UNIVERSIDADE DO ESTADO DA BAHIA – UNEB
DEPARTAMENTO DE CIÊNCIAS EXATAS E DA TERRA
CURSO DE GRADUAÇÃO EM SISTEMAS DE INFORMAÇÃO

JEVERSANDER CALIXTO DIAS GOMES

SADH - Sistema de Apoio à Decisão Vírus Linfotrópico de Células T Humanas (HTLV) na Bahia

SALVADOR
2024

JEVERSANDER CALIXTO DIAS GOMES

SADH - Sistema de Apoio à Decisão Vírus Linfotrópico de Células T Humanas (HTLV) na Bahia

Monografia apresentada ao curso de Sistemas de Informação do Departamento de Ciências Exatas e da Terra da Universidade do Estado da Bahia - UNEB, como requisito à obtenção do grau de Bacharel em Sistemas de Informação. Área de Concentração: Ciência da Computação.

Orientador: Professor Dr. Hugo Saba Pereira Cardoso
Coorientadora: Professora Dra. Mônica de Souza Massa

SALVADOR

2024

JEVERSANDER CALIXTO DIAS GOMES

SADH - Sistema de Apoio à Decisão Vírus Linfotrópico de Células T Humanas (HTLV) na Bahia

Monografia apresentada ao Curso de Bacharelado em Sistemas de Informação do Departamento de Ciências Exatas e da Terra (DCET) - Campus I, da Universidade do Estado da Bahia (UNEB), como requisito à obtenção do grau de bacharel em Sistemas de Informação. Área de concentração: Ciências da Computação.

Aprovada em: 19/12/2024

BANCA EXAMINADORA

gov.br Documento assinado digitalmente
HUGO SABA PEREIRA CARDOSO
Data: 19/12/2024 20:28:18-0300
Verifique em https://validar.iti.gov.br

Professor Dr. Hugo Saba Pereira Cardoso (Orientador)
Universidade do Estado da Bahia – UNEB

gov.br Documento assinado digitalmente
DEBORA ALCINA REGO CHAVES
Data: 19/12/2024 18:50:19-0300
Verifique em https://validar.iti.gov.br

Professora Me. Débora Alcina Rego Chaves
Universidade do Estado da Bahia – UNEB

gov.br Documento assinado digitalmente
EDUARDO MANUEL DE FREITAS JORGE
Data: 30/12/2024 12:16:37-0300
Verifique em https://validar.iti.gov.br

Professor Dr. Eduardo Manuel de Freitas Jorge
Universidade do Estado da Bahia – UNEB

AGRADECIMENTOS

Dedico este momento aos meus pais, Paulo S. Gomes e Maria de B.D. Gomes, e às minhas três irmãs, Paula M.D. Gomes, Janaina de F.D. Gomes e Jaciara N.D. Gomes, que me ofereceram apoio em cada passo, tanto nas dificuldades quanto nas vitórias. Sou profundamente grato por cada sacrifício e por cada ato de incentivo. Este trabalho é o reflexo de tudo o que aprendi com vocês. Tudo o que sou e conquistei tem um pedaço de vocês. Obrigado por estarem sempre ao meu lado.

Agradeço à Universidade do Estado da Bahia por possibilitar o acesso a uma educação digna e a todos os professores e funcionários pela dedicação, promovendo o desenvolvimento científico por meio de uma formação ampla e transformadora. Agradeço também aos meus colegas por todo o companheirismo proporcionado ao longo do curso.

Agradeço a todos os meus colegas com quem estagiei e trabalhei nesse processo de descobertas, especialmente a Darlan Jesus, que me mostrou inúmeros caminhos na área de TI e sempre me incentivou, não apenas no trabalho, mas também como ser humano, com seus conselhos e ensinamentos.

Agradeço ao meu orientador, Hugo Saba, pela amizade, pela atenção, pelos ensinamentos e pelos conselhos que foram além do projeto. A minha coorientadora, Mônica Massa, e à professora Débora Chaves, pelo caminho percorrido e pelos ensinamentos que contribuíram enormemente para este trabalho.

Por fim, agradeço a todos aqueles que, de alguma forma, contribuíram para que eu pudesse cumprir da melhor maneira possível mais essa etapa da minha vida.

Resumo

Este projeto de pesquisa apresentado nesta monografia consiste no desenvolvimento do sistema SADH - Sistema de Apoio à Decisão Vírus Linfotrópico de Células T Humanas (HTLV), integrando Modelagem Computacional e *Business Intelligence* (BI) para o monitoramento epidemiológico do HTLV na Bahia. Foi realizada uma investigação sobre o HTLV no Estado da Bahia e o uso de BI para apoiar a tomada de decisões. O projeto foi construído utilizando técnicas de Mineração de Dados e *Data Warehouse* (DW), o *Data Science Research* (DSR) foi utilizado para guiar o desenvolvimento do sistema implementado no *Power BI* (PBI) para analisar dados fornecidos pela Secretaria da Saúde do Estado da Bahia (SESAB). A partir de *dashboards* interativos, são realizadas análises detalhadas sobre a distribuição geográfica, demográfica e temporal do vírus, permitindo uma visão aprofundada por faixa etária, gênero e região. Os resultados concluíram que o uso do PBI para a análise de dados na área da saúde é viável, podendo auxiliar na tomada de decisões devido à eficiência na apresentação dos dados, possibilitando um melhor entendimento sobre o impacto do HTLV na população da Bahia.

Palavras-chave: HTLV; Mineração de dados; Business Intelligence; Power BI.

Abstract

This research project presented in this monograph involves the development of the SADH system (Decision Support System for Human T-cell Lymphotropic Virus - HTLV), integrating Computational Modeling and Business Intelligence (BI) for epidemiological monitoring of HTLV in Bahia. An investigation was conducted on HTLV in the State of Bahia and the use of BI to support decision-making. The project was built using Data Mining and Data Warehouse (DW) techniques, with Design Science Research (DSR) employed to guide the development of the system implemented in Power BI (PBI) to analyze data provided by the Health Department of the State of Bahia (SESAB). Through interactive dashboards, detailed analyses are conducted on the geographic, demographic, and temporal distribution of the virus, allowing for an in-depth view by age group, gender, and region. The results concluded that using PBI for data analysis in the health sector is feasible, as it can assist decision-making by efficiently presenting data, enabling a better understanding of the impact of HTLV on Bahia's population.

Keywords: HTLV; Data Mining; Business Intelligence; Power BI.

LISTA DE QUADROS

LISTA DE FIGURAS

LISTA DE GRÁFICOS

LISTA DE TABELAS

LISTA DE ABREVEATURAS E SIGLAS

ATLL	Leucemia de Células T do Adulto
BI	*Business Intelligence*
DAX	*Data Analysis Expressions*
DF	*DataFrame*
DSR	*Design Science Research*
DW	*Data Warehouse*
ELISA	*Enzyme-Linked Immunosorbent Assay*
ETL	*Extract, Transform, Load*
FOFA	Forças, Oportunidades, Fraquezas e Ameaças
GC	*Google Colab*
GPUs	Unidades de Processamento Gráfico
HTLV	Vírus Linfotrópico de Células T Humanas
IC	Iniciação Científica
KDD	*Knowledge Discovery in Databases*
MD	Mineração de Dados
NPAI	Núcleo de Pesquisa Aplicada e Inovação
PBI	*Power BI*
PQ	*Power Query*
SADH	Sistema de Apoio à Decisão HTLV
SESAB	Secretaria da Saúde do Estado da Bahia
SES-RS	Secretaria Estadual de Saúde do Rio Grande do Sul
SS	*Star Schema*
SWOT	*Strengths Weaknesses Opportunities Threats*
TI	Tecnologia da Informação
TPUs	Unidades de Processamento Tensorial
UNEB	Universidade do Estado da Bahia
WB	*Western Blot*

SUMÁRIO

1 INTRODUÇÃO

Este projeto teve início como uma Iniciação Científica (IC), com foco na análise epidemiológica do vírus linfotrópico de células T humanas (HTLV). O objetivo inicial foi descrever a prevalência do HTLV e explorar a eficácia das estratégias de coleta de dados pelos municípios através da criação de um Sistema de Apoio à Decisão Linfotrópico de Células T Humanas (SADH). A partir desse trabalho inicial, o projeto evoluiu para esta monografia completa, sendo feito uma revisão sistemática e um anteprojeto para melhor compreender os problemas abordados.

Os dados epidemiológicos, fundamentais para a descrição da prevalência de diversas doenças específicas, são coletados pelos municípios por meio de seus sistemas de informação, permitindo a avaliação da incidência de enfermidades na população, Araújo (2021). O vírus HTLV está associado doenças e problemas de saúde graves, como mielopatia, leucemia/linfoma de células T do adulto, dermatite e uveíte (Rosadas *et al*., 2021). A transmissão, manifestações clínicas e epidemiologia, a infecção continua sendo um desafio, transmitida por relações sexuais desprotegidas, transfusão de sangue contaminado e transmissão vertical de mãe para filho durante a gravidez ou amamentação (Rosadas *et al*., 2021).

No Brasil, entre 800 mil e 2,5 milhões de pessoas são portadoras do vírus (Ministério da Saúde do Brasil, 2023), com Salvador destacando-se como a área de maior prevalência devido à transmissão, Amoussa (2018). Com o avanço do projeto, a mineração de dados se mostrou uma estratégia promissora para preencher diversas lacunas, como a falta de conhecimento tanto entre os profissionais de saúde quanto na população em geral, o que limita a conscientização sobre o vírus. Além disso, há uma insuficiência de dados epidemiológicos, que dificulta uma análise mais detalhada da disseminação do HTLV. Outro ponto crítico é a ausência de políticas públicas efetivas no Brasil. A mineração de dados oferece informações essenciais, ajudando a superar essas limitações e a melhorar a compreensão do HTLV.

A Mineração de Dados (MD) ao descobrir padrões em grandes conjuntos de dados, fornece *insights* valiosos para a tomada de decisões (Patrício; Magnoni, 2018). Com o aumento constante na geração de dados, é fundamental utilizar técnicas e algoritmos que possam explorar esses dados brutos em busca de informações importantes, conforme destacado por Patrício e Magnoni (2018). O *Data Warehouse* (DW) consolida esses dados em informações, facilitando a análise com dados integrados. Segundo Esteves (2016), o processo Extração, Transformação e Carregamento (ETL) garante a qualidade e consistência dos dados antes de sua inserção no

DW. A organização desses dados é feita com o uso de modelos dimensionais, como o *Star Schema* (SS), promovendo consultas rápidas e eficientes. A implementação de soluções de *Business Intelligence* (BI) aproveita essa estrutura organizada, permitindo uma análise completa de dados e transformando-os em informações úteis.

Com o desenvolvimento da tecnologia, várias áreas do conhecimento e diversas organizações passaram a adotar soluções de BI, que, segundo Brito e Oliveira (2017), são utilizadas para reunir e analisar grandes volumes de dados estruturados ou não estruturados provenientes de diversas fontes internas e externas. Segundo Chen, Chiang e Storey (2012), o BI oferece tecnologias e sistemas para transformar esses dados em *insights* estratégicos, facilitando a tomada de decisões através de visualizações interativas, como gráficos e painéis. Rodrigues (2021) destaca que o *Power BI* (PBI) se sobressaiu em comparação a outras ferramentas de BI, apresentando excelência em funcionalidades críticas, como preparação de dados, painéis interativos e análises avançadas. Com isso, foi aplicado uma metodologia que envolve teoria e prática.

A metodologia *Design Science Research* (DSR) foi aplicada para estruturar o desenvolvimento dos artefatos, promovendo o avanço do conhecimento científico desta monografia. Essa abordagem assegura a confiabilidade dos resultados, facilita a replicação do estudo e contribui para a consistência das conclusões segundo Peffers *et al*. (2007). O processo segue várias etapas, incluindo a identificação do problema, a definição de carências, o desenvolvimento de artefatos, a demonstração da funcionalidade, a avaliação da estrutura de dados e a comunicação dos resultados, como pode ser observado na seção 3.

1.1 Pergunta da pesquisa

Atualmente, os dados brutos disponibilizados pela Secretaria da Saúde do Estado da Bahia (SESAB) não oferecem uma visão clara sobre o real problema do vírus HTLV. De que maneira a mineração de dados, por meio da elaboração de cenários epidemiológicos, pode auxiliar na compreensão da disseminação do HTLV e apoiar a tomada de decisão?

1.2 Justificativa

A saúde pública é importante para o bem-estar da sociedade, e compreender os problemas de saúde é fundamental para uma gestão pública eficiente. Segundo a revisão

sistemática e o anteprojeto realizada no início deste trabalho, os sistemas de saúde pública enfrentam desafios na tomada de decisões complexas em prazos limitados. O uso de ferramentas de suporte, como a BI e suas ferramentas, torna-se essencial para mitigar potenciais erros, permitindo análises rápidas e eficazes de dados.

Pesquisas na Bahia ressaltam a importância de compreender o comportamento das epidemias para combatê-las de forma eficaz, por exemplo, o estudo de Filho, Murari, Saba e Moret (2021) "Análise espaço-temporal da propagação da dengue em uma região de clima seco no Brasil", investigou a propagação da dengue na Bahia e sua correlação em 15 regiões econômicas. Utilizando diversas análises estatísticas, examinou o impacto da disseminação da infecção na região metropolitana de Salvador. Os resultados destacaram uma correlação persistente e significativa entre as regiões, indicando fatores econômicos ou climáticos. Esses achados contribuíram para a compreensão da disseminação da dengue em regiões de clima seco.

A construção de cenários epidemiológicos como os Dados Gerais, Dados HTLV, Tendencia Temporal e Razão/Sexo e Gestantes proporcionou uma visão extensa do panorama do problema de saúde pública, incluindo os municípios, as categorias de raça e cor, e o gênero mais afetado. Isso permite que as autoridades identifiquem tendências, antecipem surtos e ajam proativamente. Esse entendimento aprofundado pode ser fundamental para reduzir a contaminação do vírus e otimizar a utilização dos recursos disponíveis.

1.3 Objetivo geral

Desenvolver *dashboards* de BI utilizando dados epidemiológicos fornecidos pela SESAB para apoiar a tomada de decisões relacionadas à contaminação pelo vírus HTLV no Estado da Bahia.

1.3.1 Objetivo específico

I. Apresentar uma proposta para a construção da solução orientada pelos conceitos definidos na metodologia e a geração de cenários.
II. Coletar dados referentes ao HTLV.
III. Definir os métodos necessários para implementar um modelo que compreenda a extração, tratamento, análise e visualização de dados.

IV. Carregar a base de dados e estabelecer a estrutura de relacionamento entre os dados e seus indicadores presentes no DW.
V. Desenvolver e analisar um modelo de representação de alto nível e realizar a modelagem no PBI.

1.4 Estrutura do Trabalho

Este trabalho está estruturado em seis capítulos. O Capítulo 1 apresenta a introdução, contextualizando o tema abordado, definindo a questão de pesquisa, justificando sua relevância e estabelecendo os objetivos gerais, específicos deste estudo e metodologia. O Capítulo 2 traz a fundamentação teórica, elaborada por meio da revisão sistemática e do anteprojeto, explorando os temas HTLV, MD, DW e SS, BI, além de abordar trabalhos correlatos que sustentam esta monografia. No Capítulo 3, detalha-se a descrição do projeto, centrada na metodologia DSR, expondo suas etapas e processos aplicados, assim como os materiais, ferramentas e métodos utilizados. O Capítulo 4 descreve a análise de resultados, focando na interpretação dos dados coletados para avaliar a eficácia dos dados implementados e auxiliar na tomada de decisões futuras. O Capítulo 5 apresenta as considerações finais, enquanto o Capítulo 6 propõe as sugestões para trabalhos futuros. Por fim, são apresentadas as referências utilizadas ao longo do projeto.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo, são apresentados os fundamentos teóricos necessários para a compreensão deste projeto, que servem como base e guia para alcançar os objetivos propostos. O capítulo está dividido nas seções HTLV, MD, DW e SS, BI e artigos correlacionados.

2.1 Human T Iymphotropic Virus

O vírus linfotrópico da célula T humana (HTLV), inicialmente descoberto nos EUA na década de 1980, derivou do isolamento em um paciente afro-americano com leucemia/linfoma de células T de adultos, de acordo com Amoussa (2018). Casos similares já haviam sido identificados no Japão em 1977, denominados Leucemia de Células T do Adulto (ATLL), posteriormente compreendidos como sendo o mesmo vírus, o HTLV-I, desde então, avanços significativos no entendimento da sua transmissão e manifestações clínicas têm sido alcançados (Amoussa, 2018)

Como o pioneiro retrovírus humano associado a doenças ontogênicas. A identificação do vírus somente se tornou viável em 1983, quando foram introduzidos testes sorológicos para avaliar sua disseminação (Amoussa, 2018). No Brasil, esses testes foram adotados em 1993 e passaram a ser de cumprimento obrigatório em todas as instituições de bancos de sangue, conforme diretrizes do Ministério da Saúde do Brasil (2023).

O vírus HTLV-I, assim como o HTLV-II, são associados a doenças ou infecções como mielopatia, essa condição é caracterizada por fraqueza progressiva nas pernas, distúrbio do controle muscular em membros inferiores e dificuldades de locomoção como a leucemia/linfoma de células T do adulto é um tipo de câncer relacionado à infecção pelo vírus, afetando as células T, a dermatite pode estar relacionada a certas condições de pele, e a uveíte associada a distúrbios oculares (Rosadas *et al*., 2021). Variantes adicionais, como o HTLV-III e HTLV-IV, foram identificadas em áreas remotas das florestas de Camarões, na África Central, embora não tenham sido associadas a manifestações clínicas (Rosadas *et al*., 2021).

De acordo com Rosadas *et al*. (2021), a infecção pode ocorrer por várias vias, como relações sexuais desprotegidas, exposição a sêmen e secreção vaginal, transfusão de sangue contaminado, compartilhamento de seringas, transplante de órgãos e transmissão vertical da mãe para o filho durante a gravidez, parto ou a amamentação. Não há diferenças significativas na transmissão entre homens e mulheres.

Nunes da Silva *et al.* (2023), indica que Salvador é a cidade com o maior número de indivíduos infectados pelo HTLV-I no Brasil, onde a predominância de transmissão é feita sexualmente. Enquanto Amoussa (2018) através de estudos em doadores de sangue, revelou uma distribuição heterogênea da infecção pelo país, com ênfase na identificação de Salvador como a região com a maior incidência de indivíduos infectados no território brasileiro. Esses dados apontam para uma possível epidemia na Bahia, uma vez que a epidemia ocorre quando há um aumento significativo no número de casos de uma doença em várias regiões, estados ou cidades, mas não atinge níveis globais (Butantan, 2023). As epidemias apresentam desafios significativos, pois afetam não apenas a saúde das pessoas, mas também têm impactos sociais, econômicos e até políticos.

Estima-se que no Brasil haja entre 800 mil e 2,5 milhões de pessoas vivendo com o vírus, embora esse número possa estar subestimado, uma vez que o diagnóstico ocorre somente durante a doação de sangue (Ministério da Saúde do Brasil, 2023).

Aproximadamente 90% dos portadores do HTLV permanecem assintomáticos ao longo de suas vidas, contribuindo para uma rede de transmissão silenciosa, segundo Gonçalves *et al.* (2010), podendo não resultar em problemas de saúde graves. No entanto, entre 5% e 10% das pessoas infectadas podem desenvolver doenças graves (Gonçalves *et al.*, 2010).

Conforme apontado por Rosadas *et al. (*2021), embora a infecção pelo HTLV seja reconhecida há várias décadas, sua compreensão permanece limitada tanto para a população em geral quanto para os profissionais da área da saúde, o que, consequentemente, impede a formulação de políticas de saúde pública eficazes. A utilização de mineração de dados e da modelagem de dados emerge como uma estratégia promissora para suprir essa deficiência, ao oferecer informações essenciais destinadas à formulação de políticas de saúde pública mais eficazes.

2.2 Mineração de Dados

Segundo Fayyad, Piatetsky-Shapiro e Smyth (1996), a mineração de dados faz parte da jornada de descoberta de conhecimento, envolvendo a análise detalhada dos dados e o uso de algoritmos específicos para encontrar padrões dentro de conjuntos de dados, levando em conta limitações computacionais.

Diante do crescente volume de dados gerados, é essencial a aplicação de técnicas e algoritmos capazes de investigar esses dados em busca de informações relevantes,

especialmente para apoiar a tomada de decisões segundo Patrício e Magnoni (2018). Entre as técnicas destaca-se a MD, que, por meio da extração e análise de padrões em grandes repositórios de dados, permitem transformar essas informações em conhecimento útil, validando e apresentando, onde são convertidos em informações validadas pelo usuário (Patrício; Magnoni, 2018). A Figura 1 mostra como a mineração de dados interage no processo de *Knowledge Discovery in Databases* (KDD), ou em português, Descoberta do Conhecimento em Banco de Dados.

Figura 1– KDD Processos

Fonte: Adaptado de Fayyad, Piatetsky-Shapiro e Smyth (1996)

Os processos de KDD são divididos em etapas:

Seleção dos dados: a seleção de fontes relevantes para a análise. Dados podem ser obtidos de múltiplas origens, como bancos de dados, registros de clientes, transações comerciais e sensores, vídeos entre outros.

Pré-Processamento: após a coleta, os dados são limpos e organizados para manter a qualidade e precisão da análise. Removendo ruídos, valores ausentes ou inconsistentes, requerendo remoção de valores atípicos e normalização.

Transformação de dados: é feita a exploração e análise dos dados. Técnicas são usadas para encontrar padrões e relações nos dados, facilitando sua compreensão.

Mineração dos dados: esta fase consiste em identificar padrões e gerar diversos resultados, os quais serão analisados na etapa subsequente.

Interpretação de dados: modelos são criados para prever e classificar informações.

A mineração de dados, como parte essencial do processo de KDD, permite identificar padrões e *insights* valiosos a partir de grandes volumes de dados. No entanto, para que essa análise seja eficaz, é necessário que os dados sejam devidamente organizados e preparados. Nesse contexto, entra o papel crucial do DW e do processo de ETL.

2.3 Business Intelligence

Segundo Brito e Oliveira (2017) as soluções de BI são *softwares* destinados a reunir e analisar extensos conjuntos de dados não estruturados ou estruturados provenientes de fontes internas e externas, tais como documentos, registros, imagens, arquivos, e-mails e vídeos, entre outras fontes de dados de dados brutos. Apesar de apresentarem uma flexibilidade relativamente menor em comparação às ferramentas de análise de negócios, as ferramentas de BI oferecem um método sistemático para a agregação de dados visando à descoberta de informações, predominantemente por meio de consultas.

O BI, conforme Chen, Chiang e Storey (2012), tornou-se uma forma essencial para corporações ao oferecer um conjunto de técnicas, tecnologias e sistemas voltados para a recuperação, análise e conversão de grandes volumes de dados em *insights* estratégicos. Essas análises geralmente são apresentadas por meio de visualizações interativas, como tabelas, gráficos e painéis, o que facilita a interpretação e a tomada de decisões. As principais ferramentas de BI incluem a visualização de dados, os DW, painéis interativos e ferramentas de relatórios, proporcionando às empresas uma compreensão mais aprofundada de seus negócios e mercados (Chen; Chiang; Storey, 2012). Além disso, o BI tem se expandido para novas áreas, como a análise de mídias sociais, big data (Chen; Chiang; Storey, 2012) e outras áreas, fortalecendo assim a sua importância.

Rodrigues (2021), explora diversas ferramentas de BI para atender às necessidades específicas das organizações, desde soluções de BI de autoatendimento, destinadas a usuários corporativos, até ferramentas especializadas em visualização e armazenamento de dados. Algumas delas oferecem um conjunto completo de funcionalidades, que incluem desde o armazenamento e a limpeza de dados até a visualização e publicação dos resultados (Rodrigues, 2021).

Baseado nas métricas definidas pelo *Gartner*, que avaliam aspectos como: serviço e suporte, qualidade do suporte técnico, experiência de vendas (incluindo preços e flexibilidade contratual) e capacidade de resposta do mercado (valor percebido), de acordo com Rodrigues

(2021). Com base nesses critérios, o *Gartner Group* identificou 12 capacidades críticas que serviram como referência para a análise das ferramentas de BI e *Analytics* (Rodrigues, 2021), conforme apresentado no Quadro 1.

Quadro 1 - As 12 Capacidades Críticas que o *Gartner Group*

Item	Capacidades Críticas	Descrição
1	Segurança	Requisito crítico que engloba desde a robustez da plataforma em suportar grandes volumes de dados ao gerenciamento e administração de usuários e níveis de acesso e autenticação.
2	Capacidade de gerenciamento	Esse requisito aborda os recursos que rastreiam o uso do BI e o compartilhamento dos dados de forma granular entre os usuários.
3	Análise na nuvem	Requisito que representa a capacidade que o BI tem de realizar todo processo de análise na nuvem.
4	Conectividade da fonte de dados	São os recursos que possibilitam a conexão, consulta e coleta de dados de forma otimizada proporcionando uma melhor performance.
5	Preparação de dados	Requisitos pertinentes ao trabalho com dados oriundos de várias fontes de dados, com critérios estabelecidos pelo usuário. Possibilitando criação de modelos analíticos com medidas, agrupamentos e hierarquias personalizadas.
6	Catálogo	Requisito referente à criação de um catálogo de acesso, a fim de viabilizar a pesquisa dos conteúdos disponíveis facilitando todo o processo.
7	*Insights* automatizados	Requisito que viabiliza que os usuários desenvolvam a capacidade analítica promovendo *insights* através do uso de modelos de *Machine Learning*.
8	Visualização	Requisito que avalia o nível de interação dos usuários com os *dashboards* e a capacidade de promover cenários e gráficos interativos e dinâmicos.
9	*Storytelling*	Requisito responsável por promover e apresentar os resultados da análise do BI para os gestores e tomadores de decisão. Dessa forma, combinar a visualização de dados com uma narrativa adequada à compreensão.
10	Consultas em Linguagem Natural	Requisito que permite fazer perguntas, digitando textos ou falando e, em seguida, o BI retornar as respostas de forma gráfica.
11	Geração de Linguagem Natural	Requisito que promove respostas automáticas através de gráficos analíticos. Possibilitando um usuário interagir com os dados, e obter as respostas em linguagem natural.
12	Relatórios	Requisito que define a capacidade de criar e compartilhar os relatórios de múltiplas páginas de forma organizada.

Fonte: Rodrigues (2021).

Com base nos 12 critérios estabelecidos, Rodrigues (2021) realizou uma análise descritiva em 20 das principais ferramentas de BI e *Analytics* disponíveis do mercado. No relatório de 2021 do *Gartner*, foi realizada uma análise detalhada das principais ferramentas de BI e *Analytics*, a análise categorizou as ferramentas em quatro níveis: excelente, bom, médio e ruim, conforme os critérios pré-definidos (Rodrigues, 2021). A legenda de classificação é a seguinte:

- **Excelente**: representado pela cor verde, para ferramentas que atenderam ao requisito com excelência ou inovação.
- **Bom**: representado pela cor azul, para ferramentas que cumpriram o requisito de forma satisfatória.
- **Médio**: representado pela cor amarela, para ferramentas que atenderam parcialmente ao requisito.
- **Ruim**: representado pela cor vermelha, para ferramentas que não atenderam ao requisito ou não possuem a funcionalidade.

De acordo com a análise de Rodrigues (2021), o PBI destacou-se em comparação a outras ferramentas de BI, sendo melhor que as demais ferramentas analisadas em seu estudo. A plataforma demonstrou excelência nas 12 capacidades críticas definidas pelo *Gartner*, oferecendo um conjunto abrangente e diversificado de funcionalidades, como preparação de dados, descoberta visual, painéis interativos e análises avançadas. Esses recursos conferem ao PBI uma vantagem competitiva em relação às demais soluções disponíveis no mercado.

Abaixo, o Quadro 2 apresenta os resultados obtidos, destacando a classificação de cada ferramenta conforme os critérios estabelecidos:

Quadro 2 - Comparativo das ferramentas Business Intelligence

Fonte: Rodrigues (2021).

2.4 Data Warehouse e Star Schema

O DW é um sistema projetado para consolidar e integrar informações provenientes de diversas fontes em uma estrutura unificada. Esse processo facilita a análise e a tomada de decisões com base em dados, proporcionando uma visão consolidada e simplificada das informações (Esteves, 2016).

O processo de construção e análise de dados será fundamentado na etapa de ETL ou extração, transformação e carregamento em português. Conforme descrito por Esteves (2016), o ETL abrange a extração de dados de múltiplas fontes, sua transformação e limpeza para garantir a qualidade e consistência, e, finalmente, o carregamento desses dados em um DW.

Para a organização dos dados dentro do DW, utilizaremos modelos dimensionais que promovem uma análise eficiente. O SS, organiza os dados em torno de uma tabela central de fatos, que armazena os dados quantitativos, e tabelas dimensionais conectadas a essa tabela. As tabelas dimensionais fornecem contexto para os dados na tabela de fatos, Esteves (2016). A estrutura direta e eficiente do SS facilita consultas rápidas e uma navegação intuitiva, tornando-o uma escolha ideal para muitos projetos de DW.

2.5 Trabalhos correlatos e lacunas

Durante a elaboração deste trabalho, foi realizada uma abrangente análise de artigos acadêmicos e projetos relacionados, a fim de fornecer uma sólida estrutura teórica que possibilite a identificação de lacunas e trabalhos pertinentes, validando as novas contribuições presentes nesta monografia.

Miranda *et al.* (2022), utilizou uma análise *Strengths Weaknesses Opportunities Threats* (SWOT), também conhecido como Forças, Oportunidades, Fraquezas e Ameaças (FOFA) para identificar os desafios no controle do HTLV-I no Brasil. Entre os principais problemas destacados, está a falta de conhecimento e conscientização sobre a infecção. Além disso, a insuficiência de dados epidemiológicos atualizados e abrangentes sobre a prevalência do HTLV-1 dificulta a criação de políticas públicas eficazes.

Garcia e Hennington (2021), examinaram as ações governamentais no combate ao HTLV nos Estados da Bahia e Minas Gerais, apontando a falta de conhecimento por parte dos profissionais de saúde e da população em geral sobre o HTLV, além da insuficiência de ações educativas e preventivas eficazes.

Mai *et al.* (2017), analisaram a eficácia do Portal BI da Saúde no apoio à tomada de decisões na Secretaria Estadual de Saúde do Rio Grande do Sul (SES-RS). Através de painéis interativos e consultas personalizadas, o BI facilitou a organização de dados fragmentados e contribuiu para decisões mais informadas. Esse estudo demonstra a importância da integração de dados no processo decisório. A principal lacuna identificada é a fragmentação dos sistemas de informação em saúde, que dificulta a integração de dados e prejudica a análise e o planejamento eficaz.

A pesquisa de Rosadas *et al.* (2021), "Protocolo Brasileiro para Infecções Sexualmente Transmissíveis 2020: infecção pelo vírus linfotrópico de células T humanas (HTLV)", aborda a inclusão nas políticas públicas. Descrevendo a gravidade das complicações associadas ao HTLV e a falta de uma maior inclusão do HTLV na agenda governamental de saúde pública, especialmente no Brasil, representa uma lacuna significativa.

Rosadas *et al.* (2022), no estudo "*We Need to Translate Research Into Meaningful HTLV Health Policies and Programs*" descreve a falta de conscientização sobre o HTLV-I e II entre a população e profissionais de saúde. A ausência de dados epidemiológicos completos, especialmente na América Latina, dificulta a criação de estratégias de saúde baseadas em evidências. Sem registros confiáveis sobre a prevalência do vírus, a identificação precoce e o

tratamento adequado são limitados. A ausência de políticas públicas e globais específicas reflete a falta de ação política coordenada para combater a infecção.

A Tabela 1 apresenta um resumo das contribuições dos trabalhos correlatos e das lacunas identificadas para o desenvolvimento deste projeto de pesquisa.

Tabela 1 – Contribuições de Trabalhos Correlatos e Lacunas para o Desenvolvimento desta Pesquisa

Nome	Ano	Proposta do estudo	Lacunas Identificadas	Relação com esta Pesquisa
Miranda *et al.*	2022	Identifica desafios no controle do HTLV-I no Brasil.	Falta de conhecimento e conscientização; insuficiência de dados epidemiológicos.	Fundamenta a necessidade de reduzir a falta de conhecimento sobre o HTLV e criar cenários epidemiológicos.
Garcia e Hennington	2021	Examina ações governamentais na Bahia e Minas Gerais no combate ao HTLV.	Falta de conhecimento por parte dos profissionais de saúde e população.	Reforça a importância da integração de dados para criar cenários epidemiológicos que apoiem a tomada de decisões.
Mai *et al.*	2017	Analisa o uso do Portal BI na Secretaria de Saúde do Rio Grande do Sul para integração e tomada de decisões.	Fragmentação dos sistemas de informação em saúde, que dificulta a integração e o planejamento.	Destaca o uso do BI e PBI no SADH para unificar e analisar dados. Diminuir a fragmentação dos sistemas e facilitar as tomadas de decisões estratégicas.
Rosadas *et al.*	2021	Aborda a inclusão do HTLV nas políticas públicas e as complicações associadas ao vírus.	Falta de maior inclusão do HTLV na agenda governamental de saúde pública no Brasil.	Corrobora a importância de uma ferramenta para analisar, evidenciar a gravidade do HTLV e apoiar a tomada de decisões.
Rosadas *et al.*	2022	Discute a ausência de conhecimento sobre HTLV-I e II, especialmente na América Latina.	Dificuldade de criar estratégias de saúde devido à ausência de registros sobre a prevalência do vírus.	Sustenta o uso de um sistema para integrar dados confiáveis e facilitar estratégias baseadas em evidências para o HTLV.

Fonte: O Autoria própria (2024)

3 DESCRIÇÃO DO PROJETO

Neste capítulo, são descritos a metodologia adotada e os softwares empregados neste estudo, assim como as etapas detalhadas de sua implementação.

3.1 Metodologia

Um fator fundamental para a obtenção dos objetivos deste trabalho está relacionado à definição da metodologia para estruturar a coleta e análise de dados relacionados ao vírus HTLV. O *Design Science Research*, segundo Peffers *et al.* (2007), é direcionado à criação e avaliação de artefatos, como sistemas, métodos e modelos, para solucionar problemas complicados do mundo real. Essa abordagem é bastante utilizada em estudos na área de Tecnologia da Informação (TI), devido à sua capacidade de integrar pesquisa teórica e prática, promovendo a solução de desafios reais por meio da criação de artefatos inovadores.

A metodologia DSR forneceu a estrutura necessária para a coleta, análise e interpretação dos dados. Além disso, a utilização do DSR garantiu a validade e a confiabilidade dos resultados, facilitando a replicação do estudo e contribuindo para a consistência das conclusões. Isso ajudou a orientar a execução do projeto, assegurando a integridade científica e a relevância dos achados. No DSR, é importante validar e avaliar se o artefato é útil e eficaz para seus objetivos práticos.

3.2 Ambiente Experimental

O ambiente experimental utilizado para o desenvolvimento do ciclo experimental, baseado no DSR é composto por um conjunto de ferramentas e tecnologias configuradas para realizar a análise de dados e a construção dos *dashboards* interativos. As características do ambiente incluem:

- **Hardware:**
  - Processador: *Intel*(R) *Core* (TM) i5-1035G1 CPU @ 1.00 GHz, com frequência máxima de 1.19 GHz, com 4 núcleos.
  - GPU: *Intel(R)UHD Graphic.*
  - Memória RAM: 8 GB.
  -Sistema Operacional: *Windows 11 Pro* (64 *bits*), versão 23H2.

- **Ferramentas de Software:**

  - *Microsoft Power BI:* Versão 2.138.782.0 (64 *bits*).
  - *Microsoft Excel*: *Professional Plus* 2024, versão 2408.
  - *Google Colab:* Versão grátis.

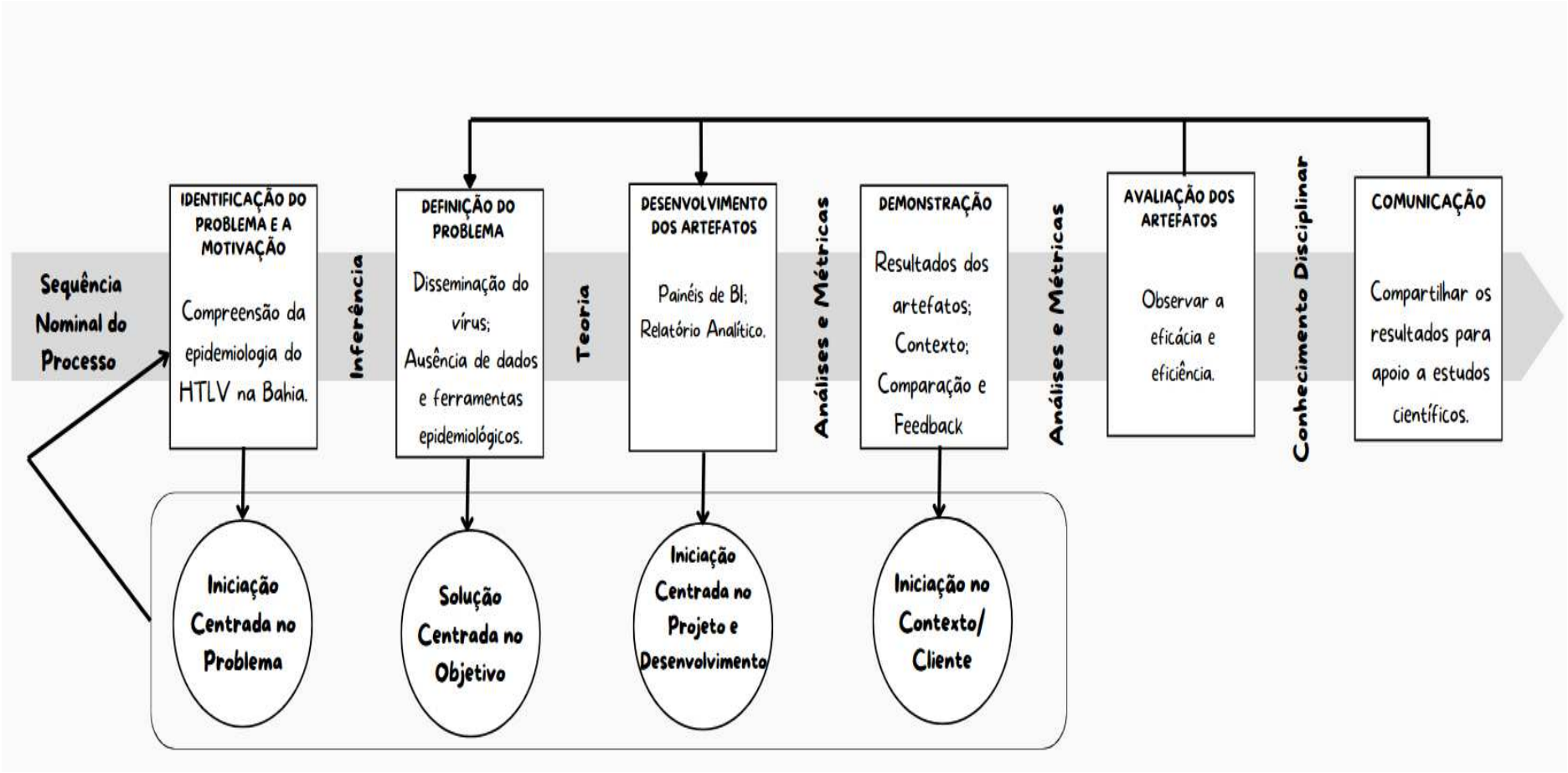
O ambiente foi configurado para permitir a integração entre as ferramentas, com foco em análises exploratórias, modelagem de dados, criação de relatórios interativos e execução de *scripts* em *Python* e em funções DAX para mineração de dados.

3.3 Ciclo de Experimental

O ciclo de experimentos foi conduzido de forma iterativa para garantir a consistência dos resultados obtidos, por meio de testes e análises controladas. O processo de análises dos resultados consistiu em comparar os dados extraídos do banco de dados *Microsoft Excel* com os resultados gerados pelas operações realizadas em *Python* e PBI. O objetivo era assegurar a integridade e consistência dos dados entre as plataformas, visando gerar cenários epidemiológicos para apoiar a gestão pública no combate à difusão do HTLV no Estado da Bahia. Segundo Pressman (2011), o teste de consistência busca garantir que as saídas de software sejam compatíveis com os resultados esperados.

A Figura 2 está representada a metodologia de investigação DSR, adaptada ao projeto SADH. Ela ilustra as diversas etapas interligadas que resumem os passos a que foram seguidos por meio da DSR a criação e desenvolvimento de artefatos.

Figura 2 – Etapas do DSR



Fonte: Adaptada de Peffers *et al.* (2007)

Seguindo esses princípios, para alcançar os objetivos propostos com o DSR, foram adotadas as seguintes etapas:

Na Etapa 1, a identificação do problema e a motivação, foi identificado a necessidade de uma melhor compreensão epidemiologia do HTLV na Bahia. Dados preliminares indicaram a falta de ferramentas de cenários epidemiológicos adequados para analisar a distribuição geográfica e temporal da infecção.

Na Etapa 2, definição do problema, foram identificadas carências significativas, como as dificuldades na avaliação precisa da disseminação do vírus no território e a ausência de dados epidemiológicos essenciais para uma compreensão mais aprofundada da situação. Em resposta a essas lacunas, identificadas na seção 2.5 Trabalhos correlatos e lacunas, iniciou-se o planejamento para estabelecer os objetivos do artefato a ser desenvolvido, bem como suas características e funcionalidades. Esse planejamento baseou-se na análise dos dados do banco de dados a ser utilizado.

Na Etapa 3, o desenvolvimento dos artefatos, foram criados logo após a coleta e o tratamento dos dados. Utilizou-se o PBI para realizar a iteração por meio da técnica de ETL, permitindo a criação de dashboards interativos que asseguram a visualização facilitada de dados sobre o HTLV. O SADH inclui os painéis com dados gerais sobre a coleta de dados (Dados Gerais), os dados específicos sobre o HTLV (Dados HTLV), os dados de tendências temporais (Tendência Temporal) e dados sobre a razão entre sexos e gestantes (Razão/Sexo e Gestantes).

Na Etapa 4, a demonstração, busca garantir a viabilidade e a funcionalidade dos artefatos desenvolvidos na análise e compreensão dos dashboards.

Na etapa 5, avaliação dos artefatos, envolve testar e analisar a estrutura das relações entre tabelas, a precisão das medidas calculadas e dos componentes do banco de dados, a clareza e a eficácia apresentadas, para garantir que as informações fossem passadas de forma eficiente.

Por fim, na etapa 6, comunicação, os resultados que foram detalhados e compartilhados no Núcleo de Pesquisa Aplicada e Inovação (NPAI), no repositório da Universidade do Estado da Bahia (UNEB) Saber Aberto e no GitHub pessoal do autor, contribuindo para o avanço da área de pesquisa.

A iteração, é o processo de revisão e refinado conforme necessário, aprimorando o artefato diante de novos desafios ou descobertas.

3.4 Matérias, Ferramentas e Métodos

Nesta seção, são descritos os materiais, ferramentas e métodos utilizados no estudo, com foco na estruturação e análise dos dados. A organização dos dados foi realizada utilizando o *Microsoft Excel*, que serviu para armazenar os dados, a visualizar e a estruturação inicial. Para a análise e manipulação dos dados, foi empregado o *Python*, executado no ambiente de desenvolvimento *Google Colab* (GC).

A partir dessas análises, os resultados foram carregados, interpretados e apresentados de forma interativa e iterativa por meio da linguagem DAX (*Data Analysis Expressions*) no PBI, uma plataforma de BI, possibilitando a criação de relatórios dinâmicos e visualizações interativas, facilitando a compreensão e interpretação dos dados por meio de gráficos e painéis personalizados.

3.4.1 Material

O material usado será a base de dados, que consiste em um conjunto de dados estruturados fornecidos pela SESAB. Contendo informações cruciais, como paciente, idade, tipoidade, sexo, idadgestacional, nacionalidade, racacor, etnia, bairro, cep.de.residencia, ibge.municipioderesidencia, estadoderesidencia, pais.de.residencia, zona, nome.da.pesquisa, exame, metodologia, ano_cadastro, mes_cadastro, mês_cadastoptexto, dia_cadastro, datacadastro, ano_coleta, mes_coleta, mês_coletatexto, dia_coleto, datadacoleta, datadaliberacao, resultado, anocad, anolib, resultadowb, mesorregiao, microrregiao, municipio, nucleoregional e regiaodesaude, organizadas de forma sistemática.

3.4.2 Ferramentas

I. Microsoft Excel

O *Microsoft Excel*, de acordo com Brito e Oliveira (2017), é um *software* utilizado em ambientes corporativos, sendo também adotado por profissionais e acadêmicos da área da informação para atividades de análise de dados. A ferramenta se destaca como uma solução de produtividade, com foco principal na visualização e armazenamento de dados em formato de

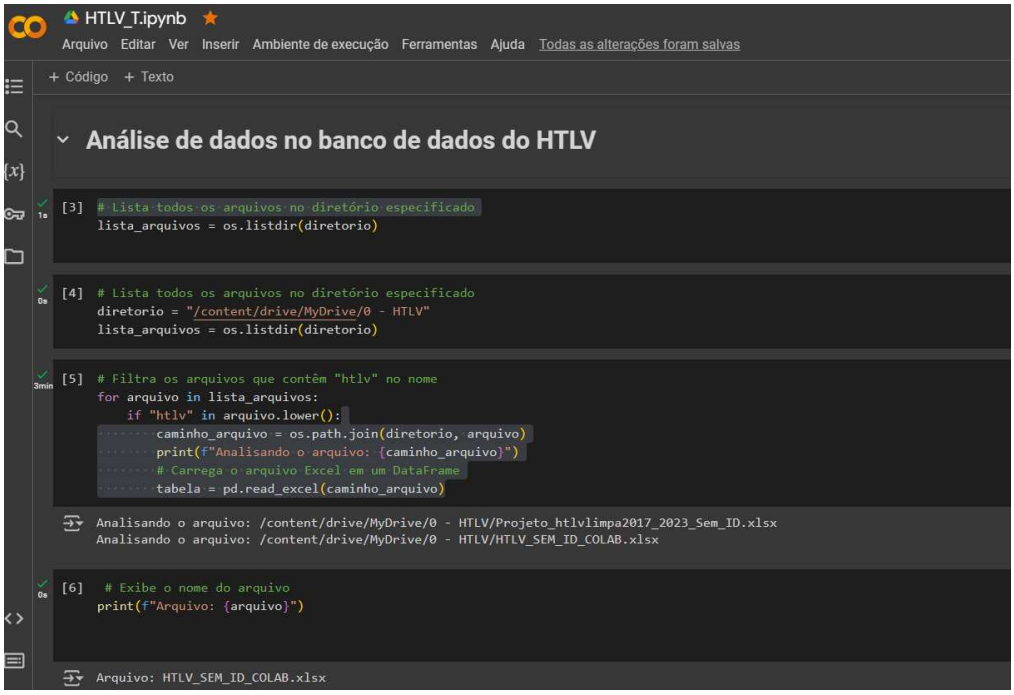
planilhas, permitindo a conexão com fontes de dados como *Analysis Services*, *SQL Server* e ferramentas de BI, aproveitando seus recursos para gerar *insights* que auxiliam na tomada de decisões.

II. Google Colab

O GC ou *Colaboratory* é um ambiente de desenvolvimento, uma ferramenta baseada em nuvem que permite executar códigos *Python* diretamente no navegador (Google, 2024). Muito utilizado para o desenvolvimento, especialmente em áreas como aprendizado de máquina, ciência de dados e inteligência artificial, facilitando a integração de código, visualizações de gráficos e explicações textuais em um único documento (Google, 2024). Oferecendo recursos computacionais avançados, como as Unidades de Processamento Gráfico (GPUs) que são processadores projetados para renderizar gráficos, mas também são utilizados em aprendizado de máquina e cálculos complexos devido à sua capacidade de processamento paralelo e as Unidades de Processamento Tensorial (TPUs), desenvolvidas pelo *Google*, que são otimizadas para acelerar tarefas de aprendizado de máquina, proporcionando desempenho superior no treinamento e na inferência de modelos de inteligência artificial, possibilitando a realização de tarefas que exigem grande poder computacional sem a necessidade de uma infraestrutura local avançada (Google, 2024).

O Código da Figura 13 irá listar todos os arquivos no diretório relacionado, filtrando os arquivos que contenham a palavra HTLV, para posteriores análises.

Figura 3 – Google Colab



Fonte: O Autoria própria (2024)

III. Python

O *Python Institute* (2023), é uma linguagem de programação interpretada, de alto nível e orientada a objetos, amplamente empregada em diversas áreas, incluindo análise de dados. Sua sintaxe clara a torna ideal tanto para iniciantes quanto para programadores experientes, sendo aplicada em desenvolvimento web, automação, inteligência artificial, computação científica e análise de dados. Assim, na área de dados, *Python* se destaca por suas poderosas bibliotecas, que facilitam a manipulação, visualização e análise de grandes volumes de informações (Python Institute, 2023).

IV. Power BI

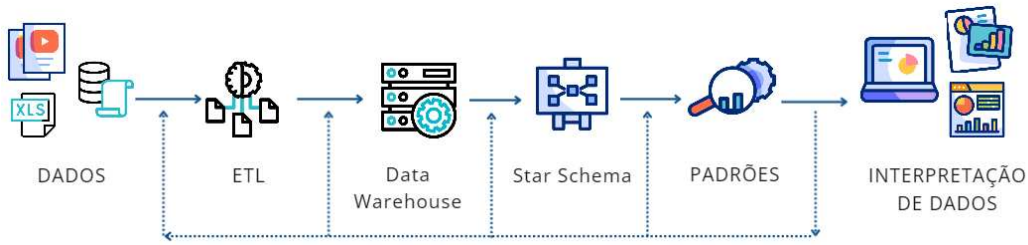
De acordo com Brito e Oliveira (2017), O PBI é uma ferramenta de BI que proporciona visualizações dinâmicas, permitindo aos usuários elaborar relatórios e *dashboards* sem exigir um conhecimento aprofundado em banco de dados. O PBI integra um conjunto de serviços de software, aplicativos e conectores que, em conjunto, transformam diversas fontes de dados independentes em informações coesas, visualmente atraentes e interativas (Microsoft, 2023).

A plataforma é subdividida em três versões: PBI *Desktop*, para computadores locais; um serviço SaaS, que permite a criação e compartilhamento de *insights* de negócios online; e

aplicativos móveis compatíveis com *Windows*, iOS e *Android* (Microsoft, 2023). Entre os principais recursos do PBI, conforme Brito e Oliveira (2017), destaca-se a capacidade de se conectar a diversas fontes de dados, como bancos de dados (SQL, Oracle), arquivos locais (*Excel*, CSV) e serviços na nuvem (Azure, *Analytics Services*), facilitando a importação e análise de dados relevantes.

As soluções de BI têm a capacidade de analisar grandes volumes de atividades, respondendo a consultas ao identificar padrões nos dados. Essas soluções abrangem uma ampla variedade de tecnologias (Rodrigues, 2021), além de incluir a execução de processos como a ETL, responsáveis por extrair, limpar, normalizar e carregar dados, que é semelhante ao processo de KDD. Elas também envolvem a criação de DWs e pode ser criado o SS, que organizam e relacionam os dados para facilitar a análise e consulta, bem como possibilitam a visualização e interpretação, exemplificada na Figura 4.

Figura 4 – Processo KDD Adaptado DW e SS



Fonte: Adaptado de Fayyad, Piatetsky-Shapiro e Smyth (1996)

V. Data Analysis Expressions

O *Data Analysis Expressions* (DAX) ou Expressões de Análise de Dados, em português, é uma linguagem de fórmulas usada para manipulação de dados em ferramentas como PBI, *Power Pivot* e *Analysis Services* da *Microsoft*. Ele permite a criação de cálculos personalizados, como medidas e colunas calculadas, usando funções que facilitam a análise de dados em modelos tabulares (Microsoft, 2023). Segundo a Microsoft (2023) o DAX é amplamente utilizado em projetos no PBI para criar fórmulas complexas, como somas, médias ponderadas, filtros condicionais e cálculos temporais, otimizando a visualização e o entendimento dos dados, desenvolvendo painéis interativos que oferecerão análises mais profundas e precisas.

3.4.3 Métodos

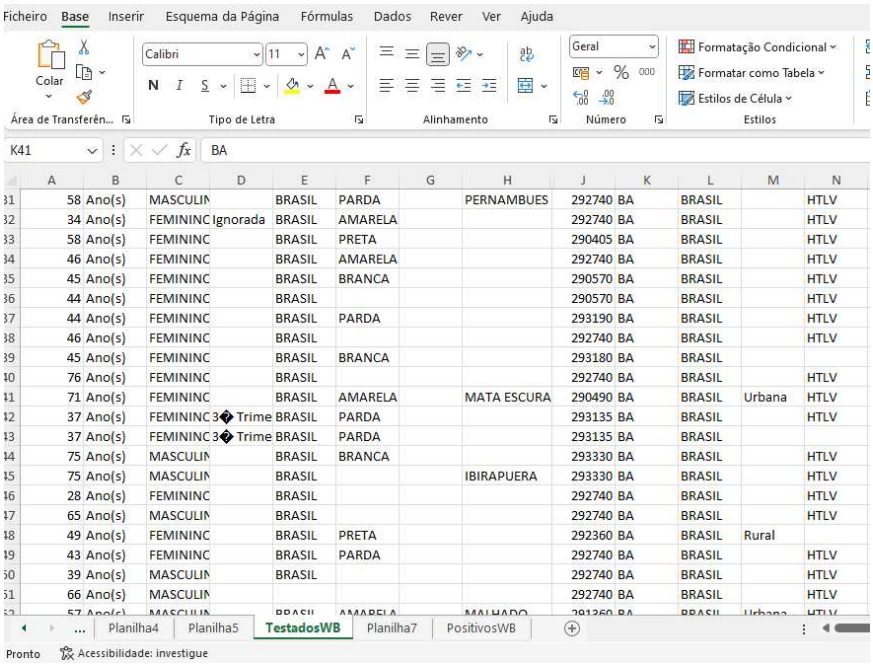
A mineração de dados é um processo que combina diferentes métodos e técnicas de descoberta de conhecimento, abrangendo a análise aprofundada dos dados e a aplicação de algoritmos especializados para identificar padrões em grandes conjuntos de informações, esse processo foi descrito detalhadamente na seção 2.2 Mineração de Dados (MD).

3.4.3.1 Extração de dados

O ETL de dados é uma integração dos dados, a primeira etapa do projeto envolveu a extração ou seleção de dados estruturados fornecidos pela SESAB. Os dados obtidos de uma planilha do *Microsoft Excel*, tendo uma amostra temporal de 2016 a 2023. Esta planilha continha variáveis relevantes como idade, tipo de idade, sexo, idade gestacional, nacionalidade, raça/cor, etnia, e outros dados geográficos e demográficos, como mostrado na Figura 5.

É possível observar a presença de espaços em branco, caracteres incomuns, além de números e textos no mesmo campo. Existem também várias abas que não são essenciais para o projeto e serão tratadas adequadamente. Além disso, os dados sensíveis, serão removidos do banco de dados para garantir a privacidade e a integridade dos pacientes.

Figura 5 – Banco de dados HTLV



	A	B	C	D	E	F	G	H	J	K	L	M	N
31	58	Ano(s)	MASCULIN		BRASIL	PARDA		PERNAMBUES	292740	BA	BRASIL		HTLV
32	34	Ano(s)	FEMININC	Ignorada	BRASIL	AMARELA			292740	BA	BRASIL		HTLV
33	58	Ano(s)	FEMININC		BRASIL	PRETA			290405	BA	BRASIL		HTLV
34	46	Ano(s)	FEMININC		BRASIL	AMARELA			292740	BA	BRASIL		HTLV
35	45	Ano(s)	FEMININC		BRASIL	BRANCA			290570	BA	BRASIL		HTLV
36	44	Ano(s)	FEMININC		BRASIL				290570	BA	BRASIL		HTLV
37	44	Ano(s)	FEMININC		BRASIL	PARDA			293190	BA	BRASIL		HTLV
38	46	Ano(s)	FEMININC		BRASIL				292740	BA	BRASIL		HTLV
39	45	Ano(s)	FEMININC		BRASIL	BRANCA			293180	BA	BRASIL		
40	76	Ano(s)	FEMININC		BRASIL				292740	BA	BRASIL		HTLV
41	71	Ano(s)	FEMININC		BRASIL	AMARELA		MATA ESCURA	290490	BA	BRASIL	Urbana	HTLV
42	37	Ano(s)	FEMININC	3ª Trime	BRASIL	PARDA			293135	BA	BRASIL		HTLV
43	37	Ano(s)	FEMININC	3ª Trime	BRASIL	PARDA			293135	BA	BRASIL		
44	75	Ano(s)	MASCULIN		BRASIL	BRANCA			293330	BA	BRASIL		HTLV
45	75	Ano(s)	MASCULIN		BRASIL			IBIRAPUERA	293330	BA	BRASIL		HTLV
46	28	Ano(s)	FEMININC		BRASIL				292740	BA	BRASIL		HTLV
47	65	Ano(s)	MASCULIN		BRASIL				292740	BA	BRASIL		HTLV
48	49	Ano(s)	FEMININC		BRASIL	PRETA			292360	BA	BRASIL	Rural	
49	43	Ano(s)	FEMININC		BRASIL	PARDA			292740	BA	BRASIL		HTLV
50	39	Ano(s)	MASCULIN		BRASIL				292740	BA	BRASIL		HTLV
51	66	Ano(s)	MASCULIN						292740	BA	BRASIL		HTLV

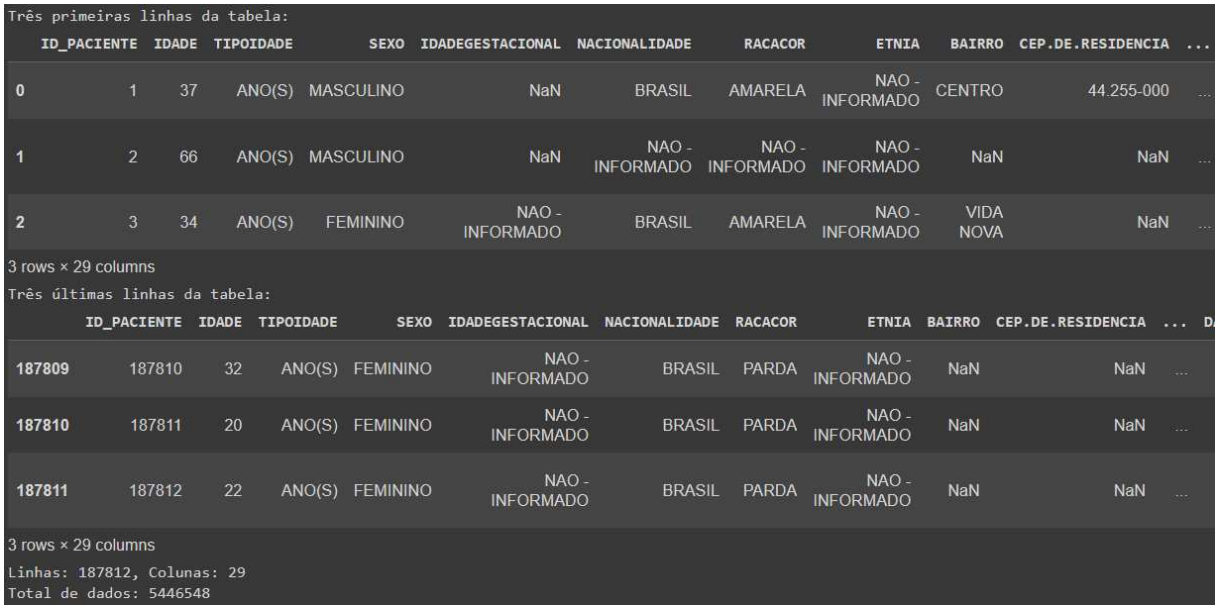
Fonte: O Autoria própria (2024)

3.4.3.2 Análise exploratória inicial

Conduzindo uma análise exploratória dos dados, conforme Medri (2011), essa abordagem é fundamental para explorar e entender o conjunto de dados, permitindo identificar padrões, verificar suposições, verificar dados que faltam, visualizar dados, detectar anomalias e testar hipóteses. Portanto, somente após essa etapa inicial é possível aplicar modelos mais complexos ou tomar decisões fundamentadas.

Para isso, a análise inicial foi realizada com *Python*, permitindo uma avaliação preliminar do tamanho, estrutura e quantidade de informações a serem processadas, além dos tipos de dados do banco. Isso proporcionou um melhor entendimento do tamanho do banco de dados, possibilitando o planejamento mais eficiente dos recursos computacionais necessários para o processamento. A estrutura dos dados também foi analisada, facilitando a otimização do desempenho e a identificação de melhorias na organização e no acesso aos dados. Além disso, a avaliação da quantidade de dados foi essencial para escolher as técnicas e algoritmos de análise mais adequados, garantindo um processamento eficiente. Essa análise inicial ajudou a detectar inconsistências e a necessidade de limpeza de dados, assegurando que as informações estivessem prontas para uma análise posterior mais precisa, conforme apresentado na Figura 6.

Figura 6 – Análise tamanho, estrutura e quantidade de dados



Três primeiras linhas da tabela:

	ID_PACIENTE	IDADE	TIPOIDADE	SEXO	IDADEGESTACIONAL	NACIONALIDADE	RACACOR	ETNIA	BAIRRO	CEP.DE.RESIDENCIA	...
0	1	37	ANO(S)	MASCULINO	NaN	BRASIL	AMARELA	NAO - INFORMADO	CENTRO	44.255-000	...
1	2	66	ANO(S)	MASCULINO	NaN	NAO - INFORMADO	NAO - INFORMADO	NAO - INFORMADO	NaN	NaN	...
2	3	34	ANO(S)	FEMININO	NAO - INFORMADO	BRASIL	AMARELA	NAO - INFORMADO	VIDA NOVA	NaN	...

3 rows × 29 columns

Três últimas linhas da tabela:

	ID_PACIENTE	IDADE	TIPOIDADE	SEXO	IDADEGESTACIONAL	NACIONALIDADE	RACACOR	ETNIA	BAIRRO	CEP.DE.RESIDENCIA	...	D
187809	187810	32	ANO(S)	FEMININO	NAO - INFORMADO	BRASIL	PARDA	NAO - INFORMADO	NaN	NaN	...	
187810	187811	20	ANO(S)	FEMININO	NAO - INFORMADO	BRASIL	PARDA	NAO - INFORMADO	NaN	NaN	...	
187811	187812	22	ANO(S)	FEMININO	NAO - INFORMADO	BRASIL	PARDA	NAO - INFORMADO	NaN	NaN	...	

3 rows × 29 columns

Linhas: 187812, Colunas: 29
Total de dados: 5446548

Fonte: O Autoria própria (2024)

Essa análise, apresenta uma maior eficiência e escalabilidade em comparação com o Excel, na gestão de grandes volumes de dados de dados, podendo ser comparado a Figura 5 e a Figura 6. A análise no GC permitiu uma visão rápida da estrutura do banco, que conta com um total de 187.812 linhas e 29 colunas, totalizando 5.446.548 pontos de dados a serem analisados, apresentando exatamente o mesmo volume de dados em comparação com banco de dados no *Excel*. Esse volume expressivo de informações proporciona obter *insights* mais detalhados e precisos, além de verificar a conformidade dos dados com o arquivo original do banco de dados.

A identificação dos tipos de dados é uma etapa fundamental na análise, desempenhando um papel crucial em várias fases do processo. Compreender os tipos de dados, como números inteiros, decimais, texto ou datas, é essencial para entender a natureza das informações contidas na tabela, além de ser vital para a correta limpeza e transformação dos dados. Essa compreensão garante que os dados estejam no formato adequado e consistente, evitando erros durante a análise e assegurando resultados precisos. Com isso, é possível identificar padrões e tendências significativas. Operações matemáticas e estatísticas devem ser aplicadas apenas em colunas numéricas, enquanto colunas de texto exigem procedimentos específicos, como concatenação

ou extração de *strings*. A escolha correta das operações é crucial para a obtenção de insights relevantes. Além disso, esse processo garante o tratamento adequado dos dados e a gestão eficaz de valores ausentes. Diferentes tipos de dados podem exigir abordagens distintas para lidar com a ausência de informações, resultando em uma análise mais eficiente, conforme ilustrado na Tabela 2.

Tabela 2 – Identificação de tipos de dados

Nome da Coluna	Tipo de Dado	Categoria
D_PACIENTE IDADE	*int64*	*Int*
IDADE	*int64*	*Int*
TIPOIDADE	*Object*	*String*
SEXO	*Object*	*String*
IDADEGESTACIONAL	*Object*	*String*
NACIONALIDADE	*Object*	*String*
RACACOR	*Object*	*String*
ETNIA	*Object*	*String*
BAIRRO	*Object*	*String*
CEP.DE.RESIDENCIA	*Object*	*String*
IBGE.MUNICIPIODERESIDENCIA	*int64*	*Int*
ESTADODERESIDENCIA	*Object*	*String*
PAIS.DE.RESIDENCIA	*Object*	*String*
ZONA	*Object*	*String*
NOME.DA.PESQUISA	*Object*	*String*
EXAME	*Object*	*String*
METODOLOGIA	*Object*	*String*
DATADECADASTRO	*datetime64[ns]*	*Data*
DATADACOLETA	*datetime64[ns]*	*Data*
DATADALIBERACAO	*datetime64[ns]*	*Data*
RESULTADO_ELISA	*Object*	*String*
ANOCAD	*int64*	*Int*
ANOLIB	*int64*	*Int*
RESULTADOWB	*Object*	*String*
MESORREGIAO	*Object*	*String*
MICRORREGIAO	*Object*	*String*
MUNICIPIO	*Object*	*String*
NUCLEOREGIONAL	*Object*	*String*
REGIAODESAUDE	*Object*	*String*

Fonte: O Autoria própria (2024)

A tabela apresentada oferece uma visão geral detalhada sobre a organização e a classificação dos dados no *Data Frame* (DF), permitindo o entendimento do tipo e a natureza dos dados em cada coluna, facilitando a análise, os relacionamentos e o processamento posterior. A coluna "Nome" indica o identificador específico para cada coluna no DF, como

"D_PACIENTE" ou "IDADE". O "Tipo de Dado" descreve o formato dos dados contidos na coluna, e respectivamente sua “categoria” que cada tipo pertencente.

3.4.3.3 Transformação e Carga

A transformação dos dados fundamental para preparar e facilitar a interpretação das informações. Na construção desse projeto, esse processo envolveu várias etapas, incluindo:

A limpeza de dados, que é o processo de identificar, corrigir ou remover dados incorretos, incompletos, duplicados ou inconsistentes de um conjunto de dados. O objetivo é garantir que os dados sejam precisos, confiáveis e prontos para análise, eliminando erros que possam comprometer os resultados.

- Remover ou corrigir entradas duplicadas, incompletas ou incorretas.
- Tratamento de valores ausentes ou valores nulos foram substituídos por "não informado" ou “0”.
- Corrigir caracteres especiais que vieram errados devido à conversão do banco de dados.
- Verificar e garantir que colunas estejam no formato correto
- Remover linhas vazias.

A transformação dos dados, que envolve a preparação e organização dos dados para análise, incluindo transformação, normalização e consolidação. Esse processo corrige inconsistências, remove erros e prepara os dados em um formato adequado para facilitar análises precisas e eficientes.

- Normalização de dados ou padronizar dados numéricos, onde havia idade em ano e meses, todas colocadas para anos.
- Verificação e correção de inconsistências nas entradas categóricas, como gênero e numéricas.

A conversão de tipos de dados é a verificação ou alteração, se necessário, onde garantimos que cada coluna esteja no formato apropriado.

E por fim, a filtragem de dados, uma etapa em que selecionamos apenas os dados relevantes para a análise, excluindo informações que não são necessárias ou que podem distorcer os resultados.

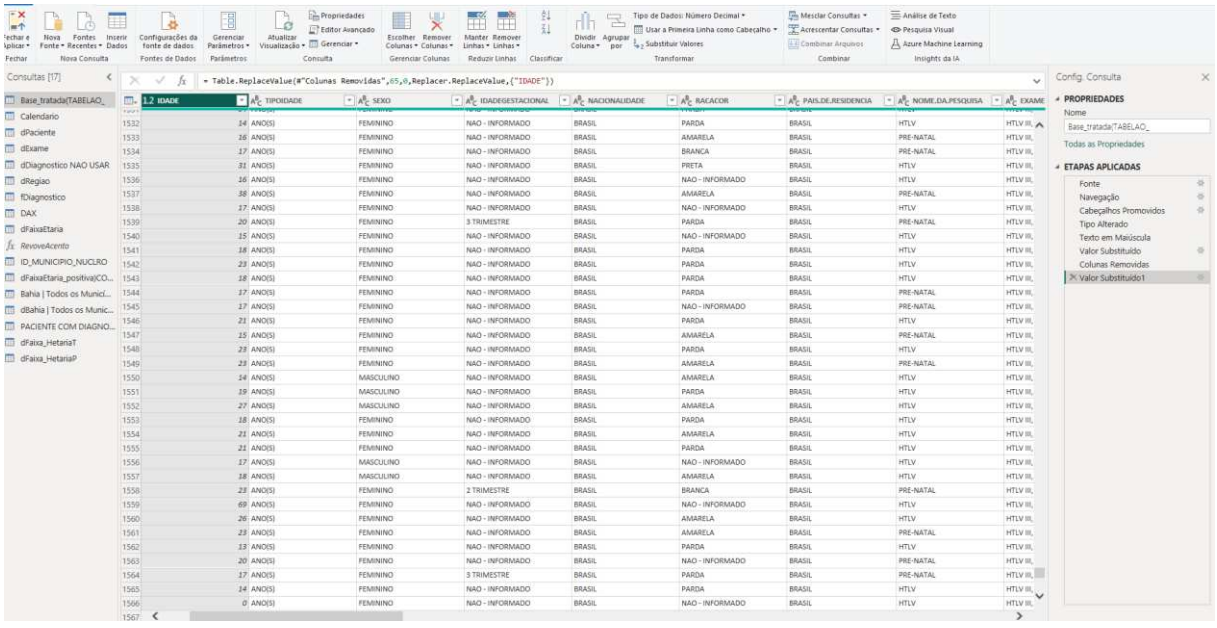
- Não foram excluídos nenhuma coluna, mantendo assim, a totalidade dos dados.

Após essas etapas e extração e tratamento inicial de dados foi carregado no PBI.

3.4.3.4 Importação e Implementação no Power BI

A etapa de importação, conhecida como carga, dos dados para o PBI *Desktop* foi conduzida com o intuito de preparar as informações para uma nova análise. Os dados foram carregados para a base do *Power Query* (PQ), uma ferramenta integrada do PBI. Nesse ambiente, foi realizado o processo de KDD, a iteração do ETL, no qual os dados foram refinados para aprimorar ainda mais qualidade e adequação para análise, uma vez que, ao carregar os dados, o PQ pode não identificar os tipos corretamente. A Figura 7, demonstra como é o ambiente do PQ.

Figura 7 – Power Query dados SADH



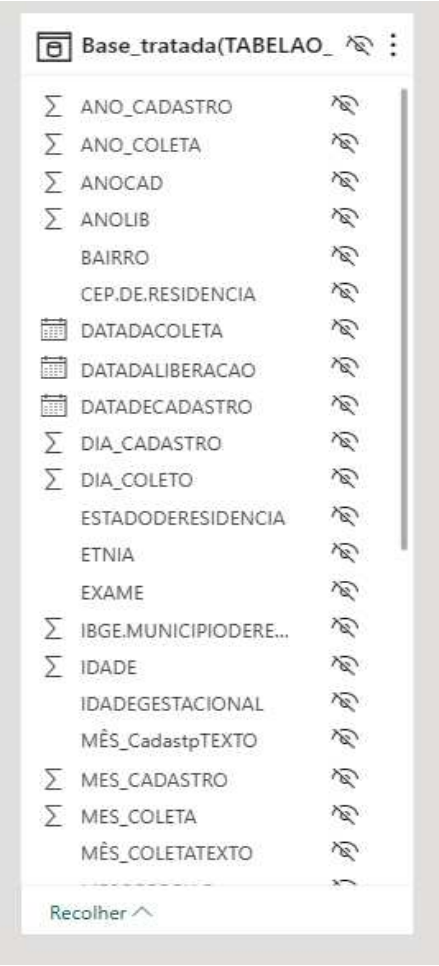
Fonte: O Autoria própria (2024)

Este procedimento assegura que os dados estejam prontos para uma análise aprofundada e também contribui para a construção de modelos de dados robustos e eficientes, preparando-

os para a criação dos *dashboards*. A utilização do PQ possibilitou uma integração fluida e eficaz dos dados, garantindo a consistência e a qualidade necessárias para a análise no PBI.

O banco de dados carregado no BI estava em uma tabela única, como demonstrado na Figura 8. Isso pode acarretar diversos problemas, o desempenho das consultas poderá ser comprometido, pois tabelas grandes exigem que o sistema percorra um volume maior de dados, tornando as análises mais lentas. Além disso, a complexidade das consultas tende a aumentar. A manutenção de um DW com tabelas extensas também se torna mais complexas.

Figura 8 – Tabela única Power BI SADH



Fonte: O Autoria própria (2024)

Portanto, será necessário criar uma modelagem dimensional, o SS, que tende a oferecer uma melhor performance, maior usabilidade e facilidade de manutenção.

3.4.3.5 Modelagem dos dados

O processo de modelagem de dados, consiste na transformação dimensional, etapa em que os dados são organizados em um formato que facilita análises e consultas em sistemas de BI, a estrutura SS das informações é composta em dois tipos principais de tabelas: tabela fato e tabela dimensão. No PBI, esse processo é realizado no PQ.

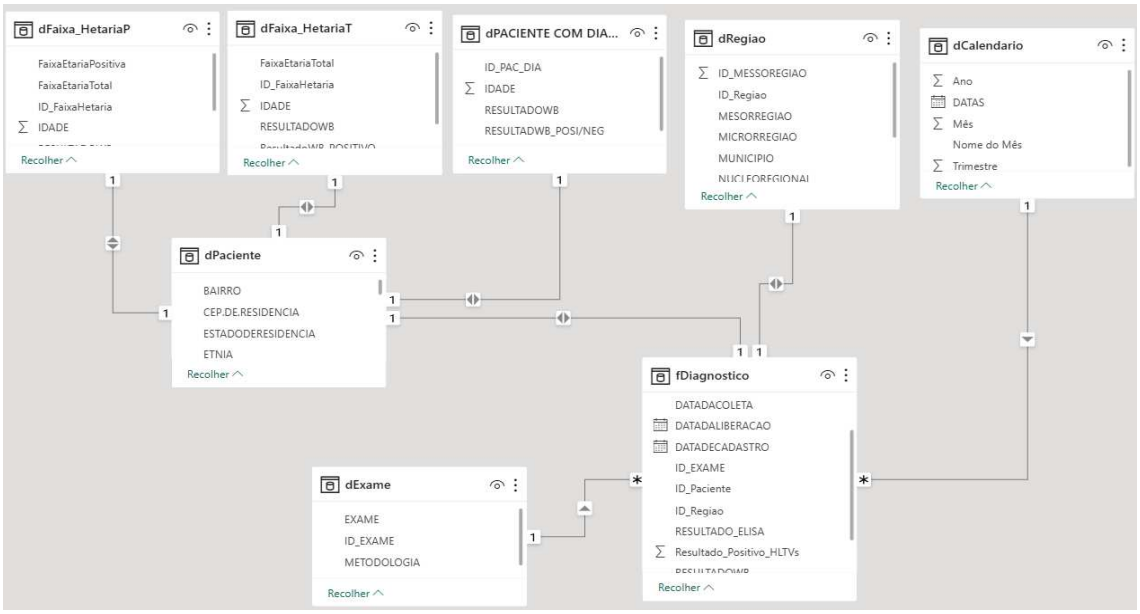
Para iniciar, acesse a página inicial do PBI e em “transformar dados”. Isso abrirá o PQ, exibindo os dados previamente carregados. No PQ, é importante preservar os dados originais. Para isso, crie uma duplicata da tabela que deseja modelar. Em seguida, selecione as colunas relevantes para a criação da tabela fato ou dimensão e utilize a funcionalidade “remover outras colunas”, mantendo apenas as colunas selecionadas.

Uma boa prática de modelagem é usar prefixos para diferenciar os tipos de tabelas: “f_” para tabelas fato e “d_” para tabelas dimensão ao renomear as tabelas criadas.

Após preparar os dados, a tabela fato deve conter métricas numéricas (como valores ou quantidades) e as chaves estrangeiras. Já a tabela dimensão armazena atributos descritivos (como categorias e nomes) e a chave primária. Ao finalizar as transformações, é aplicado processo para retornar ao PBI e carregar os dados no modelo.

No ambiente do PBI, acesse a aba “modelo” para estabelecer os relacionamentos entre as tabelas. Conecte a tabela fato e as tabelas dimensão utilizando as chaves primárias e estrangeiras apropriadas. Isso garante um modelo dimensional eficiente e bem estruturado, como demonstrado na Figura 9. Esse processo é essencial para organizar os dados de forma eficiente, fornecendo contexto e facilitando as consultas.

Figura 9 – Star Schema Power BI SADH



Fonte: O Autoria própria (2024)

Considerando essa estrutura, a tabela fato é denominada "fDiagnósticos", e as tabelas de dimensões incluem “dPaciente”, “dExames”, “dRegião” e “dCalendário”. A tabela fato "fDiagnósticos" contém informações cruciais es chaves primárias como “ID_Diagnostico”, “ID_Paciente” e “ID_Exame” e seus dados específicos, permitindo uma análise completa dos diagnósticos realizados.

As tabelas de dimensões oferecerão contextos específicos. A dimensão “dPaciente” permitirá segmentações detalhadas e análises mais granulares, como “dFaixa_Etária_Parcial” e “dFaixa_Etária_Total” criadas a partir de informações de idade, sexo, resultadoswb e estruturas condicionais e “dPaciente com Diagnóstico”, para melhor analisar esses casos. A dimensão “dExames” apresentará dados sobre os exames realizados, enquanto a dimensão “dRegião” incluirá informações geográficas relevantes. Por fim, a dimensão “dCalendário” facilitará análises temporais.

No modelo SS, a tabela fato "fDiagnósticos" se tornará o centro das relações, conectando-se às tabelas de dimensões. Isso permitirá análises detalhadas e integradas, uma vez que as diferentes dimensões se interligam por meio dos dados na tabela fato. Essa estrutura robusta não apenas proporciona uma visão clara dos dados, mas também possibilita *insights* valiosos sobre diagnósticos, pacientes, regiões e o contexto temporal, tornando-a uma abordagem eficaz para a modelagem de dados no PBI.

3.4.3.6 Análise de dados

A análise de dados tem se revelado indispensável na rotina das empresas, a importância de se manterem continuamente atualizadas e informadas, tanto a respeito de fatores externos quanto das atividades internas de seus negócios. Esse processo de transformação de dados em informações relevantes qualifica os profissionais a tomar decisões cada vez mais aprimoradas.

Segundo Gil (1999), o objetivo da análise de dados é organizar e resumir as informações de maneira que permitam responder ao problema investigado, enquanto a interpretação busca atribuir um significado mais amplo a essas respostas, relacionando-as com conhecimentos já existentes. Complementando com essa visão, Teixeira (2003), destaca que a análise de dados envolve a construção de significado a partir da interpretação das informações coletadas. Esse processo é complexo, pois envolve um movimento contínuo entre dados concretos e conceitos abstratos, alternando entre raciocínios indutivos e dedutivos, e equilibrando descrição e interpretação.

Neste projeto, após a obtenção da base de dados, foi realizada uma análise exploratória com *Python*. Os dados foram carregados no PBI, onde o ETL iterativo foi executado no PQ, gerando os modelos de SS. Por fim, foi realizada a análise dos dados com a linguagem DAX para criar as medidas, as colunas calculadas e as tabelas customizadas que ajudaram a realizar análises de dados mais avançadas, resultando nos *dashboards*, que serão apresentados a seguir na seção de Análises de Resultados.

4 ANÁLISE DE RESULTADOS

Neste capítulo, são apresentados os resultados e as análises realizadas neste trabalho, com foco nos dashboards do SADH, baseados nos dados fornecidos pela SESAB sobre as ocorrências epidemiológicas do HTLV no estado da Bahia. A seção 4.1 explica as diferentes maneiras de analisar os *dashboards*. A seção 4.2 traz uma visão geral dos dados coletados sobre o HTLV, com foco maior por Núcleo Regional. A seção 4.3 detalha a triagem do HTLV, abordando especificamente os testes *Enzyme-Linked Immunosorbent Assay* (ELISA), ou Ensaio de Imunoabsorção Enzimática, e *Western Blot* (WB), com foco por Núcleo Municipal. A seção 4.4 explora a tendência temporal, mostrando a evolução ao longo do período de coleta e identificando os tipos de casos com maior incidência. A seção 4.5 destaca os casos positivos de HTLV, segmentados por sexo e gestantes. Por fim, a secção 4.6 demonstra a análise de todos os gráficos interagindo, para fornecer um *insight* específico em uma determinada cidade.

4.1 Forma de análise

O SADH pode ser analisado de duas maneiras destintas. A primeira, e detalhada, é de forma individual, em que cada gráfico é avaliado individualmente, proporcionando uma compreensão isolada de específica de cada elemento, onde serão analisadas as seções 4.2 ao 4.5. Já a segunda forma, onde os *dashboards* são avaliados de forma interativa, que ao selecionar qualquer painel e gráfico, todos os outros se ajustam automaticamente, permitindo que os dados conversem entre si. Proporcionando uma visão mais integrada da análise e favorece uma compreensão mais dinâmica e completa, onde será exemplificado, na seção 4.6, visto a grande variação de análises que se pode obter desses painéis. A Figura 10 apresenta a tela inicial do SADH.

Figura 10 – Tela inicial: SADH

Fonte: O Autoria própria (2024)

4.2 Dados Gerais

O *dashboard*, Figura 11, apresenta a análise geral de todos os dados coletados sobre o HTLV. Seu principal foco está na distribuição geográfica dos testes realizados, o número de total exames realizados em diferentes regiões da Bahia.

A análise detalhada, além dos casos de triagem do regional, oferece a distribuição de faixa etária de exames, os resultados dos testes realizados e a distribuição por raça/cor dos indivíduos testados, ressaltando o total de exames realizados, o número total de casos reagentes e a porcentagem de triagem positivas, que serão mais detalhados nas seções 4.2.1 ao 4.2.5.

Além disso, há dois filtros integrados: um para seleção de ano (de 2016 a 2022) e um limpador de filtro, permitindo a remoção eficiente de seleções múltiplas dos gráficos.

Este *dashboard* é essencial para compreender a dinâmica geral da testagem do HTLV e monitorar a distribuição dos casos reagentes. Ele oferece uma base sólida para direcionar as tomadas de decisões. Monitorando os grupos demográficos mais afetados, priorizando áreas com maior demanda por diagnósticos ou aquelas com baixa cobertura de testagem. Além disso, permite identificar áreas que necessitam de mais atenção em termos de diagnóstico e prevenção.

Figura 11– Dados Gerais

Fonte: O Autoria própria (2024)

4.2.1 Indicadores

A Figura 12 apresenta três cartões com informações objetivas sobre os testes realizados. O primeiro cartão, “Total Testes Realizados”, indica que foram realizados 187.812 testes. O segundo cartão, "Total de Casos Reagentes", revela que 3.578 testes deram resultados reagentes, ou seja, positivos para o HTLV. Já o terceiro cartão, “% de Casos Reagentes”, informa que 1,91% de todos os testes realizados apresentaram resultados reagentes. Um resultado reagente indica a presença de anticorpos ou antígenos específicos, sugerindo que a pessoa foi exposta ou está infectada por um agente infeccioso.

Figura 12 – Indicadores Gerais

Fonte: O Autoria própria (2024)

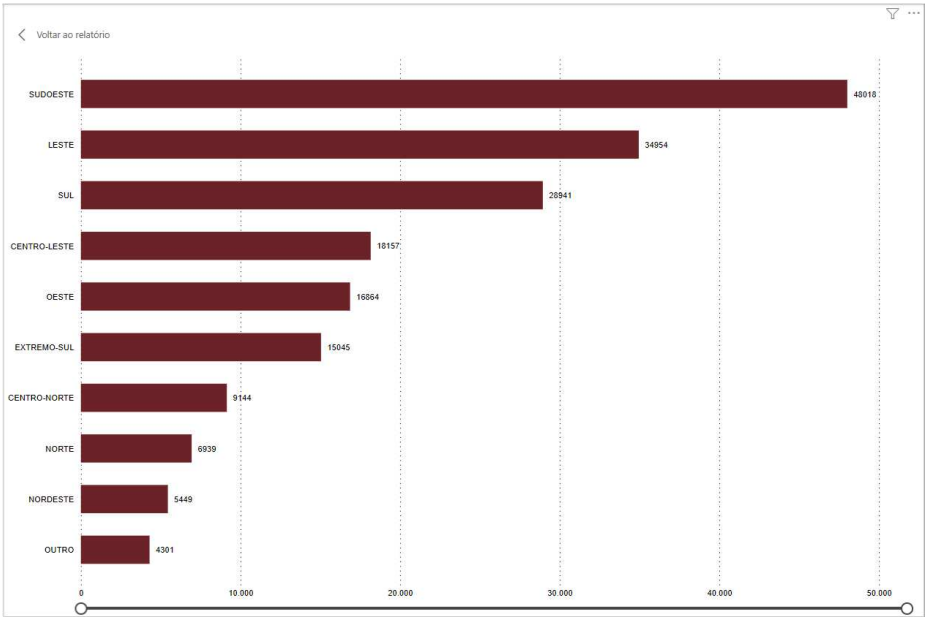
Esses indicadores fornecem uma visão geral rápida e eficiente da situação.

4.2.2 Núcleo Regional da Bahia & Testes por Região

A análise de "Núcleo Regional da Bahia & Testes por Região", com um gráfico de barras clusterizado, revela que a região Sudoeste da Bahia foi responsável por 48.018 testes, sendo a área com o maior número de testagens. Em seguida, a Região Leste realizou 34.954 testes, sendo a segunda região regional com mais testagens, enquanto a Região Sul contabilizou 28.941 testes. As Regiões Centro-Leste e Oeste realizaram 18.157 e 16.864 testes, respectivamente. A região do Extremo-Sul totalizou 15.045 testes, seguida pelo Centro-Norte com 9.144 e pelo Norte, que registrou 6.939 testes. A Região Nordeste foi responsável por 5.449 testes, sendo a região com menos testagens, e a categoria "Outro", indica que a região do paciente na Bahia não foi coletada corretamente, contabilizou 4.301 testes, visto no Gráfico 1.

O gráfico ainda dispõe do *slider*, um controle de deslizamento, que permite aos usuários ajustar visualizações, na parte inferior do gráfico.

Gráfico 1 – Núcleo Regional da Bahia & Testes por Região



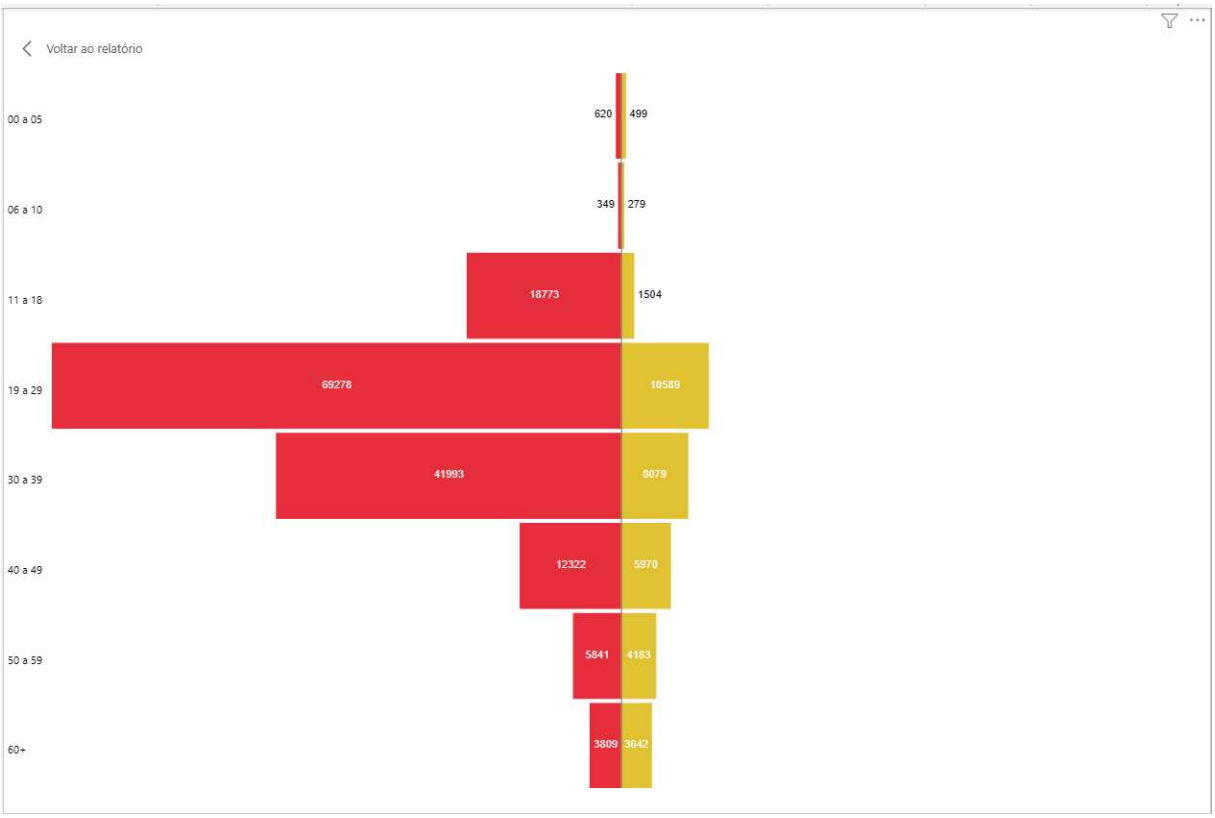
Fonte: O Autoria própria (2024)

4.2.3 Faixa Etária por Testes Realizados

Na análise por “Faixa Etária por Testes Realizados”, é um gráfico do tipo tornado, Gráfico 2, apresenta uma distribuição de casos segmentada por gênero entre o sexo feminino e o sexo masculino, e por faixas etárias de 0 a 60 (ou mais) anos. A cor vermelha representa o sexo feminino, enquanto a cor amarela representa o sexo masculino.

O grupo de 19 a 29 anos foi o mais testado, com um total de 79.856 testes, dos quais 69.278 foram realizados no sexo feminino e 10.589 no sexo masculino. Em seguida a faixa etária de 30 a 39 anos foi a segunda mais testada, com 50.802 testes, sendo 41.993 no sexo feminino e 8.079 no sexo masculino. No grupo de 11 a 18 anos, foram realizados 20.277 testes, com predominância no sexo feminino 18.773. Entre os indivíduos de 40 a 49 anos, houve 18.204 testes, com 12.322 realizados no sexo feminino e 5.970 no sexo masculino. Nas faixas etárias mais altas, observou-se uma redução na testagem, com 9.324 exames na faixa de 50 a 59 anos e 7.451 em pessoas com 60 anos ou mais. Já na faixa de 00 a 05 anos, sendo a segunda menos testada, temos 620 no sexo feminino e 499 no sexo masculino.

Gráfico 2 – Faixa Etária por Testes Realizados



Fonte: O Autoria própria (2024)

4.2.4 Resultado de Testes Realizados

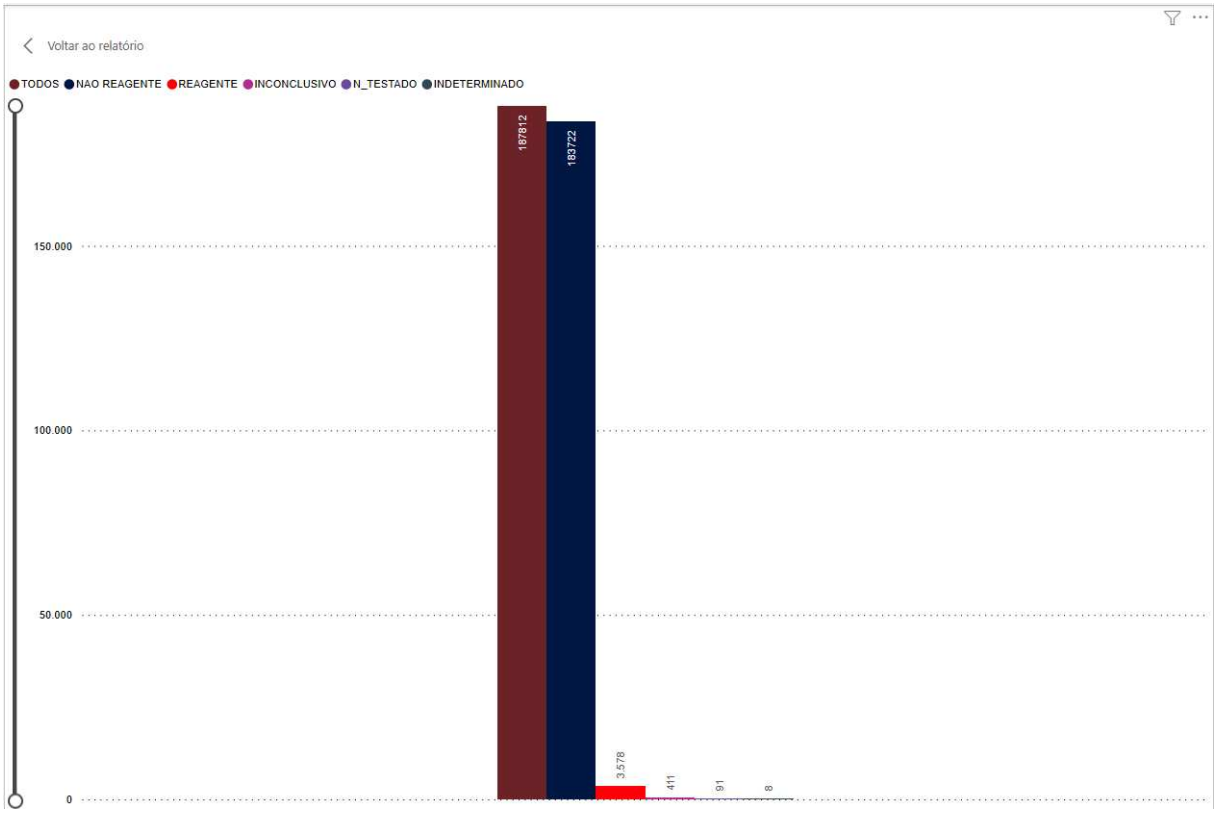
A análise “Resultado de Testes Realizados” representada por um gráfico de colunas clusterizado, Gráfico 3, demonstra os resultados de testes realizados, separados por seis categorias: "Todos", "Não Reagente", "Reagente", "Inconclusivo", “Não Testado” e “Indeterminado”. A maior parte dos resultados está concentrada nas categorias de "Todos" e "Não Reagente".

A coluna de "Todos" exibe o valor com 187.812 testes, indicando o total de testes aplicados. Em seguida a coluna de "Não Reagente", registrando um total de 183.722 testes. Indicando que a maioria dos testes realizados resultaram como "Não Reagente" para HTLV. Na categoria "Reagente", há 3.578 casos, que indica que a parcela dos testes que apresentaram um resultado positivo para o fator do HTLV. A categoria "Inconclusivo" com 411 testes, significa que não foi possível gerar um resultado definitivo. Com 91 testes “Não testados”. E por fim, os

“Indeterminados” refere-se a resultados em que o teste não conseguiu fornecer uma resposta definitiva, seja ela positiva ou negativa.

O gráfico ainda dispõe do *slider*, um controle de deslizamento, que permite aos usuários ajustar a visualização do quantitativo, na esquerda.

Gráfico 3 – Resultados De Testes Realizados



Fonte: O Autoria própria (2024)

4.2.5 Raça e Cor por testes Realizados

A análise de “Raça e Cor por testes Realizados”, é representado por um gráfico de linhas, Gráfico 4, demonstrando a variação de quantidade de casos testados ao longo dos anos, separados por raça/cor, entre 2017 e 2022. São consideradas diferentes categorias: Amarela com a cor de linha azul, Branca com a cor de linha verde, Indígena com a cor laranja, Não Informado com a cor de linha preta, Parda com a cor de linha rosa e Preta com a cor de linha amarela. Cada

linha corresponde a uma dessas classificações, podendo ser observado uma clara tendência de variação ao longo dos anos.

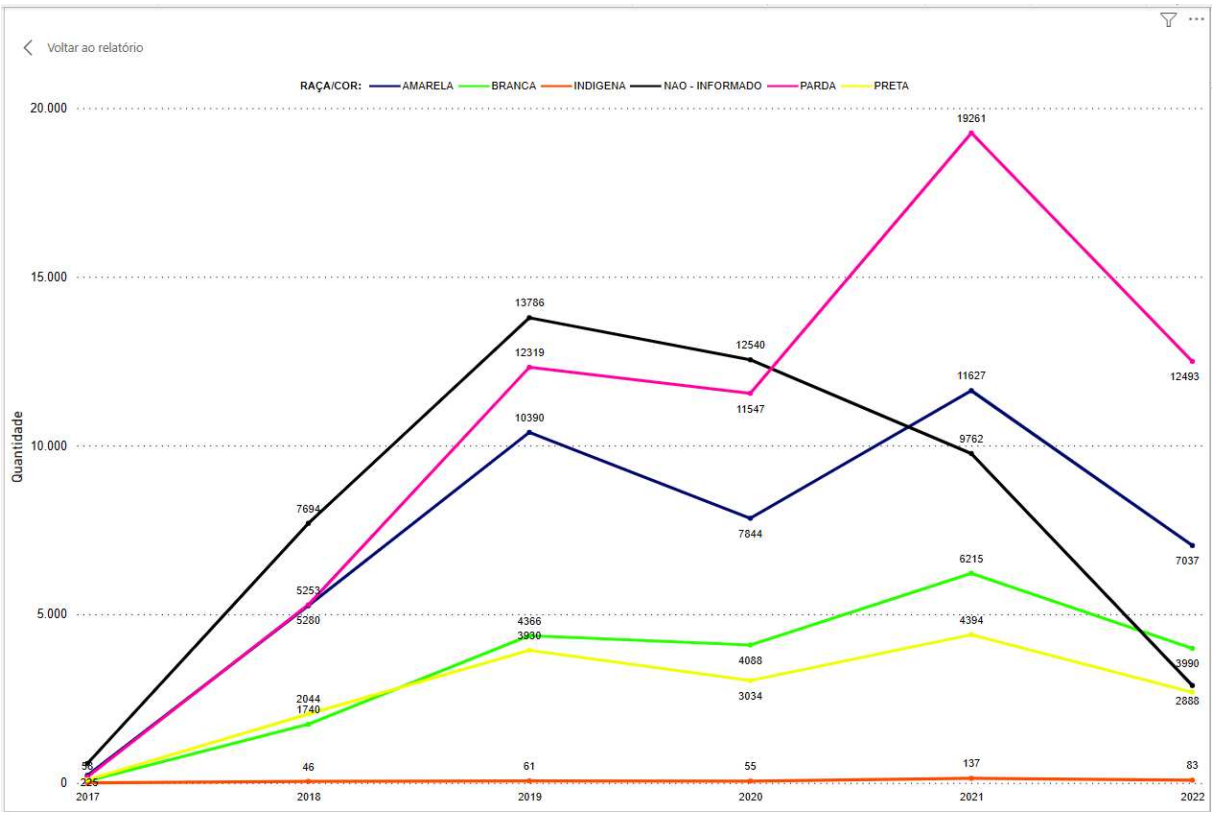
Em 2017, todas as categorias têm um número bastante baixo de registros, com algumas praticamente nulas. O caso que mais se destaca nos anos subsequentes é a referente à categoria “Parda”, que cresce rapidamente e atinge seu pico em 2021, com 19.261 casos. Esse é o maior valor registrado em qualquer das categorias durante o período analisado. No entanto, logo em seguida, essa quantidade começa a diminuir, chegando a 12.493 em 2022.

O segundo caso que mais se destaca é a da categoria “Não Informado”, que também apresenta um crescimento significativo entre 2017 e 2019, atingindo o pico de 13.786 registros em 2019. Contudo, a partir de 2020, essa categoria começa a apresentar uma queda com 2.888 casos em 2022.

Já as categorias “Preta”, “Amarela” e “Branca” mostram uma variação menor ao longo dos anos em relação a categoria “Parda”. O caso da categoria “Preta” atinge seu ponto mais alto em 2021, com 4.394 casos, mas decresce em seguida, chegando a 2.690 em 2022. O caso da categoria “Branca” tem um comportamento similar, com o maior pico em 2021, de 6.215 casos, e uma redução para 3.990 em 2022. O caso correspondente a categoria “Amarela” mostra um valor elevado, em comparação com as duas categorias acima, com o pico mais alto sendo de 11.627 em 2021 e caindo para 7.037 em 2022.

O caso referente aos “Indígenas” tem números extremamente baixos em comparação com as outras categorias, como em 2017 com 2 casos e valores máximos de 137 em 2021.

Gráfico 4 – Raça e Cor por Testes Realizados



Fonte: O Autoria própria (2024)

4.3 Dados HTLV

A Figura 13 apresenta o *dashboard* de dados com foco na infecção pelo HTLV, destacando informações epidemiológicas fundamentais para a análise e controle desse vírus. Ela contém diversos gráficos que oferecem uma visão detalhada sobre os casos, como a divisão por núcleo municipal, faixa etária e raça/cor, e os resultados de testes como ELISA e WB, onde serão mais detalhados nas seções 4.3.1 ao 4.3.5.

Além disso, há três filtros integrados: um para seleção de ano (de 2016 a 2022), um filtro de texto e um limpador de filtro, permitindo a remoção eficiente de seleções múltiplas.

A importância desse *dashboard* está na sua capacidade de fornecer uma análise detalhada dos testes positivos de triagem do HTLV, tanto ELISA quanto WB, permitindo a identificação de áreas e grupos prioritário. Isso pode facilitar a tomada de decisões baseadas em dados e o combate à disseminação do HTLV.

Figura 13 - Dados HTLV

Fonte: O Autoria própria (2024)

4.3.1 Indicadores

A Figura14 apresenta três indicadores relacionados aos testes de detecção do vírus HTLV. O primeiro indicador, "POSITIVO PARA ELISA", mostra que, por meio do teste ELISA, 3.578 casos foram identificados como positivos ou falso-positivos. O ELISA é uma técnica de triagem inicial de infecções virais, mas pode gerar resultados falso-positivos, o que torna necessário um teste de confirmação mais específico, como o WB.

O segundo indicador, "POSITIVO PARA WB", revela que, dos casos positivos no ELISA, 2.384 foram confirmados pelo teste WB. Esse exame é mais preciso e utilizado para confirmar os resultados do ELISA, eliminando a possibilidade de falsos positivos.

Por fim, o indicador "% DE CASOS POSITIVOS WB" refere-se ao percentual de casos positivos no ELISA que foram confirmados pelo WB.

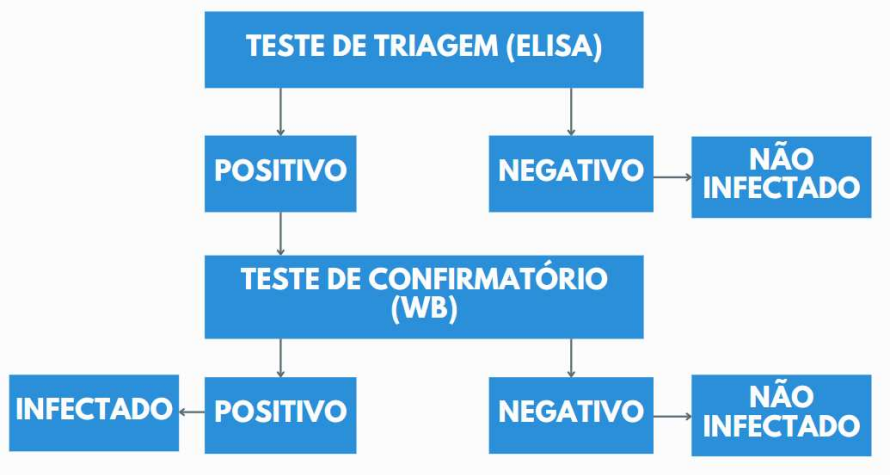
Figura 14 – Indicadores HTLV

Fonte: O Autoria própria (2024)

Os indicadores da Figura 14 são os mesmos indicadores que estão nos *dashboards* das seções 4.4 e 4.5.

No diagnóstico da infecção pelo HTLV-I e HTLV-II, um resultado negativo geralmente exclui a infecção, exceto em casos de suspeita de exposição recente, nos quais se recomenda a repetição do teste após 90 dias (Rosadas *et al*., 2021). Quando o resultado do ELISA é positivo, é necessário realizar um teste confirmatório, como o WB, para evitar falsos-positivos e se o WB resultar negativo, o resultado positivo do ELISA é considerado um falso-positivo, indicando que o paciente não está infectado, no entanto, se o WB for positivo, é confirmada a infecção, garantindo assim um diagnóstico mais preciso e confiável (Rosadas *et al*., 2021). A Figura 15 demonstra um diagrama sobre como é feita a triagem para o teste.

Figura 15 – Recomendações para o diagnóstico laboratorial da infecção pelo vírus linfotrópico de células T humanas (HTLV-1/2)



Fonte: Adaptado de Rosadas *et al*., (2021)

4.3.2 Núcleo Municipal da Bahia & Testes por Município

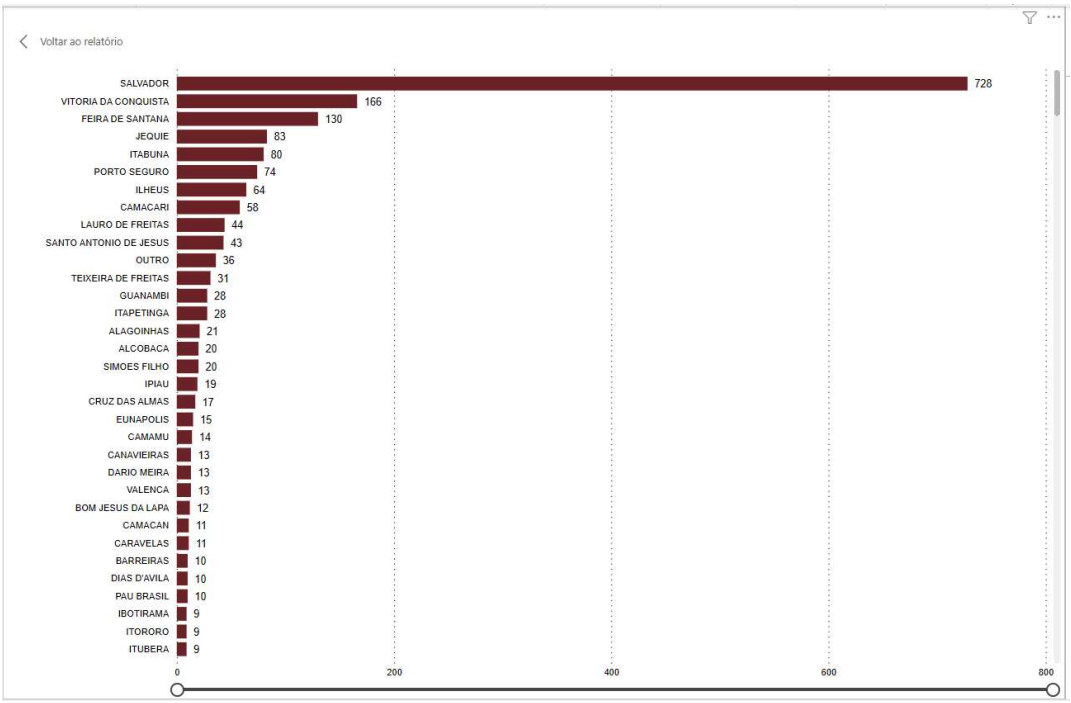
A Gráfico 5 apresenta um gráfico de barras horizontais que mostra o número de casos confirmados de WB para HTLV em diferentes municípios da Bahia, representando todos os 417 municípios do estado. Salvador se destaca como o município com o maior número de casos, com 728 confirmações, muito acima dos demais. Vitória da Conquista aparece em segundo lugar com 166 casos, seguido por Feira de Santana, que registrou 130 casos confirmados.

Outros municípios como Jequié (83 casos), Itabuna (80 casos) e Porto Seguro (74 casos) também aparecem com números expressivos, mas consideravelmente menores que os de Salvador. À medida que os municípios são listados, o número de casos diminui.

No meio da lista, há um grupo denominado "OUTRO", com 36 casos, o que sugere que pode representar casos de regiões que não foram especificadas na hora da coleta dos dados.

O gráfico ainda dispõe do *slider*, um controle de deslizamento, que permite aos usuários ajustar visualizações, na parte inferior do gráfico. Esse gráfico, é o mesmo que está nos *dashboards* das seções 4.4 e 4.5.

Gráfico 5 – Núcleo Municipal da Bahia & Testes por Município



Fonte: O Autoria própria (2024)

4.3.3 Faixa Etária por Casos Positivo Western Blot

Esta seção segue as mesmas métricas da seção 4.2.3, alterando apenas o foco da pesquisa. Enquanto o tópico anterior abordava a análise de todos os dados coletados para triagem, esta seção concentra-se na análise dos testes positivos para WB.

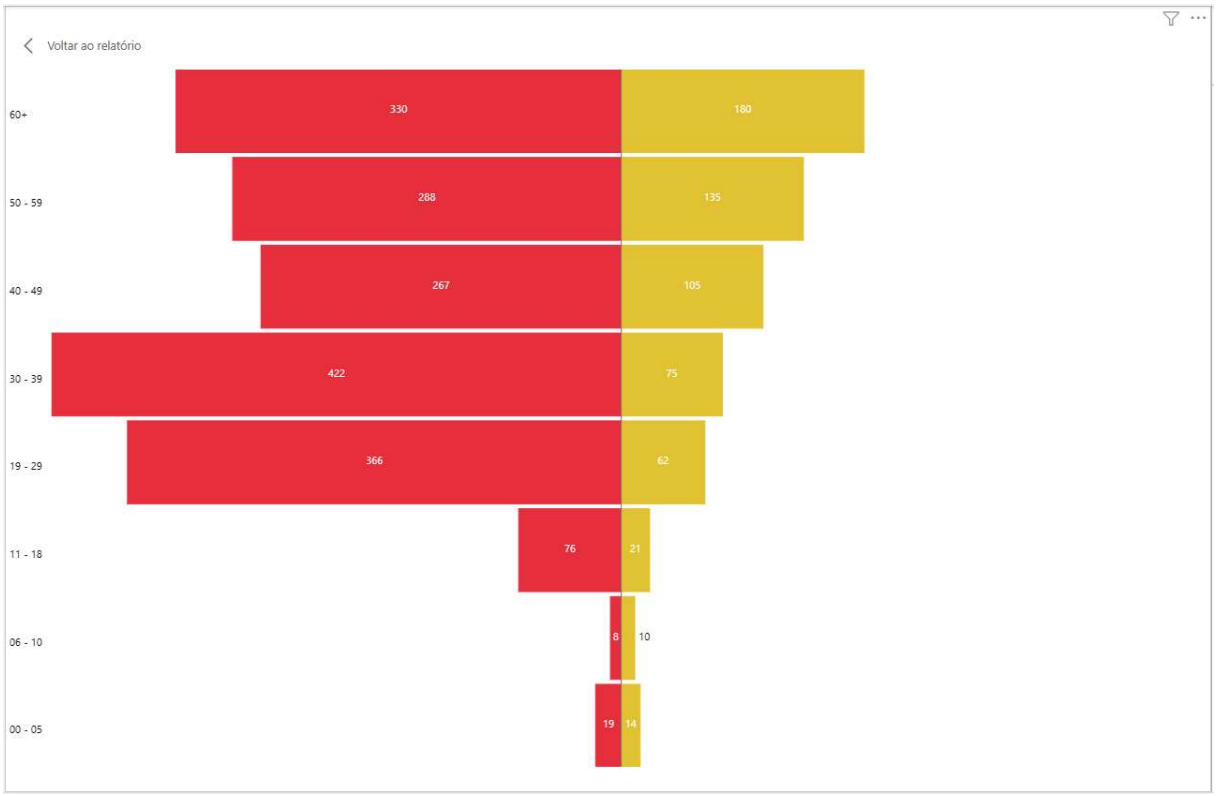
A distribuição desses testes é representada por um gráfico do tipo tornado (barras laterais), conforme a Gráfico 6, segmentado pelo gênero do sexo feminino e do sexo masculino e faixas etárias de 0 a 60 anos (ou mais). A cor vermelha representa o sexo feminino, enquanto a cor amarela representa o sexo masculino.

A maior concentração de testes ocorre na faixa etária de 30 a 39 anos, onde o número de casos no sexo feminino é de 422, significativamente maior do que o número de testes no sexo masculino, que é de 75. Essa disparidade entre os gêneros se mantém em praticamente todas as faixas etárias, especialmente nas idades mais avançadas. Por exemplo, na faixa dos 19 a 29 anos, há 366 tetes femininos contra 62 masculinos, evidenciando uma diferença notável.

A partir dos 40 anos, embora os números gerais diminuam, a tendência de maior prevalência entre o sexo feminino, continuando evidente. Na faixa de 40 a 49 anos, há 267 do sexo feminino e 105 do que no sexo masculino afetados. Já nos grupos mais velhos, como de 50 a 59 anos, essa diferença se mantém com 288 casos femininos contra 135 masculinos.

Nos extremos etários, entre os mais jovens de 0 a 18 anos, o número de testes é consideravelmente menor, tanto para o sexo masculino quanto para o sexo feminino. Há um equilíbrio nas idades de 0 a 10 anos, com um significativo aumento nas idades de 11 a 18 anos em diante.

Gráfico 6 – Faixa Etária por casos Positivos Western Bolt



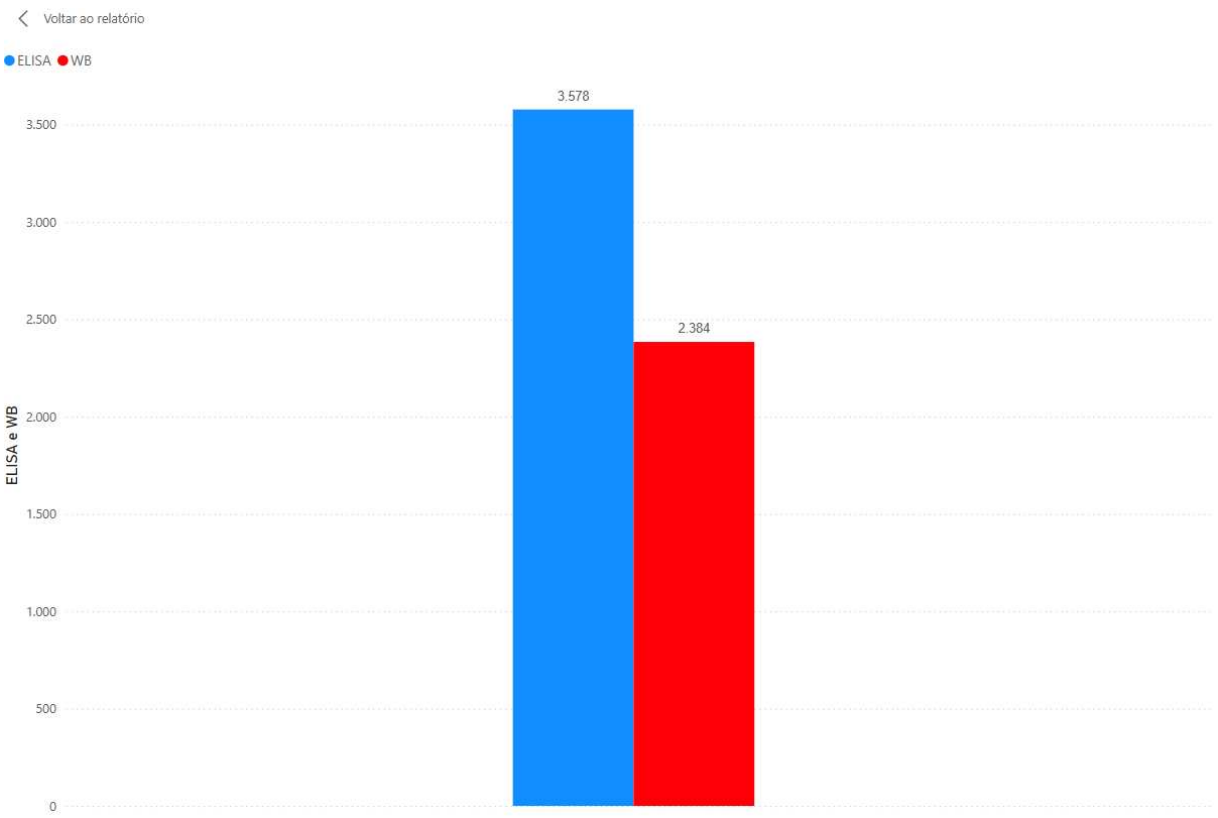
Fonte: O Autoria própria (2024)

4.3.4 Resultado Triagem

O Gráfico 7 é um gráfico de barras que mostra os resultados da triagem referente ao HTLV.

O teste “ELISA”, representado pela barra azul clara, apresentou 3.578 resultados. Esses resultados indicam casos que precisam ser confirmados com testes adicionais para determinar se são positivos ou negativos. Então, em seguida, o teste “WB”, representado pela barra vermelha, apresentou 2.384 resultados, significado que dos casos que foram positivos no teste “ELISA”, 2.384 também foram confirmados como positivos no teste “WB”, confirmando a maioria dos resultados preliminares.

Gráfico 7 – Resultado Triagem



Fonte: O Autoria própria (2024)

4.3.5 Raça e Cor por testes Realizados Western Blot

A análise de "Raça e Cor por testes Realizados WB" segue as mesmas métricas do tópico 4.2.5, alterando apenas o foco da pesquisa. Enquanto o tópico anterior abordava os dados gerais, esta seção foca nos casos positivos para WB, Gráfico 8, sendo representado por um gráfico de linhas, demonstrando a variação de quantidade de casos testados ao longo dos anos, separados por raça/cor, entre 2017 e 2022. São consideradas diferentes categorias: Amarela com a cor de linha azul, Branca com a cor de linha verde, Indígena com a cor de linha laranja, Não Informado com a cor de linha preta, Parda com a cor de linha rosa e Preta com a cor de linha amarela. Cada linha corresponde a uma dessas classificações, podendo ser observado uma clara tendência de variação ao longo dos anos.

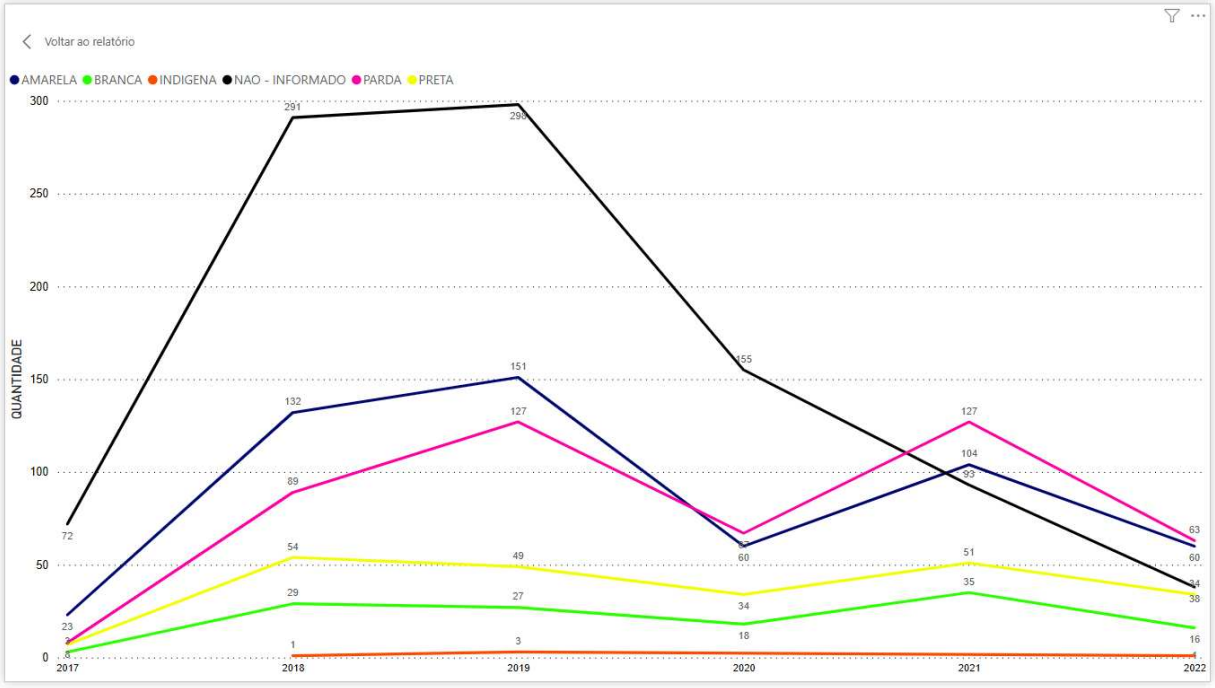
O grupo que não informou ou não foi coletado a cor/raça, “Não Informado”, teve o maior número de testes ao longo do período, com um pico de 298 testes em 2019, seguido por uma queda em 2020 para 155, continuando a decrescer em 2021, com 93 casos positivos para

WB, antes de cair novamente para 38 em 2022. O grupo “Amarelo” também realizaram uma quantidade considerável de testes, alcançando o pico de 151 em 2019, caindo para 60 em 2020, e depois subindo levemente para 104 em 2021, finalizando em 60 casos em 2022.

O grupo "Parda" apresentou um pico em 2019, com 127 casos. Houve uma diminuição em 2020, com 67 casos, mas em 2021 o número voltou a subir, novamente para 127 casos. Em 2022, houve uma queda significativa, chegando a 34 casos. O grupo "Preta" apresentou seu pico com 54 casos em 2018, com uma leve queda em 2019, para 49 casos. Em 2020, o número continuou a cair, atingindo 34 casos, mas houve uma leve alta em 2021, com 51 casos, antes de diminuir para 34 em 2022. Já os grupos "Branca" e "Indígena" tiveram um número menor de casos.

Por fim, o grupo "Branca" registrou 3 casos em 2017, aumentando para 29 casos em 2018, com pequenas oscilações até chegar a 16 casos em 2022. O grupo "Indígena" não teve nenhum caso registrado em 2017, teve 1 caso em 2018, 3 casos em 2019, caindo para 0 casos em 2020 e 2021, e novamente 1 caso em 2022.

Gráfico 8 – Raça e Cor por Testes Realizados Western Bolt



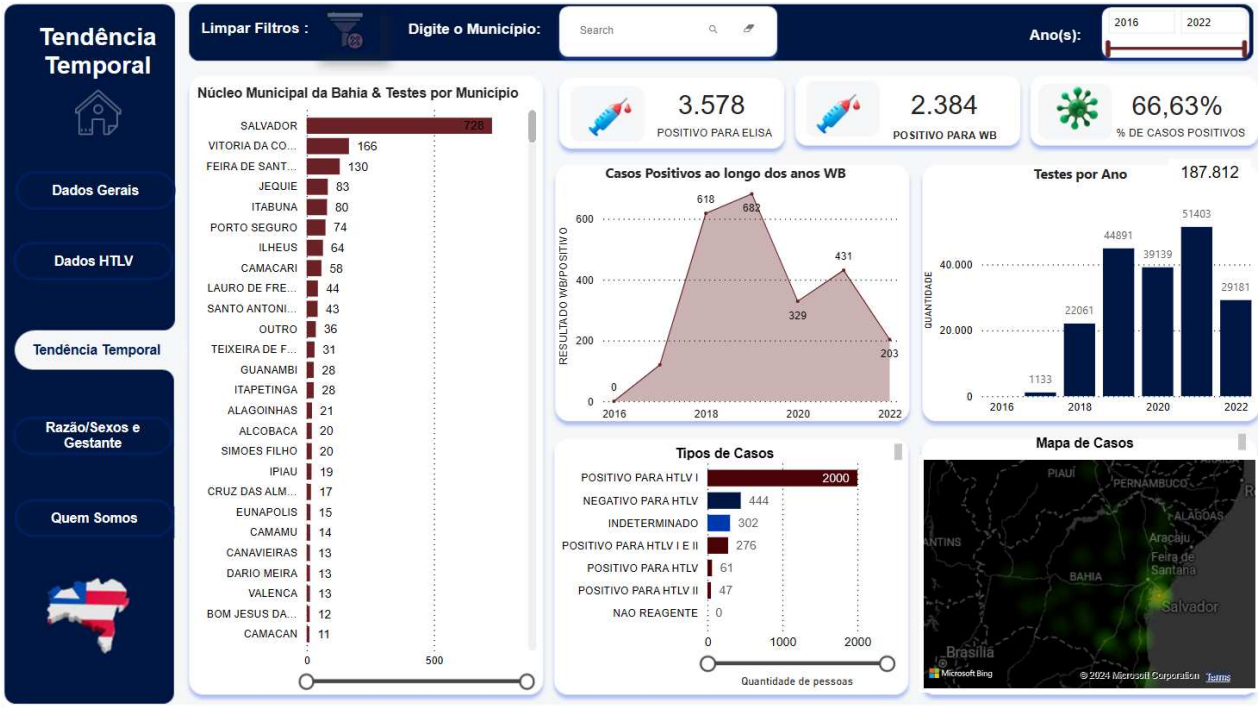
Fonte: O Autoria própria (2024)

4.4 Tendência Temporal

Esse painel, Figura 16, tendência temporal oferece uma visão detalhada da evolução dos casos positivos de HTLV na Bahia ao longo dos anos. Detalha o comportamento do vírus, como os casos positivos WB ao longo dos anos, o número de testes realizados por ano, os tipos de casos e suas variações, além de um mapa de calor que identifica as regiões com maior incidência, onde são detalhados nas seções 4.4.1 ao 4.4.5. Permitindo visualizar as mudanças de casos positivos ao longo do tempo, identificando se houve um aumento, uma diminuição ou se os números se mantiveram estáveis em determinados períodos. Ao observar os dados, é possível obter uma visão clara de como a incidência de casos positivos tem evoluído.

Esse acompanhamento temporal é essencial para compreender os padrões de disseminação do vírus e identificar possíveis mudanças na incidência da infecção em diferentes momentos. Além de auxiliar na formulação de estratégias mais direcionadas e eficazes.

Figura 16 – Tendência Temporal

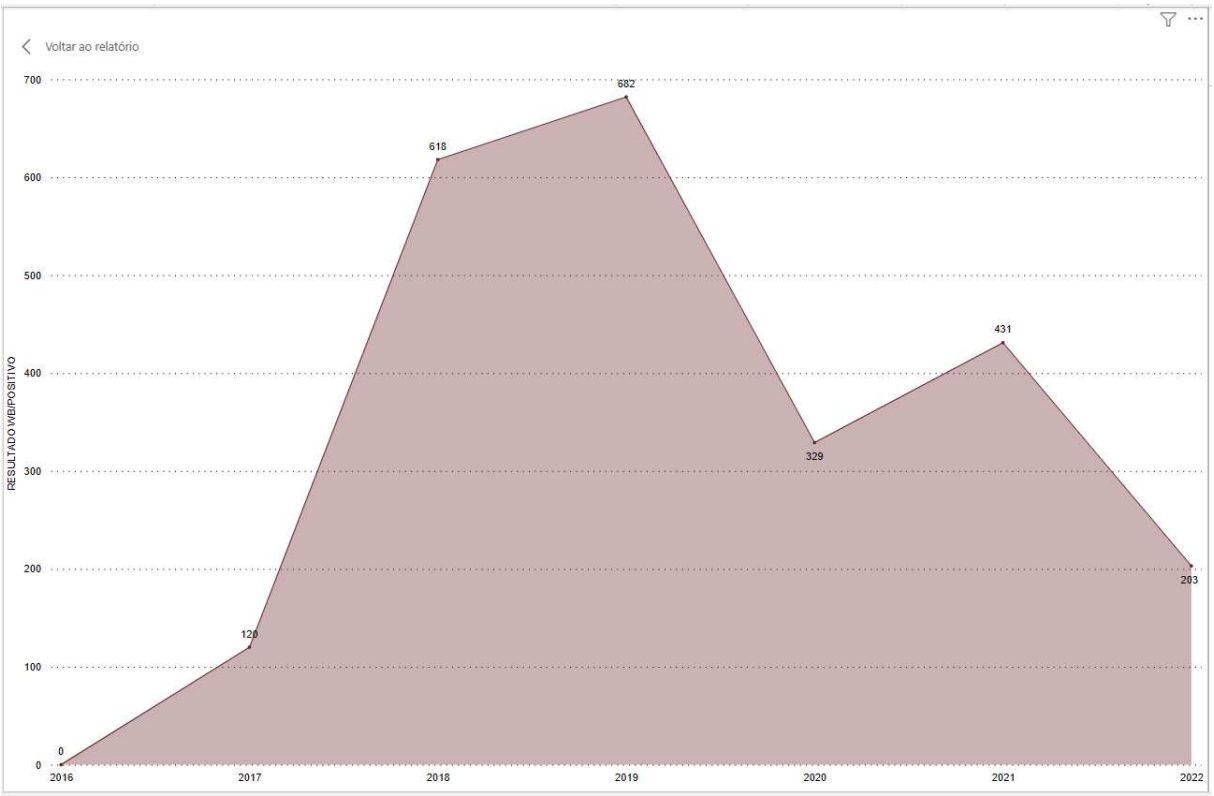


Fonte: O Autoria própria (2024)

4.4.1 Casos Positivos ao longo dos anos

O Gráfico 9 apresenta a evolução dos casos positivos de HTLV confirmados por WB entre 2016 e 2022. Em 2016, não foram registrados casos, mas em 2017 houve o início dos registros com 120 casos confirmados. Esse número aumentou significativamente em 2018, quando foram detectados 618 casos. O auge ocorreu em 2019, com 682 casos confirmados, marcando o pico do período analisado. No entanto, em 2020, houve uma queda acentuada para 329 casos, seguida por uma leve recuperação em 2021, com 431 casos confirmados. Em 2022, o número voltou a diminuir, com 203 casos registrados. Mostrando uma tendência de oscilações nos casos ao longo dos anos, destacando o pico em 2019 e variações nos anos subsequentes.

Gráfico 9 – Casos Positivos ao Longo dos Anos

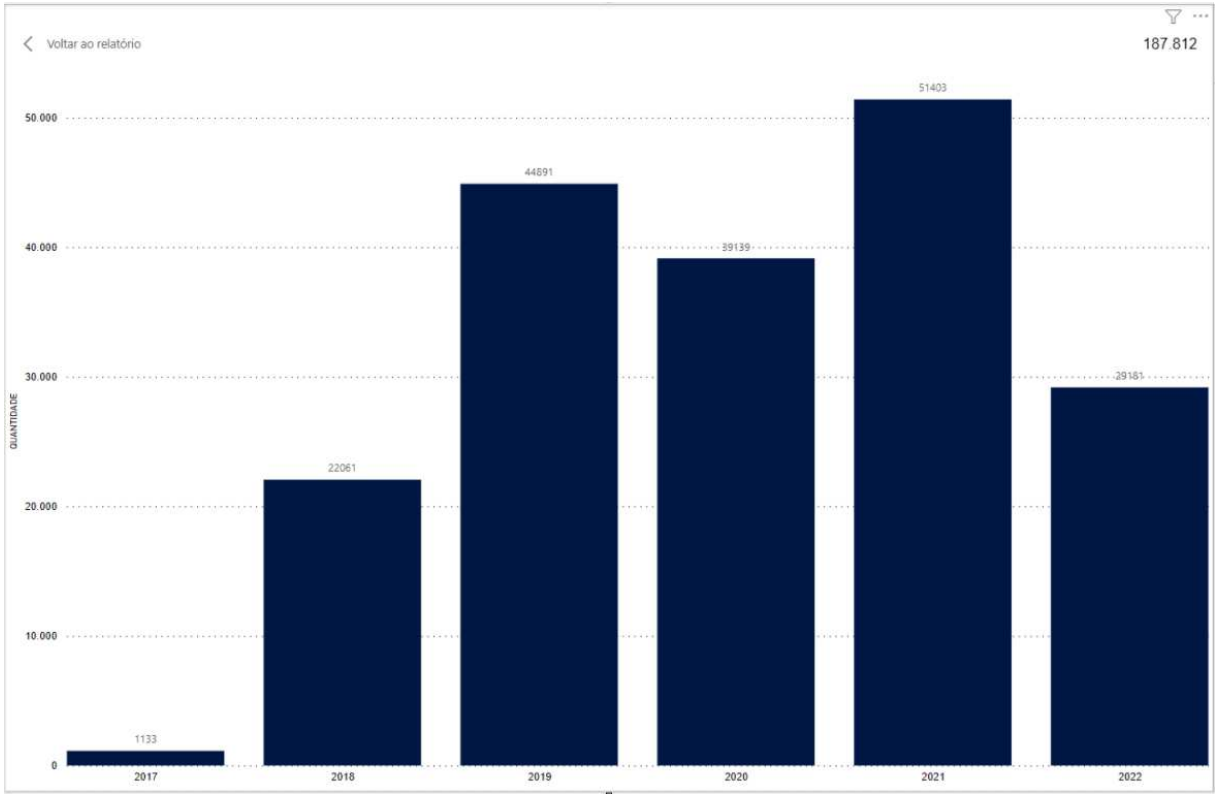


Fonte: O Autoria própria (2024)

4.4.2 Testes por Ano

O Gráfico 10, sobre a triagem de testes por ano, apresenta uma variação significativa na quantidade de testes realizados entre os anos de 2017 e 2022. O número de testes aumenta de forma constante entre 2017, quando foram realizados 1.133 testes, e 2021, que registra o pico com 51.403 testes. No entanto, em 2022, observa-se uma queda considerável, com 29.181 testes realizados.

Gráfico 10 – Teste por Ano



Fonte: O Autoria própria (2024)

4.4.3 Tipos de Casos

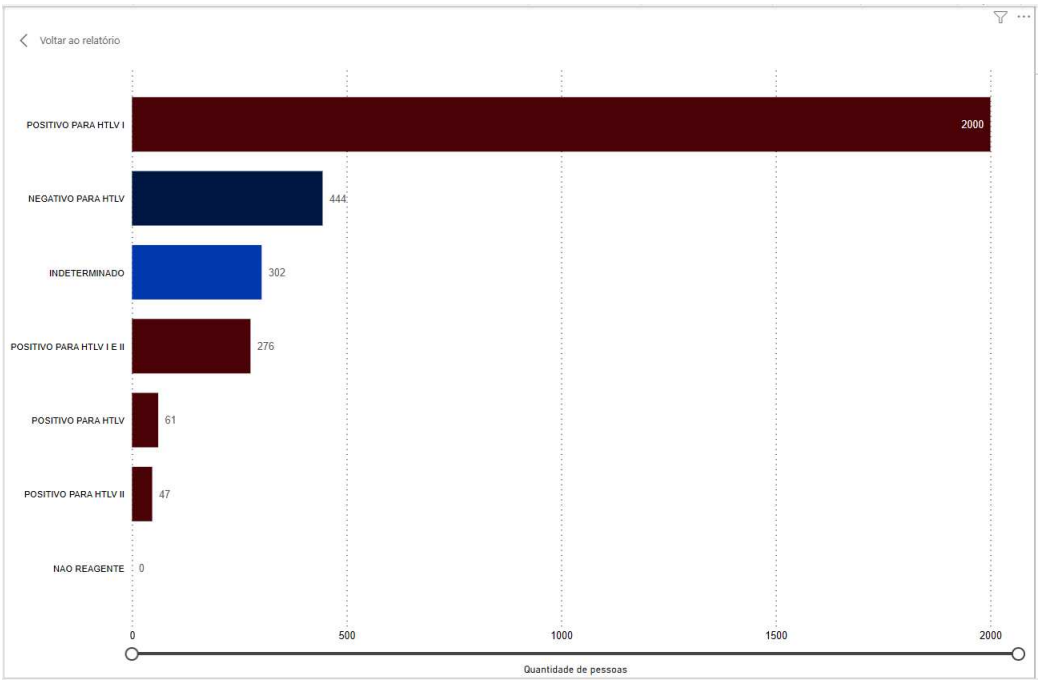
O Gráfico 11, é referente aos tipos de casos de HTLV revelando um predomínio significativo de casos positivos para o HTLV I, com um total de 2.000 pessoas diagnosticadas com essa variante do vírus, representando a maior proporção dos casos analisados e indicando a alta prevalência do HTLV I na amostra estudada.

Os resultados também mostram que 444 indivíduos apresentaram resultados negativos para HTLV, uma parcela significativa da população examinada não possui o vírus. No entanto, outros 302 casos foram classificados como "indeterminados", o que sugere a necessidade de testes adicionais para confirmar ou refutar a presença do vírus.

Além disso, foram identificados 276 casos positivos tanto para HTLV I quanto para HTLV II. Esse número reforça a coexistência dessas duas variantes em alguns pacientes, destacando a importância de um diagnóstico detalhado para diferenciar as variantes e suas possíveis implicações clínicas.

Há 61 casos identificados como "positivos para HTLV", sem distinção clara entre os tipos de variantes, indicando a necessidade de aprimoramento dos testes de especificidade, para determinar com maior se os pacientes estão infectados com HTLV I, HTLV II ou ambos. Por fim, 47 casos positivos para HTLV II foi detectado, evidenciando que essa variante é menos prevalente em comparação ao HTLV I.

Gráfico 11 – Tipo de Casos

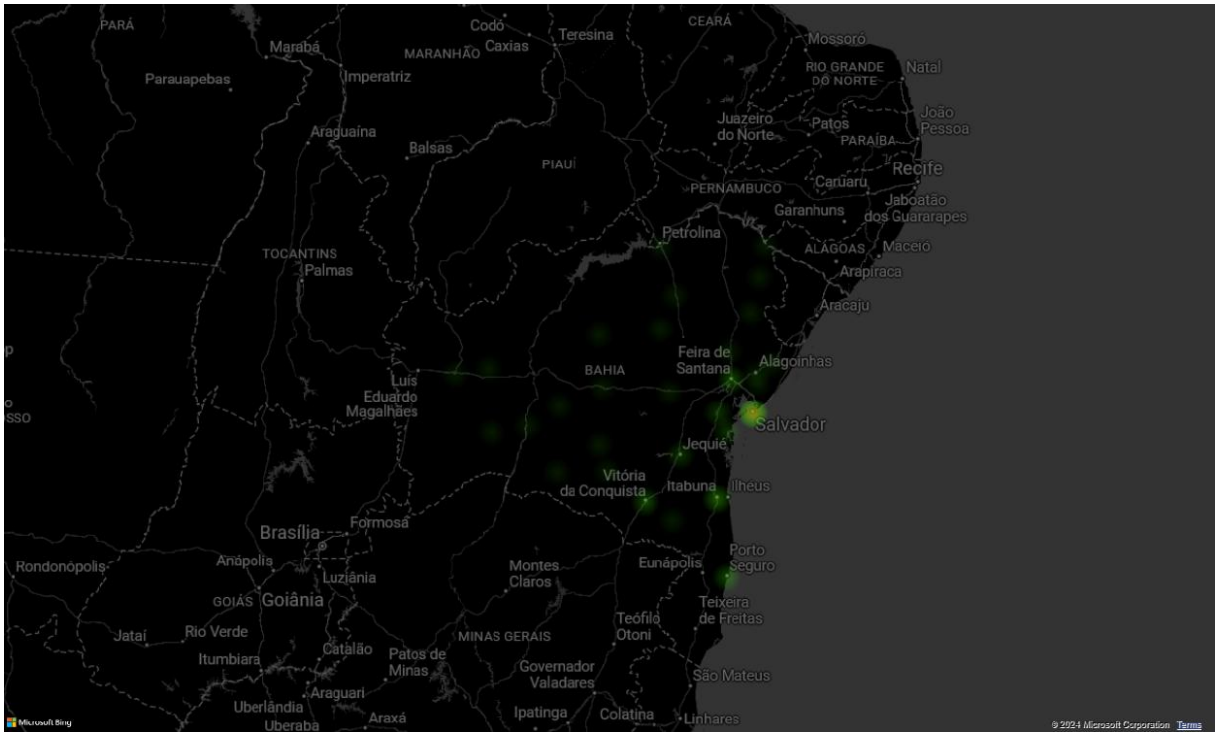


Fonte: O Autoria própria (2024)

4.4.4 Mapa de Casos

O Gráfico 12, é um mapa de calor que visualiza a distribuição geográfica de casos do HTLV no estado do nordeste brasileiro, com a maior concentração na região leste, notadamente na área que envolve Salvador. Esse *dashboard* fornece uma compreensão intuitiva de onde os casos estão mais prevalentes, com as áreas mais intensamente iluminadas representando uma maior densidade de casos. Observa-se uma distribuição desigual, com aglomerados notáveis ao longo da costa leste ou nas proximidades de cidades como Jequié e Itabuna.

Gráfico 12 – Mapa de Casos



Fonte: O Autoria própria (2024)

4.5 Razão/Sexo e Gestantes

A Figura 17, apresenta a distribuição do vírus HTLV nos municípios da Bahia, com foco na segmentação por sexo e ao grupo de gestantes. Um dos principais aspectos apresentados é a diferença entre os sexos ao longo dos anos, destacando qual grupo é mais afetado. A razão entre

os sexos é importante para comparar quantidades e avaliar proporções ou entender como uma quantidade se relaciona com outra parte, que serão mais detalhadas nas seções 4.5.1 ao 4.5.5.

Essas informações são fundamentais para identificar os grupos mais afetados e definir prioridades de intervenção. A análise da razão entre os sexos e o acompanhamento dos casos entre gestantes ajudam a entender melhor a dinâmica do vírus e a concentrar esforços nos grupos que mais precisam de atenção.

Figura 17 – Razão/Sexo e Gestantes

Fonte: O Autoria própria (2024)

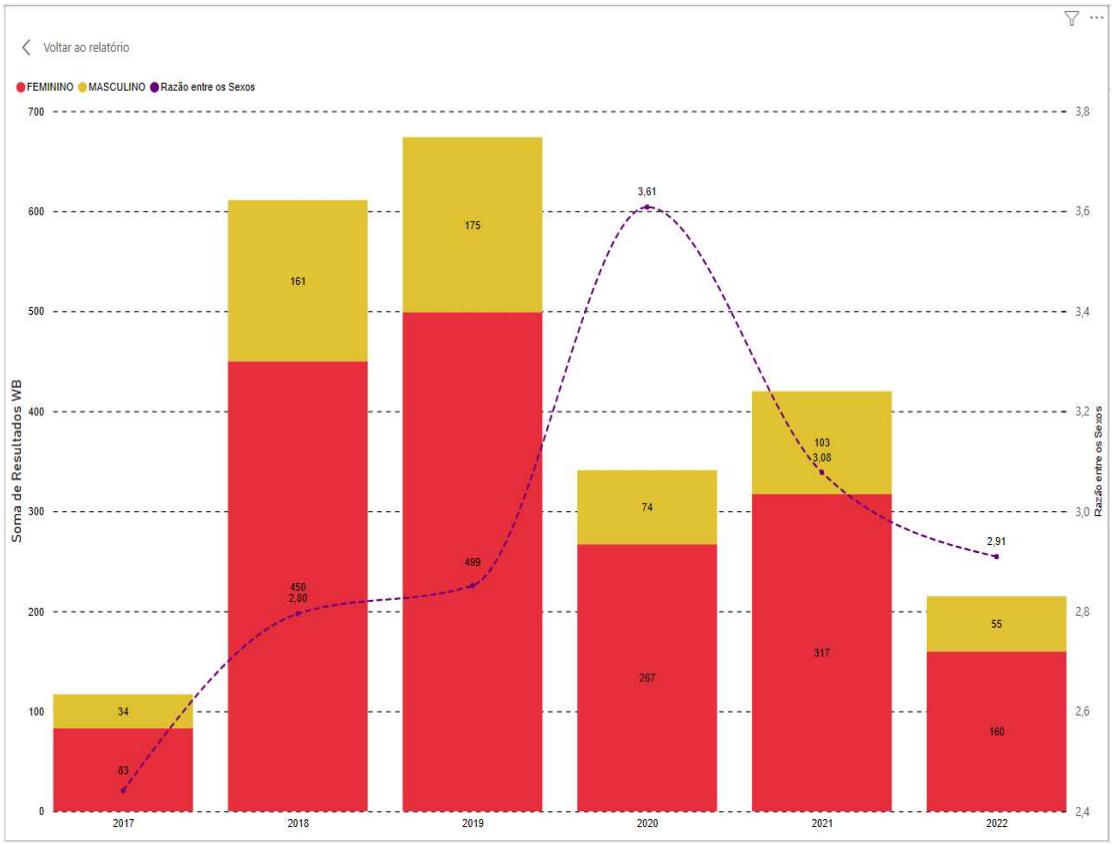
4.5.1 Razão entre os Sexos ao longo dos anos

O Gráfico 13, de colunas empilhadas e linhas, revela uma clara predominância de casos positivos para o vírus HTLV entre o sexo feminino, na cor vermelha em relação ao sexo masculino na cor amarela, ao longo dos anos de 2017 a 2022. Sendo composta por barras empilhadas, que mostram os números absolutos de infecções em cada sexo, e uma linha tracejada roxa que destaca a razão entre os sexos. Esta razão mede a proporção de mulheres infectadas em comparação aos homens, sendo representada no eixo da direita.

O número de casos positivos tem sido maior no sexo feminino durante todo o período analisado. A razão entre os sexos varia entre 2,44 em 2017 a 3,61 em 2020, indicando que, em certos momentos, o número de mulheres infectadas chegou a ser mais de três vezes superior ao de homens infectados. O pico dessa razão ocorreu em 2020, quando para cada homem com resultado positivo havia aproximadamente 3,61 mulheres infectadas. Esse dado reflete uma disparidade significativa na distribuição de infecções entre os sexos naquele ano.

Ao longo dos anos, observa-se uma tendência de aumento dessa razão entre os sexos até 2020, seguida de uma leve diminuição nos anos subsequentes, mas ainda assim permanecendo elevada em 2021 e 2022, com valores de 3,08 e 2,91, respectivamente. Esse comportamento sugere que, embora a razão tenha diminuído após 2020, a prevalência de casos positivos em mulheres continua sendo um fator relevante na dinâmica da infecção por HTLV.

Gráfico 13 – Razão entre os sexos ao longo dos Anos

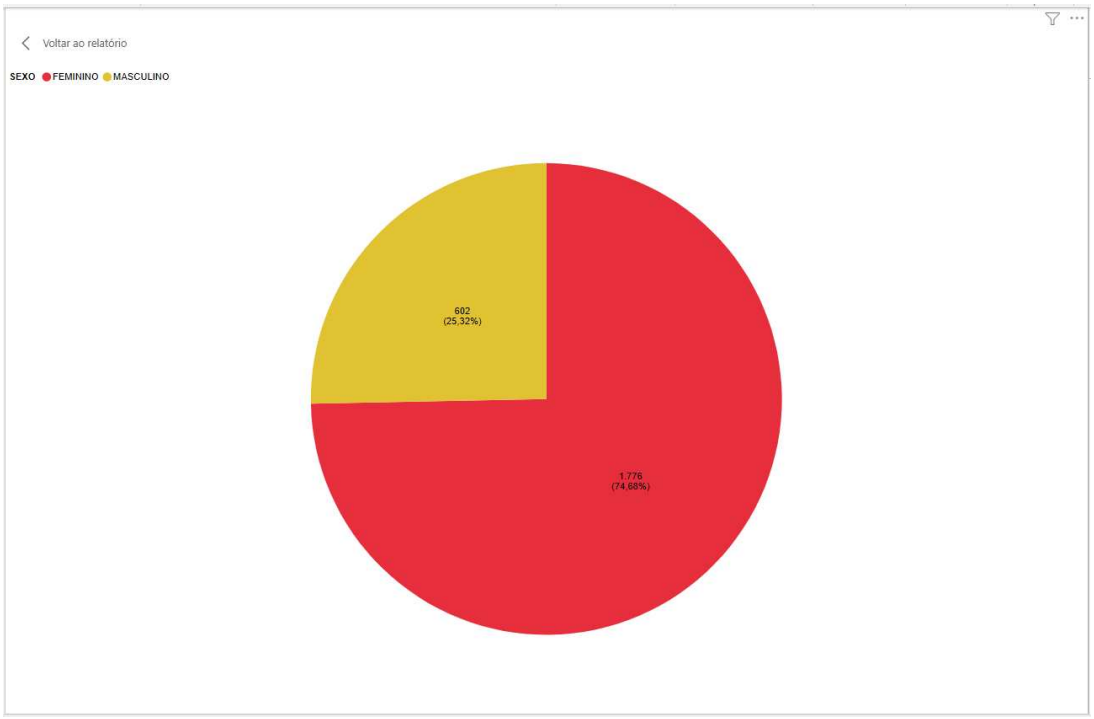


Fonte: O Autoria própria (2024)

4.5.2 Total e Porcentagem entre os Sexos

O Gráfico 14, de pizza, revela uma clara predominância de casos positivos de HTLV no sexo feminino na cor vermelha. Dos 2.378 casos totais, 1.776 correspondem a indivíduos do sexo feminino, representando 74,68% do total. Por outro lado, 602 casos são do sexo masculino na cor amarela, o que equivale a 25,32%. Essa disparidade significativa entre os gêneros indica uma maior vulnerabilidade ou prevalência da infecção pelo HTLV entre as mulheres.

Gráfico 13 – Total e Porcentagem entre os sexos



Fonte: O Autoria própria (2024)

4.5.3 Quantidade entre os sexos ao logo dos anos

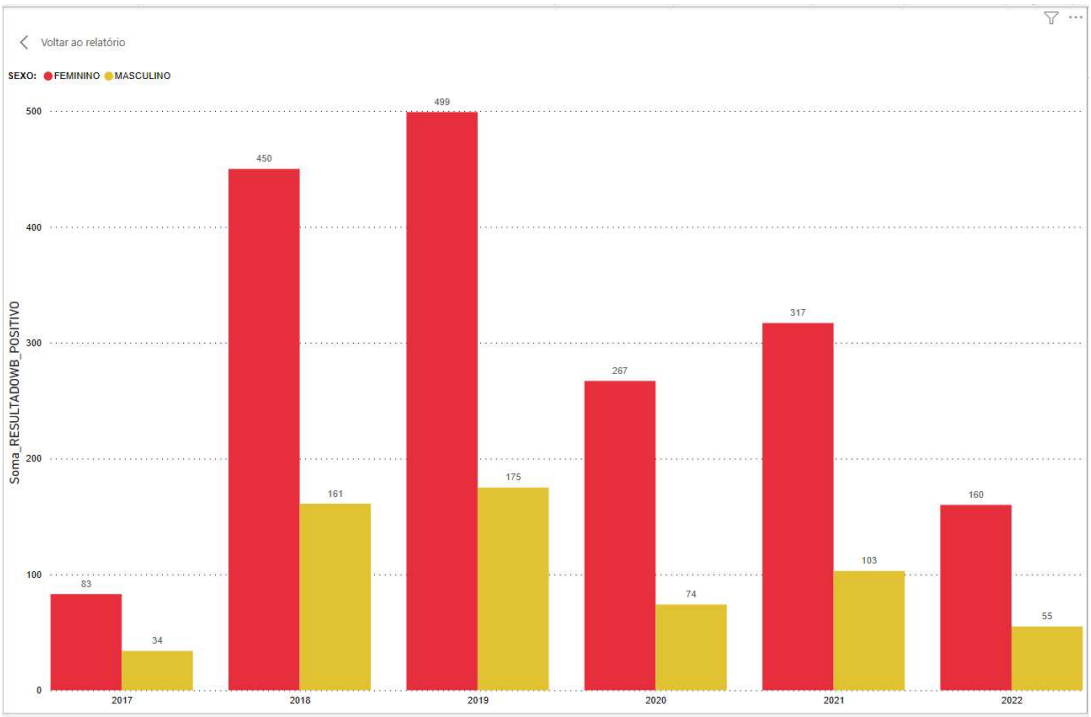
O Gráfico 15, de colunas clusterizado, evidencia uma variação significativa na quantidade de ocorrências entre os sexos ao longo dos anos. O sexo feminino, representado pela cor vermelha, mostra-se consistentemente acima do sexo masculino, representado pela cor amarela, em todos os anos analisados. Entre 2017 e 2019, o número de casos no sexo feminino aumentou notavelmente, atingindo seu ápice em 2019, com 499 ocorrências, valor significativamente superior ao observado no sexo masculino, que registrou 175 casos no mesmo ano. Esse dado revela uma disparidade clara em relação ao sexo masculino, que, embora

também tenha apresentado crescimento em 2018 e 2019, permaneceu em menor número. O pico de ocorrências masculinas em 2019, com 175 casos, contrasta fortemente com o número de caso no sexo feminino no mesmo período.

Analisando os dados, nota-se que, em 2017, o número de casos do sexo feminino (83) era inferior em termos absolutos, mas ainda superior ao do sexo masculino (34). A partir de 2018, observa-se uma ascensão acentuada nos casos do sexo feminino, que chegaram a 450, enquanto os casos do sexo masculino também aumentaram, alcançando 161. O ano de 2019 marca o pico do período analisado.

Nos anos seguintes, observa-se uma redução significativa nos números. Em 2020, os casos do sexo feminino caíram para 267, enquanto os casos do sexo masculino diminuíram para 74. Em 2021, teve um aumento significativo, em comparação com 2020, com 317 casos no sexo feminino e 103 no sexo masculino, mas voltou a cair em 2022, com os casos femininos caindo para 160 e os masculinos para 55.

Gráfico 15 – Quantidade entre os sexos ao longo dos anos



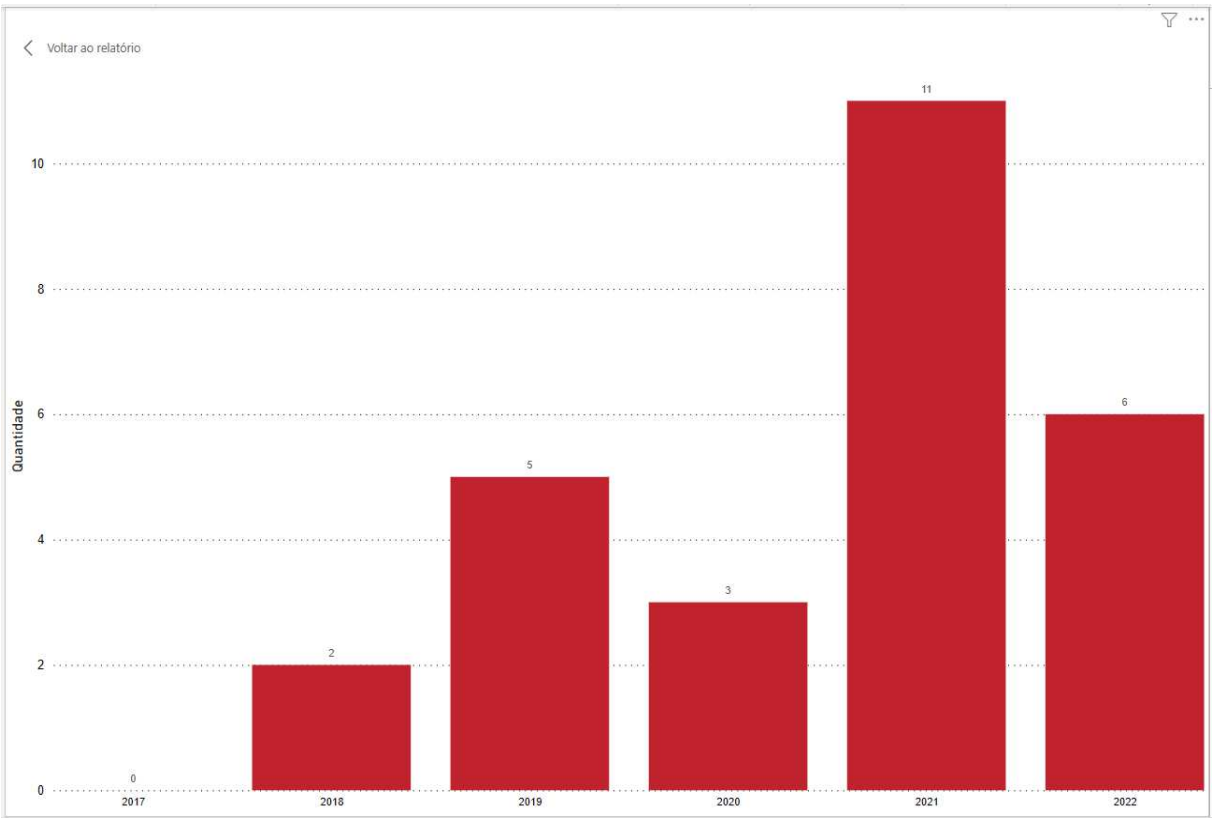
Fonte: O Autoria própria (2024)

4.5.4 Quantidade Gestantes

O Gráfico 16, de colunas empilhadas, apresenta a evolução do número de gestantes diagnosticadas com HTLV entre 2017 e 2022. Ao longo desse período, observa-se uma variação significativa nos números de casos. Em 2017, não foram registrados casos de HTLV em gestantes, enquanto em 2018 houve a identificação de 2 casos. Já em 2019, o número de diagnósticos subiu para 5 gestantes. No ano de 2020, houve uma queda, com 3 gestantes diagnosticadas.

O ano de 2021 registrou um aumento expressivo, atingindo o pico de 11 gestantes diagnosticadas com HTLV, representando o maior número de casos observados no período. Em 2022, embora ainda elevado em relação aos primeiros anos, o número de casos caiu para 6 gestantes diagnosticadas.

Gráfico 16 – Quantidade Gestantes



Fonte: O Autoria própria (2024)

4.6 Análise Interativa

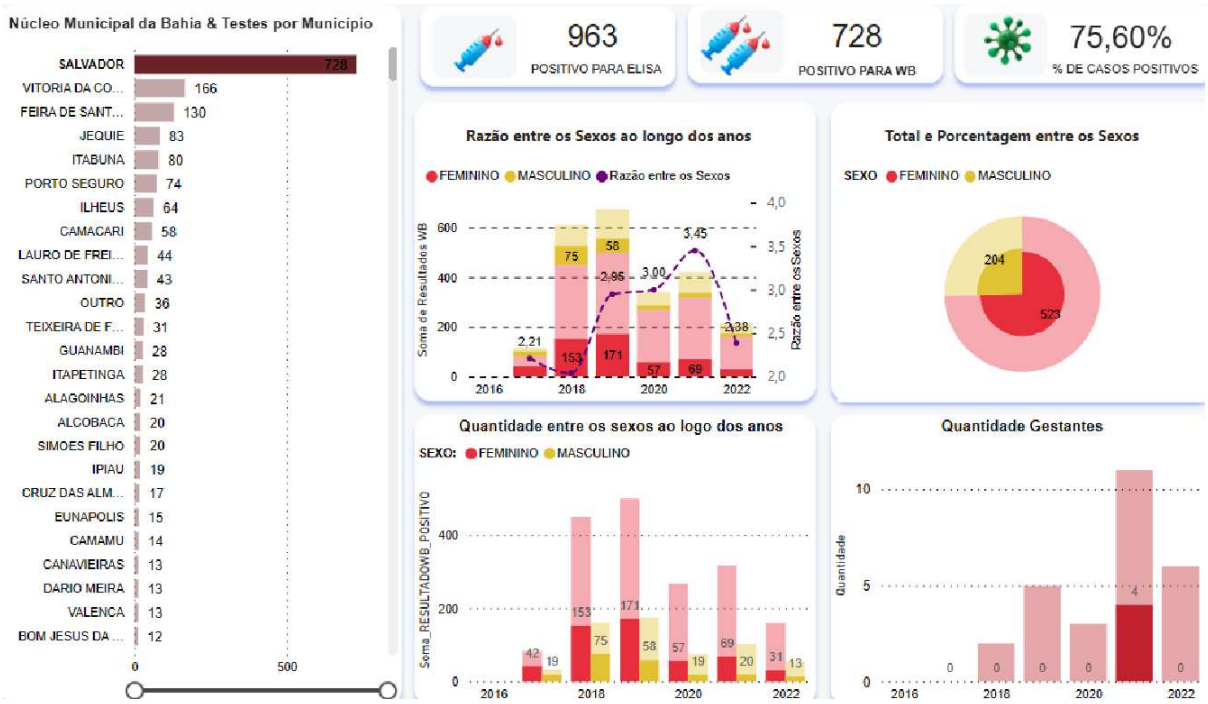
Essa análise interativa é aplicável a todos os painéis do *dashboard*, incluindo as seções 4.2, 4.3, 4.4 e 4.5, conforme demonstrado na seção 4.6.

A Figura 18 apresenta uma análise interativa do painel 'Razão/Sexo e Gestantes', composto por cinco gráficos e três indicadores que mostram a distribuição de casos positivos para HTLV. Esses casos foram identificados como 'Positivo para ELISA', com 963 casos, e confirmados como 'Positivo para WB', com 728 casos, ao selecionarmos o município de Salvador, Bahia, Brasil, no painel 'Núcleo Municipal da Bahia e Testes por Município', entre 2017 e 2022, representando 75,60% dos casos positivos de WB em relação aos casos ELISA. Em 2021, observa-se uma disparidade significativa na incidência do vírus entre os sexos, com uma razão de 3,45 mulheres infectadas para cada homem, conforme mostrado no gráfico 'Razão entre os sexos ao longo dos anos', indicando uma prevalência consideravelmente mais alta entre o sexo feminino.

Ao total, dos 728 casos confirmados por WB, o gráfico de “Total e Porcentagem entre os sexos” mostra 523 foram no sexo feminino na cor vermelha e 204 no sexo masculino na cor amarela, representando 71,99% e 28,01%, respectivamente, dos casos em Salvador.

E em “Quantidade entre os sexos ao longo dos anos”, observa-se uma oscilação nos casos de HTLV. Em Salvador, em 2017, foram registrados 42 casos em mulheres e 19 em homens. Os números aumentaram gradativamente nos anos subsequentes, atingindo um pico em 2019, com 171 casos em mulheres e 56 em homens. Houve uma queda em 2020, seguida de um aumento em 2021, e uma nova queda em 2022. O gráfico "Quantidade Gestantes" destaca que, em 2021, a incidência de HTLV entre gestantes também foi relevante, com quatro casos confirmados.

Figura 18 – Análise Interativa dos dados



Fonte: O Autoria própria (2024)

5 CONSIDERAÇÕES FINAIS

Este trabalho desenvolveu e analisou o Sistema de Apoio à Decisão Vírus Linfotrópico de Células T Humanas (SADH) na Bahia. A utilização de MD, aliado ao BI, na análise epidemiológica do HTLV visa apoiar a tomada de decisões relacionadas à disseminação do vírus no estado da Bahia, por meio da implementação de *dashboards* interativos. O estudo visou demonstrar como essas tecnologias podem ser aplicadas para fornecer uma visão mais clara e detalhada sobre a propagação do vírus.

Para compreender o impacto do HTLV na população baiana e a eficiência do uso de BI, foi definido como objetivo geral o desenvolver um dashboard de BI, com o objetivo de apoiar a tomada de decisões relacionadas à contaminação pelo vírus HTLV no estado da Bahia.

O primeiro objetivo específico foi apresentar uma proposta definida pela metodologia aplicada, o DSR, para desenvolver o sistema SADH e abordar os desafios do monitoramento epidemiológico do HTLV na Bahia. A aplicação do DSR seguiu um estudo estruturado, começando pela identificação do problema e da motivação, que envolveu compreender como e onde o vírus está disseminado no estado. Em seguida, o problema foi definido, destacando a ausência de dados e ferramentas epidemiológicas capazes de suportar análises eficazes. Na etapa de desenvolvimento dos artefatos, foram criados *dashboards* interativos que permitiram uma visualização clara e intuitiva dos dados, abordando informações gerais, tendências temporais, distribuição por sexo e dados específicos sobre gestantes. Esses artefatos foram demonstrados por meio de comparações entre os dados originais e os processados, utilizando uma análise exploratória inicial. E por fim, esse processo é iterativo, revisto se necessário ou aparecendo novos dados. Além disso, a aplicação do DSR garantiu que os artefatos desenvolvidos fossem não apenas funcionais, mas também eficazes na transmissão de informações relevantes.

O segundo objetivo específico foi coletar e tratar dados epidemiológicos sobre o HTLV na Bahia. Verificou-se que a estruturação e organização dos dados através de ferramentas como *Excel* e *Python* facilitaram a visualização e o entendimento dos dados e seus tipos.

O terceiro objetivo específico consistiu na implementação do método de ETL, essencial para assegurar que os dados fossem adequadamente extraídos, tratados, limpos e otimizados, garantindo assim uma base consistente e de qualidade para análise ao serem carregados no PBI.

O quarto objetivo específico estabeleceu a estrutura de relacionamento entre os dados e seus indicadores, criando um modelo dimensional eficaz, o SS. Esse modelo proporcionou uma visão clara e organizada dos dados, possibilitando a geração de *insights* valiosos. Como parte desse processo, foram realizadas as etapas de carregamento no DW e a modelagem dimensional na ferramenta PQ, assegurando uma estrutura consistente e otimizada para análises no PBI.

Por fim, o quinto objetivo específico desenvolveu um modelo de representação de alto nível utilizando a linguagem DAX, para análises e cálculos avançados em ferramentas de BI. Esse modelo resultou em *dashboards* dinâmicos e interativos, compostos por painéis e gráficos que facilitam tanto a visualização quanto a análise dos dados.

Com isso, a hipótese do trabalho de que a implementação de BI na saúde pública pode melhorar a compressão, a análise e gestão de dados epidemiológicos se confirmou, por permitir uma visualização clara dos dados coletados e facilitar a criação de cenários epidemiológicos.

Sendo assim, a utilização da mineração de dados e do PBI para a criação de *dashboards* epidemiológicos auxilia significativamente na compreensão da disseminação do HTLV, melhorando a tomada de decisão. Os instrumentos de coleta de dados permitiram uma análise completa e detalhada, com a limpeza e tratamento de grandes volumes de dados, garantindo precisão e consistência nos resultados apresentados através dos dashboards interativos.

Os resultados obtidos e o cumprimento integral dos objetivos reforçam a contribuição deste estudo para a compreensão da disseminação do HTLV na Bahia, oferecendo uma abordagem eficiente. Esses êxitos destacam a relevância da pesquisa no contexto científico e abrem caminhos para futuros desenvolvimentos na área.

6 TRABALHOS FUTUROS

Durante o desenvolvimento deste trabalho, novas perspectivas foram identificadas, abrindo caminho para pesquisas futuras que poderiam complementar e expandir os resultados obtidos. As seguintes possibilidades são sugeridas:

1. Expansão e Continuidade do Projeto, com modelagem preditiva:
   - Aprofundar a análise de dados usando Inteligência Artificial Generativa, como o *Chat GPT* para prever padrões epidemiológicos mais complexos. Isso permitiria a criação de novas funcionalidades e uma análise mais detalhada.

   - Explorar o uso de modelos de *Machine Learning* para prever a disseminação do HTLV em diferentes cenários epidemiológicos, permitindo a criação de alertas automáticos. Modelos preditivos como redes neurais ou regressão poderiam ser utilizados.

   - Explorar modelos híbridos que combinem aprendizado de máquina com abordagens epidemiológicas clássicas, como o modelo SIR (Susceptíveis, Infectados e Recuperados). Essa combinação permitiria gerar previsões mais robustas e adaptadas às particularidades de diferentes doenças e cenários epidemiológicos.

2. Aplicação de *Dashboards* Interativos para Outras Doenças ou Vírus:
   - Expandir o uso de painéis interativos para monitorar outras doenças ou vírus negligenciados ou de alta prevalência no Brasil, como a MPOX, utilizando os mesmos princípios do sistema atual, criando cenários para diferentes doenças e fornecendo uma ferramenta versátil para gestores públicos.

3. Criação de um Sistema Nacional Integrado:
   - Expandir o sistema atual para integrar dados epidemiológicos de outros estados do Brasil ou de países com alta prevalência de HTLV, ou outras epidemias. Isso criaria uma plataforma de monitoramento nacional ou internacional, permitindo uma visão mais ampla e a troca de dados entre instituições.

4. Análise Temporal Automatizada e Tendências a Longo Prazo:

- Desenvolver *scripts* ou algoritmos para realizar análises temporais automatizadas e contínuas, permitindo evolução dos casos ao longo de décadas. Essa análise ajudaria a entender como mudanças sociais ou políticas afetam a transmissão do vírus.

5. Desenvolvimento de uma Interface Web para Interação com o Sistema
   - Criar uma interface/site pelo qual pesquisadores, profissionais de saúde ou outros tipos de usuários interajam com o sistema dos dados do HTLV, ou outra epidemia, tornando as funcionalidades mais acessíveis.

REFERÊNCIAS

AMOUSSA, Adjile Edjide Roukiyath. **Origem do HTLV-1aA no Brasil: contribuição da população africana na introdução do vírus**. 2018. 97 f. Tese (Doutorado em Patologia) – Universidade Federal da Bahia, Faculdade de Medicina. Fundação Oswaldo Cruz, Instituto Gonçalo Moniz, Salvador, 2018

ARAÚJO, Érica Câmara. **Análise dos dados epidemiológicos e da incidência da COVID-19 no município de Assú no estado do Rio Grande do Norte**. 2021. 130 f. Trabalho de Conclusão de Curso (Bacharelado em Sistemas de Informação) – Universidade Federal Rural do Semi-Árido, Angicos, 2021

BRITO, Thiago da Silva; OLIVEIRA, Rafael da Silva. **Solução de Business Intelligence utilizando a plataforma Microsoft na área da segurança pública**. 2017. 66 f. Projeto de Graduação (Bacharelado em Sistemas de Informação) – Universidade Federal do Estado do Rio de Janeiro, Escola de Informática Aplicada, Rio de Janeiro, 2017

BUTANTAN. **Entenda o que é uma pandemia e as diferenças entre surto, epidemia e endemia**. São Paulo: Butantan, 2023. Disponível em: https://butantan.gov.br/covid/butantan-tira-duvida/tira-duvida-noticias/entenda-o-que-e-uma-pandemia-e-as-diferencas-entre-surto-epidemia-e-endemia. Acesso em: 20 ago. 2023.

CHEN, Hsinchun; CHIANG, Roger H. L.; STOREY, Veda C. **Business intelligence and analytics: From big data to big impact. MIS Quarterly**, v. 36, n. 4, p. 1165-1188, 2012

ESTEVES, Marisa Araújo**. Desenvolvimento e Exploração de uma Nova Geração de Ferramentas de Business Intelligence para o Apoio à Decisão e a Prática Clínica em Unidades Hospitalares**. 2016. Dissertação (Mestrado Integrado em Engenharia Biomédica) – Universidade do Minho, Escola de Engenharia, 2016

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. **From Data Mining to Knowledge Discovery in Databases**. AI Magazine, Menlo Park, v. 17, n. 3, p. 37-54, 1996.

FILHO, Aloísio S. Nascimento; MURARI, Thiago B.; FERREIRA, Paulo; SABA, Hugo; MORET, Marcelo A. **A spatio-temporal analysis of dengue spread in a Brazilian dry climate region**. Scientific Reports, v. 11, 2021. DOI: https://doi.org/10.1038/s41598-021-91306-z.

GARCIA, Ionara Ferreira da Silva; HENNINGTON, Élida Azevedo. **HTLV na agenda de governo: o caso da Bahia e de Minas Gerais, Brasil**. Cadernos de Saúde Pública, Rio de Janeiro, v. 37, n. 11, p. e00303420, 2021. DOI: 10.1590/0102-311X00303420.

GIL, Antônio Carlos. **Como elaborar projetos de pesquisa**. 4. ed. São Paulo: Atlas, 2002.

GONÇALVES, Denise Utsch; PROIETTI, Fernando Augusto; RIBAS, João Gabriel Ramos; ARAÚJO, Marcelo Grossi; PINHEIRO, Sônia Regina; GUEDES, Antônio Carlos; CARNEIRO-PROIETTI, Anna Bárbara F. **Epidemiology, Treatment, and Prevention of Human T-Cell Leukemia Virus Type 1-Associated Diseases**. Clinical Microbiology Reviews, v. 23, n. 3, p. 577-589, jul. 2010.

GOOGLE. **Conheça o Colab**. Disponível em: https://colab.research.google.com/. Acesso em: 01 set. 2024.

MAI, Scheila; GUIMARÃES, Cristian Fabiano; SILVA, Jarder Marques; HINKEL, José Henrique Schwanck. **O uso das tecnologias na democratização da informação em saúde.** Revista de Gestão em Sistemas de Saúde - RGSS, São Paulo, v. 6, n. 3, p. 210-218, set./dez. 2017. DOI: 10.5585/rgss.v6i3.287.

MICROSOFT. **Power BI**. Disponível em: https://powerbi.microsoft.com/pt-br/. Acesso em: 6 dez. 2023.

MICROSOFT. **Visão Geral do DAX**. 2023. Disponível em: https://learn.microsoft.com/pt-br/dax/dax-overview. Acesso em: 01 set. 2024.

MINISTÉRIO DA SAÚDE. **No Brasil, estima-se que entre 800 mil e 2,5 milhões de pessoas vivam com o vírus HTLV**. Disponível em: https://www.gov.br/aids/pt-br/assuntos/noticias/2023/fevereiro/no-brasil-estima-se-que-entre-800-mil-e-2-5-milhoes-de-pessoas-vivam-com-o-virus-htlv. Acesso em: 20 ago. 2023.

MIRANDA, Angélica Espinosa; ROSADAS, Carolina; ASSONE, Tatiane; PEREIRA, Gerson Fernando Mendes; VALLINOTO, Antonio Carlos Rosário; ISHAK, Ricardo. **Strengths, weaknesses, opportunities and threats (SWOT) analysis of the implementation of public health policies on HTLV-1 in Brazil**. Frontiers in Medicine, v. 9, art. 859115, 2022. DOI: https://doi.org/10.3389/fmed.2022.859115.

NUNES DA SILVA, Aidê; ARAÚJO, Thessika Hialla Almeida; BOA-SORTE, Ney; FARIAS, Giovanne; GALVÃO-BARROSO, Ana Karina; CARVALHO, Antônio de; VICENTE, Ana Carolina; GALVÃO-CASTRO, Bernardo; GRASSI, Maria Fernanda Rios**. Epidemiological and molecular evidence of intrafamilial transmission through sexual and vertical routes in Bahia, the state with the highest prevalence of HTLV-1 in Brazil.** PLoS Neglected Tropical Diseases, v. 17, n. 9, p. e0011005, 2023. Disponível em: https://doi.org/10.1371/journal.pntd.0011005. Acesso em: 22 ago. 2023.

PATRÍCIO, Thiago Seti; MAGNONI, Maria da Graça Mello. **Mineração de dados e big data na educação.** Revista GEMInIS, São Carlos, UFSCar, v. 9, n. 1, p. 57-75, jan./abr. 2018. DOI: 10.4322/2179-1465.0901004.

PEFFERS, Ken; TUUNANEN, Tuure; ROTHENBERGER, Marcus A.; CHATTERJEE, Samir. **A design science research methodology for information systems research**. Journal of Management Information Systems, v. 24, n. 3, p. 45–77, 2007. DOI: 10.2753/MIS0742-1222240302.

PRESSMAN, Roger S. **Engenharia de software: uma abordagem profissional**. 7. ed. Porto Alegre: AMGH, 2011.

PYTHON INSTITUTE**. Python® – a linguagem de hoje e de amanhã. [S.l.]: Python Institute, 2023.** Disponível em: https://pythoninstitute.org/about-python. Acesso em: 01 set. 2024.

RODRIGUES, Ricardo Azevedo de Oliveira. **Ferramentas de business intelligence para apoio à gestão pública no estado da Bahia na pandemia da COVID-19 2020/2021**. 2021. Dissertação (Mestrado) – Programa de Pós-Graduação Stricto Sensu em Modelagem Computacional e Tecnologia Industrial, Centro Universitário SENAI CIMATEC, Salvador, 2021. 86 p.

ROSADAS, Carolina; ASSONE, Tatiane; SERENO, Leandro; MIRANDA, Angelica Espinosa; MAYORGA-SAGASTUME, Rubén; FREITAS, Marcelo A.; TAYLOR, Graham P.; ISHAK, Ricardo. **"We Need to Translate Research Into Meaningful HTLV Health Policies and Programs"**: Webinar HTLV World Day 2021. Frontiers in Public Health, v. 10, artigo 883080, 2022. Disponível em: https://doi.org/10.3389/fpubh.2022.883080. Acesso em: 25 nov. 2023.

ROSADAS, Carolina; BRITES, Carlos; ARAKAKI-SÁNCHEZ, Denise; CASSEB, Jorge; ISHAK, Ricardo. **Protocolo Brasileiro para Infecções Sexualmente Transmissíveis 2020: infecção pelo vírus linfotrópico de células T humanas (HTLV)**. Epidemiol. Serv. Saude, Brasília, v. 30, n. esp. 1, e2020605, 2021. Disponível em: http://doi.org/10.1590/S1679-49742020000600015.esp1. Acesso em: 25 nov. 2023.

TEIXEIRA, Enise Barth. **A análise de dados na pesquisa científica: importância e desafios em estudos organizacionais.** Desenvolvimento em Questão, v. 1, n. 2, p. 177-201, jul./dez. 2003.