag-2021-a-2026>>.

Brasil. Ministério da Saúde. Secretaria de Vigilância em Saúde e Ambiente. *Guia de Vigilância em Saúde*. 6. ed. Brasília: Ministério da Saúde, 2024. Acesso em: maio 2026. Disponível em: <https://bvsms.saude.gov.br/bvs/publicacoes/guia_vigilancia_saude_6ed.pdf>.

CARVALHO, M. de. *Construindo o saber: técnicas de metodologia científica*. [S.l.]: Papirus Editora, 1989. ISBN 9788530800710.

DEBATTISTA, J.; LANGE, C.; SCERRI, S.; AUER, S. Linked’big’data: towards a manifold increase in big data value and veracity. In: IEEE. *2015 IEEE/ACM 2nd International Symposium on Big Data Computing (BDC)*. [S.l.], 2015. p. 92–98.

DuckDB. *DuckDB Documentation*. 2025. Acessado em: maio 2026. Disponível em: <<https://duckdb.org/docs/>>.

FAYYAD, U. M.; PIATETSKY-SHAPIO, G.; SMYTH, P. From data mining to knowledge discovery in databases. *AI Review*, Springer, v. 11, n. 3, p. 213–230, 1996.

Folium. *Folium Documentation*. 2025. Acessado em: maio 2026. Disponível em: <<https://python-visualization.github.io/folium/>>.

GALETSI, P.; KATSALIAKI, K.; KUMAR, S. Big data analytics in health sector: Theoretical framework, techniques and prospects. *International Journal of Information Management*, Elsevier, v. 50, p. 206–216, 2020.

GeoPandas. *GeoPandas Documentation*. 2025. Acessado em: maio 2026. Disponível em: <<https://geopandas.org/en/stable/docs.html>>.

Instituto Brasileiro de Geografia e Estatística. *Minas Gerais: Panorama*. 2022. Acesso em: 17 jun. 2026. Disponível em: <<https://cidades.ibge.gov.br/brasil/mg/panorama>>.

JONEK-KOWALSKA, I. Health care in cities perceived as smart in the context of population aging—a record from poland. *Smart Cities*, MDPI, v. 5, n. 4, p. 1267–1292, 2022.

KIRK, A. *Data Visualization: A Handbook for Data Driven Design*. London: SAGE Publications, 2016.

KNAFLIC, C. N. Storytelling com dados. *Um Guia Sobre Visualização de Dados Para Profissionais de*, 2018.

KOSTKOVA, P.; BREWER, H.; LUSIGNAN, S. D.; FOTTRELL, E.; GOLDACRE, B.; HART, G.; KOCZAN, P.; KNIGHT, P.; MARSOLIER, C.; MCKENDRY, R. A. et al. Who owns the data? open data for healthcare. *Frontiers in public health*, Frontiers Media SA, v. 4, p. 7, 2016.

LIMA, A. C.; JANUÁRIO, M. C.; LIMA, P. T.; MOURA, W. de et al. Datasus: o uso dos sistemas de informação na saúde pública. *REFAS: Revista FATEC Zona Sul*, FATEC-Faculdade de Tecnología da Zona Sul, v. 1, n. 3, p. 4, 2015.

MALIK, K. Healthcare: Current applications and issues. *Transactions on Latest Trends in Health Sector*, v. 3, n. 3, 2011.

MARR, B. *Big Data: Using Smart Big Data, Analytics and Metrics to Make Better Decisions and Improve Performance*. Chichester, UK: Wiley, 2015.

MARR, B. *Data strategy: How to profit from a world of big data, analytics and the internet of things*. [S.l.]: Kogan Page Publishers, 2017.

MATHEUS, R.; JANSSEN, M.; MAHESHWARI, D. Data science empowering the public: Data-driven dashboards for transparent and accountable decision-making in smart cities. *Government Information Quarterly*, Elsevier, v. 37, n. 3, p. 101284, 2020.

MENDES, D.; LOBATO, F.; JR, A. J. Ferramenta de pre-processamento e visualização de dados do ~datasus. *Anais do VII Workshop de Transparência em Sistemas*, 2019.

MENDONÇA, R. A. *ObservaSRAG-MG: Repositório de código-fonte*. 2025. Acesso em: 7 jul. 2026. Disponível em: <<https://github.com/rui-augusto/my-dashboard-srag>>.

MENDONÇA, R. A. *ObservaSRAG-MG: Plataforma interativa de vigilância dos dados de SRAG em Minas Gerais*. 2026. Acesso em: 7 jul. 2026. Disponível em: <<https://observa-srag-mg.streamlit.app/>>.

MIDWAY, S. R. Principles of effective data visualization. *Patterns*, Elsevier, v. 1, n. 9, 2020.

Ministério da Saúde. *Definição de caso de Síndrome Respiratória Aguda Grave*. 2026. Acesso em: 2026. Disponível em: <<https://www.gov.br/saude/pt-br/composicao/svsa/cnie/srag>>.

Ministério da Saúde. *Definição de SARS-CoV-2 - COVID-19*. 2026. Acesso em: 2026. Disponível em: <<https://www.gov.br/saude/pt-br/assuntos/saude-de-a-a-z/c/covid-19>>.

MUNISWAMAIAH, M.; AGERWALA, T.; TAPPERT, C. C. Big data and data visualization challenges. In: IEEE. *2023 IEEE International Conference on Big Data (BigData)*. [S.l.], 2023. p. 6227–6229.

NORONHA, J. C. d.; LIMA, L. D. d.; MACHADO, C. V. Sistema único de saúde-sus. In: *Políticas e sistemas de saúde no Brasil*. [S.l.]: Fiocruz, 2008. p. 435–472.

PANDAS. *Documentação do Pandas*. 2024. Acessado em: 07 out. 2024. Disponível em: <https://pandas.pydata.org/docs/>.

Plotly. *Plotly Python Open Source Graphing Library*. 2025. Acessado em: maio 2026. Disponível em: <https://plotly.com/python/>.

RAMPAZZO, L. *Metodologia científica*. [S.l.]: Edições Loyola, 2005. ISBN 9788515024988.

RANZANI, O. T.; BASTOS, L. S. L.; GELLI, J. G. M.; MARCHESI, J. F.; BAIÃO, F.; HAMACHER, S.; BOZZA, F. A. Characterisation of the first 250 000 hospital admissions for COVID-19 in Brazil: a retrospective analysis of nationwide data. *The Lancet Respiratory Medicine*, Elsevier, v. 9, n. 4, p. 407–418, 2021.

SALDANHA, R. d. F.; BASTOS, R. R.; BUSTAMANTE-TEIXEIRA, M. T.; LEITE, I. C. G.; CAMPOS, E. M. S. Proposta de um observatório epidemiológico do sistema Único de saúde. *Cadernos de Saúde Pública*, 2016.

SCHWABISH, J. *Better data visualizations: A guide for scholars, researchers, and wonks*. [S.l.]: Columbia University Press, 2021.

Secretaria de Estado de Saúde de Minas Gerais. *Painel de Monitoramento da Síndrome Respiratória Aguda Grave – Minas Gerais*. 2024. Painel desenvolvido em Microsoft Power BI pela SES-MG. Acesso em: jun. 2026. Disponível em: <https://app.powerbi.com/view?r=eyJrIjoiMzljNjllYzMtNTc1ZS00NzBILWE1YjEtZDEzM2E4MzJjOWZjIiwidCI6Ijg3ZTRkYTJiLTgyZGYt>

SKENDER, F.; ALI, I. Big data in health and the importance of data visualization tools. *Aklli Sistemler ve Uygulamaları Dergisi*, Islyra Medical and Information Technologies, v. 5, n. 1, p. 33–37, 2022.

STREAMLIT. *Documentação do Streamlit*. 2024. Acessado em: 07 out. 2024. Disponível em: <https://docs.streamlit.io/>.

TOLE, A. A. Big data challenges. *Database systems journal*, v. 4, n. 3, 2013.

TSAI, C.-W.; LAI, C.-F.; CHAO, H.-C.; VASILAKOS, A. V. Big data analytics: a survey. *Journal of Big data*, Springer, v. 2, p. 1–32, 2015.

WAZID, M.; DAS, A. K.; MOHD, N.; PARK, Y. Healthcare 5.0 security framework: applications, issues and future research directions. *IEEE Access*, IEEE, v. 10, p. 129429–129442, 2022.

WU, X.; ZHU, X.; WU, G.-Q.; DING, W. Data mining with big data. *IEEE transactions on knowledge and data engineering*, IEEE, v. 26, n. 1, p. 97–107, 2013.

ZILLNER, S.; NEURURER, S. Big data in the health sector. *New horizons for a data-driven economy: a roadmap for usage and exploitation of big data in Europe*, Springer International Publishing, p. 179–194, 2016.

Declaração de Uso de Inteligência Artificial Generativa

Durante a preparação deste documento, eu, Rui Augusto Drumond Mendonça, estudante de graduação em Ciência da Computação da Universidade Federal de Ouro Preto, declaro o uso de IAGen «Claude Opus 4.5 e Claude Sonnet 4.6» para revisão textual e geração de códigos LaTeX.

Após o uso desta ferramenta/modelo/serviço, eu revisei e editei o conteúdo em conformidade com os princípios éticos para uso de IAGen (Resolução CONPEP nº 144) e com os acordos estabelecidos com o(a) orientador(a) desta pesquisa. Dessa forma, assumo total responsabilidade pelo conteúdo da publicação.