

AUTOMAÇÃO PARA A COLETA E O TRATAMENTO DE DADOS NO SISTEMA DE TABULAÇÃO ONLINE DO DATASUS

WOLLMANN, MURILO ANTÔNIO¹; PATATT, LEONARDO ANTÔNIO¹; GOETZE, ANA PAULA¹; BAGATINI, DANIELA DUARTE DA SILVA.²; LORA, PRISCILA SCHMIDT³; MONTEIRO, GABRIELA DE ANDRADE⁴; MÜLLER, ANA PAULA WERNZ DA CUNHA⁵; SILVA, DÉBORA OLIVEIRA DA⁶; PEDRON, CRISTIANE DREBES⁷

RESUMO

A coleta e análise de dados são essenciais para monitorar doenças e desenvolver estratégias de intervenção. No entanto, o processo manual no TABNET (DATASUS) dificulta a atualização contínua. Este trabalho desenvolveu uma aplicação para automatizar a busca e tratamento desses dados. A metodologia envolveu: (1) compreensão do TABNET; (2) estudo do Selenium para automação web; (3) estudo do Pandas para análise de dados; (4) seleção de dados relevantes sobre o câncer de mama com apoio de uma especialista mastologista; (5) automação do **download** dos dados; (6) automação do tratamento correto das informações coletadas; (7) testes para validação da aplicação. O sistema mostrou-se eficiente e confiável, eliminando tarefas manuais repetitivas e reduzindo o tempo necessário para a coleta e o tratamento dos dados epidemiológicos de aproximadamente 6 horas para aproximadamente 8 minutos.

PALAVRAS-CHAVE: Extração de dados, Automação de tarefas, TABNET, DATASUS, Câncer de mama.

AUTOMATION FOR DATA COLLECTION AND PROCESSING IN THE DATASUS ONLINE TABULATION SYSTEM

ABSTRACT

Data collection and analysis are essential for monitoring diseases and developing intervention strategies. However, the manual process in TABNET (DATASUS) makes continuous updating difficult. This work developed an application to automate the search and processing of these data. The methodology involved: (1) understanding TABNET; (2) studying Selenium for web automation; (3) studying Pandas for data analysis; (4) selecting relevant data on breast cancer with the support of a specialist mastologist; (5) automating data download; (6) automating the correct processing of the collected information; (7) testing to validate the application. The system proved to be efficient and reliable, eliminating repetitive manual tasks and reducing the

¹ Acadêmicos do Curso de Ciência da Computação na Universidade de Santa Cruz do Sul (UNISC).

² Doutora e docente do Departamento de Engenharias, Arquitetura e Computação, Programa de Pós-Graduação em Sistemas e Processos Industriais, Universidade de Santa Cruz do Sul <bagatini@unisc.br>

³ Doutora e docente do Centro de Ciências da Saúde, Universidade do Vale do Rio dos Sinos <plora@unisinos.br>

⁴ Mestranda do Programa de Pós-Graduação de Engenharia de Produção e Sistemas, Universidade do Vale do Rio dos Sinos <gmonteiro1109@gmail.com>

⁵ Mastologista, CEO da ClickPalm <ana.paula.w.c.muller@gmail.com>

⁶ Doutora e docente do Programa de Pós-Graduação de Engenharia de Produção e Sistemas, Universidade do Vale do Rio dos Sinos <deboraoliveira@unisinos.br>

⁷ Doutora e docente do Programa de Pós-Graduação em Administração e Gestão de Projetos, Universidade Nove de Julho <cdpedron@uni9.pro.br>

time required to collect and process epidemiological data from approximately 6 hours to approximately 8 minutes.

KEYWORDS: Data extraction, automated tasks, TABNET, DATASUS, Breast cancer.

1 INTRODUÇÃO

A coleta e análise de dados desempenham um papel essencial na saúde pública, pois permitem o monitoramento da ocorrência de doenças, a identificação de padrões epidemiológicos e o desenvolvimento de estratégias de intervenção (CHEN *et al.*, 2022). A evidência científica é um dos três pilares da prática baseada em evidência, um modelo assistencial que comprova melhorias nos desfechos de saúde e na otimização dos gastos públicos. No entanto, para gerar evidências sólidas, é essencial contar com dados de qualidade, que muitas vezes são escassos ou de difícil acesso (NILSEN *et al.*, 2024). No Brasil, o Departamento de Informática do SUS (DATASUS) é a principal plataforma de gestão e disseminação de dados em saúde pública. Entre suas ferramentas, o TABNET se destaca como um sistema de tabulação *on-line* que possibilita a consulta a dados de diversas bases epidemiológicas (DATASUS, 2025).

No entanto, apesar da ampla disponibilização de dados pelo DATASUS, os desafios persistem em relação à sua organização e acessibilidade. A ausência de processos automatizados de extração e estruturação dos dados no TABNET torna a atualização contínua dessas informações um procedimento trabalhoso e propenso a erros. Estudos indicam que ferramentas automatizadas de extração e tratamento de dados são essenciais para facilitar análises epidemiológicas e melhorar a tomada de decisões em saúde pública (MILLER *et al.*, 2024). Dentre as vantagens, a automação de extração de dados reduz significativamente o tempo necessário para a coleta e análise de informações, aumentando a eficiência do processo e reduzindo falhas humanas (WALKER *et al.*, 2022).

Diante deste cenário, tecnologias de automação, como a biblioteca Selenium, de acesso gratuito, podem ser utilizadas para simular ações do usuário em ambientes *web*. A automação, quando associada a soluções de manipulação e análise de dados, como a biblioteca Pandas, amplamente utilizada, permite automatizar processos repetitivos e de grande volume de dados (MUTHUKADAN, 2025; DOCUMENTAÇÃO DA BIBLIOTECA PANDAS, 2025).

Identificar maneiras de facilitar a extração e a estruturação de dados pode beneficiar a implementação da prática baseada em evidências, além de contribuir para a otimização de recursos (processos e produtos), aprimoramento do acompanhamento contínuo (atenção mais eficaz), tratamentos coordenados (integração dos níveis de cuidado), diagnóstico rápido e redução de erros (com informações mais completas) (BRAGA *et al.*, 2024; LEITE *et al.*, 2024; RAHMAN *et al.*, 2024). Nesse contexto, entre as diversas condições de saúde que demandam monitoramento contínuo, o câncer de mama destaca-se pela importância do diagnóstico precoce, tornando indispensável o acesso a dados atualizados e confiáveis para subsidiar ações de prevenção, rastreamento e planejamento em saúde.

No Brasil, o câncer de mama apresenta alta incidência, o que torna essencial o diagnóstico precoce e a prevenção. Estima-se que, em 2024, ocorreram 73.610 novos casos de câncer de mama, com uma taxa de

incidência de 66,54 casos a cada 100 mil mulheres (INCA, 2024). As regiões Sul e Sudeste apresentam as maiores taxas de incidência e mortalidade (SECRETARIA ESTADUAL DE SAÚDE DO RIO GRANDE DO SUL, 2024; OBSERVATÓRIO DO CÂNCER RS, 2024), sendo o Rio Grande do Sul o terceiro maior estado em número de casos no Brasil, atrás de São Paulo e Minas Gerais (VASCONCELLOS *et al.*, 2024). Sendo assim, a obtenção de dados atualizados e confiáveis sobre essa doença, bem como sobre o acesso ao diagnóstico, é fundamental para embasar a formulação e o aprimoramento de políticas públicas. Essas informações permitem a implementação de ações mais eficazes, alinhadas à prática baseada em evidências.

Com base nesse contexto, o objetivo principal deste trabalho é desenvolver uma aplicação que automatiza a coleta, organização e tratamento dos dados do TABNET. A solução consiste em um *script* executado no ambiente Google Colab, que realiza automaticamente a navegação no sistema TABNET, seleciona os filtros e variáveis previamente definidos, efetua o download das tabelas de interesse e executa o tratamento e a padronização dos dados, gerando arquivos estruturados para análise. Essa automação permite a atualização contínua das informações relacionadas à linha de cuidado do câncer de mama, reduzindo o tempo de processamento, minimizando erros decorrentes da manipulação manual e fornecendo dados confiáveis para subsidiar pesquisas e a tomada de decisões em saúde pública.

2 FUNDAMENTAÇÃO TEÓRICA

O Departamento de Informática do SUS (DATASUS) é um dos principais recursos tecnológicos utilizados pelo governo brasileiro para armazenar e processar dados relacionados à saúde pública. Criado para apoiar a formulação de políticas públicas e a tomada de decisão baseada em evidências, o DATASUS reúne informações sobre morbidade, mortalidade, assistência hospitalar, vacinação e diversos outros aspectos da saúde da população (DATASUS, 2025).

Dentre as ferramentas disponíveis no DATASUS, o TABNET se destaca como um sistema de tabulação de dados *on-line* que permite a consulta e análise de informações epidemiológicas. Com essa ferramenta, pesquisadores podem gerar tabelas estatísticas personalizadas a partir de grandes bases de dados, facilitando o monitoramento de tendências em saúde pública (DATASUS, 2025). No entanto, o TABNET apresenta limitações relacionadas à extração automatizada de dados, exigindo que os usuários realizem consultas e *downloads* manualmente, o que pode ser um processo demorado e suscetível a erros.

A extração manual de dados de sistemas como o TABNET pode ser um processo demorado e suscetível a falhas devido à diversidade de formatos e padrões de formatação dos documentos. Segundo Zaman *et al.* (2020), a falta de um modelo universal para extração de informações é um dos principais desafios na automação desses processos, tornando essencial o desenvolvimento de metodologias padronizadas para garantir a integridade e a confiabilidade dos dados epidemiológicos.

Nesse contexto, a automação da extração e análise de dados de saúde surge como uma solução essencial para aprimorar a eficiência e confiabilidade dos sistemas de informação epidemiológica. Métodos automatizados permitem não apenas a coleta rápida e precisa de informações, mas também a padronização dos dados, reduzindo inconsistências e minimizando erros humanos (CAPPELLO *et al.*, 2024; MCGUINNESS *et al.*, 2022). A automação pode ser implementada por meio de *scripts* e ferramentas gratuitas que interagem diretamente com as bases de dados estruturadas, permitindo a extração sistemática e a organização das

informações em formatos adequados para análise. Além disso, estudos mostram que técnicas de extração de informações, como aprendizado de máquina supervisionado, ontologias e regras baseadas em conhecimento, são eficazes para estruturar e organizar grandes volumes de dados semiestruturados e não estruturados (ZAMAN *et al.*, 2020).

A automação da coleta e análise de dados, portanto, apresenta-se como uma ferramenta importante para a gestão eficiente dos sistemas de saúde. Dados epidemiológicos bem estruturados permitem a identificação precoce de surtos de doenças, a avaliação da eficácia das políticas de saúde e o planejamento de intervenções estratégicas para a melhoria dos serviços oferecidos à população. No entanto, erros na coleta ou inconsistências na organização das informações podem levar a conclusões equivocadas e comprometer a formulação de decisões (MCGUINNESS *et al.*, 2022). Por isso, a adoção de metodologias padronizadas e o uso de ferramentas tecnológicas, como ontologias para estruturar dados de forma consistente, são fundamentais para garantir a confiabilidade das informações utilizadas em estudos epidemiológicos (ZAMAN *et al.*, 2020; CAPPELLO *et al.*, 2024).

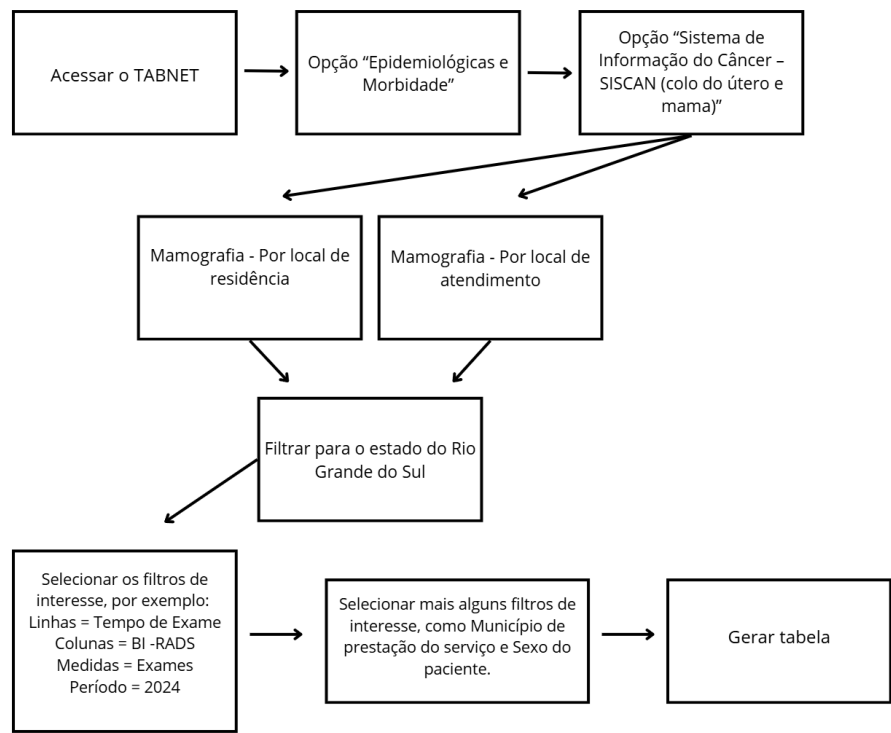
3 MATERIAIS E MÉTODOS

Considerando que se trata do desenvolvimento de uma aplicação para a área da saúde, foram utilizadas orientações do método *Design Science Research* (DSR), que utiliza práticas de *design* para a produção de um artefato inovador. Três fatores fundamentais orientam a pesquisa DSR (LACERDA, DRESCH, ANTUNES, 2013; DRESCH, LACERDA, ANTUNES JÚNIOR, 2015): (i) relevância, assegurando que os resultados atendam às necessidades da pesquisa, principalmente para a linha de cuidado em questão; (ii) rigor, garantindo que os resultados sejam confiáveis e contribuam para a identificação, extração e estruturação de dados de saúde; e (iii) viabilidade real, demonstrando resultados práticos (VAISHNAVI, KUECHLER, 2004).

Para o desenvolvimento da aplicação de automação da coleta e tratamento de dados epidemiológicos do TABNET foi estruturada sete etapas principais (adaptadas do DSR), descritas a seguir.

1. Compreensão do sistema de tabulação do DATASUS (TABNET): a primeira etapa consistiu em um estudo aprofundado do TABNET. Com o objetivo de entender como os dados são acessados e exportados, foram analisadas suas funcionalidades, estrutura de menus, dados e opções de filtro disponíveis, bem como o processo a ser percorrido para realizar o *download* das tabelas (Figura 1). O processo de busca e seleção dos dados do TABNET de interesse para a investigação foi organizado a partir de reuniões com um grupo de pesquisa multidisciplinar, formado por pesquisadores da saúde (medicina, farmácia), computação, engenharia de produção e administração. Dentre eles, a médica mastologista que orientou sobre a linha de cuidado do câncer de mama. A Figura 1 ilustra os passos necessários para que seja feito o *download* de uma única tabela com os filtros aplicados, enquanto a Figura 2 apresenta a tabela gerada nesse processo, ilustrando a quantidade de exames de mamografia por categoria BI-RADS de acordo com o tempo necessário para a obtenção dos resultados do exames, com filtros referentes ao local de atendimento, ano e sexo do paciente.

Figura 1 - Processo a ser percorrido para realizar o download de uma única tabela.



Fonte: (Autores, 2025)

Figura 2 - Tabela única gerada pelo processo descrito.

SISCAN - MAMOGRAFIA - POR LOCAL DE ATENDIMENTO - Rio Grande do Sul

Exames por BI-RADS segundo Tempo Exame

Munic.de Prest.Serviço: 431390 PANAMBI

Ano competencia: 2024

Sexo: Feminino

Tempo Exame	Categoria 0	Categoria 1	Categoria 2	Categoria 3	Categoria 4	Total
Total	346	238	922	11	17	1.534
Até 30 dias	14	5	28	1	1	49
31 - 60 dias	123	95	315	2	6	541
mais de 60	209	138	579	8	10	944

[COPIA PARA EXCEL](#) [SALVA COMO CSV](#) [COPIA PARA TABWIN](#)

Fonte: Sistema de Informações de Câncer (SISCAN)
Data de atualização dos dados: 20/02/2025

Fonte: (Autores, 2025)

2. Estudo da biblioteca Selenium (v1.0.14): para a automação da navegação no TABNET, foi estudada a biblioteca Selenium, que permite a automação de interações com navegadores **web**. A escolha dessa biblioteca ocorreu através de uma análise comparativa com outras bibliotecas de funcionalidade semelhante, que se mostraram menos adequadas ao objetivo principal do trabalho: automatizar a navegação na **web**. Foram utilizados os seguintes recursos:

- Configuração do WebDriver, para garantir compatibilidade entre o ambiente de execução (Google Colab) com o Google Chrome.
- Identificação e seleção de elementos HTML (HyperText Markup Language), para que a automação consiga se localizar dentro do site do TABNET.
- Simulação de cliques e preenchimento de caixas de texto no ambiente TABNET.
- Interação com diferentes abas do navegador, pois cada tabela é gerada em uma nova aba.
- Manipulação de **downloads** automáticos de arquivos CSV (Comma-separated values), que contém os dados referentes à linha de cuidado do câncer de mama.

3. Estudo da biblioteca Pandas (v2.1.4): a biblioteca Pandas foi utilizada para manipulação e tratamento dos dados extraídos do TABNET. A escolha por esta biblioteca justifica-se por ser uma biblioteca amplamente utilizada no contexto profissional de tratamento de dados. Assim, foram exploradas as seguintes funcionalidades:

- Leitura de arquivos CSV em formato Excel, necessário porque os arquivos são automaticamente exportados para o formato Excel após o **download**.
- Tratamento de valores ausentes e padronização dos dados, como correção de valores e colunas em branco.
- Agrupamento e reestruturação dos dados, conforme necessidades específicas, como reagrupamento das linhas da tabela por município de interesse.
- Geração de tabelas sumarizadas para análise epidemiológica, atendendo o que foi definido com o grupo multidisciplinar, na etapa 1 da metodologia.

4. Identificação dos dados de interesse: por meio de reuniões semanais com grupo de pesquisa multidisciplinar e a especialista médica mastologista, foram identificados os dados disponíveis e relevantes do TABNET para a pesquisa na linha de cuidado do câncer de mama. Essa etapa envolveu a seleção e combinação de variáveis, como:

- Faixa etária das pacientes.
- Localização geográfica.
- Classificação BI-RADS.
- Tempo médio necessário para a obtenção do diagnóstico dos exames.

5. Desenvolvimento da automação para navegação e **download** dos arquivos: com base na compreensão do TABNET e no estudo do Selenium, foi desenvolvido um **script** (sequência de instruções que o computador segue para realizar uma ação) que automatiza o acesso às tabelas por meio da seleção dos filtros e variáveis definidas, a fim de realizar a exportação dos arquivos CSV. Esse processo incluiu:

- Acesso ao **site** do TABNET.

- Criação de funções para simular interação com o site, como cliques e preenchimento de caixas de texto.
- Implementação de uma sequência lógica para acessar as opções desejadas, como foco epidemiológico, região de estudo de caso e períodos disponíveis.
- Gerenciamento de **downloads**, organizando os arquivos em pastas e ordens específicas, como uma pasta destinada aos **downloads** em ordem de chegada, e outra às tabelas resultantes.

6. Desenvolvimento da automação para tratamento e organização dos dados: após a extração, os dados foram processados para garantir sua integridade e uniformidade. O tratamento incluiu:

- Padronização de nomenclaturas e categorias, a exemplo da categoria de classificação BI-RADS, que por sua vez, não é apresentada de uma forma totalmente padronizada no site do TABNET, podendo variar de uma tabela para outra.
- Correção de inconsistências, como valores inexistentes e erros de formatação.
- Concatenação de uma ou mais tabelas, para possibilitar a comparação dos dados em diferentes condições.
- Estruturação padronizada das tabelas, conforme os critérios adotados na pesquisa, como organização das tabelas por ano e município de interesse.

7. Testes e validações: foram realizados testes para verificar a precisão e eficiência da automação. Os principais procedimentos de validação incluíram:

- Conferência manual de amostras de dados coletados, validando a automação através da comparação de resultados com os dados coletados manualmente.
- Comparação entre os dados extraídos automaticamente e aqueles obtidos através de consultas manuais ao TABNET.
- Medidas de desempenho, como tempo de processamento e taxa de sucesso na coleta realizada de forma automática, que se mostrou significativamente mais rápida e eficiente que o processo manual.

4 RESULTADOS E DISCUSSÃO

A implementação da automação trouxe melhorias significativas no processo de extração e tratamento de dados epidemiológicos do TABNET. Antes da solução proposta, a coleta era realizada manualmente, o que exigia tempo considerável para acessar o **site**, configurar os filtros, baixar os arquivos e realizar o tratamento dos dados. Tais tarefas eram executadas diversas vezes, visto que eram várias tabelas e o processo precisava ser repetido manualmente para cada uma delas.

A automação reduziu em aproximadamente 98% o tempo necessário para a realização dessas tarefas, diminuindo a possibilidade de ocorrerem falhas humanas e garantindo maior confiabilidade nas informações extraídas. Enquanto o processo manual demandava, em média, 6 horas para ser concluído por uma única pessoa, a solução desenvolvida executa a mesma atividade em aproximadamente 8 minutos, proporcionando um ganho expressivo de eficiência operacional.

A automação também garantiu a extração consistente e precisa dos dados, eliminando problemas recorrentes de formatação e inconsistências nas tabelas. A Figura 3 apresenta as tabelas geradas pela

automação e disponíveis no *drive*, logo após a execução do código. A Figura 4 mostra a tabela de ‘Exames por BI-RADS segundo tempo de exame - Local de atendimento’ do mês de setembro de 2024.

Figura 3 - Tabelas processadas disponíveis após a execução do script.

... > Dados Câncer de mama > tabelas_concatenadas_... ▾

Type ▾

People ▾

Modified ▾

Source ▾

Name	Owner
 Exames_por_Ano_segundo_Faixa_Etária_Local_de_Atendimento.xlsx 	 me
 Exames_por_Ano_segundo_Faixa_Etária_Local_de_Residência.xlsx 	 me
 Exames_por_Birads_segundo_Faixa_Etária_Local_de_Atendimento.xlsx 	 me
 Exames_por_Birads_segundo_Faixa_Etária_Local_de_Residência.xlsx 	 me
 Exames_por_Birads_segundo_Tempo_Exame_Local_de_Atendimento.xlsx 	 me
 Exames_por_Birads_segundo_Tempo_Exame_Local_de_Residência.xlsx 	 me

Fonte: (Autores, 2025)

Figura 4 - Exemplo de uma das tabelas processadas.

Município	Ano	Tempo Exame	Categoria 0	Categoria 1	Categoria 2	Categoria 3	Categoria 4	Categoria 5	Categoria 6	Total
Giruá	2019	Até 30 dias	1	1	5	0	1	2		10
		31 - 60 dias	2	2	11	1	1	0		17
		mais de 60	2	8	41	1	2	0		54
		Total	5	11	57	2	4	2		81
	2020	Até 30 dias	52	528	362	1	4			947
		31 - 60 dias	5	15	16	0	0			36
		mais de 60	0	5	4	0	0			9
		Total	57	548	382	1	4			992
	2021	Até 30 dias	85	532	539	6	1			1163
		31 - 60 dias	1	14	9	0	0			24
		mais de 60	3	9	9	2	0			23
		Total	89	555	557	8	1			1210
	2022	Até 30 dias	46	426	401	12	3	2	1	891
		31 - 60 dias	2	76	61	1	1	1	0	142
		mais de 60	1	1	4	0	0	0	0	6
		Total	49	503	466	13	4	3	1	1039
	2023	Até 30 dias	19	149	118	3	3	1	0	293
		31 - 60 dias	12	89	124	0	0	0	0	225
		mais de 60	13	120	109	4	0	1	1	248
		Total	44	358	351	7	3	2	1	766
	2024	Até 30 dias	42	210	222	7	3	4	2	490
		31 - 60 dias	11	86	96	2	4	0	1	200
		mais de 60	30	144	208	2	1	0	0	385
		Total	83	440	526	11	8	4	3	1075
Panambi	2019	Até 30 dias	36	149	176	5	3	3	2	374
		31 - 60 dias	4	15	10	1	0	0	0	30
		mais de 60	2	3	9	1	1	0	0	16
		Total	42	167	195	7	4	3	2	420
	2021	Até 30 dias	5	33	76	0	2	1		117
		31 - 60 dias	24	70	133	0	4	2		233
		mais de 60	18	39	140	1	4	0		202
		Total	47	142	349	1	10	3		552
	2022	Até 30 dias	15	10	29	2	2	0		58
		31 - 60 dias	62	68	162	5	4	2		303
		mais de 60	186	241	616	8	7	7		1065
		Total	263	319	807	15	13	9		1426
	2023	Até 30 dias	8	17	43	1	3	0		72
		31 - 60 dias	192	141	422	8	12	0		775
		mais de 60	156	127	447	5	1	1		737
		Total	356	285	912	14	16	1		1584
	2024	Até 30 dias	11	5	26	1	1			44
		31 - 60 dias	68	65	212	1	5			351
		mais de 60	160	105	463	6	4			738
		Total	239	175	701	8	10			1133

Fonte: (Autores, 2025)

A aplicação desenvolvida é capaz de gerar automaticamente as seguintes tabelas:

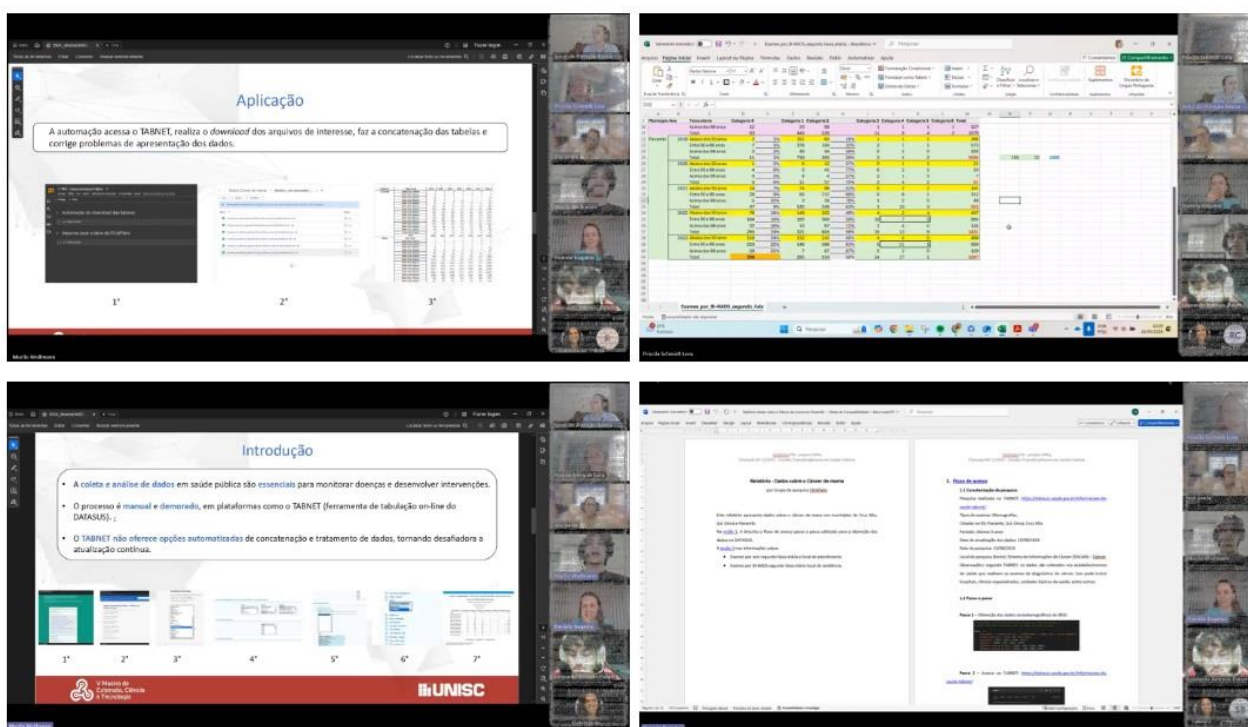
- Exames por ano segundo faixa etária - Local de atendimento;
- Exames por ano segundo faixa etária - Local de residência;
- Exames por BI-RADS segundo faixa etária - Local de atendimento;
- Exames por BI-RADS segundo faixa etária - Local de residência;
- Exames por BI-RADS segundo tempo de exame - Local de atendimento;
- Exames por BI-RADS segundo tempo de exame - Local de residência.

O código-fonte é de acesso público, podendo ser utilizado e adaptado conforme necessário. A documentação e o código-fonte foram disponibilizados em um projeto no GitHub, podendo ser acessado através do seguinte link: <<https://github.com/testeclickpalm/Automacao-para-a-coleta-e-tratamento-de-dados-no-TABNET>>.

Além disso, como contribuição do trabalho, o **script** desenvolvido, executado no ambiente Google Colab, foi disponibilizado para a Secretaria Municipal de Saúde de Panambi e para o Hospital de Panambi (SHP). Por meio desse ambiente, os usuários podem executar a automação para coletar, organizar e tratar os dados do TABNET diretamente no navegador, sem necessidade de instalação local., obtendo acesso contínuo a informações atualizadas sobre o câncer de mama na região.. O intuito foi apoiar a gestão do município e hospital para a tomada de decisões baseadas em dados, contribuindo para o planejamento de políticas públicas e aprimoramento da gestão da saúde.

Também, foi organizado e realizado um **workshop on-line** para capacitar os profissionais no uso da ferramenta, garantindo sua correta utilização e a possibilitando sua adaptação para outras necessidades dentro do setor de saúde pública (Figura 5). Um documento de orientação para facilitar o uso da ferramenta foi elaborado e entregue aos interessados.

Figura 5 - Workshop realizado com o Hospital de Panambi (SHP).



Fonte: (Autores, 2025)

Como principais benefícios observados com a implementação e disponibilização da automação, constatou-se:

- **Redução de tempo e esforço:** a coleta e o tratamento dos dados passaram a ser realizados de forma automática, otimizando a rotina dos profissionais envolvidos. As tabelas podem ser geradas acessando a aplicação, e com um clique o **script** faz o **download** das tabelas, segue para o tratamento e reestruturação e, por fim, exportar as tabelas resultantes para o drive escolhido.
- **Maior precisão e confiabilidade:** a eliminação de processos manuais reduziu significativamente a incidência de erros na manipulação dos dados, já que para reestruturar uma tabela manualmente era necessário dedicar muito tempo e prestar bastante atenção para não errar na hora de realocar os dados de forma imprecisa. Portanto, um processo manual exaustivo. Com isso, o processo, que

antes exigia cerca de seis horas para ser concluído por uma única pessoa, agora é realizado em apenas oito minutos com a automação.

- Acesso facilitado às informações: os dados passaram a ser organizados de forma padronizada, sempre no mesmo formato de linhas e colunas e filtros aplicados, possibilitando análises mais ágeis e eficientes.

5 CONCLUSÃO

Este trabalho apresentou o desenvolvimento e a implementação de uma solução computacional para a automação da coleta e tratamento de dados epidemiológicos extraídos do TABNET. A partir da utilização de ferramentas como Selenium e Pandas, foi possível criar uma aplicação eficiente e confiável, eliminando a necessidade de constantes intervenções manuais e reduzindo significativamente o tempo necessário para o processamento dos dados. Os resultados demonstraram que a automação possibilita maior precisão e padronização das informações coletadas, além de reduzir significativamente o tempo necessário para tal, permitindo uma melhor análise epidemiológica e facilitando a tomada de decisões na gestão da saúde pública.

A principal contribuição deste trabalho é a disponibilização de uma ferramenta que otimiza a obtenção de dados para pesquisas na área da saúde pública, especificamente na linha de cuidado do câncer de mama. A solução desenvolvida permite que instituições de saúde tenham acesso ágil e estruturado a informações atualizadas, podendo impactar positivamente na formulação de estratégias e políticas de saúde.

No entanto, algumas limitações devem ser consideradas. A automação desenvolvida é suscetível a possíveis alterações na estrutura do *site* do TABNET, o que pode exigir manutenções para garantir a continuidade do funcionamento do sistema. Além disso, os dados disponíveis no TABNET podem conter inconsistências e lacunas, como informações incompletas e colunas não padronizadas, o que pode impactar na análise e interpretação das informações.

Para trabalhos futuros, busca-se a ampliação da automação para outros conjuntos de dados epidemiológicos do TABNET, além da incorporação de funcionalidades adicionais que aprimorem a eficiência e a precisão no tratamento das informações.

AGRADECIMENTOS

Este projeto conta com o apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq e Departamento de Ciência e Tecnologia da Secretaria de Ciência, Tecnologia, Inovação e Complexo da Saúde do Ministério da Saúde, edital Chamada Nº 21/2023 - Estudos Transdisciplinares em Saúde Coletiva. Agradecemos ao Hospital de Panambi (SHP) e a Secretaria Municipal de Saúde de Panambi.

REFERÊNCIAS

BRAGA, T. M.; SOARES, L. M.; FERREIRA, A. C. G.; SIQUEIRA, A. C. F. L. DE; HELD, J. M. V. DE M. W.; GAMA, M. E. C. N. DA; FRIZON, A. B.; RIBEIRO, R. B.; MEDEIROS, G. L. P. DE; SILVESTRI, L. M. O. O uso de inteligência artificial na interpretação de exames médicos. *Brazilian Journal of Health Review*, v. 7, n. 3, p. e70932, 2024.

- CAPPELLO, A.; MURGIA, Y.; GIACOBBE, D. R.; MORA, S.; GAZZARATA, R.; ROSSO, N.; GIACOMINI, M.; BASSETTI, M. Automated extraction of standardized antibiotic resistance and prescription data from laboratory information systems and electronic health records: a narrative review. *Front Antibiot.*, v. 3, p. 1380380, 2024.
- CHEN, Y.; HAO, L.; ZOU, V. Z. et al. Automated medical chart review for breast cancer outcomes research: a novel natural language processing extraction system. *BMC Med Res Methodol*, v. 22, p. 136, 2022.
- DEPARTAMENTO DE INFORMÁTICA DO SUS (DATASUS). Informações de Saúde. Disponível em: <https://datasus.saude.gov.br/>. Acesso em: 3 fev. 2025.
- DOCUMENTAÇÃO DA BIBLIOTECA PANDAS. Disponível em: <https://pandas.pydata.org/>. Acesso em: 5 fev. 2025.
- DRESCH, A.; LACERDA, D. P.; ANTUNES JÚNIOR, J. A. V. A. *Design science research: método de pesquisa para avanço da ciência e tecnologia*. Porto Alegre: Bookman, 2015.
- INSTITUTO NACIONAL DE CÂNCER (INCA). *Controle do câncer de mama no Brasil: dados e números 2024*. Rio de Janeiro: INCA, 2024.
- LACERDA, D. P.; DRESCH, A.; PROENÇA, A.; ANTUNES JÚNIOR, J. A. V. A. Design science research: método de pesquisa para a engenharia de produção. *Gest. Prod.*, v. 20, n. 4, p. 741–761, 2013.
- LEITE, K. R. V.; PASQUALI, M. A. B.; SILVA, T. A. S. N. DA; RODRIGUES, L. V.-R. V.; MELO, G. D. X.; PACHECO, A. P. Q. D. F. et al. Transformação da saúde: o impacto da inteligência artificial na medicina moderna. *Contribuciones a las Ciencias Sociales*, v. 17, n. 5, p. e6726, 2024.
- MCGUINNESS, J. E.; ZHANG, T. M.; COOPER, K.; KELKAR, A.; DIMOND, J.; LORENZI, V. et al. Extraction of Electronic Health Record Data using Fast Healthcare Interoperability Resources for Automated Breast Cancer Risk Assessment. *AMIA Annu Symp Proc.*, v. 2021, p. 843-852, 2022.
- MILLER, T. P.; GETZ, K. D.; KRAUSE, E.; JO, Y. G.; CHARAPALA, S.; GRAMATAGES, M. M. et al. Automated Electronic Health Record Data Extraction and Curation Using ExtractEHR. *JCO Clin Cancer Inform.*, v. 8, p. e2400100, 2024.
- MUTHUKADAN, B. *Selenium with Python*. Disponível em: <https://selenium-python.readthedocs.io/>. Acesso em: 5 fev. 2025.
- NILSEN, P.; SUNDEMO, D.; HEINTZ, F.; NEHER, M.; NYGREN, J.; SVEDBERG, P.; PETERSSON, L. Towards evidence-based practice 2.0: leveraging artificial intelligence in healthcare. *Front Health Serv.*, v. 4, p. 1368030, 2024.
- OBSERVATÓRIO DO CÂNCER RS. Disponível em: <https://observatoriodocancer.saude.rs.gov.br/>. Acesso em: 21 ago. 2024.
- RAHMAN, M. A.; VICTOROS, E.; ERNEST, J.; DAVIS, R.; SHANJANA, Y.; ISLAM, M. R. Impact of Artificial Intelligence (AI) Technology in Healthcare Sector: A Critical Evaluation of Both Sides of the Coin. *Clin Pathol.*, v. 17, p. 2632010X241226887, 2024.
- SECRETARIA ESTADUAL DE SAÚDE DO RIO GRANDE DO SUL. *Boletim Epidemiológico*. v. 1, n. 1, 2024. Disponível em: <https://admin.estado.rs.gov.br/upload/arquivos/202410/boletim-epidemiologico-cancer-de-mama-no-estado-do-rio-grande-do-sul-versao-final-2.pdf>. Acesso em: 13 out. 2024.
- VAISHNAVI, V.; KUECHLER, B. *Design Science Research in Information Systems*. 2004/21. Disponível em: <http://desrist.org/desrist/content/design-science-research-in-information-systems.pdf>. Acesso em: 21 dez. 2024.

- VASCONCELLOS, F. M.; LUTZ, K. C. C.; MORAIS, E. C.; MELO, B. A.; VARGAS, I. H. (orgs.). *Boletim Epidemiológico da Situação do Câncer de Mama no Estado do Rio Grande do Sul*. Porto Alegre: Secretaria de Estado da Saúde/RS, 2024. Disponível em: <https://admin.estado.rs.gov.br/upload/arquivos/202410/boletim-epidemiologico-cancer-de-mama-no-estado-do-rio-grande-do-sul-versao-final-2.pdf>. Acesso em: 13 jan. 2025.
- WALKER, V. R.; SCHMITT, C. P.; WOLFE, M. S.; NOWAK, A. J.; KULESZA, K.; WILLIAMS, A. R. et al. Evaluation of a semi-automated data extraction tool for public health literature-based reviews: Dextr. *Environ Int.*, v. 159, p. 107025, 2022.
- ZAMAN, G.; MAHDIN, H.; HUSSAIN, K.; RAHMAN, A. Information extraction from semi and unstructured data sources: A systematic literature review. *ICIC Express Letters*, v. 14, n. 6, p. 593-603, 2020. Disponível em: <http://www.icicel.org/ell/contents/2020/6/el-14-06-09.pdf>. Acesso em: 27 fev. 2025.