Uso de Aprendizado de Máquina para Identificar Perfis de Mulheres com Câncer de Mama no Brasil

Alice Cerbino Soares[1], **Laura Bianca Oliveira Gomes**[1], **Luis Enrique Zárate**[2]

[1]Curso Ciência de Dados e Inteligência Artificial – Pontifícia Universidade Católica de Minas Gerais (PUC-Minas)
CEP — 30140-100 — Belo Horizonte — MG — Brasil

alice.cerbino.me@gmail.com, laurabog2@icloud.com, zarate@pucminas.brr

Abstract. *This study aims to investigate the profile of women diagnosed with breast cancer in Brazil through the application of machine learning techniques. Using data from the 2019 Brazilian National Health Survey (PNS), sociodemographic and health variables were analyzed to identify regional patterns among patient profiles. The Decision Tree algorithm was employed to model the data and facilitate the interpretation of results. The conclusions indicate that the rules generated are mainly related to medical follow-up rather than sociodemographic and behavioral aspects.*

Resumo. *Este trabalho tem como objetivo investigar o perfil de mulheres diagnosticadas com câncer de mama no Brasil por meio da aplicação de técnicas de aprendizado de máquina. Utilizando dados da Pesquisa Nacional de Saúde (PNS) de 2019, foram analisadas variáveis sociodemográficas e de saúde com o intuito de identificar padrões regionais entre os perfis das pacientes. Os algoritmos de Árvore de Decisão e Random Forest foram empregados para modelar os dados e facilitar a interpretação dos resultados. As conclusões indicam que regras geradas estão principalmente relacionadas a acompanhamento médico, antes de aspectos sociodemográficos e comportamentais.*

1. Introdução

O câncer de mama é o tipo de câncer mais comum entre as mulheres e uma das principais preocupações de saúde pública no Brasil, com estimativas indicando um aumento significativo nos próximos anos. Segundo [INCA 2022], estima-se que 73.610 novos casos de câncer de mama sejam registrados até o final de 2025, com uma taxa de incidência de 66 casos por 100 mil mulheres. De acordo com o INCA, a neoplasia maligna da mama é a primeira causa de morte por câncer em mulheres no Brasil. Em 2022, as maiores taxas de mortalidade por câncer de mama foram registradas no Sul (12,69) e Sudeste (12,43), seguidas pelo Centro-Oeste (10,90), Nordeste (10,75) e Norte (8,59) óbitos por 100 mil mulheres.

Pesquisas recentes apontam desafios na capacitação de profissionais da Estratégia Saúde da Família (ESF) no controle do câncer de mama [Paixão et al. 2023]. Diante desse desafio, [Rocha et al. 2023] discute os principais métodos que podem contribuir na detecção precoce. Os resultados reforçam a importância da conscientização sobre sinais e sintomas, contribuindo para um tratamento mais eficaz.

O presente trabalho visa identificar os fatores que melhor caracterizam o perfil de mulheres diagnosticadas com câncer de mama, por meio de uma abordagem baseada na descoberta de conhecimento em bases de dados e na construção de modelos de aprendizado. Para isso, foi considerado a mais recente pesquisa em saúde realizada pelo IBGE, a Pesquisa Nacional em Saúde, PNS 2019 que recolhe por meio de questionários, informações acerca da saúde da população brasileira. Para este trabalho, foram utilizadas técnicas de preparação de dados e modelagem descritiva/preditiva, buscando compreender os principais determinantes associados à incidência da doença. Além disso, foi realizada uma análise comparativa entre dois modelos de classificação, com o objetivo de avaliar o desempenho preditivo e fornecer subsídios para estratégias de políticas públicas relacionadas ao diagnóstico precoce.

2. Trabalhos Relacionados

Estudos evidenciam desafios na capacitação de médicos e enfermeiros no controle do câncer de mama, o que afeta negativamente a qualidade da prevenção e do diagnóstico precoce [Paixão et al. 2023]. Além da formação profissional, fatores socioeconômicos e demográficos também contribuem para a detecção tardia. Segundo [Santos et al. 2022], mulheres jovens (20-49 anos), negras e pardas, com menor escolaridade, sem companheiro(a), encaminhadas pelo SUS e residentes em áreas não metropolitanas apresentam maior probabilidade de diagnóstico em estágio avançado.

Uma revisão de [Costa et al. 2021] apontou como fatores de risco a idade avançada, menarca precoce, menopausa tardia, ausência de filhos, primeira gravidez após os 30 anos e alterações hormonais. Há também consenso sobre o impacto de hábitos não saudáveis, como etilismo, tabagismo, sedentarismo e consumo frequente de alimentos industrializados. [Dourado et al. 2022] analisa mulheres jovens, muitas sem histórico familiar de câncer, e observaram maior prevalência entre 50 e 69 anos. Já [Lima et al. 2023] indicam que 90% dos casos são esporádicos, com forte influência de fatores hormonais e ambientais, e pouca relação com fatores genéticos.

Diante desse cenário, surgem iniciativas que aliam inteligência artificial ao diagnóstico. [Silva and Monteiro 2023] desenvolve um sistema baseado em redes neurais artificiais, com alta precisão para distinguir casos benignos e malignos, enquanto [Neto et al. 2021] propõe um classificador binário com dados de exames de sangue e características fisiológicas, atingindo 83,3% de acurácia. Esses avanços reforçam a importância da integração entre tecnologia e prática clínica. Ainda assim, é essencial fortalecer ações de conscientização que esclareçam sinais, sintomas e mitos sobre a doença, ampliando as chances de diagnóstico precoce e sucesso no tratamento.

3. Materiais e Métodos

A Figura 1 ilustra as etapas da metodologia proposta para este trabalho. As etapas foram aplicadas para construir o modelo de predição para caracterizar o perfil de mulheres com diagnóstico positivo e negativo para o câncer de mama.

3.1. Descrição da Base de Dados

Para a realização deste estudo, foi utilizada a base de dados da Pesquisa Nacional de Saúde (PNS) de 2019, conduzida pelo Instituto Brasileiro de Geografia e Estatística (IBGE).

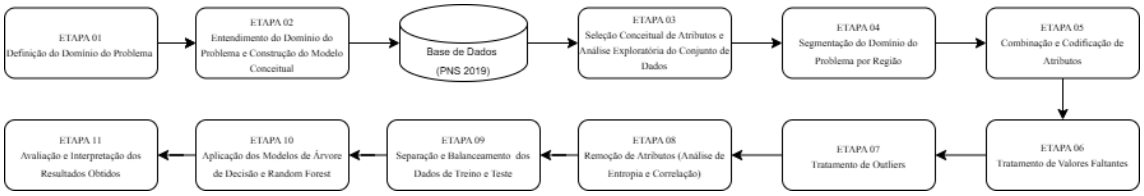

Figure 1. Fluxograma da Metodologia Proposta para Extração do Conhecimento

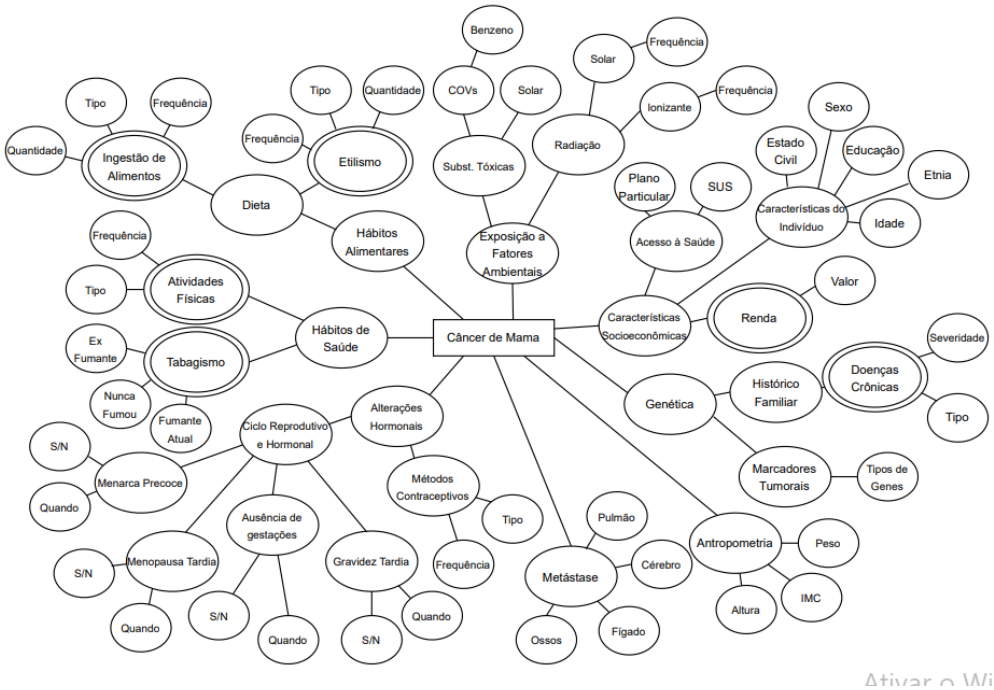

Figure 2. Modelo Conceitual do Câncer de Mama

Essa pesquisa abrange dados relacionados às condições de trabalho, rendimentos domiciliares, uso de serviços de saúde e informações demográficas dos participantes. A base possui 293.726 registros e 1.088 atributos. Para este estudo, foi realizado um corte para o câncer de mama de mulheres entre 40 e 75 anos, contendo 24725 registros referentes a mulheres que não foram diagnosticadas e 363 para aquelas que de fato apresentaram um diagnóstico positivo para a doença.

3.2. Entendimento do Problema e seleção conceitual de atributos

A compreensão do domínio do problema é uma etapa essencial no processo de descoberta de conhecimento. Um entendimento prévio permite reconhecer a complexidade envolvida no problema e identificar conhecimento útil e não óbvio sobre o domínio em estudo. Dada a alta dimensionalidade da base de dados da PNS 2019, a seleção conceitual de atributos torna-se uma estratégia particularmente relevante. Para este estudo, foi desenvolvido um modelo conceitual, apresentado na Figura 2, construído com base no método CAPTO, recentemente proposto por [Gonçalves et al. 2024].

O método CAPTO propõe uma abordagem que integra o conhecimento tácito com o conhecimento explícito. Essa integração visa fornecer uma compreensão aprofundada do domínio antes da aplicação de algoritmos de aprendizado de máquina, possibilitando a

Table 1: **Atributos selecionados a partir do mapa conceitual**

Dimensão	**Aspecto**	**Variáveis PNS**
Hábitos de saúde	Tabagismo	P50 e P06701
Hábitos de saúde	Atividades Físicas	P03904
Hábitos Alimentares	Dieta - Etilismo	P027
Hábitos Alimentares	Dieta - Ingestão de Alimentos	P02801, P05401, P035, P00901, P01101, P013, P015, P018
Genética	Marcadores Tumorais	Indisponível
Genética	Histórico Familiar - Doenças Crônicas	Q00201, Q03001, Q060, Q06306
Características Socioeconômicas	Sexo	C006
Características Socioeconômicas	Idade	C008
Características Socioeconômicas	Educação	D00301, D00901, D014
Características Socioeconômicas	Estado Civil	C01001, C011, C014
Características Socioeconômicas	Raça	C009
Características Socioeconômicas	Renda	E01602 e E01802
Características Socioeconômicas	Acesso á Saúde	R013, R014, R015, R01701, R019, R020, R02101, R022, R023, I00102
Antropometria	Peso e Altura	P00103 e P00403
Alterações Hormonais	Métodos Contraceptivos	R034, R03601, R03607, R03608, R03610
Alterações Hormonais	Ciclo Reprodutivo e Hormonal - Gravidez Tardia	S066 e S06703
Alterações Hormonais	Ciclo Reprodutivo e Hormonal - Ausência de Gestações	S065
Alterações Hormonais	Ciclo Reprodutivo e Hormonal - Menopausa Tardia	R028
Alterações Hormonais	Ciclo Reprodutivo e Hormonal - Menarca Precoce	R025
Exposição e Fatores Ambientais	Substâncias Tóxicas - COVs	M011011
Exposição e Fatores Ambientais	Substâncias Tóxicas - Hormônios	M011011
Exposição e Fatores Ambientais	Radiação - Solar	M011031
Exposição e Fatores Ambientais	Radiação - Ionizante	M011041
Metástase	Câncer nos ossos	Indisponível
Metástase	Câncer de pulmão	Q12104
Metástase	Câncer no fígado	Indisponível
Metástase	Câncer no cérebro	Q121013

redução da dimensionalidade por meio da seleção dos atributos mais relevantes. A Tabela 1 apresenta os atributos selecionados com base no modelo conceitual resultante dessa abordagem.

3.3. Pré-Processamento e preparação de dados

Este trabalho aplicou o teste Qui-quadrado de independência para verificar se há associação significativa entre as regiões do Brasil e a condição de saúde (doente ou não doente) da população. A decisão estatística seguiu os seguintes critérios: para $Q < \chi^2$ crítico ou p-valor $> 0,01$, aceita-se H_0 (sem associação significativa); para $Q > \chi^2$ crítico ou p-valor $< 0,01$, rejeita-se H_0 (há associação significativa). Com base nos resultados obtidos, foi possível identificar quais regiões possuem distribuição proporcionalmente semelhante de casos, possibilitando a união de grupos com comportamento estatístico equivalente. Conforme mostra a Figura 3, a análise foi direcionada às regiões Norte e Sudeste, que tiveram associação significativa e diferenças relevantes para o estudo.

Devido à alta taxa de dados ausentes nos atributos selecionados, foi necessário realizar combinações entre eles a fim de minimizar o impacto causado por essas ausências. Para isso, nove atributos foram combinados com base no produto cartesiano, considerando suas relações semânticas, além do atributo classe (diagnóstico). Foram eles: frequencia_bebida, alimentação, escolaridade, consulta, metodo_contraceptivo, class_frequencia_tabaco, interpretacao_imc e pagou_ou_sus_mamografia. A seguir, apresentamos alguns desses atributos combinados:

Diagnóstico: A variável diretamente relacionada ao diagnóstico do câncer de mama (Q12701) é um complemento da variável que diz respeito ao diagnóstico de câncer em geral (Q120). Portanto, analisando Q12701 individualmente, havia poucos registros de pessoas totalmente saudáveis.

Se Q120 = 1 e Q12107 = 1: 'possui diagnóstico' Se Q120 = 2 e Q12107 = 2: 'não possui diagnóstico'
Se Q120 = 1 e Q12107 = 2: 'não possui diagnóstico' Se Q120 = 2 e Q12107 = 1: 'Incoerente' (Não houveram casos)

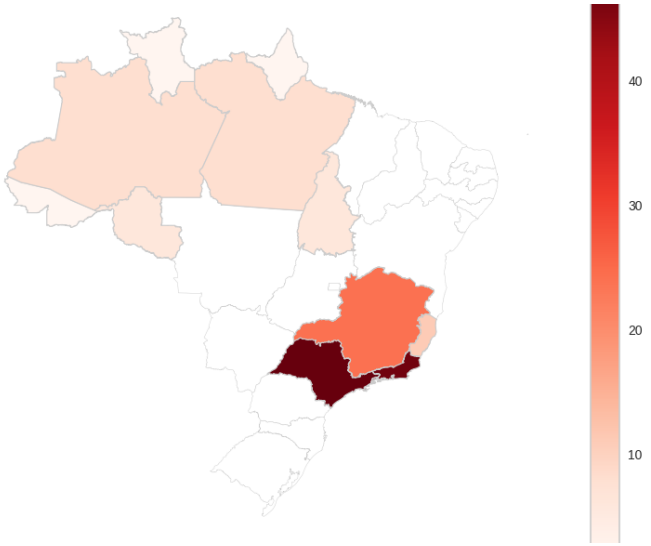

Figure 3. Casos diagnosticados nas regiões Norte e Sudeste do Brasil.

Frequencia_bebida: A estratégia utilizada para a criação deste atributo combina duas respostas: a frequência geral (P027) e a semanal (P02801) de consumo.

Se P027 = '1' / Se P027 = '3' e P02801 = '0': Nunca
Se P027 = '3' e P02801 em ['1', '2', '3']: Eventual
Se P027 = '3' e P02801 em ['6', '7']: Quase diariamente
Se P027 = '2': Menos de uma vez por mês
Se P027 = '3' e P02801 em ['4', '5']: Frequente
Se P027 = '3' e P02801 com outro valor: ignorado

Alimentação: Une os indicadores de consumo de verduras e legumes (P00901), carne (P01101 e P013) e frutas (P018) para a construção de uma dieta saudável. Considera-se frequência saudável quando $P00901 \geq 3$, $P01101 \geq 3$ ou $P013 \geq 3$ dias na semana. Por outro lado, caracteriza-se excesso de carne quando $P015 \geq 4$ ou $P018 \geq 4$ dias na semana.

Se frequência saudável ≥ 2 e excesso de carne = 0: Muito saudável
Se frequência saudável = 0 e excesso de carne ≥ 1: Ruim
Se frequência saudável ≥ 1 e excesso de carne ≤ 1: Saudável
Em qualquer outro caso: Regular

Após uma análise de proporcionalidade dos valores de atributos categóricos, quando há proporcionalidade semelhante entre valores categóricos adjacentes, essas respostas foram agrupadas para reduzir a dimensionalidade e facilitar a interpretação do modelo. Durante esse processo, atributos como atividade_fisica sofreram essa redução de valores categóricos.

A fim de eliminar redundâncias, foram identificadas e removidas instâncias duplicadas garantindo a unicidade dos registros. Além disso, os atributos foram organizadas conforme sua natureza — categóricas nominais, ordinais, numéricas e binárias — e codificadas adequadamente. Os atributos categóricas foram codificados por rótulo, mantendo a estrutura original sem aplicação de One-Hot Encoding, dado o impacto negativo no número de atributos e a compatibilidade dos modelos utilizados. Não foi necessário o tratamento de outliers devido ao fato de não haver atributos numéricos após as junções e transformações citadas acima. Em geral, para valores ausentes, aplicou-se a imputação por k-nearest neighbors (KNN). Para atributos relacionados à exposição e atividade física foi adotada uma estratégia distinta: os valores ausentes foram preenchidos com zero, assumindo que a ausência de resposta indica ausência de exposição e da prática.

Atributos relacionados a doenças crônicas foram excluídos por apresentarem entropia muito baixa, apesar de serem fatores de risco. Após o pré-processamento, restaram 16 atributos, listados na Tabela 2. A base de dados está disponível em https://github.com/licapLaboratory/Database-C-ncer-Mama

Table 2: Atributos da base após o pré-processamento

Atributos	**Descrição**	**Tipo do dado**	**Entropia Padrão**	**Entropia Condicional**	**Possíveis valores**
diagnostico	Diagnóstico de câncer de mama da entrevistada.	Dicotômica	-	0.1197	'1: Sim', '0: Não'
tempo_resultado_mamografia	Tempo que a mulher levou para receber o resultado da mamografia.	Categórico ordinal	0.9896	0.1193	'0: Tempo razoável', '1: Demorou / Nunca recebeu', '2: Não foi buscar / Ainda não recebeu'
entrou_menopausa	Se a mulher já passou pela menopausa.	Dicotômica	0.2273	0.1194	'1: Sim', '0: Não'
exp_subst_quimicas	Exposição a substâncias químicas.	Dicotômica	0.9934	0.1186	'1: Sim', '0: Não'
exp_sol	Exposição frequente ao sol.	Dicotômica	0.9909	0.1191	'1: Sim', '0: Não'
exp_material_radioativo	Indica exposição a materiais radioativos.	Dicotômica	0.9989	0.1188	'1: Sim', '0: Não'
alimentacao	Alimentação em termos de qualidade nutricional.	Categórico nominal	1.5051	0.1192	'0: Muito saudável', '1: Regular', '2: Ruim', '3: Saudável'
consulta	Se a mulher foi encaminhada para consulta médica.	Categórico nominal	0.9030	0.1173	'0: Encaminhada e consultou', '1: Encaminhada, mas não consultou', '2: Não encaminhada'
metodo_contraceptivo	Tipo de método contraceptivo utilizado.	Categórico nominal	0.8360	0.1188	'0: Não usa', '1: Usa hormonal', '2: Usa não hormonal'
frequencia_tabaco	Frequência de uso de produtos com tabaco.	Dicotômica	0.5104	0.1196	'0: Fuma', '1: Não fuma'
raca	Autodeclaração da raça/cor da entrevistada.	Categórico nominal	1.0738	0.1190	'0: Amarela', '1: Branca', '2: Indígena', '3: Preta'
ultimo_exame_mama	Se a mulher fez ou não o exame de mama.	Dicotômica	0.6927	0.1160	'0: Nunca fez', '1: Fez'

Atributos	Descrição	Tipo do dado	Entropia Padrão	Entropia Condicional	Possíveis valores
ultima_mamografia	Tempo desde a última mamografia realizada.	Categórico ordinal	1.7360	0.1157	'0: Menos de 1 ano', '1: Entre 1 a 3 anos', '2: 3 anos ou mais'
partos	Número de partos que a mulher teve ao longo da vida.	Categórico ordinal	1.0350	0.1176	'0: Nenhum parto', '1: 1 a 3 partos', '2: 4 ou mais partos'
bebida	Frequência com que a mulher consome bebidas alcoólicas.	Categórico ordinal	1.0960	0.1193	'0: Nunca', '1: Às vezes', '2: Frequentemente'
escolaridade	Nível de escolaridade mais alto alcançado pela mulher.	Categórico ordinal	1.6059	0.1187	'0: Educação Infantil', '1: Ensino Fundamental', '2: Ensino Médio', '3: Educação Superior', '4: Pós-graduação'

4. Resultados

4.1. Treinamento e validação do modelo

O conjunto de dados foi separado em treino e teste (estratégia *hold-out*) numa proporção de 70% e 30%, respectivamente. Para o processo de treinamento foi realizada validação cruzada com 10 dobras, e aplicado o balanceamento de classes por undersampling. Para a construção dos modelos, foi utilizado o algoritmo *ensemble*, caixa-preta Floresta Aleatória e o algoritmo caixa-branca Árvore de Decisão, pela sua capacidade interpretativa. Com o intuito de otimizar o desempenho dos classificadores, foi utilizado o método Grid Search para a seleção dos hiperparâmetros, garantindo uma busca sistemática pelas melhores combinações possíveis dentro de um espaço pré-definido. Os parâmetros utilizados pela Floresta Aleatória foram: Critério = entropy, max_depth = 10, max_features = 2, min_samples_leaf = 1, min_samples_split = 10 e n_estimators = 70, apresentando 0.75 como maior pontuação da validação cruzada. Enquanto para a árvore de decisão foram Critério = entropy, max_depth = 10, max_features = 2, min_samples_leaf = 1, min_samples_split = 5.

4.2. Análise dos resultados

Conforme apresentado na Tabela 5, o modelo de Árvore de Decisão apresentou uma acurácia de 59%, com desempenho não equilibrado entre as classes. Para a classe "Sem Câncer", o modelo obteve uma precisão de 64%, indicando que uma boa parte das previsões dessa classe foram corretas, porém com um recall de apenas 43%, revelando dificuldade em identificar todos os casos negativos. Já para a classe "Com Câncer", o modelo apresentou um recall mais elevado, de 76%, mostrando que conseguiu identificar a maioria dos casos positivos, embora com uma precisão mais baixa, de 57%.

A avaliação do modelo de Random Forest, conforme delineado na Tabela 6, revela uma acurácia global de aproximadamente 65%, com uma performance mais equilibrada entre as classes. Para a classe 0 (Sem Câncer), o modelo apresentou uma precisão de 66%

e um recall de 63%, indicando um desempenho razoável tanto na identificação correta quanto na cobertura dos casos negativos. Para a classe 1 (Com Câncer), o modelo obteve precisão de 65% e recall de 67%, demonstrando boa capacidade de identificar corretamente os casos positivos. O equilíbrio entre precisão e recall nos dois grupos se refletiu nos valores de F1-score, que foram 65% para a classe 0 e 66% para a classe 1.

Table 3. Métricas dos modelos no teste

Modelo	Classe	Precisão	Recall	F1-Measure	Acurácia
Árvore de Decisão	0: Não possui câncer	0.64	0.43	0.51	0.59
	1: Possui câncer	0.54	0.76	0.65	0.59
Random Forest	0: Não possui câncer	0.66	0.63	0.65	0.65
	1: Possui câncer	0.65	0.67	0.66	0.65

A matriz de confusão (Tabela 6) mostra que o modelo acertou 31 dos 49 casos positivos, mas classificou 18 negativos como positivos. Isso indica boa sensibilidade, porém com geração de falsos positivos, o que pode causar preocupações e exames desnecessários no diagnóstico de câncer de mama. Para a Árvore de Decisão (Tabela 5), o modelo também acertou 28 casos positivos reais, mas errou em 21 negativos, reforçando a presença de alarmes falsos que podem impactar a confiança no diagnóstico e levar a procedimentos adicionais.

Table 4. Matriz de confusão - Random Forest para Teste

	0 - Não possui câncer	1 - Possui câncer
0	21	28
1	12	37

Table 5. Matriz de confusão - Árvore de Decisão Teste

	0 - Não possui câncer	1 - Possui câncer
0	31	18
1	16	33

A principal limitação dos modelos está na classe ”possui câncer”. Apesar do recall elevado, a baixa precisão indica uma alta taxa de falsos positivos, como confirmado pelas matrizes de confusão.

4.3. Interpretação das regras

A fim de entender o comportamento de cada modelo em relação aos fatores mais relevantes, foram extraídas as cinco principais regras, aquelas que classificaram os maiores números de instâncias.

Table 6: Regras com as maiores coberturas geradas pelos modelos

Random Forest	Árvore de Decisão
SE ultimo_exame_mama = Nunca fez ENTÃO diagnostico = Saudável.	SE partos = 1 a 3 partos E escolaridade = Ensino fundamental E consulta = Encaminhada, mas não consultou ENTÃO diagnostico = Doente.
SE metodo_contraceptivo = Usa não hormonal E exp_subst_quimicas = Não E exp_material_radioativo = Não E partos = 1 a 3 partos E ultimo_exame_mama = Fez E raca = Branca E tempo_resultado_mamografia = Tempo razoável E frequencia_tabaco = Fuma E bebida = Nunca ENTÃO diagnostico = Doente.	SE partos = 4 ou mais partos E ultimo_exame_mama = Fez E entrou_menopausa = Não E exp_subst_quimicas = Não E bebida = Nunca E frequencia_tabaco = Fuma E escolaridade = Ensino fundamental E exp_material_radioativo = Não E tempo_resultado_mamografia = Tempo razoável E metodo_contraceptivo = Usa hormonal ENTÃO diagnostico = Doente.

Continua na próxima página

Random Forest	**Árvore de Decisão**
SE bebida = Às vezes E ultimo_exame_mama = Nunca fez ENTÃO diagnostico = Saudável.	SE partos = 1 a 3 partos E ultimo_exame_mama = Fez ENTÃO diagnostico = Saudável.
SE exp_material_radioativo = Não E escolaridade = Educação superior E raca = Branca E bebida = Nunca E frequencia_tabaco = Fuma E ultima_mamografia = Menos de 1 ano E metodo_contraceptivo = Usa hormonal E partos = 1 a 3 partos E tempo_resultado_mamografia = Tempo razoável ENTÃO diagnostico = Doente.	SE partos = 1 a 3 partos E escolaridade = Educação superior E exp_sol = Sim E bebida = Nunca E ultima_mamografia = Menos de 1 ano E entrou_menopausa = Não E ultimo_exame_mama = Fez E frequencia_tabaco = Fuma E alimentacao = Saudável ENTÃO diagnostico = Saudável.
SE ultima_mamografia = Menos de 1 ano E entrou_menopausa = Não E ultimo_exame_mama = Fez E exp_subst_quimicas = Não E tempo_resultado_mamografia = Tempo razoável E metodo_contraceptivo = Usa não hormonal E alimentacao = Regular E exp_sol = Não E raca = Indígena ENTÃO diagnostico = Doente.	SE partos = 1 a 3 partos E escolaridade = Ensino Fundamental E exp_sol = Sim E bebida = Às vezes E raca = Indígena E frequencia_tabaco = Não fuma E ultima_mamografia = Entre 1 a 3 anos E consulta = Encaminhada, mas não consultou E ultimo_exame_mama = Fez ENTÃO diagnostico = Doente.

As regras extraídas pelos modelos Random Forest e Árvore de Decisão revelam que os diagnósticos de câncer estão fortemente associados a fatores comportamentais e ao acesso a práticas preventivas. A ausência de exames de rotina, como mamografia e exame de mama, associada ao tabagismo, maior número de partos e uso de métodos contraceptivos hormonais, aparece com frequência em regras que levam ao diagnóstico da doença. Esses padrões sugerem que o risco não está apenas em fatores biológicos, mas na falta de acompanhamento médico regular.

Por outro lado, as regras que resultam em diagnóstico Saudável costumam incluir mulheres com hábitos de vida saudáveis, como boa alimentação, ausência de tabagismo e consumo moderado ou nulo de álcool, além de nível educacional mais elevado e acesso a exames periódicos. Isso reforça que condições sociais mais favoráveis contribuem para maior adesão a cuidados preventivos, influenciando positivamente o diagnóstico.

Entre os atributos mais relevantes estão o último exame de mama e a última mamografia, seguidos por fatores como consulta médica, escolaridade, número de partos e exposição a agentes nocivos. Conclui-se que, embora características sociodemográficas tenham papel importante, as práticas preventivas são determinantes no desfecho clínico, indicando caminhos para políticas públicas que incentivem o cuidado contínuo com a saúde feminina.

5. Conclusões

Embora tenha apresentado resultados relevantes, este estudo possui algumas limitações que devem ser consideradas. O forte desbalanceamento da base de dados impactou diretamente o desempenho dos modelos, contribuindo para um número elevado de falsos positivos, especialmente no Random Forest. Além disso, a pesquisa foi restrita à aplicação de dois algoritmos — Random Forest e Árvore de Decisão —, sem a exploração de outros modelos de aprendizado de máquina que poderiam proporcionar um desempenho mais equilibrado entre precisão e recall, principalmente na detecção dos casos positivos de câncer de mama. Diante disso, recomenda-se que trabalhos futuros contemplem a utilização de algoritmos mais robustos, como redes neurais ou métodos de ensemble, além da adoção de técnicas específicas para correção do desbalanceamento dos dados.

A análise das regras extraídas também permitiu observar padrões importantes no

perfil dos indivíduos diagnosticados, como a falta de mamografia recente, histórico de exposição a agentes químicos e ausência de acompanhamento médico periódico. Esses fatores estão frequentemente associados às regras que indicam maior risco de câncer, destacando a relevância de ações preventivas e do acesso facilitado aos exames de rastreamento. Dessa forma, este trabalho reforça a necessidade de fortalecer campanhas de prevenção e de garantir, por meio de políticas públicas, o acesso aos exames de detecção precoce e a profissionais da área da saúde devidamente capacitados. Além disso, fica evidente a importância de pesquisas futuras, tanto para o aprimoramento dos modelos preditivos quanto para a ampliação e diversificação das bases de dados, visando minimizar vieses e melhorar a acurácia dos sistemas de suporte ao diagnóstico.

Agradecimentos

Os autores agradecem o apoio recebido do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Processo No 303133/2021-0, e da Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG), Processo PCE-00349-25.

6. Referências Bibliográficas

Costa, L. S. et al. (2021). Fatores de risco relacionados ao câncer de mama e a importância da detecção precoce para a saúde da mulher. *Revista Eletrônica Acervo Científico*, 31.

Dourado, C. A. R. d. O. et al. (2022). Câncer de mama e análise dos fatores relacionados aos métodos de detecção e estadiamento da doença. *Cogitare Enfermagem*, 27.

Gonçalves, L., Franca, D., and Zarate, L. (2024). Relevância do entendimento do domínio de problema na construção de modelos computacionais de aprendizado. In *Anais do XVIII Brazilian e-Science Workshop*, pages 135–142, Porto Alegre, RS, Brasil. SBC.

INCA (2022). Atlas da mortalidade, instituto nacional de câncer. Acesso em: 01 jul. 2025.

Lima, R. F., Silva, M. T., and Souza, J. R. (2023). Câncer de mama em mulheres no brasil: epidemiologia, fisiopatologia, diagnóstico e tratamento: uma revisão narrativa. *Brazilian Journal of Development*.

Neto, J. d. D. E. S. et al. (2021). Modelos de aprendizado de máquina aplicados à detecção de câncer de mama. In *Anais do 15º Congresso Brasileiro de Inteligência Computacional*, pages 1–7. Sociedade Brasileira de Inteligência Computacional.

Paixão, M. C. et al. (2023). Enfrentamento da problemática do câncer de mama na estratégia da saúde da família. *Brazilian Journal of Implantology and Health Sciences*, 5(5):1501–1509.

Rocha, J. R. B. d., Santos, J. B., et al. (2023). Rastreamento, diagnóstico e tratamento do câncer de mama: cenário do brasil. *Revista Brasileira de Cancerologia*, 72(4):347–358.

Santos, T. B. d. et al. (2022). Prevalência e fatores associados ao diagnóstico de câncer de mama em estágio avançado. *Ciência Saúde Coletiva*, 27(2):471–482.

Silva, L. C. A. d. and Monteiro, J. R. (2023). Sistema para apoio ao diagnóstico de câncer de mama baseado em redes neurais artificiais. In *Engenharias - Automação, Robótica, Metrologia e Energia: Estudos e Tendências*, volume 1, pages 49–65.