Identificação de Fatores Associados à Artrite Utilizando Técnicas de Mineração de Dados

Anderson Babetto, Cristiane Neri Nobre, Luis Enrique Zárate

[1]Pontifícia Universidade Católica de Minas Gerais (PUC Minas) Belo Horizonte MG Brasil

[2]

anderbabetto@gmail.com, zarate@pucminas.br, nobre@pucminas.br

Abstract. *Arthritis is a public health challenge whose prevalence increases with aging, requiring effective population-level screening strategies. This study analyzes factors associated with arthritis in the Brazilian population aged 40 to 60 using 2019 National Health Survey (PNS) microdata to develop a classification model for public health screening. The pipeline integrates missing data imputation (KNNImputer), hybrid class balancing (RUS + SMOTE), and feature selection via Ant Colony Optimization (ACO). Bayesian search optimized the models, with Bernoulli Naive Bayes achieving the best performance (0.66 recall), prioritizing the minimization of false negatives. Results show that depression and chronic comorbidities have higher predictive power for arthritis than isolated dietary habits.*

Resumo. *A artrite é um desafio de saúde pública cuja prevalência aumenta com o envelhecimento, exigindo estratégias eficazes de triagem populacional. Este estudo analisa fatores associados à artrite na população brasileira entre 40 e 60 anos, utilizando microdados da Pesquisa Nacional de Saúde (PNS) de 2019, para desenvolver um modelo de classificação voltado à triagem. O pipeline integra imputação de dados (KNNImputer), balanceamento híbrido (RUS + SMOTE) e seleção de atributos via Ant Colony Optimization (ACO). A otimização por busca bayesiana indicou o Bernoulli Naive Bayes como o melhor modelo, alcançando recall de 0,66 e priorizando a redução de falsos negativos. Os resultados evidenciam que a depressão e comorbidades crônicas possuem maior poder preditivo que hábitos alimentares isolados.*

1. Introdução

A artrite representa um dos maiores desafios epidemiológicos da atualidade, sendo uma das principais causas de incapacidade funcional e redução da qualidade de vida no Brasil. Na faixa etária entre 40 e 60 anos, o impacto da doença é particularmente severo, pois coincide com o pico da produtividade laboral. O diagnóstico tardio nesse grupo não apenas resulta em danos articulares irreversíveis, mas frequentemente está associado a comorbidades psíquicas, como a depressão, gerando um elevado custo previdenciário e social [IBGE 2020, Skare et al. 2014].

Apesar da relevância clínica, o sistema de saúde brasileiro carece de ferramentas automatizadas que permitam a triagem populacional em larga escala [Silva et al. 2024].

A literatura recente [Gupta et al. 2024, Zhang et al. 2025] tem explorado o uso de Inteligência Artificial para identificar padrões inflamatórios, mas há uma lacuna na aplicação de modelos de mineração de dados que integrem variáveis socioeconômicas e de estilo de vida em grandes bases nacionais, como a Pesquisa Nacional de Saúde (PNS). O desafio técnico reside na complexidade desses dados, caracterizados por um volume massivo de registros, dados ausentes e um desbalanceamento severo entre indivíduos saudáveis e doentes.

Neste contexto, o presente trabalho propõe o desenvolvimento de um pipeline de mineração de dados para a classificação da artrite em brasileiros. Utilizando os microdados da PNS 2019, o estudo aplica técnicas de imputação, balanceamento híbrido e seleção de atributos via meta-heurística (Ant Colony Optimization). O objetivo é avaliar diferentes algoritmos de classificação e determinar o modelo com maior sensibilidade (Recall), priorizando a identificação eficaz de grupos de risco para intervenções precoces no âmbito da saúde pública.

Este artigo está organizado da seguinte forma: a Seção 2 apresenta uma breve revisão de trabalhos relacionados; a Seção 3 detalha a metodologia adotada, subdividida em materiais, métodos e métricas de avaliação; a Seção 4 descreve a modelagem e a análise inicial de resultados; a Seção 5 aprofunda a discussão por meio das avaliações detalhadas dos experimentos; e, por fim, a Seção 6 apresentam a explicabilidade do modelo, as limitações encontradas, as conclusões deste estudo e as propostas para trabalhos futuros.

2. Trabalhos Relacionados

A literatura explora a ciência de dados em doenças reumáticas sob diversas óticas. Internacionalmente, o aprendizado de máquina identifica biomarcadores moleculares [Zhang et al. 2025], enquanto modelos baseados em árvores e técnicas de interpretabilidade (SHAP) superam métodos tradicionais na distinção clínica entre artrite e osteoartrite [Khamparia et al. 2022, Gupta et al. 2024]. No Brasil, as pesquisas focam no impacto de perfis nutricionais e dietas pró-inflamatórias no quadro clínico [Oliveira et al. 2014, Skare et al. 2014], destacando a necessidade de técnicas de pré-processamento para lidar com a alta dimensionalidade de bases governamentais [Silva et al. 2024].

Este trabalho diferencia-se ao utilizar a escala nacional da PNS 2019 em vez de amostras clínicas restritas. O diferencial reside em um pipeline que integra imputação (KNNImputer) e balanceamento híbrido (RUS + SMOTE), técnica pouco explorada na epidemiologia nacional. Além disso, o uso da meta-heurística Ant Colony Optimization (ACO) otimiza a seleção de atributos, focando nas variáveis de maior impacto para a saúde pública.

3. Metodologia

Esta seção descreve os recursos computacionais e os procedimentos estatísticos e de mineração de dados adotados para a construção dos modelos preditivos.

3.1. Materiais

O estudo utiliza os microdados da Pesquisa Nacional de Saúde (PNS) de 2019, conduzida pelo IBGE em parceria com o Ministério da Saúde. A base de dados original é composta

por 279.382 registros e de 1081 atributos.

Para esta pesquisa, foi extraído um subconjunto focado na população brasileira entre 40 e 60 anos. 30 (trinta) atributos foram selecionados abrangendo dimensões socioeconômicas (escolaridade, renda), demográficas (sexo, idade), estilo de vida (frequência de exercícios, consumo de feijão/frutas) e histórico de saúde (presença de depressão). Estas dimensão estão representadas na Figura 1.

O pipeline foi desenvolvido em linguagem Python, utilizando bibliotecas especializadas como Scikit-learn para modelagem, Imbalanced- learn para balanceamento e Skopt para otimização bayesiana.

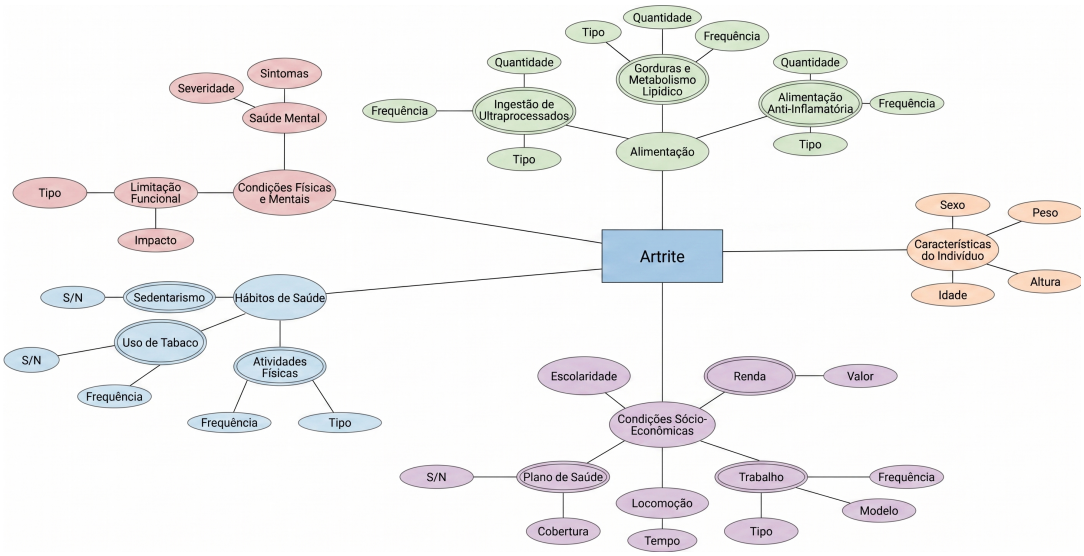

Figura 1. Diagrama Entidade-Relacionamento para o domínio de Artrite.

3.2. Métodos

O pipeline de processamento foi estruturado de forma a garantir a integridade dos dados e a reprodutibilidade dos resultados, seguindo as etapas descritas abaixo:

1. *Seleção Conceitual de Atributos:*
   Diferente da seleção meramente técnica, aqui os atributos foram escolhidos com base em evidências epidemiológicas que associam hábitos de vida e comorbidades a processos inflamatórios [Liu et al. 2024, Zhang et al. 2025]. Foram selecionadas cerca de 30 variáveis da PNS 2019 [IBGE 2020], incluindo fatores de risco com relevância clínica comprovada, como histórico de depressão [Khamparia et al. 2022] e hábitos alimentares [Wastyk et al. 2018] (veja Tabela 1).
2. *Preparação de Dados e Tratamento de Dados Ausentes:*
   A etapa de preparação iniciou-se pela engenharia de atributos, em que variáveis de peso e altura foram fundidas para o cálculo do Índice de Massa Corporal (IMC), enquanto métricas de frequência de exercícios foram agregadas para consolidar o perfil de atividade física dos indivíduos. Para otimizar o desempenho de classificadores não lineares, aplicou-se a técnica de discretização (*binning*) em variáveis

Tabela 1. Mapeamento conceitual e atributos finais mantidos para a modelagem.

Dimensão	**Atributos e Descrição (*Features*)**
Alvo (Target)	`Q079` (Diagnóstico médico de Artrite: 0 = Sim, 1 = Não).
Características do Indivíduo	`C006` (Sexo); `C008` (Idade); `P00104` (Peso); `W00203` (Altura).
Hábitos de Saúde	`P034` (Praticou exercício?); `P035` (Dias/semana de prática); `P068` (Frequência de fumo no domicílio - fumo passivo).
Alimentação	`P006;P018;P00901` (Alimentação Anti-inflamatória); `P01101;P02401;P015` (Gorduras e Metabolismo Lipídico); `P02002;P02501;P02602` (Ingestão de Ultra-processados).

contínuas, transformando-as em faixas categóricas que facilitam a captura de padrões de comportamento. No tratamento de inconsistências, códigos de erro presentes nos microdados originais foram mapeados e convertidos em valores nulos (NaN).

Para a recomposição da base, realizou-se uma comparação rigorosa de desempenho entre os algoritmos de imputação MissForest e K-Nearest Neighbors (KNNImputer). O KNNImputer foi o método selecionado, pois os testes de aderência comprovaram sua eficácia superior em preservar a distribuição estatística original dos dados da PNS, evitando a exclusão de registros fundamentais para a análise epidemiológica. Como etapa final do tratamento, implementou-se um ajuste de arredondamento dos dados imputados para números inteiros. Esta transformação foi imprescindível para manter a integridade semântica e lógica das variáveis provenientes dos questionários de saúde (como número de dias de exercício físico ou frequência semanal de consumo de certos alimentos), as quais não possuem significado prático se representadas por valores fracionários.

3. *Análise de Outliers:*

   A remoção de outliers foi realizada nas features numéricas utilizando a técnica do Z-Score (desvio-padrão). Esta técnica foi aplicada exclusivamente ao conjunto de treinamento após a divisão inicial da base. Foram removidos os registros cujos valores excedem um limiar de 3 desvios-padrão em relação à média de sua respectiva variável, minimizando o impacto de valores extremos no aprendizado do modelo sem introduzir vazamento de dados (*data leakage*) no conjunto de teste.

4. *Balanceamento do Conjunto de Dados:*

   A base de dados de treino, após encoding e scaling, apresentava um desbalanceamento significativo. Para resolver essa assimetria, foi implementada uma estratégia híbrida em duas etapas: Undersampling Seletivo (RUS): Inicialmente, o Random Under-Sampler (RUS) foi aplicado com um ratio de 0.1 (ou seja, reduzindo a classe majoritária até que ela representasse 10% do seu tamanho original em relação à minoritária). Esta etapa teve como objetivo reduzir apenas parcialmente a classe majoritária, preservando a maior parte de sua informação e preparando o dataset para a próxima fase. Oversampling (SMOTE): Em seguida, a técnica SMOTE (Synthetic Minority Oversampling Technique) foi aplicada. O SMOTE

gerou novos registros sintéticos da classe minoritária até que ela atingisse o tamanho da classe majoritária após a primeira etapa de RUS, resultando em uma distribuição de classes balanceada.

5. *Modelagem e Parametrização:*
   Foram selecionados seis algoritmos com diferentes paradigmas: Bernoulli Naive Bayes (NB), Árvore de Decisão (DT), Extra Tree (ET), Random Forest (RF), XGBoost e Redes Neurais Artificiais (RNA). A diversidade de modelos visa comparar abordagens probabilísticas, baseadas em árvores, *ensembles* e aprendizado profundo.
   A otimização de hiperparâmetros foi realizada via *BayesSearchCV* (`skopt`), utilizando busca bayesiana para maior eficiência computacional. O espaço de busca para cada modelo é detalhado na Tabela 2.

Tabela 2. Espaço de busca para otimização dos hiperparâmetros dos modelos implementados.

Modelo	Hiperparâmetro	Valores / Intervalo
Bernoulli NB	Alpha	[10^{-5}, 1000.0]
	fit_prior	[True, False]
Decision Tree (DT)	max_depth	[3, 15]
	min_samples_split	[2, 20]
	min_samples_leaf	[1, 10]
	criterion	gini, entropy
Random Forest (RF)	n_estimators	[50, 300]
	max_depth	[5, 20]
	min_samples_split	[2, 10]
	bootstrap	[True, False]
XGBoost	n_estimators	[50, 500]
	learning_rate	[0.01, 0.3] (log-uniform)
	max_depth	[3, 10]
	subsample	[0.5, 1.0]
	gamma	[0, 5]
Rede Neural (RNA)	hidden_layer_sizes	(50,), (100,), (50, 50)
	activation	relu, tanh
	solver	adam, sgd
	alpha	[10^{-5}, 10^{-2}]
Extra Tree (ET)	n_estimators	[20, 150]
	max_depth	[3, 15]
	min_samples_split	[2, 20]
	max_features	'sqrt', 'log2', None
	criterion	gini, entropy, log_loss

6. *Visualização e Estruturas de Dados:*
   Nesta etapa, aplicou-se a técnica de t-Distributed Stochastic Neighbor Embedding (t-SNE) para visualizar a estrutura intrínseca dos dados em duas dimensões. Este gráfico foi gerado a partir do conjunto de treino para observar a separabilidade das classes (0 = indivíduos saudáveis vs. 1 = indivíduos com Artrite) no espaço

de features. Pela Figura 2 observa-se que o problema é complexo, possuindo uma dificuldade de separação entre as classes.

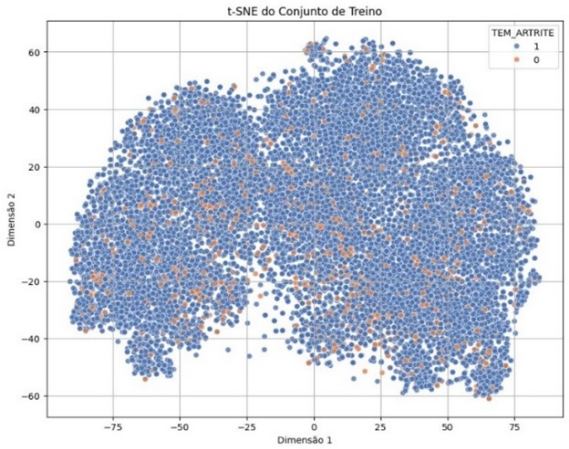

Figura 2. Visualização t-SNE do conjunto de Treino, destacando a distribuição das classes.

7. *Seleção de Atributos (Meta-heurística):*
   A redução de dimensionalidade foi realizada utilizando a meta-heurística *Ant Colony Optimization* (ACO), especificamente a variante *MAX-MIN Ant System* (MMAS) [Stützle and Hoos 2000]. Diferente de métodos de busca exaustiva, o MMAS utiliza uma lógica inspirada no comportamento de colônias de formigas para encontrar subconjuntos de atributos que maximizam o desempenho do classificador.
   No contexto deste projeto, o algoritmo opera através de agentes estocásticos (formigas) que percorrem o espaço de características, depositando feromônios nas combinações de atributos que resultam em melhor precisão preditiva. A variante MMAS foi escolhida por introduzir limites superiores e inferiores nas trilhas de feromônio, evitando a convergência prematura para ótimos locais e garantindo uma exploração mais robusta do espaço de busca [Dorigo et al. 2006]. Para a implementação, utilizou-se uma abordagem de filtragem baseada em métricas de importância, simulando o comportamento de seleção do MMAS para manter os 50% dos atributos com maior relevância estatística para a classificação da artrite.
8. *Processos de Treinamento e Avaliação de Qualidade:*
   O conjunto de dados foi dividido em 70% para treinamento e 30% para teste, utilizando estratificação para preservar a distribuição das classes. Para garantir a robustez dos hiperparâmetros e a generalização dos modelos, aplicou-se a técnica de validação cruzada com 10 dobras sobre o conjunto de treino.
   A avaliação do desempenho baseou-se em métricas de classificação: *Recall*[1], Precisão[2], e *F1-Score*[3]. Adicionalmente, utilizou-se o Índice de Confiabilidade (pHi)

[1]Recall = $\frac{VP}{VP+FN}$
[2]Precision = $\frac{VP}{VP+FP}$
[3]F1-Score = $2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$

para mensurar a estabilidade das probabilidades preditas nos acertos, e a métrica de acerto global (HIT) para avaliar a eficácia geral. O teste final foi realizado em dados inéditos e desbalanceados, simulando um cenário real de diagnóstico clínico onde o custo de falsos negativos é crítico.

4. Resultados e Discussões

Nesta seção, apresentam-se os desempenhos obtidos pelos modelos após a otimização de hiperparâmetros e a seleção de atributos via ACO. Os experimentos foram desenhados para validar a capacidade preditiva em um cenário de desbalanceamento real

4.1. Avaliação dos Modelos de Treinamento

Os modelos foram submetidos à Otimização Bayesiana com validação cruzada ($k = 10$), visando assegurar a estabilidade dos hiperparâmetros. Para aferir a consistência das predições, utilizou-se o Índice de Confiabilidade (pHi), que mensura a segurança do modelo nos acertos, e a métrica de acerto global (HIT). Os resultados consolidados da fase de treinamento são apresentados na Tabela 3.

Tabela 3. Comparação de consistência e acerto global entre os modelos (Média).

Modelo	pHi (Confiabilidade)	HIT (Precisão de Acerto)
XGBoost	**0,9233**	0,9184
Random Forest	0,9227	0,9235
Rede Neural (RNA)	0,8913	**0,9518**
Bernoulli Naive Bayes	0,8700	0,9300
Extra Tree	0,8400	0,9400
Decision Tree	0,8400	0,9500

Na fase de treinamento, o XGBoost destacou-se com o maior índice de confiabilidade ($pHi = 0,9233$), indicando que, quando o modelo classifica corretamente, ele o faz com alta probabilidade de certeza. Por outro lado, a Rede Neural apresentou o maior HIT $(0,9518)$, demonstrando uma eficácia superior na classificação geral da amostra de treino. Embora os modelos de *ensemble* e redes neurais tenham dominado as métricas de treino, a decisão final considerará o desempenho frente ao desbalanceamento do conjunto de teste.

4.2. Modelo Final e Conjunto de Treinamento

O modelo final foi consolidado utilizando o total do conjunto de treinamento, após a aplicação do pipeline de balanceamento híbrido (RUS + SMOTE) e a seleção do subconjunto ótimo de atributos via MMAS. Esta abordagem garantiu que o classificador final absorvesse toda a variabilidade disponível nos dados de treino antes de ser confrontado com o conjunto de teste.

4.3. Medidas Finais e Desempenho Comparativo

A avaliação definitiva foi conduzida no conjunto de teste (30% dos dados), mantido isolado e com o desbalanceamento original para simular a aplicação prática. A Tabela 4

Tabela 4. Quadro Comparativo de Métricas no Conjunto de Teste (Classe: Com Artrite)

Modelo	Precisão (%)	Recall (%)	F1-Score (%)	Acurácia Global
Bernoulli Naive Bayes	20,0	**66,0**	31,0	0,75
XGBoost	20,0	62,0	31,0	0,76
Extra Tree	23,0	56,0	33,0	0,88
Decision Tree	24,0	54,0	33,0	0,81
Rede Neural (RNA)	20,0	30,0	24,0	0,84
Random Forest	**24,0**	20,0	22,0	**0,88**

consolida o desempenho dos modelos, evidenciando o *trade-off* entre a precisão global e a sensibilidade na detecção da patologia.

Os resultados revelam que, embora o **Random Forest** e a **Rede Neural** apresentem maior robustez na classificação da classe majoritária (saudáveis), resultando em acurácias globais elevadas, eles falham em identificar uma parcela significativa dos indivíduos doentes. Em contrapartida, o **Bernoulli Naive Bayes** e o **XGBoost** demonstraram maior utilidade clínica para triagem, atingindo os maiores índices de *Recall* (66% e 62%, respectivamente), garantindo que mais casos positivos sejam capturados pelo sistema de saúde.

4.4. Discussão e Escolha do Modelo

A comparação revela um contraste: embora modelos como Random Forest e Redes Neurais obtenham maior acurácia, apresentam *Recall* insuficiente (20% e 30%), falhando em identificar a classe minoritária. Como o custo de um falso negativo em triagens públicas é altíssimo, tais modelos mostram-se inadequados ao objetivo clínico.

Assim, o Bernoulli Naive Bayes foi eleito o modelo definitivo, justificando-se por seu *Recall* superior (0,66), que supera inclusive o XGBoost. Priorizar a sensibilidade maximiza a detecção precoce da patologia, viabilizando intervenções que desoneram o sistema de saúde a longo prazo.

Destaca-se, contudo, o paradoxo técnico de um algoritmo simples e baseado em independência condicional (Bernoulli NB) superar modelos complexos e não lineares. Isso decorre do próprio *pipeline*: a forte seleção via ACO e a síntese de dados pelo SMOTE possivelmente reduziram a variância natural da base. Essa homogeneização tende a provocar *overfitting* local nas árvores e redes neurais durante o treinamento, prejudicando a generalização no teste isolado e favorecendo classificadores probabilísticos mais generalistas.

5. Análise do Conhecimento Extraído e Regras de Decisão

A extração de padrões revelou que os fatores de risco não atuam isoladamente. A coexistência de doenças crônicas e o diagnóstico de depressão emergiram como o indicador mais robusto, sugerindo que a inflamação sistêmica e as comorbidades aumentam substancialmente a probabilidade diagnóstica, corroborando preditores mapeados na literatura [Khamparia et al. 2022, Wastyk et al. 2018].

Quanto ao gênero, observou-se uma associação de risco mais acentuada no sexo masculino nesta subpopulação, atuando o sexo feminino como um fator de proteção relativo.

Em relação à idade, a probabilidade de diagnóstico foi superior na faixa dos 40-49 anos em comparação aos indivíduos de 55-60 anos. Para afastar potenciais vieses populacionais da base [IBGE 2020], avaliou-se a proporção de casos em relação ao total de respondentes de cada faixa, validando a tendência. Esse achado alerta para a necessidade de triagem precoce, desmistificando a doença como uma condição exclusiva da idade avançada [Gupta et al. 2024].

6. Conclusão

Este estudo desenvolveu um pipeline de mineração de dados especializado para a classificação da artrite, oferecendo uma ferramenta fundamentada para a saúde pública brasileira. A integração da meta-heurística *MAX-MIN Ant System* (MMAS) foi determinante para a redução de dimensionalidade, permitindo a seleção de um subconjunto de atributos com alto poder discriminatório e otimização do custo computacional. Dentre os seis modelos avaliados após a otimização Bayesiana, o Bernoulli Naive Bayes consolidou-se como a escolha mais estratégica para triagem populacional. Ao priorizar o *Recall* (0,66) em detrimento da acurácia global, o modelo garante que uma maior parcela da população portadora da patologia seja capturada pelo sistema de saúde, superando o desempenho de modelos mais complexos, como Random Forest e Redes Neurais, na identificação da classe minoritária.

No âmbito do conhecimento extraído, os resultados confirmaram que a saúde mental, especificamente o diagnóstico de depressão, e a presença de comorbidades crônicas possuem peso preditivo superior aos hábitos alimentares isolados na detecção da artrite. Contudo, devido à natureza transversal (*cross-sectional*) dos dados da PNS 2019, é crucial destacar que o modelo captura apenas associações estatísticas e não causalidade. É clinicamente plausível que a depressão seja uma consequência das dores crônicas da artrite, e não um fator de risco preditivo antecedente. Adicionalmente, a análise das regras de decisão revelou um importante paradoxo etário e de gênero na base, indicando que a patologia apresenta uma associação maior com indivíduos do sexo masculino em faixas etárias mais jovens (40-49 anos). Estes achados são fundamentais para o redirecionamento de políticas públicas, permitindo que as estratégias sejam calibradas para grupos que tradicionalmente não são o foco principal em diagnósticos de doenças osteoarticulares.

Apesar do rigor metodológico, o estudo enfrenta limitações inerentes à natureza dos dados autorreferidos da PNS, que podem introduzir vieses de percepção. O desbalanceamento extremo da base também impõe um limite natural à precisão, sendo um desafio persistente em triagens populacionais. Para trabalhos futuros, vislumbra-se a inclusão de variáveis geográficas e regionais aos modelos de classificação. Essa expansão permitiria compreender como fatores climáticos e disparidades no acesso ao saneamento influenciam a prevalência da artrite no Brasil, aprofundando o potencial preditivo e o impacto social da ferramenta desenvolvida.

Referências

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*.

Dorigo, M., Birattari, M., Blum, C., Clerc, M., Stützle, T., and Winfield, A. (2006). *Ant colony optimization*. Springer.

Fix, E. and Hodges, J. L. (1951). Discriminatory analysis. nonparametric discrimination: Consistency properties. *US Air Force School of Aviation Medicine*.

Gupta, S., Mehta, R., et al. (2024). Innovative machine learning approach for distinguishing rheumatoid arthritis and osteoarthritis. *Journal of Scientific Innovation in Medicine*.

IBGE (2020). *Pesquisa Nacional de Saúde 2019: percepção do estado de saúde, estilos de vida, doenças crônicas e saúde bucal*. Instituto Brasileiro de Geografia e Estatística, Rio de Janeiro.

Khamparia, A., Singh, P. K., et al. (2022). Prediction of hidden patterns in rheumatoid arthritis patients records using data mining. *Multimedia Tools and Applications*, 81.

Liu, X., Sun, M., et al. (2024). Diet affects inflammatory arthritis: a mendelian randomization study of 30 dietary patterns causally associated with inflammatory arthritis. *Frontiers in Nutrition*.

Mockus, J. (1978). The application of bayesian methods for seeking the extremum. *Journal of Global Optimization*.

Oliveira, A. K. et al. (2014). Avaliações dietética e antropométrica em artrite reumatoide juvenil. *Revista da Associação Médica Brasileira*.

Quinlan, J. R. (1993). C4.5: Programs for machine learning. *Morgan Kaufmann Publishers*.

Silva, J., Santos, M., et al. (2024). Técnicas de pré-processamento e seleção de atributos em bases de dados governamentais. *Anais do XXXIX Simpósio Brasileiro de Bancos de Dados*.

Skare, T. L. et al. (2014). Perfil nutricional do paciente com artrite reumatoide. *Revista Brasileira de Reumatologia*.

Stützle, T. and Hoos, H. H. (2000). Max–min ant system. *Future generation computer systems*, 16(8):889–914.

Vadell, A. K. et al. (2018). Anti-inflammatory diet in rheumatoid arthritis (adira): protocolo de ecr cruzado. *Nutrition Journal*.

Wastyk, H. C. et al. (2018). Machine learning with ingredient-level food trees reveals contributors to systemic inflammation. *American Journal of Clinical Nutrition*.

Zeng, P. et al. (2020). The relationship between dietary patterns and rheumatoid arthritis: casecontrol. *Nutrition & Metabolism*.

Zhang, Y., Wang, L., Liu, J., et al. (2025). Machine learning-driven insights into lipid metabolism and inflammatory pathways in knee osteoarthritis. *Frontiers in Nutrition*.