Qual a influência de hábitos de vida e fatores socioeconômicos na ocorrência de câncer de próstata no Brasil?

Marco Antonio de Souza, Camila Nascimento Monteiro, Cláudia Renata dos Santos Barros

https://doi.org/10.1590/SciELOPreprints.7566

Submetido em: 2023-11-30
Postado em: 2023-12-08 (versão 1)
(AAAA-MM-DD)

Qual a influência de hábitos de vida e fatores socioeconômicos na ocorrência de câncer de próstata no Brasil?

What is the influence of lifestyle habits and socioeconomic factors on the occurrence of prostate cancer in Brazil?

¿Cuál es la influencia de los hábitos de vida y de los factores socioeconómicos en la aparición del cáncer de próstata en Brasil?

Marco Antonio de Souza[1] - https://orcid.org/0000-0003-3340-5912
Camila Nascimento Monteiro[2] - https://orcid.org/0000-0002-0121-0398
Cláudia Renata dos Santos Barros[3] - https://orcid.org/0000-0002-1582-2010

[1]Universidade de São Paulo, Instituto de Física, São Paulo, SP, Brasil
[2]Hospital Sírio-Libanês, Saúde Populacional, São Paulo, SP, Brasil
[3]Instituto Butantan, São Paulo, SP, Brasil

CORRESPONDÊNCIA
Marco Antonio de Souza | e-mail: marsouza@if.usp.br

RESUMO

Objetivo: investigar características físicas, de hábitos de vida e socioeconômicas que podem estar associadas à ocorrência de câncer de próstata no Brasil. **Métodos:** uma base de microdados referente à Pesquisa Nacional de Saúde 2019 foi utilizada, com a seleção de 42.799 indivíduos do sexo masculino; este grupo foi analisado por meio de métodos estatísticos e modelagem por *machine learning* (regressão logística e árvore de decisão). **Resultados:** os modelos aplicados permitiram identificar com bom nível de acurácia os indivíduos que receberam o diagnóstico de câncer de próstata (DCP), além de grupos com características específicas mais fortemente associados a esta doença. **Conclusão:** os modelos indicam uma influência significativa de fatores socioeconômicos, físicos e alimentares na frequência de DCP no grupo analisado. O alto nível de acurácia e sensibilidade dos modelos demonstra o potencial dos métodos de *machine learning* para a previsão de DCP.

Palavras-chave: câncer de próstata, estilo de vida, estudo transversal, machine learning

ABSTRACT

Objective: the investigation of physical, lifestyle and socioeconomic features that may be associated with the occurrence of prostate cancer in Brazil. **Methods:** a microdata base referring to the 2019 National Health Survey in Brazil was used, with the selection of 42,799 male individuals; this group was analyzed using statistical methods and machine learning modeling (logistic regression and decision tree). **Results:** the models applied allowed us to identify with a good level of accuracy individuals with prostate cancer diagnosis (DCP), in addition to groups with specific features more strongly associated with such a disease. **Conclusion:** the models indicate a significant influence of socioeconomic, physical and dietary factors on the frequency of DCP in the analyzed group. The high level of accuracy and sensitivity of the models demonstrates the potential of machine learning methods for predicting DCP.

Keywords: prostate cancer, lifestyle, cross-sectional study, machine learning

RESUMÉN

Objetivo: investigar características físicas, de estilo de vida y socioeconómicas que pueden estar asociadas con la aparición de cáncer de próstata en Brasil. **Métodos:** se utilizó una base de microdatos referente a la Encuesta Nacional de Salud de 2019, con la selección de 42.799 individuos del sexo masculino; este grupo fue analizado mediante métodos estadísticos y modelado de *machine learning* (regresión logística y árbol de decisión). **Resultados:** los modelos aplicados permitieron identificar con buen nivel de exactitud a los individuos con diagnóstico de cáncer de próstata (DCP), además de grupos con características específicas más fuertemente asociadas a esta enfermedad. **Conclusión:** los modelos indican influencia significativa de factores socioeconómicos, físicos y dietéticos sobre la frecuencia de DCP en el grupo analizado. El alto nivel de exactitud y sensibilidad de los modelos demuestra el potencial de los métodos de *machine learning* para predecir la DCP.

Palabras clave: cáncer de próstata, estilo de vida, estudio transversal, machine learning

INTRODUÇÃO

No Brasil, o câncer de próstata é o segundo mais comum entre os homens, atrás apenas do câncer de pele não melanoma[1], com estimativa de 1 em cada 8 homens com câncer de próstata no decorrer da vida em outros países como nos Estados Unidos.[2] Esse tipo de câncer é multicausal e os principais fatores de risco identificados são idade, cor/etnia, nacionalidade, histórico familiar e alterações genéticas. Há também outros fatores associados ao hábito de vida

relacionados ao câncer de próstata que têm sido estudados para melhor conhecimento de possível relação causal que são: dieta, obesidade, tabagismo, exposição ocupacional, inflamação da próstata, doenças sexualmente transmissíveis e vasectomia.[3] Além disso, fatores sociais e econômicos têm uma forte influência sobre os hábitos de vida da população e as suas condições de acesso aos serviços de saúde, podendo, em tese, influenciar a ocorrência e diagnóstico de câncer de próstata (DCP) na população masculina. Assim, o objetivo deste estudo foi analisar os fatores associados ao câncer de próstata no Brasil.

MÉTODOS

Desenho do estudo

Estudo transversal com dados secundários provenientes da Pesquisa Nacional de Saúde (PNS), realizada em 2019.

População e amostra

A base de dados contém 42.799 indivíduos, destes 339 (0,79 %) disseram "sim" para DCP em algum momento da sua vida e 42.460 (99,21%) disseram "não".

Variáveis

Entre as 58 variáveis selecionadas inicialmente, estavam a variável dependente "câncer de próstata" e 57 variáveis descritivas. Após a aplicação de critérios de eliminação de variáveis (detalhes nas Subseções "*Filtragem das variáveis pelo IV"* e "*Regressão logística e árvore de decisão"*), restaram 13 variáveis independentes as quais foram aplicadas nos modelos preditivos. Elas são: idade; há quantos anos a pessoa consultou um médico pela última vez; como a pessoa avalia a sua própria saúde (muito boa, boa, ruim, etc.); quantos dias da semana consome frutas; quantos dias da semana consome verduras e/ou legumes; quantos dias da semana consome sucos artificiais; se já recebeu diagnóstico de nível alto de colesterol (sim ou não); se manuseia ou manuseava substâncias químicas no trabalho que são potencialmente prejudiciais à saúde (sim ou não); se possui plano de saúde (sim ou não); cor/etnia (branca, preta, parda, etc.); nível de instrução (fundamental completo, médio completo, superior incompleto, etc.); se fuma algum produto do tabaco, e com qual frequência (não fuma, diariamente, menos que diariamente); se já recebeu diagnóstico de depressão (sim ou não).

Pesquisa Nacional de Saúde (PNS)

A base original foi obtida no site do IBGE, na seção da PNS, subseção de microdados.[4] A PNS, inquérito domiciliar de saúde, foi realizada em 2019, com amostra representativa da

população brasileira. Esta base original, contida no arquivo PNS_2019.txt, possui 346 Mb, um total de 279.382 linhas (indivíduos entrevistados) e 817 colunas (características).

Preparação da base de dados

No 1° processo de filtragem, feito através da biblioteca PNS-IBGE do R, foi possível extrair um *dataframe* da base original com as variáveis de interesse. Foi então extraído um arquivo csv de 279.382 linhas (indivíduos entrevistados) e 68 colunas (características). No 2º processo de filtragem, a base de dados foi reduzida para conter apenas os indivíduos do sexo masculino que responderam 2 perguntas que compõem a variável target: a) se o indivíduo recebeu diagnóstico de algum tipo de câncer na vida (1- sim, 2- não); b) se o indivíduo recebeu DCP ao longo da vida (1- sim, 2- não). Após o 2º processo de filtragem, foi obtido um *dataframe* de 42.799 linhas (indivíduos) e 58 colunas (características). A base de dados então foi devidamente preparada para ser usada nas modelagens, incluindo o tratamento de dados faltantes (*missings*) e a organização de variáveis categóricas, resultando numa base de dados final com 42.799 linhas (indivíduos) e 58 variáveis.

Análise dos dados

Com a base de dados devidamente preparada, foram realizados os procedimentos de filtragem das variáveis de maior relevância, análise estatística, e modelagem por métodos de *machine learning*. No caso deste trabalho, os modelos de classificação escolhidos para a descrição da variável dependente "DCP" foram: *regressão logística* e *árvore de decisão*. Tais procedimentos são detalhados a seguir.

Filtragem das variáveis pelo IV

Antes da aplicação da base de dados nos modelos de regressão logística e árvore de decisão, foi feita uma filtragem inicial das variáveis descritivas através do *IV* (*information value*), calculado através do RStudio.[5] Todas as variáveis com *IV* considerado fraquíssimo ($\leq$ 0,02) foram excluídas dos modelos previamente, restando 42 variáveis descritivas nesta etapa.

Regressão logística e árvore de decisão

Após filtragem pelo *IV*, foi aplicado o modelo de regressão logística com o uso do RStudio, considerando como variável *target* a ocorrência de câncer de próstata (variável binária, com **1** para “sim”, e **0** para “não”). Para isso, a base de dados foi dividida em base de treino (70 % dos indivíduos) e base de teste (30 %). O modelo de regressão logística permite prever se a variável *target* terá resultado positivo ou negativo para um certo indivíduo da base.

O modelo inicialmente tem os seus parâmetros determinados com o uso da base de treino, e então a sua eficiência é testada com o uso da base de teste.

Usando a base de treino, foi aplicado o método *backward* e um nível de significância de 0,10 (ou seja, p-valor $\leq 0{,}10$) para a seleção das variáveis. Após este processo, chegou-se a um modelo final com um total de 13 variáveis descritivas, todas categóricas, as quais são citadas antecipadamente na Subseção "*Variáveis"*.

O modelo de árvore de decisão foi executado no RStudio, com as mesmas 13 variáveis descritivas, considerando árvores de 2 e 3 níveis. A árvore de decisão também é utilizada como um modelo preditivo para a variável *target* "câncer de próstata", e além disso, permite identificar grupos específicos com maior frequência de casos positivos para a variável *target*, de acordo com o nível de associação com as variáveis descritivas.

RESULTADOS

Análise exploratória das variáveis aplicadas na modelagem

Dentre as variáveis descritivas selecionadas, a que apresentou o valor mais alto de *IV* foi a variável "idade" (*IV* = 2,86), o que representa uma capacidade preditiva muito forte em relação ao câncer de próstata. Verifica-se que a idade média dos homens entrevistados é de 45,9 anos, e a mediana é de 45 anos; isso mostra que o grupo de 50% dos indivíduos acima da mediana contém a faixa etária de maior propensão a receber DCP (Figura 1).

A **Tabela 1** mostra a distribuição dos indivíduos nas categorias das variáveis descritivas e o cruzamento com a variável *target* "câncer de próstata"; os resultados que podem ser contextualizados mais claramente são:

Idade: aumento da frequência de DCP a partir de 50 anos de idade, com frequência mais alta na categoria "≥ 80 anos";

Consultas médicas: maior frequência de DCP nos indivíduos que tiveram sua última consulta médica mais próxima do momento da pesquisa (até 2 anos antes);

Autoavaliação de saúde: maior frequência de DCP nos indivíduos que afirmaram ter uma qualidade de saúde "ruim ou muito ruim".

Diagnóstico de colesterol alto: maior frequência de DCP nos indivíduos que receberam diagnóstico de nível alto de colesterol;

Plano de saúde: maior frequência de DCP nos indivíduos que afirmam ter plano de saúde;

Diagnóstico de depressão: maior frequência de DCP nos indivíduos que receberam diagnóstico de depressão.

A seleção de variáveis descritivas como "plano de saúde" e "nível de instrução" indica uma influência da situação socioeconômica nos modelos preditivos. De fato, tal indicação pode ser verificada através do cruzamento da variável dependente DCP com a renda per capita familiar dos indivíduos. Considerando a faixa de idade a partir de 50 anos, observam-se taxas de DCP de 0,84 %, 1,82 %, 2,49 % e 3,07 % para as respectivas faixas de renda per capita familiar (em salários mínimos [sm]): até ½ sm; mais de ½ sm até 2 sm; mais de 2 sm até 5 sm; mais de 5 sm. Isto é, a faixa de renda mais alta (mais de 5 sm) tem uma taxa de DCP ≈ 3,7 × maior do que na faixa de renda mais baixa (até ½ sm).

Com a premissa descrita acima, o possível efeito socioeconômico sobre as variáveis descritivas foi investigado; para isso, foi determinado o nível de associação entre cada variável descritiva e a renda per capita familiar através das medidas V de Cramer e ω de Cohen. Verificou-se um nível alto de associação da renda per capita familiar com as variáveis "plano de saúde" (V de Cramer = 0,487 e ω de Cohen = 0,487) e "nível de instrução" (V de Cramer = 0,314 e ω de Cohen = 0,544). Um nível médio de associação com a renda per capita familiar foi observado na variável "cor/etnia" (V de Cramer = 0,150 e ω de Cohen = 0,260), e um nível razoável foi observado em "consumo de frutas" e "consumo de verduras e legumes" (ω de Cohen = 0,218 e 0,229, respectivamente). Portanto, a influência da renda familiar nas variáveis descritivas deve ser levada em conta na interpretação dos resultados da Tabela 1.

Apesar do nível importante de associação descrito acima, é conveniente salientar que a variável "renda per capita familiar" não foi incluída nos modelos preditivos finais, pois esta variável foi naturalmente excluída pelo método *backward.*

Resultados da regressão logística

Os resultados de acurácia, sensibilidade, especificidade e ROC-AUC para a regressão logística são mostrados na **Tabela 2,** referentes às bases de treino e teste. Observa-se que a acurácia, sensibilidade e especificidade para as bases de treino e teste são muito satisfatórios, pois todos estão acima de 80 %. Especificamente na base de teste, houve um pequeno aumento na acurácia e especificidade, e uma pequena diminuição na sensibilidade. O resultado obtido para o ROC-AUC na base de treino (0,822) pode ser classificado como excelente, enquanto que o resultado na base de teste (0,780) encontra-se muito próximo da mesma condição.

Resultados da árvore de decisão (2 e 3 níveis)

Os resultados obtidos para as árvores de decisão de 2 e 3 níveis são mostrados graficamente nas **Figuras 2 e 3**, respectivamente. Observando as frequências de casos de câncer de próstata nos nós finais da árvore de **2 níveis**, verifica-se:

- Os **nós 3 e 4** sugerem que há uma probabilidade maior de ocorrência de câncer de próstata em homens negros comparativamente ao conjunto das outras etnias, considerando a faixa etária abaixo de 50 anos;
- Os **nós 6 e 7** indicam que o grupo de homens com nível de colesterol alto tem uma maior frequência de DCP (≈ 3× maior) em comparação com o grupo de nível de colesterol normal ou desconhecido, considerando a faixa etária de 50 a 65 anos;
- Os **nós 9 e 10** indicam que o grupo de homens com nível de instrução entre o ensino fundamental completo e superior completo tem uma maior frequência de DCP (≈ 2× maior) em comparação com o grupo de menor nível de instrução, considerando a faixa etária a partir de 65 anos e abaixo de 80 anos;
- Os **nós 12 e 13** indicam que o grupo de homens que possuem plano de saúde tem uma maior frequência de DCP (≈ 2× maior) em comparação com o grupo que não possui plano de saúde, considerando a faixa etária a partir de 80 anos.

Com relação aos nós 9 e 10, deve-se levar em conta a forte associação entre nível de instrução e renda per capita familiar discutida anteriormente.[6]

A respeito dos nós 12 e 13, deve-se levar em conta as diferentes características dos sistemas público e privado de saúde no Brasil, além da forte associação entre a variável "plano de saúde" e a renda per capita familiar discutida anteriormente.[7]

Os principais resultados da árvore de decisão de **3 níveis** são descritos a seguir:

- Entre os homens com idade a partir de 50 e < 65 anos, e que possuem nível de colesterol normal ou desconhecido, observa-se que a frequência de DCP é ≈ 2,5× maior no subgrupo que possui plano de saúde em comparação com o subgrupo que não possui plano de saúde;
- Entre os homens com idade a partir de 50 e < 65 anos, e que possuem nível de colesterol alto, observa-se que o subgrupo de homens que toma sucos artificiais todos os dias tem uma frequência de DCP ≈ 4× maior do que o subgrupo de homens que diz não tomar ou tomar muito pouco sucos artificiais.
- Entre os homens com idade a partir de 65 e < 80 anos, e que possuem pelo menos o nível fundamental completo de escolaridade, observa-se que o subgrupo de homens com nível de colesterol alto tem uma frequência de DCP ≈ 2× maior do que o subgrupo de homens que têm nível de colesterol normal ou desconhecido.

- Entre os homens com idade a partir de 65 e < 80 anos, e que possuem nível de escolaridade abaixo do fundamental completo, observa-se que o subgrupo de homens que possui plano de saúde tem uma frequência de DCP ≈ 2,5× maior do que o subgrupo de homens que não possui plano de saúde.

A árvore de decisão de 3 níveis mostra resultados que reforçam aspectos discutidos sobre a árvore de 2 níveis, contudo, em grupos mais específicos. Os pares de nós {9,10}, {20,21} e {23,24} reforçam a maior probabilidade de DCP para os indivíduos que afirmam ter plano de saúde. O par de nós {17,18} reforça a maior probabilidade de DCP para os indivíduos com alto nível de colesterol, contudo, na faixa etária a partir de 65 e < 80 anos e que possuem pelo menos o nível fundamental completo de escolaridade. Uma informação relevante ocorre no trio de nós {12,13,14}, o qual indica uma probabilidade significativamente maior de DCP para os indivíduos que afirmam consumir sucos artificiais todos os dias, dentro da faixa etária a partir de 50 e < 65 anos e que possuem alto nível de colesterol.

DISCUSSÃO

A mediana de idade dos homens que responderam “sim” para DCP (72 anos) está bem acima da mediana dos homens que responderam “não” (44 anos), resultado que é compatível com a tendência geral na qual os homens desenvolvem o câncer de próstata a partir de aproximadamente 50 anos.[3]

Com relação aos nós 3 e 4 da árvore de 2 níveis, sugere-se que as pessoas negras têm uma maior tendência para o desenvolvimento de câncer de próstata em idades abaixo de 50 anos; tal constatação parece corroborar estudos anteriores sobre o câncer de próstata[3], onde verificou-se que esta doença é mais frequente em homens com ascendência africana e caribenha. Contudo, os casos de câncer de próstata abaixo de 50 anos são raros na base (2 casos entre 15.493 indivíduos no nó 3, e 3 casos entre 2.175 indivíduos no nó 4) e, por isso, sugere-se que conclusões a respeito desta faixa etária devam ser corroboradas com uma base de dados mais numerosa abaixo de 50 anos.

A respeito dos nós 6 e 7 da árvore de 2 níveis, deve-se considerar estudos como de Pelton *et al.*[8], o qual afirma que altos níveis de colesterol no sangue estão relacionados a casos de câncer de próstata mais agressivos, e de Jamnagerwalla *et al.*[9], o qual aponta que altos níveis de colesterol sérico total e HDL estão associados a um risco aumentado de câncer de próstata de alto grau. Uma comunicação do Johns Hopkins Medicine[10] descreve pesquisas mais recentes que apontam para conclusões semelhantes. Como o presente estudo é transversal, não se pode

afirmar com segurança que ele corrobora os resultados dos trabalhos anteriores, mas indica-se a importância de estudos adicionais sobre a relação colesterol/DCP com a inclusão de indivíduos do Brasil.

O fato do presente estudo ser transversal limita a possibilidade de conclusões sobre o efeito dos sucos artificiais na incidência de câncer de próstata; contudo, há estudos anteriores que sugerem a relação entre o alto consumo de bebidas açucaradas e uma maior incidência de câncer de próstata, como o de Miles *et al.*[11] e de Llaha *et al*.[12] Além disso, o estudo de Makarem *et al.*[13] sobre o consumo de alimentos açucarados sugere um aumento do risco de câncer de próstata nos homens que consomem sucos de frutas com maior frequência. Os resultados deste trabalho indicam a importância de estudos complementares sobre a influência do consumo de bebidas açucaradas e sucos artificiais na taxa de DCP no contexto brasileiro. Portanto, estudos longitudinais são necessários para analisar o possível efeito de fatores como o alto nível de colesterol e o consumo de sucos artificiais na taxa de ocorrência de câncer de próstata.

As árvores de decisão também foram usadas como modelos para a previsão de casos de câncer de próstata. A base de treino foi usada na parametrização do modelo, o qual foi aplicado em seguida na base de teste. Os resultados de acurácia, sensibilidade, especificidade e ROC-AUC para as árvores de decisão de 2 e 3 níveis são mostrados na **Tabela 2,** referentes às bases de treino e teste. O modelo de 2 níveis é muito satisfatório, tanto na base de treino quanto na base de teste. Há uma leve tendência para uma melhor reprodução dos eventos positivos de câncer de próstata (dado pela sensibilidade) em relação aos eventos negativos (dado pela especificidade). No caso da árvore de 3 níveis, houve um aumento considerável na sensibilidade em relação à árvore de 2 níveis, sendo um resultado excelente no aspecto da previsão dos casos positivos de câncer de próstata; contudo, a árvore de 3 níveis é um pouco menos eficiente na previsão dos casos negativos, o que se verifica pela leve diminuição da especificidade.

Os modelos apresentados podem ser usados para identificar homens com características físicas, socioeconômicas e de hábitos de vida que os tornam mais propensos a receber DCP. Porém, deve-se levar em conta que há uma diferença importante entre desenvolver uma doença e receber o diagnóstico da doença. Existem variáveis socioeconômicas como renda familiar, possuir ou não plano de saúde, nível de instrução, etc. que claramente influenciam a obtenção do diagnóstico precoce do câncer de próstata. Por isso, se os modelos aqui apresentados forem aplicados com o objetivo restrito de identificar homens com tendência a desenvolver o câncer de próstata, então deve-se atentar que o modelo terá naturalmente limitações, dependendo do grupo socioeconômico analisado.

Os resultados deste estudo podem ser investigados mais profundamente em pesquisas futuras, incluindo a influência do consumo de certos alimentos na ocorrência do câncer de próstata em estudos longitudinais, as diferenças estatísticas entre grupos sociais no diagnóstico da doença, a relação entre o câncer de próstata e outras doenças, entre outros aspectos.

Forças e Limitações

O cruzamento da variável target "câncer de próstata" com as variáveis descritivas nem sempre permite uma interpretação clara dos resultados, pois algumas variáveis podem sofrer uma influência significativa de fatores socioeconômicos, como é descrito na análise exploratória. Além disso, há algumas variáveis descritivas cujas perguntas correspondentes da PNS não foram respondidas pela totalidade dos indivíduos (usada a sigla *N.A.* no caso de perguntas não aplicadas).

Deve-se levar em conta que o questionário da PNS não foi preparado especificamente para a análise estatística do câncer de próstata. Uma pergunta importante não incluída no questionário é referente aos casos de câncer de próstata em membros da família. Pesquisas na área[3] mostram que ter um parente de primeiro grau com DCP aumenta significativamente o risco de um homem desenvolver a doença. Em geral, o questionário da PNS faz perguntas que não esclarecem como foram os hábitos de vida e estado de saúde da pessoa ao longo da sua vida, isto é, a maioria das perguntas faz somente um "retrato" do entrevistado no momento daquela pesquisa. Por isso, os modelos aqui apresentados podem ser aperfeiçoados futuramente se houver uma base de dados histórica da pessoa entrevistada e da sua família, com perguntas que se referem à evolução da sua saúde e hábitos de vida.

A análise exploratória desta base de dados e os modelos desenvolvidos podem ser usados para estimativas das demandas do sistema público ou privado de saúde, como recursos humanos e infraestrutura, para a prevenção e tratamento do câncer de próstata em grupos específicos ou regiões específicas do Brasil, bem como para identificar grupos de indivíduos que sejam alvos preferenciais de campanhas de prevenção ao câncer de próstata.

Os modelos indicam uma forte influência de fatores socioeconômicos no diagnóstico de câncer de próstata no Brasil. O alto nível de acurácia e sensibilidade dos modelos mostra o potencial dos métodos de *machine learning* para a prevenção do câncer de próstata no contexto brasileiro.

Esta pesquisa apresenta resultados úteis para o planejamento no uso de recursos públicos ou privados no tratamento ou prevenção do câncer de próstata no contexto brasileiro.

REFERÊNCIAS

1. Instituto Nacional de Câncer – INCA. Câncer de próstata. [consultado em 2023 Jul 14] Disponível em: https://www.gov.br/inca/pt-br/assuntos/cancer/tipos/prostata
2. Prostate Cancer Foundation. Prostate Cancer Survival Rates. [consultado em 2023 Jul 14] Disponível em: https://www.pcf.org/about-prostate-cancer/what-is-prostate-cancer/prostate-cancer-survival-rates/
3. Instituto Oncoguia. Fatores de Risco para Câncer de Próstata. [consultado em 2023 Jul 14] Disponível em: http://www.oncoguia.org.br/conteudo/fatores-de-risco-para-cancer-de-prostata/5850/1130/
4. IBGE - Instituto Brasileiro de Geografia e Estatística. PNS - Pesquisa Nacional de Saúde. Microdados. [consultado em 2023 Jul 14] Disponível em: https://www.ibge.gov.br/estatisticas/sociais/saude/9160-pesquisa-nacional-de-saude.html?=&t=microdados
5. Website do RStudio. [consultado em 2023 Jul 14] Disponível em: http://www.rstudio.com/ide
6. Victora C *et al*. Association between breastfeeding and intelligence, educational attainment, and income at 30 years of age: a prospective birth cohort study from Brazil. The Lancet Global Health 2015;3(4): E199-E205. doi:10.1016/S2214-109X(15)70002-1
7. Paim J *et al*. The Brazilian health system: history, advances, and challenges. The Lancet 2011;377(9779): 1778-97. doi:10.1016/S0140-6736(11)60054-8
8. Pelton K, Freeman MR, Solomon KR. Cholesterol and Prostate Cancer. Curr Opin Pharmacol. 2012;12(6):751-9. doi:10.1016/j.coph.2012.07.006
9. Jamnagerwalla J, Howard LE, Allott EH, Vidal AC, Moreira DM, Castro-Santamaria R, Andriole GL, Freeman MR, Freedland SJ. Serum cholesterol and risk of high-grade prostate cancer: results from the REDUCE study. Prostate Cancer Prostatic Dis. 2018;21(2): 252-59. doi:10.1038/s41391-017-0030-9
10. Johns Hopkins Medicine. Cholesterol, Prostate Cancer, and Race. [consultado em 2023 Jul 14] Disponível em: https://www.hopkinsmedicine.org/news/articles/cholesterol-prostate-cancer-and-race
11. Miles FL, Neuhouser ML, Zhang Z-F. Concentrated sugars and incidence of prostate cancer in a prospective cohort. Br J Nutr. 2018;120(6):703-10. doi: 10.1017/S0007114518001812

12. Llaha F, Gil-Lespinard M, Unal P, de Villasante I, Castañeda J, Zamora-Ros R. Consumption of Sweet Beverages and Cancer Risk. A Systematic Review and Meta-Analysis of Observational Studies. Nutrients 2021;13(2):516. doi: 10.3390/nu13020516
13. Makarem N, Bandera EV, Lin Y, Jacques PF, Hayes RB, Parekh N. Consumption of Sugars, Sugary Foods, and Sugary Beverages in Relation to Adiposity-Related Cancer Risk in the Framingham Offspring Cohort (1991–2013). Cancer Prev Res 2018;11(6):347-58. doi: 10.1158/1940-6207.CAPR-17-0218

CONTRIBUIÇÃO DOS AUTORES

Souza MA contribuiu com a extração da base de dados da PNS, preparação da base de dados, preparação e execução dos modelos de *machine learning*, análise estatística dos dados, delineamento do estudo, interpretação dos resultados, redação e revisão crítica do manuscrito. Monteiro CN contribuiu com a interpretação dos resultados, redação e revisão crítica do manuscrito. Barros CRS contribuiu com a interpretação dos resultados, redação e revisão crítica do manuscrito. Os autores aprovaram a versão final do manuscrito e são responsáveis por todos os seus aspectos, incluindo a garantia de sua precisão e integridade.

CONFLITOS DE INTERESSE

Os autores declararam não possuir conflitos de interesse.

TRABALHO ACADÊMICO ASSOCIADO

Artigo derivado do trabalho de conclusão de curso intitulado “Influência de características físicas, hábitos de vida e fatores socioeconômicos na ocorrência e diagnóstico de câncer de próstata: um estudo baseado nos dados da PNS 2019”, defendido por Marco Antonio de Souza na Pós-Graduação "Lato Sensu" Especialização em Análise de Dados, Data Mining e Inteligência Artificial, da Faculdade FIA de Administração e Negócios, em 2022.

AGRADECIMENTO

Não se aplica.

TABELAS, QUADROS E FIGURAS

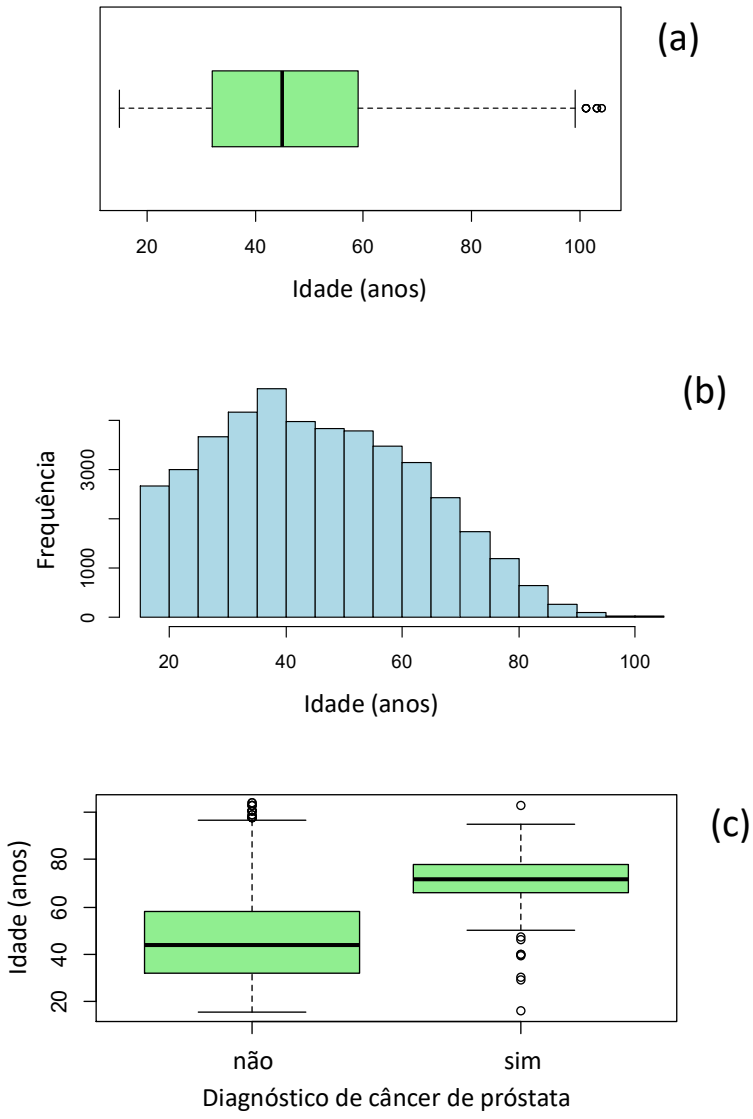

Figura 1. Descrição da variável "idade" dos indivíduos do sexo masculino selecionados na base da PNS. **(a)** Boxplot mostrando a distribuição de idade. **(b)** Histograma mostrando a distribuição de idade. **(c)** Comparação em boxplot das distribuições de idade nos grupos: com diagnóstico de câncer de próstata em algum momento da vida (sim); sem diagnóstico de câncer de próstata (não).

Tabela 1. Variáveis descritivas usadas nos modelos finais de regressão logística e árvore de decisão. São mostradas as categorias de cada variável, a frequência de indivíduos em cada categoria, e a frequência de casos positivos (sim) e negativos (não) de câncer de próstata em cada categoria. *N.A.*: pergunta não aplicada.

VARIÁVEL DESCRITIVA			**Câncer de próstata (%)**	
Idade	**Faixa de idade (anos)**	**Frequência (%)**	**Não**	**Sim**
	< 35	29,5	99,98	0,02
	≥ 35 e < 50	29,5	99,96	0,04
	≥ 50 e < 65	24,8	99,35	0,65
	≥ 65 e < 80	13,5	96,63	3,37
	≥ 80	2,8	94,48	5,52
Consultas médicas	**Tempo desde a última consulta médica (anos)**	**Frequência (%)**	**Não**	**Sim**
	até 2 anos	83,0	99,05	0,95
	mais de 2 anos	15,8	99,96	0,04
	nunca foi	1,2	100,00	0,00
Autoavaliação de saúde	**Autoavaliação**	**Frequência (%)**	**Não**	**Sim**
	muito boa ou boa	66,8	99,5	0,5
	regular	27,8	98,6	1,4
	ruim ou muito ruim	5,4	98,4	1,6
Consumo de frutas	**Consumo semanal**	**Frequência (%)**	**Não**	**Sim**
	1 a 3 dias	40,5	99,5	0,5
	4 a 6 dias	20,6	99,3	0,7
	nunca ou muito pouco	13,1	99,5	0,5
	todos os dias	25,8	98,5	1,5
Consumo de suco artificial [a]	**Consumo semanal**	**Frequência (%)**	**Não**	**Sim**
	1 a 3 dias	21,4	99,7	0,3
	4 a 6 dias	7,9	99,6	0,4
	nunca ou muito pouco	64,1	99,0	1,0
	todos os dias	6,7	99,2	0,8
Diagnóstico de colesterol alto	**Resposta**	**Frequência (%)**	**Não**	**Sim**
	não ou *N.A.*	89,1	99,4	0,6
	sim	10,9	98,0	2,0
Exposição a produtos químicos no trabalho [b]	**Resposta**	**Frequência (%)**	**Não**	**Sim**
	não ou *N.A.*	87,2	99,1	0,9
	sim	12,8	99,8	0,2

Tabela 1. (continuação)

VARIÁVEL DESCRITIVA			Câncer de próstata – frequência na categoria (%)	
Possui plano de saúde	**Resposta**	**Frequência (%)**	**Não**	**Sim**
	não	78,8	99,4	0,6
	sim	21,2	98,6	1,4
Cor/etnia	**Cor/etnia**	**Frequência (%)**	**Não**	**Sim**
	amarela	0,73	98,72	1,28
	branca	36,14	98,89	1,11
	ignorado	0,02	100,00	0,00
	indígena	0,78	100,00	0,00
	parda	50,51	99,45	0,55
	preta	11,82	99,11	0,89
Nível de instrução	**Nível de instrução**	**Frequência (%)**	**Não**	**Sim**
	sem instr. ou fund. inc.	42,7	99,0	1,0
	fund. compl. ou méd. inc.	15,5	99,4	0,6
	méd. compl. ou sup. inc.	28,9	99,4	0,6
	sup. compl.	12,9	99,1	0,9
Consumo de verduras e legumes	**Consumo semanal**	**Frequência (%)**	**Não**	**Sim**
	1 a 3 dias	35,7	99,3	0,7
	4 a 6 dias	21,2	99,2	0,8
	nunca ou muito pouco	9,5	99,5	0,5
	todos os dias	33,5	99,0	1,0
Fuma atualmente	**Uso semanal**	**Frequência (%)**	**Não**	**Sim**
	diariamente	14,3	99,4	0,6
	menos que diário	1,8	99,7	0,3
	não fuma	83,9	99,2	0,8
Teve diagnóstico de depressão	**Resposta**	**Frequência (%)**	**Não**	**Sim**
	não	95,4	99,2	0,8
	sim	4,6	98,5	1,5

[a] De acordo com a PNS, esta variável refere-se ao consumo dos chamados sucos de "caixinha", em lata ou refresco em pó.

[b] De acordo com a PNS, esta variável refere-se ao manuseio de produtos químicos como: agrotóxicos, gasolina, diesel, formol, chumbo, mercúrio, cromo, quimioterápicos, etc.

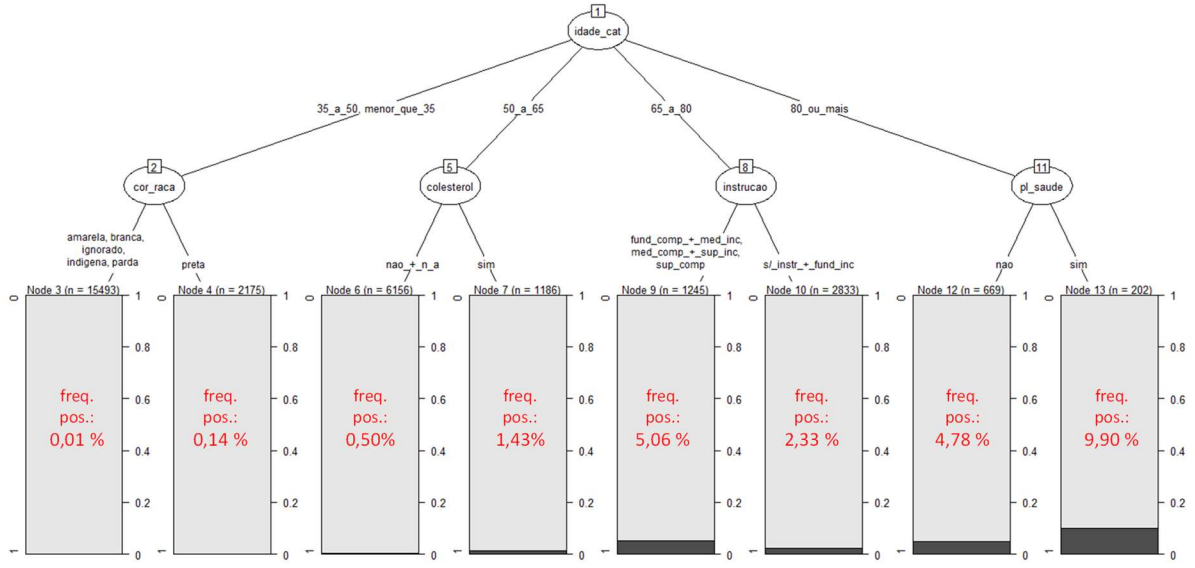

Figura 2. Árvore de decisão de 2 níveis, considerando a base de treino com 70% dos indivíduos selecionados. As porcentagens em vermelho indicam a frequência de casos positivos de câncer de próstata nos nós finais da árvore. Descrição das variáveis: *idade_cat*: idade na forma categórica (anos); *cor_raca*: cor/etnia; *colesterol*: nível de colesterol alto; *instrucao*: nível de instrução; *pl_saude*: plano de saúde. Abreviações mencionadas: *n_a*: pergunta não aplicada; *fund_comp*: nível fundamental completo; *med_inc*: nível médio incompleto; *med_comp*: nível médio completo; *sup_inc*: nível superior incompleto; *sup_comp*: nível superior completo; *s/_instr*: sem instrução; *fund_inc*: nível fundamental incompleto.

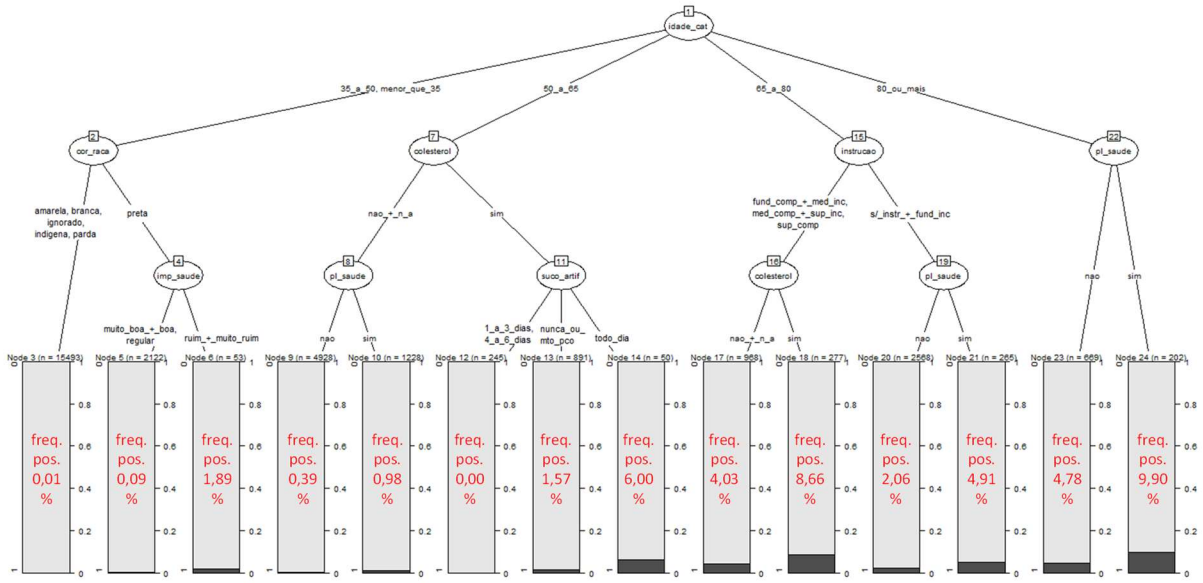

Figura 3. Árvore de decisão de 3 níveis, considerando a base de treino com 70% dos indivíduos selecionados. As porcentagens em vermelho indicam a frequência de casos positivos de câncer de próstata nos nós finais da árvore. Descrição das variáveis: *idade_cat*: idade na forma categórica (anos); *cor_raca*: cor/etnia; *colesterol*: nível de colesterol alto; *instrucao*: nível de instrução; *pl_saude*: plano de saúde; *imp_saude*: impressão sobre a própria saúde; *suco_artif*: consumo semanal de suco artificial. Abreviações mencionadas: *n_a*: pergunta não aplicada; *fund_comp*: nível fundamental completo; *med_inc*: nível médio incompleto; *med_comp*: nível médio completo; *sup_inc*: nível superior incompleto; *sup_comp*: nível superior completo; *s/_instr*: sem instrução; *fund_inc*: nível fundamental incompleto; *mto_pco*: muito pouco.

Tabela 2. Acurácia, sensibilidade, especificidade e ROC-AUC obtidos nos modelos de regressão logística, árvore de decisão de 2 níveis e árvore de decisão de 3 níveis, considerando as bases de treino (70% do total) e teste (30% do total).

Modelo	Base	Acurácia	Sensibilidade	Especific.	ROC-AUC
Regressão logística	treino	0,828	0,828	0,828	0,822
	teste	0,835	0,802	0,835	0,780
Árvore de decisão (2 níveis)	treino	0,801	0,850	0,800	0,823
	teste	0,805	0,821	0,804	0,812
Árvore de decisão (3 níveis)	treino	0,767	0,906	0,766	0,832
	teste	0,773	0,887	0,772	0,813

Este preprint foi submetido sob as seguintes condições:

- Os autores declaram que estão cientes que são os únicos responsáveis pelo conteúdo do preprint e que o depósito no SciELO Preprints não significa nenhum compromisso de parte do SciELO, exceto sua preservação e disseminação.
- Os autores declaram que os necessários Termos de Consentimento Livre e Esclarecido de participantes ou pacientes na pesquisa foram obtidos e estão descritos no manuscrito, quando aplicável.
- Os autores declaram que a elaboração do manuscrito seguiu as normas éticas de comunicação científica.
- Os autores declaram que os dados, aplicativos e outros conteúdos subjacentes ao manuscrito estão referenciados.
- O manuscrito depositado está no formato PDF.
- Os autores declaram que a pesquisa que deu origem ao manuscrito seguiu as boas práticas éticas e que as necessárias aprovações de comitês de ética de pesquisa, quando aplicável, estão descritas no manuscrito.
- Os autores declaram que uma vez que um manuscrito é postado no servidor SciELO Preprints, o mesmo só poderá ser retirado mediante pedido à Secretaria Editorial do SciELO Preprints, que afixará um aviso de retratação no seu lugar.
- Os autores concordam que o manuscrito aprovado será disponibilizado sob licença Creative Commons CC-BY.
- O autor submissor declara que as contribuições de todos os autores e declaração de conflito de interesses estão incluídas de maneira explícita e em seções específicas do manuscrito.
- Os autores declaram que o manuscrito não foi depositado e/ou disponibilizado previamente em outro servidor de preprints ou publicado em um periódico.
- Caso o manuscrito esteja em processo de avaliação ou sendo preparado para publicação mas ainda não publicado por um periódico, os autores declaram que receberam autorização do periódico para realizar este depósito.
- O autor submissor declara que todos os autores do manuscrito concordam com a submissão ao SciELO Preprints.