UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE INFORMÁTICA
GRADUAÇÃO EM SISTEMAS DE INFORMAÇÃO

PEDRO HENRIQUE SANTIAGO DE LUNA

PIPELINE DE DADOS PARA ANÁLISE EPIDEMIOLÓGICA DE CASOS SOBRE TRANSTORNOS MENTAIS RELACIONADOS AO TRABALHO NO BRASIL

Recife,
2025

CURSO DE BACHARELADO EM SISTEMAS DE INFORMAÇÃO

PIPELINE DE DADOS PARA ANÁLISE EPIDEMIOLÓGICA DE CASOS SOBRE TRANSTORNOS MENTAIS RELACIONADOS AO TRABALHO NO BRASIL

Trabalho apresentado ao programa de Graduação em Sistemas de Informação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para a obtenção do grau de Bacharel em Sistemas de Informação.

Orientador(a): Adiel Teixeira de Almeida Filho

Aprovado em: 06/08/2025

Recife

2025

Ficha de identificação da obra elaborada pelo autor,
através do programa de geração automática do SIB/UFPE

Luna, Pedro Henrique Santiago de.
Pipeline de dados para análise epidemiológicas de casos sobre transtornos mentais relacionados ao trabalho no Brasil / Pedro Henrique Santiago de Luna. - Recife, 2025.
40 p. : il., tab.

Orientador(a): Adiel Teixeira de Almeida Filho
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal de Pernambuco, Centro de Informática, Sistemas de Informação - Bacharelado, 2025.
Inclui referências.

1. Python. 2. Pyspark. 3. Power BI. 4. ETL. 5. Pandas. I. Almeida Filho, Adiel Teixeira de. (Orientação). II. Título.

000 CDD (22.ed.)

PIPELINE DE DADOS PARA ANÁLISE EPIDEMIOLÓGICA DE CASOS SOBRE TRANSTORNOS MENTAIS RELACIONADOS AO TRABALHO NO BRASIL

Trabalho apresentado ao programa de Graduação em Sistemas de Informação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para a obtenção do grau de Bacharel em Sistemas de Informação.

Aprovado em: 06/08/2025

BANCA EXAMINADORA

_____

Prof. Adiel Teixeira de Almeida Filho (Orientador)

Universidade Federal de Pernambuco

_____

Prof. Jamilson Ramalho Dantas (Examinador Interno)

Universidade Federal de Pernambuco

RESUMO

Este trabalho tem como objetivo desenvolver um pipeline completo de dados para a análise de casos de Transtornos Mentais Relacionados ao Trabalho (TMRT) no Brasil, utilizando a base do Departamento de Informática do Sistema Único de Saúde (DATASUS). Foi implementado um processo de ETL em Python, com o uso de PySpark e Pandas, para tratar dados do período de 2006 a 2023. Os indicadores epidemiológicos e as tendências temporais foram apresentados por meio de um *dashboard* interativo no Power BI, facilitando a visualização e a interpretação das informações. A solução permite ao usuário realizar análises como a evolução anual dos casos, a distribuição geográfica por municípios, o perfil demográfico dos trabalhadores afetados, os tipos de ocupação mais recorrentes, os desfechos registrados e os fatores associados. As visualizações atendem a tarefas como monitoramento temporal, comparação entre regiões, identificação de grupos vulneráveis e apoio à formulação de políticas públicas voltadas à saúde mental no trabalho.

Palavras-chave: Python, Pyspark, Pandas, Power BI, ETL, TMRT, DATASUS. Sistema Único de Saúde, Transtorno mental relacionados ao trabalho.

ABSTRACT

This study aims to develop a complete data pipeline for analyzing cases of Work-Related Mental Disorders (WRMD) in Brazil, using data from the Department of Informatics of the Unified Health System (DATASUS). An ETL process was implemented in Python, using PySpark and Pandas, to process data from the period between 2006 and 2023. Epidemiological indicators and temporal trends were presented through an interactive dashboard built in Microsoft Power BI, facilitating data visualization and interpretation. The solution enables users to perform analyses such as yearly trends in case notifications, geographic distribution by municipality, demographic profiles of affected workers, most frequent occupations, reported outcomes, and associated factors. The visualizations support tasks such as temporal monitoring, regional comparisons, identification of vulnerable groups, and guidance for public policy planning focused on mental health in the workplace.

Keywords: PySpark, Pandas, Power BI, ETL, TMRT, DATASUS, Unified Health System, Work-Related Mental Disorders.

LISTA DE FIGURAS

LISTA DE QUADROS

SUMÁRIO

1. INTRODUÇÃO

1.1. Motivação e justificativa

Os Transtornos Mentais Relacionados ao Trabalho (TMRT) há alguns anos vêm se tornando uma preocupação crescente na saúde ocupacional e na gestão de recursos humanos e são fatores como a intensificação das demandas profissionais, aumento da carga horária e precarização das condições de trabalho que têm contribuído para esse aumento significativo de casos, incluindo ansiedade, depressão e síndrome de *burnout* [1]. Esses transtornos não afetam apenas a produtividade individual, mas também a eficiência de vários setores de uma organização, resultando em afastamentos prolongados e custos elevados para empresas e sistemas de saúde [2].

A Organização Mundial da Saúde (OMS) estima que, globalmente, aproximadamente 280 milhões de pessoas sofrem de depressão, e uma parcela significativa desses casos está relacionada ao ambiente de trabalho [3]. Ratificando isso, um levantamento da *International Labour Organization* (ILO) aponta que cerca de 12 bilhões de dias de trabalho são perdidos anualmente devido a problemas de saúde mental, resultando em um impacto econômico de aproximadamente 1 trilhão de dólares por ano em perda de produtividade [4].

No contexto brasileiro, entre 2007 e 2017, os transtornos mentais e comportamentais foram identificados como a terceira principal causa de afastamento do trabalho e, durante esse período, as licenças médicas representaram 52% de todos os benefícios sociais concedidos, sendo que cerca de 50% desses pedidos estavam relacionados a transtornos do humor, como depressão e ansiedade, evidenciando a relevância do impacto da saúde mental no ambiente ocupacional [5]. Esse cenário reforça a necessidade de existirem medidas preventivas e políticas públicas voltadas para a promoção do bem-estar psicológico dos trabalhadores.

Diante disso, o monitoramento dessa problemática e a análise de dados são essenciais para compreender a dinâmica dos TMRT ao longo do tempo. No Brasil, os dados sobre esses transtornos são registrados pelo Sistema de Informação de Agravos de Notificação (SINAN), acessíveis através do DATASUS que permite uma consulta detalhada sobre a distribuição e evolução dos casos ao longo dos anos de 2006 a 2023. Diante da crescente que poderá ser observada ao longo do trabalho, a

análise desses dados é de extrema importância para a formulação de políticas públicas eficazes e intervenções preventivas no ambiente de trabalho [6].

Com o aumento do volume de informações disponíveis ao passar dos anos, é necessário utilizar soluções tecnológicas avançadas para processar e visualizar esses dados de forma eficiente. PySpark e Pandas são algumas das ferramentas que têm se destacado por sua capacidade de lidar com grandes volumes de dados, enquanto plataformas de Business Intelligence, como Power BI, permitem a criação de *dashboards* interativos que facilitam a tomada de decisões baseadas em evidências [7].

1.2. Objetivos

1.2.1. Objetivo Geral

Desenvolver um pipeline completo de análise de dados sobre os casos de TMRT no Brasil, utilizando a base oficial do DATASUS e a biblioteca Pysus para extração dos dados.

1.2.2. Objetivos Específicos

- Implementar um pipeline de ETL utilizando PySpark e Pandas para a manipulação e o processamento de grandes volumes de dados;
- Construir *dashboard* interativo na plataforma Power BI para visualização e interpretação dos indicadores gerados;
- Identificar padrões temporais e regionais nos casos de TMRT registrados no SINAN;
- Fornecer subsídios para o desenvolvimento de estratégias preventivas e políticas públicas voltadas à saúde mental dos trabalhadores.

1.3. Estrutura do trabalho

Este trabalho está organizado em seis seções, conforme descrito a seguir:

- **Introdução:** Apresenta o tema, a justificativa, os objetivos e a organização do trabalho.
- **Base Conceitual**: Aborda os principais conceitos teóricos relacionados a Transtornos Mentais Relacionados ao Trabalho, análise de dados e ferramentas utilizadas.
- **Desenvolvimento e Metodologia**: Detalha o processo de construção do pipeline de dados, as ferramentas aplicadas (PySpark, Pandas, Power BI) e as etapas do processo de ETL.

- **Resultados**: Apresenta as análises realizadas, os indicadores gerados e os principais achados extraídos do *dashboard*.
- **Conclusão**: Resume os resultados alcançados, discute as limitações e propõe possíveis caminhos para trabalhos futuros.
- **Referências**: Lista todas as fontes bibliográficas utilizadas ao longo do desenvolvimento do trabalho, conforme as normas da ABNT.

2. BASE CONCEITUAL

A seguir, são descritos os conceitos essenciais para o desenvolvimento do estudo em questão. Cada tópico busca evidenciar sua relevância dentro da proposta de análise de dados epidemiológicos aplicada aos Transtornos Mentais Relacionados ao Trabalho no Brasil.

2.1. Bibliotecas Python

Python é uma linguagem de programação interpretada, de alto nível, bastante utilizada em diversas áreas como desenvolvimento web, automação, ciência de dados, inteligência artificial e aprendizado de máquina. Essa linguagem, foi criada por Guido van Rossum no final da década de 1980 e ela se destaca, principalmente, por sua sintaxe simples e legível, tornando-a um ponto de partida para muitas pessoas que se interessam por programação, visto que a escrita e manutenção de código é bem mais fácil de ser entendida por iniciantes. Outro ponto extremamente positivo do Python é sua vasta biblioteca padrão e ecossistema de pacotes que tornam a linguagem extremamente versátil para análise de dados, por exemplo-principal uso da linguagem ao longo do trabalho [8].

Diante disso, a popularidade do Python na análise de dados deve-se, justamente, à existência de bibliotecas robustas como Pandas, NumPy e SciPy, que permitem uma manipulação eficiente de grandes volumes de dados e informações. Além disso, a integração com frameworks como PySpark amplia mais ainda seu uso no processamento distribuído, possibilitando o trabalho com big data de maneira escalável e eficiente [9]. Para esta análise, serão utilizadas as bibliotecas de Pandas, Pyspark e Pysus.

O "Pandas" é uma das bibliotecas mais utilizadas para manipulação e análise de dados, sendo criada com o objetivo de fornecer estruturas de dados flexíveis e eficientes que permitissem operações avançadas em tabelas e/ou séries temporais, sendo essencial para aplicações e desenvolvimento na engenharia e ciência de dados. A principal estrutura de dados da biblioteca é o Data Frame que possibilita a organização dos dados em formato tabular, como uma planilha do Excel ou tabelas SQL. Dentro da biblioteca existem diversas funcionalidades para serem utilizadas em cima dessas estruturas, como por exemplo filtragem de dados, agregação, junção e transformação do tipo do dado, facilitando a análise exploratória e construção de indicadores [10].

Outra biblioteca é a “PySpark” que é uma interface do Apache Spark para Python, permitindo a manipulação de grandes volumes de dados em ambientes distribuídos. Ele combina a facilidade e flexibilidade do Python com a robustez que o Spark possui. O Spark utiliza processamento em memória para aumentar a velocidade e eficiência das operações analíticas, mantendo os os dados armazenados sempre que possível, reduzindo significativamente o tempo de

execução de tarefas mais complexas. A principal vantagem do PySpark é sua capacidade de escalar para petabytes de dados, o que o torna essencial para aplicações que envolvem *Big Data*. Além disso, a sua compatibilidade com bibliotecas tipo Pandas e SQL permite uma integração fácil para análises avançadas, otimizando os fluxos de trabalho [11].

Por fim, o “Pysus” é uma biblioteca voltada para a manipulação e análise dos dados disponibilizados no Datasus pelo Ministério da Saúde. Essa biblioteca permite a extração e conversão de arquivos em formatos específicos, como DBC, facilitando a integração desses dados em pipelines analíticos, como é o caso deste trabalho. A biblioteca oferece funcionalidades para baixar os arquivos das diversas fontes, agregação e visualização de dados através do Pandas/Spark, tornando-se uma ferramenta essencial para a consulta das informações estruturadas para análises epidemiológicas [12].

2.2. **ETL (*Extract, Transform, Load*)**

O processo de ETL (*Extract, Transform, Load*) é fundamental na engenharia de dados, sendo responsável pela extração, transformação e carga de dados para um repositório final, como um banco de dados ou até mesmo um arquivo Excel. Esse processo permite que grandes volumes de dados sejam consolidados, limpos e otimizados para serem analisados e utilizados como informação para a tomada de decisão, por exemplo [13]. Atualmente, esse processo é diretamente impactado pelo avanço das tecnologias de *Big Data* e ferramentas como Apache Spark e Apache Hadoop são bastante utilizadas durante o processo para lidar em tempo real com esse grande volume de informação [11]. A Figura 1, ilustra o framework do processo durante essa integração dos dados.

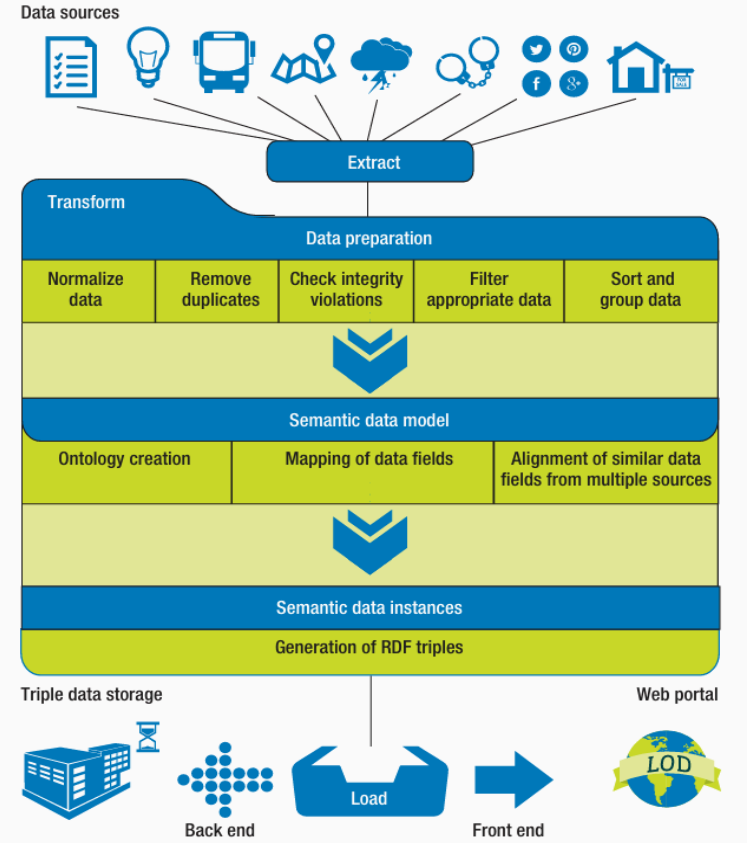


Figura 1: Framework de integração de dados por ETL [13]

Por um lado, a extração é a etapa inicial e fundamental para o início do processo, é nela que toda a base do pipeline de dados é definida. Durante essa fase, os dados brutos podem ser coletados desde bancos de dados relacionais até fontes não estruturadas como Logs de erros, documentos de texto ou APIs. Esses dados não sofrem nenhum tipo de transformação durante essa etapa, são apenas coletados e armazenados na base do usuário.

Partindo para a segunda etapa, a transformação ela abrange a limpeza, padronização e enriquecimento desses dados que foram coletados na extração, garantindo sua integridade e qualidade. Essa fase é de extrema importância para a transformação de dado em informação, uma vez que os dados extraídos podem conter erros, valores ausentes, formatos inconsistentes e duplicações e a transformação faz justamente o tratamento desses e outros casos.

Por fim, a carga faz referência à inserção dos dados, agora extraídos e transformados, em um destino estruturado para uso e análises futuras. Esse carregamento pode ser realizado de formas diversas dependendo da necessidade do usuário e/ou cliente, podendo ser feita para *Data Warehouses*, bancos de dados relacionais, *Data lakes* ou um arquivo Excel.

2.3. *Business Intelligence* (BI)

Business Intelligence (BI) faz referência ao conjunto de estratégias e tecnologias utilizadas para a análise de dados empresariais, permitindo uma tomada de decisões mais assertiva, visto que ela será baseada em informações mais estruturadas. Tal conceito surgiu na década de 1970 com os primeiros sistemas de suporte à decisão e evoluiu para soluções mais robustas que incluem OLAP (Processamento Analítico Online), mineração de dados e análises preditivas [10]. O uso dessas estratégias tem se tornado um diferencial competitivo em setores como varejo, finanças e saúde, uma vez que ele permite identificar tendências e otimizar processos internos de uma forma que antes não era tão simples [14].

O BI vem acompanhado de diversas ferramentas que possibilitam a coleta, processamento e visualização de grandes volumes de dados, gerando relatórios interativos e *dashboards* dinâmicos que facilitam o entendimento da informação por parte do usuário. Como por exemplo Tableau e Power BI, uma das principais ferramentas atualmente no mercado de dados, que foi desenvolvida pela Microsoft e permite a conexão entre múltiplas fontes de dados, transformação e modelagem de dados e, principalmente, a criação de visualizações interativas [15]. A Figura 2 do quadrante da Gartner sobre as plataformas de *Business Intelligence* confirma a liderança do Power BI [15].

Figura 2: Quadrante Mágico da Gartner Group [15]

De forma resumida, a plataforma do Power BI conta com funcionalidades avançadas como DAX (*Data Analysis Expressions*) para cálculos personalizados e

Power Query para ETL. Além disso, sua integração com o Azure e o Microsoft 365 facilita a implementação da plataforma em ambientes corporativos [16].

3. METODOLOGIA E DESENVOLVIMENTO

Para a realização deste trabalho, foi adotada uma abordagem baseada na construção de um pipeline de dados estruturado em etapas, visando o tratamento e a análise de grandes volumes de informações sobre TMRT no Brasil. A metodologia contempla desde a seleção das bases de dados até a disponibilização de *dashboards* interativos que possibilitam a interpretação dos resultados de forma clara e acessível. Esta seção está organizada seguindo a ordem destes cinco tópicos para detalhar os métodos e o desenvolvimento:

- **Pesquisa**, etapa inicial de todo o trabalho para a definição do tema, bases e processos a serem realizados;
- **Bases utilizadas**, que apresenta as fontes de dados e sua relevância para o estudo;
- **Extração**, que descreve o processo de coleta e leitura dos dados brutos via biblioteca Pysus;
- **Transformações**, onde são aplicadas técnicas de limpeza, padronização e enriquecimento dos dados utilizando PySpark e Pandas;
- **Carga**, que aborda o armazenamento e a estruturação final dos dados para análise; e
- **Power BI**, que trata da criação de visualizações interativas para apoio à análise epidemiológica e tomada de decisão.

Cada etapa foi desenvolvida com foco na consistência, escalabilidade e reprodutibilidade do processo, permitindo gerar indicadores relevantes para a compreensão dos TMRT e contribuir com análises e a formulação de políticas públicas em saúde mental no trabalho.

3.1. Pesquisa

A etapa inicial consistiu em uma pesquisa exploratória e bibliográfica, fundamental para o desenvolvimento do trabalho. Essa fase teve como propósito identificar, selecionar e analisar artigos, relatórios institucionais e outras publicações que abordassem diretamente ou indiretamente a problemática em questão. A pesquisa não se restringiu a um momento pontual, mas sim se manteve de forma ao longo de todo o desenvolvimento, permitindo atualizações constantes e incorporando novos referenciais e bases à medida que surgiam informações relevantes. Essa estratégia assegurou uma compreensão aprofundada do contexto, dos conceitos-chave, das bases existentes para análises e das bibliotecas Python que ajudariam no processo de ETL..

3.2. **Bases utilizadas**

3.2.1. **Bases TMRT**

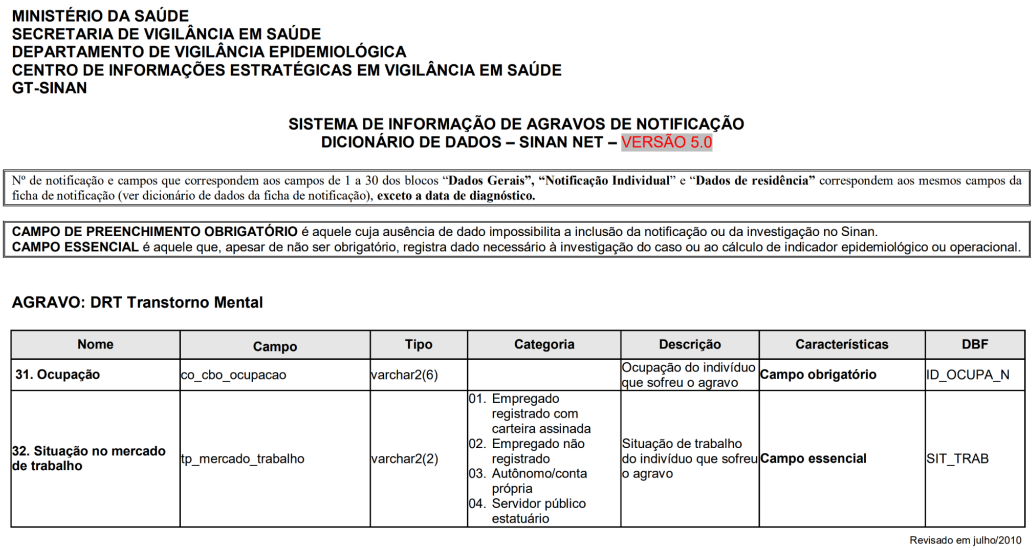
Primeiramente, houve uma etapa de pesquisa que se manteve ao longo de todo o desenvolvimento, com o objetivo de achar trabalhos relacionados, entender o tema e buscar bases que abordassem a problemática almejada.

As primeiras bases de dados selecionadas para realizar a extração foram as voltadas para o TMRT: DataSus e a SINAN. Primeiramente, o Datasus é o sistema de informações em saúde do Ministério da Saúde do Brasil e é nele que tem-se o armazenamento e disponibilização de dados epidemiológicos e administrativos do setor. Ele fornece informações sobre mortalidade, hospitalizações, imunizações e outros indicadores fundamentais para a gestão pública de saúde, por exemplo [12]. O acesso aos dados do Datasus possibilita que profissionais realizem análises epidemiológicas detalhadas, auxiliando na formulação de políticas públicas e na identificação de padrões das doenças listadas no sistema. No contexto da ciência de dados do projeto, ferramentas como Pandas e PySUS facilitam a extração e tratamento dessas informações, permitindo a modelagem estatística e *machine learning*, por exemplo [13].

Segundamente, o SINAN é um sistema nacional utilizado para registro e monitoramento de doenças e agravos de notificação compulsória no Brasil que é disponibilizado através do Datasus. Ele tem como principal objetivo subsidiar análises epidemiológicas e apoiar a tomada de decisões relacionadas à saúde pública [17]. O fluxo de dados no Sinan inicia-se nas unidades de saúde, onde é preenchida a Ficha Individual de Notificação (FIN) para cada caso suspeito de doença e essas fichas são encaminhadas às Secretarias Municipais de Saúde que são as responsáveis por inserir os dados no sistema. Cada município deve enviar os dados semanalmente às Secretarias Estaduais de Saúde, que repassam essa informações quinzenalmente ao Ministério da Saúde, seguindo um cronograma nacional. Além da FIN, também são utilizados instrumentos como a Ficha de Investigação, Planilhas de Surtos e Boletins de Acompanhamento. Mesmo na ausência de casos, é obrigatória a notificação negativa, indicando que o sistema está ativo e essa regularidade no envio das informações é extremamente essencial e sua ausência pode acarretar a suspensão de repasses financeiros do Ministério da Saúde.

Para a construção do pipeline do trabalho, foram selecionados todos os dados disponíveis sobre o problema em análise, resultando em uma amostragem dos anos de 2006 a 2023, até o momento desse trabalho- contendo 22399 registros, que foram extraídos através da biblioteca Pysus desenvolvida pela comunidade que facilita a extração desses arquivos. Atualmente, no total são 22399 linhas e 60 colunas de informação e todas as colunas apresentam códigos que referenciam o

dicionário de dados da base disponibilizado pelo portal do governo brasileiro, como apresentado na Figura 3.



MINISTÉRIO DA SAÚDE
SECRETARIA DE VIGILÂNCIA EM SAÚDE
DEPARTAMENTO DE VIGILÂNCIA EPIDEMIOLÓGICA
CENTRO DE INFORMAÇÕES ESTRATÉGICAS EM VIGILÂNCIA EM SAÚDE
GT-SINAN

SISTEMA DE INFORMAÇÃO DE AGRAVOS DE NOTIFICAÇÃO
DICIONÁRIO DE DADOS – SINAN NET – VERSÃO 5.0

Nº de notificação e campos que correspondem aos campos de 1 a 30 dos blocos "**Dados Gerais", "Notificação Individual**" e "**Dados de residência"** correspondem aos mesmos campos da ficha de notificação (ver dicionário de dados da ficha de notificação), **exceto a data de diagnóstico.**

CAMPO DE PREENCHIMENTO OBRIGATÓRIO é aquele cuja ausência de dado impossibilita a inclusão da notificação ou da investigação no Sinan.
CAMPO ESSENCIAL é aquele que, apesar de não ser obrigatório, registra dado necessário à investigação do caso ou ao cálculo de indicador epidemiológico ou operacional.

AGRAVO: DRT Transtorno Mental

Nome	Campo	Tipo	Categoria	Descrição	Características	DBF
31. Ocupação	co_cbo_ocupacao	varchar2(6)		Ocupação do indivíduo que sofreu o agravo	**Campo obrigatório**	ID_OCUPA_N
32. Situação no mercado de trabalho	tp_mercado_trabalho	varchar2(2)	01. Empregado registrado com carteira assinada 02. Empregado não registrado 03. Autônomo/conta própria 04. Servidor público estatuário	Situação de trabalho do indivíduo que sofreu o agravo	**Campo essencial**	SIT_TRAB

Revisado em julho/2010

Figura 3: Dicionário de dados SINAN sobre TMRT [17]

3.2.2. **Bases complementares**

Para complementar os dados do SINAN, que possui apenas dados em IDs códigos, e permitir análises mais completas por municípios, estados e ocupações, foram selecionadas bases oficiais sobre os temas. Para localização, utilizou-se duas bases em formato ".CSV", salvas pela base de dados do IBGE, contendo informações sobre unidades federativas e municípios do Brasil (27 UFs e 5.570 municípios). O uso dessas bases foi essencial para conseguir a padronização dos dados geográficos, permitindo análises espaciais mais precisas e amigáveis de serem entendidas. E, para ocupações, foi utilizada a de Classificação Brasileira de Ocupações (CBO 2002), nela está contida a classificação oficial das ocupações no Brasil, com códigos e descrições das atividades profissionais. Essa base está disponível em PDF no site oficial do Ministério do Trabalho, mas para fins acelerar o desenvolvimento, foi utilizado um arquivo CSV disponibilizado pelo usuário "lucassmacedo" no GitHub [18].

3.3. **Extração**

A etapa de extração consistiu na coleta de dados brutos referentes a TMRT, utilizando como principal fonte o SINAN, mantido pelo Ministério da Saúde. A extração foi realizada utilizando a biblioteca Pysus, que permite justamente um acesso extremamente mais fácil e simplificado aos dados públicos presentes na base do DATASUS.

O ambiente de desenvolvimento foi configurado utilizando o Google Colab, devido a sua facilidade de desenvolvimento em um ambiente gratuito e na nuvem e

a sua integração ao Google Drive para armazenamento dos arquivos e organização do pipeline de dados. A coleta desses dados iniciais abrangeu todos os anos presentes no SINAN até o momento- 2006 a 2023-, filtrando os registros associados ao código de doença "MENT", que representa os transtornos mentais relacionados ao trabalho. Para cada ano, os arquivos foram baixados utilizando a função "get_files()" e armazenados após serem convertidos para objetos DataFrame com a adição de uma coluna de ano. Os dados de todos os anos foram então concatenados em um único DataFrame denominado "df_geral". Por fim, o arquivo consolidado foi salvo em formato de "Parquet" dentro do diretório do Google Drive reservado para o projeto, representando a camada Bronze do pipeline de dados, onde os dados permanecem em estado bruto.

Além dessa base principal extraída do SINAN, foram extraídos para o projeto dois arquivos auxiliares no formato, também, de "csv", com o objetivo de tornar a análise mais amigável e fácil de entender, através de cruzamentos relevantes com variáveis contextuais:

- **Classificação Brasileira de Ocupações (CBO2002)**: arquivo extraído manualmente a partir do repositório do "lucassmacedo" no GitHub [18], contendo a descrição e estrutura hierárquica dos códigos de ocupação. Esse dado permite analisar a distribuição das notificações por categoria profissional;
- **Municípios, Unidades Federativas e Regiões**: arquivo obtido manualmente a partir do site oficial do IBGE, contendo informações territoriais necessárias para mapeamento geográfico das notificações analisadas, permitindo que a análise agora tenha informações textuais sobre o município e a unidade federativa da notificação [19].

Todos os arquivos foram armazenados na camada bronze no Google Drive e lidos no ambiente do Colab por meio de comandos de leitura de arquivos "CSV". Essa etapa garantiu a centralização e organização dos dados brutos necessários para as fases seguintes de transformação (prata) e análise (ouro), estabelecendo uma base sólida e padronizada para o desenvolvimento do projeto.

3.4. **Transformações**

Após a etapa de extração e consolidação dos dados brutos, foi realizada a fase de transformação que corresponde à camada prata do pipeline de dados, utilizando puramente Pyspark para transformações e Pandas. Esta etapa teve como objetivo padronizar, enriquecer e preparar as informações para análises posteriores, visando garantir uma maior integridade e coerência dos dados.

Inicialmente, o arquivo bruto sobre TMRT, salvo na etapa anterior no Google Drive, foi carregado como um Data Frame do PySpark. Em seguida, foi aplicado um processo de renomeação das colunas, substituindo os identificadores técnicos por

nomes mais descritivos e padronizados, facilitando a interpretação das variáveis ao longo do projeto. Essa etapa do processo é possível ser visualizada no recorte presente no Quadro 1.

Campo Original	**Novo Nome**	**Descrição**
ID_AGRAVO	id_codigo_agravo	Código do agravo notificado segundo CID-10
DT_NOTIFIC	dat_data_notificacao	Data de preenchimento da ficha de notificação
SEM_NOT	vlr_semana_notificacao	Semana epidemiológica que o caso foi notificado (AAAASS)
NU_ANO	vlr_ano_notificacao	Ano da notificação, preenchido a partir da data de notificação
SG_UF_NOT	id_uf_notificacao	Sigla da Unidade Federativa onde está localizada a unidade notificadora
ID_MUNICIP	id_municipio_notificacao	Código do município onde está localizada a unidade notificadora
ID_REGIONA	id_regiao_notificacao	Código da regional de saúde da unidade notificadora
DT_DIAG	dat_data_diagnostico	Data do diagnóstico ou primeiros sintomas
NU_IDADE_N	vlr_idade_numero	Idade composta por código de tempo e valor: 1=h, 2=d, 3=m, 4=a
CS_SEXO	id_sexo	Sexo do paciente: M – Masculino, F – Feminino, I – Ignorado
CS_RACA	id_raca_cor	Raça/Cor autodeclarada: 1 – Branca, 2 – Preta, 3 – Amarela, 4 – Parda, 5 – Indígena, 9 – Ignorado
CS_ESCOL_N	id_escolaridade	Grau de instrução,

		considerando a última série concluída
ID_OCUPA_N	id_ocupacao	Ocupação do indivíduo segundo tabela CBO
SIT_TRAB	id_situacao_trabalho	Situação no mercado de trabalho do indivíduo
CNAE	id_codigo_atividade_economica	Classificação Nacional da Atividade Econômica do contratante

Quadro 1: Renomeações com suas respectivas descrições

Após a leitura dos dados da bronze e renomeação, utilizando o dicionário de dados do SINAN, algumas variáveis categóricas foram transformadas para incluir descrições textuais em português, substituindo códigos por rótulos compreensíveis. Como a Quadro 2 que apresenta o mapeamento “código-descrição” do campo de “situação_trabalho” que no arquivo original era marcado apenas com o código que dificulta o entendimento da informação e Quadro 3 que apresenta o mapeamento de evolução do caso, seguindo a mesma lógica do Quadro 2.

Código	**Descrição**
1	Empregado registrado com carteira assinada
2	Empregado não registrado
3	Autônomo/Conta própria
4	Servidor público estatutário
5	Servidor público celetista
6	Aposentado
7	Desempregado
8	Trabalho temporário
9	Cooperativado
10	Trabalhador avulso
11	Empregador
12	Outros
99	Ignorado
Outro	Desconhecido

Quadro 2: Mapeamento do campo “id_situacao_trabalho”

Código	**Descrição**

1	Cura
2	Cura não confirmada
3	Incapacidade temporária
4	Incapacidade permanente parcial
5	Incapacidade permanente total
6	Óbito por doença relacionada ao trabalho
7	Óbito por outra causa
8	Outro
9	Ignorado
Outro	Desconhecido

Quadro 3: Mapeamento do campo “id_evolucao_caso”

No caso da idade, a variável original era composta por um código que indicava a unidade de medida (anos, meses, dias, horas), como por exemplo 3009 indicava nove meses e 4018, dezoito anos. Diante disso, o tratamento envolveu a separação e conversão desses valores para idade estimada em anos, com criação adicional da variável “faixa_idade”, agrupando os indivíduos por intervalos de 10 anos, como é possível observar no código da Figura 4.

```
#IDADE EM ANOS
df_transtornos_final = df_transtornos_final.withColumn("tipo_idade",
                                                substring("vlr_idade_numero", 1, 1))
df_transtornos_final = df_transtornos_final.withColumn("valor_idade",
                                                substring("vlr_idade_numero", 2, 3).cast("int"))

df_transtornos_final = df_transtornos_final.withColumn(
    "idade_em_anos",
    when(col("tipo_idade") == "1", col("valor_idade") / 8760)     # horas -> anos
    .when(col("tipo_idade") == "2", col("valor_idade") / 365.0)   # dias -> anos
    .when(col("tipo_idade") == "3", col("valor_idade") / 12.0)    # meses -> anos
    .when(col("tipo_idade") == "4", col("valor_idade"))           # anos -> anos
    .otherwise(None)
)

df_transtornos_final = df_transtornos_final.withColumn("faixa_idade", floor(col("idade_em_anos") / 10) * 10)
```

Figura 4: Código tratamento de idade

Por fim, o *DataFrame* final foi salvo no formato parquet, compondo a camada prata do projeto e estando pronto para ser analisado na etapa seguinte de enriquecimento (ouro) dos dados.

3.5. **Carga**

Por fim, foi preparada a última etapa do pipeline de dados, correspondente à camada ouro, que consistiu na preparação e organização final das informações com foco no consumo através das ferramentas de visualização e análise, nesse caso o Power BI. Essa fase é responsável por consolidar os dados tratados e enriquecidos

nas etapas anteriores, garantindo que estejam estruturados de maneira analítica, padronizada e com semântica compreensível para os usuários finais.

Inicialmente, houve a leitura na sessão do arquivo das bases auxiliares e a base principal, que foram extraídas e tratadas anteriormente, com as informações de ocupação, municípios e unidades federativas e transtornos mentais relacionados ao trabalho que estavam salvos no Google Drive. Logo após a leitura , foi realizado o processo de junção (*join*) entre a base consolidada de transtornos mentais e os dados auxiliares, enriquecendo a base final com as descrições dos municípios e ocupações que antes se encontravam apenas através de IDs e códigos. Detalhes sobre os *joins* realizados podem ser vistos no Quadro 4.

Base 1	**Base 2**	**Tipo do Join**	**Chave Utilizada**
Base TMRT	Base municípios e UF	left	ID municípios da notificação = ID município
Base TMRT + Base UF	Base de ocupação CBO	left	ID da ocupação = ID da ocupação CBO

Quadro 4: *Joins* realizados na camada ouro

Com os dados integrados, foi construída a visão analítica final que seria carregada para o Power BI e lá ser consumida para a criação dos *dashboards*. Essa construção se deu através de um agrupamento e seleção de campos considerando diversas dimensões de análises e use-tasks consideradas relevantes, como:

- **Ano de notificação**: permite realizar análises temporais, identificando tendências, padrões sazonais e variações entre os anos na ocorrência de transtornos mentais relacionados ao trabalho. Essa dimensão é fundamental para entender a progressão histórica dos casos e avaliar o impacto de políticas públicas ao longo do tempo;
- **UF e município**: possibilitam análises espaciais detalhadas, permitindo a identificação de regiões com maior concentração de casos. Essa informação é relevante para direcionar estratégias regionais de prevenção, alocação de recursos e fortalecimento da rede de atenção psicossocial em áreas prioritárias;
- **Faixa etária e sexo**: fornecem o perfil demográfico dos trabalhadores afetados. Essas variáveis são importantes para identificar grupos etários mais vulneráveis, bem como possíveis desigualdades de gênero na problemática;
- **Escolaridade e raça/cor**: permitem análises de cunho socioeconômico, evidenciando desigualdades estruturais que podem influenciar na exposição a riscos psicossociais. Esses dados contribuem para compreender como fatores sociais e raciais se relacionam como adoecimento mental;

- **Ocupação e situação de trabalho**: são essenciais para correlacionar os transtornos mentais com as condições laborais. Essas dimensões permitem identificar quais categorias profissionais e tipos de vínculo empregatício estão mais associados à ocorrência dos casos, subsidiando ações preventivas em setores específicos;
- **Uso de álcool, drogas e psicofármacos**: indicam fatores associados ou agravantes dos casos notificados. A análise desses comportamentos pode revelar padrões de autocuidado, automedicação ou vulnerabilidade adicional entre os trabalhadores afetados;
- **Evolução do caso e encaminhamento ao CAPS**: possibilitam avaliar os desfechos dos casos e a efetividade do fluxo de cuidado em saúde mental. Informações sobre cura, incapacidade ou óbito, bem como o encaminhamento para atendimento especializado, são cruciais para monitorar a resposta dos serviços de saúde.

Finalmente, essa visualização final foi exportada como arquivo ".csv" ao diretório do Google Drive para servir posteriormente como fonte de dados para as análises no Power BI.

3.6. **Power BI**

Após a consolidação da base final na camada ouro, o arquivo foi carregado no Power BI como fonte de dados, através do Google Drive para um setup que faz a atualização automática dos dados da nuvem para a máquina, para a construção dos *dashboards* interativos. Cada análise foi organizada em uma aba individual no relatório, facilitando a navegação e interpretação por parte do usuário.

Foram utilizados diferentes tipos de gráficos, selecionados conforme a natureza dos dados e o tipo de análise desejada. A seguir, estão listadas as visualizações que foram desenvolvidas:

- **Linha temporal de casos por ano**: gráfico de linha simples, exibindo a evolução do total de casos notificados por ano. Permite visualizar tendências e variações ao longo do tempo;
- **Distribuição geográfica por estado e município**: gráfico de mapa coroplético e de calor, utilizando campos de UF e município. Essa visualização identifica regiões com maior incidência de transtornos mentais relacionados ao trabalho;
- **Perfil demográfico dos afetados**: gráficos de barra, coluna de pizza foram utilizados para representar variáveis como sexo, faixa etária, raça/cor e escolaridade. Essas visualizações permitem identificar os principais grupos demográficos impactados;
- **Por ocupação e situação de trabalho**: gráficos de barra mostram a distribuição dos casos por categoria ocupacional e tipo de vínculo de trabalho, evidenciando os grupos mais expostos;

- **Fatores associados**: uso de álcool, drogas e psicofármacos: gráficos de barra simples foram usados para representar o comportamento dos trabalhadores em relação ao uso dessas substâncias;
- **Evolução do caso e encaminhamento ao CAPS**: gráfico de barras e pizza com o total de casos por tipo de desfecho (cura, óbito, incapacidade etc.) e se houve ou não encaminhamento para atendimento especializado.

Todas as análises foram desenvolvidas com base no arquivo de saída da camada ouro, utilizando os campos agregados e enriquecidos previamente tratados, com o objetivo de oferecer uma visão clara, segmentada e exploratória dos dados de transtornos mentais relacionados ao trabalho. No capítulo seguinte os resultados dessa construção serão mostrados através dos *dashboards* construídos.

4. Resultados e Discussões

Neste capítulo será apresentado o *dashboard* e cada uma das visualizações construídas. Cada parte do painel foi dedicada a uma dimensão de análise relevante, com o objetivo de facilitar a interpretação dos dados e apoiar análises, ações de prevenção e intervenção em saúde mental no trabalho.

De forma mais macro, as abas principais do *dashboard* possuem todas as informações consolidadas permitindo uma visualização geral das informações sobre TMRT, nessas abas o usuário pode realizar filtros por ano de registro do caso, por sexo, por escolaridade, por raça/cor, por faixa etária, por municipio e entre outros. A Figura 5 mostra a primeira aba do *dashboard* com foco na visualização geral e interativa dos casos, com filtros que permitem ao usuário explorar os dados conforme o ano de notificação, sexo, faixa etária, escolaridade, raça/cor, município e ocupação. Os gráficos foram dispostos de forma a facilitar a identificação de padrões e concentrações de casos ao longo do tempo e por grupo populacional. A Figura 6 mostra a segunda aba que aprofunda a análise, trazendo informações mais clínicas e circunstanciais dos casos registrados, com a intenção foi fornecer uma perspectiva voltada ao contexto de trabalho e evolução do adoecimento. A separação em abas e a escolha dos gráficos foram guiadas pensando na facilidade de leitura e interpretação visual, organização temática e aderência às boas práticas de visualização de dados.

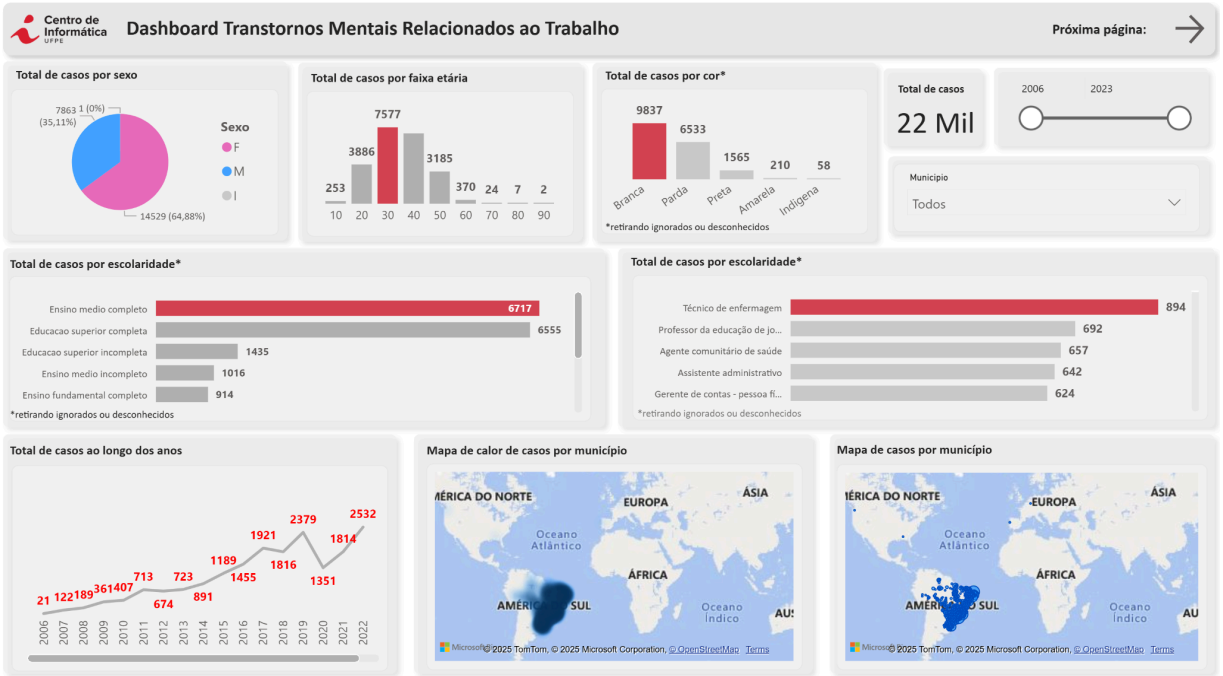


Figura 5: Página 1 *Dashboard*

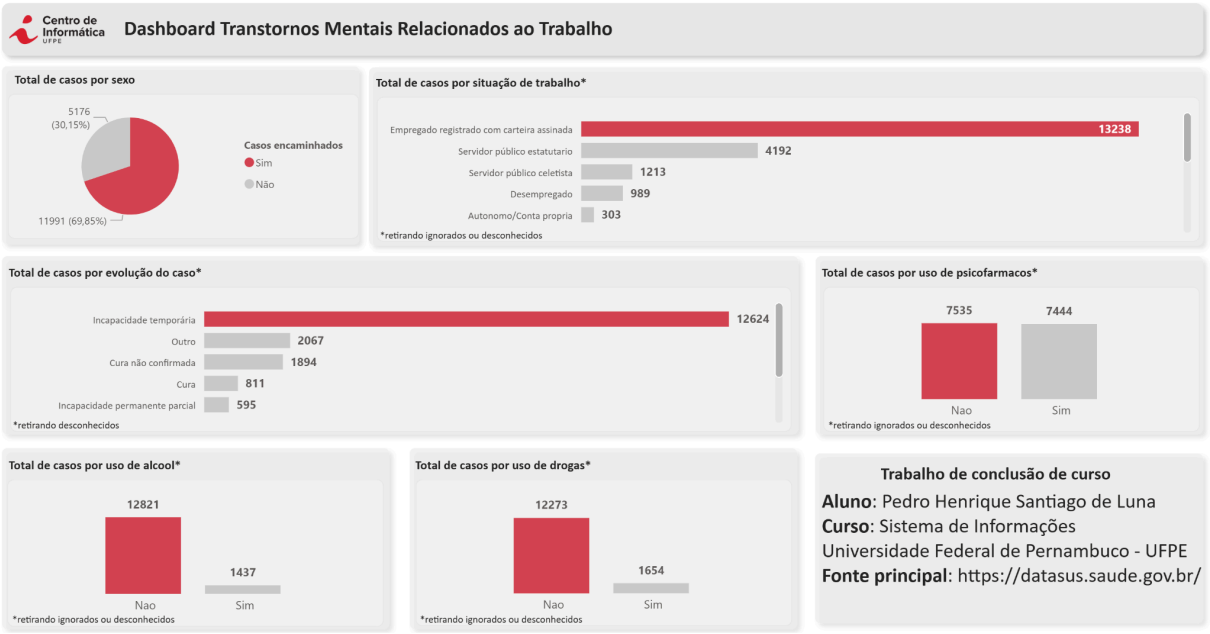


Figura 6: Página 2 *Dashboard*

De forma mais micro será apresentado individualmente cada parte do *dashboard*, detalhando possíveis análises e use-tasks para aquela visualização.

4.1.1. Casos ao longo do tempo

A linha do tempo apresentada na Figura 7 permite ao usuário realizar análises temporais sobre a evolução dos casos de TMRT entre 2006 e 2023. Por meio desta visualização, é possível identificar padrões de crescimento, quedas pontuais e comportamentos atípicos ao longo dos anos. Como exemplo de uso, o usuário pode avaliar o aumento significativo no número de notificações, partindo de 21 casos em 2006 para 3.835 em 2023, e investigar variações abruptas como a queda observada em 2020, possivelmente relacionada à pandemia da COVID-19. Além disso, é possível calcular taxas de crescimento entre anos consecutivos — como os aumentos de 39,6% entre 2021 e 2022, e de 51,5% entre 2022 e 2023 — e utilizar esses dados para projeções futuras, como uma estimativa de mais de 5.500 casos em 2024, caso a tendência se mantenha. Essa visualização atende a use-tasks como: análise de tendência ao longo do tempo, identificação de anos críticos, comparação entre períodos e apoio a decisões baseadas em séries históricas. Ela oferece uma base exploratória para gestores públicos, pesquisadores e analistas que desejam compreender a evolução do fenômeno e orientar ações futuras.

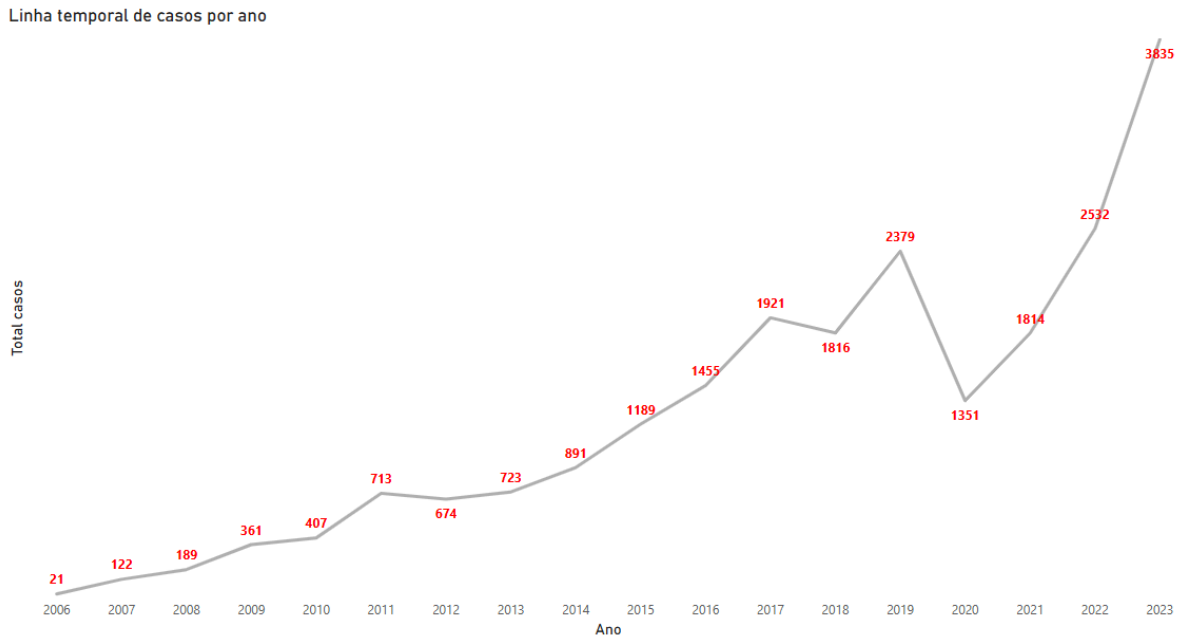


Figura 7: Gráfico de casos ao longo do tempo

4.1.2. **Análise geográfica**

As visualizações apresentadas na Figura 8 permitem a análise geográfica da distribuição de casos de TMRT em nível municipal. A primeira visualização, em formato de mapa de calor, evidencia as regiões com maior concentração de notificações, enquanto a segunda, em formato de mapa de pontos, apresenta a localização exata dos municípios com registros, com o tamanho do marcador proporcional ao volume de casos. Essas representações atendem a diferentes use-tasks, como:

- Identificação de áreas com maior densidade de notificações;
- Comparação entre municípios e estados;
- Reconhecimento de padrões regionais e suporte à priorização de políticas públicas em determinadas regiões.

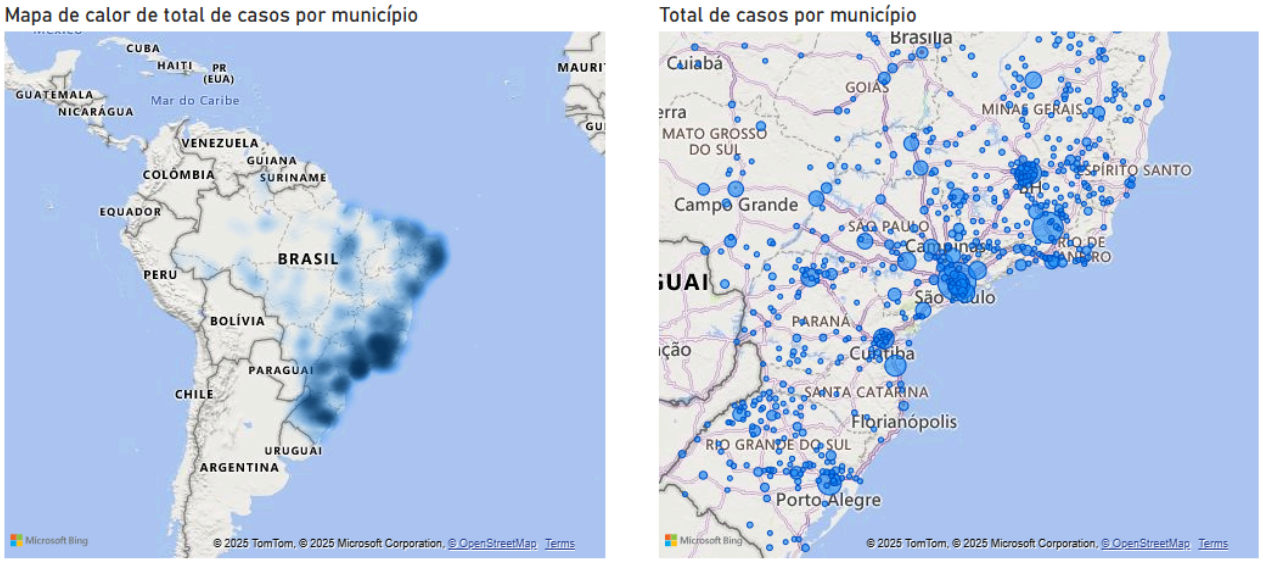


Figura 8: Mapas de casos por município

Por meio dessas visualizações, o usuário pode, por exemplo, observar agrupamentos de casos em determinadas regiões do país, identificar municípios com elevado número de notificações ou ainda comparar concentrações em áreas urbanas versus rurais. A combinação entre os dois mapas fornece uma visão complementar- uma mais sintética e agregada (calor), e outra mais detalhada e pontual (dispersão)-, permitindo uma exploração espacial rica e interativa dos dados.

4.1.3. **Perfil demográfico**

As visualizações apresentadas nas Figuras 9, 10, 11 e 12 permitem a análise do perfil demográfico dos indivíduos com notificações de TMRT, com base em quatro variáveis principais: sexo, faixa etária, raça/cor e escolaridade. Esses gráficos possibilitam a realização de use-tasks como:

- Segmentação da população afetada por características sociodemográficas;
- Comparação entre grupos específicos;
- Identificação de perfis mais frequentemente associados às notificações e apoio a políticas públicas direcionadas a grupos vulneráveis;
- Cruzamentos com outras variáveis no *dashboard*, como localidade ou período.

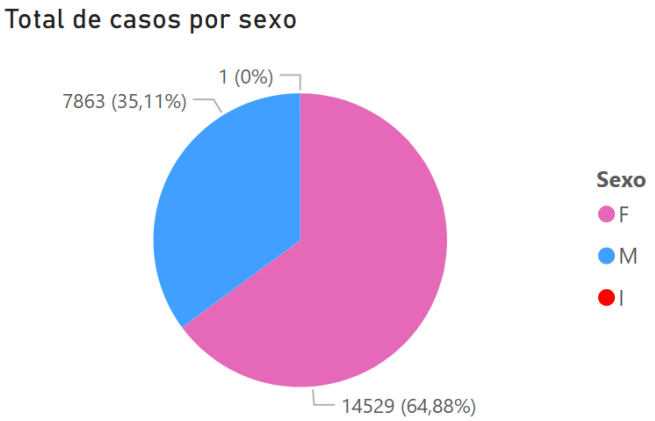


Figura 9: Gráfico de casos por sexo

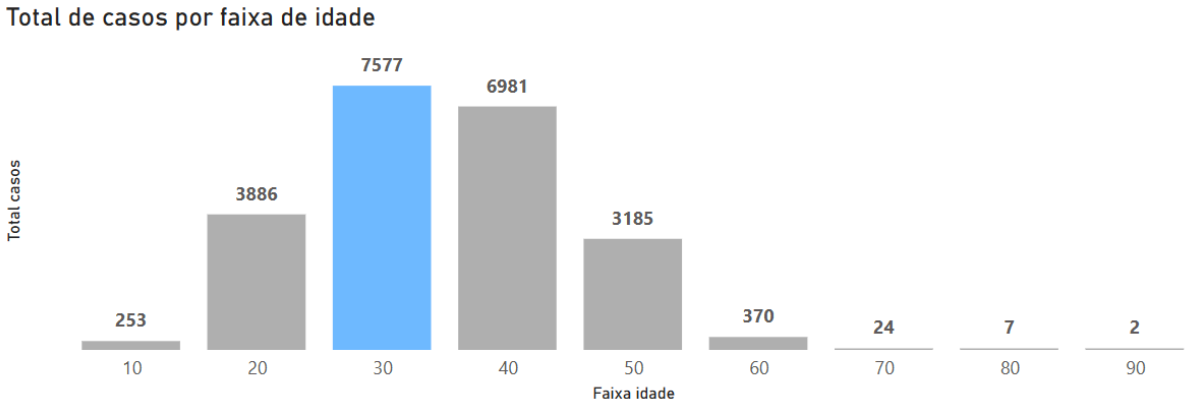


Figura 10: Gráfico de casos por faixa etária

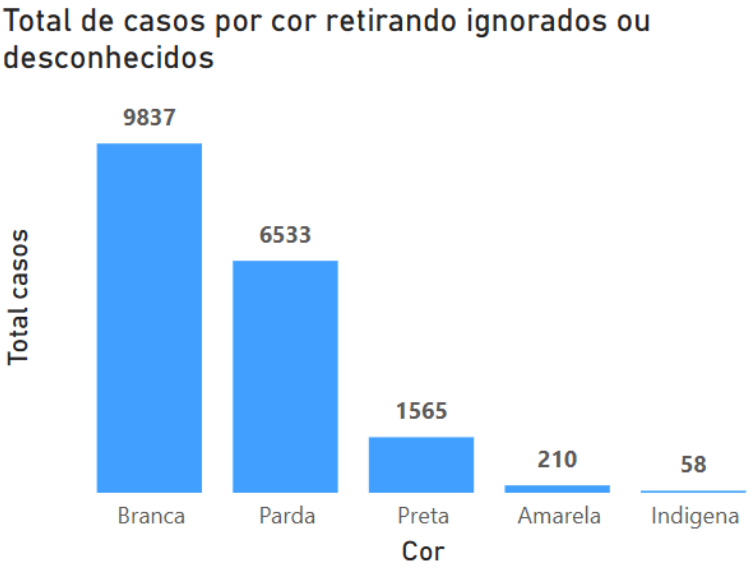


Figura 11: Gráfico de casos por raça/cor

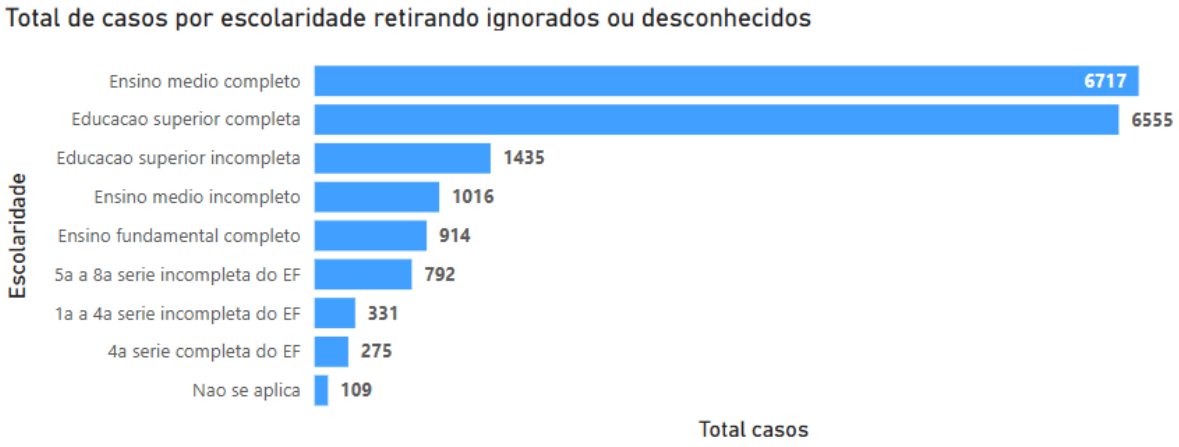


Figura 12: Gráfico de casos por escolaridade

Na visualização por sexo na Figura 9, por exemplo, o usuário pode verificar a proporção entre os casos femininos e masculinos. A distribuição por faixa etária na Figura 10 permite analisar a concentração de notificações em diferentes ciclos de vida, enquanto o gráfico de raça/cor na Figura 11 mostra a composição dos casos por autodeclaração. Já a escolaridade na Figura 12 permite observar o nível de instrução dos indivíduos registrados, com a possibilidade de avaliar se o problema atinge mais pessoas com baixa ou alta formação.

Essas visualizações oferecem uma base exploratória relevante para profissionais de saúde e pesquisadores interessados em compreender os perfis sociais mais impactados, orientar campanhas de prevenção ou identificar possíveis desigualdades no processo de notificação.

4.1.4. Situação de trabalho e ocupação

As visualizações das Figuras 13 e 14 permitem explorar os dados com base em duas variáveis: a situação de trabalho do indivíduo e a sua ocupação declarada. Essas visualizações oferecem suporte a use-tasks como:

- Identificação de vínculos empregatícios mais frequentes entre os casos notificados;
- Análise da presença de registros entre diferentes tipos de trabalhadores (formais, informais, autônomos, desempregados etc.);
- Investigação das ocupações mais associadas aos casos de TMRT e avaliação de padrões de risco ocupacional por setor de atividade.

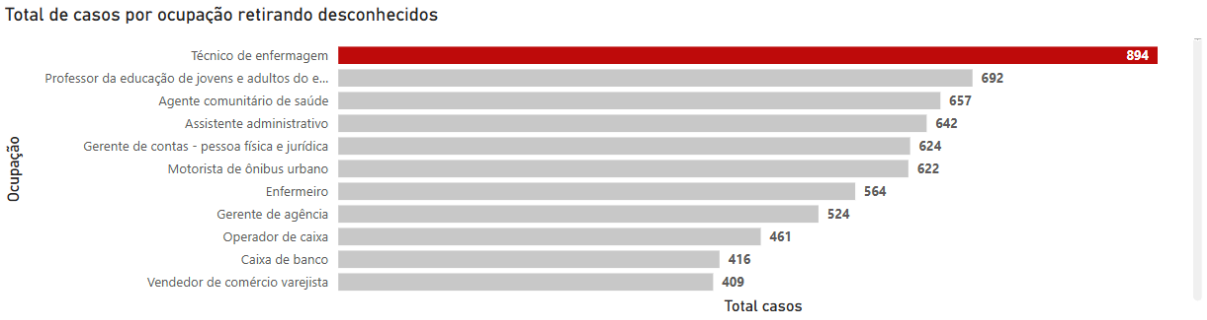


Figura 13: Gráfico de casos por ocupação

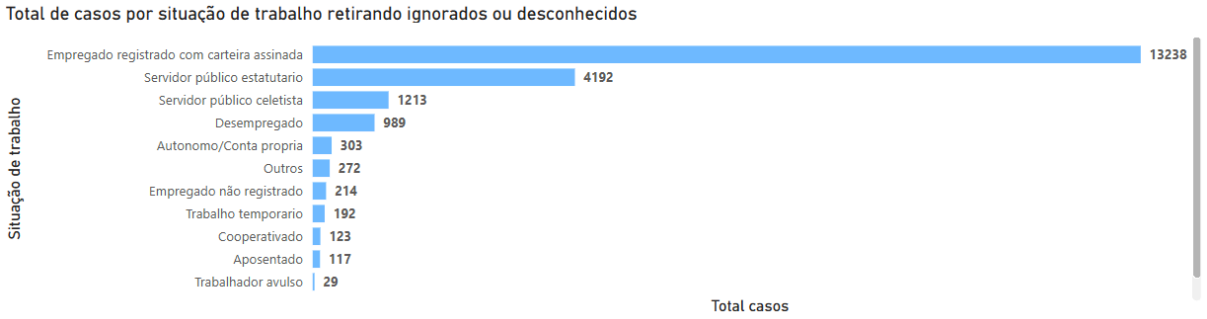


Figura 14: Gráfico de casos por contrato de trabalho

A visualização por situação de trabalho na Figura 13 possibilita, por exemplo, que o usuário compare diferentes vínculos empregatícios, como trabalhadores com carteira assinada, servidores públicos ou autônomos, e observe sua representação no total de casos. Já o gráfico por ocupação na Figura 14 permite uma análise detalhada das categorias profissionais mais frequentes entre as notificações, destacando grupos com maior número de registros. Essas análises podem apoiar na priorização de grupos ocupacionais em ações preventivas, além de fornecer subsídios para o entendimento do contexto profissional em que os transtornos estão sendo mais frequentemente notificados.

4.1.5. Desfecho dos casos

As Figuras 15 e 16 apresentam duas dimensões relevantes para o acompanhamento dos TMRT: a evolução clínica do caso e o encaminhamento para serviços especializados, como o CAPS. Essas visualizações permitem a realização de tasks como:

- Análise da gravidade e desfecho dos casos notificados;
- Quantificação de incapacidades temporárias e permanentes;
- Identificação de óbitos;
- Verificação de ações subsequentes à notificação, como encaminhamentos, e avaliação da articulação com a rede de saúde mental.

Figura 15: Gráfico de desfecho dos casos

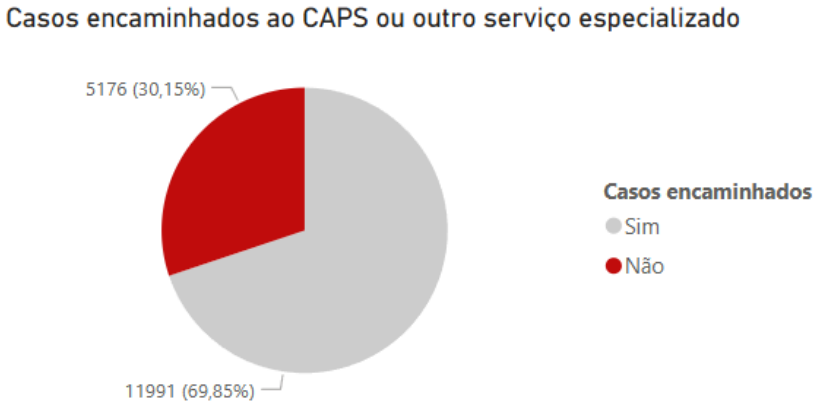


Figura 16: Gráfico de encaminhamento dos casos

Na visualização de desfecho na Figura 15, o usuário pode observar a proporção entre diferentes tipos de evolução clínica, como cura, incapacidade, óbito ou outros desfechos, e identificar padrões entre os registros. Já o gráfico de encaminhamento na Figura 16 possibilita verificar a efetividade da resposta institucional à notificação, permitindo comparar a quantidade de casos que geraram desdobramentos em termos de cuidado psicossocial com aqueles que não foram encaminhados. Essas informações apoiam na avaliação do fluxo pós-notificação e no fortalecimento da rede de atenção à saúde mental dos trabalhadores, especialmente em casos de maior gravidade ou reincidência.

4.1.6. **Fatores associados**

A Figura 17 apresenta uma análise de fatores associados aos casos analisados, com foco no uso de álcool, drogas e psicofármacos entre os indivíduos notificados. Essa visualização permite a realização de tarefas como:

- Análise da presença de consumo de substâncias entre os casos registrados;
- Comparação entre usuários e não usuários por tipo de substância;

- Verificação de possíveis correlações entre uso de substâncias e ocorrência de TMRT;
- Apoio a ações de prevenção, cuidado e suporte psicossocial voltadas a grupos mais vulneráveis.

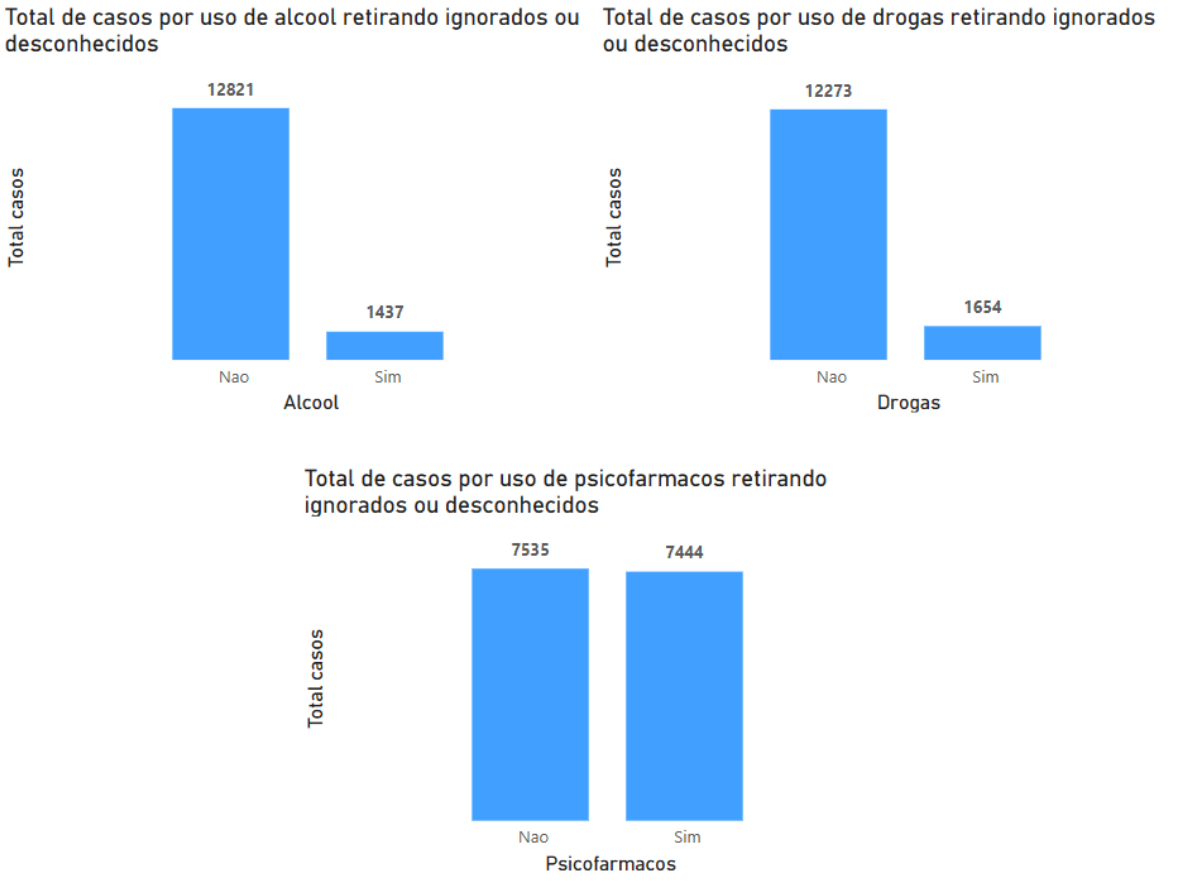


Figura 17: Gráfico de casos por faixa etária

Por meio desse conjunto de gráficos, o usuário pode identificar a frequência com que os indivíduos declararam o uso de álcool, drogas ou psicofármacos no momento da notificação. A separação por tipo de substância permite uma análise mais detalhada, facilitando a segmentação de casos e possibilitando futuros cruzamentos com outras variáveis como idade, ocupação ou desfecho. Essa visualização fornece insumos para que políticas públicas possam explorar a influência de fatores associados no agravamento ou na manifestação de transtornos mentais relacionados ao trabalho.

Com base nas visualizações construídas, foi possível disponibilizar um conjunto de análises exploratórias sobre os casos de TMRT no Brasil, abrangendo aspectos temporais, geográficos, sociodemográficos, ocupacionais, clínicos e contextuais. Os dashboards desenvolvidos possibilitam múltiplas combinações de filtros e cruzamentos, o que amplia a capacidade de interpretação e direcionamento das informações. Dessa forma, a entrega proposta neste trabalho oferece uma

ferramenta interativa e acessível, que pode apoiar pesquisadores, profissionais de saúde e gestores públicos na formulação de estratégias mais direcionadas de prevenção, acolhimento e monitoramento da saúde mental no contexto de trabalho.

5. CONCLUSÃO

Neste trabalho, foi realizada a análise de TMRT no Brasil por meio da construção de um pipeline de dados que viabilizasse a organização, o processamento e a visualização das informações de forma eficiente e acessível. Propôs-se uma abordagem baseada nas etapas de Extração, Transformação e Carga (ETL), utilizando tecnologias como PySpark, Pandas e Power BI, com o objetivo de tornar o acesso aos dados mais ágil e compreensível. A solução desenvolvida demonstrou resultados positivos nos seguintes aspectos: organização do fluxo de dados, capacidade de manipulação de grandes volumes de informação e geração de insights relevantes para subsidiar decisões em políticas públicas voltadas à saúde mental dos trabalhadores.

5.1. Considerações Finais

A construção do pipeline e dos dashboards interativos evidenciou o potencial da análise de dados epidemiológicos como ferramenta de apoio à gestão em saúde mental ocupacional. Ao longo do projeto, observou-se que a estruturação de um processo de ETL bem definido, somada ao uso de ferramentas analíticas, contribui para ampliar a transparência dos dados, facilitar a interpretação por diferentes perfis de usuários e promover ações mais assertivas. A possibilidade de realizar análises temporais, geográficas, demográficas e ocupacionais diretamente na interface do Power BI representa um avanço na democratização da informação sobre TMRT no Brasil. Ainda que o escopo do trabalho tenha se concentrado em visualizações exploratórias, a arquitetura desenvolvida permite futuras expansões com análises mais preditivas e prescritivas.

5.2. Limitações

Este estudo apresenta algumas limitações. O processo de extração e padronização de dados depende da qualidade da base original do DATASUS pode afetar a completude e a confiabilidade dos registros. Além disso, o trabalho não incluiu a aplicação de modelos estatísticos ou de aprendizado de máquina, que poderiam aprofundar a análise dos fatores associados aos casos.

5.3. Trabalhos Futuros

Com base na estrutura implementada, há diversas possibilidades para trabalhos futuros. A inclusão de análises multivariadas, modelagem estatística e machine learning pode gerar previsões e identificação de padrões não triviais nos dados. Além disso, a integração com outras bases de dados- como previdência social, registros de afastamento, CNAE e dados socioeconômicos - pode ampliar a contextualização dos casos. Também se recomenda o desenvolvimento de relatórios automatizados, painéis temáticos por setor produtivo e alertas de vigilância que possam ser utilizados diretamente por equipes de saúde pública e gestores. A

continuidade desse tipo de projeto representa um avanço na construção de sistemas mais inteligentes e responsivos para a promoção da saúde mental dos trabalhadores.

REFERÊNCIAS

[1] SHIMAOKA, Andre Massahiro; DUARTE, José Marcio; SILVA JUNIOR, Antonio Carlos da; LOPES, Luciano Rodrigo; et al. **Ansiedade no trabalho em tempos de mudança: tendências e perfis durante e pós-pandemia no estado de São Paulo 2020-2023**. SciELO Preprints, jan. 2025. Disponível em: https://doi.org/10.1590/SciELOPreprints.11019. Acesso em: 12 fev. 2025.

[2] TORRES, Gabriella Maria Schr; BACKSTROM, Jessica; DUFFY, Vincent G. **A systematic review of workplace stress and its impact on mental health and safety. In: Late Breaking Papers: 25th International Conference on Human-Computer Interaction**, p. 610–627, dez. 2023. Disponível em: https://doi.org/10.1007/978-3-031-48041-641. Acesso em: 01 mar. 2025.

[3] ORGANIZAÇÃO MUNDIAL DA SAÚDE (OMS). Depression. [S.l.]: **World Health Organization, 2023**. Disponível em: https://www.who.int/news-room/fact-sheets/detail/depression. Acesso em: 12 fev. 2025.

[4] INTERNATIONAL LABOUR ORGANIZATION (ILO). **Mental health in the workplace: a global concern**. 2022. Acesso em: 10 fev. 2025.

[5] SILVA, Antônio G. da; SERPA, Alexandre L.; NARDI, Antonio E.; KESSLER, H. P.; et al. **Mental illnesses and their impact on the Brazilian workforce: an analysis of the cost of sick leave and pensions. Brazilian Journal of Psychiatry**, p. 567–569, nov. 2021. Disponível em: https://doi.org/10.1590/1516-4446-2020-1652. Acesso em: 10 fev. 2025.

[6] MASRI, Ghinwa; AL-SHARGIE, Fares; TARIQ, Usman; ALMUGHAIRBI, Fadwa; BABILONI, Fabio; AL-NASHASH, Hasan. **Mental stress assessment in the workplace: a review. IEEE Transactions on Affective Computing**, v. 15, n. 3, p. 958–976, jul. 2024. Disponível em: https://doi.org/10.1109/TAFFC.2023.3312762. Acesso em: 10 fev. 2025.

[7] SHARMA, Koustubh; SHETTY, Aditya; JAIN, Arnish; DHANARE, Ritesh Kumar. **A comparative analysis on various Business Intelligence (BI), data science and data analytics tools. In: 2021 International Conference on Computer Communication and Informatics**, p. 1–11, jan. 2021. DOI: 10.1109/ICCCI50826.2021.9402226.

[8] PENG, Yun; ZHANG, Yu; HU, Mingzhe. An empirical study for common language features used in Python projects. In: 2021 **IEEE International Conference on**

Software Analysis, Evolution and Reengineering (SANER), p. 24–35, 2021. DOI: 10.1109/SANER50967.2021.00012.

[9] NAGPAL, Abhinav; GABRANI, Goldie. **Python for data analytics, scientific and technical applications**. In: 2019 Amity International Conference on Artificial Intelligence (AICAI), p. 140–145, fev. 2019. DOI: 10.1109/AICAI.2019.8701341.

[10] WATSON, Hugh; WIXOM, Barb. **The current state of business intelligence. Computer**, v. 40, p. 96–99, out. 2007. DOI: 10.1109/MC.2007.331.

[11] SINGH, Amritpal; KHAMPARIA, Aditya; LUHACH, Ashish Kr. **Performance comparison of Apache Hadoop and Apache Spark. In: Proceedings of the Third International Conference on Advanced Informatics for Computing Research (ICAICR '19)**, p. 1–5, jun. 2019. DOI: https://doi.org/10.1145/3339311.3339329.

[12] COELHO, Flávio Codeço; BARON, Bernardo Chrispim; FONSECA, Gabriel Machado de Castro; RECK, Pedro; PALUMBO, Daniela. **AlertaDengue/PySUS: Vaccine**. Zenodo, maio 2021. Versão 0.5.17. Disponível em: https://doi.org/10.5281/zenodo.4883502. Acesso em: 12 fev. 2025.

[13] BANSAL, Srividya; KAGEMANN, Sebastian. **Integrating big data: a semantic extract-transform-load framework. Computer**, v. 48, p. 42–50, mar. 2015. DOI: 10.1109/MC.2015.76.

[14] **Análise do Power BI Embedded | Microsoft Azure**. Disponível em: <https://azure.microsoft.com/pt-br/overview/what-are-business-intelligence-tools/>. Acesso em: 12 fev. 2025.

[15] GARTNER. **Gartner Magic Quadrant for Analytics and Business Intelligence Platforms, 2024**. Acesso em: 10 mar. 2025.

[16] JULCSC. **Tutorial: Introdução à criação no serviço do Power BI - Power BI**. Disponível em: <https://docs.microsoft.com/pt-br/power-bi/fundamentals/service-get-started>. Acesso em: 12 ago. 2025.

[17] **SINANWEB - Funcionamento**. Disponível em: <https://portalsinan.saude.gov.br/funcionamentos>. Acesso em: 10 fev. 2025.

[18] MACEDO, Lucas. **CBO Brasil – Ocupações segundo a Classificação Brasileira de Ocupações (CBO-2002)**. GitHub. Disponível em: https://github.com/lucassmacedo/cbo-brasil/blob/master/csv/CBO2002%20-%20Ocupacao.csv. Acesso em: 10 fev. 2025.

[19] **Códigos dos Municípios | IBGE**. Disponível em: <https://www.ibge.gov.br/explica/codigos-dos-municipios.php>. Acesso em: 10 fev. 2025.