

Integração de bases de dados para estudos de esquistossomose para o estado de Minas Gerais

Integration of databases for schistosomiasis studies for the state of Minas Gerais

Integración de banco de datos para estudios de esquistosomiasis en el estado de Minas Gerais

Bruno Petrocchi de Sena Azevedo¹, Daniel Rocha Franca¹, Henri Gabriel Viana Ramos¹, Ligia Ferreira de Carvalho Gonçalves¹, Rafael Romualdo Pinto Rodrigues¹, Samuel Augusto Barbosa Santos¹, Marco Paulo Soares Gomes², Luis Enrique Zárate²

1 Bac., Ciência de Dados, PUC Minas, Belo Horizonte (MG), Brasil

2 Dr., Ciência de Dados, PUC Minas, Belo Horizonte (MG), Brasil

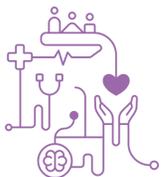
Autor correspondente: Prof. Dr. Luis Enrique Zárate

E-mail: zarate@pucminas.br

Resumo

Objetivo: este artigo propõe uma integração de fontes de dados disponíveis para criação de uma base de dados, em nível municipal, que permita a análise da prevalência da esquistossomose no estado de Minas Gerais, Brasil. Método: por meio de um processo para entendimento do domínio, foram identificadas as principais dimensões e aspectos (fatores) que podem influenciar a prevalência da doença em um município. O modelo conceitual gerado pelo processo de entendimento auxiliou o processo de integração de dados. Resultados: a base de dados integrada permitiu realizar análises descritivas para o estado, e são uma fonte para identificação de padrões dos municípios propensos à doença. Conclusões: A base de dados resultante possibilita a construção de modelos preditivos para a formulação de políticas públicas e de programas de intervenção municipal.

Descritores: Esquistossomose; Saúde pública; Integração de Dados.



Abstract

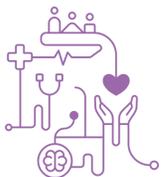
Objective: this article proposes an integration of available data sources to create a database, at the municipal level, that allows the analysis of the prevalence of schistosomiasis in the state of Minas Gerais, Brazil. **Method:** through a process to understand the domain, the main dimensions and aspects (factors) that can influence the prevalence of the disease in a municipality were identified. The conceptual model generated by the understanding process helped the data integration process. **Results:** the integrated database made it possible to carry out descriptive analyses for the state, and it is a source for identifying patterns in municipalities prone to the disease. **Conclusions:** The resulting database will allow the construction of predictive models for the formulation of public policies and municipal intervention programs.

Keywords: Schistosomiasis; Public Healthcare; Data Integration.

Resumen

Objetivo: este artículo propone una integración de fuentes de datos disponibles para crear una base de datos, a nivel municipal, que permita el análisis de la prevalencia de esquistosomiasis en el estado de Minas Gerais, Brasil. **Método:** a través de un proceso de comprensión del dominio, se identificaron las principales dimensiones y aspectos (factores) que pueden influir en la prevalencia de la enfermedad en un municipio. El modelo conceptual generado por el proceso de comprensión ayudó al proceso de integración de datos. **Resultados:** la base de datos integrada permitió realizar análisis descriptivos para el estado y son una fuente para identificar patrones en municipios propensos a la enfermedad. **Conclusiones:** La base de datos resultante permitirá la construcción de modelos predictivos para la formulación de políticas públicas y programas de intervención municipal.

Descriptores: Esquistosomiasis; Salud Pública; Integración de Datos.



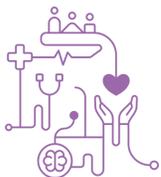
Introdução

A Esquistossomose, uma doença parasitária causada pelo *Schistosoma mansoni*, representa um desafio significativo em termos de saúde pública em diversas regiões do mundo¹. No Brasil, especialmente em áreas endêmicas como o Nordeste, a prevalência da doença é notável². A transmissão ocorre principalmente através do contato com água doce contaminada por larvas do parasita, liberadas por caramujos³. Os fatores que contribuem para a ocorrência da doença são locais sem saneamento básico ou com saneamento básico precário, as condições socioeconômicas desfavoráveis, o nível educacional da região e a exposição da população ao risco⁴. A prevalência da parasitose também está associada ao crescimento dos centros urbanos, que leva ao estabelecimento de assentamentos humanos em áreas periféricas, carentes de infraestrutura sanitária mínima⁵.

Do ponto de vista econômico, os custos vinculados à prevenção e tratamento da esquistossomose sobrecarregam os sistemas de saúde, já fragilizados⁶. A perda de produtividade no trabalho devido à doença também contribui, significativamente, para os impactos negativos na economia local, criando um ciclo de pobreza e debilidade financeira difícil de quebrar.

Além dos impactos diretos na saúde, a doença pode levar ao estigma e à exclusão social. O fenômeno do estigma social em relação às doenças tropicais negligenciadas (DTNs) é bem documentado^{5,7}. Resultados indicam que pessoas afetadas pela esquistossomose são frequentemente alvo de estigmatização social, contribuindo para sua exclusão e isolamento⁷. Essa estigmatização pode ser atribuída à falta de compreensão sobre a transmissão da doença, resultando em dificuldades adicionais para sua prevenção.

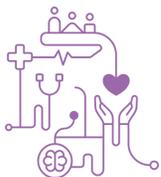
Para abordar esses desafios complexos, é necessária uma abordagem abrangente que inclui a identificação de regiões de risco, medidas de prevenção, diagnóstico precoce, tratamento eficaz e educação comunitária. Para essas ações serem preventivas e efetivas, é necessário, primeiramente, identificar/prever as



regiões de risco, caracterizando o perfil dos municípios propensos a sofrer com a doença.

Diante do desafio representado, este trabalho apresenta uma abordagem para integrar bases de dados disponíveis, em nível de município, para caracterizar e compreender os fatores determinantes da doença. A base de dados integrada pode auxiliar na análise descritiva do problema por município, como também compor um conjunto de dados que podem ser utilizados em projetos de Ciência de Dados. Com aplicação de técnicas de aprendizado de máquina e mineração de dados, seria possível identificar os padrões dos municípios que podem sofrer com a doença e realizar ações preditivas de prevenção.

Para construção da base de dados integrada e seleção dos fatores de risco, este estudo se fundamenta na literatura^{4,9,10}. Braga et al.⁹ oferecem uma contribuição fundamental ao apresentar um modelo preliminar de registro de informações de saúde voltado para a atenção primária no Brasil. Por meio de uma abordagem de pesquisa-ação exploratória, especialistas de diversas áreas da saúde colaboraram na identificação das informações essenciais para a avaliação da saúde geral do indivíduo. O resultado foi um modelo informativo abrangente, que não apenas segue as diretrizes semiológicas, mas também transcende as abordagens convencionais, destacando a importância da transdisciplinaridade na prestação de cuidados de saúde primários. Harrison et al.¹⁰ apresentam uma análise detalhada dos fatores que limitam a qualidade dos dados no programa ampliado de imunização em países de baixa e média renda. O trabalho destaca os desafios enfrentados na obtenção de dados precisos sobre a cobertura de imunização nessas regiões. Problemas como inflação e deficiências nos sistemas de informação em saúde foram identificados como obstáculos importantes. O estudo ressalta a importância crítica da qualidade dos dados para o sucesso de programas de saúde pública em contextos de recursos limitados, destacando a necessidade de intervenções eficazes para melhorar a qualidade dos dados e, conseqüentemente, a eficácia dos programas de imunização.



A fonte de dados resultante deste processo de integração visa ser uma fonte para fornecer *insights* valiosos para formuladores de políticas públicas, profissionais de saúde e pesquisadores interessados em abordar a Esquistossomose de forma mais eficaz. Ao entender os fatores determinantes e as interrelações entre eles, pode-se desenvolver estratégias mais direcionadas e eficientes para prevenção, controle e tratamento da doença, contribuindo, assim, para melhorar a saúde e o bem-estar das populações afetadas.

Como estudo de caso, e para mostrar o processo de integração das fontes de dados, será considerado o estado de Minas Gerais, que possui 853 municípios. Para guiar o processo de integração das fontes de dados, a metodologia adota uma etapa de entendimento do domínio de problema. Para isso, foi utilizado o método CAPTO⁸ de captura do conhecimento. O método é baseado em modelos de gestão do conhecimento e, junto com o conhecimento explícito disponível/adquirido, propõe uma estratégia para construção de modelos conceituais para representação de domínios de problemas. Os modelos são constituídos a partir da identificação de dimensões (perspectivas), aspectos e atributos que podem ser relevantes a um domínio. O entendimento formal acerca do domínio de esquistossomose pode auxiliar na construção de uma base de dados orientados a domínio.

A criação de uma base de dados integrada, utilizando o método CAPTO, permite uma análise descritiva para os municípios considerados e serve como entrada para construção de modelos de aprendizado computacional que permitam identificar/prever áreas com maior prevalência da doença, os grupos populacionais mais afetados e os fatores de risco associados à sua transmissão.

Materiais e Métodos

Para a realização deste estudo, foram utilizadas diversas fontes de dados para coleta de informações relevantes acerca dos municípios do estado de Minas Gerais. A Tabela 1 descreve as fontes de dados consideradas.

Tabela 1 - Fontes de dados por município para o estado de Minas Gerais



Indicadores	Descrição	Fonte	Links
Código do Município	Município e código de sete dígitos em que os dois dígitos referem ao estado da federação.	IBGE	https://www.ibge.gov.br/cidades-e-estados/mg/
Salário-mínimo médio mensal dos trabalhadores formais	Média mensal de salário do trabalhador regido pela Consolidação das Leis de Trabalho (CLT).	IBGE	https://cidades.ibge.gov.br/brasil/sintese/mg?indicadores=96385,96386,29765,29763,60036,60037
Índice de Desenvolvimento Humano Municipal (IDHM)	Medida que avalia o desenvolvimento humano em diferentes municípios.	Atlas do Desenv. Humano GOV.BR	https://dados.gov.br/dados/conjuntos-dados/atlasbrasil
Sistema de Esgoto sanitário adequado	Porcentagem do esgoto doméstico gerado no município que é devidamente coletado, tratado e/ou disposto de forma segura pelo sistema de esgotamento sanitário local.	Instituto Água e Saneamento IAS	https://www.aguaesaneamento.org.br/municipios-e-saneamento/explore-compare
Bioma	Indica qual é o tipo de vegetação predominante do município.	IBGE	https://cidades.ibge.gov.br/brasil/mg/panorama
Área urbanizada (km²)	Média de área (km ²) que contém edifícios, ruas pavimentadas, redes de água e esgoto, sistemas de transporte público	IBGE	https://www.ibge.gov.br/geociencias/organizacao-do-territorio/tipologias-do-territorio/15789-areas-urbanizadas.html?edicao=35569&t=downloads
Área da unidade territorial (km²)	Área (km ²) que representa o tamanho de cada cidade em escala territorial.	IBGE	https://www.ibge.gov.br/geociencias/organizacao-do-territorio/tipologias-do-territorio/15789-areas-urbanizadas.html?edicao=35569&t=downloads
Mesorregião	Representa a região composta por diversos municípios de uma área geográfica.	IBGE	https://cidades.ibge.gov.br/brasil/mg/panorama
Casos em Revisão	Casos notificados em revisão para esquistossomose	SUS	http://tabnet.datasus.gov.br/cgi/tabcgi.exe?sinannet/cnv/esquistomg.def

A seguir serão descritas cada uma das etapas envolvidas no processo de integração aplicadas ao domínio considerado.

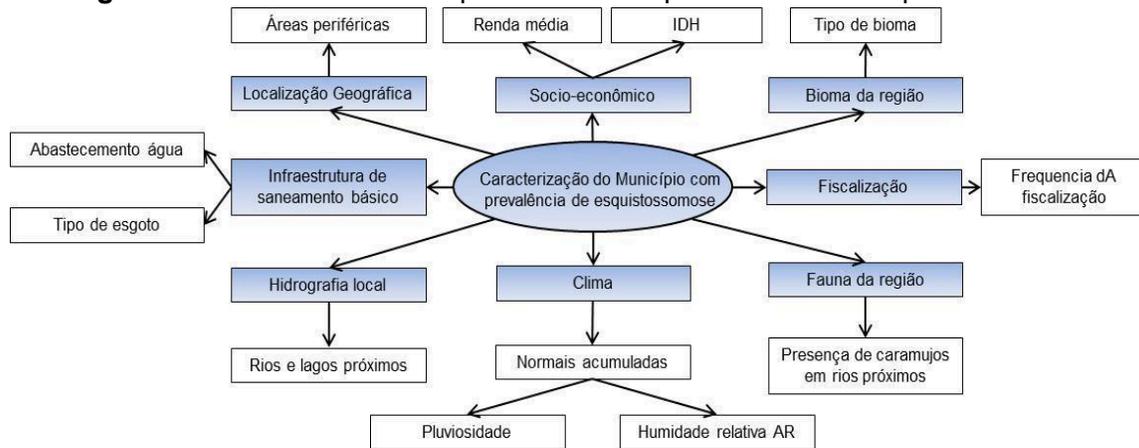
Etapas 1 - Entendimento do domínio de problema: Para caracterizar a doença esquistossomose em nível de município, foi aplicado o método CAPTO



para entendimento do domínio. O método identifica as principais dimensões e aspectos vinculados ao domínio e serve como um guia para a seleção conceitual de indicadores/fatores, a partir das fontes de dados consideradas.

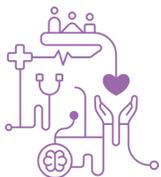
Seguindo as etapas do método, foram identificados, por meio da captura do conhecimento tácito/explicito, dimensões e aspectos relevantes ao domínio^{2,6,9,10}. Como resultado, foi obtido um modelo conceitual para o domínio considerado (ver Figura 1). O modelo representa as diferentes dimensões e aspectos relacionados com a doença, identificados durante a etapa de entendimento do domínio e fundamentados pela literatura.

Figura 1 - Modelo Conceitual para fatores de prevalência da esquistossomose



Dentre os indicadores apontados citamos: socio-econômicos, condições ambientais (tipo de bioma e fauna), clima, características hidrográficas, infraestrutura e aspectos relacionados com a saúde pública.

Etapa 2 - Seleção e coleta de dados: A partir do modelo conceitual, foram procuradas as fontes de dados que atendessem ao domínio considerado. Essas fontes de dados são descritas na Tabela 1. Porém, para a dimensão clima, dados das normais de pluviosidade acumulada e umidade relativa do ar por mês e para cada município não foram encontrados dados disponíveis. O período médio de integração corresponde ao ano de 2021. Devido à falta de dados atualizados do IDH, foram incluídos, na base de dados integrada, os índices de desenvolvimento humano municipal (IDHM) dos anos 1991, 2000, e 2010. Durante o processo de



seleção e coleta de indicadores dos municípios, foram priorizadas as informações que não apresentaram dados faltantes, garantindo, assim, uma representação abrangente do escopo estudado. Isso foi fundamental para garantir a validade e a confiabilidade do processo de integração de dados.

Foram extraídos os seguintes indicadores: Salário médio mensal dos trabalhadores formais, Índice IDHM, Área urbanizada, Área da unidade territorial, Sistema de saneamento, Bioma, Mesorregião e casos registrados (em revisão) para 2021, 2022 e 2023.

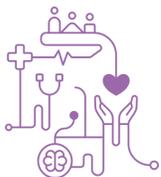
Etapa 3 - Integração dos dados: Este procedimento consistiu na integração dos dados coletados a partir das diferentes fontes, com formatos de arquivos distintos. Utilizamos técnicas computacionais adequadas para extrair os dados das fontes originais e para integração dos conjuntos de dados, de forma a garantir sua compatibilidade, integridade e coesão. Foi, também, calculada a porcentagem da área urbanizada por município, com dados da área urbanizada total e da unidade disponíveis no IBGE (ver Tabela 1).

Etapa 4 - Preparação do conjunto de dados: Os dados foram normalizados/padronizados para garantir que todas as variáveis estivessem na mesma escala, evitando, assim, o viés causado por diferenças nas unidades de medida. Essas etapas são essenciais para assegurar a qualidade e a eficácia dos modelos desenvolvidos a partir do conjunto de dados integrado. A Tabela 2 mostra um segmento do conjunto de dados resultante do processo de integração.

Etapa 5 - Análise Exploratória de Dados: Uma vez finalizado o processo de integração, foi realizada uma análise exploratória da base de dados integrada de forma a mostrar potencialidade para análise da esquistossomose no estado de Minas Gerais. A análise exploratória inclui a aplicação de técnicas de visualização, estatística descritiva e a construção de modelo de aprendizado de máquina, baseado na tarefa de classificação.

Tabela 2 – Segmento de dados do resultado da integração das fontes de dados

Município	Código	Salário- mínimo médio	IDHM 1991	IDHM 2000	IDHM 2010	Sistema de Esgoto	Bioma	Área urbanizada (km ²)	Área da unidade	Porcentagem de área urbanizada
-----------	--------	-----------------------------	--------------	--------------	--------------	-------------------------	-------	--	--------------------	--------------------------------------



CBIS'24 Congresso Brasileiro de Informática em Saúde

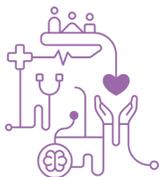
8/10 a 11/10 de 2024 - Belo Horizonte/MG - Brasil

		mensal formais x1000				sanitário adequado (%)			territorial (km²)		
Abadia dos Dourados	3100104	1,4	0,449	0,575	0,689	59,9	Cerrado	1,41	880,461	1,60	
Abaeté	3100203	1,7	0,417	0,587	0,698	85	Cerrado	6,96	1817,067	3,83	
Abre Campo	3100302	1,5	0,379	0,54	0,654	48,7	Mata Atlântica	2,47	470,551	5,25	
Acaiaca	3100401	1,6	0,336	0,518	0,63	63	Mata Atlântica	0,8	101,886	7,85	
Açucena	3100500	1,5	0,267	0,443	0,61	43,8	Mata Atlântica	2,09	815,422	2,56	

Código	Jan_21	Fe_21	Mar_21	Abr_21	Mai_21	Jun_21	Jul_21	Ago_21	Set_21	Out_21	Nov_21	Dez_21
3100104												
3100203												
3100302												
3100401	1				4		5		1	1		
3100500					1	1	2	2	4	2	2	3

Código	Jan_22	Fev_22	Mar_22	Abr_22	Mai_22	Jun_22	Jul_22	Ago_22	Set_22	Out_22	Nov_22	Dez_22
3100104												
3100203												
3100302												
3100401	1					4	1	2	1			
3100500	1		1		1	1		1	1	4	4	7

Código	Jan_23	Fev_23	Mar_23	Abr_23	Mai_23	Jun_23	Jul_23	Ago_23	Set_23	Out_23	Nov_23	Dez_23
3100104												
3100203												
3100302												
3100401				1		1				3		
3100500		2	1		4	2	1	1				



Resultados e Discussão

A base de dados resultante possui 46 atributos (Ver Tabela 2) e está disponível no link: (<https://github.com/licapLaboratory/BaseDadosIntegrada-Esquistossomose>). Os gráficos da Figura 2 mostram a incidência de casos de esquistossomose para o estado de Minas Gerais por município e por região. As Figuras 3 e 4 mostram a quantidade de municípios por mesorregião e bioma para o estado de Minas Gerais, destacando-se as mesorregiões Sul, Sudoeste e Zona da Mata, e a preponderância da Mata Atlântica para o bioma.

A Tabela 3 mostra uma análise estatística descritiva para as variáveis numéricas da base de dados, indicando uma média salarial para o estado de Minas Gerais variante de 1.500 a 2.100 para 65% dos municípios.

Figura 2 - Quantidade de notificações por municípios e regiões para Minas Gerais

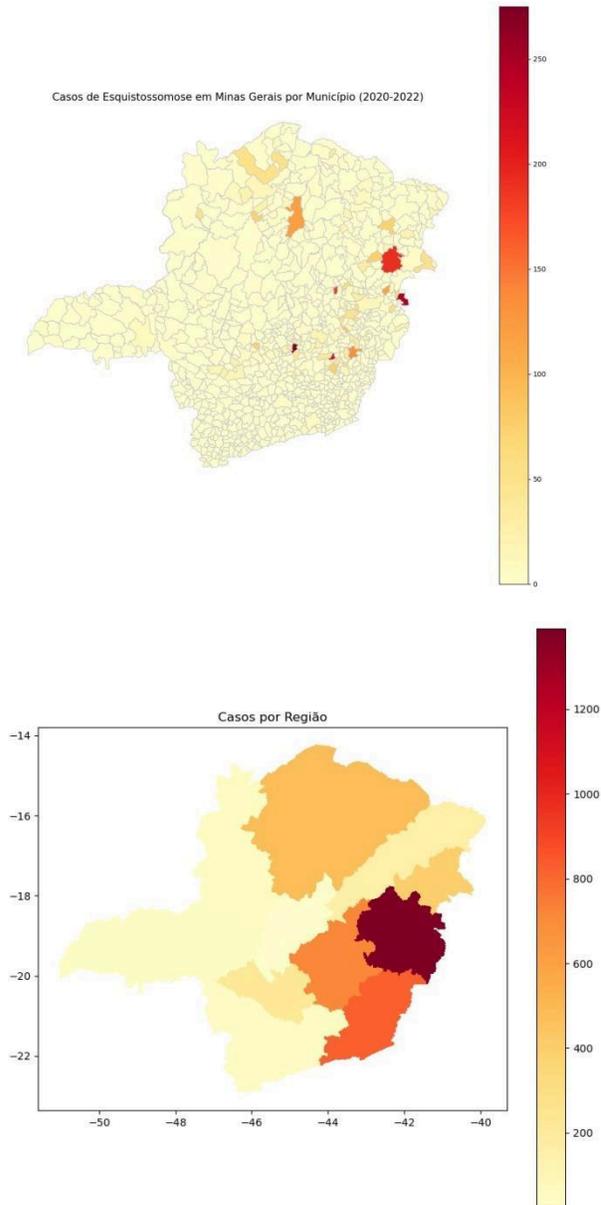


Figura 3 - Quantidade de municípios por tipo de mesorregião para Minas Gerais

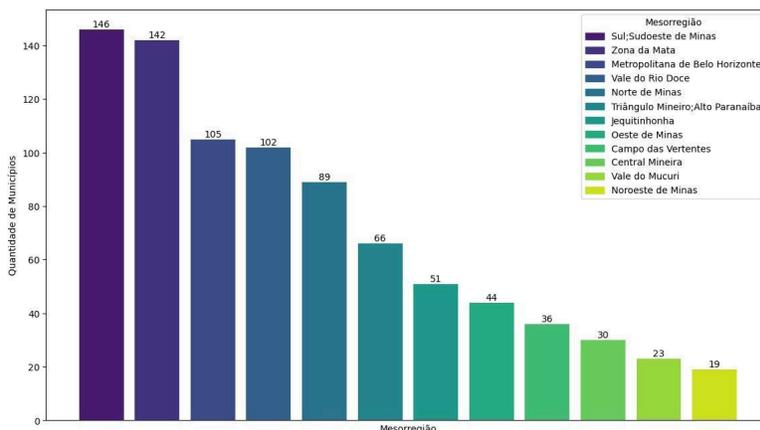


Figura 4 - Quantidade de municípios por tipo de bioma para Minas Gerais

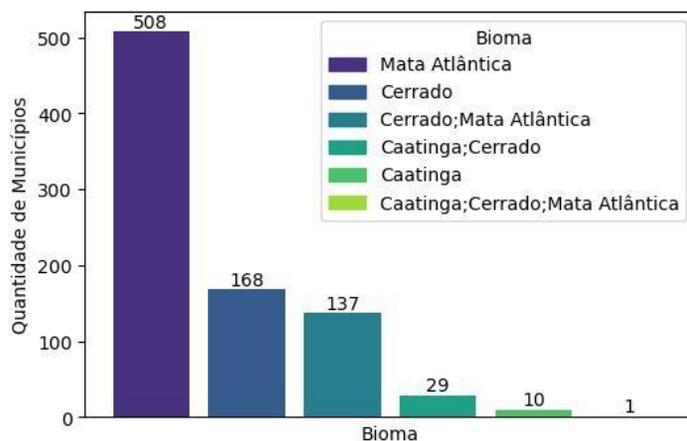
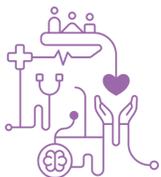


Tabela 3 - Descrição da base de dados integrada para esquistossomose para MG

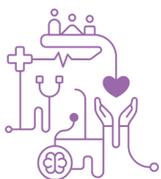
	Salário-mínimo médio mensal dos trabalhadores formais (x1000)	Índice de Desenvolvimento Humano Municipal (IDHM)-2010	Sistema de Esgoto sanitário adequado (%)	(%) da área urbanizada
Média	1.8	0.667	59.1	15.4
Desvio Padrão	0.3	0.049	25.6	47.5
Mínimo	1.2	0.529	0.3	0.2
Máximo	5.2	0.813	98.4	827
1º Quartil (25%)	1.6	0.633	42.1	2.8
2º Quartil (50%)	1.7	0.670	64	5.1
3º Quartil (75%)	1.9	0.702	79.6	10.5



A integração dos dados utilizando o método CAPTO⁸, permite também aplicar técnicas de aprendizado de máquina de forma mais efetiva para extrair padrões, *insights* ou hipóteses, e descobrir conhecimento útil para auxiliar na elaboração de políticas públicas. Note que o entendimento do problema baseado no método CAPTO permite aplicar um processo de descoberta de conhecimento mais consistente, evitando aplicar técnicas computacionais sobre bases de dados “simplesmente disponíveis”, sem um prévio entendimento do domínio de problema.

Para demonstrar a potencialidade do conjunto de dados integrado, é apresentada, como estudo de caso, a construção de um modelo de aprendizado de máquina de classificação, baseado em árvore de decisão. Para isso, municípios com casos de esquistossomose registrados foram rotulados como 1 e sem casos registrados foram rotulados como 2. Para desenvolvimento do modelo, foram considerados os seguintes atributos como entrada: Salário-mínimo, Sistema de esgoto sanitário adequado, IDHM-2010, Bioma, Mesorregião, e Porcentagem de área urbanizada. Não foi aplicado qualquer processo de preparação ou transformação dos dados. A Figura 5 mostra o esquema da implementação utilizando a ferramenta workflow Knime (<https://www.knime.com/>). Para ajustar o modelo, foi aplicada a técnica *Hold-out* (80% para treino e 20% para teste). A Tabela 4 mostra a métrica F-measure para os resultados de treinamento e teste.

A partir dos resultados é possível observar um desempenho regular do modelo tanto para o treinamento como para o teste, 68% para traçar o perfil de municípios com casos de esquistossomose e 61% para os municípios sem a doença. Comportamento similar é observado para o conjunto de teste (69% e 66%, respectivamente). É importante ressaltar que este trabalho não tem por objetivo a construção de modelos de aprendizado computacional para traçar o perfil dos municípios, com e sem casos da doença, e sim mostrar a potencialidade do conjunto de dados. O objetivo é disponibilizar uma fonte de dados para



aplicações em projetos de Ciência de Dados, onde procedimentos mais adequados de preparação de dados possam ser aplicados.

Figura 5 - Esquema da implementação utilizando a ferramenta workflow Knime

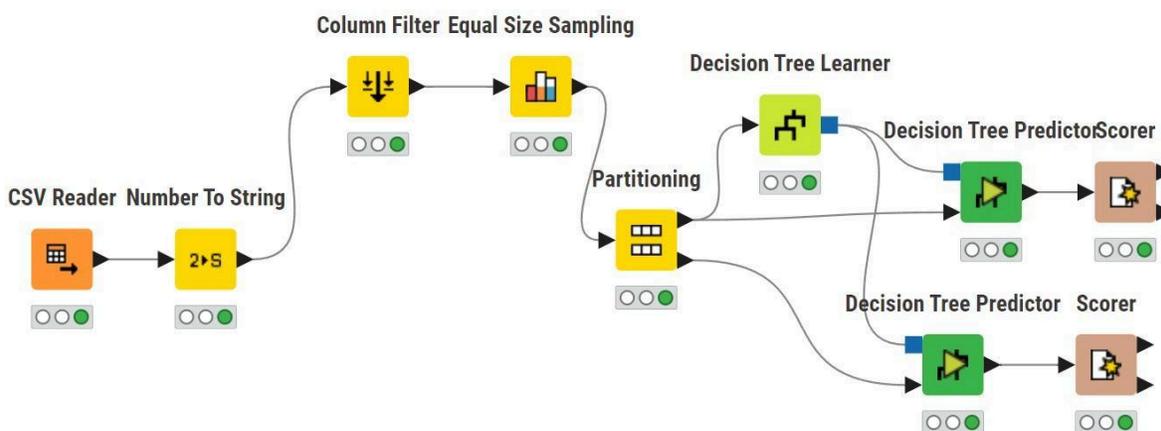
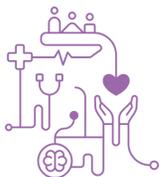


Tabela 4 - Descrição da base de dados integrada para esquistossomose para MG

	Classe	Treino	Teste
F-Measure	1: Município com Notificação	0.69	0.68
	2: Município sem Notificação	0.66	0.61

A Tabela 5 mostra as regras de decisão obtidas pelo modelo de classificação. Nota-se que a árvore de decisão e as regras dela extraídas são uma representação do conteúdo da base de dados, podendo ser consideradas hipóteses a serem comprovadas ou rejeitadas.

A partir das Regras 4, 5, 6 e 9, observa-se que as mesorregiões Vale do Rio Doce, Oeste de Minas, Vale do Mucuri e Jequitinhonha são determinantes para a doença. Pela Regra 10, se a mesorregião é Zona da Mata e o sistema de esgoto sanitário é menor que 58,5%, há presença da doença. Porém, pela Regra 11, sendo a mesorregião Zona da Mata, se o sistema de esgoto sanitário está compreendido entre 58,5% e 81,6%, não há presença da doença. Pela Regra 16,

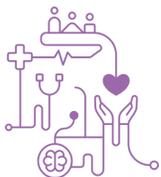


se o sistema de esgoto sanitário adequado é menor ou igual a 17,6 e a mesorregião é Norte de Minas, há notificação para a doença.

Tabela 5 - Regras de decisão obtidas pelo modelo de classificação

Regra 1:	SE a porcentagem da área urbanizada é menor ou igual a 3.41 e a mesorregião é Triângulo Mineiro/Alto Paranaíba, ENTÃO Classe= "2".
Regra 2:	SE a porcentagem da área urbanizada é maior que 3.41 e a mesorregião é Triângulo Mineiro/Alto Paranaíba, ENTÃO Classe= "1".
Regra 3:	SE a mesorregião é Central Mineira, ENTÃO Classe= "2".
Regra 4:	SE a mesorregião é Vale do Rio Doce, ENTÃO Classe= "1".
Regra 5:	SE a mesorregião é Oeste de Minas, ENTÃO Classe= "1".
Regra 6:	SE a mesorregião é Vale do Mucuri, ENTÃO Classe= "1".
Regra 7:	SE a mesorregião é Sul/Sudoeste de Minas, ENTÃO Classe= "2".
Regra 8:	SE a mesorregião é Campo das Vertentes, ENTÃO Classe= "2".
Regra 9:	SE a mesorregião é Jequitinhonha, ENTÃO Classe= "1".
Regra 10:	SE o sistema de esgoto sanitário adequado é menor ou igual a 58.5 e a mesorregião é Zona da Mata, ENTÃO Classe= "1".
Regra 11:	SE o sistema de esgoto sanitário adequado é maior que 58.5 e menor ou igual a 81.6 e a mesorregião é Zona da Mata, ENTÃO Classe= "2".
Regra 12:	SE o sistema de esgoto sanitário adequado é maior que 81.6 e maior que 58.5 e a mesorregião é Zona da Mata, ENTÃO Classe= "1".
Regra 13:	SE o sistema de esgoto sanitário adequado é menor ou igual a 42.8 e a mesorregião é Metropolitana de Belo Horizonte, ENTÃO Classe= "2".
Regra 14:	SE o sistema de esgoto sanitário adequado é maior que 42.8 e a mesorregião é Metropolitana de Belo Horizonte, ENTÃO Classe= "1".
Regra 15:	SE a mesorregião é Noroeste de Minas, então Classe= "2".
Regra 16:	SE o sistema de esgoto sanitário adequado é menor ou igual a 17.6 e a mesorregião é Norte de Minas, ENTÃO Classe= "1".
Regra 17:	SE o sistema de esgoto sanitário adequado é maior que 17.6 e a mesorregião é Norte de Minas, ENTÃO Classe= "2".

1: Município com Notificação 2: Município sem Notificação



Conclusão

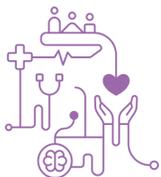
Descrever e analisar, profundamente, os fatores da prevalência da esquistossomose em uma região ou município é fundamental para desenvolver estratégias eficazes de prevenção e controle. Ao caracterizar os municípios mais afetados pela doença, seria possível identificar outras regiões vulneráveis, direcionando recursos e realizando intervenções preventivas mais eficazes.

A partir da base de dados integrada, proposta neste trabalho, é possível realizar análises estatísticas mais detalhadas e aplicar diversas técnicas de aprendizado de máquina, como agrupamentos e classificação, para a descoberta de padrões e, com isso, aumentar a capacidade preditiva para identificar municípios propensos a contrair casos da doença. Com base nesses resultados, seria possível sugerir estratégias específicas de prevenção, controle e tratamento da doença.

É importante ressaltar que os procedimentos adotados neste trabalho, especialmente a construção do modelo conceitual, via método CAPTO, como guia para integração das bases de dados para o estado de Minas Gerais, podem ser generalizados para outras regiões e estados e, desta forma, construir um atlas da esquistossomose no Brasil.

Em relação à replicação deste trabalho para outros estados, os procedimentos de integração devem buscar garantir a qualidade e a confiabilidade do processo de extração e transformação dos dados. Por exemplo, as bases de dados disponíveis podem estar fragmentadas, inconsistentes ou desatualizadas, o que pode afetar a análise que possa ser aplicada à base de dados integrada. Além disso, os dados podem variar entre diferentes fontes e regiões/municípios do Brasil, limitando sua utilidade para fins de análise e tomada de decisões. Essas limitações podem representar um desafio para a integração e a compreensão dos fatores da doença, para qualquer região, e para a implementação de medidas eficazes de prevenção e controle.

Outro aspecto não considerado nesta integração refere-se à inserção de dados climáticos, especificamente da pluviosidade acumulada e da umidade



relativa por mês para cada município. Estas informações devem corresponder às normais, que são tipicamente calculadas com históricos de dados de aproximadamente 30 anos. Considera-se, então, como trabalho futuro, a incorporação, mediante tratamento de dados disponíveis, de repositórios públicos como INMET (<http://pclima.inpe.br/analise/>).

Referências

1. LoVerde PT. Schistosomiasis. *Adv Exp Med Biol.* 2019;1154:45-70. doi: 10.1007/978-3-030-18616-6_3. PMID: 31297759.
2. Dutra , A. S. de S., Dutra , R. F. F., Costa , G. F. da, Cruz , B. I. L. da, Oliveira , K. de, Souza , N. K. F. de, Jugo , K. E. L., Souza, L. A. de, Abeiya, K.-A., Alexandre , A. sandriely A., Farias , G. de H., & Farias , E. C. M. de H. (2024). Análise da Esquistossomose na região nordeste de 2020 a 2023: *Brazilian Journal of Implantology and Health Sciences*, 6(2), 691–711. <https://doi.org/10.36557/2674-8169.2024v6n2p691-711>
3. FRANÇA, F. S.; SILVA, A. S.; MAGALHÃES, C. M. M.; BENEVIDES, K. S. Esquistossomose: uma endemia de importância no Brasil. *RBAC*, v.52, p.224-7, 2020.
4. Santos MCS, Oliveira GL, Mingoti SA, Heller L. Effect of environmental factors in reducing the prevalence of schistosomiasis in schoolchildren: An analysis of three extensive national prevalence surveys in Brazil (1950-2018). *PLoS Negl Trop Dis.* 2023 Jul 17;17(7):e0010804. doi: 10.1371/journal.pntd.0010804. PMID: 37459358; PMCID: PMC10374055.
5. Teresinha Schröder N, Fraga Da Silveira E, Thomasi Janhke Botton L, Périco E. Neglected diseases in Brazil: space-temporal trends and public policies [Internet]. *Neglected Tropical Diseases - Unsolved Debts for the One Health Approach.* IntechOpen; 2023. Available from: <http://dx.doi.org/10.5772/intechopen.1003000>
6. Silva, J. de P., Ramos, S. B., & Andrade, M. de. (2018). Multivariate analysis of schistosomiasis in the state of Minas Gerais: principal component analysis. *ABCS Health Sciences*, 43(2). <https://doi.org/10.7322/abcshs.v43i2.995>
7. Hofstraat, K., & van Brakel, W. H. (2016). Social stigma towards neglected tropical diseases: a systematic review. *International Health*, 8(suppl_1), i53-i70.
8. Zárate, L.E., Petrocchi , B., Dias Maia, C., Felix, C., & Gomes, M. P. (2023). CAPTO - A method for understanding problem domains for data science projects: CAPTO - Um método para entendimento de domínio de problema para projetos em ciência de dados. *Concilium*, 23(15), 922–941. <https://doi.org/10.53660/CLM-1815-23M33>.
9. Braga RD, Lucena FN, Ribeiro-Rotta RF. A preliminary multiprofessional information model for Brazilian primary care. *J Health Inform.* 2020 Apr/Jun; 14(2):319-326.



10. Harrison K, Rahimi N, Danovaro-Holliday MC. Factors limiting data quality in the expanded programme on immunization in low and middle-income countries: A scoping review. *J Health Inform.* 2020 Apr/Jun; 14(2):319-326.